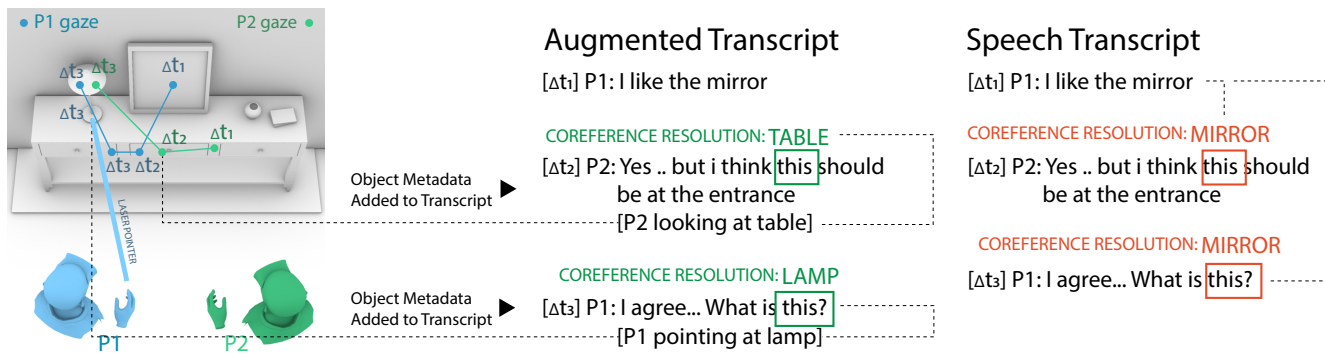# Augmenting speech transcripts of VR recordings with gaze, pointing, and visual context for multimodal coreference resolution

Riccardo Bovo
Imperial College London
London, United Kingdom
rb1619@ic.ac.uk

Frederik Brudy
frederik.brudy@autodesk.com
Autodesk Research
Toronto, Ontario, Canada

George Fitzmaurice
george.fitzmaurice@autodesk.com
Autodesk Research
Toronto, Ontario, Canada

Fraser Anderson
fraser.anderson@autodesk.com
Autodesk Research
Toronto, Ontario, Canada

Figure 1: Depiction of our system performing coreference resolution by leveraging non verbal cues such as pointing and visual attention. The system uses user's pointing behaviour and eye-gaze to determine the referent of ambiguous referring expressions, by correlating the location of eye gaze and pointing targets with pronoun "this" used by users.

Abstract

Understanding transcripts of immersive multimodal conversations is challenging because speakers frequently rely on visual context and non-verbal cues, such as gestures and visual attention, which are not captured in speech alone. This lack of information makes coreferences resolution-the task of linking ambiguous expressions like "it" or "there" to their intended referents-particularly challenging. In this paper we present a system that augments VR speech transcript with eye-tracking laser pointing data, and scene metadata to generate textual descriptions of non-verbal communication and the corresponding objects of interest. To evaluate the system, we collected gaze, gesture, and voice data from 12 participants (6 pairs) engaged in an open-ended design critique of a 3D model of an apartment. Our results show a 26.5% improvement in coreference resolution accuracy by a GPT model when using our multimodal transcript compared to a speech-only baseline.

CCS Concepts

• **Human-centered computing** → **Natural language interfaces**; **Virtual reality**; **Collaborative interaction**;

Keywords

multimodal coreference resolution, virtual reality (VR), speech, gaze, pointing

1 Introduction

The use of VR and AR for collaborative applications is rapidly growing, enabling immersive recordings of activities, such as design reviews [19, 20, 21, 2, 3, 1], remote support [33, 14], and social interactions [40]. These recordings are increasingly revisited to support collaboration, analysis, and follow-up tasks. As adoption grows, machines will need to better understand such recordings in order to summarize discussions, extract insights, and provide accessibility support.

However, immersive conversations (both, in real-world or mixed reality environments) are inherently multi-modal. Besides speech, conversation participants frequently use non-verbal cues–such as gaze and pointing–to establish shared understanding (Figure 1). This poses a difficulty for machine comprehension. *Referring expressions* (REs), such as "this" or "that," are common in natural conversation, yet their meaning often depends on visual or gestural context. For example, in the utterance "I like it!" the RE "it" can only be understood when paired with a gesture or gaze toward the referent. Current Large Language Models (LLMs) and

Visual Large Language Models (VLMs) can process transcripts of meetings to generate summaries and insights, but they struggle with these ambiguous Referring Expressions, leading to incorrect or incomplete identification of referenced objects. This limits the ability of intelligent systems to support tasks such as summarizing VR conversations [47], extracting information from design review meetings [23], and offering real-time accessibility support.

Prior research has explored grounding conversation in shared visual contexts by linking dialogue to entities in a scene using neural models [46, 9, 15, 10, 17]. However, these approaches are ineffective when speakers never explicitly name the object they are referring to (exophora). For example, in the dialogue 'P1: "What do you think about this?"' 'P2: "It does not look comfortable!"', the referent remains unclear without additional cues. Even when entities are explicitly named (endophora), coreference resolution is often unreliable in natural conversations [47]. For instance, after 'P1: "There is a funny coffee table."' a response like 'P2: "We should move this."' could refer either to the table or another object introduced later. While seminal HCI work such as "Put That There" [5] and follow-ups [34, 27, 7] leveraged non-verbal cues in human–machine interaction, they did not address multi-speaker conversations or use such cues for transcript comprehension. Similarly, studies on gaze synchrony in collaboration [41, 39, 28, 31] highlight its role in establishing shared focus, but have not quantified how effectively it identifies the precise object under discussion.

This paper proposes a system (Figure 2) that augments VR speech transcripts to improve coreference resolution, by leveraging non-verbal cues. By inferring speakers' attention through their gaze and pointing behavior, the system disambiguates spatial implicit referring expressions (REs) such as 'P1: "What do you think about this?"' or 'P2: "It does not look comfortable."' The system augments the transcript with contextual cues, linking these expressions to the intended object in the scene (e.g., the living room sofa), enabling more accurate downstream processing, such as summarization and information extraction. Our primary contribution is a system that integrates gaze and pointing into VR speech transcripts to address the fundamental problem of coreference resolution. We built a multi-user VR application that captures verbal interactions, eye-tracking data, and laser-pointing behavior. VR provides a controlled environment where pre-segmented 3D objects simplify linking verbal references and non-verbal behavior, avoiding the challenges of object detection and segmentation that arise in AR or real-world settings. VR is increasingly adopted for collaborative design review in both academia [19, 20, 21] and commercial applications [2, 3, 1], making it a natural setting to systematically study how non-verbal cues can be integrated into transcript comprehension. To evaluate our system, we conducted a study with 12 participants (six pairs) performing a collaborative design review in VR. We compared coreference resolution performance performed by Chat GPT4 on our augmented transcripts against a baseline using only speech transcripts.

The primary contributions of this work are twofold:

- A novel **system** that augments speech transcripts from collaborative VR sessions with non-verbal cues (gaze and pointing) to resolve ambiguous referring expressions, validated through a 12-participant study that demonstrates a 26.5% improvement in coreference resolution accuracy over a speech-only baseline.
- A **quantitative analysis** of how different non-verbal behaviors—gaze and pointing—and their synergies (individual, concurrent, and recurrent patterns) contribute to identifying the object of interest, establishing a clear hierarchy of cues for resolving ambiguity in immersive collaborative dialogue.

2 Related Work

We review previous work on *coreference resolution*, *visual attention* as well as *pointing based communication* in collaborative settings, and how HCI research leveraged comprehending deictic behaviour such as speech+gaze and speech+gestures through *non verbal and multimodal interaction*.

2.1 Coreference Resolution

Coreference resolution involves identifying words and phrases in a text that refer to the same entity, a crucial task in natural language processing [9]. This task is particularly challenging in conversational contexts due to their fluid structure, dynamic topic shifts, ambiguous pronoun use, and implicit shared knowledge, which can obscure clear reference resolution [46]. For example, speakers often use pronouns instead of specific names, complicating entity linking [4].

2.1.1 Visual Coreference Resolution. Recently, coreference resolution has significantly advanced through machine learning, especially when integrating visual and textual data to link text expressions to entities in images. This involves a system's ability to understand linguistic cues and visual features, using Neural Networks to process visual scenes and identify entities connected to text mentions [9, 45, 46]. Yu et al. developed *VisCoref*, a model for visual pronoun coreference resolution using deep learning techniques [46]. Goel et al. utilized "weak supervision" to train a model that identifies coreferences in text-image pairs and determines pronoun referents [9]. Yu et al. introduced *VD-PCR*, a framework to improve Visual Dialog comprehension through Pronoun Coreference Resolution by training a multi-modal BERT to understand pronouns in image-dialogue pairs and pruning dialogue to retain relevant input [45]. These approaches involve resolving coreference by training neural networks to analyze visual scenes, using one or multiple neural networks to process 2D images for visual scene analysis.

2.1.2 Immersive Visual Coreference Resolution. Several studies have approached visual coreference resolution using 3D visual representations instead of 2D. Kong et al. [15] presented a method that uses natural language descriptions of RGB-D scenes to enhance visual scene comprehension. Hong et al. advanced this by embedding 3D world knowledge into expansive language models in their *3D-LLM* [12], providing insights into coreference resolution in three-dimensional contexts. These methods resolve coreferences by analyzing the visual scene, disambiguate entities, and correlating them with textual mentions. Kottur et al. introduced the SIMMC 2.0 dataset, which includes immersive multi-modal conversations and a baseline model for coreference resolution and multi-modal disambiguation to improve AI assistants [17]. Dynamics of coreference resolution in human-AI interactions inherently differ from

those in human-human exchanges, where in human-AI dialogues, the AI system can request clarifications when visual coreference resolution is uncertain, while seeking clarification is inherently limited when analysing human-human dialogues, especially when the original speakers are no longer accessible for further context. Using the SIMMC 2.0 dataset, Guo et al. proposed a framework that uses metadata in the field of view to help disambiguate objects and correlate them with textual dialogue information [10].

