Glass Chirolytics: Reciprocal Compositing and Shared Gestural Control for Face-to-Face Collaborative Visualization at a Distance

Dion Barja*
University of Manitoba
Winnipeg, Manitoba, Canada
University of Waterloo
Waterloo, Ontario, Canada
barjad@myumanitoba.ca

Matthew Brehmer
University of Waterloo
Waterloo, Ontario, Canada
mbrehmer@uwaterloo.ca

Figure 1: A collaborative visualization interface incorporating our Glass Chirolytics approach, captured from both David and Abhi's perspectives as they jointly decide upon an travel itinerary. In Scene 1, David selects an origin (Madrid) and destination (Stockholm) with their left and right hands, respectively. In Scene 2, David and Abhi both spread a hand to select multiple origin and destination airports in Western Europe and India, respectively. In Scene 3, David scrolls a list of flights with their left hand while pointing at one listed flight with their other hand. Abhi continues to select destinations in India while selecting a listed flight with their left hand.

Abstract

Videoconference conversations about data often entail screen sharing visualization artifacts, in which nonverbal communication goes largely ignored. Beyond presentation use cases, conversations supported by visualization also arise in collaborative decision making, technical interviews, and tutoring: use cases that benefit from participants being able to see one another as they exchange questions about the data. In this paper, we employ a reciprocal compositing of visualization and interface widgets over the mirrored video of one's conversation partner, suggestive of a pane of glass, in which both parties can simultaneously manipulate composited elements via bimanual gestures. We demonstrate our approach with implementations of several visualization interfaces spanning the aforementioned use cases, and we evaluate our approach in a study (N = 16) comparing it to videoconferencing while using a mouse to interact with a collaborative web application. Our findings suggest that our approach promotes feelings of presence and mutual awareness of analytical intent.

CCS Concepts

• **Human-centered computing** → **Visualization**; *Gestural input*; *Mixed / augmented reality*; *Collaborative and social computing*.

Keywords

Collaborative visualization, synchronous collaboration, augmented reality, gestural interaction

*Barja performed this work while at the University of Waterloo.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791122

ACM Reference Format:
Dion Barja and Matthew Brehmer. 2026. Glass Chirolytics: Reciprocal Compositing and Shared Gestural Control for Face-to-Face Collaborative Visualization at a Distance. In *Proceedings of the 2026 CHI Conference on Human*

Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3772318.3791122

1 Introduction

Videoconference applications continue to be a prevalent medium for synchronous remote communication between knowledge workers [6, 86], communication that often involves the sharing and discussion of data visualization and other analytics artifacts [15]. When using applications such as Microsoft Teams [59] or Zoom [91], a single participant assumes a presenter role and screen shares slide presentations or dashboard applications. When screen sharing, others cannot interact with these artifacts, and participants' webcam feeds are relegated to thumbnail videos in the periphery. As a consequence, participants' attention is split and they lose awareness of nonverbal communication cues omnipresent in face-to-face conversations, such as deictic gestures and eye contact. Recent work [28, 37, 54] has explored the potential of augmenting a presenter's webcam feed by compositing visual artifacts atop it, which are in turn revealed or highlighted by a presenter's hands gestures. With computer vision-based recognition, these expressive gestures can now also be functional. However, by following a strict *what-you-see-is-what-I-see* (WYSIWIS) approach [76], this augmentation perpetuates a pattern of one-way communication from presenter to audience and does not speak to the need for nonverbal communication in scenarios *beyond presentation*, such as collaborative ideation and analytical decision-making [19]. Moreover, many effective presentations that communicate observations grounded in data [72] can be supported with conventional statistical charts modified with simple reveal and highlight animations that draw attention to individual observations. In contrast, open-ended collaborative analytical scenarios may entail a greater breadth and complexity of visualization artifacts, with a commensurately richer set of interactions to modify them [87].

In this paper, we extend augmented videoconferencing to support rich analytical conversations with a reciprocated form of WYSIWIS augmentation applied to commodity webcam video. We call our approach *Glass Chirolytics*, as it suggests a single pane of glass separating the participants: our approach composites visualization artifacts and interface widgets atop the mirrored webcam video of a conversation partner. Unlike prior approaches to augmented video, which prioritized one-to-many presentation scenarios, ours prioritizes paired data analysis with inherently relational data and allows for either participant to control composited and visually-interconnected elements with bimanual mid-air gestures.

Our **research contributions** are as follows. First, we propose *Glass Chirolytics* as a novel approach to supporting *face-to-face* paired analytics conversations between remote participants, combining reciprocal WYSIWIS video compositing with gestural interaction, allowing either participant to manipulate shared visualization artifacts. We assemble a bimanual gestural vocabulary for supporting analytical conversations, which features two novel gestures as well as logic for supporting gestures performed simultaneously by two sets of hands. This vocabulary prioritizes the analytically useful actions of navigating coordinate spaces as well as the manipulation and persistent selection of visual elements, so as to reveal connecting relationships between them. Next, we demonstrate the potential utility of our approach through the implementation of seven application interfaces, spanning decision making, collaborative analysis, one-on-one tutoring, and technical interviewing. Lastly, we report and reflect on findings from an evaluation with 16 participants, one based around a collaborative decision-making scenario.

2 Background & Related Work

Earlier work involving the augmentation of videoconference experiences for delivering presentations about data via gestural interaction had theoretical foundations in the *rhetorical* use of gesture for public speaking and in *communicative* use cases for data visualization. In contrast, the foundations of this work draw more from the *conversational* utility of gesture in face-to-face exchanges and in *collaborative* use cases for visualization.

2.1 The Utility of Hand Gestures in Conversation and in Problem Solving

Voluntary nonverbal communication expressed via facial expression as well as via head and hand movement are essential to face-to-face conversation. In particular, hand gestures amplify communicative intentions [26, 82], and in educational contexts, they promote discussion engagement and improve learning outcomes [53]. Moreover, they benefit listener and speaker alike [21]: for the former, gestures provide extra context when speech is unclear, improving overall comprehension, while for the latter, gestures support the production of speech. Despite these benefits and a consensus that interpersonal communication is more effective when gestures complement speech [47], our hands are seldom visible in videoconferencing applications: common webcam placements tend to capture only a person's head and upper torso [4]. As a result, the human mirror neuron system (hMNS), a single overarching system governing both linguistic and gestural communication, is engaged to a lesser extent during videoconference calls than in face-to-face conversations [24]. Also less apparent during videoconferencing is cross-brain synchrony [89]: the neural of alignment activity between people during social interactions and a signifier of coordination and empathy.

Deictic gestures in particular serve to direct a listener's attention [49], such as pointing at a location or object. However, with the convention of screen sharing visual aids in videoconferencing applications, deictic hand gestures, even when they appear in a speaker's video frame, are often ambiguous given the relative placement of speaker video and screen shared content on the viewer's display.

Non-communicative gesturing, in the meantime, can assist with problem-solving [30] and learning as these gestures can produce embodied representations of problems and concepts [45]. These gestures may therefore be good candidates for manipulating virtual objects in mid-air gestural interfaces, and accordingly we use this justification for some of the gestural vocabulary defined below in Section 3.2.

2.2 Augmenting Video Communication

Augmenting synchronous video communication between remote peers has long been an area of interest in HCI [27], and the lineage of this research includes work dedicated to restoring an awareness of gestural communication (e.g., [79]) and other non-verbal cues such as facial expressions (e.g., [43]) in collaborative tasks. In recent years, both commercial tools and research projects have introduced approaches to augmented video experiences with composited visual aids, and beyond making gestures visible, some of this work allows for the manipulation of these visual aids via gestural interaction.

Compositing webcam video and visual aids. Several existing commercial videoconferencing tools can composite screen shared content with a speaker's webcam video, producing either a picture-in-picture experience akin to how visual aids appear over the shoulder of a broadcast news anchor, or an experience in which the speaker's outline is segmented and composited atop the content, akin to a broadcast weather reporter. These include virtual camera applications such as OBS [65] and Airtime [60], operating system features like macOS's Presenter Overlay [1], and recent webcam segmentation features in Microsoft Teams [59] and Zoom [91]. These tools assume a designated presenter role, in which a single participant updates their visual aids by interacting with the screen shared application's interface. A limitation of these approaches is that the use of deictic gestures to direct an audience's attention to elements in the composited visual aids requires coordination and practice, particularly if the content is composited behind the speaker. To address this, recent research (e.g., [23, 37, 54]) composites content atop of webcam video, a design choice that we also make in our work.

Looking beyond presentation use cases, recent research has looked to support remote and hybrid videoconference meetings by exploring approaches that composite webcam video from multiple participants with shared content. Grønbæk et al.'s MirrorBlender [35] and Mirrorverse [34], for instance, incorporate a *what-you-see-is-what-I-see* (WYSIWIS) [76] composited experience wherein all participants have a common perspective on the shared content; each participant can manipulate the position, scale, and translucency of their webcam video feeds, thereby adapting the interface in real-time to diverse meeting situations and allowing them to direct the attention of their peers. Hu et al.'s OpenMic [41] is another point in this design space, one that applies proxemic metaphors to conversational turn-taking in multi-party teleconference meetings. In our work, we employ a reciprocal WYSIWIS experience to support face-to-face meetings between two people.

Gestural interaction with composited visual aids. Several of the aforementioned commercial tools (e.g., [1, 91]) also incorporate computer vision and specifically basic pose recognition to trigger animations when a participant performs a static gesture, such as by raising a hand or giving a thumbs-up. For continuous hand tracking and the recognition of dynamic mid-air hand gestures, depth cameras such as the Microsoft Kinect [73] can be used to reveal and animate visuals composited in the video foreground [70]. More recently, researchers have used frameworks such as OpenPose [18] and MediaPipe [81] to achieve similar results with commodity webcams [17, 54]. These developments have led to augmented video presentation experiences that allow presenters to directly manipulate individual multimedia objects [23] and even the contents of entire web browser tabs composited behind or in front of a presenter [22], thereby replacing indirect manipulation with pointing devices. Thus far, interfaces that composite webcam video with visual aids grant the gestural manipulation of these aids to a single presenter, meaning that when these interfaces are used in videoconferencing scenarios, only one participant can reveal, move, or draw content. Moreover, the gestural vocabulary of these interfaces, while likely appropriate for manipulating multimedia and web content, is unlikely to support the breadth and complexity of interaction typical of visual data analysis applications [87].

2.3 Collaborative Visualization

Isenberg et al. [42] define *collaborative visualization* as "*the shared use of computer-supported, (interactive,) visual representations of data by more than one person with the common goal of contribution to joint information processing activities.*" The parenthetical *'interactive'* is notable in this definition, as it serves to distinguish typical presentation scenarios from analytical ones; in the former category, it is likely that a single presenter interacts with visual representations to satisfy the information processing needs of an audience, whereas in the latter category, there is arguably a more equitable access to interactive affordances. The collaborative visualization design space can be characterized by the two established axes of computer-supported cooperative work [46]: *space* (co-located vs. distributed) and *time* (asynchronous vs. synchronous). Brehmer and Kosara [15] further characterize synchronous collaboration with and around visualization artifacts along a spectrum of *formality*, from informal data analysis sessions between peers to semi-improvised briefings and formal presentations with designated presenter roles. We therefore position our work within the *distributed* (i.e., remote or hybrid) and *synchronous* quadrant of this design space, focusing on *informal* data analysis scenarios involving a pair of collaborators.

Infrastructure for collaborative visualization. Early work in remote and synchronous collaborative visualization [5] showed that the ability for remote peers to concurrently see and manipulate a shared interactive visualization artifact can improve group performance on a collaborative problem-solving task. More recently, infrastructural support for remote and synchronous collaborative visualization includes Badam et al.'s Vistrates [3], which presents a workflow for creating, interacting with, and presenting visualization artifacts for distributed peers, and Schwab et al.'s VisConnect [71], which supports synchronizing low-level interaction events when many peers interact with these artifacts. Infrastructure for synchronous collaboration is also appearing in commercial tools, such as Observable's Canvases [66]. Complementing these projects are Neogy et al.'s design space [64] for representing remote peers' concurrent interactions, as well as Han and Isaacs's deictic approach [38] to cursor-based gestural interaction for annotating and drawing peers' attention to visualization elements. In general, these frameworks and techniques are agnostic as to whether remote collaborators can also be seen and heard, such as via a separate videoconference application.

Collaborative visualization beyond the desktop / beyond mouse and keyboard. While much of the extant work in remote

and synchronous collaborative visualization to date addresses keyboard and mouse interaction with web-based visualization, some explicitly addresses asymmetric interaction and display modalities, such as Tong et al. [80]'s recent finding that collaborators with asymmetric device capabilities (i.e., desktop and head-mounted displays) can achieve similar task performance relative to those with symmetric device capabilities. However, irrespective of visualization display and interaction modalities and the possible asymmetries between collaborators, prior synchronous and remote collaborative visualization does not directly acknowledge the value of being able to see (or hear) one's collaborators and the nonverbal communication cues they make while performing a collaborative task, a value that is central to our current work. A recent exception is Borowski et al.'s DashSpace [10], which enables synchronous collaborative immersive analytics by those using either head-mounted displays or desktop devices; those using the former appear as silhouette avatars and those using the latter appear as floating video thumbnails in the periphery of a 3D scene, both serving to provide a sense of collaborator presence. In our work, we concentrate on low-cost commodity webcams rather than head-mounted displays, and by compositing visualization artifacts with collaborator video, our goal is not only presence, but also to promote the reception of nonverbal communication that a face-to-face perspective offers.

