# A Deixis-Centered Approach for Documenting Remote Synchronous Communication around Data Visualizations

Chang Han and Katherine E. Isaacs

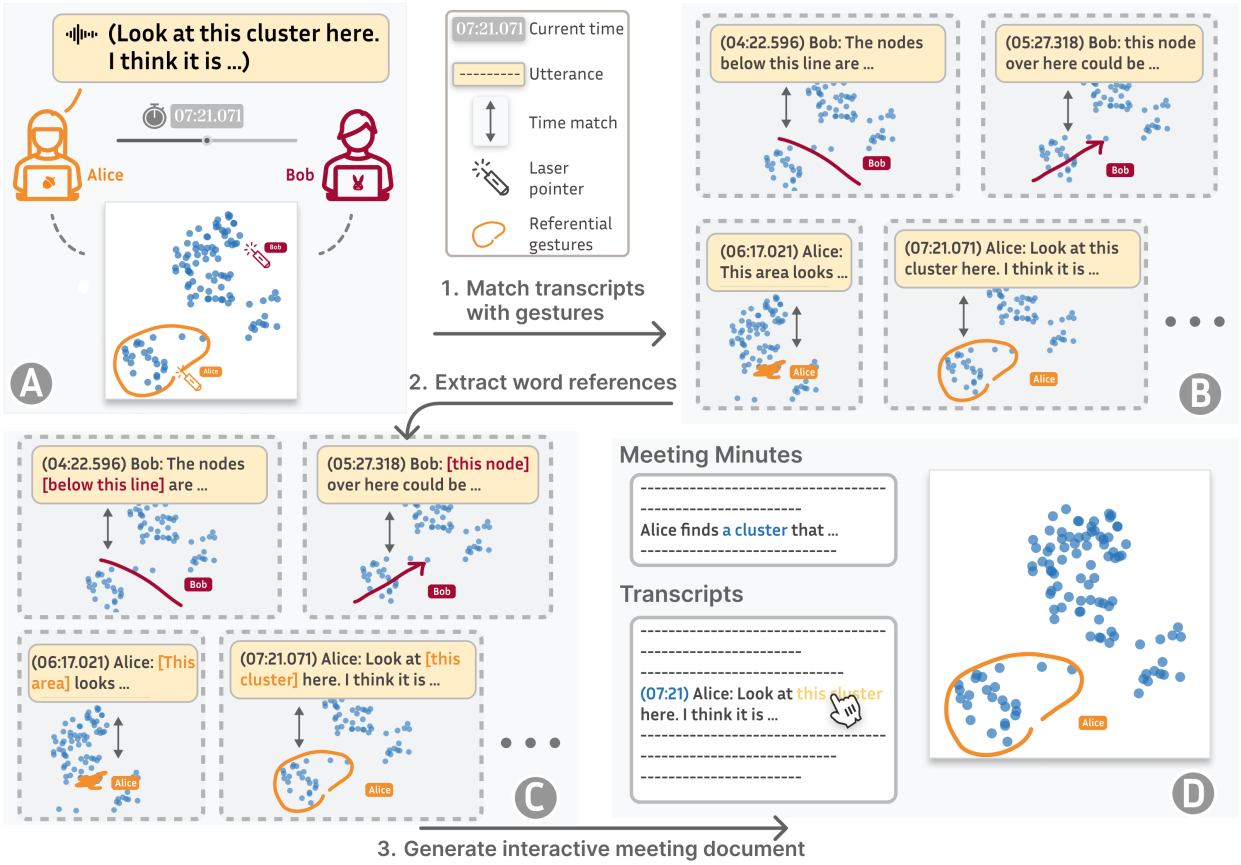



Fig. 1: An overview of the proposed framework. It has four stages: (A) *Data collection* that captures both deictic and verbal portions of the meetings. (B) *Utterance matching* that associates referential gestures with the corresponding transcribed utterances. (C) *Reference Extraction* that further extracts connections within the matched utterance. (D) *Interactive Notes Generation* that represents collaborative data meetings with automatically generated transcripts and meeting minutes, both augmented with annotations.

**Abstract**— Referential gestures, or as termed in linguistics, *deixis*, are an essential part of communication around data visualizations. Despite their importance, such gestures are often overlooked when documenting data analysis meetings. Transcripts, for instance, fail to capture gestures, and video recordings may not adequately capture or emphasize them. We introduce a novel method for documenting collaborative data meetings that treats deixis as a first-class citizen. Our proposed framework captures cursor-based gestural data along with audio and converts them into interactive documents. The framework leverages a large language model to identify word correspondences with gestures. These identified references are used to create context-based annotations in the resulting interactive document. We assess the effectiveness of our proposed method through a user study, finding that participants preferred our automated interactive documentation over recordings, transcripts, and manual note-taking. Furthermore, we derive a preliminary taxonomy of cursor-based deictic gestures from participant actions during the study. This taxonomy offers further opportunities for better utilizing cursor-based deixis in collaborative data analysis scenarios.

**Index Terms**—Taxonomy, Models, Frameworks, Theory ; Collaboration ; Communication/Presentation, Storytelling

## 1 Introduction

When people meet to present and discuss data and data visualizations, significant communication occurs through referential gestures such as pointing and indications of movement and flow [35]. These gestures enrich the *context* within which statements are made, playing a crucial role in shaping the meaning of those statements. This meaning-through-context is referred to as *deixis* by the linguistics community [57].

• *Chang Han and Katherine E. Isaacs are with the University of Utah. E-mail: {changhan, kisaacs}@sci.utah.edu.*

*Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx*

Among the various forms of communication, deixis holds a particularly important role in interchanges concerning visualizations [27], providing clarity and depth to the explanation of insights. As shown in Figure 1(A), someone may verbally refer to "this cluster" while using a laser pointer to lasso points in the scatterplot for others to observe. In this example, verbal expression alone would fail to convey a clear interpretation without the accompanying referential gestures.

Despite being essential in communication around data, deixis is often overlooked when documenting collaborative visualization meetings. Traditional methods of documenting meetings each have their limitations: transcripts do not store referential gestures, while screen and video recordings may be missing pointer movements or lack fidelity to understand context from them. Additionally, manual note-taking can be distracting and must often be selective in what context it captures to keep up with the pace of the meeting [51]. Existing methods of documenting insights during visual analysis are either not designed for synchronous communication [26, 59] or demand significant manual annotation and input [11, 64], posing considerable challenges for their use in highly interactive settings like collaborative visualization meetings.

To facilitate future examination and review of communications around data, we propose a framework to document collaborative visualization meetings that underscores the importance of deixis. We focus particularly on online meetings, where people synchronously discuss data visualizations in a distributed setting. In video conferencing, gestures to a digital visualization are often performed with the mouse or touch pointer. Video conferencing software [43, 65] further allows people to annotate what is shown on screen with tools like the virtual laser pointer or pencil tool while interactive visualizations have deictic interactions such as brushing and highlighting [5, 41]. We design our framework to capture these pointer-based deictic behaviors, storing them alongside the utterances.

Participants in virtual synchronous meetings *encode*, in the semiotics sense [10, 23], their deictic communication with a combination of audio and pointer gestures. To better understand how we can identify these communication pairs, we first conduct a formative study of the use of pointer-based gestures in online synchronous meetings around data. We use the findings to inform the design of our pipeline. We associate recorded gestures from annotation tools and visualization interactions with audio transcripts based on both temporal overlap and utterance semantics, utilizing the in-context learning capability of a large language model (LLM) [18, 44] to perform the latter. With these matched gestures and dialogue, we generate an interactive document that provides both meeting minutes, a compressed narrative of the meeting, and the full transcript, both linked with their matching referential gestures represented as animated annotations or interaction states.

To evaluate the effectiveness of our approach, we conducted a second user study involving our developed prototype, which is available on Github[1]. Overall, participants acknowledged the usefulness of our generated documents and expressed a preference for using them for meeting documentation over screen recording and audio transcription methods. Given that our study replicates gestural tools commonly used in collaborative meeting systems, we believe that the proposed concepts and design principles can be readily integrated into existing systems.

Our framework aims to *identify* deictic communication and re-encode it into documentation without decoding or changing the meaning, expecting the user to decode as they would have in the meeting. Recognizing the potential of decoding these deictic communications, we code the observed gestures from both of our studies and derive the first taxonomy for referential gestures in online synchronous communication. Though preliminary, this taxonomy can aid in understanding deictic pointing behaviors in online communication and provide a foundation for future projects that seek to recognize, capture, and design for such gestures.

In summary, our contributions are two fold: (1) A framework for generating deixis-informed interactive notes from synchronous online audio-visual meetings and (2) A preliminary taxonomy of pointer-based deictic gestures used in collaboratively exploring data visualizations.

[1] https://github.com/hconhisway/vitraexample

## 2 Background and Related Work

We present relevant background in deixis, referential gestures, communication, and collaborative visualization. Then we discuss related work regarding insight documentation in visual analysis and methods for connecting text with visual components.

