

Real-time Activity-Aware Video Streaming While Preserving Privacy

Kaiya Shimura

kaiya@ucl.nuee.nagoya-u.ac.jp
Graduate School of Engineering,
Nagoya University
Nagoya, Japan

Kazuma Kano

Graduate School of Engineering,
Nagoya University
Nagoya, Japan

Tahera Hossain

Graduate School of Engineering,
Nagoya University
Nagoya, Japan

Shin Katayama

Graduate School of Engineering,
Nagoya University
Nagoya, Japan

Kenta Urano

Graduate School of Engineering,
Nagoya University
Nagoya, Japan

Shun Taguchi

Collaborative Intelligence Research
Domain, Toyota Central R&D
Laboratories
Tokyo, Japan

Hideki Deguchi

Collaborative Intelligence Research
Domain, Toyota Central R&D
Laboratories
Tokyo, Japan

Hiroyuki Sakai

Collaborative Intelligence Research
Domain, Toyota Central R&D
Laboratories
Tokyo, Japan

Takuro Yonezawa

Graduate School of Engineering,
Nagoya University
Nagoya, Japan

Nobuo Kawaguchi

Graduate School of Engineering,
Nagoya University
Nagoya, Japan

Abstract

We present ActStream, a remote communication system that selectively shares the user and objects relevant to their current activities. By combining vision-language models and real-time segmentation, ActStream detects and highlights only activity-relevant elements, enhancing awareness while preserving privacy. The system identifies key objects through user focus or interaction, synthesizes them with the user's image, and streams the result to a remote partner. ActStream aims to provide useful context without exposing unnecessary information, highlighting the potential of selective sharing in privacy-sensitive remote collaboration.

CCS Concepts

• **Human-centered computing** → **Collaborative and social computing systems and tools.**

Keywords

human-computer interaction, remote communication, video-mediated communication, object sharing, privacy-aware systems, activity recognition

ACM Reference Format:

Kaiya Shimura, Kazuma Kano, Tahera Hossain, Shin Katayama, Kenta Urano, Shun Taguchi, Hideki Deguchi, Hiroyuki Sakai, Takuro Yonezawa, and Nobuo Kawaguchi. 2018. Real-time Activity-Aware Video Streaming While Preserving Privacy. In *Proceedings of Adjunct Proceedings of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2025 ACM International Symposium on Wearable Computers (UbiComp/ISWC '25 Adjunct)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

The advancement of communication technologies and changing societal needs have led to the widespread adoption of online meeting systems and video calling applications in daily life [1]. These technologies enable people to interact across spatial and temporal boundaries, fostering new forms of personal and social connection. Informal digital events, such as “online drinking parties” via Zoom [2] and virtual gatherings on VR platforms [20], highlight how individuals creatively adapt to remote social interaction. Beyond conventional tools, experimental systems using large shared displays have also been proposed to support communication between physically separated spaces.

At the same time, live streaming platforms such as YouTube and Twitch have become integral to everyday social and creative

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 components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

UbiComp/ISWC '25 Adjunct, Espoo, Finland

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06

<https://doi.org/XXXXXXX.XXXXXXX>

expression [5, 7, 15]. These platforms allow individuals to broadcast personal activities in real time, contributing to a growing culture of casual, self-directed communication.

However, the increasing use of video-based remote communication and personal livestreaming raises important privacy concerns [22, 25, 31]. Users may unintentionally reveal private aspects of their environment or expose others without consent. Although many tools provide background blurring or virtual backgrounds, such methods often obscure not only the environment but also objects relevant to the user's current activity. This can hinder mutual understanding during interaction. For instance, when an object being held, viewed, or manipulated is hidden, it becomes difficult for the viewer to understand the user's focus or intention.

To address this, prior HCI research has explored object-sharing techniques in video communication, including manual selection, sensor-based detection, and predefined item lists [10, 21, 24]. However, these approaches often present practical limitations: manual control disrupts natural interaction [25], sensors may fail in cluttered or dynamic environments, and static lists lack flexibility to capture the diversity of everyday activities. Consequently, many existing systems focus only on sharing the user's body, which helps protect privacy but often fails to provide the contextual cues necessary for rich communication. For example, people frequently interpret and mimic others' actions—such as drinking water—based on their interaction with visible objects.