2.4 Visualization in Augmented Video Presentations

In recent years, several research projects [28, 37, 52, 78] have demonstrated gestural interaction with composited visualization artifacts for video-based presentations, each incorporating a WYSIWIS approach in which both a presenter and their audience see artifacts composited over the presenter's webcam video. The common inspiration for these tools can be traced to a 2010 BBC documentary hosted by the late Hans Rosling [9], in which a semi-transparent animated bubble chart appears in the foreground; meanwhile, Rosling's gesticulations give the impression that he is controlling the dynamic chart animation with his movements. The types of visualization artifacts and gestural vocabulary featured in these projects understandably prioritize presentation scenarios. First, Hall et al. [37] demonstrated bimanual rhetorical gestures for revealing, annotating, and comparing elements in common statistical charts such as bar, line, and proportion charts. Femi-Gege et al.'s VisConductor [28] then extended this vocabulary with continuous gestures for controlling playback in animated bubble and rank bar charts, however this more complex vocabulary necessitated the use of a secondary presenter display annotated with gestural hints and detection feedback. Most recently, Takahira et al.'s InSituTale [78] is also reminiscent of Rosling with the integration of physical objects in presentations about data, with gestures that manipulate common household objects that trigger updates to composited visualization artifacts based on the position and orientation of the held objects. In this work, we explicitly depart from presentation use cases [28, 37, 78] with an augmented video approach for addressing collaborative data analysis scenarios. Our contribution extends the design space of this emerging medium by incorporating a gestural vocabulary grounded in an interaction taxonomy for visual analysis [87], with composited visualization artifacts reflecting complex spatial and relational structures, such as node-link network graphs and origin-destination trajectory maps.

3 The Design of Glass Chirolytics

We propose *Glass Chirolytics* (Fig. 2) as an approach to face-to-face pair analytics [2] between two remote collaborators, an approach developed for the medium of gesture-aware augmented videoconferencing. The name combines the metaphor of a glass partition between participants with a neologism referencing the use of our hands (*'chiro–'*) for collaboration around shared visual *analytics* artifacts.

3.1 Key Design Decisions

Informed by prior work in augmented video (Sections 2.2, 2.4), we made three governing design decisions:

D1: Provide a common perspective on shared visual aids while coordinating face-to-face conversation. We depart from a strict *what-you-see-is-what-I-see* (WYSIWIS) approach [76] in which all participants see the same interface. Instead, by focusing on the give and take of analytical conversations between two people, we ensure that one can always see their conversation partner, rather than a reflection of themselves. While this mutual visibility is also achieved in conventional videoconferencing, as soon as those conversations require visual aids, this visibility is compromised with the shift to screen sharing. We therefore composite visual aids over the webcam video of one's conversation partner, with the latter shown in grayscale to boost the salience of the visual aids (Figures 1, 3 — 6). The effect is not unlike two people facing one another with a pane of glass or a semi-transparent display material [31] separating them. However, an issue that arises when using these display materials in a co-located setting is that content must be legible from both sides; in other words, it must be symmetric, which precludes the display of text and many visualization design conventions. In remote settings, this issue is resolved simply by compositing content over a horizontally-mirrored video feed of one's conversation partner, which ensures that both participants are looking (and gesturing) at the same visual aids.

While we do away with the 'self-view' video of conventional videoconferencing, we nevertheless provide both participants with visual feedback (Fig. 2). Composited atop the visual aids is feedback for the local collaborator's hands, appearing as a bright green skeletal mesh, while the successful recognition of their gestures appears as ephemeral activation icons. Altogether, we characterize our approach to augmented videoconferencing as employing a *reciprocal compositing* of remote video, shared visual aids, and gestural feedback.

D2: Support the collaborative analysis of complex data abstractions. Prior work exploring the use of augmented video to talk about data with remote audiences (Section 2.4) understandably prioritized presentation-oriented visualization idioms [51]. This includes a palette of familiar statistical charts for presenting observations about tabular data [37] or animated approaches for suspenseful storytelling about time-varying data [28]. However, analytical use cases involving visualization [62] often entail a targeted interest beyond individual values, attribute distributions, or trends, and towards more complex structural constructs such as

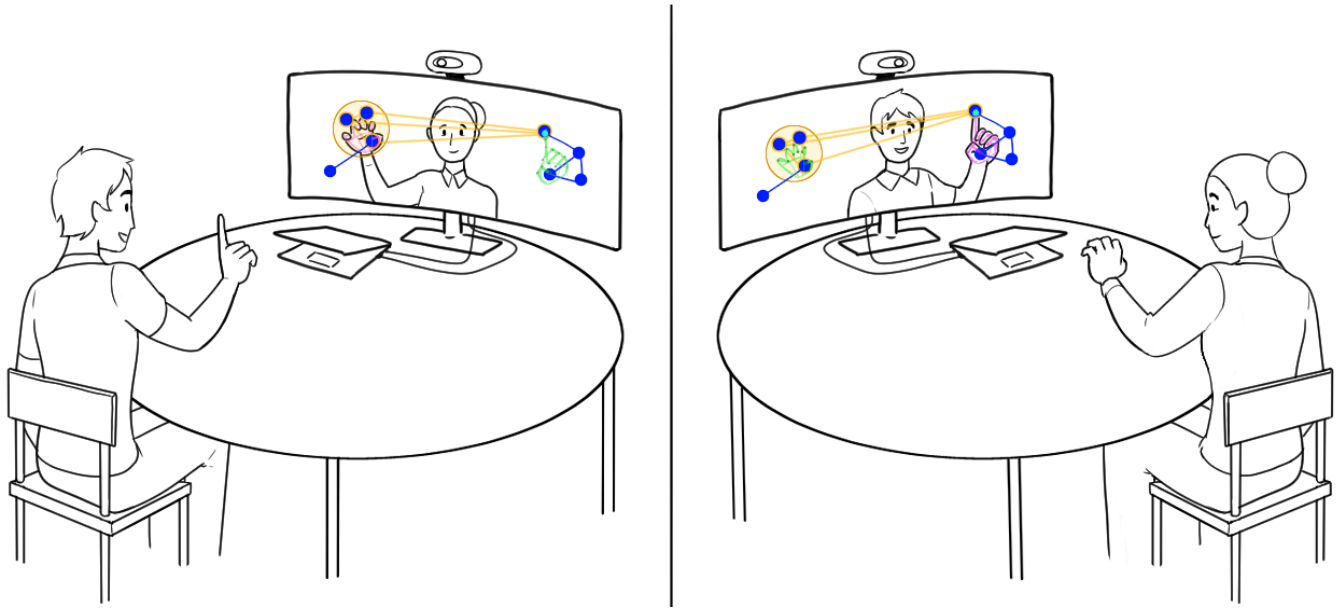

Figure 2: In this illustrated instance of the Glass Chirolytics approach, a remote collaborator appears composited behind a synchronized node-link diagram. The collaborator on the left points with their right hand to select one node while the collaborator on the right selects multiple nodes with their left hand; both can see a reflection of their hands as a green skeletal mesh composited in their local foreground. Lastly, the combination of their gestures triggers the reveal of highlighted links between the set of selected nodes.

correlations, (spatial) relationships, and topologies. We therefore prioritize the categories of visualization artifacts associated with these higher-level abstractions, such as origin-destination network maps (Fig. 1 and Fig. 4: Scene 2) and node-link diagrams (Fig. 3 and Fig. 5: Scene 1).

D3: Synchronize interaction with a gestural vocabulary commensurate with analytical use cases. In data presentation scenarios, seldom do presenters need to depart from showing and comparing values in common statistical charts, manipulated through progressive reveal and ephemeral highlighting [15]. Cordeil et al.'s Hanstreamer [52] is therefore notable for its showcasing of dense node-link diagrams along with a gestural vocabulary incorporating persistent selection, zooming, and the repositioning of visual elements, suggesting that this medium could accommodate more complex analytical use cases. Our gestural vocabulary, characterized below in Section 3.2, allows two people to concurrently ask questions of the data, questions that require a broader palette of interactions than value highlighting and pairwise value comparisons [87]. Given our aforementioned interest in structural data abstractions (**D2**), our gestures can correspond with structural elements at varying levels of specificity: a single element, a set of elements, or multiple sets. As our approach synchronizes interaction and the state of the composited visual aids, each participant can contribute one or both of their hands to the interaction.

3.2 Gestural Vocabulary

To realize **D3**, we began by considering common mouse and keyboard actions typical in interactive exploratory data analysis tools and reflected in taxonomies of interaction for information visualization [87]: clicking or lassoing to *select* elements, scrolling or clicking and dragging on navigation widgets to *explore* the spatial distribution of elements, clicking elements and dragging them to *reconfigure* their placements, clicking on interface widgets to conditionally *filter* elements, and clicking while pressing a modifier key to *connect* elements. Unlike prior work in augmented video presentation [28, 37] prioritizing rhetorical flourish and allowing for ambiguity, we sought a vocabulary that would prioritize functional clarity to differentiate these five actions. We also required gestures that could be easily distinguished from commodity webcam video input.

***Point*: ephemeral fine *selection* of an individual element (Fig. 3a).** Pointing the index finger at a single visual element will ephemerally highlight it, such as by modifying the stroke, drop shadow, or fill color of an element in a node-link network representation, or by showing a text label. As this deictic gesture primarily serves to quickly draw the attention of one's conversation partner, moving the finger away will revert the element's appearance. We adopt this gesture from prior work in augmented videoconferencing [37, 52].

***Point-and-tap*: persistent fine *selection* of an individual element (Fig. 3c).** In the first of two novel gestures introduced in this work, we detect a quick transition away from and immediately back

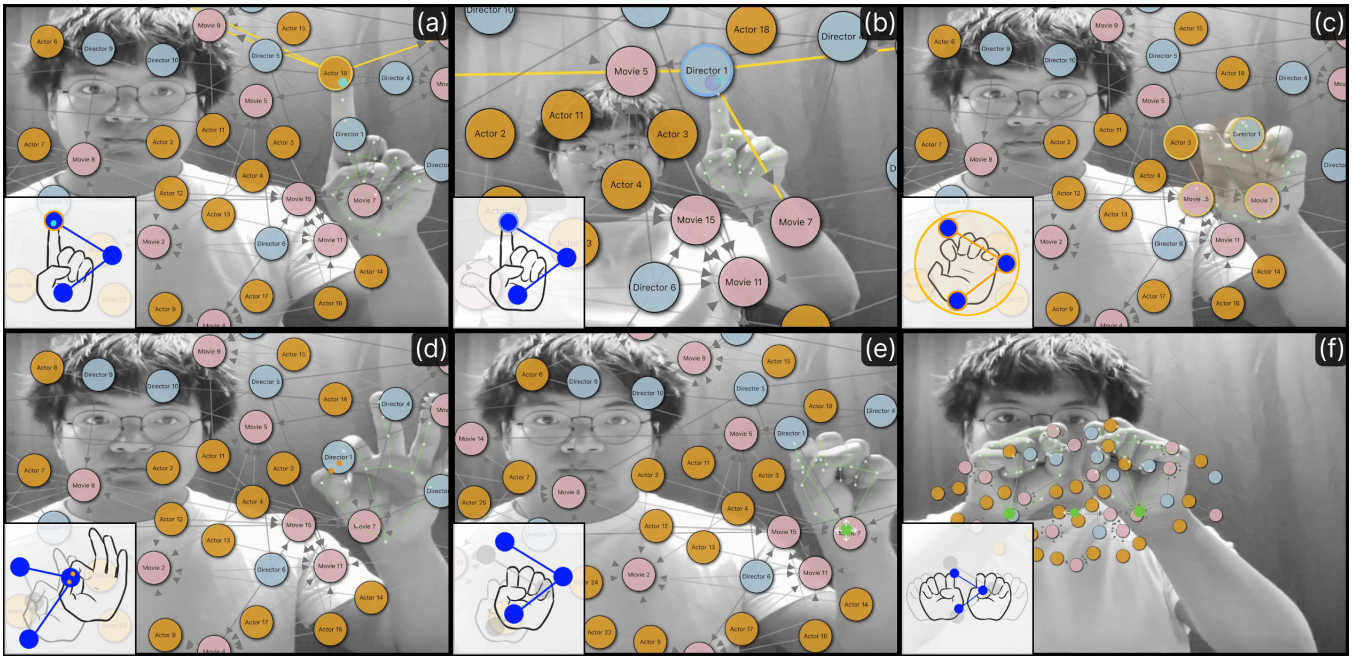

Figure 3: Our gestural vocabulary includes: (a) *pointing* to ephemerally select an individual element; (b) *point-and-tap* to persist an individual selection; (c) *spread* to coarsely select multiple elements (ephemeral unless held for one second); (d) *pinch-and-move* to reconfigure the positions of elements; (e) *grab-and-move* panning; and (f) *separate-or-join* zooming. Recognition of the latter two gestures follows a one-second delay communicated with an ephemeral green indicator icon appearing at the base of the wrists.

to the fine selection gesture (Fig. 3a) by briefly extending the thumb. This is analogous to a click event performed with the thumb when using a mouse with a side button. We indicate that a persistent selection has occurred with a different treatment of an element's visual attributes, such as by applying a different stroke color relative to ephemeral selection. This is also the gesture we employ to interact with composited action widgets emulating buttons or switchers, which we illustrate in Fig. 4 with the selection of a *filter* switcher.