### 2.1 Background on Deixis and Referential Gestures

In linguistics, *deixis* refers to the phenomenon where the meaning of certain words is dependent on the context in which they are used [57]. For example, the statement "that apple tree over there" cannot be fully understood without observing the speaker's index finger pointing towards the tree. Deixis can take on different forms, such as personal, spatial, and temporal. We focus on spatial deixis, where language is used to indicate location or direction within the speaker's contextual environment. We use the term *referential gestures* interchangeably with deictic pointing behavior.

In the encoder/decoder model of Stuart Hall [23], the communicator encodes their *meaning* as a message which is then *decoded* into a perceived meaning by the receiver. In our case, attention is directed through a combination of speech and gesture. We seek to identify and document both parts of the encoded message, essentially re-encoding the message for later decoding by the the meeting participants.

Previous work has underscored the critical role of referential gestures in human communication. Through surveying various types of communicative gestures, Clark [13] categorized them into two categories: pointing and placing. Pointing directs attention towards an object or location, while placing involves positioning objects in a way that communicates meaning. Clark argues these nonverbal actions are foundational to the way we convey and interpret meaning. Brennan et al. [7] conducted an experiment with paired participants working together to analyze a map. By analyzing the movement of mouse cursors, they concluded that successful conversation depends on the dynamic interplay of both verbal and non-verbal (i.e., visual) cues. Building upon these works, our study acknowledges the critical role of non-verbal cues in communication. Particularly, we focused on the context of communication around visual representations.

Hill and Hollan [27] explored how deictic pointing behaviors encompass more than the simple directive of "look here,"A discussing various hand gestures that convey diverse meanings. Heer et al. [24] further investigated the application of spatial referential gestures within the domain of asynchronous collaborative visualization. They distinguished between two primary types of referential gestures in spatial contexts. The first type encompasses brushing and dynamic querying, which are directly tied to data. They can support various automated tasks [63] and are applicable through different data views. The second type, *graphical annotations* are more expressive, but are view-dependent due to their lack of data awareness. Building upon the taxonomy proposed by previous studies and our observations, we identify three types of referential gestures pertinent to remote synchronous communication around data visualizations: transient gestures expressed with a laser pen [48, 49], durable annotations made with a pencil tool [17, 26], and manipulations of the visual interface activated by mouse actions [22].

A critical issue related to referential gestures is ambiguity. In some scenarios, people may successfully communicate without using gestures. Clark et al. [14] demonstrated how ambiguity resolution is influenced by the familiarity among individuals. For instance, two individuals sharing a common understanding about certain flowers might effectively communicate by simply referring to "this flower," while a third party, lacking this shared context, might be confused. The ambiguity problem also poses a threat to our documentation framework, as we cannot document information that is not explicitly expressed and relies on external context. We also discuss this limitation in Section 7.

### 2.2 Background on Collaborative visualization

*Collaborative visualization* is defined 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"* [28]. Based on early works in the Computer-Supported Collaborative Work (CSCW) community [1, 30], collabo-

rative visualization scenarios are characterized by space (co-located vs. distributed) and time (asynchronous vs. synchronous) axes [28, 58]. When designing our documentation framework, we focus on synchronous distributed collaborations, with an emphasis on the critical communication and discussion phases. Specifically, we envision small collaborative team meetings, described as "Jam Sessions" by Brehmer and Kosara [6], as primary use cases for our framework.

Although our work does not aim to directly enhance collaborative visualization, we built the collaborative interface we use for data collection (Section 3.1) based on the previous efforts of collaborative visualization [2, 4, 33, 56]. For example, VisConnect [56] introduces support for synchronizing low-level events in web-based interactive visualizations for collaboration. Neogy et al. [46] explored the design space of interfaces for remote synchronous collaborative visualizations, with the aim of enabling effective collaboration among users while preserving the autonomy necessary for independent work. We use these works to inform our testbed so that our framework operates on ecologically valid scenarios.

### 2.3 Other Related Work

**Insight documentation in visual analysis.** Previous research has recognized the important role of notes made during visual analysis [39], though not in the context of remote synchronous collaborative visualization. In these other contexts, observed benefits included that analysts can review insights previously recorded and use the records as materials for deeper analysis and organization [39, 40]. Apart from taking notes in separate medium, annotating visualizations can also be considered a form of knowledge externalization and can also facilitate the analysis process [64]. Kim et al. [34] found that free-form annotations, as opposed to note-taking in a separate medium, may encourage participants to maintain a state of "flow," which is associated with heightened creativity. Walny et al. [60] also used a free-form pen in their study to ask participants to draw on visualizations while reading. Their results suggest that active reading behaviors transfer from documents to visualizations. We draw an analogy between our provided pencil tool (shown in Section 3.1) and the free-form pen in non-collaborative scenarios. Our framework also aims to document insights from visualizations. We focus on remote synchronous meetings as the primary use case, prioritizing the automatic generation of notes, as manual note-taking can be challenging in such highly interactive scenarios.

**Connecting text to visual media.** In the digital era, many documents produced and consumed online incorporate both textual and visual elements. Many works have sought to harness computational technologies to enhance the readability and user experience of online document text. For example, Kong et al. [36] introduced a crowdsourcing pipeline for extracting references between texts and charts and an interactive application that highlights these correspondences on selection. Kim et al. [32] developed an interactive document reader that automatically links document text with corresponding spreadsheet cells. This linking was shown to enhance users' ability to match text with tables. Badam et al. [3] further introduced a contextual visualization technique that can automatically link text contents with tables and provide visualizations based on the reader's current focus. Kori [38] presented a mixed-initiative approach for building references between charts and text through both manual input and algorithmic suggestions. EmphasisChecker [31] can detect and highlight mismatches in emphasis between line charts and textual descriptions. In addition to utilizing visualization to enhance the reading experience, textual descriptions (annotations) are also employed to augment visualizations. Lai et al. [37] developed a pipeline that uses computer vision techniques to automatically annotate visualizations based on provided textual descriptions. Similarly, ChartText [50] links text with statistical charts. They discussed a use case of automatically annotating charts to a presenter's description during a video conference as motivation, but did not demonstrate it in practice. Our work resonates with previous efforts to connect text to visual media, but targets a different type of reference between text and media: those encoded by participants in virtual meetings, expressed through a combination of deictic references and spoken words.

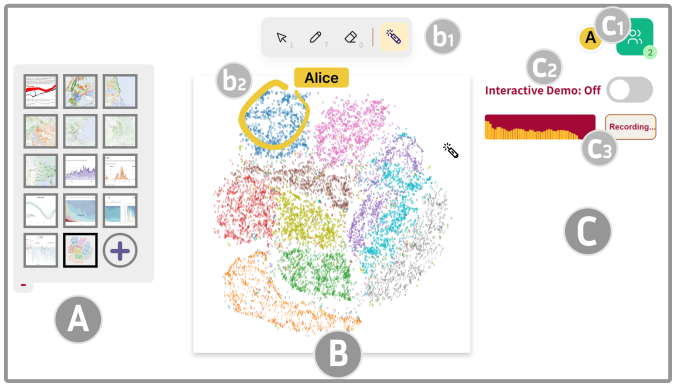



Fig. 2: An overview of the collaborative interface, serving to enable collaborative visualization and data collections. It consists of (A) a gallery of meeting materials, (B) a collaborative board, and (C) room controls.

## 3 Formative Study & Design Rationale

To inform our design, we conducted a formative study of how people use referential gestures in remote synchronous meetings around visualizations. We first designed and implemented a collaborative interface for use in this study, representing such intefaces in general, based on a survey of existing collaborative working tools. Then, we recruited four participants with expertise in the design and analysis of visualizations. Participants were asked to share and discuss their own visualization projects as well as visualizations they found engaging through the collaborative interface. These visualizations included common charts (e.g., Gantt charts, line charts, scatter plots), scientific visualizations (including volume rendering results), and visual analytics systems. We detail the findings from the observations and discussions, and explain how they influenced our design decisions. We first describe our collaborative interface. Then we present the findings of our formative study and the design guidance we interpreted from them.

### 3.1 Representative Collaborative Interface

In designing our representative interface to collect deictic gestures, we encountered limited discussion and guidance regarding how they are expressed in remote environments. Early works focused on the use of the mouse pointer to direct viewer's intentions [7, 27]. However, the volume of data generated by the trajectory of mouse movements is substantial and does not always indicate a referential gesture, limiting its capacity for preserving records.

We surveyed how referential gestures are expressed in popular collaborative working tools, including videoconferencing tools, such as Zoom [65] and Microsoft Teams [43], and whiteboard tools, such as Mural [45] and Excalidraw [20]. We found that referential gestures are often expressed using a laser pointer metaphor. Compared to the mouse pointer, the virtual laser pointer is more expressive as it maintains temporary trails that can express deixis more naturally with diverse meanings beyond just "look here." In addition to the laser pointer, we also found these applications have a free-drawing pencil tool to create durable annotations in collaborative visual analysis that last until deliberately erased, such as, for example, marking a location as "Area A" and referring to it afterward. We thus chose to implement these two common tools in our interface.