We use a similar approach by extracting metadata from the scene based on a user's nonverbal cues (e.g., gaze, pointing). Previous methods resolve coreferences by analyzing the visual scene with neural networks, which involves understanding the scene, segmenting entities, and modeling relationships with mentions in the accompanying text. In contrast we adopt a simpler approach to visual coreference resolution by leveraging non-verbal cues (e.g., users' visual attention and pointing) to identify the target in a 3D scene, enhancing a speech transcript with contextual information.

2.2   Visual attention during communication

Synergies of gaze are common during human-human collaboration, stemming from the shared visual context but extending beyond visual alignment. Previous research highlights how this fosters mutual understanding, reduces the likelihood of misunderstandings, and enhances collaboration. Vrzakova et al. used recurrence quantification analyses (RQA) to identify patterns in visual attention during collaborative tasks, showing alignment with screen activity correlated with team performance and collaboration [41]. Villamor and Rodrigo found concurrent visual attention crucial in pair-programming tasks [39], while Moulder et al. quantified team-level gaze dynamics using RQA, finding them predictive of task success [28]. Awareness of visual attention helps ground referring expressions (REs) within a visual scene. Schneider and Pea demonstrated enhanced collaboration with mutual gaze perception in 2D tasks [36], and Zhang et al. showed gaze cursors facilitated communication in 2D screen interactions [48]. D'Angelo and Begel found visual attention cues reduced communication complexity in pair programming, aiding implicit referring expressions [8]. Similarly, visual attention can help to ground context in immersive scenarios where the visual context spans 360 degrees, creating blind spots for interlocutors known as the fragmentation problem. Hindmarsh et al. introduced this in collaborative environments [11], and Bovo et al. showed bidirectional head-based cues in VR increased mutual awareness and reduced cognitive load during data analysis [6]. Jing et al. found bidirectional eye gaze cues improved co-presence, gaze awareness, and collaboration in MR environments [13].

This prior work has shown the importance of visual attention awareness in simplifying human-human dialogue comprehension by enabling more implicit communication. In our work, we leverage visual attention to enhance machine comprehension of immersive human-human dialogues. Visual attention patterns, especially in tasks where participants work closely together, have been shown to predict various aspects of the collaboration, ranging from task success to the quality of the collaborative experience. However, to the best of our knowledge, visual attention synergism, such as concurrent and recurrence of visual attention, has never been used

as a retrieval method to identify the object of attention during collaborative discussion.

2.3   Pointing Based Communication

The absence of visual attention awareness, the complexity of a visual scene, and the difficulty of verbally describing a referent prompt the use of pointing gestures with utterances, known as deictic expressions [44]. Research shows that pointing gestures, particularly in immersive multi-modal conversations, significantly aid in referent disambiguation [44]. In AR/VR, pointing gestures become even more effective due to the use of lasers, enabling users to specify the referent of a pointing gesture [43].

Piumsomboon et al. found that virtual awareness cues, including pointing gestures, field of view, and eye gaze, significantly improved user performance, usability, and subjective preferences in MR collaboration [32]. Similarly, Bovo et al. demonstrated that users prefer pointing at complex referents, even if it requires movement, over verbal descriptions [6].The importance of pointing in collaborative VR environments is emphasized by research focused on enhancing pointing gestures accuracy. Techniques like warping or distorting deictic gestures have been proposed to improve collaboration [37]. Mayer et al. explored offset correction and cursor effects on mid-air pointing, finding that subtle redirection of a user's arm to align with their gaze can significantly improve pointing accuracy from an observer's perspective [26].

This collection of research emphasises that when the gold standard for immersive multi-modal dialogue comprehension (i.e., the human) faces uncertainty regarding referent clarity, it consistently resorts to *pointing* for disambiguation, whether in the speaker or observer scenario. Therefore, our work uses pointing behaviour to enhance machine comprehension of immersive human-human dialogues.

2.4   Non Verbal and Multimodal Interaction

Prior work in HCI leverages natural nonverbal communication, such as deictic pointing gestures or interpreting visual attention concurrently with speech commands. In both cases, speech+pointing or speech+gaze, the use of visual attention or pointing gestures aid the process of understanding the referent intended by the user. Bolt's seminal work "Put-That-There" [5] explored the integration of voice commands and hand gestures to enhance user interaction within graphical interfaces. Similarly Miniotas et al. studied the integration of eye gaze with speech, especially for interactions with small, closely spaced on-screen targets [27]. Recent works have pivoted towards understanding the visual context and harnessing natural collaborative communication for enhanced interactions. Specifically, Bovo et al. explored how head direction serves as an indicator of visual attention and speech [7].

Similarly, Mayer et al. [25] enhanced mobile voice assistants' understanding of nearby buildings by incorporating GPS location and user's head gaze (i.e. user's location and visual direction). Romaniak et al. [34] introduced 'Nimble,' a mobile interface combining visual question-answering models with gesture recognition for more intuitive user interactions. Similarly, GazePointAR, a wearable AR system, resolves speech query ambiguities using eye gaze, pointing gestures, and Human-AI conversation history [18]. Penzkofer et al.

explored collaborative behavior in multimodal conversations to extract metrics like collaboration quality and the impact of technology usage [30]. However, none of these prior works explored the collaborative patterns of visual attention or pointing behaviour specifically towards the problem of identifying a referent of a REs; instead, they focus on interactions rooted in single-user nonverbal communication and single-event interactions. In our work, we leverage similar information for coreference resolution in recorded speech transcripts of immersive multimodal conversations.

3 System and Implementation

Our system processes a VR recording of an immersive spatial conversation to identify implicit referring expressions (RE) and perform coreference resolution for each. The input includes the VR session's audio, eye-gaze, and laser pointer data. The output is an explicit referent for each spatial RE identified.

The process involves multiple steps: converting speech into a diarized transcript, detecting implicit spatial REs, and analyzing non-verbal cues (gazing and pointing) to identify the object of interest in a 3D scene for each RE. The system uses the transcript, non-verbal cues, and visual scene metadata (objects' names) to generate an augmented transcript. Finally, coreference resolution is performed for each spatial RE.

Implemented on an Intel Core i7 with 16GB RAM, the system processes each 15-minute session in about 5 minutes, including multimodal data processing and transcript generation.

3.1 Core Concepts

To contextualize our system's design, we first define the core linguistic concepts it is built to address. The primary challenge is **coreference resolution**: the task of associating ambiguous references, such as pronouns like "it" or "this," with the specific entities they refer to. These ambiguous references are a type of **Referring Expression (RE)**, which is any phrase used by a speaker to identify an entity in their environment. Our system specifically targets **implicit spatial REs**, where the referent is not explicitly named within the phrase (e.g., "I like this"), as opposed to **explicit spatial REs** that clearly name the referent (e.g., "I like this *mirror*"). Implicit REs introduce two distinct challenges that guide our system's design. The first is **exophora**, where the referent is completely absent from the dialogue and can only be identified through non-verbal cues and the shared visual context. The second, more frequent, challenge is **endophora**, where the referent is named elsewhere in the transcript, creating textual ambiguity between potential candidates mentioned earlier (**anaphora**) or later (**cataphora**) in the conversation. Our system is therefore designed to resolve both exophoric and endophoric ambiguity by augmenting the transcript with non-verbal data, providing the necessary context for accurate coreference resolution.

3.2 Transcript

We utilize the timestamped Whisper AI model [A] to transcribe VR session recordings. Each participant's audio track is transcribed separately, preserving speaker identity and enabling diarization. The individual transcriptions are then merged, appending speaker

[A]https://github.com/linto-ai/whisper-timestamped

identifiers to each segment and arranging them chronologically. The resulting transcript of the collaborative communication includes temporal timestamps for each word and sentence, along with speaker identity information.