To fill these gaps, we propose *ActStream*, a video-sharing method that selectively reveals objects associated with the user's current activity while keeping unrelated parts of the scene hidden. The system automatically detects user behavior from video input and segments objects that are visually or behaviorally relevant. For example, when a user is looking at a monitor, *ActStream* shares both the user and the monitor, allowing the communication partner to understand the user's focus without exposing the full environment.

2 RELATED WORK

2.1 Sensor-Based and Proximity-Driven Object Sharing

Several systems support remote collaboration by detecting physical objects in the user's environment, often using spatial sensing, RGB-D cameras, or projection. These approaches typically prioritize near-field objects based on visibility or proximity [3, 12]. Projection-based systems [17] enhance co-presence by projecting gestures, tools, and documents into remote workspaces, but are constrained by static hardware and limited projection areas. RGB-D systems [8, 9] leverage depth data to identify relevant objects and activities, integrating cues such as voice, gesture, and proximity. For instance, FocalSpace [8] uses multimodal inputs to reduce distractions and protect privacy in mobile contexts. However, these systems often operate within fixed spatial bounds and lack adaptability when user focus shifts beyond the tracking zone. Spatial augmented reality systems [30] combine projection and tracking to support co-located collaboration, but assume stable environments and do not accommodate interactions with untracked objects. While focusing on proximate content simplifies design, it limits flexibility in dynamic tasks. Future systems should aim to identify contextually relevant objects beyond mere proximity.

2.2 Manual Object Selection Interfaces

Other systems enable users to manually select or highlight objects during remote interactions, often using live video. These interfaces are common in collaboration, education, and live streaming contexts. A core challenge is minimizing user effort while maintaining communication clarity. ThingShare [14] allows users to drag and drop objects from the video feed, reducing the need to reposition physical items and providing stable visual references. Mixed reality systems explore ways to reduce manual input through automatic guidance. Johnson et al. [18] highlight objects in the user's view to assist remote collaborators, reducing reliance on precise pointing or verbal instructions. Feick et al. [11] use gesture-based interactions to manipulate virtual representations of objects, supporting shared understanding. However, manual interfaces may increase operational overhead. Ludwig et al. [19] found that users often avoid complex annotation tools, preferring verbal descriptions once a shared view is established. These findings suggest that ease of use and low interaction cost are essential for adoption in practice.

2.3 Static Object Sharing Techniques

Some systems rely on predefined content to support collaboration [4, 6, 16]. For instance, several approaches use pre-scanned models of physical objects to build shared virtual environments. Barroso et al. [6] provide pre-built object libraries for mixed reality collaboration. While these offer high-quality representations, they require advance preparation and cannot support spontaneous references to unscanned items. Standard video conferencing tools often rely on fixed-window sharing (e.g., slides or documents). While reliable, this limits spontaneous interaction with physical objects. Studies by Standaert et al. [27] and Hu et al. [14] report that such constraints often lead to awkward workarounds. Other systems use predefined labels or markers to identify objects. For example, Villanueva et al. [29] demonstrate AR-based labs with pre-tagged components. However, unlabeled objects cannot be integrated, and maintaining these tags adds overhead. Overall, while static object sharing techniques are effective in controlled settings with stable content, they lack the flexibility required for dynamic and long-term interactions, where spontaneous and adaptive collaboration is often essential.

3 CONTEXTUAL REMOTE COMMUNICATION

This section revisits conventional remote communication models and highlights their limitations in conveying contextual elements in users' environments. We then introduce alternative models that extend beyond body-centric interaction, incorporating activity-relevant surroundings. Finally, we present practical scenarios that benefit from such context-bearing communication, particularly in long-term, low-interruption settings.

3.1 Limitations of Conventional Models

Traditionally, remote communication systems have been categorized into two primary models: *Meet-type* and *Visit-type*, as shown in Figure 1. While these models support diverse forms of mediated interaction, they often overlook the user's surrounding environment, limiting their expressive capacity.