#### 3.1.1 Collaborative Interface for Capturing Gestures

Based on the above investigation, we designed a collaborative interface (Figure 2). It consists of three main parts: (A) the gallery of meeting materials , (B) the collaborative board, and (C)) the room controls.

**(A) Gallery of meeting materials.** Anticipating multiple visualizations may be examined in a meeting, we designed a gallery of meeting materials. This gallery is synchronized among all users and acts as a menu for selecting the main visualization being viewed. The gallery concept draws a parallel to the slide gallery in presentation tools, but is

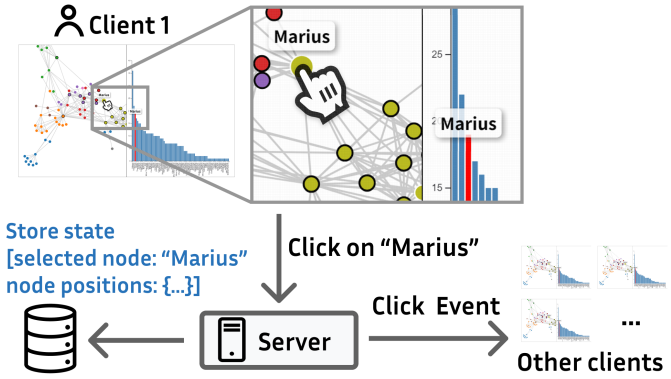



Fig. 3: An illustration of the operational mechanisms underpinning the interactive visualization demo within our collaborative interface. In this example, "Client 1" clicked on the node "Marius". This event is relayed to the server and then broadcast to all connected clients. Concurrently, the change of the current selected node and node positions are stored in the state file.

designed to offer greater flexibility in a more informal structure. It is intended for housing raw materials that require analysis and discussion within the meeting context.

The gallery supports direct import of static visualizations in JPEG, PNG, and SVG formats. As the collaborative interface's support for interactive web-based visualizations requires some manual effort to work with our system, we do not yet support automatic upload. Other types, such as videos, we expect will require other design considerations for appropriately capturing referential gestures.

**(B) Collaborative board.** The collaborative board is the main area of the collaborative interface and where we expect in-depth analysis of visualizations and the referential gestures of the participants to occur. We base our design on the Excaldraw whiteboard tool [20].

We incorporate the virtual laser pointer and pencil tools in the toolbar at the top of the screen (Figure 2$b_1$). From left to right, the toolbar includes the mouse pointer for initiating interactions (further details in Section 3.1.2), the pencil tool for durable annotations, an eraser for removing them, and the laser pointer for transient gestures. Figure 2$b_2$ shows an example of Alice, a user remote to the viewer, using the laser pointer to circle a cluster in the scatterplot. The yellow trail is synchronized with the view of the viewer (local user). It remains visible for a brief period before gradually fading away.

**(C) Room controls.** We enable users to initiate a new session by starting a 'room' and sharing its link with others (Figure 2$c_1$). This feature is also inspired by Exalidraw. It includes a real-time display of participants currently active within the room. These controls also include the audio recording functionality, as indicated by the button in Figure 2$c_3$, which is essential for the data collection process.

#### 3.1.2 Supporting Interactive Visualizations

We developed a proof-of-concept interactive visualization within our interface to demonstrate our deixis-centered documentation for remote synchronous meetings in interactive scenarios. As illustrated in Figure 3, we employ the event synchronization mechanisms provided by VisConnect [56] to ensure that changes within the visualization are coherently synchronized across different clients. Moreover, each user interaction with the visualization triggers the recording of the current interaction state. These interactions represent another form of deixis.

Currently, there is no existing solution that can import interactive visualizations directly into the collaborative interface without some coding effort. While VisConnect enables collaborative visualization through event synchronization, it still requires some implementation changes to the non-collaborative interactive visualizations.

Capturing the deictic actions on the interactive visualization requires provenance tracking, which has been extensively researched in the visualization community [47, 53, 62]. Recently, tools like SIMProv [9] and Trrack [16] have emerged, offering capabilities for capturing provenance in web-based interactive visualizations. We employ trrack [16] to collect provenance data related to user interactions. Trrack can record the state of the interactive visualization at each interaction event, enabling the reconstruction of the visualization's state at any point in the user's journey. This feature is helpful for generating interactive notes that accurately reflect the current visualization state.

### 3.2 Formative Study Observations & Outcomes

We make the following observations based on the activities of the participants during our formative study:

**Observation: The laser pointer and pencil tool engendered different gesture types, frequency of use, and degrees of precision.** As the laser pointer was transient and the pencil tool durable, they were used in different ways by participants. First, the laser pointer was primarily used to direct attention, often accompanied by verbal cues such as "look here." In contrast, the pencil marks were made to last a long time. One participant noted that they would opt for the pencil tool to highlight discussion points they deemed important. So durable annotations made with pencil tool often span a large portion of a discussion while gestures made with laser pointer are typically associated with one sentence.

Second, the laser pointer was used significantly more than the pencil tool. Two participants mentioned they preferred staying with the laser pointer and only switched to the pencil tool when necessary. They would switch back to the laser pointer after completing their tasks with the pencil. This behavior underscores the laser pointer's role as the primary tool for most users.

Third, referential gestures with the laser pointer tended to lack precision, which means drawings may not precisely cover intended area, and rely heavily on verbal clarification to resolve. For example, a participant might use the laser pointer to indicate a growing segment in a line chart, saying, "look at this growing period..." The line drawn might not exactly cover the intended segment, so the verbal explanation of "growing period" helped convey the intended meaning. In contrast, annotations made with the pencil tool were characterized by greater precision. Participants tended to be more cautious with the pencil, sometimes using the eraser to correct inaccurate annotations.

**Design Outcome:** Given the vastly different natures of how the transient laser pointer and durable pencil tool are used, we developed distinct strategies for linking them with text in interactive documents. We choose to link transient gestures in a fine-grained fashion, like phrases or keywords. We choose to treat durable annotations more like image changes in the interactive document.

These choices influence our technical requirements, particularly in the utterance matching and reference extraction steps described in Section 4. For transient gestures, we match utterances to the laser pointer's actions based on the timestamps captured at the start (mouse down) and end (mouse up) of its activation. In contrast, durable annotations are linked to every sentence articulated from their creation (indicated by the pencil tool's mouse down action) until their deletion (erased with the eraser).

Furthermore, due to the imprecise nature of transient gestures, we implement an additional reference extraction step. This process aims to connect these gestures with specific words and phrases in the corresponding sentence, thereby creating more intuitive and clearer linkages.

**Observation: People use the laser pointer for purposes beyond gesturing.** The use of referential gestures among participants showed significant variation. Some individuals used the laser pointer beyond gesturing, occasionally creating drawings that were meaningless. One participant characterized these actions as involuntary and an unconscious byproduct of their thought process. Generally, these behaviors do not hinder interpersonal communication, as others tend to disregard these non-essential gestures. However, these indiscriminate doodles can introduce inaccuracies in the resulting interactive document by generating irrelevant annotations.

**Design Outcome:** To mitigate the inclusion of spurious or non-deictic gestures in our interactive notes, we take additional measures to identify and filter out extraneous gestures.

**Observation: The motion of the laser pointer is used to attract attention, indicate direction, and convey emotional intensity.** The motion of the laser pointer can convey additional meaning. First, motion can be used to better direct attention. Especially when referring to small areas, people tended to use the laser pointer to draw back and forth over the same location. Second, motion can convey critical information such as reading direction. For example, a person can draw a line to guide viewers to follow a path from a certain direction. In the absence of animation, the indication of direction becomes ambiguous. Third, motion can be used to convey a sense of emotional intensity. For example, by rapidly accelerating the movement of the pointer to highlight a rapidly growing area on a line chart.

**Design Outcome:** In light of these insights, we adapted our interface to record the timestamps of each point along the laser pointer's trails. This data allows us to recreate the referential gestures through animations, accurately mirroring the original motions.

## 4 Automatic Generation of Interactive Notes

We present our framework for automatically generating interactive meeting notes with both verbal and non-verbal communication cues. We first present a brief overview. We then go over the techniques used to transform the recorded audio and gestures into interactive notes.

### 4.1 Framework Overview

The goal of our framework is to enhance documentation of collaborative meetings around visualization. The framework comprises four main components: i) data collection (Figure 1A), ii) utterance matching (Figure 1B), iii) reference extraction (Figure 1C), and iv) interactive notes generation (Figure 1D). The first these steps identify messages with deixis-encoded aspects. The final step re-encodes the identified message as a persistent digital document. Figure 1 illustrates this workflow. We describe the framework in overview here and present details in the subsequent sections.