3.3 Identify Implicit Spatial Referring Expressions (RE)

Implicit spatial referring expressions (REs) reference a location or spatial relationship indirectly without explicitly naming the location or object. For instance, "right on top of it!" is an implicit spatial RE, whereas "right on top of the couch!" is an explicit spatial RE. To identify these, we first *identify all spatial REs* and then *classify each as either implicit or explicit.*

3.3.1 Identify Spatial referring expressions. GPT-4 is used to identify spatial referring expressions (REs) in the VR transcript. The prompt defines the system's role: *"you are a system that identifies spatial referring expressions related to objects/places in a given sentence".* The prompt also specifies the response format: *"list them in the following JSON format:{ "spatial_referring_expressions": ["referring_expression",...]}".* We further refine the prompt to clarify what constitutes a spatial RE to an object or place and what does not. We exclude REs where the referent is a person (e.g., you, me, we, guests) or temporal REs (e.g., now, then, today, tomorrow). Additionally, we specify not to list REs related to objects not currently present in the scene (e.g., "there is no oven") or abstract/metaphorical entities.
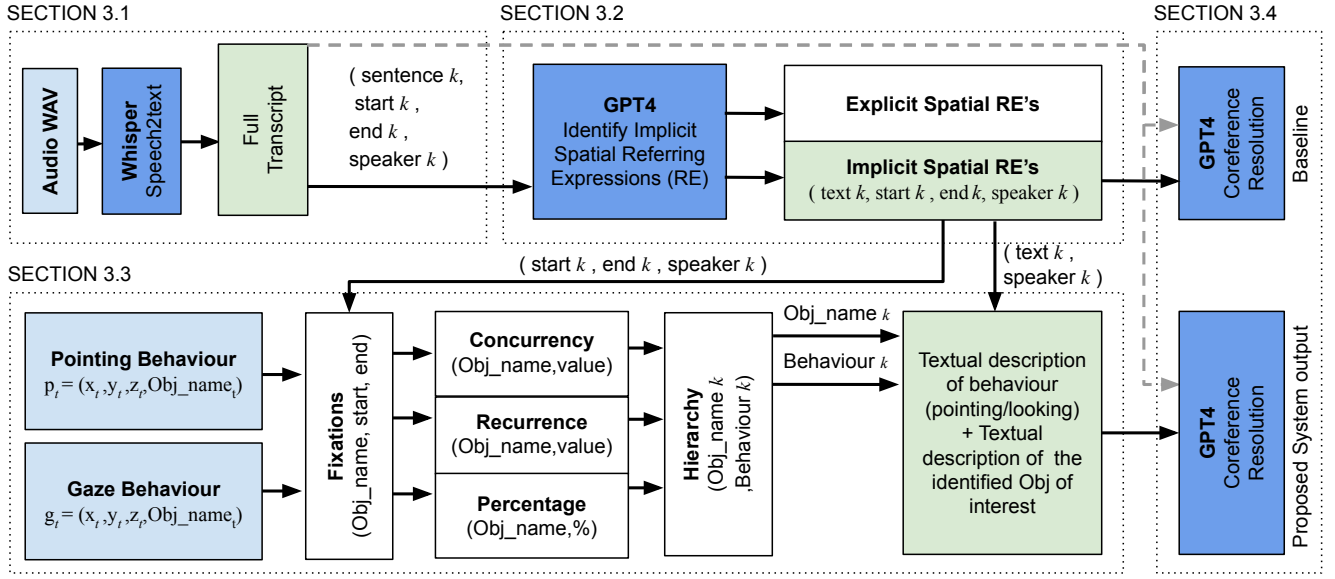
Next, we iterate through each sentence in the transcript. For each, we send the refined prompt to GPT-4. The model's responses (i.e., identified REs) are saved within the JSON file of the transcript. The full prompt can be seen in the supplemental material 5.3, Listing 1.

3.3.2 Classify implicit and explicit referring expression (RE). After identifying each spatial RE, the system further classifies them to identify those requiring coreference resolution (i.e., implicit spatial REs). We generate a prompt for GPT-4 containing the spatial RE and its sentence, asking the system to determine if the RE's referent noun is present within the sentence. If a spatial RE's referent is contained within the sentence, we classify it as *explicit*; if not, we classify it as *implicit*. The full request sent to the API can be seen in the supplemental material 5.3, Listing 2.

3.4 Identify Object of Interest

To pinpoint the object of interest within a scene, we analyze the spatial behaviors of people, using time series data for gaze and pointing actions. Each sample includes a 3D vector (x, y, z) representing the gaze position or pointing direction, along with the name of the intersected object in the 3D model. We process this data by identifying the *gaze and pointing fixations* on objects within the 3D scene, calculating *concurrent, recurrent,* and *individual* behaviors of pointing and gazing at objects, and finally *prioritizing the identified objects hierarchically* based on these behaviors.

3.4.1 Gaze and Pointing Fixations. Fixations refer to periods during which the person's attention (eye gaze or laser pointer) remains steady on a specific point in space. By analyzing fixations, we can discern meaningful samples within a signal—such as those indicating the person is pointing at or looking at an object—from noise, like when the person is merely glancing around the room.

Figure 2: The system architecture, as depicted in the diagram, consists of four main components: Transcript Generation (Section 3.2), Spatial Referring Expression (RE) Identification (Section 3.3.1), Object of Interest Identification (Section 3.4), and Coreference Resolution (Section 3.5). In Transcript Generation (Section 3.2), the Whisper AI model transcribes audio with time-stamping. Spatial RE Identification (Section 3.3.1) uses GPT-4 to detect Implicit Spatial REs for coreference resolution, with performance detailed in Fig 8(b). Object of Interest Identification (Section 3.4) analyzes pointing and gaze to identify objects in Implicit REs, employing fixation calculation, concurrence recurrence, and a hierarchical selection method; results shown in Fig 7, and the fourth step generates descriptions of non-verbal behaviours and identified objects. Finally, Coreference Resolution (Section 3.5) contrasts a baseline system (which uses only the speech transcript) with our proposed system (which integrates the transcript with the non-verbal behaviour and the object that it entails).

We calculate fixations using the I-DT (Dispersion-Threshold Identification) algorithm [35], extended to use additional scene information. Our gaze samples, recorded as points (x, y, z) in a virtual reality (VR) environment, are computed by casting a ray from the eye position along the eye tracker's recorded direction. This includes identifying the geometry hit by the ray in the VR environment. This helps determine whether the person is fixating on an object or moving toward a new target. If the pointer moves to a different object, the fixation is either completed or reset.

When using a laser pointer in a virtual environment, people often initially point incorrectly and then adjust. To address this, we calculate both eye and 'laser fixations'—periods where the laser pointer is steadily aimed at one location. This filters out data related to adjustments, keeping only the informative parts where the laser is fixed on the intended object. By considering both eye and laser fixations, we more accurately determine the person's attention and interaction within the VR environment. We use a 0.5-degree threshold and a 100-millisecond duration threshold as proposed by Salvucci and Goldberg [35].

3.4.2 Fixation Concurrency over object. Previous research highlights that synchronizing verbal communication with visual attention reduces misunderstandings during collaboration [41]. Building on this, we believe that measuring simultaneous visual attention

or pointing provides better insights into conversation focus than analyzing individual behavior alone.

Our algorithm detects overlapping fixations on the same object by two individuals. Each fixation is represented as a tuple with the object and start and end times. We measure concurrency by counting how often both individuals fixate on the same object at overlapping times, normalizing this count by the maximum possible concurrent fixations. This approach quantifies the extent of shared attention on objects during collaborative interactions.

3.4.3 Fixation Recurrence over object. A variation of concurrent attention is recurrent fixation, where two people focus on the same object in a 3D scene at different times. One person makes a comment and fixates on an object momentarily before looking away. The second person, guided by cues, then shifts their attention to the same object later. Although not simultaneous, this shows synchronization of verbal communication with visual attention.

To detect recurrent fixations between two people, we use their fixations to compute the recurrence of an object. The recurrence value is the total duration of fixations on the object by both individuals, divided by twice the total time of the REs. This approach measures how often both people direct their attention to the same object, indicating shared focus. Recurrent fixations are calculated for both gaze and pointing.

3.4.4 Individual fixations over object. We also calculate the percentage of time an object *o* was fixated on. This percentage is determined by dividing the total duration of fixations on the object $D(o)$ by the total duration of all fixations T, and then multiplying by 100. This gives us the proportion of the experiment time that each object held the participants' attention.

3.4.5 Hierarchical Selection. Since LLMs are primarily designed to predict and generate human language based on probabilistic models, their handling of numbers [42, 22] is not as precise or reliable as their processing of textual information. Consequently, they are less effective at handling numerical data, such as behavioural signals (i.e., non-verbal cues). To address this limitation, we supply the LLMs with deterministic answers instead of numerical quantities by developing a Hierarchical Selection algorithm. This algorithm establishes a hierarchy of importance for the metrics identified previously, using it as a fallback mechanism if a behaviour measure combination is not occurring.

Pointing, a deliberate action requiring effort, strongly indicates intention and attention, while gazing is more reflexive and influenced by various factors. Thus, pointing behaviour is prioritized over gazing. Additionally, the context of these behaviours is crucial; synergistic behaviours (concurrent or recurrent pointing/gazing) are more informative than individual behaviours, indicating shared focus and likely discussion topics.

This hierarchy guides the identification of the object of interest. Synergistic pointing behaviour is prioritized, selecting objects pointed at concurrently or recurrently for the most time. If absent, individual pointing by the speaker of the implicit spatial reference is considered. If no individual pointing occurs, visual attention (gazing) is analyzed, prioritizing objects gazed at concurrently or recurrently, followed by objects the speaker gazed at the most.

The output includes the selected behaviour measure (concurrent/recurrent/individual with gaze/pointing), the object with the highest percentage of fixation for that measure, and the person performing the implicit spatial RE.

3.4.6 Generate a description of the object of interest and the behaviour used to determine it. To make the implicit spatial RE less ambiguous and, therefore, facilitate coreference resolution, we integrate the object identified in the previous step into the transcript in textual form, forming an *augmented transcript*. The previous steps provided details such as the name of the recognised object, the behaviour measure leading to its identification, and the person performing the implicit spatial RE. Using this information we enhance the transcript by adding contextual information, such as "Person X was pointing/looking at Object Y" or "Both people concurrently observed Object Z" to the end of the sentences containing the implicit spatial RE. For example, a sentence in the augmented transcript might appear as: "[01:00] P1: This looks weird. [P1 was pointing at the sofa]."