In our initial design of a persistent selection gesture, we sought to replicate a click of the primary mouse button under the index finger or a single finger tap on a touchscreen interface. However, when facing a standard webcam, this gesture would map to a small motion of the finger briefly toward the camera and immediately recoiling from it. In other words, tapping toward the camera exhibited the Heisenberg effect [13], in which the finger motion changed the position of the selection.

***Spread*: ephemeral and persistent coarse *selection* of multiple elements (Fig. 3b).** Our second novel gesture is detected by spreading all the fingers with one's palm facing the camera, which will trigger a circular selection with a diameter mapped to the extent of the spread, a diameter that will continuously change as the fingers spread or contract and as the distance of the hand from the camera changes. This gesture is analogous to a wedge-shaped gesture introduced by Xia et al. [84] for coarse selections on large touchscreen displays, in which the dynamic aperture between the thumb and forefinger determines the size of the projected selection. In our case, this dynamic projection extends toward the camera rather than parallel to a 2D display. When we detect this hand shape, we visually indicate the selection area with a subtle orange circular marquee, and if visualization elements intersect with this marquee, they are ephemerally highlighted as they would be with fine selection until the hand moves away or the hand shape changes. If, however, one maintains this hand shape for longer than a second, we infer that this held pose reflects a deliberate intent to persist the selection of any elements within the circular marquee.

Combinations of coarse and fine selection gestures for two sets of hands. While the three gestures described thus far are single-handed, they can be combined in various ways. Between two collaborators, there are four hands in play, meaning that there can be up to four fine or coarse selections, and each selection can be ephemeral or persistent. Multiple selections can, for instance, *connect* selected elements by highlighting links between them (Fig. 1, Fig. 2). Moreover, since it is possible to determine which hand (left or right, local or remote) is making a selection, applications employing our approach can use this information in their selection logic, which we demonstrate in two scenarios (Sections 4.1 and 4.2).

***Pinch-and-move*: *reconfigure* the positions of individual elements (Fig. 3d).** Pinching the index finger and thumb together allows one to 'grab' a visual element and reposition it by moving the hand while maintaining the pinch. As with point-and-tap, the detection of this gesture needs to be visible when facing the camera, so the orientation of the pinch must be parallel to the display,

recalling the 'OK' gesture. Like our ephemeral pointing gesture, we adopt this gesture from prior work in augmented videoconferencing [37]. Meanwhile, we associate a two-handed *pinch-and-hold* variant of this gesture with *connecting* elements, such as by drawing a connection between elements if one did not exist previously.

***Grab-and-move*: panning to *explore* the spatial distribution of elements (Fig. 3e).** Forming and holding a fist allows one to 'grab' the underlying coordinate space on which elements are drawn and shift it vertically or horizontally. Visual feedback in the form of a circular green progress icon at the base of the wrist (visible only to who is performing the gesture) indicates a recognition of a deliberate intent to pan; releasing the fist disables 'pan' mode. We adopt this gesture from Hosseini et al.'s consensus survey of mid-air gestures across application domains [40].

***Separate-or-join*: zooming to *explore* the spatial distribution of elements (Fig. 3f).** Finally, forming and holding a fist with both hands triggers a 'zoom' mode, in which separating the fists zooms in, while bringing the fists together zooms out, analogous to a touchscreen zooming. The visual feedback to enter 'zoom' mode is similar to when panning, and likewise releasing either fist disables the mode. Like panning, we adopt this gesture from Hosseini et al.'s survey [40].

While *point*, *spread*, and *pinch-and-move* are simultaneously deictic and functional, *grab-and-move* and *separate-or-join* are non-communicative according to the classification discussed above in Section 2.1. However, the latter two may yet be useful for a remote collaborator to witness, particularly if both collaborators have an intent to explore the spatial distribution of elements.

3.3 Implementation Details

Our approach uses Yjs [44], a conflict-free replicated datatype (CRDT), to sync the underlying data for the composited visualization elements, synchronizing a shared state document used to render both collaborators' views. WebRTC [32] synchronizes this shared state document, as it needs to be updated in real time, in parallel to the remote video feeds. We use commodity webcam video to track hand shape and movement. Our gesture classification pipeline starts with the MediaPipe [81] hand landmark detection model, identifying the hand and its key points. The outputs of that model are the inputs for the MediaPipe hand gesture classification model, which identifies the gesture of a given hand. We trained a custom hand gesture classification model by combining of a subset of the HaGRID hand gesture dataset [48] and our own training data. Finally, we built the interfaces featured throughout this paper and in our supplemental video in React [58], and we used D3.js to generate the SVG-based visualization elements [12]. Our implementation is available under an open-source license at github.com/ubixgroup/Glass-Chirolytics.

4 Application Scenarios

We implemented seven visualization interfaces that demonstrate the depth and breadth of the Glass Chirolytics approach across four application scenarios: *decision making*, *exploratory analysis*, *tutoring*, and *technical interviewing*, exhibiting a variety of structural data abstractions and corresponding analytical visualization idioms (**D2**). These interfaces appear in Figures 1, 4 — 6, and in the supplemental video. Throughout these demonstrations, actors Abhi and David use the interfaces and execute the gestures in our vocabulary (**D3**). We discuss the *decision making* scenario (Fig. 1) in greater depth, as this scenario also forms the basis for our study described below in Section 5. For each scenario, we assume that any requisite data is prepared and visualization design and implementation is complete. In other words, the scope of our approach is the set of synchronous collaboration and communication episodes taking place within a longer data analysis and decision-making life cycle. This life cycle also includes work performed asynchronously and individually, from data cleaning and visualization implementation to confirmatory data analysis and documentation.

4.1 Decision Making

People use visualization artifacts as decision aids [25] when tasked either to make selections from a set of options given their attribute values, to determine threshold values to inform future selections, or to create new options and attributes [16]. When these tasks are preformed collaboratively, maintaining a shared awareness of the options and values that collaborators select can reduce redundant effort and ideally accelerate progress toward a state of consensus [57].

Collaborative itinerary and event planning is one manifestation of a decision-making scenario, one in which collaborators evaluate and make selections from a set of options. Fig. 1 presents an interface for evaluating and selecting flights between origin and destination airports on a world map, with blue circles representing airports, and viable flights between them represented as purple arcs. Such an interface could be used by a pair of family members or friends to plan a travel itinerary, or by the planning team of a conference or event, so as to identify venues that are logistically viable. As with all of the example interfaces to follow in this section, this interface allows a local collaborator to see their remote peer's intentions as well as their peer's reactions to their own interactions with the interface.

In Fig. 1a, David *points* at Madrid with their left index finger and at Stockholm with their right index finger, ephemerally selecting both locations and revealing a flight path between them. This particular interface associates left-hand selections with origin locations and right-hand selections with destination locations, and these selections are reflected in respective list widgets composited in the top left corner of the interface. After observing this selection behavior, Abhi is ready to evaluate some options with David, however this will require some negotiation and panning of the map via the *bimanual grab* gesture (see video). In Fig. 1b, David performs a *spread* gesture with their left hand to make an ephemeral coarse selection of five candidate origins in Western Europe, while Abhi performs the same gesture with their right hand, resulting in a similar coarse selection of three destinations in India. Given this coordinated selection of origins and destinations, a list widget containing candidate flights now appears to the left of the composited interface, along with histograms reflecting the cost, travel time, and departure dates of these flights below them. Finally, in Fig. 1c, David uses their left hand to perform the *pinch-and-move* gesture over this widget, which scrolls the list of flights. During this scroll, David spots an affordable flight leaving on an acceptable date between

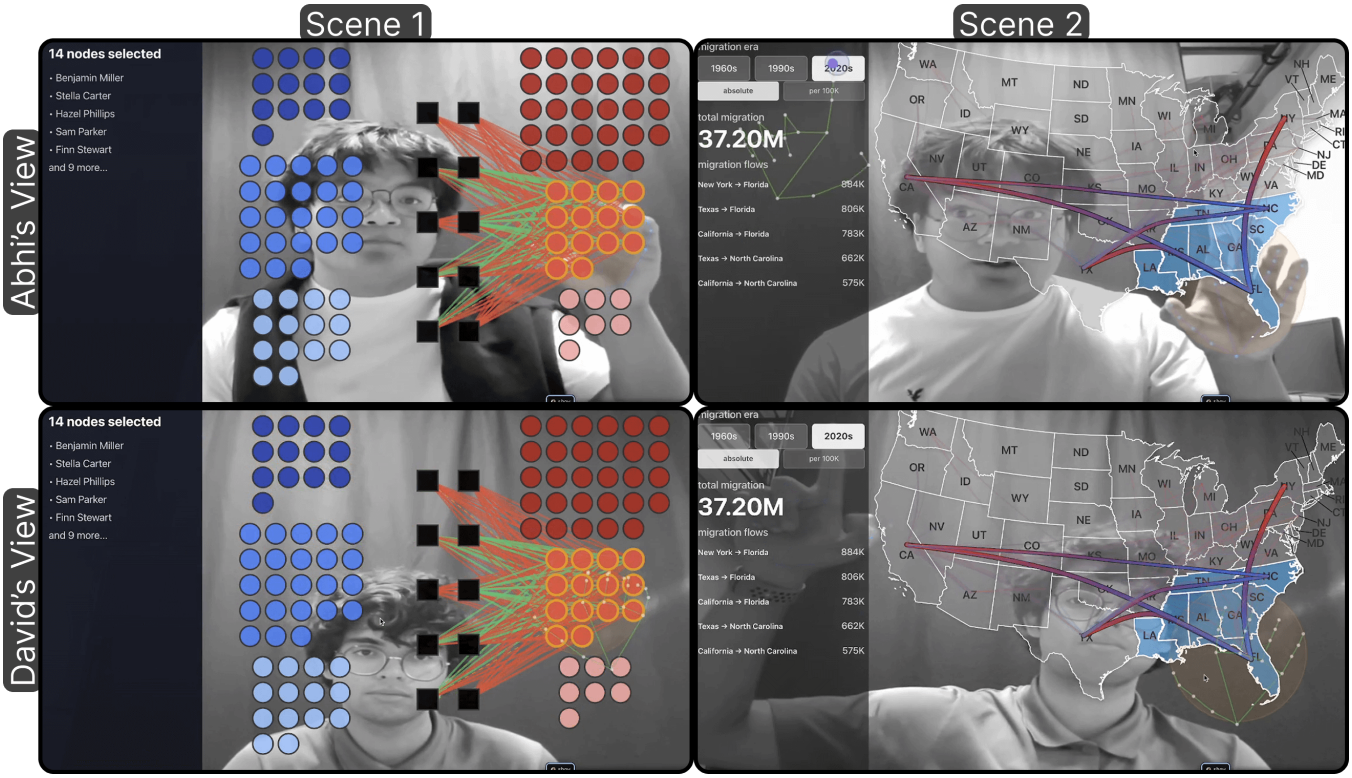

Figure 4: *Exploratory analysis* scenarios: in Scene 1, David spreads their hand to select a group of establishment legislators on the right, revealing their voting patterns. In Scene 2, David spreads their right hand to select states in the American Southeast, revealing the largest sources of incoming domestic migration to those states, while Abhi filters the data with their left hand to reflect the 2020s.

Madrid and Delhi, pointing toward it, but without activating its ephemeral selection. Abhi maintains their coarse selection within India while using their other hand to ephemerally select the flight of interest.

4.2 Exploratory Analysis

Collaborative visualization scenarios involving two people is sometimes characterized as *paired analytics* [2]; like paired programming, this could include situations where a pair of visual analysts work together, or when a visual analytics application expert works alongside a subject matter expert, translating the latter's domain questions into interactions with an interface.

Two example interfaces speak to this scenario. First, in Fig. 4 (Scene 1), Abhi and David take on the roles of political scientists analyzing the voting patterns of elected officials with respect to proposed legislative bills. Abhi asks whether the voting patterns of each party's establishment contingents are similarly homogeneous; in response, David performs a *spread* gesture with their right hand to make an ephemeral coarse selection of moderate legislators in the right party, revealing their votes for and against the ten bills, represented as green and red links, respectively.