**i) Data Collection.** To accurately document both deictic and verbal portions of collaborative visualization meetings, we collect three types of data: audio recordings, referential gestures (using pointer-based laser pointer and pencil tool), and interaction provenance data. Each data type is labeled with timestamps to facilitate subsequent matching processes. We use the collaborative interface from our formative study, though the framework could be applied to any interface that collects this data. In our implementation, the recorded audio is transcribed using Whisper [52], an open-source automatic speech recognition (ASR) system. Whisper further provides word-level precision timestamp prediction, which helps in the later utterance matching step. We also utilize the implementation of speaker diarization using Whisper [19] to distinguish different speakers.

**ii) Utterance Matching.** Once data is collected, the audio is transcribed and the transcribed utterances are aligned with corresponding referential gestures based on timestamps. For example, in Figure 1B, the transcribed utterance "Alice: Look at this cluster here..." is matched with the referential gesture of Alice circling a cluster in the scatterplot using the timestamps of the gesture and speech. Drawing from our formative study with visualization experts, we developed specific matching strategies for different types of referential gestures. These strategies are discussed in detail below.

**iii) Reference Extraction.** We further winnow the audio associated with a gesture to a more precise meaning using a large language model (LLM). We design prompts to extract connections between words or phrases within the matched utterances and referential gesture pairs. For example, in Figure 1C, the LLM identifies "this cluster" from the transcribed utterance "Alice: Look at this cluster here..." as the object most likely to be the focus of the gesture. The design and implementation of this reference extraction process are described in Section 4 and the practical limitations are discussed in Section 7. We note this step partially decodes meaning through the LLM to refine the identification. We hypothesize further understanding of deixis encoding could improve this and other applications and thus present a preliminary taxonomy of pointer-based deictic gestures in Section 6.

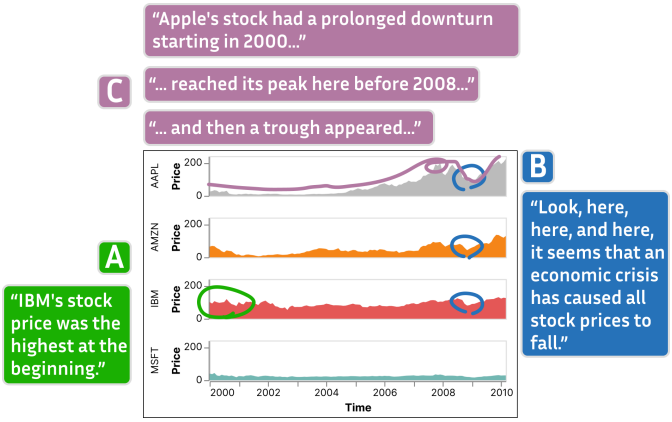



Fig. 4: Illustration of various scenarios for matching utterances with transient referential gestures. (A) One gesture matches with one sentence. (B) Several gestures match with one sentence. (C) One gesture matches with several sentences.

**iv) Interactive Notes Generation.** We create an interactive document representing the collaborative data meeting with meeting minutes, a transcript, and annotated visualizations. The meeting minutes are generated by the LLM. Both the meeting minutes and transcripts are augmented by creating links in the text to animated annotations that display the referential gesture in the visualization. For example, in Figure 1D, hovering over the word "cluster" in the interactive note triggers a replay of the referential gesture made by Alice—the circle she made around the cluster replays in the scatterplot. We describe the design and implementation of interactive notes generation below.

### 4.2 Gesture Filtering & Matching

We have different strategies for matching transient, durable, and interactive visualization gestures to sentence-sized utterances. However, first we filter out easily identifiable non-referential gestures. As noted in Section 3, some people tended to doodle when not speaking. Keeping only gestures made by the active speaker is a simple yet effective strategy to filter out most non-communicative gestures.

**Transient gestures.** Referential gestures made with the laser pointer are the most common. They typically have a short duration, usually with one gesture corresponding to a single sentence. However, this heuristic is not sufficient. During our experiment, we also observed many cases where people used multiple gestures within a single sentence. Drawing a long-lasting trajectory while speaking multiple sentences was a very rare case. Figure 4 illustrates these three cases. In case A and case B, where a gesture is associated with at most one sentence, it is sufficient to record the match by assigning each gesture to the single sentence it appears during.

Conversely, scenarios like C, where multiple sentences are linked to a single gesture, typically arise when the speaker uses the laser pointer continuously without releasing the mouse button. In such instances, a single gesture may convey different meanings at different junctures, thus presenting a challenge to the following reference extraction task. This complexity arises because extracting meaningful references from the gesture becomes more difficult as the size of the text (utterance) grows. To mitigate this issue, we divide the continuous gesture into distinct segments based on the timestamps associated with each sentence. This process ensures that each gesture is paired with a single sentence at most.

**Durable gestures.** As discussed in Section 3, durable annotations made with the pencil tool persist from the creation to the deletion. Rather than associate these with all of the sentences uttered during their existence, we instead associate them with timestamps. The intent is to match them with a state of the visualization rather than specific utterances. Thus, in the output interactive notes document, users can reveal durable annotations by timestamp. Figure 6 shows these timestamps in the

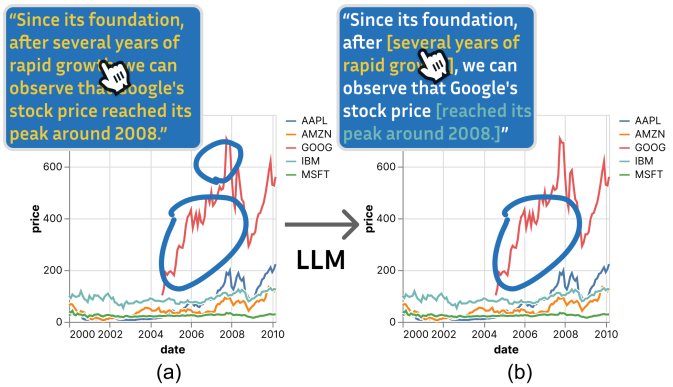



Fig. 5: Illustration of reference extraction: (a) Two imprecise gestures are matched with one sentence. (b) After reference extraction, gestures are sequentially associated with two separate phrases.

interactive notes, enclosed by red rectangles.

**Interactive visualization gestures.** Multiple factors could be considered when handling stored manipulations of visual interface (interaction provenance). These factors include the presence of animated transitions, the duration of the transitions, the type of interaction, and the extent of changes made to the visualization. Some interactions, like selection, highlighting, and brushing, are analogous to the mouse pointer gestures in terms of focusing attention. We decided not to consider animated transitions as these are properties of the interactive visualizations rather than a choice of the person manipulating the system. It would be unclear whether the animated transition is an intended part of the encoding of the speaker or an aspect they could not control and therefore we rely on the instigating action, e.g., selection or brushing, instead. We also chose not to consider interactions that alter the visual encoding, as those are more akin to looking at another visualization rather than gesturing. We chose to include dynamic queries that do not change the visual encoding as gestures for their attention and focus aspects. Given the durability of the changes, we do timestamp matching instead of utterance matching, similar to durable gestures. We further discuss the nuance of different interaction types and possible ways to capture them in (interactive) documentation in Section 7.

### 4.3 Reference Extraction

The casual and imprecise characteristics of transient gestures, which we discovered in our formative study, pose additional challenges to associating those gestures with the most meaningful speech. We cannot typically discern the annotator's intent based on the gesture alone. Reviewing the speech (text) is necessary to grasp the intended meaning fully. This issue is compounded in instances where multiple gestures are made within a single sentence, as demonstrated in Figure 5(a). To make the generated notes more comprehensible and reduce the cognitive load on users, the reference extraction step matches gestures to relevant words and phrases from their previously matched sentences. As illustrated in Figure 5(b), two gestures are identified and sequentially associated with "several years of rapid growth" and "reached its peak around 2008", thereby making both user interaction and reading clearer and easier to understand.

We employ the in-context learning capabilities of LLMs [18, 44] for reference extraction. In-context learning allows the LLM to adapt to new tasks through inference alone, eliminating the need for additional training or fine-tuning. To enhance the precision and reliability of the extraction process, we adopt the "chain-of-thought" [61] prompting strategy coupled with a few-shot learning approach [8]. This approach can enhance the model's robustness by explicitly demonstrating the thought process required for the task, encouraging the model to follow a step-by-step cognitive approach. The "chain-of-thought" prompting, in particular, aids in making the model's decision-making process more transparent and interpretable. The few-shot examples in the prompt can be found in the supplemental materials.

### 4.4 Design and Implementation of Interactive Meeting Note

We design the interactive note document following the design outcomes we identified during the formative study (Section 3). As shown in Figure 6, the document consists of two primary views: the *Transcripts & Minutes view* (Figure 6A) and the the *Gallery & Visualization view* (Figure 6B).