3.5 Coreference Resolution

Coreference resolution in GPT-4 uses two methods: the baseline (speech-only transcript) and our proposed system (augmented transcript). Initially, a system description defines GPT-4's role in resolving implicit spatial references (REs) within sentences. The full transcript is included for each coreference resolution attempt. GPT-4 processes up to 8,192 tokens, sufficient for our transcripts averaging 1,841.625 tokens (max 3,066, min 1,001). Each implicit spatial RE is addressed individually within the transcript, updating the prompt with the specific RE sentence. (See supplemental material 5.3, Listing 3 for the complete prompt.)

4 Data collection for System Evaluation

There are currently no existing datasets that encompass collaborative speech, eye-gaze, and contextual information. Therefore, to assess the effectiveness of our proposed system, we compiled a dataset by recording 6 pairs of participants who were asked to collaboratively review a virtual apartment [A] [B] scene in VR (Figure 3) capturing their speech and eye-gaze within the 3D space. This study received review and approval from our institution's internal ethics review process.

4.1 Apparatus

Each participant used an Oculus Quest Pro, which rendered the scene and collected audio, gaze, and gesture data. Audio was recorded via an internal microphone, gestures via controllers, and gaze via built-in cameras at 120Hz with 0.5 degrees accuracy. We developed a custom application using Unity 2022.3.2f1 and the Oculus Unity SDK. This application rendered a 3D scene featuring two apartments in a real-time session where participants' avatars, represented using the Oculus Avatar SDK, interacted. Participants navigated using thumbstick controllers and used a laser pointer tool activated via controller buttons. The embedded microphone and speakers enabled verbal communication between participants. Avatar movements, including head and hand gestures, were streamed with low latency using the Photon Fusion v2 Network API to synchronize their positions and behaviors across all participants' scenes. A moderator could communicate audibly but was not visually rendered within the VR session.

4.2 Participants

We recruited 12 participants (4 women, 8 men) with the following inclusion criteria: being a fluent English speaker and having normal, or corrected-to-normal vision. Participants received compensation of $75CAD, and sessions lasted approximately 60 minutes.

4.3 Task

Participants engaged in an open-ended collaborative task involving navigating a virtual apartment scene, where they identified and discussed design aspects such as issues, considerations, and personal preferences. This task draws inspiration from recent VR applications in architectural reviews, as evidenced by studies [19,

---

[A]https://sketchfab.com/3d-models/vr-loft-living-room-baked-f3e6f16527af4465858a34cc1e9e7a2b
[B]https://sketchfab.com/3d-models/vr-morden-loft-apartment-baked-dd252381b69d41f883083677e56a7f3e

Figure 3: The two apartment scenes reviewed by participants. Participants were asked to collaboratively review these two virtual apartments in VR. [A] [B]

20, 21] and commercial uses [2, 3, 1] aimed at pre-construction evaluation of architectural designs. Experimenters guided participants encouraging them to examine spaces, furnishings, fixtures, and equipment, and to discuss observations with their collaborator. Participants had the freedom to comment on any design aspect without specific requirements. It was emphasized that consensus was not necessary, there were no right or wrong solutions, and the main goal was to engage in an environment-focused discussion. Supplemental material includes five conversation excerpts with labeling, co-reference resolution, and associated videos.

4.4 Procedure

Each data collection session followed a structured procedure. Initially, participants were presented with an informed consent form detailing the study's purpose, participant expectations, and data handling procedures, which they read and signed. Participants were informed that they would navigate a virtual 3D model of an apartment, collaborating to identify and discuss design elements. Next, participants received an orientation to the head-mounted display (HMD), focusing on proper fitting and adjustment of the HMD. They also received instruction on using the device's controllers for navigation and interaction within the application. Participants then engaged in the main task, which lasted approximately 15 minutes. Additionally, they performed an unrelated VR task for approximately 15 minutes per apartment. After completing both tasks, participants underwent a comprehensive debriefing session where the study's purpose and data collection rationale were explained.

4.5 Collected Data

Various data were collected during the task, all recorded at 120Hz. Eye-tracking data was collected recording the x, y, and z coordinates of gaze locations on the scene's objects. Laser pointing data was similarly collected, recording the x, y, and z coordinates of the laser's location on the scene's objects. Additionally, participant conversations were recorded. All participant data, was recorded by the monitoring application. Data synchronization and time stamping were ensured using a common time reference. An example snapshot of the collected data is shown in Figure 4, which displays a view from a participant's perspective alongside a segment of the transcribed conversation.

4.6 Data Labelling

Three of the authors manually labeled transcribed audio recordings to identify spatial referring expressions (REs) and classify them as implicit or explicit spatial REs. Implicit REs were further categorized based on whether they referred to an entity present in the transcript (endophora) or one absent, relying on visual context and non-verbal communication (exophora). Subsequently, we determined the referent and corresponding target geometry in the 3D scene for all references. The manual labeling process was carried out in order to identify the ground truth and compare the results of our proposed system. Given that three distinct raters participated in this labeling task, we assessed inter-rater reliability to ensure consistency, as discussed in Section 4.6.5. An example of labeled data is shown in Figure 5, which displays implicit RE (exophora/endophora) and explicit RE. Further examples of labeled data are available in the supplemental material 1–4.

4.6.1 Labeling Spatial Referring Expressions. Manual labeling begins by identifying all expressions and phrases which reference objects or places in the scene. These phrases or expressions might contain demonstrative deixis "this chair" or "the other one", adverbs of place "it would be better there" and prepositional phrases "...under the counter-top". We exclude certain RE from the process: objects not currently present (e.g., 'there is no oven'), temporal deixis (e.g., 'in the room we were before'), and abstract or metaphorical entities (i.e. in the dialogue 'P1: "We should isolate this shower."' 'P2: "I'd like that!"' in the second sentence, P2 refers to the idea of isolating the shower, not the shower itself).

4.6.2 Labeling Implicit, Explicit, Endophora, Exophora. Once all the spatial referring expressions are identified, we categorize each of them as either implicit or explicit. For example, an explicit spatial reference would be "this chair" as it clearly identifies a specific object in the scene. An implicit spatial reference might be "it looks weird," because "it" might refer to an object previously mentioned or an object that the participant is pointing at in the VR environment, without explicitly naming the object. For each implicit RE it is also labeled based on its reference type: whether it alludes to an entity explicitly named within the transcript (endophora) or to an entity not mentioned in the transcript (exophora).

4.6.3 Labeling Resolved Referent. For all the labeled implicit RE, we then annotate their intended target by watching the video corresponding to when the reference was made. The video provides

```
[02:34 --> 02:36] u7 : There are no handles on the cabinets.
[02:37 --> 02:38] u7 : Or maybe these are like push ones.
[02:39 --> 02:42] u7 : And...
[02:44 --> 02:46] u8 : But it's a good design overall.
[02:47 --> 02:51] u7 : I know. Like I got a lot of time
                         picking on it .
[02:52 --> 02:53] u8 : Maybe if you go upstairs, we'll find...
```

Figure 4: This snapshot shows participant's in the environment while the transcript captures the dialogue.

```
| implicit RE Exophora | implicit RE Endophora | explicit RE | non sptial RE
| ( referent ) -->  resolved by human labeler

[02:34 --> 02:36] u7 :  There are no handles on the cabinets 1.
[02:37 --> 02:38] u7 :  Or maybe these (cabinets) 2 are like push ones.
[02:39 --> 02:42] u7 :  And...
[02:44 --> 02:46] u8 :  But it's (kitchen design) 3 a good design overall.
[02:47 --> 02:51] u7 :  I know. Like I got a lot of time picking on it (kitchen design) 4.
[02:52 --> 02:53] u8 :  Maybe if you go upstairs, we'll find...
```

Figure 5: An example of the manual annotation process. The transcript corresponds to the snapshot in Figure 4 and the manual labels constitutes the ground truth used to test the system.

```
| correctly identified implicit spatial RE  | incorrectly identified implicit spatial RE
| augmented transcrip | correctly resolved  | incorrectly resolved

[02:34 --> 02:36] u7 :  There are no handles on the cabinets1.
[02:37 --> 02:38] u7 :  Or maybe these 2 are like push ones.
                        [u7 and u8 concurrently  looking at the kitchen cabinets]
                        these2 --> baseline: cabinets , system: kitchen cabinets

[02:39 --> 02:42] u7 :  And...
[02:44 --> 02:46] u8 :  But it's 3 a good design overall.
                        [u8 looking at the kitchen cabinets]
                        it's 3 --> baseline: design overall , system: kitchen overall design
[02:47 --> 02:51] u7 :  I know. Like I got a lot of time picking on it 4.
                        [u8 looking at the lamp]
                        it 4 --> baseline: the apartment , system: the lamp
[02:52 --> 02:53] u8 :  Maybe if you go upstairs, we'll find...
```

Figure 6: An example of the system's output showing the transcript augmentation corresponding to the snapshot in Figure 4, together with the final coreference resolution results for both the baseline and the proposed system.

context for where the users were and what they were watching, as well as understanding if they were performing pointing gestures, in order to understand what participants were referring to.

4.6.4 Labeling Target Geometry. Once all the implicit references' target objects are identified, we then annotate the corresponding

geometry name in the 3D scene by manually inspecting the geometry.

4.6.5 Inter Rater Reliability. All labeling was conducted by authors, where each label was initially assigned by a Labeler and

subsequently reviewed by a Rater. The Rater either agreed or disagreed with the Labeler's assignment, adding additional labels in cases of disagreement. Disagreements were documented with comments for each specific label and resolved through discussion at a later stage. To evaluate the consistency of the labeling process, we performed an inter-rater reliability analysis, initially resulting in an agreement percentage of 80.7%. Following discussion and resolution of disagreements, a final complete agreement on all labels was achieved among all labellers.