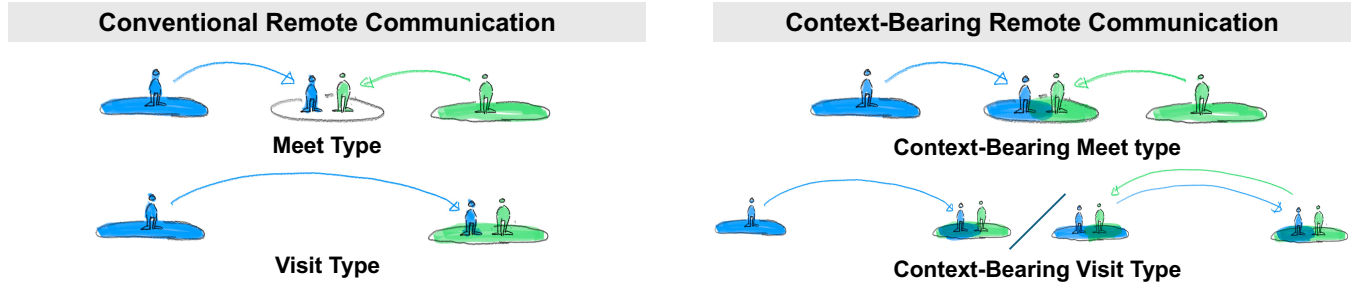


Figure 1: Comparison of conventional and context-bearing remote communication models.

Meet-type models (e.g., online meetings, metaverse platforms) connect participants in a shared digital space through audio and video streams. These systems primarily transmit facial and bodily cues through video or avatars while omitting contextual elements such as nearby tools or documents.

Visit-type models (e.g., telepresence robots) allow one user to virtually “enter” another’s space, sometimes with control over navigation and viewpoint. However, they typically require manual operation and do not automatically highlight semantically relevant aspects of the environment.

A common limitation across these models is the absence of contextual awareness. In face-to-face interactions, people infer intent and attention through shared visual access to surrounding objects [13, 28]. For example, cockpit communication studies have shown that pilots spend up to 60% of their time monitoring their partner’s control panel [26], underscoring the importance of shared environmental reference points for coordination. This indicates that the ability to selectively share tools, screens, or other relevant objects is not a secondary enhancement but rather a fundamental mechanism for nonverbal understanding. Conventional systems, by focusing solely on user’s body, fail to support this level of implicit communication.

3.2 Context-Bearing Communication Models

Conventional remote communication technologies have primarily focused on transmitting users’ bodily presence, often excluding environmental elements such as objects or tools associated with ongoing activities. However, in practice, user behavior and attention are closely tied to surrounding objects. Sharing such contextual information can lead to more accurate and effective communication.

In this section, we introduce a set of remote communication models that extend beyond bodily representation to include objects relevant to users’ current activities.

3.2.1 Context-Bearing Meet Type. This model extends traditional Meet-type interactions by incorporating not only the user’s body but also activity-relevant objects such as documents, tools, or digital devices into the shared video feed. By visually representing a user’s immediate environment, this approach enables communication with higher contextual resolution, facilitating a better understanding of user focus and actions.

3.2.2 One-Way Visit Type. In this asymmetric model, one user shares both their body and surrounding environment, while the

other passively observes. This configuration is well suited for scenarios involving remote guidance or support, where unidirectional information transfer is sufficient.

3.2.3 Two-Way Visit Type. This model enables both users to simultaneously share their bodies and contextual environments. While spatial alignment between differing physical spaces presents a challenge, the mutual visibility of each user’s activities and interaction targets can support visual synchrony and enhance mutual understanding.

3.3 Use Cases: Lightweight and Persistent Connections

Although the proposed models can be applied to task-oriented use cases such as remote collaboration and work support, their unique value emerges in communication contexts that do not rely on explicit conversation or coordinated tasks.

Examples include “study-with-me” livestreams, always-on connections via large shared displays between remote locations, casual interactions in virtual environments, or integrated streaming features such as the simultaneous sharing of facial video and game screen in consumer devices like the Nintendo Switch 2. In these scenarios, simply sharing one’s presence and activity state provides social value.

To support such forms of connection, remote communication systems should satisfy the following technical requirements:

- Avoid disrupting or requiring changes to users’ natural behavior.
- Dynamically and intelligently select shared content based on user activity or attention.
- Eliminate the need to predefine or manually select shared elements.

3.4 Application Scenarios

The proposed models offer a foundation for a variety of context-sensitive applications that go beyond traditional video or screen sharing.

