

Imperial College London
Department of Computing

Leveraging the multi-modal nature of communication in immersive XR environments

Riccardo Bovo BSc, MSc.

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of Imperial College London.
February 2026.

Abstract

In the evolving landscape of immersive spatial computing, such as extended reality (XR), effective collaboration remains a cornerstone of productivity and innovation. In traditional co-located work, collaborators naturally align both their visual and cognitive attention through a blend of gaze, gesture, and referring expressions, enabling rapid establishment of common ground and precise coordination.

This thesis investigates how such attentional cues can be reconstructed and visualised in immersive collaboration settings, particularly on low-cost VR/AR headsets that lack eye-tracking. Rather than treating raw gaze as the sole proxy for intent, the work adopts a *complementary reconstruction* approach that combines naturally occurring collaborative behaviours (head movement and verbal communication) with the shared spatial context of the task to approximate a partner’s attention and expose it through adaptive visual cues.

Across a series of empirical studies, the thesis introduces and evaluates three components. First, a probabilistic *Cone-of-Vision* derived from head orientation supports broad mutual awareness without requiring eye-tracking. Second, a deep learning model refines this coarse signal, reducing spatial ambiguity while avoiding the computational cost of full computer vision pipelines. Third, a *Speech-Augmented Cone-of-Vision* integrates referring expressions in real time, dynamically scaling from scene-level awareness to object-level focus as collaborators speak.

These findings challenge the assumption that raw sensor data is the gold standard for visual attention cues. By demonstrating the importance of broad attention cues for collaborative alignment and the role of speech in reducing ambiguity, this thesis proposes a shift from merely replicating gaze mechanics to visualizing the higher-order of cognitive attention. Ultimately, this work underscores that leveraging naturally occurring behaviours is not just a hardware workaround, but a fundamental requirement for the hyper-productive, augmented collaboration promised by spatial computing.

Statement of Originality

I hereby declare that the work presented in this thesis is the result of my own research, unless explicitly stated otherwise. This thesis has not been submitted, either in whole or in part, for any degree or diploma at any other university or institution. All sources of information, assistance, and materials have been appropriately acknowledged and cited within the text.

Where contributions from other researchers have been included, their input is clearly specified in the relevant sections of the thesis. The research, design, and development of the ideas presented herein are, to the best of my knowledge, original and a product of my own independent work.

Riccardo Bovo
Imperial College London
2025

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International License (CC BY-NC). Under this license, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work on condition that you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the license terms clear to others by naming the license and linking to the license text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for any uses of this work that are not included in this license or permitted under UK Copyright Law.

Acknowledgements

I would like to express my deepest gratitude to the many people who have supported me throughout this journey.

My supervisor, Thomas Heinis, for his continuous encouragement, trust in me, which granted great freedom in my research, and his unconditional support throughout this journey, all of which have been invaluable and have shaped both my research and personal growth during the PhD. My second supervisor, Enrico Costanza, for their support and critical eye throughout my work. Their contributions were essential to the progress of this thesis. The other researchers and collaborators who contributed to this work such as Anthony Steed, Hans Gellersen and Ludwig Sidenmark whether through insightful discussions, collaborative projects, or sharing knowledge. I especially want to mention Daniele Giunchi for his specific contributions and support.

Thank you all for making this accomplishment possible.

Dedication

I dedicate this to my wife Petra Vukmirovic, whose unwavering support, patience, and belief in me provided the strength I needed to complete this journey. Her encouragement has been a constant source of motivation, and I could not have done this without her.

Glossary

- 6DOF** Six Degrees of Freedom. Refers to the ability to move freely in three-dimensional space (translation in x, y, and z axes) and rotate around those axes (pitch, yaw, and roll). viii
- AFM** Average Fixation Map. A visualisation technique that aggregates gaze data to show the average distribution of visual attention across a scene. viii
- AI** Artificial Intelligence. The simulation of human intelligence processes by machines, especially computer systems. viii
- ANOVA** Analysis of Variance. A statistical method used to analyse the differences between group means and their associated procedures. viii
- AOI** Area of Interest. A specific region or element within a visual scene that is relevant to a task or analysis. viii
- AR** Augmented Reality. A technology that overlays digital information onto the real world, enhancing the user's perception of reality. viii
- Avatar** A digital representation of a user in a virtual environment. viii
- bi-directional** Involving or operating in two directions. In the context of visual attention cues, it means that both the user and their collaborator can see each other's focus of attention. viii
- bi-gram** A sequence of two adjacent elements from a string of tokens, typically used in natural language processing. viii

CAVE Cave Automatic Virtual Environment. A large, immersive display environment where projectors display images on the walls, floor, and ceiling, surrounding the user in a virtual world. viii

CNN Convolutional Neural Network. A deep learning algorithm commonly used for image recognition and processing. viii

CoV Cone of Vision. A visualisation technique that displays the area where a user is most likely focusing their attention based on head movements. viii

CSCW Computer-Supported Cooperative Work. A field of study that focuses on how computer technology can support collaborative activities and social interactions. viii

CVE Collaborative Virtual Environment. A virtual space where multiple users can interact and collaborate in real-time. viii

deictic Relating to or denoting a word or expression whose meaning is dependent on the context in which it is used (e.g., here, you, me, that one there). viii

EDA Exploratory Data Analysis. An approach to analyzing data sets to summarise their main characteristics, often with visual methods. viii

egocentric Centered on or from the perspective of the individual. In VR, an egocentric view shows the virtual world from the user's own perspective. viii

explicit Clearly stated or defined; leaving no room for confusion or doubt. In the context of referring expressions, it means that the referent is directly named or identified. viii

eye-gaze The direction in which the eyes are looking. viii

eye-tracking The process of measuring eye movements and gaze direction, typically using specialized cameras or sensors. viii

f1-score The harmonic mean of precision and recall, used to evaluate the performance of a classification model. viii

fixation A period of time during which the eyes remain relatively still on a particular point in space. viii

FoV Field of View. The extent of the observable world that is seen at any given moment. viii

HCI Human-Computer Interaction. The study of how people interact with computers and the design of computer systems that are easy and enjoyable to use. viii

HMD Head-Mounted Display. A device worn on the head that presents visual information to the user, often used in VR and AR systems. viii

ICVE Immersive Collaborative Virtual Environment. A virtual environment that supports collaboration and interaction among multiple users in an immersive setting. viii

iso-contour A line on a map or chart connecting points of equal value. viii

iso-hypes Contour lines on a map representing points of equal elevation. viii

LTR Left-to-Right. Refers to the direction in which text is written and read in many languages. viii

MGF Multivariate Gaussian Fitting. A statistical method used to fit a multivariate Gaussian distribution to a set of data points. viii

MLP Multi-Layer Perceptron. A type of artificial neural network with multiple layers of interconnected nodes, used for various machine learning tasks. viii

MR Mixed Reality. A technology that blends the real and virtual worlds, allowing users to interact with both physical and digital objects. viii

Contents

Abstract	i
Statement of Originality	ii
Copyright Declaration	iii
Acknowledgements	iv
1 Introduction	1
1.1 Summary	1
1.2 Research Objectives	6
1.3 Contributions	7
1.4 Thesis Structure	11
1.5 Contributing Papers	13
2 Literature Review	16
2.1 Evolution of XVA Environments	17
2.2 Mutual Awareness of Visual Attention in Collaborative Work	19

2.3	Pointing-Based Communication in Collaborative Work	26
2.4	Gaze Prediction in XR	30
2.5	Gaps in the Literature Addressed by this Thesis	34
2.6	Conclusion	37
3	Cone-of-Vision	39
3.1	Study	40
3.2	Results	50
3.3	Discussion	60
4	Model for gaze probability regions	64
4.1	Model	66
4.2	Results	73
4.3	Discussion	76
5	Speech-Augmented Cone-of-Vision	78
5.1	Study	79
5.2	Results	88
5.3	Qualitative Analysis	96
5.4	Discussion	100
6	Conclusion	104
6.1	Discussion	105
6.2	Impact and Implications	108

6.3 Limitations and Future Work 109

6.4 Closing Remarks 114

Bibliography **114**

List of Tables

3.1	Questionnaires	49
3.2	Statistical Analysis	51
3.3	Statistical Analysis of Questionnaires	54
5.1	Verbal communication taxonomy.	85