5 Results

We evaluate our system's ability to identify implicit spatial REs (Figure 8d), to identify the specific geometry referred to (Section 5.2), and to perform coreference resolution for the identified implicit spatial REs (Section 5.1). The described manual labeling procedure provide the ground truth for the evaluation. Performance measures of precision and F1-score are calculated for each individual user. Data aggregation is done per user rather than per conversation, considering each user's unique communication style, preference in using non-verbal communication, and vocabulary. Because the collected data does not conform to a normal distribution, we use a nonparametric bootstrap approach with 1,000 resamples to estimate the sampling distribution and derive 95% confidence intervals for evaluating differences.

5.1 Identify Implicit Spatial RE

To evaluate the performance of the implicit spatial RE identification phase, the F1-score was calculated (Figure 8 (d)). Out of the 350 Implicit REs we manually labelled, our system correctly identified 318 and misclassified 82. Therefore when interpreting this F1-score it is important to take into account that errors generated in the identification phase directly propagate to the coreference resolution phase.

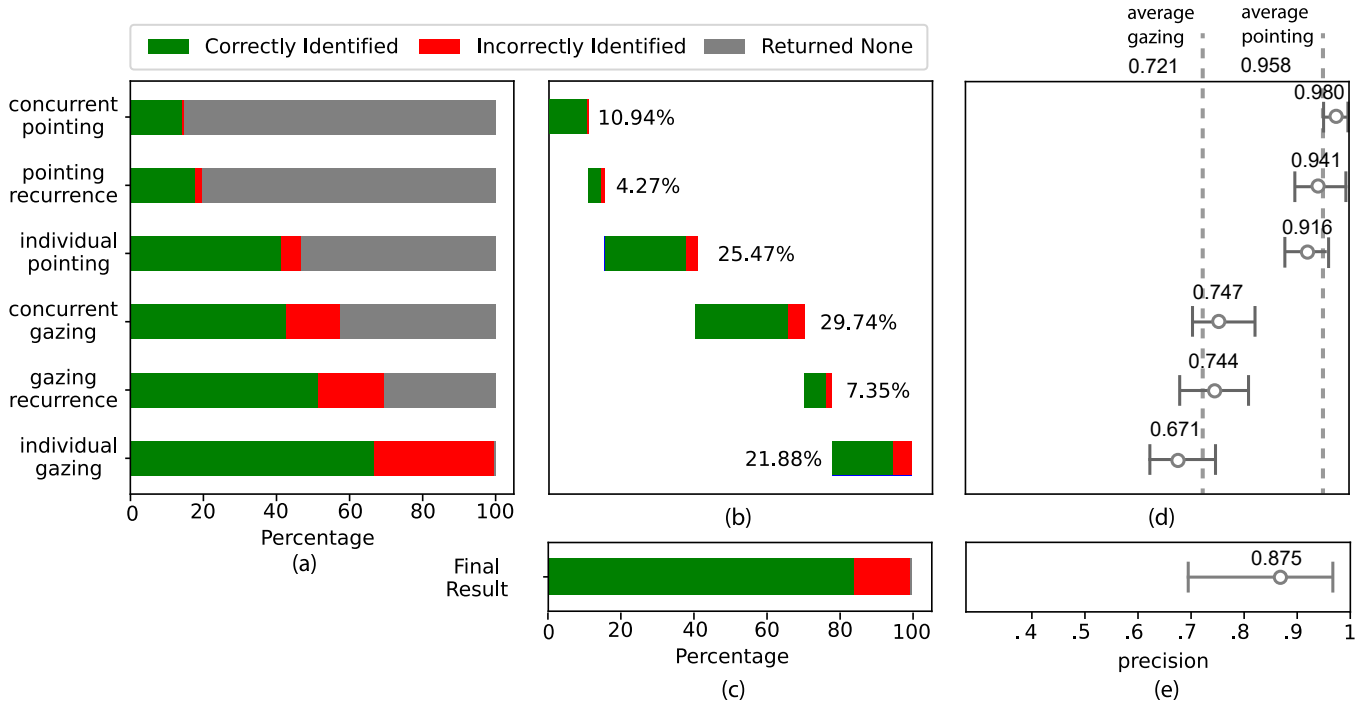### 5.2 Identifying Object of Interest

We evaluate the effectiveness of different behaviour measures, such as concurrent pointing, pointing recurrence, individual pointing, concurrent gazing, gazing recurrence, and individual gazing, in identifying the object of interest for each implicit RE. Since no established techniques utilize collaborative speech, gaze, and world semantics, direct comparisons with other methods are not feasible. Therefore, we evaluate the merit of each behavior combination in detecting the object of interest through an ablation study. This study assesses the contribution of each individual component to the overall effectiveness of our system (Figure 7 (a)). We do so by comparing the results with the ground truth to determine how many correct and incorrect objects each behaviour measure returns (Figure 7 (a,d)). While we compare all behaviours with one another, our system then selects a single one via the hierarchical selection mechanism described previously(section 3.4.5). We also calculate how many correct and incorrect objects our overall system identifies (Figure 7 (b,c,e)). Using the same ground truth as a reference, we assess our system's capability in recognising the object of interest for the implicit REs.

5.3 Effectiveness of Behaviours in Identifying Object of Interest

To better understand the impact of the various non-verbal behaviours (gaze, pointing), we compute the number of correctly identified objects for each behaviour, as well as how frequently a behaviour did not return any result because it did not happen in tandem with the RE (see Figure 7). The data reveals that pointing behaviours, with concurrent pointing, recurrent pointing, and individual pointing occurrences at 15%, 19.3%, and 46.4% respectively, are less frequent than gaze behaviour, which occurs at 56.8% for concurrent gazing, 69.3% for gazing recurrence, and 99.4% for individual gazing instances. Note that individual gazing is 99.4% rather than 100% due to low eye tracking confidence or blinking. Furthermore, both concurrent and recurrent behaviours manifest less often than their corresponding individual behaviours (see Figure 7). For each combination, we computed precision by dividing the number of correctly identified objects by the total number of returned objects. The final precision of our system in identifying the object of interest is 0.875. The results highlight that pointing is more precise than gazing, showing a .16 increment in the precision towards identifying the object of interest (Figure 7 (d)). There is a large difference between gazing and pointing (gazing: 95% CI [0.667, 0.766]; pointing: 95% CI [0.926, 0.980]), when comparing their precision.A comparison of concurrent pointing (Mean = 0.988, 95% CI [0.963,1.000]) and individual pointing (Mean = 0.926, 95% CI [0.881,0.967]) also indicates a strong difference in precision. Additionally, we assess how each combination influenced our final result through its selection in our hierarchical process (Section3.4.5). This is achieved by determining the frequency with which each behavior is chosen by our hierarchical selection process. Such frequency highlights the contribution of each behavior measure towards the goal of determining the object of interest. The results indicate that the combination contributing most to identifying the object of interest is concurrent gazing (representing concurrent visual attention) at 29.74%. This is followed by individual pointing at 25.47% and individual gazing at 21.88%. Moreover, it's evident that the majority of errors stem from individual gazing at 5.12% and concurrent gazing at 4.44%, with individual pointing contributing 3.07% of errors.

5.4 Coreference Resolution

We define a baseline consisting of GPT-4 performing coreference resolution on the speech-transcript. We compare it to GPT-4 performing coreference resolution on our system's augmented transcript. We compare both against the manually labelled ground truth (Section 4.6.3).We calculated F1-score for both baseline and system (Figure 8 (c)). When interpreting the F1-score, it's important to take into account that errors generated in the identification phase (Figure 8 (b)) directly propagate to the coreference resolution phase (Figure 8 (c)). This is because the errors between the identification system and the coreference are independent. For reference we plotted the F1-score from our implicit spatial RE's identification in Figure 8 (c) and Figure 9 (b)(d). Coreference resolution examples are available in the supplemental material 1–4.
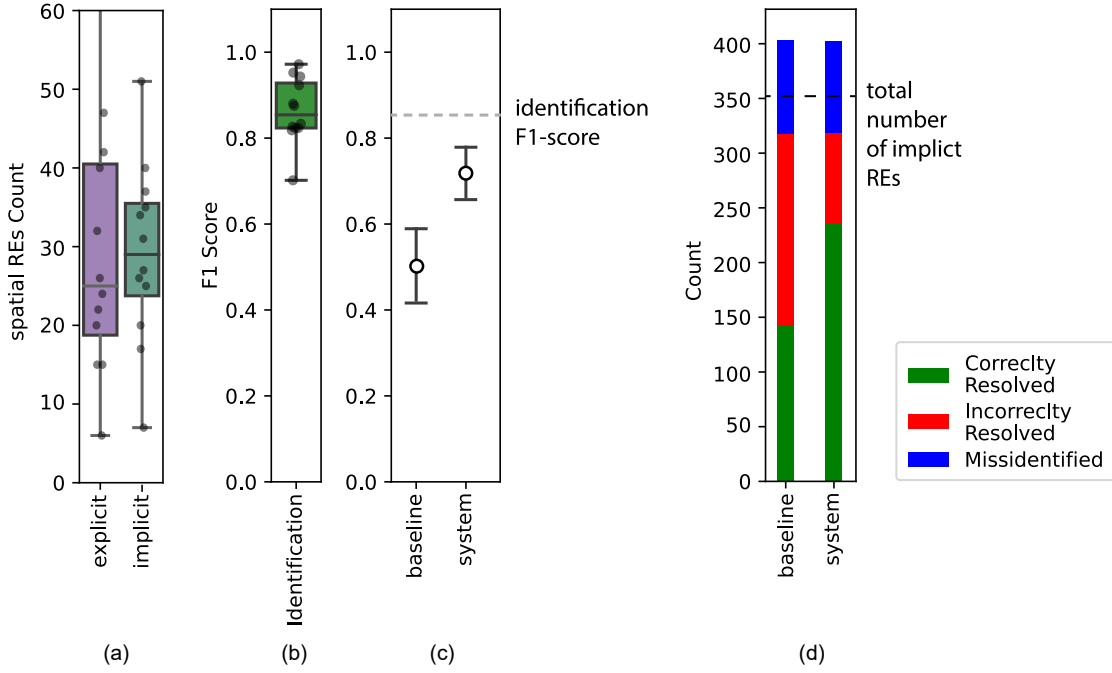
Figure 7: Identified Object of Interest: (a) Plot depicting the performance of each behaviour combination in terms of identifying the correct object, the incorrect object, or not identifying an object at all. The percentages for correct, incorrect, or none are calculated for each implicit spatial RE (y-axis: behaviour measures , x-axis: percentage of correct/incorrect/none). (b) Plot showcasing the frequency with which each behaviour measure was selected during our hierarchical selection process. Each behaviour's percentage bar is divided into 'correct' and 'incorrect'. The scenario of not identifying anything is excluded since this condition triggers our hierarchical selection process to fallback onto the subsequent behaviour measure(y-axis: behaviour measures , x-axis: percentage of correct/incorrect). (c) Plot presenting the percentages for correct,incorrect and none as results of the hierarchical selection process (y-axis: behaviour measures , x-axis: percentage of correct/incorrect). (d) Plot displaying the object identification precision, defined as the ratio of correctly predicted positives to total predicted positives, for each behaviour measure (y-axis: behaviour measures , x-axis: precision on a unit scale). (e) Plot highlighting the final precision of our object identification process.