Livestreaming Contextualization. In platforms such as YouTube or Twitch, automatically identifying and displaying only the objects relevant to a creator’s current activity—such as ingredients, tools,

or devices—can enhance audience understanding while preserving privacy. This contrasts with background blurring, which often removes crucial visual information along with sensitive content.

Casual Remote Hangouts. Context-bearing sharing can enhance informal video calls—like virtual hangouts or online drinking parties—by showing only relevant items, such as your drink or a snack, while keeping private or cluttered background areas hidden. This allows people to share parts of their environment that add to the experience, without exposing more than they intend. It offers a more relaxed and socially comfortable way to stay connected.

Mixed Reality Collaboration. In MR settings, combining bodily presence with selective object sharing provides high-context visual feedback while minimizing cognitive load. This enables more effective remote support, co-creation, and tutoring compared to avatar-based or whole-scene streaming approaches.

In summary, by enabling selective and adaptive sharing of contextual elements, the proposed models expand the design space of remote communication. They support more natural, grounded, and privacy-conscious interactions, particularly in informal or ambient connection scenarios.

4 DESIGN AND IMPLEMENTATION

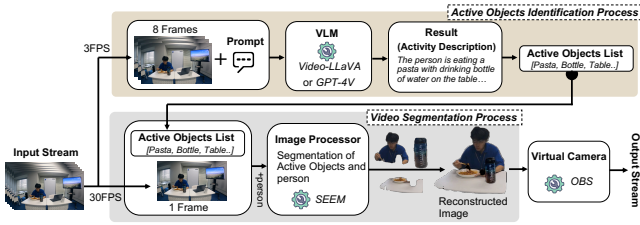


Figure 2: Overview of the ActStream

An overview of the ActStream system is shown in Figure 2. The system captures the activities of a remote individual—referred to as the user—and enhances situational awareness for remote observers while preserving privacy. It achieves this by identifying and sharing only the physical objects that are contextually relevant to the user’s current activity—referred to as active objects. ActStream operates through two parallel processes: (1) Active Object Identification and (2) Video Segmentation. After capturing a video stream from a camera, the system analyzes user activity by detecting and segmenting only the objects relevant to that activity—referred to as active objects. The identified objects are then used to filter the scene, preserving only the user and contextually important elements. This filtered output is streamed to downstream applications, aiming to enhance remote situational awareness while preserving privacy.

4.1 Active Objects Identification Process

To detect objects related to the user’s activity, we use a vision–language model (VLM) that integrates visual and textual understanding. Unlike many UbiComp approaches that rely on wearable sensors or pose estimation, our method requires no additional devices, reducing user burden. We tested two VLMs: Video-LLaVA (local) and GPT-4V (API-based). Initially, we explored various prompts, such as

asking the model to describe the user’s activity or select from predefined options. However, these yielded inconsistent results. Through iterative testing, we found that the prompt “What is the person looking at?” produced more reliable descriptions of relevant objects. The model processes 8 frames sampled at 3 FPS (2.66 seconds). It returns a natural language description, which we parse into an active objects list. Due to VLM latency, processing is asynchronous: a new task starts only after the previous one finishes, and the most recent result is reused between updates. The object list is sent to the segmentation module via UDP. This component operates on a high-end desktop with an Intel i9 CPU, 192GB RAM, and an RTX A6000 GPU.

4.2 Video Segmentation Process

This process takes the latest video frame and the corresponding list of active objects to produce a filtered image showing only the user and relevant items. We use SEEM [32], a multimodal segmentation model that generates pixel-level masks from text prompts. Each object label is input as a separate prompt, and SEEM returns binary masks for each, which are merged into a single composite mask. At the same time, panoptic segmentation extracts the user’s figure. The output is an RGBA image where the user and active objects are fully opaque and the background is transparent. This image is encoded as a PNG and streamed through a virtual camera using OBS Studio [23]. We also provide a WebSocket interface that transmits base64-encoded pixel data asynchronously, allowing segmentation, encoding, and transmission to run in parallel without blocking. This component operates on a machine with an AMD Ryzen 9 CPU, 96GB RAM, and an NVIDIA RTX 4090 GPU.

5 CONCLUSION

In this paper, we explored how conventional remote communication models fall short in conveying contextual elements of user environments, which are important for nonverbal understanding and situational awareness. Motivated by these limitations, we introduced a context-bearing communication approach that selectively shares users and activity-relevant objects.