List of Figures

3.1	Cone of Vision	39
3.2	First Experiment Conditions	41
3.3	Layout of the first experiment	42
3.4	Cone of Vision Validation	43
3.5	Cone of Vision Iso-Hypses Metaphor	44
3.6	Field of View —Visualisation History	45
3.7	Cone of Vision Dimensions and Intersection with 2D Screens	45
3.8	First Experiment Procedure	48
3.9	Behavioural Questionnaire Results	51
3.10	Questionnaires Answers	54
4.1	Model for Gaze Probability Regions in Collaborative VR	64
4.2	Data Distributions and Regions	68
4.3	Deep Learning Pipeline	69
4.4	Standard deviation and Percentage of Data Radial Diagram	71
4.5	Standard Deviation and Percentage of Data Histograms	71

4.6	RMSE and Percentage of Data Captured Radial Diagram	74
4.7	Histogram of RMSE and Percentage of Data Captured	74
5.1	Speech-Augmented Cone-of-Vision visual explanation diagram	78
5.2	Second Experiment Conditions	81
5.3	Second Experiment Layout	82
5.4	Speech-Augmented Cone-of-Vision Algorithm's Diagram	83
5.5	Visual Coupling in Area of Interest	88
5.6	Individual Visual Attention	90
5.7	Analysis of Verbal References	92
5.8	Speech Augmented Gaze Approximation	94
5.9	Questionnaire Results	96

Chapter 1

Introduction

1.1 Summary

Virtual Reality (VR) has moved beyond its origins in entertainment and gaming into sectors such as education [1], healthcare [2], and manufacturing [3, 4], where it enhances spatial understanding and memory retention. It is now increasingly adopted for knowledge work: companies such as Apple position devices like the Vision Pro [5] as productivity tools, and market forecasts predict substantial workplace uptake [6]. Academic research highlights portable virtual offices as productivity enhancers that provide extensive screen “real estate” [7, 8] and align with flexible working styles, including commuting and remote work [9, 10].

VR environments can replicate verbal and non-verbal cues essential in traditional collaboration settings—such as gestures, facial expressions, and eye contact [11, 12, 13]—and can augment these cues by leveraging user-facing sensors such as eye-tracking, face cameras [14], or body-facing cameras that capture and reconstruct user movements [15]. However, the market is dominated by affordable, untethered headsets that lack eye-tracking [16], restricting the transmission of precise visual attention cues that are critical for mutual awareness and shared understanding in collaborative environments [17, 18, 19].

Yet the challenge extends beyond the mere absence of sensors. Even when eye-tracking is

available, raw gaze data is not a perfect proxy for user intent [20, 21]. Effective alignment during collaboration depends on *cognitive* attention—what a partner is thinking about and referring to—rather than just visual attention, which only indicates where their eyes are pointed [22, 23]. Visual attention is noisy: gaze can be drawn reflexively to motion or brightness [24, 25, 26], attention shifts between broad scene-level context and narrow detail [27, 28], and cognitive attention may decouple from visual gaze entirely—for example, when a user looks at one object while mentally referring to another element seen previously [29, 30, 31, 32]. Consequently, simply displaying a raw gaze cursor (even if hardware permitted) is not sufficient; collaborative attention cues must filter noise, adapt their spatial scale, and respect user agency.

To address both the hardware limitations of low-cost devices and the conceptual need for cognitively aligned cues, this thesis adopts a complementary reconstruction approach. Modern XR systems can capture multiple naturally occurring signals related to attention: head motion provides a broad, low-resolution indication of attentional direction, while verbal communication adds a semantic layer, revealing referential intent through spoken keywords and descriptions.

Core Hypothesis: This thesis posits that naturally occurring collaborative behaviours, specifically head movement and verbal communication, can be leveraged as implicit inputs to reconstruct and display users’ attention.

This thesis operationalises the principle of complementary reconstruction through three progressive contributions. First, I establish Head Orientation as a foundational proxy for broad awareness via the *Cone-of-Vision (CoV)* (Chapter 3), a probabilistic visualisation that filters reflexive noise and reduces clutter compared to traditional head based visual cues. Second, I refine this signal with a Deep Learning Model (Chapter 4) that reduces spatial ambiguity without the computational overhead of full gaze tracking. Third, I integrate Verbal Communication (Chapter 5) to achieve dynamic adaptation: the *Speech-Augmented CoV* scales focus from broad awareness to specific targets in real-time, effectively reconstructing the narrowing of cognitive attention. Together, these methods form a hardware-independent framework that enhances mutual awareness across the full spectrum of XR devices, narrowing the gap between raw sensors and cognitive attention representation.

1.1.1 Spatial Productivity and Visual Attention Cues

The XR landscape is defined by staggering investment, with major players pouring billions into infrastructure to position spatial computing not just as an entertainment medium, but as a critical platform for knowledge work [16, 5]. To realize this vision, XR is evolving from immersive simulation into Cross-Virtuality Analytics (XVA), where physical monitors are replaced by dense arrangements of 2D virtual screens tailored for complex data analysis [33, 34]. In these information-rich environments, the ambition extends beyond merely replicating physical office mechanics to *augmenting* human collaboration [35, 36]. However, unlike traditional co-located work where gaze and gesture naturally align, XVA environments require explicit digital cues to support the high-fidelity coordination needed for complex knowledge work.

To fulfil the promise of a hyper-productive virtual workspace, systems must transcend sensor replication to reconstruct the higher-order cognitive state of “seeing.” This distinction is paramount because mutual awareness of attention serves as the primary coordination mechanism in collaborative work, enabling teams to establish common ground and anticipate actions [17, 19]. Yet, bridging the critical gap between raw gaze data—where a user looks—and their actual communicative intent remains an unresolved challenge [36, 37].

Current XR collaboration relies on two primary cues to support this awareness, but both face limitations within these dense XVA settings. Gaze cursors offer precision for referencing specific targets but suffer from high-frequency sensory noise (jitter) and calibration drift, often oversignalling micro-saccades rather than stable intent [36, 38]. Alternatively, Field-of-View (FoV) visualisations approximate attention from head orientation, providing a vital fallback for low-cost, eye-tracker-less headsets [39, 40]. However, standard FoV frustums are often too coarse to disambiguate specific data elements on 2D screens and lack the bi-directional feedback required for users to monitor their own communicative intent [41, 42]. Consequently, current cues tend to be either overly erratic (gaze) or overly vague (head).

1.1.2 The Dual Challenge

Despite the promise of spatial productivity, immersive collaboration in XR is constrained by a dual challenge:

The Hardware Challenge (The Absent Signal): A significant barrier remains for the widespread realisation of this vision: the hardware limitations of mass-market devices. Low-cost VR devices face significant limitations, primarily due to the absence of advanced user-facing sensors such as eye-tracking, which is costly [43, 44]. Devices such as the Oculus Quest dominate the market due to their affordability [16]. This hardware gap creates a form of “interaction blindness”; without sensor data to drive gaze, the system strips away the subtle, continuous visual cues that human collaborators rely on to establish mutual awareness. Mutual awareness of visual focus plays a vital role in effective collaboration, as it helps team members to establish a shared understanding and thus coordinate their actions [17, 18, 45, 36, 46]. As a result, users are forced to compensate for this missing non-verbal signal through explicit verbal descriptions of their focus, increasing communicative effort and disrupting the low-friction coordination characteristic of face-to-face collaboration [37].

The Conceptual Challenge (Looking $\neq$ Seeing): Yet, the challenge extends beyond the mere absence of sensors. Even when eye-tracking is available, raw gaze data is not always a perfect proxy for user intent [20, 21]. In immersive collaboration, effective communication relies on understanding a partner’s cognitive attention, which reflects their thoughts, rather than merely tracking their visual attention, which only indicates where their eyes are physically pointed [22, 23]. Visual attention is noisy; gaze can be drawn reflexively to motion or brightness [24, 25, 26]. Moreover, attention is dynamic in scale, fluctuating between a ‘global’ scope—encompassing the broader scene context—and a ‘narrow’ focus on specific details [27, 28]. A raw gaze point fails to capture this variability, often misrepresenting broader contextual awareness as precise focus. Furthermore, cognitive attention may decouple from visual gaze entirely; for instance, a user might look at one object while mentally referring to another element seen previously [29, 30, 31]. This decoupling is especially relevant during collaboration, where shared understanding depends on what is being mentally referenced [32].

1.1.3 The Principle of Complementary Reconstruction

Taken together, these constraints create a tension for which the solution is to transcend the availability of the raw sensor and reason directly on the reconstruction of the higher-order cognitive state of “seeing.”—inferring communicative intent from the noisy, multimodal signals that remain available.