The Transcripts & Minutes view depicts the verbal communication of the meeting. The transcript is generated directly from the audio as described in Section 4.1 i). However, raw transcripts can be lengthy, unstructured, and noisy, making it challenging to quickly identify key points and actionable items within the text. Recently, researchers have explored how to leverage an LLM's in-context learning capacity to understand complex, nuanced meeting content effectively [54, 55]. Following a similar strategy to Schneider et al. [55], we produce meeting minutes (long-form summaries) by segmenting meetings into topics and generating meeting minutes for each topic separately. This method can produce meeting minutes that are cost-effective compared to using large language models across the whole transcript.

During the generation of meeting minutes, we apply an additional step to instruct the LLM to preserve referential gestures when transforming transcripts, merging multiple gestures when necessary. As shown in Figure 6, $a_1$ and $a_2$, five referential gestures were preserved in the meeting minutes out of nine total gestures in the transcripts. ①, ②, and ③ all depicted the Asian communities in Boston, and are semantically merged into 1-3 in the meeting minutes. When a mouse hovers over 1-3, the visualization view (Figure 6 $b_2$) displays the three gestures made by *Bob*, which are a combination of gestures ①, ②, and ③. Of the next five gestures, ④ and ⑨ are preserved and included in the meeting minutes, while the intervening gestures are only available in the transcript.

We build our interactive notes using prior work on authoring interactive documents [15, 25]. Specifically, we use the interaction features of Living Papers [25], a framework for integrating executable code, interactive components, and traditional text into a unified document. We write the reference extraction results into Living Papers' custom markdown. When the mouse hovers over the text links (light blue), as shown in Figure 6(A), the parameters (data regarding the referential gestures) are passed to the *Gallery & Visualization view* (Figure 6B), which then switches between different charts and plays the annotations based on the provided parameters. This approach provides the additional feature that users can edit the interactive notes directly through this markdown file.

## 5 User Evaluation

We conducted an evaluation with users to assess the utility of our approach and to further examine the use of gestures in remote synchronous meetings with data visualizations. The study was approved by the University of Utah Institutional Review Board (00175137).

### 5.1 Study Design

Our study was conducted in one-hour sessions with paired individuals who collaboratively discussed visualizations through our system and then reviewed the generated documents. While we expect collaboration dynamics may be different with larger groups, we chose to conduct the evaluation in pairs as we were focusing on the gesture-capture aspects of our framework. We consider our sessions as a simplified version of small collaborative team meetings described by Brehmer and Kosara [6]. In these "Jam Sessions" people meet and discuss the data and data visualizations casually.

#### 5.1.1 Participants

We recruited 18 participants (10 men, 8 women, all fluent English speakers) to evaluate our note generating tool. Based on our recruitment channels, all participants are graduate students in biology or computer science from a university in North America. Participants were matched into pairs based on their availability. Participants were compensated 30 USD for their time.

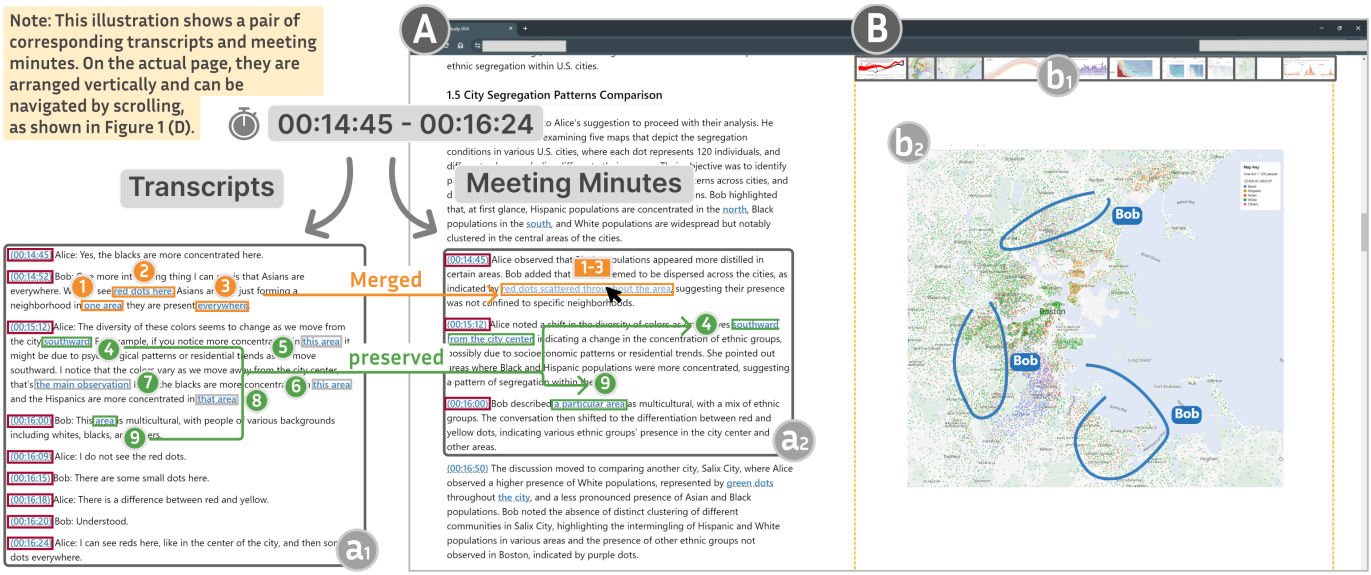



Fig. 6: An overview of the interactive notes, with: (A) Interactive text, comprising transcripts from audio and the LLM-generated meeting minutes, includes interactive text components based on the results of utterance matching and reference extraction. (B) Visual media from the meetings are presented with annotations based on parameters transmitted by the interactive text on the left. This operation can change the underlying visualization, add annotations, and alter interactive states.

We sought participants that would be likely to perform exploratory data analysis. As such, we advertised to the following groups: (1) a data science course at our institution, (2) a data science meetup at a second institution, (3) a visualization course at a third institution, and (4) the graduate student group at our institution.

#### 5.1.2 Procedure

We began our sessions with briefing and consenting. The facilitator (first author) then gave a demonstration of our collaborative UI and its features (∼10 minutes). Participants were then asked to turn on their audio recordings, view the visualizations already in the interface, and discuss them with each other based on prompt analysis questions in the interface. (∼30 minutes). The visualizations and prompts are available in supplemental materials.

After the analysis session, the interactive documents were then generated and provided to the participants. Before reviewing the documents, the participants were asked to fill out a background survey about their experiences with collaborative data meetings and note-taking. The facilitator then verbally verified with each participant that the linked gesture annotations were available to them. The facilitator then asked the participants to review the documents and fill out a second post-survey (described below). This process took about 20 minutes. Participants were then debriefed and the session ended.

#### 5.1.3 Post-Survey

The post survey consisted of seven 5-point scale questions and three open response questions. It was administered through Google Forms. We asked participants to (1) rate the quality of the interactive documents and their specific components, (2) compare the documents to video recording and transcripts, and (3) rate their likelihood of using a framework like ours in the future. Each of these sections was followed by an open response asking them to elaborate. The full survey with responses is available in the supplemental material.

#### 5.1.4 Visualizations Used in the Study

We chose a variety of visual idioms for the participants to explore during the session as different idioms might encourage different gestures. The static visualizations included the Minard map of Napoleon's march, segregation maps of several cities from the New York Times [42], and (stacked) line charts, node-link diagrams with topics of global warming and virus evolution from Reuters [12, 29]. The interactive visualization is a Les Misérables co-occurrence graph with a bar chart displaying the frequency of occurrence for each character's name, shown in Figure 3. We asked participants to select the visualization that they deemed most likely to spark discussion. Each visualization was discussed by at least two groups.

### 5.2 Results

We present the results of both post-surveys.

#### 5.2.1 Data Meeting Background Survey Results

In our background survey, we asked the participants to report their experiences with collaborative data meetings and note-taking. We discuss the findings from this survey. The raw survey results are available in supplemental materials.

Most participants (17/18) have experience with data analysis, of which 13 reported having done so with visual aids. Six participants claimed that they engage in meetings that are similar to our setting (collaborative data meeting with visual aids) frequently at school or during work.

When asked about their experiences with note-taking, the majority (11/18) reported jotting down some key points from their meetings. Three participants reported they do not take notes during meetings, with two adding they do not have time during the meeting. Three participants said they generate notes after meetings by watching video recordings or just using their memory. Fourteen participants said they use digital or physical writing tools (e.g., iPad, pen and paper) to take notes. Eight participants use a text editor and keyboard input (e.g., Notepad and Visual Studio Code).

We then asked the participants how they incorporate visual data into their notes. Eleven participants said they would put screenshots in their notes. Nine participants said they would draw a sketch. Five participants said they would write text description of the visual items.