To put this approach into practice, we developed ActStream, a system that automatically identifies and visualizes semantically important elements based on user behavior and attention. By combining a vision–language model with real-time segmentation, ActStream generates a focused view that highlights relevant content while reducing unnecessary exposure.

This design aims to support lightweight, persistent, and socially comfortable connections by representing not only the user’s body, but also the surrounding context of their ongoing activities. Through this selective and adaptive sharing, ActStream aims to support remote communication that balances situational awareness and privacy.

ACKNOWLEDGMENTS

This research was partially supported by JST CREST (Grant No. JP-MJCR22M4), JST RISTEX (Grant No. JPMJRS23K), and the NEDO SIP program “Development of foundational technologies and rules for expansion of the virtual economy” (Grant No. JPJ012495).

References

- [1] 2020. Microsoft Teams. <https://www.microsoft.com/en/microsoft-teams/group-chat-software> Accessed: 2025-04-11.
- [2] 2020. Zoom. <https://www.zoom.com/en> Accessed: 2025-04-11.
- [3] Rawan Alharbi, Mariam Tolba, Lucia C. Petito, Josiah Hester, and Nabil Alshurafa. 2019. To Mask or Not to Mask? Balancing Privacy with Visual Confirmation Utility in Activity-Oriented Wearable Cameras. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 72 (Sept. 2019), 29 pages. doi:10.1145/3351230
- [4] Tatsuya Amano, Teruhiro Mizumoto, Srikant Manas Kala, Hirozumi Yamaguchi, Tomokazu Matsui, and Keiichi Yasumoto. 2024. Visual Privacy Control for Metaverse and the Beyond. *IEEE Pervasive Computing* 23, 01 (Jan. 2024), 10–17. doi:10.1109/MPRV.2024.3365989
- [5] Backlinko. 2025. Twitch Users: Growth and Engagement Statistics (2025). <https://backlinko.com/twitch-users> Accessed: 2025-04-11.
- [6] Joao Barroso et al. 2020. Mixed Reality platform for remote collaboration. *Journal of Collaborative Computing* (2020).
- [7] Business of Apps. 2024. YouTube Revenue and Usage Statistics (2025). <https://www.businessofapps.com/data/youtube-statistics/> Accessed: 2025-04-21.
- [8] Zhihao Chen, David A. Shamma, and Walter S. Lasecki. 2020. FocalSpace: Sharing Near-Field Objects by Depth-Aware Video Cropping. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. doi:10.1145/3313831.3376498
- [9] Hang Chi, Yu Wang, et al. 2021. ARTEMIS: Mixed-Reality Environment for Immersive Surgical Telementoring. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3334480.3383169
- [10] Martin Feick, Terrance Mok, Anthony Tang, Lora Oehlberg, and Ehud Sharlin. 2018. Perspective on and re-orientation of physical proxies in object-focused remote collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [11] Matthew Feick, Anthony Tang, and Scott Bateman. 2018. Mixed-Reality for Object-Focused Remote Collaboration. In *Adjunct Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*.
- [12] Glenn J. Fernandes, Jiayi Zheng, Mahdi Pedram, Christopher Romano, Farzad Shahabi, Blaine Rothrock, Thomas Cohen, Helen Zhu, Tanmeet S. Butani, Josiah Hester, Aggelos K. Katsaggelos, and Nabil Alshurafa. 2024. HabitSense: A Privacy-Aware, AI-Enhanced Multimodal Wearable Platform for mHealth Applications. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 101 (Sept. 2024), 48 pages. doi:10.1145/3678591
- [13] Carl Gutwin and Saul Greenberg. 2002. A Descriptive Framework of Workspace Awareness for Real-Time Groupware. *Computer Supported Cooperative Work (CSCW)* 11 (09 2002), 411–. doi:10.1023/A:1021271517844
- [14] Xiao Hu, Jens Emil Grønbaek, et al. 2023. ThingShare: Ad-Hoc Digital Copies of Physical Objects for Sharing Things in Video Meetings. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*.
- [15] Global Media Insight. 2025. YouTube Statistics 2025: Users and Usage Trends. <https://www.globalmediainsight.com/blog/youtube-users-statistics/> Accessed: 2025-04-11.