As a second example, in Fig. 4 (Scene 2) Abhi and David are social scientists analyzing trends in inter-state migration patterns. As in the decision making scenario above, this interface shows data on a map and distinguishes a selection based on which hand performs the gesture. Here, David performs a *spread* gesture with their right hand, making an ephemeral coarse selection of the Southeast states as migration destinations, painting them blue and revealing their largest sources of migration as arcs that transition from origin states (red) to destination states (blue). Meanwhile Abhi expresses an interest in recent migration patterns, *pointing and tapping* to persist a selection in a radio widget composited in the top left, filtering the migration data to reflect the 2020s. Their coordinated selection reveals a ranked list of state-to-state migration to the Southeast during this period.

4.3 Tutoring

One-on-one peer tutoring scenarios, particularly in science, technology, engineering, and mathematics (STEM) fields, often involve the use of visualization artifacts, diagrams, and other visual abstractions to communicate concepts. One inspiration for this scenario is a particular instantiation [67] of Perlin et al.'s Chalktalk [68], a sketch-based presentation tool in which pen-based gestures are used to draw and control dynamic physical and mathematical models; in this demonstration [67], Perlin sketches using a light pen, compositing the sketches over his webcam video as he explains and animates them. We anticipate that our approach could be similarly employed, albeit without requiring a specialized pointing device.

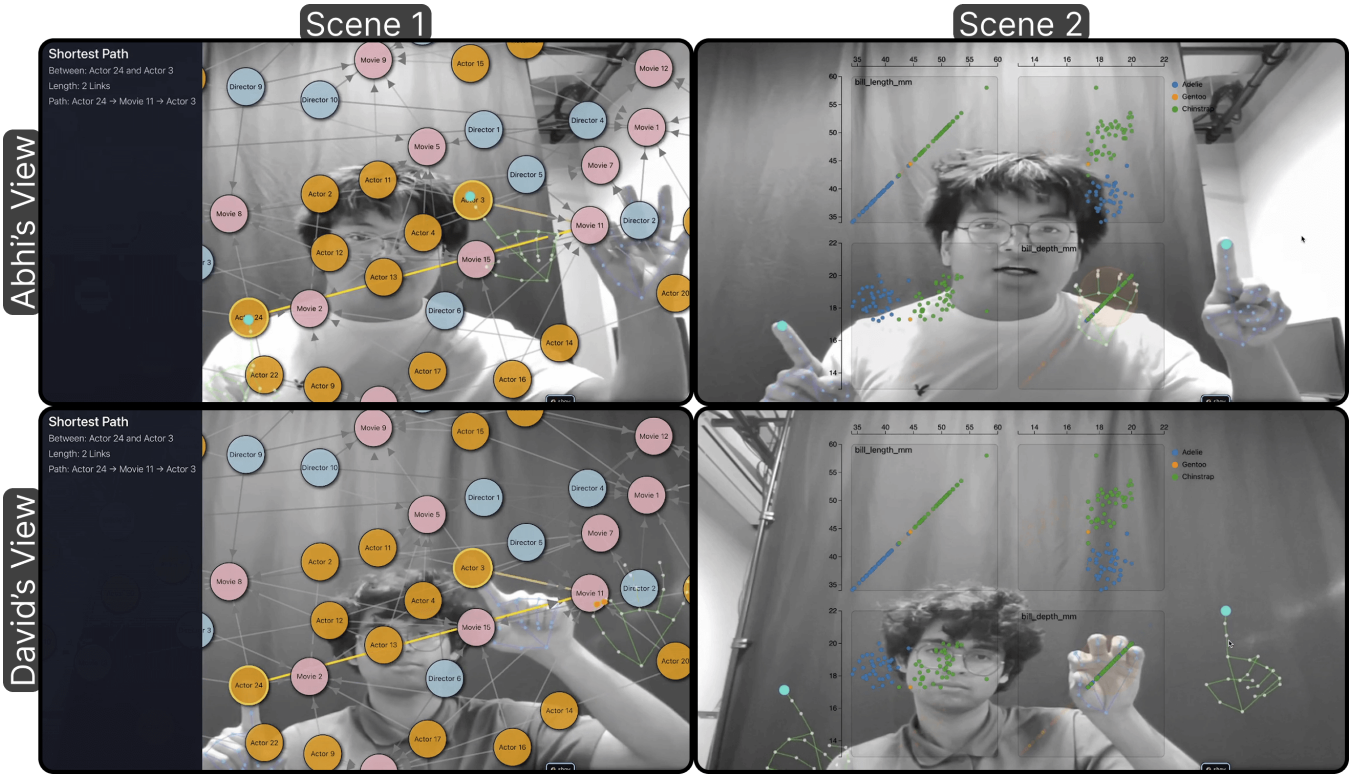

Figure 5: *Tutoring* scenarios: In Scene 1, Abhi points to different nodes with both hands to reveal the shortest path between them, while David moves the path's intermediary node a new position via *pinch-and-move*. In Scene 2, Abhi spreads their right hand to select points in the bottom-right plot of a SPLOM, revealing the selected data points' positions in the other plots.

Moreover, our approach ensures a mutual visibility between tutor and student, allowing the former to ask questions about visual elements via deictic gestures. As for shared gestural control of composited visual aids, the tutor might selectively grant interactive privileges to the student to evaluate their learning.

Two of the implemented interfaces reflect this scenario. First, in Fig. 5 (Scene 1), Abhi is teaching David about graph topologies using a node-link representation. He illustrates the shortest path between two nodes by performing a *point* gesture with both index fingers, which ephemerally highlights the path between them and lists the nodes along the path in the left margin of the interface. To better see the connecting node along this path, David *pinches and moves* the node to more clearly see the highlighted path. Second, in Fig. 5 (Scene 2), Abhi is illustrating pairwise relationships between attributes in a tabular dataset using a scatterplot matrix (SPLOM). When they perform a *spread* gesture over a set of data points in one of the scatterplots, they ephemerally highlight the same data points in the other scatterplots while dimming out unselected data points, drawing David's attention to the different correlation relationships between attribute pairs.

4.4 Technical Interviewing

Lastly, our approach could also be used in technical assessments of job applicants, such as in system design and data analysis interviews. Figure 6 (Scene 1) presents an example of the former, in which Abhi provides David with a dynamic shared whiteboard interface, one where David can *pinch and move* architectural components from a component menu on the left, generating clones of them and placing them where they deem appropriate. After placing them, David can perform *pinch and hold* gestures on any two placed elements to add a visual connection between them. At any point, Abhi can interject with *point* gestures of their own, ephemerally highlighting elements, asking David to explain the rationale for their architectural design choices. Meanwhile, Figure 6 (Scene 2) presents an example of the latter, in which Abhi tests David's visualization literacy by providing him with a Sankey diagram depicting the propagation of energy from natural sources to points of consumption. Here, David *points* at two nodes in the diagram to ephemerally highlight the proportion of natural gas passing through the electric grid.

5 Evaluation

We evaluated our approach in a study designed around the *decision-making* scenario described above in Section 4.1 and depicted in Fig. 1. Our baseline for comparison was a collaborative visualization application (Fig. 7), one in which the local and remote webcam feeds appear as thumbnail videos in the top right corner of the interface, replicating a conventional videoconferencing experience. However, this baseline application diverges from the convention of one-way screen sharing, a convention that we deem to be unfairly disadvantaged relative to our approach. Instead, our baseline

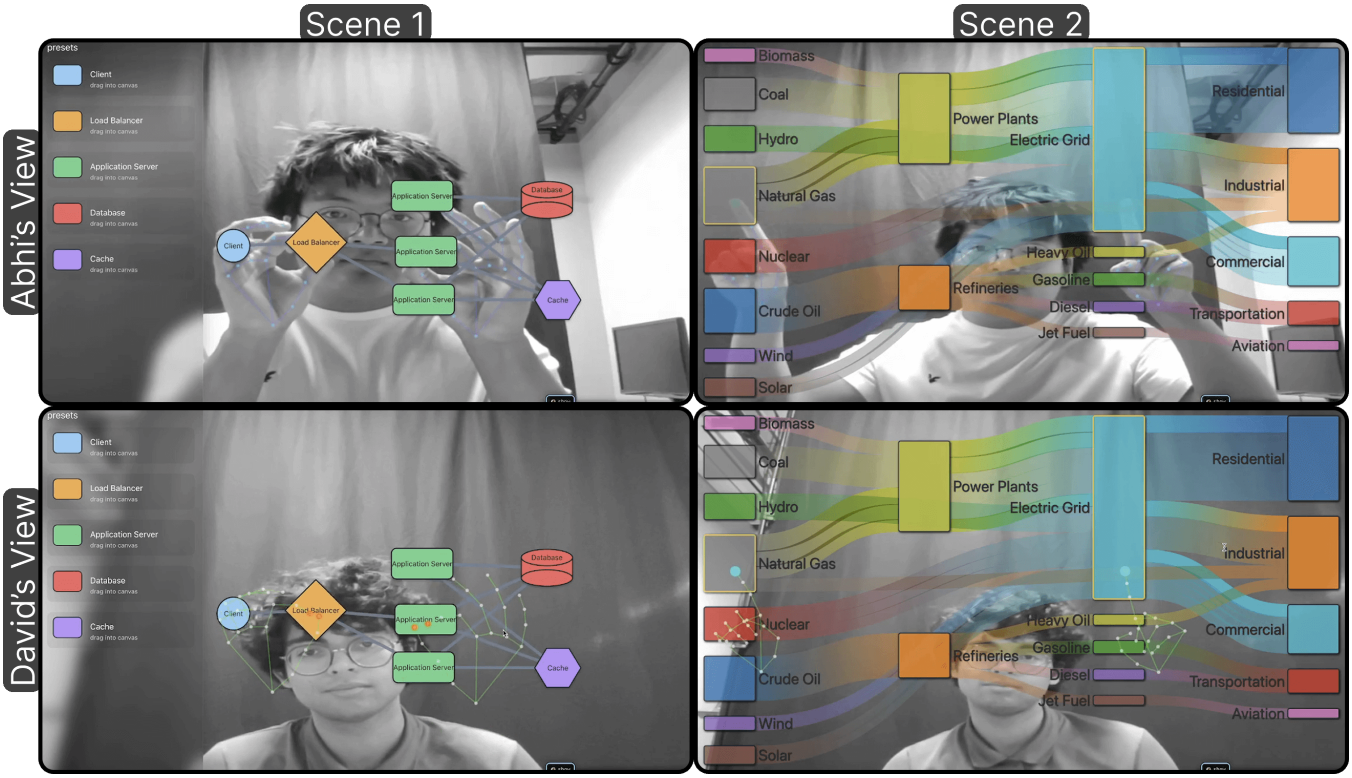

Figure 6: *Technical interviewing* scenarios: In Scene 1, David repositions two elements in a system design diagram. In Scene 2, David points at two nodes in a Sankey energy diagram to highlight the flow of natural gas to the electric grid.

application allows for shared mouse cursor awareness and control of visualization and interface widgets by either collaborator using mouse input, similar to previous work on synchronous collaborative visualization [38, 64, 71].

As a result, both baseline and Glass Chirolytics applications satisfy design goal **D2** (*support the collaborative analysis of complex data abstractions*; see Section 3.1), whereas only the latter supports **D1** (*provide a common perspective on shared visual aids while coordinating face-to-face conversation*) and **D3** (*synchronize interaction with a gestural vocabulary commensurate with analytical use cases*).

The University of Waterloo research ethics board approved this study.

5.1 Participants

We recruited 16 participants, a group reflecting diversity in terms of gender and age from the student and professional population in the locality of our institution. We were permitted to advertise the study via university mailing lists and on physical poster boards around campus, spanning multiple faculties. We did not specify explicit inclusion or exclusion criteria, such as having specialized knowledge about network data or prior data analysis experience. As a result, the recruited participants reflect a range of education and prior experience with respect to gestural interfaces. Eleven participants were between the ages of 21 and 25 and five were between the ages of 26 and 40. Eleven participants had an academic or professional background in computer science, three had a background in engineering, and two had a background in environmental science. 14 participants used videoconferencing applications at least weekly. Nine participants reported having a limited experience with gestural interaction, while five reported a moderate level of experience, and two reported extensive experience. All of the participants reported having experience with visualization and analytics applications, with eight reporting a moderate level of experience and eight reporting extensive experience. The study took an hour to complete, and we remunerated participants with a $20 CDN multi-retailer gift card.

5.2 Setting, Apparatus, & Procedure

We conducted each study session in a controlled lab setting, with two participants per session. Upon arrival, participants signed a consent form, completed a questionnaire on demographics and prior experience with teleconferencing, gestural interaction, and visualization, and received a brief introduction to the study's purpose and procedures. To emulate a remote videoconferencing experience, we asked participants to sit at desks in separated partitions of the lab space. We provided each participant with an M1 MacBook Air laptop connected to a 27" monitor perched on an adjustable stand, a USB mouse, a 1080p Logitech USB webcam positioned atop the monitor, and a Logitech combined headphone-microphone USB headset. The physical separation between participants ensured that they could not directly see or hear each other beyond the videoconferencing experience.