In this thesis, I propose a shift in the way attention in XR is approached, which I term the Principle of Complementary Reconstruction. Rather than defining attention solely by where the eyes are pointed, this principle leverages the multi-modal nature of collaborative communication. When humans collaborate, they generate a redundant trail of clues across behavioural channels of varying specificity (gaze, head, speech) that are temporally coupled by cognitive intent and physically grounded in the semantic layout of the virtual world. The availability of a subset of these signals enables the system to reconstruct the others.

The Redundancy of Intent: Collaborative signals are rarely isolated. Head orientation provides spatial redundancy for gaze due to physiological synergy [47, 48]. When eye-tracking is absent, the head serves as a viable, albeit diluted, proxy for communicative intent [49, 41].

The Hierarchy of Specificity: Signals operate at varying resolutions. Head orientation establishes broad awareness [40], while speech enables precise selection [13]. Accuracy is achieved not by a single sensor, but by layering low-precision signals (head) with high-precision semantic filters (speech).

Cognitive Coupling: Visual and verbal behaviours are driven by a shared cognitive state [17]. During *perceptual decoupling*, where gaze diverges from thought [30], speech remains a robust anchor for maintaining the validity of the attentional cue [50].

Contextual Grounding: Attention targets meaningful elements. Reconstructing intent requires intersecting behavioural vectors with a *semantically segmented context* to resolve ambiguity. Just as gaze is drawn to informative features [51, 52], referring expressions rely on spatial context to convey meaning [11, 12].

1.2 Research Objectives

Building on the dual challenge and the Principle of Complementary Reconstruction, this thesis is organised around three research objectives, which Together form an end-to-end pipeline: deriving attention from available behavioural signals, externalising it as usable visual cues, and testing whether those cues actually support collaboration.

Objective 1 – Reconstruct Visual Attention (The Inference Problem) The first objective addresses the inference problem: developing methods to reconstruct visual attention in real time using only accessible signals (head tracking, microphones) on low-cost headsets. As discussed in Section 2.4.1, most consumer devices lack eye-tracking, while existing inference models are often too imprecise for dense 2D data or computationally heavy for mobile processors (Section 2.5.5, Section 2.5.6).

Objective 2 – Display Cognitive Attention (The Representation Problem) The second objective addresses the representation problem: transforming inferred attention into visual cues that reflect cognitive intent rather than raw sensor data. As detailed in Section 2.5.2, raw gaze is noisy and prone to decoupling, while traditional FoV visualisations often lack the specificity required for 2D-in-VR screens (Section 2.5.4, Section 2.5.1). This objective therefore seeks to design adaptive representations that filter noise, dynamically scale to match the locus of attention, and remain informative even when visual and cognitive focus diverge.

Objective 3 – Evaluate Collaborative Impact (The Validation) The third objective addresses the validation problem by evaluating the impact of these methods on collaboration, user experience, and task efficiency through empirical studies. Building on prior work identifying mutual awareness as critical (Section 2.2) and gaps in FoV evaluation (Section 2.5.1, Section 2.5.3), this objective uses systematic user studies to test the central claim that multi-modal, hardware-independent cues can effectively compensate for the lack of eye tracking on low-cost devices.

1.3 Contributions

This thesis advances the field of collaborative VR in two ways. First, it develops novel methods to infer and display visual attention, enhancing mutual awareness and communication among users. Second, it investigates the impact of FoV-style visualisations on collaboration dynamics in productivity scenarios. Additionally, it contributes valuable resources for future research in immersive environments. The key contributions are:

1.3.1 Investigating FoV-style Visualisations effects on Collaboration Dynamics in Productivity Scenarios

1.3.1.1 FoV-style Visualisation effects on Collaboration Dynamics

To investigate the efficacy of head-based signals as a proxy for attention, I conducted exploratory studies in the context of collaborative productivity tasks focused on the analysis of 2D data visualisations in a virtual office environment. The results revealed a reduction in self-reported cognitive load when FoV-style visualisations were present. This reduction was attributed to participants' improved ability to align broad visual attention without relying heavily on verbal communication or interpreting collaborator avatars. Furthermore, FoV-style visualisations facilitated smoother transitions between independent and collaborative phases and reduced ambiguity in relation to collaborators' focus and direction. These findings provide insights into how showing users FoV-style visualisations in a productivity environment affects collaboration dynamics. This contribution addresses the gap regarding bi-directional FoV-style visualisations in collaborative VR (Section 2.5.1).

1.3.1.2 A comparison of Mono-directional and Bi-directional FoV-style Visualisation effects on Collaboration Dynamics

Traditional FoV-style visualisations are typically displayed only to collaborators and not to the person generating the cue, as they are already aware of the direction of their own head.

This thesis investigates the impact of self-awareness of FoV-style visualisations by comparing mono-directional and bi-directional FoV-style visualisations. The results showed that the bi-directional FoV improved mutual awareness of visual attention and visual attention coupling, reduced reliance on collaborator avatars for understanding visual focus, and led to better pacing during collaboration. Participants were able to switch from independent to collaborative work more efficiently, and the bi-directional FoV condition resulted in reduced cognitive load as measured by the NASA TLX. These findings provide insights into how showing users their own visual focus affects collaboration dynamics, highlighting how self-awareness acts as a feedback mechanism for attention containment. This contribution addresses the gap regarding bi-directional FoV-style visualisations in collaborative VR (Section 2.5.3).

1.3.2 Developing Novel Methods to Infer and Visualise Visual Attention

1.3.2.1 Cone of Vision as a Novel Method to Infer and Display Visual Attention

This thesis introduces the "Cone of Vision" (CoV), a novel FoV-style visualisation technique specifically designed for 2D content displayed within VR environments (Chapter 3). Serving as the foundation for broad visual awareness, the CoV 2D visualisation provides a two-dimensional representation tailored for 2D screens embedded in VR environments, unlike traditional 3D FoV frustums. The CoV is visualised only at the intersection with the 2D surface. This approach is better suited for 2D panels or VR screens as compared to previous FoV techniques that display the field of view volume or wireframe, effectively causing occlusion with the display data. The geometry of the CoV is derived and validated from datasets of eye tracking data and represents a narrower area of focus (i.e., the probabilistic area where visual attention is more likely to be) compared to the traditional field of view, providing a probabilistic visualisation of visual attention that filters out high-frequency noise. This approach is particularly suitable for 2D panels or VR screens, offering more specificity than previous FoV techniques that displayed a broader, less focused area and avoiding occlusions from the visualized volume or wireframe.

This contribution directly addresses the lack of tailored FoV-style visualisations for VR screens displaying 2D content in symmetric collaborative settings (Section 2.5.1).

1.3.2.2 A Real-Time Deep Learning Model for Probabilistic Refinement

The thesis introduces a computationally efficient deep-learning model that predicts gaze probability regions based on head movement data (Chapter 4). Instead of pinpointing an exact gaze location, the model generates a probability density function (PDF) identifying the most likely gaze regions, which are visualised as percentile contours. Predicting gaze probability regions refines the coarse head signal provided by the standard CoV, addressing the inherent inaccuracies of using head orientation to infer gaze, especially in environments with high-density information. When trying to infer gaze with head movements alone, small prediction errors can result in pointing at entirely different objects, leading to confusion during collaboration. By predicting a region rather than a specific point, the model manages the trade-off between spatial ambiguity and precision, providing a more reliable indication of visual focus. This approach enhances the usability of XR environments, particularly when fine-grained eye movement tracking is not available in low-cost VR headsets. By utilising head orientation and rotational velocity as primary inputs this model operates in real-time, even on low-power devices, making it particularly suitable for untethered and low-cost VR headsets. This contribution addresses the gaps in real-time gaze inference for untethered devices and improves precision in gaze prediction models (Section 2.5.5 and Section 2.5.6).

1.3.2.3 Speech-Augmented CoV for Dynamic Adaptation and Semantic Precision

The thesis explored the integration of verbal communication with the CoV, dynamically refining its focus based on spoken cues to improve the specificity and accuracy of visual attention representation (Chapter 5). This method enhances the effectiveness of the CoV by incorporating speech cues, allowing the visualisation to scale dynamically from a broad scene-level view to an object-level focus. This contribution directly addresses the lack of dynamic representations

for cognitive attention (Section 2.5.2) by mimicking the "zoom lens" of attention, while also resolving the gap in specificity of FoV-style visualisations (Section 2.5.4).