5.4.1 Comparing our system to the baseline. We compare baseline and system F1-score and we observed improved coreference resolution performance, resulting in a .21 increase in the F1 score when comparing baseline with a precision of .507 to system with a precision of .723. Comparing baseline and the system, we see a clear difference baseline (95% CI: 0.507–0.584) and system (95% CI: 0.675–0.770) (Figure 8 (c)). Results are further analyzed by categorizing them into 'Correctly Resolved', 'Incorrectly Resolved', and 'Miss-identified' (Figure 8 (d)). The latter category includes instances where implicit REs were classified as explicit, non-REs were classified as implicit REs, or explicit REs were classified as implicit. Out of the 350 Implicit REs, there were 318 correctly identified. The baseline approach correctly resolved 142 (40.6%), while our proposed system accurately resolved 235 (67.1%) of them.

5.4.2 Understanding endophora and exophora. We compute the performance of both the baseline and system when resolving both endophora and exophora Figure 9 (a, b). Results show that for both baseline and system, there was an average increase of

0.44 in the F1 score for the endophora group (baseline and system combined). This effect is more pronounced in the baseline, which showed a 0.475 point increase in the F1 score. A comparison between the baseline endophora (95% CI: 0.133–0.211) and baseline exophora (95% CI: 0.544–0.696) revealed a clear difference, with the baseline exophora group achieving substantially higher performance. This more substantial increase aligns with expectations of endophora being more challenging to resolve than the exophora. Furthermore we observed a 0.25 increase in F1 score from baseline exophora to system exophora, (baseline exophora: 95% CI 0.544–0.696; system exophora: 95% CI 0.809–0.882), indicating a substantial improvement. This is consistent with expectations, as the system benefits from additional contextual information and enriched features that are particularly effective for resolving explicit references, such as those present in exophora scenarios. Finally the system endophora improved by 0.15 in F1 score over baseline endophora, (baseline endophora: 95% CI 0.133–0.211; system endophora: 95% CI 0.342–0.576), indicating a large improvement. This indicates that the supplementary information introduced by

Figure 8: (a) Plot depicting the count of explicit and implicit REs labelled for each participant (y-axis: count per participant, x-axis: implicit/explicit). (b) Plot presenting the f1-score for identifying implicit spatial REs. (c) Plot comparing the f1-score for the coreference resolution task executed by GPT-4 using the speech transcript (baseline) versus our augmented transcript with metadata (y-axis: f1 score, x-axis: baseline/system). (d) Plot depicting the count of correctly resolved, incorrectly resolved and miss-identified (y-axis: REs count, x-axis: baseline and our proposed method).

our pipeline aids the GPT-4 model in differentiating between previously mentioned entities and subsequent entities, enhancing its coreference resolution capability. Furthermore, by observing the higher number of endophora occurrences compared to exophora (Figure 9 (a)), we can infer that this is where the system achieves most of its performance improvements.

5.4.3 Understanding object and place references. Lastly, we calculated the F1-score based on whether the reference's target entity was labeled as an `Object` or a `Place` (Figure 9 (c, d)). We observed that coreference resolution tends to be less accurate across all models when the reference's target entity is a "place." However, this difference is more pronounced in the `system` model. Specifically, `system object` exhibited a 0.18 increase in the F1 score, with moderately-overlapping confidence intervals (95% CI: 0.717–0.861 for `system object` vs. 0.555–0.745 for `system place`). This finding aligns with our expectations, as references to `places` may lack clear geometric boundaries in the 3D scene, with identified objects (e.g., a fridge) representing only a part of the place (e.g., the kitchen) rather than encapsulating it entirely.
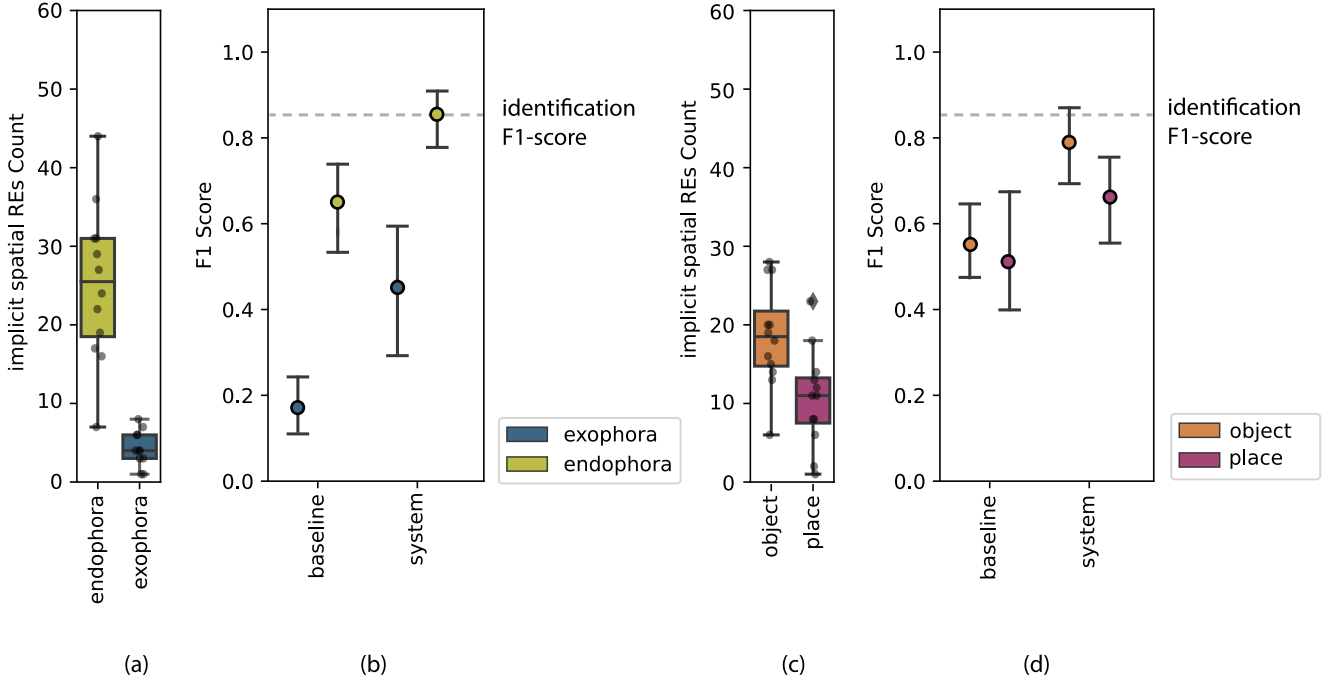
6 Discussion

We discuss our system's ability to identify and resolve spatial referring expressions (REs) through a transcript that has been augmented

with contextual metadata about the scene via nonverbal communication (i.e., objects of interest, gaze, and pointing behaviour in relation to implicit spatial REs).