- [16] Andrew Irlitti, Mesut Latifoglu, Thuong Hoang, Brandon Victor Syiem, and Frank Vetere. 2024. Volumetric Hybrid Workspaces: Interactions with Objects in Remote and Co-located Telepresence. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. doi:10.1145/3613904.3642814
- [17] Daisuke Iwai, Kosuke Sato, et al. 2018. Geometrically Consistent Projection-Based Tabletop Sharing for Remote Collaboration. *IEEE Access* 6 (2018), 11242–11253. doi:10.1109/ACCESS.2017.2781699
- [18] Steven Johnson, Danilo Gasques, et al. 2021. Do You Really Need to Know Where “That” Is? Enhancing Support for Referencing in Collaborative Mixed Reality Environments. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*.
- [19] Thomas Ludwig, Oliver Stickel, et al. 2021. shAR-IT: Ad hoc Remote Troubleshooting through Augmented Reality. *Computer Supported Cooperative Work (CSCW)* 30, 3–4 (2021), 341–375.
- [20] Divine Maloney and Guo Freeman. 2020. Falling Asleep Together: What Makes Activities in Social Virtual Reality Meaningful to Users. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play (Virtual Event, Canada) (CHI PLAY '20)*. Association for Computing Machinery, New York, NY, USA, 510–521. doi:10.1145/3410404.3414266
- [21] James Norris, Holger Schnädelbach, and Guoping Qiu. 2012. CamBlend: an object focused collaboration tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 627–636.
- [22] Joseph O'Hagan, Pejman Saeghe, Jan Gugenheimer, Daniel Medeiros, Karola Marky, Mohamed Khamis, and Mark McGill. 2023. Privacy-Enhancing Technology and Everyday Augmented Reality: Understanding Bystanders' Varying Needs for Awareness and Consent. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 177 (Jan. 2023), 35 pages. doi:10.1145/3569501
- [23] Open Broadcaster Software. 2024. OBS Studio – Open Broadcaster Software. <https://obsproject.com/> Version 30.0 or later. Accessed April 30, 2025.
- [24] Mehul S. Raval, Khai N. Truong, and David Dearman. 2014. MarkIt: Privacy Markers for Protecting Visual Secrets. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 1289–1295. doi:10.1145/2638728.2641707
- [25] Nisarg Raval, Animesh Srivastava, Kiron Lebeck, Landon Cox, and Ashwin Machanavajjhala. 2014. MarkIt: privacy markers for protecting visual secrets. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (Seattle, Washington) (*UbiComp '14 Adjunct*). Association for Computing Machinery, New York, NY, USA, 1289–1295. doi:10.1145/2638728.2641707
- [26] Leon D Segal. 1994. *Effects of checklist interface on non-verbal crew communications*. Technical Report.
- [27] Wim Standaert et al. 2021. Virtual Work Meetings During the COVID-19 Pandemic: The Good, Bad, and Ugly. *International Journal of Information Management* (2021).
- [28] Anthony Tang, Carman Neustaedter, and Saul Greenberg. 2007. VideoArms: Embodiments for Mixed Presence Groupware. In *People and Computers XX – Engage*, Nick Bryan-Kinns, Ann Blanford, Paul Curzon, and Laurence Nigay (Eds.). Springer London, London, 85–102.
- [29] Alejandro Villanueva et al. 2022. Tangible Remote Laboratories: A Toolkit for AR Marker Tracking. *Journal of Educational Technology* (2022).
- [30] Yuhang Wang, Jiahui Luo, et al. 2023. BeHere: A VR/SAR Remote Collaboration System Based on Virtual Replicas Sharing Gesture and Avatar in a Procedural Task. *Virtual Reality* 27, 1 (2023), 55–76. doi:10.1007/s10055-023-00748-5
- [31] Lydia Weinberger, Christian Eichenmüller, and Zinaida Benenson. 2023. Interplay of Security, Privacy and Usability in Videoconferencing. 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 185, 10 pages. doi:10.1145/3544549.3585683
- [32] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. 2023. Segment Everything Everywhere All at Once. In *Advances in Neural Information Processing Systems (NeurIPS)*. <https://openreview.net/forum?id=UHBWwFWIL>.