Integration of Verbal Communication to Enhance Gaze Prediction: Explored the integration of verbal communication to further enhance gaze prediction accuracy and reconstruct cognitive attention. By interpreting referring expressions from collaborative conversations, this approach leverages the natural link between verbal communication and visual attention. For example, during collaborative tasks, when a user says "check the user engagement metrics," the system correlates this verbal input with elements within the user's field of view, enabling a more informed prediction of the user's gaze. This method is particularly beneficial in scenarios involving data analysis. In collaborative data visualisation tasks, such as exploratory data analysis in VR, participants often use verbal references to describe specific data points or trends, which helps direct collaborators' attention to relevant aspects of the visualisation. This approach offers a more accurate approximation of the user's gaze as compared to traditional head gaze methods, particularly in multi-user XR environments. This contribution helps to enhance the specificity of gaze inference by incorporating additional verbal cues, thereby addressing the issues related to gaze specificity and the inaccuracy of current models (Section 2.5.5). Furthermore, particularly in cases where verbal communication does not match visual attention, it is utilised as a semantic proxy for cognitive attention; by grounding the visual cue in spoken referents, the system allows the representation to scale dynamically and maintain validity during gaze aversion, thereby directly addressing the limitations of static scale and decoupling identified in Section 2.5.2.

1.3.3 Contributing Resources for Future Research

Released the first collaborative dataset that includes head movements, eye gaze, transcribed verbal communication, and semantically segmented context during VR collaboration¹. This dataset provides valuable insights into the interplay between visual and verbal cues during joint attention and serves as a foundation for training multimodal inference models. It is an

¹<https://github.com/Collaborative-Immersive-Visual-Toolkit/Speech-Gaze-Head-Datataset>

important resource for researchers studying collaborative behaviour in immersive environments. This contribution addresses the gap in publicly available datasets that combine head, eye, and verbal behaviour data for collaborative VR settings, facilitating future research in this area.

1.4 Thesis Structure

1.4.1 Chapter 2 — Literature Review

Chapter 2 presents a comprehensive Literature Review on the coordination of visual attention, verbal communication, and pointing-based gestures in collaborative environments. It examines how these elements function independently and interdependently in both co-located and remote collaboration. It reviews the evolution of XVA environments and the complementary roles of visual, verbal, and non-verbal cues in enhancing communication. The review also identifies several gaps in the existing literature, particularly in leveraging these modalities to enhance collaboration in XR environments, which this thesis aims to address.

1.4.2 Chapter 3 — Cone of Vision

In Chapter 3, I introduce the "Cone of Vision" (CoV), a novel Field-of-View (FoV) visualisation technique for 2D screens in VR. Unlike previous methods which focused primarily on 3D environments and were mono-directional, the CoV provides a bi-directional visualisation of visual attention, enhancing mutual awareness (Section 2.5.3). By estimating attention based on head movements, the CoV supports effective collaboration in scenarios where eye-tracking technology is unavailable, particularly in 2D-in-VR applications (Section 2.5.1). Empirical studies demonstrate that the CoV improves collaboration by supporting mutual awareness (Section 3.3), with bi-directional FoV-style visualisations fostering better communication compared to mono-directional ones (Section 3.3.2). This chapter showcases an accessible solution for enhancing collaboration in immersive VR without the need for specialised hardware.

1.4.3 Chapter 4 — Real-Time Deep Learning Model for Gaze Probability Regions

In Chapter 4, introduces a real-time deep learning model that uses head movements to infer gaze probability regions in VR. Current methods for gaze inference often lack precision (Section 2.5.5) and are resource-intensive for low-cost head-mounted displays (Section 2.5.6). Our proposed model addresses these issues by analysing head movement data to generate a probability density function of gaze, which enhances the accuracy of visual attention inference. The model employs an autoencoder for probability density generation (Section 4.1.4.1) and a contour encoder for real-time inference (Section 4.1.4.4). Results show improved gaze prediction accuracy compared to head-based baselines, offering a nuanced understanding of visual attention in improving collaboration in VR, especially in settings without eye-tracking technology.

1.4.4 Chapter 5 — Speech-Augmented Cone of Vision

In Chapter 5, I introduce the "Speech-Augmented Cone of Vision" (CoV+Speech), which combines head orientation with speech to refine the representation of visual attention in VR. Existing visualisations, such as FoV and gaze-based methods, have limitations, including ambiguity and distraction due to natural eye behaviour (Section 2.5.4). CoV+Speech addresses these gaps by dynamically adjusting the CoV to highlight spoken referents, thereby enhancing mutual awareness and improving collaboration. Empirical studies show that CoV+Speech offers more precise visual attention representation and influences gaze behaviour by encouraging users to focus within the visualized region (Section 2.2.3). This approach leverages natural verbal behaviours as an input for gaze inference, where referring expressions enhance visual attention inference in immersive environments.

1.4.5 Chapter 6 — Conclusion

In Chapter 6, this thesis concludes by summarising its contributions in enhancing collaboration and gaze inference in XR environments through novel gaze inference and visualisation methods.

It reflects on the incremental development of techniques introduced across the preceding chapters, emphasising their role in improving mutual awareness and gaze inference in immersive settings. The chapter also explores the broader impact and implications of these findings on XR applications beyond virtual offices, including education, healthcare, and creative industries. Finally, it outlines limitations and future directions, identifying key areas for refinement, such as integrating multi-modal attention cues, scaling CoV visualisations for larger groups, and extending its applicability to complex 3D environments.

1.5 Contributing Papers

The primary contributions of this thesis are grounded in the following peer-reviewed papers, which have been accepted and published in reputable, high-impact venues.

Cone of vision as a behavioural cue for VR collaboration

Riccardo Bovo, Daniele Giunchi, Muna Alebri, Anthony Steed, Enrico Costanza, and Thomas Heinis. Cone of vision as a behavioural cue for vr collaboration. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 2022. doi: 10.1145/3555615

Real-time head-based deep-learning model for gaze probability regions in collaborative VR

Riccardo Bovo, Daniele Giunchi, Ludwig Sidenmark, Hans Gellersen, Enrico Costanza, and Thomas Heinis. Real-time head-based deep-learning model for gaze probability regions in collaborative vr. In *Proceedings of the Eye Tracking Research and Applications Symposium (ETRA)*, 2022. doi: 10.1145/3517031.3529642

Speech-augmented Cone-of-Vision for exploratory data analysis

Riccardo Bovo, Daniele Giunchi, Ludwig Sidenmark, Joshua Newn, Hans Gellersen, Enrico Costanza, and Thomas Heinis. Speech-augmented cone-of-vision for exploratory data analysis. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2023. doi:

10.1145/3544548.3581283

1.5.1 Other papers authored and co-authored during this PhD

In addition to the papers contributing to this thesis highlighted earlier, during my PhD, I was fortunate to author and co-author several other papers that explored related aspects of VR technology.

Papers that contributed to the field of VR collaboration dynamics:

Shall I Describe it or Shall I Move Closer? Verbal references and locomotion in VR collaborative search tasks

Riccardo Bovo, Daniele Giunchi, Enrico Costanza, Anthony Steed, and Thomas Heinis. Shall i describe it or shall i move closer? verbal references and locomotion in vr collaborative search tasks. In *Proceedings of 20th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET), 2022

Papers that contributed to the field of VR interfaces for knowledge workers, specifically focusing on integrating traditional 2D desktop applications into immersive environments:

EmBARDiment: an Embodied AI Agent for Productivity in XR

Riccardo Bovo, Steven Abreu, Karan Ahuja, Eric J. Gonzalez, Li-Te Cheng, and Mar Gonzalez-Franco. Embardiment: An embodied ai agent for productivity in xr. *arXiv preprint arXiv:2408.08158*, 2024. doi: 10.48550/arXiv.2408.08158

Streamspace: A Framework for Window Streaming in Collaborative Mixed Reality Environments

Daniele Giunchi, Riccardo Bovo, Nels Numan, and Anthony Steed. Streamspace: A framework for window streaming in collaborative mixed reality environments. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (IEEEVR)*, 2024

WindowMirror: an opensource toolkit to bring interactive multi-window views into XR

Riccardo Bovo, Eric J. Gonzalez, Li-Te Cheng, and Mar Gonzalez-Franco. Windowmirror: An open-source toolkit to bring interactive multi-window views into xr. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (IEEEVR)*, 2024

EmBARDiment: an Embodied AI Agent for Productivity in XR

Riccardo Bovo, Steven Abreu, Karan Ahuja, Eric J. Gonzalez, Li-Te Cheng, and Mar Gonzalez-Franco. Embardiment: An embodied ai agent for productivity in xr. *arXiv preprint arXiv:2408.08158*, 2024. doi: 10.48550/arXiv.2408.08158

Browsing Internet Content in Multiple Dimensions.