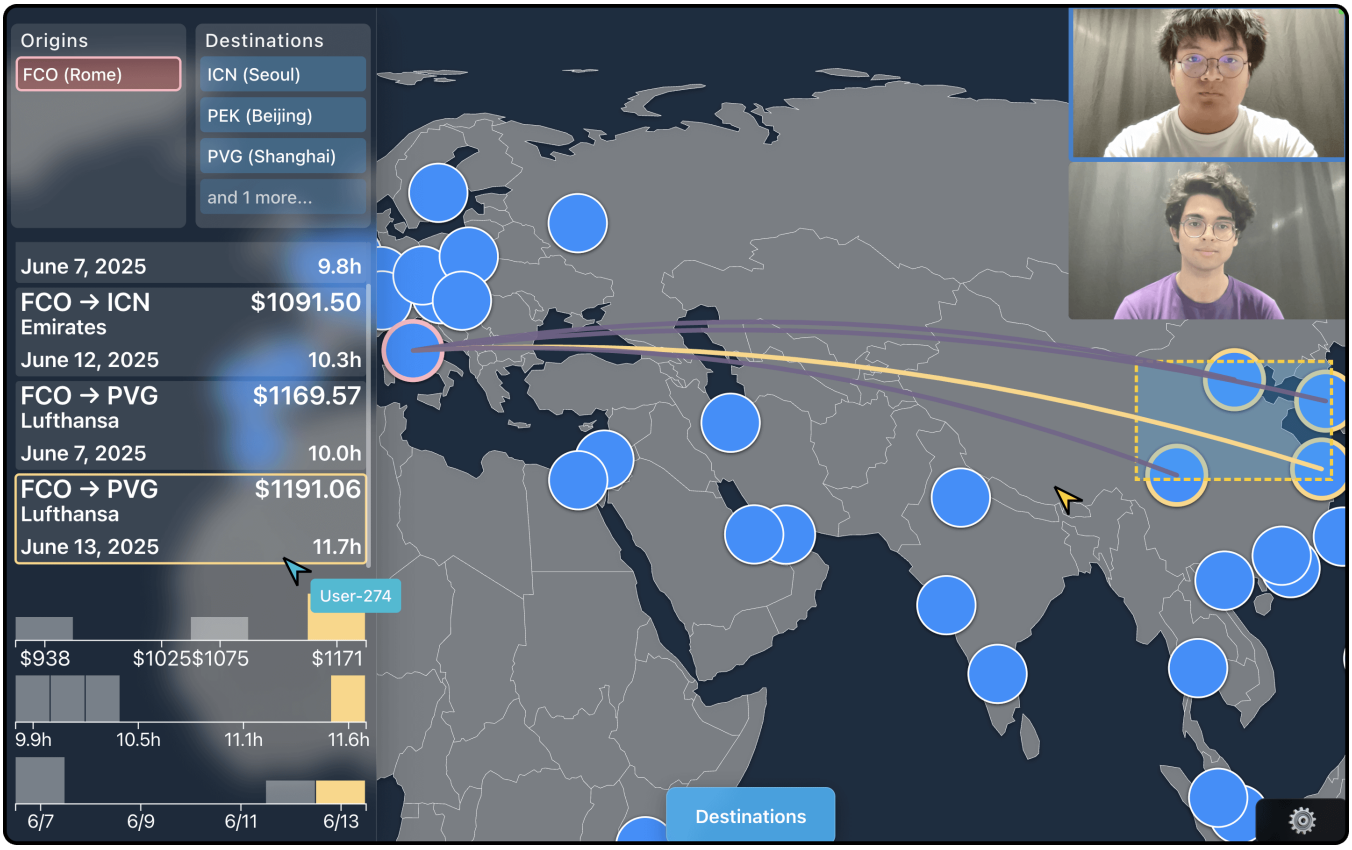

Figure 7: A screenshot of the baseline flight search application used in the evaluation. Remote collaborators can jointly control the interface via mouse interaction and see each other's cursor position.

The task that we assigned each pair of participants mirrors the *decision-making* scenario described in Section 4.1. We selected this scenario as its visualization is representative of a high level of visual complexity relative to the other scenarios, reflecting highly interconnected data and featuring multiple coordinated views, with tasks necessitating a greater employ of our gestural vocabulary. Also unique among the others, this scenario required no specialized domain expertise, thereby avoiding situations in which there is an asymmetry of prior subject matter knowledge between the two participants.

The participants used an interactive flight search interface to arrive at a consensus travel decision. To achieve this, the participants role-played as a pair of travelers planning a trip to a shared destination that accommodated their individual constraints. We indicated a small set of travel origins (e.g., New York City, Washington D.C.) as well as a deliberately vague *cognitive region* [61] as the travel destination, such as *Western Europe* or *Southeast Asia*, so as to promote deliberation between participants regarding the bounds of this region, from which they would narrow down to specific destinations (e.g., Paris, Bangkok). We also gave each participant a unique set of travel constraints encompassing a budget, a departure date window, and a set of preferred airlines. Solutions for the task were flight itineraries that reflected the intersection of constraints given to both participants. For this study, we procedurally generated a dataset containing two thousand flights, and ensured that both participants' constraints overlapped in such a way that would lead them to identify up to three solution itineraries satisfying their constraints. The task ended either after 15 minutes or when the participants had found all of the solution itineraries.

Participants performed this task once with the baseline application and once with the Glass Chirolytics application. We counterbalanced the ordering of the applications as well as the assignments of origins, destinations, and constraints across all participants. Prior to using each application, the researcher provided a tutorial consisting of a short prerecorded video and an interactive training session for practicing how to control the interface via mouse or gesture; we include the former as supplemental material.

We recorded conversation audio as well as a video of the applications from both participants' perspectives during the study. After using either the baseline or the Glass Chirolytics application, we asked participants to complete the Temple Presence Inventory (TPI) [56], an assessment used in prior remote collaboration research (e.g., [69]) that reflects impressions of a collaborator's presence. We also asked participants to complete the NASA-TLX [39], which reflects subjective mental workload. For both instruments, participants answered with respect to a seven-point Likert-scale.

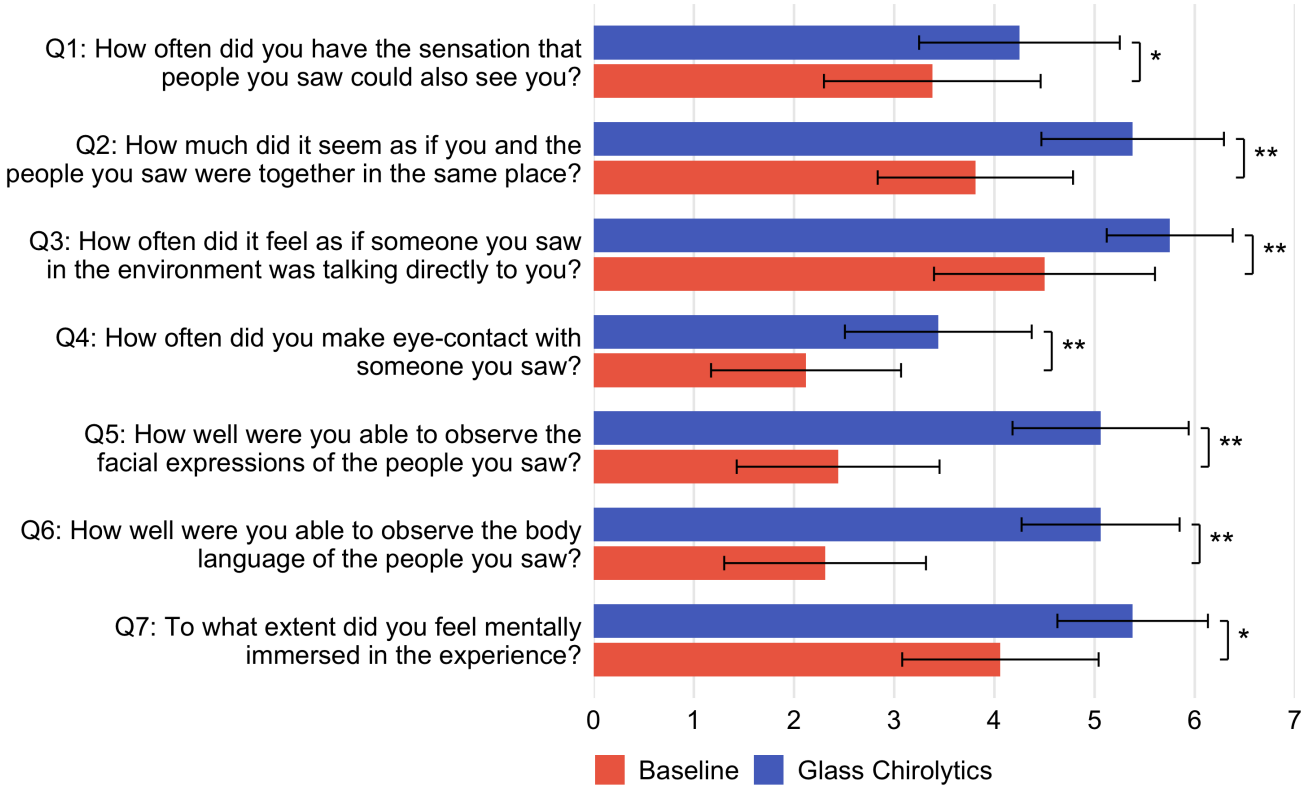

Figure 8: Temple Presence Inventory (TPI) [56] presence scores for the baseline and Glass Chirolytics applications. ($* : p < .05, ** : p < .01, *** : p < .001$)**. Error bars indicate 95% confidence intervals.**

Finally, the researcher conducted an open-ended group interview with both participants to gather qualitative feedback on their experience with both applications and their suggestions with respect to other application scenarios.

5.3 Results: Performance, Presence, & Workload

Performance. Each pair of participants identified at least one solution itinerary before the time elapsed in both conditions, and we observed no significant difference in terms of the number of solutions found. Due to the existence of multiple solutions and varying levels of rapport and conversation between participants (the experimenter encouraged participants to deliberate), the time taken to identify solutions is impractical to directly measure and compare.

For the TPI and NASA-TLX, we analyzed responses to the Likert-scale questions using the Wilcoxon Signed-Rank Test [83]. Following recommendations from previous research [74], we ensured that the sample size for this test exceeded 15 pairs.

Presence. Fig. 8 shows the distribution of scores across the TPI [56] questions. Overall, we found that participants reported a significantly higher sense of presence with the Glass Chirolytics application relative to the baseline application across all seven questions ($p < 0.05$), encompassing mutual visibility, eye-contact, observation of facial expressions and body language, and a sense of immersion.

Workload. Fig. 9 shows the distribution of scores across the NASA-TLX [39] questions. Relative to the baseline application, we found that participants reported a significantly lower level of temporal demand ($p = 0.0454$) and a significantly higher level of physical demand ($p = 0.002$) with the Glass Chirolytics application. We revisit the significantly lower perceived temporal demand below in our analysis of video observations and interview responses. We did not find different levels of reported mental demand, perceived performance, perceived effort, or perceived frustration.

5.4 Results: Video Observations

In the baseline condition, interaction was predominantly serial, where one participant would monopolize interacting with the application while the other assumed a spectator role. For instance, P5 performed nearly all interactions while P6 watched and provided feedback; we observed a similar dynamic between P7 and P8 and between P11 and P12. Pairs often resorted to territorial strategies to avoid cursor conflicts, such as P3 explicitly asking P4, *"do you want to hover over airports while I scroll on the list?"*, effectively assigning distinct interface areas. Furthermore, turn-taking relied on explicit verbal coordination, leading to awkward transitions and apparent friction. P13 and P14, for example, negotiated control using phrases like *"let me do it,"* referring to list navigation, or *"you go, you go"* when changing destinations.