Lastly, we asked participants to discuss the main challenge they face in taking and using notes in meetings with data analysis. Eight participants (P3, P5, P7, P8, P9, P10, P12, P17) mentioned time constraints and the fact that talking while writing can be distracting. Seven participants (P1, P4, P6, P9, P12, P13, P16, P18) wrote about the difficulty of relating data (visualizations) with text. These responses validate the motivation for our framework.

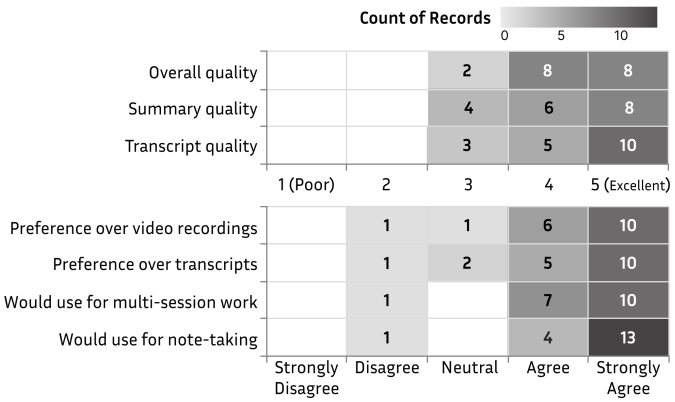



Fig. 7: Participant responses to scale-based survey questions. Most participants preferred our interactive notes to video recordings and transcripts and would use them again, despite mixed responses to the quality.

#### 5.2.2 Interactive Notes Evaluation Survey Results

Figure 7 shows the results from our scale-based questions. We note that most participants preferred the interactive notes over reviewing video or the transcript alone and most would use a tool like our framework for both reviewing during a multi-session project or documenting a collaborative online meeting. Participants were more mixed on the quality of the framework. We look at the open responses to better interpret these results. The second author coded the open responses and coalesced them into the insights below:

**Participants noticed errors in both the transcript and meeting minutes, but many considered both good enough despite these errors.** Ten participants (P2, P4-5, P7, P9-10, P14, P16-18) wrote that there were errors in either the transcript, specifically incorrect or dropped words, or the meeting minutes, specifically incorrect or missing details. Four (P5, P10, P14, P18) also mentioned the notes were good despite these errors. Three participants (P5, P7, P14) additionally noted what they considered matching errors in the gesture correspondences.

**Most participants found utility in the interactive documents, citing reasons like the speed of reviewing the notes and labor saved during meetings.** Nine participants (P1-3, P5-6, P16-18) suggested the interactive documents were better than video, with all but P17 alluding to the fact that they could read and seek much faster than in video. P17 noted instead that the context is clearer through interactive documents:

> The advantage of the tool is that it has annotated visualizations, which makes it a lot easier to get the context. I guess you can still achieve that with a screen recording to some extent but you may need to hit pause a couple of times. -P17

Eight participants (P1, P3, P6, P8, P13, P15-16, P18) picked out the value of the meeting minutes as useful for situations where they did not need specific details. Three participants (P1, P13, P18) specifically mentioned the annotations.

Seven participants (P2-3, P5, P10, P13, P15, P17) mentioned the framework is labor-saving: they did not have to take notes themselves or summarize them after the fact. For some, this would allow them to focus more on the meeting:

> I do collaborative meetings and having this tool during my discussion session would be helpful. It means I wouldn't have to concentrate about taking notes. Instead I can focus fully on the discussion with my collaborators. -P13

**Some participants noted video has advantages of higher accuracy and the inclusion of human expressions.** When comparing with video, three participants (P4, P14, P18) noted the video would not have errors. Two (P8, P16) suggested the video may contain precise details the automated system might note capture. Two participants (P7, P12) noted that video captured human cues from voice and expression where text notes do not, suggesting these cues enhance understanding. Additionally, P14 wrote they were used to searching videos from their coursework and thus would prefer videos over transcripts.

**Results Summary.** Overall participants generally found utility in the framework, with many wanting to use it again and noting its features over other automatic methods such as video recordings and transcripts. Views on transcription and subsequent summary quality were mixed, which limited the utility of the framework.

### 5.3 Limitations

We did not pair participants based on their mutual familiarity. However, previous work in non-pointer-based settings has found that the familiarity between two people can affect the referential gestures they use [14]. People with greater familiarity may successfully communicate using fewer gestures. Whether this holds true for online meetings with pointer-based gestures remains to be investigated.

We chose to limit the evaluation sessions to collaborative pairs instead of larger groups. While we did not expect this choice to have an large effect on overall transcription quality or the kinds of gestures used, we do not know the extent to which it may have affected utterances, repetitions, or audio collisions and our framework's subsequent performance. Other studies are needed to further explore the use of pointer-based gestures in remote synchronous multi-person meetings around data.

The facilitator knew seven of the participants before they volunteered as they work in the same building. Those participants may have been more generous in their responses. Participants acquainted with the facilitator rated an average quality of 4.43/4.14/4.43 (overall, summary, transcript), in contrast to the total averages of 4.33/4.32/4.39.

## 6 Preliminary Taxonomy of Pointer-based Deictic (Referential) Gestures

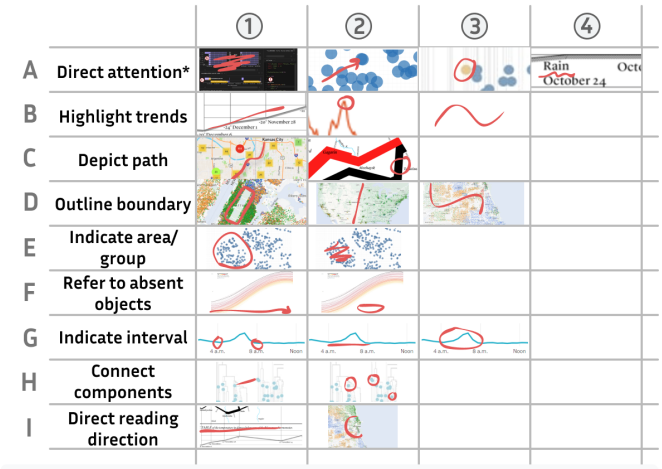
In our participant studies, we observed that gestures using the laser pointer and pencil were more expressive compared to those made with the mouse pointer alone. Participants made a variety of deictic gestures beyond just directing attention. We thus sought to better understand the nuances of these behaviors to better recognize, process, potentially decode, and design for these gestures in this and future tools.

The first author reviewed all study recordings to identify each gesture and assess its intention. From this review, we found patterns in the types of gestures used and their purpose. Such patterns underscore the complexity and richness of non-verbal communication facilitated by digital tools, thus providing a foundation for advancing our understanding of non-verbal communication through digital media.

Following the previous efforts of categorizing hand gestures [27] and the exploration of referential behavior in asynchronous communication [24], we describe the first preliminary taxonomy for referential gestures based on simulated laser pointers. We classify the gestures based on user intentions, emphasizing the critical notion that the significance of these gestures is deeply rooted in the context of communication.

Our taxonomy categorizes the observed gestures into nine intentions, each served by multiple gestures. Figure 8 lists these intentions and the gestures observed. While each observed intention-gesture pair is supported by numerous examples from our study, our dataset is limited in scale and scope and thus this taxonomy should be considered *preliminary*.It can be used to inform additional, more comprehensive studies. In assembling this preliminary taxonomy, we observed:

**i) Intentions can be served by multiple distinct gestures.** In Figure 8, we note that each intention, from A to I, can be served by multiple different gestures. Besides personal preference, this appears to be related to the type and attributes of the targeted object. For example, when trying to direct attention, people tend to draw wavy lines under text (A④), use Z/N shape scanning for large objects (A①), and employ arrows to point to smaller objects (A②). This finding underscores the inherent diversity and adaptability in non-verbal communication, providing insights for understanding deictic behavior in visualization




**A** ①:Z/N-shaped scanning at target. ②: Draw an arrow to point to the target. ③:Circle the target with the laser pointer. ④: Draw (wavy) lines under the target.

**B** ①: Trace trends along a line chart. ②: Highlighting a pivot as a narrative anchor, e.g., 'after this point'. ③: Draw a polyline in a blank space to convey intent, such as 'rising and then falling'.

**C** ①:Laser pointer moves along the path. ②:Draw circles on the path to denote it, suitable for charts with fewer paths.

**D** ①: Move the laser pointer along the target's outer boundary, may slow or pause at certain spots for emphasis. ②: Draw a line to differentiate content on both sides. ③: Use more complex shapes to differentiate multiple areas.

**E** ①:Lasso-select the area/group to be highlighted. ②:Z-shaped or N-shaped scanning to denote area/group, works when the group boundary is obvious, or combined with verbal description.

**F** ①:Redraw the contour of the missing part. For example, draw a line with arrow to denote a missing axis. ②:Circle the location of the missing part.

**G** ①:Direct attention to both ends of an interval to indicate its range. ②:Mark interval with line segments. ③:Encircle the interval with a circular frame, supplemented by verbal description.