Tiger Cross, Riccardo Bovo, and Thomas Heinis. Browsing internet content in multiple dimensions. In *Proceedings of the EDBT/ICDT Workshops*, 2021

Papers that contributed to the field of pedestrian simulation in VR:

Perceived realism of pedestrian crowd trajectories in VR

Daniele Giunchi, Riccardo Bovo, Panayiotis Charalambous, Fotis Liarokapis, Alastair Shipman, Stuart James, Anthony Steed, and Thomas Heinis. Perceived realism of pedestrian crowds trajectories in vr. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology (VRST)*, 2021

VR Toolkit for Identifying Group Characteristics

Hugo Mayo, Alastair Shipman, Daniele Giunchi, Riccardo Bovo, Anthony Steed, and Thomas Heinis. Vr toolkit for identifying group characteristics. *Collective Dynamics*, 2021

Chapter 2

Literature Review

This chapter provides a comprehensive review of the literature on the coordination of visual attention, verbal communication, and pointing-based gestures in collaborative environments. It examines how these elements function both independently and interdependently in co-located and remote collaboration, highlighting their complementary roles in enhancing communication and coordination. The review begins by discussing the evolution of XVA environments (Section 2.1), particularly focusing on their adaptation for 2D content as a replacement for physical screens. These environments have become crucial in supporting collaboration and data visualisation within multi-purpose applications, laying a robust foundation for their use in modern hybrid work settings. Understanding XVA environments is essential as they represent the application context for the contributions proposed in this thesis.

Following this, the chapter explores the complementary nature of visual and verbal communication, along with non-verbal cues. Section 2.2 examines visual attention in collaboration, highlighting joint attention as a key factor in shared understanding. The discussion then shifts to the interaction between verbal communication and visual attention (Section 2.2.6), examining how each influences the other to enhance collaborative effectiveness. Finally, the role of non-verbal communication, particularly pointing gestures (Section 2.3), is analysed to illustrate how gestures during collaboration can impact both visual attention and verbal exchanges. These sections will cover both the theoretical foundations and practical applications of these

interactions as explored in the literature.

Crucially, this chapter identifies several gaps in the existing literature that this thesis aims to address (Section 2.5). Through this exploration, the chapter lays the groundwork for understanding how visual attention, verbal communication, and non-verbal cues can be leveraged as implicit inputs to enhance human-human collaboration within XR environments.

2.1 Evolution of XVA Environments

The rapid advancement of VR and XR technologies has significantly reshaped the approach to collaborative work and data analysis. These immersive technologies have paved the way for XVA, which integrates data analytics across the reality-virtuality continuum, allowing users to interact with data and collaborate effectively in both physical and virtual environments. This section will delve into the specific emergence of XVA tailored for 2D content, examining how these technologies have evolved to support collaboration and data visualisation within standard multi-purpose applications, providing a robust foundation for their application in modern hybrid work settings.

2.1.1 The Emergence of XVA Tailored for 2D Content

XVA can support users simultaneously via collaborative interfaces encompassing the reality-virtuality continuum Fröhler et al. [33] and has been adopted by a large set of works [41, 63, 64, 65, 66, 67, 68, 34]. XVA has recently gained interest from researchers due to the COVID pandemic, driven by the increase in demand for remote working and collaboration. In this context, VR facilitates portable, personalised working environments that can be used at home, on the go, or in hybrid setups with standard screens [69, 70, 71]. While some of the XVAs are tailored for specific 3D data applications [63, 67, 65], there is a growing trend to support 2D content, which is cross-compatible with the standard multi-purpose applications available on desktop PCs [72, 66, 34, 73, 74, 75]. This approach is also followed by commercial applica-

tions [76, 77, 78, 79, 80, 81], which either enable the display of standard 2D documents (e.g., web pages, calendar layout, office documents) or capture arbitrary 2D windows from a desktop PC and display them as a 2D surface in the virtual environment. Such software allows users to set up their own layout of 2D windows in the 3D environment (Figure 5.3 a) and b)), which is a problem that previous XVA studies have explored (Figure 5.3 c) and d)).

2.1.2 2D Screen Layouts for XVA

Environment configuration and the positioning and orientation of 2D panels can be arranged in different ways in an immersive virtual scenario. Such setups may influence participants' perceptions, impacting both the sense-making process and collaboration.

Previous works related to HMD headsets for supporting everyday computing tasks with desks and keyboards have opted for retaining a cylindrical display layout similar to that found on physical working desktops, shaped according to the ergonomic mechanics of the eyes and head [74, 75]. This semi-cylindrical layout approach is also used by applications such as Oculus Infinite Office [82] and has been explored in AR studies such as in Pavanatto et al. [75], McGill et al. [74]. The work of Google Daydream, when introducing the virtual screen concept, exploited the idea of screens within a cylindrical contour, providing a less claustrophobic feeling and an improved parallax effect [83].

On the other hand, Hybrid Reality environments for immersive data analysis, such as those used by Cavallo et al. [84] and Febretti et al. [85], utilise a more cylindrical layout in which screens might completely surround users. In these hybrid environments, the size of the physical space may limit the virtual environment. For example, the environment used by Lee et al. [66] is limited by the physical space available (which is estimated to be a 3m x 3m square room), while the Hybrid Reality Environment for Collaborative Information Analysis used by Cavallo et al. [84] spans a circle with a 5m radius, and the environment used by Febretti et al. [85] spans a 6m radius. Despite the variance in the size of collaborative environments, they all adhere to a similar approach where 2D screens define the walls/boundaries of the environments.

This approach is consistent with the findings of Satriadi et al. [34] and Lee et al. [66]. Satriadi et al. [34] highlight how users arrange the layout of 2D surfaces in 3D space, with findings showing a preference for semi-spherical and spherical layouts. Further studies by Lee et al. [66] highlight considerations related to the placement of 2D VR screens in 3D space when collaborating with other users. They suggest that, when 2D visualisations are placed for collaboration purposes, participants neatly arrange them on the walls of the virtual environments, making the 2D screen content easily visible to others. Moreover, they show that participants can effectively present their visualisations to others when this wall layout is used.

Both cylindrical and semi-cylindrical layouts adhere to these requirements and, in this thesis, both configurations are explored. In the first experiment (Chapter 3), the virtual environment was designed to match the same size and circular shape (3m radius) as the Hybrid Reality Environment described by Cavallo et al. [84] and Febretti et al. [85], as shown in Figure 3.3. Without being limited by physical space, this design choice ensured consistency with prior work. Although Satriadi et al. [34] describe the layout as spherical, due to the increase in size caused by the collaborative context, the spherical curvature can be approximated to a cylinder.

In the second study (Chapter 5), a semi-cylindrical layout is used, one which is more in line with users constrained in a sitting position, fixed by a desk or a keyboard, as highlighted in studies by McGill et al. [74] and Pavanatto et al. [75].

2.2 Mutual Awareness of Visual Attention in Collaborative Work

Visual attention plays a vital role in collaborative work by enabling team members to develop a shared understanding and coordinate their actions. The ability to perceive where others are focusing their attention is crucial for effective teamwork, often leading to synchronised gaze behaviour where collaborators observe the same elements within a shared environment.

This section explores the importance of visual attention in collaboration, focusing on mutual

gaze awareness, synchronised gaze behaviour, and their impact on collaboration quality (Section 2.2.1). It then examines XR visual attention cues, such as gaze cursors and field of view (FoV) visualisations (Section 2.2.2). Next, the section discusses the impact of self-awareness of visual attention in XR and how it influences coordination and task performance (Section 2.2.3). Finally, it examines how joint attention, measured via eye or head gaze, is considered a key indicator of collaboration quality when assessed through either gaze tracking or head direction (Section 2.2.4).