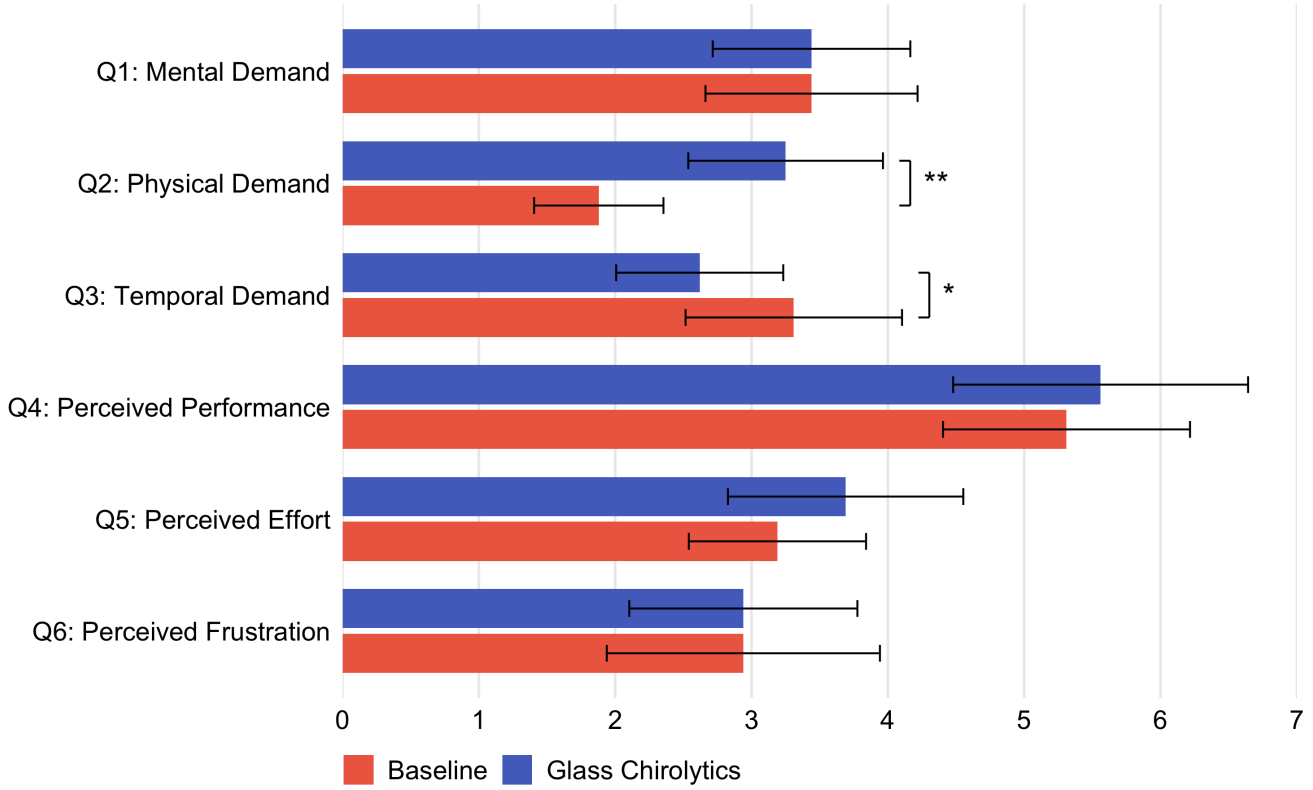

Figure 9: NASA-TLX [39] workload scores for the baseline and the Glass Chirolytics applications. ($* : p < .05, ** : p < .01, *** : p < .001$)**. Error bars indicate 95% confidence intervals.**

In contrast, the Glass Chirolytics interface appeared to foster more fluid and simultaneous collaboration where rigid roles were less apparent. P5 and P6, who interacted serially with the baseline application, worked together simultaneously, with P6 moving the map in anticipation of a destination change while P5 scrolled the list. The embodied nature of the approach served as an inherent collision avoidance mechanism, reducing the need for territoriality and encouraging participants to work closely in the same screen space. This allowed P3 and P4 to take turns, quickly hovering over potential destinations in the same region, stopping only when they identified a destination that had a favorable flight distribution. Instead of a verbal coordination of turn-taking, we observed turn-taking initiated via gestural interjections and nonverbal cues. For example, unlike their baseline behavior, P13 and P14 worked together in the same geographic area, using a combination of their partner's hand position, gaze, and conversation context to determine how to avoid collisions. Altogether, the parallelization of interaction and reduced verbal coordination effort observed when participants used the Glass Chirolytics interface may account for the perception of reduced temporal demand mentioned above.

5.5 Results: Interview Responses

We transcribed the open-ended group interview audio recordings using Microsoft Word's transcription tool, cleaned the resulting transcripts manually, and added our observations based on the study task video and audio recordings. Given this dataset of utterances and observations, we conducted a reflexive thematic analysis [14].

Glass Chirolytics provides a shared awareness of analytical intent and provenance. A recurring perceived benefit of our approach was that it provided foresight into their partner's analytical intent via their gaze and hand positioning, with P12 stating that they "*could easily see where* [their peer] *would be pointing and even before she selects the origin or the destination,*" whereas in the baseline application, "*only after she selects one of them, I could see which one she selected.*" This foresight promotes guidance, with P10 remarking that by "*being able to see people's hand motions, you can actually see what they're trying to do and guide them, versus what it's like on a mouse:* [...] *you can't see anything.*" Despite a mutual visibility of each other's cursor in the baseline application, participants found that this offered minimal utility beyond a mere position. P4 commented that a cursor "*feels insufficient to capture the entirety of my intentions.*" From the observer perspective, P6 noted that in the baseline application, "*even though you could see their mouse, it was a lot less evident what they were about to do.*"

Performing and witnessing gestural commands also increased a sense of accountability, with P5 stating that "*it's easier to track with gestures what you* [selected] *because you remember.*" In contrast, they cited an instance when using the baseline application: "*we*

clicked on two flights unexpectedly… and then we had to find out which one was wrong." This apparent embodied sense of interaction provenance made it easier to identify and rectify mistakes.

Beyond gesture, the visibility of nonverbal cues like gaze and facial expression also improved communication flow according to several participants (P3, P8, P9). P9 cited an instance in which their peer was momentarily distracted while reading their assigned travel constraints offscreen, and in this moment P9 waited for their peer to stop reading before resuming the task. Overall, participants generally paid less attention to each other when using the baseline application, with multiple individuals (P3, P11, P12) reporting to have never looked at the other participants' thumbnail video feed.

Shared gestural control: potentially distracting at first, but quickly learned and even deemed to be 'fun'. While most participants quickly gained proficiency with the gestural vocabulary, some remarked that gestures detracted from the conversation, at least initially. For instance, P4 reflected to P3 that at when first using the Glass Chirolytics interface, "*I had not enough brain power to observe you.* […] *I was trying to get used to the hand motions.*" Other participants noted a need to be continuously cognizant of the camera as well as the gesture classifier and its limitations throughout the task. P9 disliked having to "*make sure my hand is always flat and facing the camera*" when interacting with the application, wishing it was "*able to identify even if your hands* [are] *off axis.*"

Despite any initial difficulties with the gestures, most participants were able to quickly learn the gestures and were proficient enough to complete the task, despite the brevity of the training exercises before the task commenced. Discussing the gestural learning curve, P8 commented that it was "*not very big*" and took around "*two to three minutes… after that, it becomes easier.*" P9 told us that learning how to use the gestural interface *"was easier than it seemed,"* whereas with the baseline interface, they said: *"I thought would be easy, but I actually found it more difficult.*" Participants reported learning by observing their peers, as reflected on by P10: "*I got to see him do it first, so it made it a little easier seeing someone else do it.*" Similarly, several participants (P1, P13, P14, P16) described the experience of using this interface as being more fun, playful, and enjoyable than the baseline interface.

Reactions to shared gestural control suggest a rich interaction design space. Multiple participants (P5, P6, P8, P14) stated that they found gestures to be more intuitive than using a keyboard and mouse. For instance, P5 and P6 agreed that using their left and right hands for selecting origins and destinations, respectively, felt more natural than mode-switching with a mouse and keyboard. P8 would similarly characterize mouse- and keyboard-based mode-switching as abstract, while having the modes embodied in the hands felt like a tangible control scheme.

Discussion in the open-ended group interviews often turned to expanding the gestural vocabulary. One common suggestion (P1, P9, P11, P12, P13) was the ability to draw annotations with the hands. P1 suggested repurposing the *spread* gesture; instead of coarsely selecting, this gesture could spawn a circular annotation.

Participants also bemoaned the lack of an idle state in which no gestures are recognized: a mode in which one can keep their hands within the camera's view without inadvertently triggering changes to the visualization interface, such as by scratching one's head (P6). To this point, P5 requested "*a way to filter out other hand movements and then catch only when you mean to use the system,*" suggesting a potentially automated means of detecting interactive intent.

Other participants suggested expanding the gestural vocabulary by using gestures in tandem with other modalities. For instance, P2 and P15 suggested overloading gestures with multiple commands associated with different modes (e.g., idle, selection, navigation, annotation), and that this mode-switching would be done with the mouse or keyboard. Meanwhile, P16 suggested voice as an additional input modality, similarly complementing or modulating gestural input.

Meanwhile, other participants suggested that the gestural vocabulary need not be limited to one's hands. Citing large wall-mounted displays, P3 and P4 proposed incorporating the entirety of one's body, with P4 remarking that "*half of my body is kind of useless and half of my body is not,*" reflecting an interest to control the interface with their whole body, such as by walking toward and away from different parts of the display.

Participants suggested other collaborative scenarios involving complex data artifacts. While our approach elicited a range of application scenarios, we focus on three scenarios involving visually complex data artifacts:

Codebases. First, P2 described the challenges of remote code collaboration in Zoom meetings where one person shares their screen, and "*the other person doesn't understand what part* [of the code] *you are you talking about.*" They suggested that the ability to "*point out exactly to different places*" may accelerate this work, particularly given the many alternative visual representations of code bases and system logs.

Tables. Another suggestion from P2 reflects a need for data analysts to look at spreadsheets or tables together, an omnipresent category of artifact in data science workflows [7]: "*we could even work in Excel* […] *being able to see the other person and see how* [they] *select different parts of the screen* […] *I think that could be useful.*"

3D models. P8 suggested applying the Glass Chirolytics approach to remote conversations involving a 3D model, where either participant could pull apart or reassemble parts of it via coordinated gestures. They also envisioned applications in educational settings, such as those involving complex engineering drawings or anatomical models, where an intent is either to teach or learn about part-whole relationships.

6 Discussion

By reflecting on the Glass Chirolytics approach and our study results, we now present design implications for tools that aim to support synchronous and remote collaboration around data, propose new research directions, and discuss the limitations of our implementation and evaluation.

6.1 From Design Decisions to Design Implications

First, we revisit and extend our key decision decisions (**D1** — **D3**: Section 3.1) into design implications for future collaborative visualization applications.

Extend the mutual awareness of analytical intent with additional visual cues. Participants in our evaluation valued the visibility of nonverbal cues prioritized by our approach. Like in Tang et al.'s projection of remote collaborators' hands over a tabletop display [79], our approach also appeared to give collaborators insight into their remote peers' intents, insight that may lead to fewer collisions, such as simultaneous attempts to manipulate the same element. Given this awareness, it may not be necessary to employ locking mechanisms on interface elements, such as in cursor-based collaborative visualization approaches [71]. Nor would it be necessary to highlight recently manipulated elements, such as with ephemeral afterglow effects [8], because the remote collaborator's hand movements already accomplish this.

While we may not need to draw further attention to a remote collaborator's hands, it nevertheless may be fruitful to bring awareness to other nonverbal cues. Given our study participants' hesitation at times to use their hands for fear of unintentional interaction, we could track and communicate a remote collaborators' gaze [50], so as to determine if they are looking at the same elements as the local collaborator. Going further, we could recognize a remote collaborator's facial expressions as a means of inferring their affective state [63], which might be particularly helpful to communicate to a tutor, interviewer, or anyone aiming to reach a consensus decision with their peer. However, returning to **D1**, any accentuated nonverbal cues or inferences made from them should not detract from face-to-face conversation.

Accommodate visualization artifacts of further complexity. Across our scenarios (Section 4), the visualization elements reflected multidimensional, relational, and (geo-)spatial data abstractions. Although this scope might satisfy many collaborative visual analysis scenarios, truly satisfying **D2** means accommodating greater complexity. For two-dimensional visual representations, either collaborator should be able to load or paste any hierarchical SVG-based object on to the shared 'glass'. However, a complication creeps in when we begin to consider rich spatial data abstractions represented visually as 3D volumes, such as simulation models (as suggested by study participant P8) or isometric-perspective topographical maps. With 2D visuals, the reciprocal compositing over mirrored video produces a shared perspective, whereas with 3D visuals, a remote collaborator will appear as though they have a different perspective, thereby conflicting with **D1**. Despite this complication, we expect that aspects of Glass Chirolytics could be applied in collaboration around 3D visualization artifacts. At a minimum, we could accept that the local and remote participant will have different perspectives, looking at 3D content from opposing sides, such as in the Spatialstrates collaborative mixed reality ecosystem [11]. Alternatively, we could allow for toggling between multiple perspectives, akin to the Blended Whiteboard approach [36] for mixed reality collaboration, in which a face-to-face orientation facilitates rapid turn-taking in brainstorming, presenting, and sensemaking, while a side-by-side orientation ensures a common perspective on the content. The latter orientation would likely require compositing, segmenting, and juxtaposing video input from both participants. Future work could therefore study how participants freely toggle between orientations when performing open-ended collaborative analysis tasks with complex representations of spatial data.

Support the journey from exploration to communication and back again. Informal collaborative visualization is bound to alternate between exploration and explanation, between data analysis and storytelling with data [33]. This alternation might be particularly evident in pair analytics [2] scenarios in which there is role asymmetry between analyst and domain expert. A question this elicits is whether our gestural vocabulary, motivated by analytical use cases (**D3**), should be expanded to recognize and act on the rhetorical and expressive gestures associated with persuasive data presentations [37].

A larger gestural vocabulary might compromise the learnability of our approach, although it is possible, based on our study participants' comments, that watching others perform these gestures may accelerate learning. In particular, a local collaborator watching a remote peer who is more experienced with the gestural vocabulary may learn at a faster pace than our study participants, who were on equal footing at the beginning of the study. Nevertheless, if the gestural vocabulary of our approach is to be expanded, the temptation to introduce mode-switching (e.g., between idle, analysis, and storytelling modes) via keyboard or mouse commands is problematic because these mode switches are not likely to be visually apparent to remote collaborators in the same way that hand gestures are. Alternatively, recent work suggests that in lieu of a fixed gestural vocabulary, the meaning of gestures could be inferred in real time with approaches like the LLM-based GestureGPT [88], which links untrained gestures to predicted interface actions.