**H** ①:Draw a line connecting two components. ②:Continuously direct attention across multiple objects to indicate connection.

**I** ①:Draw straight line in specific direction to indicate reading order. ②:Use arcs or other shapes to denote more complex reading directions.

**Other categories often serve to direct attention as well*

Fig. 8: Our preliminary taxonomy derived from observations made during our two studies. Letters A to I correspond to different intentions, with each intention associated with multiple distinct gestures, denoted by ①-④. Text descriptions of each gesture are in the key in the lower half.

context. It also suggests that intention might be derivable when taking mark type and other visual features into account.

**ii) Gestures often convey dual meanings.** On one hand, they all serve to direct attention, guiding the gaze of other participants. On the other hand, many of them (all categories except A, which is just "direct attention") serve to convey their own specific meanings, such as highlighting trends (B) or indicating intervals (G). This demonstrates the rich expressiveness inherent in the deictic gestures.

**iii) The meaning of referential gestures relies heavily on the accompanied verbal expression.** From the table, it is evident that many gestures employed to convey various intentions share fundamentally similar forms, such as A③, B②, and C② all appearing in the form of a small circle. This reveals another critical characteristic of referential gestures: their meaning is contingent upon the accompanying verbal expressions, underscoring the inseparable relationship between these non-verbal cues and verbal expressions.

## 7 Discussion and Future Work

We proposed a new framework for capturing referential gestures together with utterances to document collaborative data meetings. To the best of our knowledge, it is the first tool that tries to document meetings following a deixis-centered approach. As shown in the evaluation, the utility of the generated notes was recognized by most users. Compared to traditional recording methods such as screen recording and audio transcription, our approach was more preferred.

The interactive notes generated by our framework resemble the interactive documents discussed in Section 2.3. Our work differs from these methods primarily because it does not seek automated extraction of references, but instead capitalizes on natural human input—referential gestures used in communication—to obtain the references between texts and charts.

Our approach tends to result in fewer annotations than fully automated methods like ChartText [50], as gestures are not constantly used with every verbal expression. In many scenarios, the speakers do not need to use the laser pointer to direct people's attention. For example, they can just use the labeled names or colors to refer to a certain entity and the audience will be able to quickly identify the referred target. Speech without gestures frequently occur when the visualization is less information-rich or when the cognition process is pre-attentive (such as color pop-out).

We designed our system to only have annotations in deictic situations to preserve the original modes of directing attention, such as when pure verbal cues utilize salient labels or features alone. However, this does not account for the impact of ambiguity in communication. The lack of referential gestures could also be due to a high degree of familiarity and common ground regarding the topic between the two parties. For such cases, further investigation is needed.

Our work also has limitations. First, the quality of the reference extraction step could be improved. It is limited by both the in-context learning capabilities of the LLM and situations where a single utterance can contain multiple potential objects that could be referenced by gestures, making it difficult to determine the intended referent. Solving this problem requires improved techniques to further extract information from gestures and the visualization. The multimodal capabilities of an LLM may potentially resolve this issue, but at present, such improvements would require substantial computational resources. An alternative approach would be to refine and extend the taxonomy we have provided, thereby enabling further decoding of meaning. By combining contextual information about the target visualization, such as chart and mark types, with their accompanied utterances, we might better infer the type of gesture, thus providing more reliable links. As such, more data regarding pointer-based gestures in collaborative settings is needed.

Another limitation is that our framework only considers a limited set of visualization interactions that can serve as referential gestures. We implemented a highlight interaction in our proof-of-concept demonstration, as such interactions are clearly analogous to other deictic gestures. However, interactive visualizations in the wild typically involve multiple different types of interactions. Some interactions, such as highlighting and querying, do not alter the encoding method of the visualization, while others may change the visualization entirely, such as transitioning from a node-link diagram to an adjacency matrix. This type of interaction is distinct from deixis and more akin to image switching in static visualization discussions. Further study is needed to understand the intersection of deixis and interactive visual manipulations.

Our framework focuses on the automated documentation of collaborative meetings around visualization. Compared to manual note-taking, automated documentation allows participants to focus on the meeting with fewer distractions, a point several participants brought up in our evaluation. However, research has shown that structured note-taking can benefit cognitive engagement in online classes [21]. Our design goal was to aid more interactive settings where participants frequently communicate and discuss with each other. The trade-offs in using such an automated approach versus manual note-taking in a more one-way setting such as online classes would require further investigation.

## Acknowledgments

We thank the study participants, anonymous reviewers, Md Dilshadur Rahman, Zach Culter, Alexander Lex, Kiran Gadhave, Connor Scully-Allison, Sayef Sakin, Shadmaan Hye, Kalina Borkiewicz, and Utah SCI members for their valuable feedback.

The work reported here was supported by the Defense Advanced Research Projects Agency (DARPA), under agreement HR00112290092. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Defense Advanced Research Projects Agency (DARPA) or the U.S. Government.

## References

[1] L. M. Applegate. Technology support for cooperative work: A framework for studying introduction and assimilation in organizations. *Journal of Organizational Computing and Electronic Commerce, Taylor & Francis*, 1(1):11–39, 1991. doi: 10.1080/10919399109540148 2

[2] S. K. Badam and N. Elmqvist. Polychrome: A cross-device framework for collaborative web visualization. In *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces*, pp. 109–118. ACM, New York, 2014. doi: 10.1145/2669485.2669518 3

[3] S. K. Badam, Z. Liu, and N. Elmqvist. Elastic documents: Coupling text and tables through contextual visualizations for enhanced document reading. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):661–671, 2018. doi: 10.1109/TVCG.2018.2865119 3

[4] S. K. Badam, A. Mathisen, R. Rädle, C. N. Klokmose, and N. Elmqvist. Vistrates: A component model for ubiquitous analytics. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):586–596, 2018. doi: 10.1109/TVCG.2018.2865144 3

[5] R. A. Becker and W. S. Cleveland. Brushing scatterplots. *Technometrics, Taylor & Francis*, 29(2):127–142, 1987. doi: 10.2307/1269768 2

[6] M. Brehmer and R. Kosara. From jam session to recital: Synchronous communication and collaboration around data in organizations. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1139–1149, 2021. doi: 10.1109/TVCG.2021.3114760 3, 6

[7] S. E. Brennan et al. How conversation is shaped by visual and spoken evidence. *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions, MIT Press*, pp. 95–129, 2005. 2, 3

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. doi: 10.5555/3495724.3495883 6

[9] A. Camisetty, C. Chandurkar, M. Sun, and D. Koop. Enhancing web-based analytics applications through provenance. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):131–141, 2018. doi: 10.1109/TVCG.2018.2865039 4

[10] D. Chandler. *Semiotics: The Basics*. Taylor & Francis, 2022. doi: 10.4324/9781003155744 2

[11] Y. Chen, S. Barlowe, and J. Yang. Click2annotate: Automated insight externalization with rich semantics. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pp. 155–162, 2010. doi: 10.1109/VAST.2010.5652885 2

[12] M. G. Chris Canipe and S. Hart. Wild weather, warming planet. https://www.reuters.com/graphics/ENVIRONMENT-2020/WARMING/qzjpqdadnvx/, 2022. 7

[13] H. H. Clark. Pointing and placing. pointing: Where language, culture and cognition meet, 2003. 2

[14] H. H. Clark, R. Schreuder, and S. Buttrick. Common ground at the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, 22(2):245–258, 1983. doi: 10.1016/S0022-5371(83)90189-5 2, 8

[15] M. Conlen and J. Heer. Idyll: A markup language for authoring and publishing interactive articles on the web. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pp. 977–989. ACM, New York, 2018. doi: 10.1145/3242587.3242600 6

[16] Z. Cutler, K. Gadhave, and A. Lex. Trrack: A library for provenance-tracking in web-based visualizations. In *2020 IEEE Visualization Conference (VIS)*, pp. 116–120. IEEE, 2020. doi: 10.1109/VIS47514.2020.00030 4

[17] I. Denisovich. Software support for annotation of visualized data using hand-drawn marks. In *Ninth International Conference on Information Visualisation (IV'05)*, pp. 807–813. IEEE, 2005. doi: 10.1109/IV.2005.118 2

[18] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 2, 6

[19] M. A. et al. Speaker diarization using openai whisper. https://github.com/MahmoudAshraf97/whisper-diarization, 2023. 5

[20] Excalidraw community. Excalidraw. https://excalidraw.com, 2023. 3, 4

[21] J. Fang, Y. Wang, C.-L. Yang, C. Liu, and H.-C. Wang. Understanding the effects of structured note-taking systems for video-based learners in individual and social learning contexts. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–21, 2022. doi: 10.1145/3492840 9

[22] W. Fikkert, M. D'Ambros, T. Bierz, and T. J. Jankun-Kelly. *Interacting with Visualizations*, pp. 77–162. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. doi: 10.1007/978-3-540-71949-6_3 2