2.2.1 Mutual awareness of visual attention during collaboration

When people collaborate in immersive spaces (other in-person or virtual environments), their gaze behaviours often become synchronised or coordinated. This means that they tend to look at the same things or follow similar visual patterns, which helps them to understand each other better and work together more effectively. This is explained by the fact that environments or tasks people are engaged in during collaboration provide shared visual information which is explored synchronously. For example, if two people are working on a computer together, the screen becomes a shared visual context that influences their gaze behaviour.

Research has shown that synergies in gaze behaviour correlate with performances by fostering mutual understanding, reducing misunderstandings, and enhancing collaborative efficiency. For instance, Vrzakova et al. utilised recurrence quantification analysis (RQA) to examine patterns of visual attention in teams during collaborative problem-solving tasks. Their findings revealed that synergies in gaze behaviour aligned with screen activity, and the number of synergies in gaze behaviour directly correlated with team performance and the overall collaboration process [17]. Similarly, in a pair-programming task, Villamor and Rodrigo demonstrated the critical role of concurrent visual attention in achieving successful collaborative coding outcomes [18]. Expanding on this, Moulder et al. investigated visual attention synergies during collaborative problem-solving, using RQA to quantify team-level gaze dynamics, which they found to be predictive of task success [19].

2.2.2 Visual Attention Cues in Collaborative VR

The ability of collaborators to be aware of where others are looking or focusing their attention is crucial for effective collaboration because it helps individuals anticipate each other's actions and intentions. In AR and VR settings, tools such as gaze cursors (indicators showing where someone is looking) are used to help collaborators understand each other's focus.

Several studies have shown that visualising this attention can enhance teamwork by allowing individuals to anticipate the intentions of others and increase mutual awareness [86, 87, 88, 89, 12]. In AR and VR settings, visualisations such as gaze cursors are commonly used and have been extensively researched and have been proven to enhance collaboration [36, 90, 46, 91, 92]. However, modern VR equipment often lacks built-in eye-tracking capabilities. To address this, several studies propose using field of view (FoV) visualisations as an approximation of gaze [93, 39, 94, 95, 96, 41, 40]. These methods have demonstrated that FoV-style visualisations can help collaborators establish mutual awareness and maintain shared attention [93, 41, 39, 94, 41, 36, 97]. Furthermore, Piumsomboon et al. suggested that these visualisations could be dynamically displayed or hidden, depending on the context of collaboration, thereby minimising visual clutter while maximising collaborative efficiency.

2.2.3 Self-Awareness of Visual Attention cues

In collaborative virtual environments, self-awareness plays a critical role in how individuals coordinate their actions and interactions with others. Understanding one's own visual attention cues—such as where one's gaze is directed—can significantly influence the effectiveness of teamwork and the smoothness of coordination. This subsection explores the concept of self-awareness in VR settings, particularly focusing on how being conscious of one's gaze and other visual cues can enhance collaborative efforts.

Gaze behavioural cues have been explored in physical manipulation tasks in an MR setup by Jing et al. [98]. This study is particularly notable for its bi-directional approach to gaze visualisation, which is absent in studies implementing the FoV frustum visualisation (e.g., [41]).

Their findings highlight how their visualisation increases gaze awareness and joint focus (i.e., visual coordination). In addition, they extend previous research by exploring the self-awareness of visual cues. Such exploration leads to interesting outcomes, as they report that self-awareness in gaze visualisations increased confidence and the certainty that their gaze location is being accurately delivered.

2.2.4 Joint Attention as a Measure of Collaboration Quality

Joint attention, characterised by synchronised eye gaze and head direction, is a critical indicator of collaboration quality [17, 18, 19]. The ideal measure of joint attention is through eye gaze tracking, as it directly reveals where an individual’s focus lies and how it aligns with others. However, in the absence of eye-tracking technology, head direction can serve as a viable proxy for attention [49, 99, 100]. While head direction may not pinpoint the exact focus of attention as precisely as eye gaze, it generally correlates with a person’s gaze direction and can be effectively used to measure joint attention. This approach allows for the assessment of collaborative quality, even in those environments where advanced eye-tracking equipment is not available.

2.2.5 The Disconnect Between Visual and Cognitive Attention

While eye-tracking is often regarded as the gold standard for attention inference in XR, a significant body of research suggests that raw gaze data is an imperfect proxy for user intent. In collaborative environments, effective coordination relies on joint attention a shared mental focus on an object or task rather than merely shared gaze the physical intersection of visual axes [101, 102]. Understanding the limitations of raw gaze is crucial for designing robust collaborative cues. These limitations can be categorised into three theoretical disconnects: sensory noise, scale mismatch, and perceptual decoupling.

Sensory Noise: Bottom-Up vs. Top-Down Attention: Visual attention is governed by a competition between ‘top-down’ goals (internal intent) and ‘bottom-up’ saliency (external stimuli) [25]. Computational models of visual attention, such as the saliency map proposed by Itti

and Koch, demonstrate that gaze is reflexively drawn to low-level features—such as high contrast, brightness, or motion—often before semantic processing occurs [24]. This phenomenon, known as attentional capture, means that a user’s gaze may momentarily fixate on a distraction (e.g., a flashing notification or moving avatar) even if their cognitive attention remains focused on a static data point [26]. Consequently, raw gaze data contains high-frequency ‘noise’.

Scale Mismatch: The Zoom Lens of Attention: A fundamental limitation of standard eye-tracking cursors is their representation of attention as a single coordinate point (x, y) . However, cognitive psychology posits that attention is dynamic in scale. The ‘Zoom Lens’ model suggests that the focus of attention can expand to encompass a global scene or contract to scrutinise specific details [27]. Similarly, the ‘Spotlight’ model argues that attention operates over a spatial region rather than a dimensionless point [28]. By representing gaze as a precise point, current XR systems often misrepresent instances of ‘broad’ or ‘global’ attention (e.g., scanning a dashboard layout) as specific, narrow focus. This variability suggests that region-based visualisations, such as Field-of-View cones or probabilistic density maps, may offer a more cognitively aligned representation of attention during tasks requiring broad situational awareness.

Perceptual Decoupling and Gaze Aversion: Finally, visual gaze can decouple from cognitive attention entirely. This phenomenon, often termed ‘perceptual decoupling’, occurs during states of internal cognitive processing, such as mind-wandering or complex problem-solving, where the brain attenuates external sensory input [30, 29]. Research distinguishes between *externally directed cognition* (EDC), where gaze aligns with the task, and *internally directed cognition* (IDC), where users may stare blankly or actively avert their gaze to reduce cognitive load [31]. For instance, Glenberg et al. demonstrated that individuals often look away from a target (gaze aversion) when answering difficult questions to prevent visual interference [103]. In collaborative XR, this implies that a user looking at a specific graph may actually be mentally processing a previous data point or planning a future sentence. This decoupling highlights the necessity of multimodal signals—such as verbal references—to disambiguate whether a user is ‘looking’ or truly ‘seeing’ [32].

2.2.6 Verbal Communication in Collaborative Work:

Building on the idea of synchronised gaze behaviour in collaborative environments, this section explores how verbal communication and visual attention interact and influence each other. In immersive settings, whether in person or in virtual environments, the way individuals communicate verbally can significantly affect how they direct and share their visual focus. Verbal cues, such as spoken references to visual elements, work together with the awareness of others' visual focus to enhance understanding and coordination within teams. The interplay between verbal communication and visual attention is crucial in ensuring that team members are aligned and engaged with the same elements of their shared environment.

This section delves into these dynamics, highlighting how verbal communication and visual attention interact and influence each other (2.2.7). The discussion also extends to the evolution and application of speech-to-text technologies as inputs for XR interactions, emphasising their impact on user interaction (2.2.8).

2.2.7 Verbal Communication and Visual Attention During Collaboration

Verbal communication is essential in aligning and coordinating visual attention in collaborative scenarios, both in person and within remote immersive environments like XR. For instance, Pettersson et al. [13] identified strategies such as colour and size statements that collaborators use to coordinate visual attention during collaborative map analysis on tabletop displays. Similarly, Zagermann et al. [104] demonstrated that the size of tabletop displays influences the interplay between verbal and visual communication, affecting how collaborators interact. Furthermore, Stark et al. [105] emphasised the critical role of verbal communication in coordinating decision-making within complex systems, highlighting its importance in collaborative environments.

By effectively referencing visual elements, verbal communication helps maintain mutual awareness and improves overall coordination through the use of referring expressions (REs). These

expressions, which can be explicit or implicit, enable effective identification and understanding of specific entities or objects within a given context. For instance, implicit REs, such as pronouns, rely heavily on the surrounding context and visual cues to convey meaning without explicitly naming the entities.