Our gestural vocabulary thus far maps to selecting elements, reconfiguring their placements, and exploring their spatial distribution, with some selection gestures triggering filter events and the generation of connections between elements. Other common interactions with conventional desktop visualization applications [87] do not have obvious gestural counterparts, such as changing the mapping of data attributes to visual variables, or changing the level of abstraction or detail. For these interactions, we might look to additional modalities such as speech input [75], sketch input [68], or interaction with tangible icons [77] to complement our existing set of deictic gestures.

We remarked in Section 4 how the scope of our approach focuses on synchronous and collaborative analysis and decision-making activities, and that these episodes form part of a larger fabric of data work, much of which being conducted asynchronously and individually. This invites the question of whether and how the scope of our approach could be expanded to support activities that precede our scenarios, such as data preparation and shaping interfaces or visualization and dashboard design, activities in which having another set of eyes (and an another pair of hands) could have downstream benefits.

6.2 Limitations

An obvious limitation of our approach is that it is intended for face-to-face conversations between two people with the same role, and thus it is unclear whether it is also suitable for small groups, such as a team of data scientists [15]. Solutions could involve relegating all but one or two members of the group to a passive status, in which they see a side-by-side compositing of the active speakers. Alternatively, we might adopt an approach like MirrorBlender [35],

in which participants can scale and position their video feeds in a unified composited display; however, this approach may limit the reach of any one group member's gestural control, as well as the mutual visibility of these gestures and other nonverbal communication. Expanding to larger groups may necessitate other forms of awareness of remote collaborator activity, such as via aggregated visual annotation layers on composited visualization and interface elements [20]. A limitation with this interaction approach is the lack of precision in mid-air hand gestures relative to mouse and keyboard input; as a consequence, interactive elements must be sufficiently large to ensure reliable manipulation.

Our evaluation also has four noteworthy limitations. First, while one or more of the four scenarios described in Section 4 are likely to be familiar to those with STEM education and work experiences, we opted to focus on a single scenario of collaborative decision making in our study, a scenario that may not occur with similar frequency for all of our participants. Second, we did not explicitly seek out participant pairs with existing levels of rapport and trust, and we acknowledge that these factors may impact the interpersonal awareness of nonverbal cues and collaborative working styles. Third, given the length of our evaluation and the time spent working with the Glass Chirolytics interface, we are unsure how fatiguing it would be to engage in longer conversations supported by our approach. Finally, the individual gestures in our vocabulary (Section 3.2) were not systematically evaluated with respect to their accuracy, precision, or memorability relative to alternative candidate gestures, and thus further experimental study is warranted.

6.3 Future Research Opportunities

Report on longitudinal use by analysts and tutors. We are eager to better understand any long-term benefits of our approach in scenarios spanning multiple usage sessions, such as exploratory data analysis and STEM tutoring. In particular, we call for investigations into whether this approach improves comprehension and information retention, and whether the dynamics and metrics of collaboration change over time. The impact of role and expertise asymmetries would also be worthy to study longitudinally. Meanwhile, in collaboration with cognitive scientists, we are curious to determine whether the observation of a conversation partner's noncommunicative and manipulative gestures has a learning benefit akin to what is exhibited with co-located communication through activation of the human mirror neuron system (hMNS) [24].

Investigate asymmetries of technology. Another question is how the collaborative experience changes in cases where one collaborator casts their video and interacts via gesture while the other remains off-camera and interacts via standard input devices, or if they collaborate via avatar while wearing a head-mounted display [80]. Collaborators with large displays [90] and multiple webcams also introduce asymmetries. For instance, those working from large rooms outfitted with large displays could interact via full-body gestures [70], as suggested by study participants P3 and P4. If their peers are limited to small displays and conventional distances from their cameras [4], future adaptations of our approach must reconcile disparities in the scale of composited elements and the reach of gestures. Unsurprisingly, our study participants reported that the physical demand imposed by the Glass Chirolytics interface was significantly higher than that of the baseline interface. Using multiple webcams might alleviate this demand by positioning one webcam in a conventional position to capture the face and another positioned above the work surface in front of the monitor, pointing down to capture the hands gesturing on that surface. This camera perspective has made for compelling data journalism with hand-drawn charts [29], which suggests that it could also be a viable perspective for compositing visual aids. Furthermore, this approach may represent a middle ground between conventional screen sharing and Glass Chirolytics, one that retains some benefits of nonverbal communication without inducing considerable levels of fatigue.

Augment our approach with speech-generated visual aids. Finally, our approach could be complemented by the addition of speech input. Speech input could, for instance, be used to trigger mode-switching, particularly as mode-switching gestures expand the gestural vocabulary and may be distracting to the remote observer. Beyond the aforementioned prospect of using speech to manipulate existing composited visual aids [75], conversation speech could also be used to generate, retrieve, and recommend other context-appropriate visual aids. With systems like Liu et al.'s Visual Captions [55] or Xia et al.'s CrossTalk [85] bringing contextually-appropriate images and documents into the conversation, we are hopeful that similar approaches could be taken to generate contextually-appropriate, valid, and interactive data visualization artifacts for the shared glass.

7 Conclusion

We introduced Glass Chirolytics, an augmented videoconferencing approach designed to foster engaging and multimodal face-to-face analytical conversations involving complex data abstractions. It employs a reciprocal compositing of shared visualization artifacts over the mirrored webcam video of one's conversation partner, which we pair with bimanual mid-air gestures for manipulating visualization and interface elements. These design decisions bring collaborators' hands back into the conversation by virtue of them appearing in the video frame: a step toward restoring the benefits of nonverbal communication lost in videoconference meetings relative to co-located meetings.

We evaluated our approach with eight pairs of participants in a comparative study, in which participants completed travel itinerary decision-making tasks given a large flight schedule dataset, both with a Glass Chirolytics application and with a baseline application reflecting conventional videoconferencing with mouse control of a shared interface. Our findings suggest that our approach significantly enhances a sense of presence while reducing the temporal demand of collaborative analysis. Participants' remarks also suggest that our approach provided an awareness of the analytical intent of one's conversation partner, an awareness that was lacking in the baseline experience. Given these results, we are optimistic about the potential of gesture-controlled applications for enriching remote collaboration around data, and for accelerating learning in cases where there is an asymmetry of knowledge and tool experience between conversation partners.

Acknowledgments

We thank Anchit Mishra, Mohammad Abolnejadian, Jenny Zhang, Jessica Chen, and Skylar Ji for their feedback on the project. This research was supported by a University of Waterloo Cheriton School of Computer Science Undergraduate Research Fellowship and a Natural Sciences and Engineering Research Council of Canada (NSERC) Undergraduate Student Research Award.

References

[1] Apple. 2025. Use Reactions, Presenter Overlay, and other effects when videoconferencing on Mac. support.apple.com/105117.

[2] Richard Arias-Hernandez, Linda T Kaastra, Tera M Green, and Brian Fisher. 2011. Pair analytics: Capturing reasoning processes in collaborative visual analytics. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*. doi:10.1109/HICSS.2011.339

[3] Sriram Karthik Badam, Andreas Mathisen, Roman Rädle, Clemens N Klokmose, and Niklas Elmqvist. 2019. Vistrates: A component model for ubiquitous analytics. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 25, 1 (2019). doi:10.1109/TVCG.2018.2865144

[4] Jeremy N Bailenson. 2021. Nonverbal overload: A theoretical argument for the causes of Zoom fatigue. *Technology, Mind, and Behavior* 2, 1 (2021). doi:10.1037/tmb0000030

[5] Aruna D. Balakrishnan, Susan R. Fussell, and Sara Kiesler. 2008. Do visualizations improve synchronous remote collaboration?. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/1357054.1357246

[6] José María Barrero, Nicholas Bloom, and Steven J Davis. 2023. The evolution of work from home. *Journal of Economic Perspectives* 37, 4 (2023). doi:10.1257/jep.37.4.23

[7] Lyn Bartram, Michael Correll, and Melanie Tory. 2022. Untidy data: The unreasonable effectiveness of tables. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 28, 1 (2022). doi:10.1109/TVCG.2021.3114830

[8] Patrick Baudisch, Desney Tan, Maxime Collomb, Dan Robbins, Ken Hinckley, Maneesh Agrawala, Shengdong Zhao, and Gonzalo Ramos. 2006. Phosphor: Explaining transitions in the user interface using afterglow effects. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*. doi:10.1145/1166253.1166280

[9] BBC. 2010. Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats. https://www.youtube.com/watch?v=jbkSRLYSojo

[10] Marcel Borowski, Peter WS Butcher, Janus Bager Kristensen, Jonas Oxenbøll Petersen, Panagiotis D Ritsos, Clemens N Klokmose, and Niklas Elmqvist. 2025. DashSpace: A live collaborative platform for immersive and ubiquitous analytics. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 31, 10 (2025). doi:10.1109/TVCG.2025.3537679

[11] Marcel Borowski, Jens Emil Sloth Grønbæk, Peter W. S. Butcher, Panagiotis D. Ritsos, Clemens Nylandsted Klokmose, and Niklas Elmqvist. 2025. Spatialstrates: Cross-reality collaboration through spatial hypermedia. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*. doi:10.1145/3746059.3747708

[12] Mike Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 17, 12 (2011). doi:10.1109/TVCG.2011.185

[13] Doug Bowman, Chadwick Wingrave, Joshua Campbell, and Vinh Ly. 2001. Using Pinch Gloves (TM) for both natural and abstract interaction techniques in virtual environments. Technical Report TR-01-23, Computer Science, Virginia Tech. eprints.cs.vt.edu/archive/00000547.

[14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006). doi:10.1191/1478088706qp063oa

[15] Matthew Brehmer and Robert Kosara. 2022. From jam session to recital: Synchronous communication and collaboration around data in organizations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 28, 1 (2022). doi:10.1109/TVCG.2021.3114760

[16] Camelia D Brumar, Sam Molnar, Gabriel Appleby, Kristi Potter, and Remco Chang. 2025. A typology of decision-making tasks for visualization. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 31, 10 (2025). doi:10.1109/TVCG.2025.3572842

[17] Yining Cao, Rubaiat Habib Kazi, Li-Yi Wei, Deepali Aneja, and Haijun Xia. 2024. Elastica: Adaptive live augmented presentations with elastic mappings across modalities. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3613904.3642725

[18] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 43, 01 (2021). doi:10.1109/TPAMI.2019.2929257

[19] Peter Cappelli. 2021. *The Future of the Office: Work from Home, Remote Work, and the Hard Choices, We All Face*. Wharton School Press.

[20] John Joon Young Chung, Hijung Valentina Shin, Haijun Xia, Li-yi Wei, and Rubaiat Habib Kazi. 2021. Beyond show of hands: Engaging viewers via expressive and scalable visual communication in live streaming. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3411764.3445419

[21] Sharice Clough and Melissa C Duff. 2020. The role of gesture in communication and cognition: Implications for understanding and treating neurogenic communication disorders. *Frontiers in Human Neuroscience* 14 (2020). doi:10.3389/fnhum.2020.00323

[22] Maxime Cordeil, Anais Servais, Guillaume Truong, Tim Dwyer, Dhaval Vyas, and Christophe Hurter. 2025. The presenter in the browser: Design and evaluation of human interactive overlays with web content. *Multimodal Technologies and Interaction* 9, 2 (2025). doi:10.3390/mti9020010

[23] Josh Urban Davis, Paul Asente, and Xing-Dong Yang. 2023. Multimodal direct manipulation in video conferencing: challenges and opportunities. In *Proceedings of the ACM Conference on Designing Interactive Systems (DIS)*. doi:10.1145/3563657.3596099

[24] Kelly Dickerson, Peter Gerhardstein, and Alecia Moser. 2017. The role of the human mirror neuron system in supporting communication in a digital world. *Frontiers in Psychology* 8 (2017). doi:10.3389/fpsyg.2017.00698

[25] Evanthia Dimara, Harry Zhang, Melanie Tory, and Steven Franconeri. 2022. The unmet data visualization needs of decision makers within organizations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 28, 12 (2022). doi:10.1109/TVCG.2021.3074023

[26] Marion Dohen and Benjamin Roustan. 2017. Co-production of speech and pointing gestures in clear and perturbed interactive tasks: Multimodal designation strategies. In *Proceedings of the Conference of the International Speech Communication Association*. hal.science/hal-02367749/.

[27] Douglas C Engelbart and William K English. 1968. A research center for augmenting human intellect. In *Proceedings of Fall Joint Computer Conference (AFIPS)*. doi:10.1145/1476589.1476645

[28] Temiloluwa Paul Femi-Gege, Matthew Brehmer, and Jian Zhao. 2024. VisConductor: Affect-varying widgets for animated data storytelling in gesture-aware augmented video presentation. *Proceedings of the ACM on Human-Computer Interaction* 8, ISS (2024). doi:10.1145/3698131

[29] Joss Fong. 2022. How American conservatives turned against the vaccine. Vox News. youtu.be/sv0dQfRRrEQ.