6.1 Identifying implicit spatial referring expressions

Our system's first step is to identify implicit spatial referring expressions (REs) using GPT-4. Out of 350 implicit REs, the system successfully identified 318 but misclassified 63. Miss-classifications primarily involved two types: firstly, explicit REs were mistakenly classified as implicit, such as in "This is a cool little spot," where "spot" accompanied the RE, leading to incorrect classification. Secondly, expressions like "I like that." in contexts such as "There is no TV, this means digital detox." meant non-physical contexts and wrongly labeled as spatial REs. These errors suggest a need for future systems to incorporate accuracy estimates or verify if verbally expressed entities exist in the 3D environment. Several alternative methods for parsing text effectively and identifying implicit spatial REs exist, such as Stanford's CoreNLP parser [24]. However, we chose a GPT model because our intention was to identify implicit spatial referring expressions (REs) and avoid implicit REs that refer to abstract ideas and hypothetical objects. Given the GPT model's nuanced approach and its ability to detect subtle patterns in language a GPT model is more effective for this task compared to rule-based or traditional machine learning systems, which might

Figure 9: (a) Plot illustrating the count of implicit REs categorized as endophora or exophora for each participant (y-axis: count per participant, x-axis: endophora/exophora). (b) Plot of the F1-score for the coreference resolution task by GPT-4, comparing the baseline with our system for two identified groups exophora/endophora (y-axis: f1 score, x-axis: transcript vs. transcript+metadata). (c) Plot showing the count of implicit spatial REs targeting either an object or place (y-axis: count per participant, x-axis: object/place). (d) Plot of the F1-score for the coreference resolution task by GPT-4, comparing the baseline with our system for the two groups object/place (y-axis: f1 score, x-axis: baseline/system).

overlook such patterns. While there were several possible Large Language Models (LLM) that we could have chosen such as Llama [38] or Falcon [29] we used the GPT-4 API for the convenience of not having to run the model locally.

6.2 The merit of non-verbal synergies towards identifying the referent

Our findings underscore the value of analyzing non-verbal synergies—collaborative patterns in gaze and pointing—to resolve referents in complex human-human dialogues. This approach advances beyond prior work, which has largely focused on single-user, human-machine interactions with isolated deictic gestures [34, 25, 7, 27]. While other research has acknowledged the significance of synchronized gaze dynamics in collaborative tasks [28, 31, 41, 39], this information had not been leveraged as a direct retrieval method for identifying an object of interest. Our work expands on this by demonstrating that synergistic behaviors in both pointing (concurrent and recurrent) and gaze (concurrent and recurrent visual attention) are effective retrieval methods during collaborative discussions. Furthermore, our results revealed a critical distinction between these two modalities: pointing is a significantly more accurate predictor of the intended referent than gaze (Figure 7 (e)). This key finding was foundational to our system's design. We established a hierarchical selection method that prioritizes the more

reliable, volitional behavior (pointing) over the more reflexive, less precise behavior (gazing). This ensures that our system leverages the most accurate non-verbal cue available to resolve ambiguity.

6.3 Improving resolution of both endophora and exophora

Our system performs coreference resolution on each of the identified implicit spatial REs, resulting in an improved coreference resolution process for all implicit REs. While the results for *exophora* showed significant improvements compared to the baseline, there were very few cases of them in our dataset (occurrences Figure 9 (a)). However, for those cases in the exophora group, it is evident that our system's advantage lies in incorporating novel information from the 3D scene metadata. The most common implicit REs were *endophora* (referents present in the text but in a different sentence). Results underscore that, for the endophora group, the system's contribution is primarily in enabling the disambiguation of existing entities within the text. Therefore, our technique improves the comprehension of conversations not only by introducing new information but also by facilitating the coreference resolution process of selecting existing entities within the text through the augmented transcript.

6.4 Extending Vision-Language Models

Vision-language models (VLMs), aim to understand visual context using neural networks like CNNs, MultiModal BERT, ResNet, or RetinaNet [12, 16, 17, 46, 45]. These models analyze the visual scene, segment entities, and model relationships based on textual mentions. However, they can be prone to errors in scenarios where multiple objects within the user's field of view could relate to an ambiguous referring expression, such as saying "I like that cake" in a cake shop. In contrast, our approach offers an alternative method to model relationships between entities in the visual scene without relying on detailed visual analysis like VLMs. We emphasize the modeling of relationships between verbal communication and scene objects by leveraging dynamics in individual and collaborative non-verbal communication. For instance, collaborative eye-gaze and pointing can serve as additional inputs to neural networks, providing crucial information for accurately modeling relationships. This includes using eye gaze and pointing to indicate where visual attention aligns during verbal communication, akin to the temporal parsing of mouse traces demonstrated by Goel et al. [9]. This paper extends prior work on visual coreference resolution [46, 45, 15, 12, 17, 10], highlighting novel inputs such as synergistic gaze and pointing that can operate alongside or independently of visual scene analysis, thereby contributing to higher accuracy in coreference resolution.

6.5 Reliance on the 3D model and its granularity

It is important to note that referents varied in granularity, as no specific constraints were imposed on the participants (i.e. to only refer to objects). When a spatial reference refers to a place (e.g., kitchen, bathroom, or other area), which is an aggregate of objects, the metadata (3d model name) might not represents the entity the speaker is referencing. While some object names might include the location's name (e.g., "kitchen cabinets" or "bathroom sinks"), others might not (for instance, "faucet" could be located either in the kitchen or bathroom). Our analysis shows a lower F1 score when resolving place-related entities compared to objects. For instance, if the referent is the kitchen but the speaker is pointing the laser at the fridge, while the fridge is part of the kitchen, it is not the *kitchen* itself. By describing this verbally (e.g., "[P1 pointing at the fridge]"), we do not constrain GPT-4, which, being an abstractive generative model, can infer, based on the richness of the RE and surrounding dialogue, that the referent might be the kitchen and not the fridge. Nevertheless, our results indicate that our system performs statistically worse when the referent is a place than when it is an object. We acknowledge that our 3D model could also be set up with more granular object names and using a hierarchical structure (e.g., a "faucet" would be a descendant of a "sink", which would be part of "kitchen"). Future work should explore the appropriate level of abstraction and granularity in the 3D model to address this limitation.

6.6 Utility of coreference resolution for immersive conversations recordings

There is a growing interest in spatial computing devices such as VR/AR for professional collaborative applications. Consequently, as user adoption increase, there will be an increasing number of conversations, such as meetings and design reviews, that will be conducted in VR and recorded for later review, summarization, and archiving. As more and more conversations get recorded, the desire to automatically process these conversations and extract salient moments will increase. However, human-human dialogues present challenges for machine comprehension for various reasons, such as determining the object being referred to using implicit referring expressions. Therefore, we argue that the impact of this work, by leveraging synergies of non-verbal cues to detect the referent of REs, will be instrumental in enhancing the accuracy and reliability of machine comprehension in human-human interactions, ultimately contributing to more effective and meaningful use of the recording capabilities of these spatial devices.

7 Conclusion

To address the issue of identifying objects of interest during an immersive conversation, we developed a system that leverages transcribed text, eye tracking, and laser-pointing data to resolve coreferences. It detects implicit REs, identifies the object of attention in the scene using non-verbal cues, generates a textual description of the object of interest, and performs coreference resolution using the textual description generated. By analyzing the data collected during a 12-participant user study, we find that gaze and pointing data add value, with pointing data often providing highly precise (though infrequent) information about which objects share focus. Compared to a baseline with only speech information which resolved 142 cases, our system resolved 235 implicit REs showing an improvement of 26.5%.

References

[1] Arkio ehf. 2023. Arkio. https://www.arkio.is/. Accessed: [2023-12-09]. (2023).

[2] Autodesk. 2023. The Wild. https://thewild.com/. Accessed: [2023-12-09]. (2023).

[3] Autodesk. 2023. WorkshopXR. https://workshopxr.autodesk.com/. Accessed: [2023-12-09]. (2023).

[4] Jiaxin Bai, Hongming Zhang, Yangqiu Song, and Kun Xu. 2021. Joint Coreference Resolution and Character Linking for Multiparty Conversation. In *EACL*. https://aclanthology.org/2021.eacl-main.43.pdf.

[5] Richard A. Bolt. 1980. "Put-That-There": Voice and Gesture at the Graphics Interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques* (SIGGRAPH '80). Association for Computing Machinery, Seattle, Washington, USA, 262–270. ISBN: 0897910214. doi:10.1145/800250.807503.

[6] Riccardo Bovo, Daniele Giunchi, Alebri Muna, Anthony Steed, Enrico Costanza, and Thomas Heinis. 2022. Cone of Vision as a Behavioural Cue for VR Collaboration. *Taipei 2022: Conference on Computer Supported Cooperative Work and Social Computing, November 12-16, 2022, Taipei, Taiwan*, 1, 1. doi:10.1145/3555615.

[7] Riccardo Bovo, Daniele Giunchi, Ludwig Sidenmark, Joshua Newn, Hans Gellersen, Enrico Costanza, and Thomas Heinis. 2023. Speech-Augmented Cone-of-Vision for Exploratory Data Analysis. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, (Apr. 2023). ISBN: 9781450394215. doi:10.1145/3544548.3581283.

[8] Sarah D'Angelo and Andrew Begel. 2017. Improving communication between pair programmers using shared gaze awareness. *Conference on Human Factors in Computing Systems - Proceedings*, 2017-Janua, 6245–6255. ISBN: 9781450346559. doi:10.1145/3025453.3025573.

[9] Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. 2023. Who Are You Referring To? Coreference Resolution In Image Narrations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. (Oct. 2023), 15247–15258.

[10] Danfeng Guo, Arpit Gupta, Sanchit Agarwal, Jiun-Yu Kao, Shuyang Gao, Arijit Biswas, Chien-Wei Lin, Tagyoung Chung, and Mohit Bansal. 2022. GRAVL-BERT: Graphical Visual-Linguistic Representations for Multimodal Coreference Resolution. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju,

Republic of Korea, (Oct. 2022), 285–297. https://aclanthology.org/2022.coling-1.22.