Visual cues, such as gaze, significantly enhance collaboration by clarifying these REs. Ishii and Kobayashi [11] demonstrated that integrating visual attention cues with verbal communication resolves ambiguities in spatial references, leading to more effective teamwork. Additionally, Monk et al. [106] found that combining visual cues with verbal instructions reduces cognitive load on collaborators, resulting in a smoother and more intuitive task flow.

Moreover, D'Angelo and Begel [12] explored how eye-gaze visualisations influence verbal communication, particularly in remote pair programming. Her research measured the implicitness and explicitness of spatial REs, finding that visual cues facilitate more implicit communication, allowing collaborators to engage in simpler, more streamlined exchanges.

These studies collectively underscore the importance of verbal communication in enhancing visual coordination and the impact that visual coordination has on communication. By explicitly referencing visualised elements, collaborators can achieve precise coordination and build a shared understanding. This is crucial for effective teamwork in both physical and virtual environments. Conversely, relying on implicit references supported by visual attention awareness allows for more efficient communication, thereby optimising the collaborative process.

2.2.8 Evolution of Speech-to-Text Technology and Its XR Implementations

The real-time conversion of speech to text has evolved significantly over the years. Early speech-to-text systems faced limitations in accuracy and reliability, often requiring extensive training and speaker-dependent models. Only in recent years have highly accurate models emerged, enabling near-perfect real-time transcription. This advancement has had a profound impact on various domains, including XR. Speech interfaces have been widely adopted in a

wide variety of interactive contexts [107, 108, 109, 110, 111, 112, 113, 114, 115, 116]. Speech interfaces present various challenges, including speech recognition, phrase interpretation, and interaction. Speech interaction has been used in numerous works, and how the user interacts is highly dependent on the task and the functionalities of the various released systems [117, 118, 119, 120, 121]. Speech interaction and XR met decades ago [122, 123, 124, 125] with the implementation of multi-modal systems with two possible approaches: fully interactive speech or "command and control". The first type was speaker-dependent because the variety of words and sentences forced the user into a training phase called enrolment [126]. Before the recent revolution of natural language processing (NLP), researchers adopted the paradigm of the 'Wizard of Oz' [127, 128, 129] to avoid the technical limitations of free-speech interfaces for both the recognition and process phases. However, the command and control system was speaker-independent, as the limited number of words to be converted into commands did not require an enrolment stage. In particular, this approach presents advantages over keyboard input or gesture input [130], as the last one necessitates practice. Further, the interaction style derived from these two approaches originates from studies illustrating that vocabulary size can impact interaction [131, 132] as well as the awareness with which a machine or human interprets speech [133]. Given these advantages, speech interaction was experimented with in medicine to treat social phobias [134], in civil engineering to help with architectural design [135], and in dealing with the digital twins of complex machines such as airplanes [136]. With the advent of more powerful deep-learning models for recognition and NLP, speaker-dependent systems became obsolete, as various services [137, 138, 139, 140, 141] can receive and process audio streaming that almost provides transcription in real-time.

2.3 Pointing-Based Communication in Collaborative Work

Non-verbal communication, particularly through gestures, complements verbal communication by adding layers of meaning and context that significantly influence collaborative interactions. In both face-to-face and virtual environments, non-verbal cues, such as pointing gestures and body language, play a crucial role in directing visual attention and clarifying implicit references

to visual elements (Section 2.3.1). A nuanced examination of proximal and distal pointing gestures is presented, particularly within XR environments, where tools like virtual lasers and avatar hands refine these dynamics (Section 2.3.2). Additionally, the impact of multi-modal interaction—combining gestures with speech and gaze—is explored within its role in enhancing communication and collaboration across both traditional and XR settings. Through a detailed review of existing research, this section provides a comprehensive understanding of how pointing-based communication can be leveraged to facilitate more effective collaborations.

2.3.1 Pointing Gestures During Collaboration

Non-verbal communication, such as pointing gestures in VR environments, also plays a critical role in negotiating attention dynamics during collaboration. The use of gestures or gesture aids, including laser pointers, helps to synchronise visual attention as well as clarify REs, thereby enhancing overall collaboration efficiency. The integration of avatar hands enables pointing gestures, as well as the ability to perform accurate distance pointing with a laser in XR, further refines these dynamics, making collaboration more intuitive and effective. The absence of visual attention awareness, the intricacy of a visual scene, and the challenge of verbally describing a referent all serve as catalysts for using pointing gestures coupled with utterances, also known as deictic expressions [37]. Pointing gestures have been extensively researched in the context of immersive multi-modal conversations and are considered highly beneficial in facilitating referent disambiguation [37].

In AR/VR, pointing gestures become even more effective due to the use of virtual lasers, which enable users to specify the referent of a pointing gesture with greater accuracy [50]. Piumsomboon et al. explored MR collaboration and the benefits of providing virtual awareness cues, such as pointing gestures, field of view, and eye gaze, revealing that these cues significantly enhance user performance, usability, and subjective preferences [142]. Further, Bovo et al. demonstrated that, when users encounter a referent that is difficult to describe verbally, they prefer to point at it, even if it requires physical movement, rather than providing a verbal description [143]. The importance of pointing in collaborative VR environments is further underscored by several

studies that have focused on improving the accuracy of pointing gestures. Techniques such as warping or distorting deictic gestures have been proposed to enhance collaboration [144]. Additionally, Mayer et al. investigated the effects of offset correction and cursor feedback on mid-air pointing in both real and virtual environments, suggesting that the subtle redirection of a user's arm embodiment to align with their gaze can significantly improve pointing accuracy from an observer's perspective [145].

2.3.2 Pointing Gestures and Proxemics

Pointing gestures can be categorised as either proximal or distal, depending on the spatial relationship between the producer of the gesture and the referent [146]. In the case of proximal pointing, the gesture producer can directly touch the target, allowing observers to identify the target with a high degree of confidence [147]. This proximity minimises ambiguity, making proximal pointing highly accurate as it leaves little room for misinterpretation. Conversely, distal pointing requires the observer to extrapolate the intended direction based on the vector defined by the pointer's posture [147, 148]. In XR environments, the precision of distal pointing can be enhanced through the use of augmentative tools such as cursors. These tools improve mid-air pointing accuracy by providing visual feedback, thereby eliminating the need for observers to infer the direction of the pointing gesture [145].

The literature allows us to define accurate pointing (both from the points of view of the producer and observer) and inaccurate pointing. Pointing gestures can be either proximal or distal [146]. When indicating proximal referents, the producer of a pointing gesture can touch the target, and observers can identify targets with confidence [147]. Therefore, proximal pointing is considered accurate, as there is no room for misinterpretation. With distal pointing, the observer instead needs to extrapolate the vector direction as defined by the pointer's posture [147, 148]. However, previous studies have found that using a cursor improves mid-air pointing precision thanks to visual feedback and removes the need to infer the pointing direction [145]. Therefore, distal pointing with a cursor is considered more accurate, whereas distal pointing without a cursor is more prone to misinterpretation.

While previous works offer several methods to improve the accuracy of pointing via machine learning models, research has also explored how users adapt their behaviour when faced with inaccurate pointing within collaborative environments. Studies suggest that, in immersive virtual environments, users may compensate for imprecise pointing by adjusting their position relative to the referent, modifying their pointing technique, or enriching their verbal references to reduce ambiguity.

2.3.3 Leveraging Multi-Modal Interaction in CSCW

Multi-modal communication in collaboration leverages multiple forms of interaction to enhance understanding, alignment, and efficiency among team members, making collaborative efforts more effective and cohesive. As multi-modality in communication is a naturally learned behaviour for humans, previous HCI studies have explored the use of multi-modal interaction techniques to create more natural interactions. Prior work in HCI leverages natural non-verbal 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 aids the process of understanding the referent intended by the user. Bolt's seminal work "Put-That-There" [149] 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 [150].

Similarly Mayer et al. conducted a study that incorporated GPS location and user's head gaze into mobile voice assistants, enhancing their comprehension of nearby shops and buildings by understanding a user's location and visual direction [151]. Ahuja et al. et al. made use of speech as a direction communication channel, akin to employing user-gaze to facilitate interactions across smart environment ecosystems [152]. Romaniak et al. introduced *Nimble* [153], a mobile interface that integrates visual question-answering models with gesture recognition, specifically directing gestures, for more intuitive user interactions. Marwecki et al. exploited the implicit attention cues of gaze in *MiseUnseen* to facilitate novel interactions inside XR environments,

including creating or arranging objects covertly at runtime [154].