[23] S. Hall. *Encoding and Decoding in the Television Discourse*. Media series: 1972. Centre for Contemporary Cultural Studies, University of Birmingham, 1973. 2

[24] J. Heer and M. Agrawala. Design considerations for collaborative visual analytics. *Information Visualization*, 7(1):49–62, 2008. doi: 10.1145/1391107.1391112 2, 8

[25] J. Heer, M. Conlen, V. Devireddy, T. Nguyen, and J. Horowitz. Living papers: A language toolkit for augmented scholarly communication. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–13, 2023. doi: 10.1145/3586183.3606791 6

[26] J. Heer, F. B. Viégas, and M. Wattenberg. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 1029–1038, 2007. doi: 10.1145/1240624.1240781 2

[27] W. C. Hill and J. D. Hollan. Deixis and the future of visualization excellence. In *IEEE Visualization*, vol. 91, pp. 314–320, 1991. 2, 3, 8

[28] P. Isenberg, N. Elmqvist, J. Scholtz, D. Cernea, K.-L. Ma, and H. Hagen. Collaborative visualization: Definition, challenges, and research agenda. *Information Visualization*, 10(4):310–326, 2011. doi: 10.1177/1473871611412817 2, 3

[29] S. S. Jitesh Chowdhury and J. Wardell. How the novel coronavirus has evolved. https://www.reuters.com/graphics/HEALTH-CORONAVIRUS/EVOLUTION/yxmpjqkdzvr/index.html, 2020. 7

[30] R. Johansen. *Groupware: Computer support for business teams*. The Free Press, 1988. doi: 10.5555/542298 2

[31] D. H. Kim, S. Choi, J. Kim, V. Setlur, and M. q. Agrawala. Emphasis-checker: A tool for guiding chart and caption emphasis. *IEEE Transactions on Visualization and Computer Graphics*, 2023. doi: 10.1109/TVCG.2023.3327150 3

[32] D. H. Kim, E. Hoque, J. Kim, and M. Agrawala. Facilitating document reading by linking text and tables. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pp. 423–434, 2018. doi: 10.1145/3242587.3242617 3

[33] K. Kim, W. Javed, C. Williams, N. Elmqvist, and P. Irani. Hugin: A framework for awareness and coordination in mixed-presence collaborative information visualization. In *ACM International Conference on Interactive Tabletops and Surfaces*, pp. 231–240, 2010. doi: 10.1145/1936652.1936694 3

[34] Y.-S. Kim, N. Henry Riche, B. Lee, M. Brehmer, M. Pahud, K. Hinckley, and J. Hullman. Inking your insights: Investigating digital externalization behaviors during data analysis. In *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces*, pp. 255–267. ACM, New York, 2019. doi: 10.1145/3343055.3359714 3

[35] S. Kita. *Pointing: Where language, culture, and cognition meet*. Psychology Press, 2003. doi: 10.1016/j.cogsys.2004.01.002 1

[36] N. Kong, M. A. Hearst, and M. Agrawala. Extracting references between text and charts via crowdsourcing. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 31–40. ACM, New York, 2014. doi: 10.1145/2556288.2557241 3

[37] C. Lai, Z. Lin, R. Jiang, Y. Han, C. Liu, and X. Yuan. Automatic annotation synchronizing with textual description for visualization. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13. ACM, New York, 2020. doi: 10.1145/3313831.3376443 3

[38] S. Latif, Z. Zhou, Y. Kim, F. Beck, and N. W. Kim. Kori: Interactive

synthesis of text and charts in data documents. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):184–194, 2021. doi: 10.1109/TVCG.2021.3114802 3

[39] N. Mahyar, A. Sarvghad, and M. Tory. Note-taking in co-located collaborative visual analytics: Analysis of an observational study. *Information Visualization*, 11(3):190–204, 2012. doi: 10.1177/1473871611433713 3

[40] N. Mahyar and M. Tory. Supporting communication and coordination in collaborative sensemaking. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1633–1642, 2014. doi: 10.1109/TVCG.2014.2346573 3

[41] A. R. Martin and M. O. Ward. High dimensional brushing for interactive exploration of multivariate data. In *Proceedings of the 6th Conference on Visualization '95*, p. 271, 1995. doi: 10.5555/832271.833844 2

[42] A. C. MATTHEW BLOCH and T. GIRATIKANON. Mapping segregation. https://www.nytimes.com/interactive/2015/07/08/us/census-race-map.html, 2015. 7

[43] Microsoft Corporation. Microsoft teams: Meet, chat, call, and collaborate in just one place. https://www.microsoft.com/en-us/microsoft-teams, 2024. 2, 3

[44] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064. ACL, Abu Dhabi, Dec. 2022. doi: 10.18653/v1/2022.emnlp-main.759 2, 6

[45] Mural. Mural. https://www.mural.co, 2024. 3

[46] R. Neogy, J. Zong, and A. Satyanarayan. Representing real-time multi-user collaboration in visualizations. In *2020 IEEE Visualization Conference (VIS)*, pp. 146–150, 2020. doi: 10.1109/VIS47514.2020.00036 3

[47] C. North, R. Chang, A. Endert, W. Dou, R. May, B. Pike, and G. Fink. Analytic provenance: process+ interaction+ insight. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pp. 33–36. ACM, New York, 2011. doi: 10.1145/1979742.1979570 4

[48] J.-Y. Oh and W. Stuerzlinger. Laser pointers as collaborative pointing devices. In *Graphics Interface*, vol. 2002, pp. 141–149. Citeseer, 2002. doi: 10.20380/GI2002.17 2

[49] J. K. Parker, R. L. Mandryk, and K. M. Inkpen. Tractorbeam: seamless integration of local and remote pointing for tabletop displays. In *Proceedings of Graphics interface 2005*, pp. 33–40. ACM, New York, 2005. doi: 10.5555/1089508.1089515 2

[50] J. Pinheiro and J. Poco. Charttext: Linking text with charts in documents. *arXiv preprint arXiv:2201.05043*, 2022. 3, 9

[51] A. Piolat, T. Olive, and R. T. Kellogg. Cognitive effort during note taking. *Applied cognitive psychology*, 19(3):291–312, 2005. doi: 10.1002/acp.1086 2

[52] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023. doi: 10.5555/3618408.3619590 5

[53] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen. Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):31–40, 2015. doi: 10.1109/TVCG.2015.2467551 4

[54] I. Rousseau, L. Fosse, Y. Dkhissi, G. Damnati, and G. Lecorvé. Darbarer@ automin2023: Transcription simplification for concise minute generation from multi-party conversations. In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pp. 121–131, 2023. 6

[55] F. Schneider and M. Turchi. Team zoom@ automin 2023: Utilizing topic segmentation and llm data augmentation for long-form meeting summarization. In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pp. 101–107, 2023. 6

[56] M. Schwab, D. Saffo, Y. Zhang, S. Sinha, C. Nita-Rotaru, J. Tompkin, C. Dunne, and M. A. Borkin. Visconnect: Distributed event synchronization for collaborative visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):347–357, 2021. doi: 10.1109/TVCG.2020.3030366 3, 4

[57] A. Stapleton. Deixis in modern linguistics. *Essex Student Journal*, 9, 1 2017. doi: 10.5526/esj23 1, 2

[58] F. B. Viegas and M. Wattenberg. Communication-minded visualization: A call to action [technical forum]. *IBM Systems Journal*, 45(4):801–812, 2006. doi: 10.1147/sj.454.0801 3

[59] F. B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon. Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, 2007. doi: 10.1109/TVCG.2007.70577 2

[60] J. Walny, S. Huron, C. Perin, T. Wun, R. Pusch, and S. Carpendale. Active reading of visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):770–780, 2017. doi: 10.1109/TVCG.2017.2745958 3

[61] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. doi: 10.5555/3600270.3602070 6

[62] K. Xu, A. Ottley, C. Walchshofer, M. Streit, R. Chang, and J. Wenskovitch. Survey on the analysis of user interactions and visualization provenance. In *Computer Graphics Forum*, vol. 39, pp. 757–783. Wiley Online Library, 2020. doi: 10.1111/cgf.14035 4

[63] D. Yang, E. A. Rundensteiner, and M. O. Ward. Analysis guided visual exploration of multivariate data. In *2007 IEEE Symposium on Visual Analytics Science and Technology*, pp. 83–90. IEEE, 2007. doi: 10.1109/VAST.2007.4389000 2

[64] J. Zhao, M. Glueck, S. Breslav, F. Chevalier, and A. Khan. Annotation graphs: A graph-based visualization for meta-analysis of data based on user-authored annotations. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):261–270, 2016. doi: 10.1109/TVCG.2016.2598543 2, 3

[65] Zoom Video Communications, Inc. Zoom: Video conferencing, web conferencing, webinars, screen sharing. https://zoom.us, 2024. 2, 3