[30] Susan Goldin-Meadow, Susan Wagner Cook, and Zachary A Mitchell. 2009. Gesturing gives children new ideas about math. *Psychological science* 20, 3 (2009). doi:10.1111/j.1467-9280.2009.02297.x

[31] Jiangtao Gong, Jingjing Sun, Mengdi Chu, Xiaoye Wang, Minghao Luo, Yi Lu, Liuxin Zhang, Yaqiang Wu, Qianying Wang, and Can Liu. 2023. Side-by-side vs face-to-face: Evaluating colocated collaboration via a transparent wall-sized display. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023). doi:10.1145/3579623

[32] Google. 2024. WebRTC: Real-Time Communication for the Web. https://webrtc.org/.

[33] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Nicola Cosgrove, and Marc Streit. 2016. From visual exploration to storytelling and back again. *Computer Graphics Forum* 35, 3 (2016). doi:10.1111/cgf.12925

[34] Jens Emil Sloth Grønbæk, Marcel Borowski, Eve Hoggan, Wendy E. Mackay, Michel Beaudouin-Lafon, and Clemens Nylandsted Klokmose. 2023. Mirrorverse: Live tailoring of video conferencing interfaces. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*. doi:10.1145/3586183.3606767

[35] Jens Emil Sloth Grønbæk, Banu Saatçi, Carla F. Griggio, and Clemens Nylandsted Klokmose. 2021. MirrorBlender: Supporting hybrid meetings with a malleable video-conferencing system. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3411764.3445698

[36] Jens Emil Sloth Grønbæk, Juan Sánchez Esquivel, Germán Leiva, Eduardo Velloso, Hans Gellersen, and Ken Pfeuffer. 2024. Blended Whiteboard: Physicality and reconfigurability in remote mixed reality collaboration. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3613904.3642293

[37] Brian D. Hall, Lyn Bartram, and Matthew Brehmer. 2022. Augmented chironomia for presenting data to remote audiences. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*. doi:10.1145/3526113.3545614

[38] Chang Han and Katherine E Isaacs. 2025. A deixis-centered approach for documenting remote synchronous communication around data visualizations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 31, 1 (2025). doi:10.1109/TVCG.2024.3456351

[39] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Human Mental Workload* 52 (1988). doi:10.1016/S0166-4115(08)62386-9

[40] Masoumehsadat Hosseini, Tjado Ihmels, Ziqian Chen, Marion Koelle, Heiko Müller, and Susanne Boll. 2023. Towards a consensus gesture set: A survey of

mid-air gestures in HCI for maximized agreement across domains. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3544548.3581420

[41] Erzhen Hu, Jens Emil Sloth Grønbæk, Austin Houck, and Seongkook Heo. 2023. OpenMic: Utilizing proxemic metaphors for conversational floor transitions in multiparty video meetings. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3544548.3581013

[42] Petra Isenberg, Niklas Elmqvist, Jean Scholtz, Daniel Cernea, Kwan-Liu Ma, and Hans Hagen. 2011. Collaborative visualization: Definition, challenges, and research agenda. *Information Visualization* 10, 4 (2011). doi:10.1177/1473871611412817

[43] Hiroshi Ishii and Minoru Kobayashi. 1992. ClearBoard: a seamless medium for shared drawing and conversation with eye contact. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. doi:10.1145/142750.142977

[44] Kevin Jahns. 2023. Yjs: A CRDT Framework for Building Collaborative Applications. https://github.com/yjs/yjs.

[45] Azadeh Jamalian, Valeria Giardino, and Barbara Tversky. 2013. Gestures for thinking. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 35. escholarship.org/uc/item/0zk7z5h9.

[46] Robert Johansen. 1988. *Groupware: Computer Support for Business Teams*. The Free Press.

[47] Seokmin Kang and Barbara Tversky. 2016. From hands to minds: Gestures promote understanding. *Cognitive Research: Principles and Implications* 1, 1 (2016). doi:10.1186/s41235-016-0004-9

[48] Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhliarchuk. 2024. HaGRID – HAnd Gesture Recognition Image Dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. doi:10.1109/WACV57701.2024.00451

[49] Adam Kendon. 2004. *Gesture Visible Action as Utterance*. Cambridge University Press. doi:10.1017/CBO9780511807572

[50] Maurice Koch, Nelusa Pathmanathan, Daniel Weiskopf, and Kuno Kurzhals. 2026. A multimodal framework for understanding collaborative design processes. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 32, 1 (2026). doi:10.1109/TVCG.2025.3634232

[51] Robert Kosara. 2016. Presentation-oriented visualization techniques. *IEEE Computer Graphics and Applications (CG&A)* 36, 1 (2016). doi:10.1109/MCG.2016.2

[52] Adrian Kristanto, Maxime Cordeil, Benjamin Tag, Nathalie Henry Riche, and Tim Dwyer. 2023. Hanstreamer: An open-source webcam-based live data presentation system. In *Proceedings of MERCADO Workshop at IEEE VIS 2023: Multimodal Experiences for Remote Communication Around Data Online*. https://arxiv.org/abs/2309.12538

[53] Turgay Kucuk. 2023. The power of body language in education: A study of teachers' perceptions. *International Journal of Social Sciences and Educational Studies* 10 (2023). doi:10.23918/ijsses.v10i3p275

[54] Jian Liao, Adnan Karim, S. Jadon, Rubaiat Habib Kazi, and Ryo Suzuki. 2022. RealityTalk: Real-time speech-driven augmented presentation for AR live storytelling. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*. doi:10.1145/3526113.3545702

[55] Xingyu "Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, and Ruofei Du. 2023. Visual captions: Augmenting verbal communication with on-the-fly visuals. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3544548.3581566

[56] Matthew Lombard, Theresa B Ditton, and Lisa Weinstein. 2009. Measuring presence: The Temple Presence Inventory. In *Proceedings of the Annual International Workshop on Presence*.

[57] Narges Mahyar and Melanie Tory. 2014. Supporting communication and coordination in collaborative sensemaking. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 20, 12 (2014). doi:10.1109/TVCG.2014.2346573

[58] Meta. 2023. React: A JavaScript Library for Building User Interfaces. https://react.dev/.

[59] Microsoft Teams 2022. Microsoft Teams. https://www.microsoft.com/en-us/microsoft-teams/group-chat-software.

[60] mmhmm inc. 2025. Airtime. https://airtimetools.com/.

[61] Daniel R Montello. 2003. Regions in Geography: Process and Content. In *Foundations of Geographic Information Science*, Matt Duckham and Michael F Goodchild (Eds.). CRC Press.

[62] Tamara Munzner. 2014. *Visualization Analysis and Design*. CRC Press.

[63] Prasanth Murali, Javier Hernandez, Daniel McDuff, Kael Rowan, Jina Suh, and Mary Czerwinski. 2021. AffectiveSpotlight: Facilitating the communication of affective responses from audience members during online presentations. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3411764.3445235

[64] Rupayan Neogy, Jonathan Zong, and Arvind Satyanarayan. 2020. Representing real-time multi-user collaboration in visualizations. In *Proceedings of the IEEE Visualization Conference (VIS)*. doi:10.1109/VIS47514.2020.00036

[65] OBS 2022. OBS Studio. https://obsproject.com.

[66] Observable. 2025. Observable Canvases. https://observablehq.com/platform/canvases.

[67] Ken Perlin. 2018. Chalktalk in Augmented Reality. vimeo.com/232230096.

[68] Ken Perlin, Zhenyi He, and Karl Rosenberg. 2018. Chalktalk: A visualization and communication language – As a tool in the domain of computer science education. In *SPLASH LIVE workshop 2018*. doi:10.48550/arXiv.1809.07166

[69] Xun Qian, Feitong Tan, Yinda Zhang, Brian Moreno Collins, David Kim, Alex Olwal, Karthik Ramani, and Ruofei Du. 2024. ChatDirector: Enhancing video conferencing with space-aware scene rendering and speech-driven layout transition. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3613904.3642110

[70] Nazmus Saquib, Rubaiat Habib Kazi, Li-yi Wei, and Wilmot Li. 2019. Interactive body-driven graphics for augmented video performance. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3290605.3300852

[71] Michail Schwab, David Saffo, Yixuan Zhang, Shash Sinha, Cristina Nita-Rotaru, James Tompkin, Cody Dunne, and Michelle A Borkin. 2021. VisConnect: Distributed event synchronization for collaborative visualization. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 27, 02 (2021). doi:10.1109/TVCG.2020.3030366

[72] Jonathan Schwabish. 2016. *Better Presentations: A Guide for Scholars, Researchers, and Wonks*. Columbia University Press.

[73] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. 2013. Real-time human pose recognition in parts from single depth images. *Commun. ACM* 56, 1 (2013). doi:10.1145/2398356.2398381

[74] Sidney Siegel. 1957. Nonparametric Statistics for the Behavioral Sciences. *The Journal of Nervous and Mental Disease* 125, 3 (1957).

[75] Arjun Srinivasan and Matthew Brehmer. 2023. Combining voice and gesture for presenting data to remote audiences. In *Proceedings of MERCADO Workshop at IEEE VIS 2023: Multimodal Experiences for Remote Communication Around Data Online*.

[76] Mark Stefik, Daniel G Bobrow, Gregg Foster, Stan Lanning, and Deborah Tatar. 1987. WYSIWIS revised: Early experiences with multiuser interfaces. *ACM Transactions on Information Systems* 5, 2 (1987). doi:10.1145/27636.28056

[77] Kentaro Takahira, Wong Kam-Kwai, Leni Yang, Xian Xu, Takanori Fujiwara, and Huamin Qu. 2025. TangibleNet: Synchronous network data storytelling through tangible interactions in augmented reality. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3706598.3714265

[78] Kentaro Takahira, Yue Yu, Takanori Fujiwara, Ryo Suzuki, and Huamin Qu. 2025. InSituTale: Enhancing augmented data storytelling with physical objects. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*. doi:doi.org/10.1145/3746059.3747678

[79] Anthony Tang, Michel Pahud, Kori Inkpen, Hrvoje Benko, John C. Tang, and Bill Buxton. 2010. Three's company: understanding communication channels in three-way distributed collaboration. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*. doi:10.1145/1718918.1718969

[80] Wai Tong, Meng Xia, Kam Kwai Wong, Doug A Bowman, Ting-Chuen Pong, Huamin Qu, and Yalong Yang. 2023. Towards an understanding of distributed asymmetric collaborative visualization on problem-solving. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. doi:10.1109/VR55154.2023.00054

[81] Julien Valentin, Adarsh Kowdle, Suhil Nassar, Ameesh Makadia, Angjoo Kanazawa, Michael Ratasich, and Caroline Pantofaru. 2020. MediaPipe: A framework for building perception pipelines. doi:10.48550/arXiv.1906.08172

[82] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication* 57 (2014). doi:10.1016/j.specom.2013.09.008

[83] Frank Wilcoxon. 1992. *Individual Comparisons by Ranking Methods*. Springer.

[84] Haijun Xia, Ken Hinckley, Michel Pahud, Xiao Tu, and Bill Buxton. 2017. WritLarge: Ink unleashed by unified scope, action, & zoom. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3025453.3025664

[85] Haijun Xia, Tony Wang, Aditya Gunturu, Peiling Jiang, William Duan, and Xiaoshuo Yao. 2023. CrossTalk: Intelligent substrates for language-oriented interaction in video-based communication and collaboration. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*. doi:10.1145/3586183.3606773

[86] Longqi Yang, David Holtz, Sonia Jaffe, Siddharth Suri, Shilpi Sinha, Jeffrey Weston, Connor Joyce, Neha Shah, Kevin Sherman, Brent Hecht, et al. 2022. The effects of remote work on collaboration among information workers. *Nature Human Behavior* 6, 1 (2022). doi:10.1038/s41562-021-01196-4

[87] Ji Soo Yi, Youn ah Kang, John Stasko, and Julie A Jacko. 2007. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 13, 6 (2007). doi:10.1109/TVCG.2007.70515

[88] Xin Zeng, Xiaoyu Wang, Tengxiang Zhang, Chun Yu, Shengdong Zhao, and Yiqiang Chen. 2024. GestureGPT: Toward zero-shot free-form hand gesture understanding with large language model agents. *Proceedings of the ACM on Human-Computer Interaction* 8, ISS (2024). doi:10.1145/3698145

[89] Nan Zhao, Xian Zhang, J Adam Noah, Mark Tiede, and Joy Hirsch. 2023. Separable processes for live in-person and live "Zoom-like" faces. *Imaging Neuroscience* 1 (2023). doi:10.1162/imag_a_00027

[90] Jakob Zillner, Christoph Rhemann, Shahram Izadi, and Michael Haller. 2014. 3D-board: a whole-body remote collaborative whiteboard. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*. doi:10.1145/2642918.2647393

[91] Zoom 2022. Zoom. https://zoom.us/.