[11] Jon Hindmarsh, Mike Fraser, Christian Heath, Steve Benford, and Chris Greenhalgh. 1998. Fragmented Interaction: Establishing Mutual Orientation in Virtual Environments. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work* (CSCW '98). Association for Computing Machinery, Seattle, Washington, USA, 217–226. ISBN: 1581130090. doi:10.1145/289444.289496.

[12] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3D-LLM: Injecting the 3D World into Large Language Models, (July 2023). http://arxiv.org/abs/2307.12981.

[13] Allison Jing, Kieran William May, Mahnoor Naeem, Gun Lee, and Mark Billinghurst. 2021. EyemR-Vis: Using Bi-Directional Gaze Behavioural Cues to Improve Mixed Reality Remote Collaboration. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, (May 2021). ISBN: 9781450380959. doi:10.1145/3411763.3451844.

[14] Seungwon Kim, Gun Lee, Mark Billinghurst, and Weidong Huang. 2020. The Combination of Visual Communication Cues in Mixed Reality Remote Collaboration. *Journal on Multimodal User Interfaces*, (July 2020), 1–15. doi:10.1007/s12193-020-00335-x.

[15] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What Are You Talking About? Text-to-Image Coreference. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 3558–3565. doi:10.1109/CVPR.2014.455.

[16] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. [n. d.] What are you talking about? Text-to-Image Coreference. Tech. rep.

[17] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations, (Apr. 2021). http://arxiv.org/abs/2104.08667.

[18] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S Rodriguez, and Jon E Froehlich. 2024. Gazepointar: a context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality.

[19] Yifan Liu, Fadi Castronovo, John Messner, and Robert Leicht. 2020. Evaluating the Impact of Virtual Reality on Design Review Meetings. *Journal of Computing in Civil Engineering*, 34, 1, 04019045.

[20] Yifan Liu, Jennifer Lather, and John Messner. 2014. Virtual Reality to Support the Integrated Design Process: A Retrofit Case Study. In *Computing in civil and building engineering (2014)*, 801–808.

[21] Juan López-Tarruella Maldonado, Juan Luis Higuera Trujillo, Susana Iñarra Abad, M ª Carmen Llinares Millán, Jaime Guixeres Provinciales, and Mariano Alcañiz Raya. 2018. Virtual Reality as a Tool for Emotional Evaluation of Architectural Environments. In *Architectural Draughtsmanship: From Analog to Digital Narratives 16*. Springer, 889–903.

[22] Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2022. A survey of deep learning for mathematical reasoning. *arXiv preprint arXiv:2212.10535*.

[23] Karthik Mahadevan, Qian Zhou, George Fitzmaurice, Tovi Grossman, and Fraser Anderson. 2023. Tesseract: Querying Spatial Design Recordings by Manipulating Worlds in Miniature. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, (Apr. 2023). ISBN: 9781450394215. doi:10.1145/3544548.3580876.

[24] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010.

[25] Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing Mobile Voice Assistants with WorldGaze. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, (Apr. 2020). ISBN: 9781450367080. doi:10.1145/3313831.3376479.

[26] Sven Mayer, Valentin Schwind, Robin Schweigert, and Niels Henze. 2018. The Effect of Offset Correction and Cursor on Mid-Air Pointing in Real and Virtual Environments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18). Association for Computing Machinery, Montreal QC, Canada, 1–13. ISBN: 9781450356206. doi:10.1145/3173574.3174227.

[27] Darius Miniotas, Oleg Špakov, Ivan Tugoy, and I. Scott MacKenzie. 2006. Speech-Augmented Eye Gaze Interaction with Small Closely Spaced Targets. In *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications* (ETRA '06). Association for Computing Machinery, San Diego, California, 67–72. ISBN: 1595933050. doi:10.1145/1117309.1117345.

[28] Robert Moulder, Brandon Booth, Angelina Abitino, and Sidney D'Mello. 2023. Recurrence Quantification Analysis of Eye Gaze Dynamics during Team Collaboration. In Association for Computing Machinery, (Mar. 2023), 430–440. ISBN: 9781450398657. doi:10.1145/3576050.3576113.

[29] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only, (June 2023). http://arxiv.org/abs/2306.01116.

[30] Anna Penzkofer, Philipp Müller, Felix Bühler, Sven Mayer, and Andreas Bulling. 2021. Conan: a usable tool for multimodal conversation analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, 341–351.

[31] Sami Pietinen, Roman Bednarik, Tatiana Glotova, Vesa Tenhunen, and Markku Tukiainen. 2008. A Method to Study Visual Attention Aspects of Collaboration: Eye-Tracking Pair Programmers Simultaneously. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (ETRA '08). Association for Computing Machinery, Savannah, Georgia, 39–42. ISBN: 9781595939821. doi:10.1145/1344471.1344480.

[32] Thammathip Piumsomboon, Arindam Dey, Barrett Ens, Gun Lee, and Mark Billinghurst. 2019. The Effects of Sharing Awareness Cues in Collaborative Mixed Reality. *Frontiers Robotics AI*, 6, FEB. doi:10.3389/frobt.2019.00005.

[33] RealWear, Inc. 2023. RealWear. https://www.realwear.com/. Accessed: [2023-12-09]. (2023).

[34] Yevhen Romaniak, Anastasiia Smielova, Yevhenii Yakishyn, Valerii Dziubliuk, Mykhailo Zlotnyk, and Oleksandr Viatchaninov. 2020. Nimble: Mobile Interface for a Visual Question Answering Augmented by Gestures. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (UIST '20 Adjunct). Association for Computing Machinery, Virtual Event, USA, 129–131. ISBN: 9781450375153. doi:10.1145/3379350.3416153.

[35] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying Fixations and Saccades in Eye-Tracking Protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (ETRA '00). Association for Computing Machinery, Palm Beach Gardens, Florida, USA, 71–78. ISBN: 1581132808. doi:10.1145/355017.355028.

[36] Bertrand Schneider and Roy Pea. 2013. Real-Time Mutual Gaze Perception Enhances Collaborative Learning and Collaboration Quality. *International Journal of Computer-Supported Collaborative Learning*, 8, 4, 375–397. ISBN: 1141201391814. doi:10.1007/s11412-013-9181-4.

[37] Maurício Sousa, Rafael Kuffner Dos Anjos, Daniel Mendes, Mark Billinghurst, and Joaquim Jorge. 2019. Warping deixis: Distorting Gestures to Enhance Collaboration. In *Conference on Human Factors in Computing Systems - Proceedings*. Vol. 12. Association for Computing Machinery, New York, NY, USA, (May 2019), 1–12. ISBN: 9781450359702. doi:10.1145/3290605.3300838.

[38] Hugo Touvron et al. 2023. LLaMA: Open and Efficient Foundation Language Models, (Feb. 2023). http://arxiv.org/abs/2302.13971.

[39] Maureen Villamor and Ma Mercedes Rodrigo. 2018. Predicting Successful Collaboration in a Pair Programming Eye Tracking Experiment. In *UMAP 2018 - Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* number July, 263–268. ISBN: 9781450357845. doi:10.1145/3213586.3225234.

[40] VRChat Inc. 2023. VRChat. https://vrchat.com/. Accessed: [2023-12-09]. (2023).

[41] Hana Vrzakova, Mary Jean Amon, Angela E.B. Stewart, and Sidney K. D'Mello. 2019. Dynamics of Visual Aention in Multiparty Collaborative Problem Solving using Multidimensional Recurrence antification Analysis. *Conference on Human Factors in Computing Systems - Proceedings*, 14. ISBN: 9781450359702. doi:10.1145/3290605.3300572.

[42] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do nlp models know numbers? probing numeracy in embeddings. *arXiv preprint arXiv:1909.07940*.

[43] Nelson Wong and Carl Gutwin. 2014. Support for Deictic Pointing in CVEs: Still Fragmented after All These Years'. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '14). Association for Computing Machinery, Baltimore, Maryland, USA, 1377–1387. ISBN: 9781450325400. doi:10.1145/2531602.2531691.

[44] Nelson Wong and Carl Gutwin. 2010. Where Are You Pointing? The Accuracy of Deictic Pointing in CVEs. In *Conference on Human Factors in Computing Systems - Proceedings*. Vol. 2. ACM Press, New York, New York, USA, 1029–1038. ISBN: 9781605589299. doi:10.1145/1753326.1753480.

[45] Xintong Yu, Hongming Zhang, Ruixin Hong, Yangqiu Song, and Changshui Zhang. 2022. VD-PCR: Improving Visual Dialog with Pronoun Coreference Resolution. *Pattern Recognition*, 125, (May 2022). doi:10.1016/j.patcog.2022.108540.

[46] Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. What You See is What You Get: Visual Pronoun Coreference Resolution in Dialogues, (Sept. 2019). http://arxiv.org/abs/1909.00421.

[47] Xiaoyu Zhang, Jianping Li, Po Wei Chi, Senthil Chandrasegaran, and Kwan Liu Ma. 2023. ConceptEVA: Concept-Based Interactive Exploration and Customization of Document Summaries. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, (Apr. 2023). ISBN: 9781450394215. doi:10.1145/3544548.3581260.

[48] Yanxia Zhang, Ken Pfeuffer, Ming Ki Chong, Jason Alexander, Andreas Bulling, and Hans Gellersen. 2017. Look together: using gaze for assisting co-located collaborative search. *Personal and Ubiquitous Computing*, 21, 1, 173–186. doi:10.1007/s00779-016-0969-x.