These works highlight the effectiveness of using visual attention or pointing gestures to aid in understanding the user’s intended referent during verbal requests. However, none of these prior works explored the collaborative patterns of visual attention or pointing behaviour to enrich immersive multi-modal conversations. Instead, they focused on interactions rooted in single-user non-verbal communication and single-event interactions.

2.4 Gaze Prediction in XR

Eye-tracking in VR headsets offers valuable insights into user attention but comes with significant challenges, including high costs and accuracy issues, which hinder the widespread adoption of eye-tracking in consumer VR devices. These challenges have prompted the need for alternative gaze prediction methods that do not rely on eye-tracking technology.

In XR environments where eye-tracking is not available or reliable, gaze prediction models aim to infer the user’s gaze location using alternative data, primarily head orientation and movement, which are readily accessible through most VR headsets. More computationally intensive models incorporate richer features, such as visual saliency—distinctive elements within a visual scene that naturally attract attention—to enhance prediction accuracy. This section will examine the cost and limitations associated with eye-tracking technology in VR devices (2.4.1) and explore how these constraints have led to the development of alternative gaze inference models. These models utilise head movement (2.4.2) and visual saliency (2.4.3) as proxies for eye-tracking, addressing the challenges of accuracy, affordability, and widespread adoption across XR systems.

2.4.1 Cost and Limitations of Eye-tracking on VR Devices

Depicting the direction of eye gaze would be the most obvious way to represent visual attention in collaborative settings [98]. Currently, including eye-tracking technology in HMDs is feasible,

although it impacts pricing considerably. For example, at the time of writing, the Pico Neo 3 with eye-tracking costs 50% more than the basic version (699\$ vs. 899\$), and the HTC Vive Pro has a 100% increment in price for the version with an eye-tracker (699\$ to 1,399\$). Owing to such cost differentials, HMD devices with eye tracking are not popular within the consumer market. For example, the volume of VR headset unit sales worldwide in the 4th quarters of 2019 and 2020 indicates that the vast majority of VR devices sold do not have eye-tracking [16]. While such a lack of popularity in the consumer market of HMDs with eye-tracking affects the potential pool size of participants for a remote VR study during the pandemic, it is nonetheless plausible that eye-tracking devices will become cheaper and more popular over time. However, whether all VR HMDs will eventually feature eye-tracking technologies remains questionable. Additional hardware costs are not the only reason for exploring alternative methods of eye-tracking.

Eye-tracking can suffer from low accuracy due to calibration errors and the drift of wearable eye-tracking sensors [155]. While eye-tracking was used in HCI studies with different purposes such as selection [156], object manipulation [157], and collaboration [35], previous work shows that participant data is often discarded due to calibration and tracking issues [158, 159, 160]. Participant behaviour may cause the eye tracker to slip on the participant's head, potentially strongly affecting data quality [38]. Moreover, the eye characteristics of ethnic groups may affect the accuracy of eye trackers. For example, a study by [161] shows that Asian participants' eye-trackers' accuracy was worse than that of African and Caucasian participants. The problems mentioned above affect the accuracy in such a way that, in a collaborative scenario with dense data visualisation, a small angular displacement may convey the wrong information between participants.

Indeed, the studies discussed above suggest that the HCI research community and VR developers should explore alternative methods for providing visual attention information in eye-tracker-less VR HMDs, as well as fallback mechanisms for scenarios where eye-tracking loses accuracy or fails.

2.4.2 Gaze Prediction Based on Head Velocity

All 6DOF XR headsets are equipped with head-tracking capabilities. These features are crucial for calculating the user’s frustum of vision and rendering images accordingly based on head orientation. Head direction is closely linked to eye movements due to low-level neurological connections [162, 163]. This relationship has been extensively studied [164, 100], including in the context of XR head-tracking capabilities [47, 48].

Physiological nature of Eye-Head Coordination: The coordination between eye and head movements is largely driven by low-level neural reflexes designed to stabilize vision, most notably the Vestibulo-Ocular Reflex (VOR) [162, 163]. When a user initiates a head movement to look at a target, the eyes typically move in a “ballistic” (pre-programmed) manner relative to the head’s velocity to maintain fixation or shift gaze [100, 164, 163]. Research indicates that eyes often anticipate head motion, shifting in the direction of the turn milliseconds before the head begins to rotate [165, 166, 99].

Owing to this underlying ballistic and anticipatory behaviour, head dynamics such as head velocity can be effectively utilised in visual attention and gaze prediction models. Recent work by Sidenmark and Gellersen [47] quantifies the reliability of head direction as a proxy for gaze direction. This coordination creates a predictable pattern where head velocity serves as an informative cue for likely gaze targets, a property that has been exploited in models for egocentric video and VR environments [48, 167, 168, 51]. However, it is important to note that head-based gaze predictions remain approximate, with typical angular errors of around 10° [51].

2.4.3 Gaze Prediction with Alternative Data

Visual saliency refers to the distinctiveness or prominence of certain elements within a visual scene that naturally attract a viewer’s attention. These elements often stand out due to factors such as colour, contrast, movement, or unique patterns. Incorporating visual saliency into gaze prediction models can significantly enhance their accuracy, as it helps predict where a user is likely to focus their attention in complex visual environments.

For example, Yamada et al. [169] demonstrated that incorporating visual saliency in gaze prediction models significantly improves the prediction of where a user will look in complex visual scenes. Similarly, Matsuo et al. [170] showed that integrating saliency features with head movement data can lead to more accurate gaze prediction in video content. Li et al. [171] also explored the role of visual saliency, finding that models that account for saliency along with head direction achieve better performance in predicting gaze in dynamic environments. Nakashima et al. [172] further supported these findings by highlighting that saliency-based models are particularly effective in directing attention based on head movement in egocentric videos. Hu et al. [51] developed a model that integrates saliency cues with head movement, which was shown to outperform traditional models that rely solely on head direction. In a follow-up study, Hu et al. [168] refined this approach, demonstrating that further information, such as object movement with saliency features, leads to even more precise gaze predictions in VR and AR settings.

However, while head movements can be easily retrieved from HMD sensors in real time, calculating visual saliency is a resource-intensive process [51, 168]. State-of-the-art saliency models are typically implemented as deep neural networks (e.g., convolutional or transformer-based architectures) that operate on full-resolution video frames and must be evaluated on every frame in interactive settings. For instance, the top-performing DeepGaze IIE configuration [173] necessitates the simultaneous execution of four heavy CNN backbones to generate a single prediction. This massive per-frame inference cost competes directly with rendering resources on standalone XR devices, rendering such models impractical for real-time interaction and motivating our reliance on lightweight head-pose features instead.

2.5 Gaps in the Literature Addressed by this Thesis

2.5.1 Lack of Field-of-View Visualisations Tailored for VR Screens

The current literature on FoV techniques for collaborative virtual environments (CVEs) reveals several gaps that this study seeks to address. While prior research has demonstrated the benefits of FoV-style visualisations in enhancing mutual awareness of visual focus [39, 40, 41], these studies have primarily focused on 3D environments and asymmetric scenarios (Section 2.2.2). However, with the rise of VR office applications and XVA (Section 2.1.1), there is an emergent need to address the FoV-style visualisations of attention on 2D content (XR screens), particularly in symmetric collaborative settings (VR/VR) such as those found in VR offices and XVA applications [75].

Existing FoV-style visualisations may not be well-suited for these emerging 2D-in-VR scenarios. They could lead to occlusion of the 2D content, ambiguity in pinpointing the focus on the screen, or visual clutter when multiple users are present [40, 41]. Therefore, there is a clear gap in understanding how to effectively visualise and share visual focus when collaborators are interacting with 2D data (such as spreadsheets, web pages, or presentations) on VR screens. This necessitates the development of tailored FoV visualisation techniques specifically designed for 2D content in VR to enhance mutual awareness and support effective collaboration in these increasingly prevalent use cases. This gap will be addressed in Chapter 3, where a novel FoV visualisation technique tailored for 2D content in VR will be introduced.

2.5.2 Lack of Dynamic Representations for Cognitive Attention

While the literature identifies theoretical disconnects between visual sensors and cognitive intent (Section 2.2.5), existing collaboration cues in XR fail to bridge this gap practically. Current representations—whether raw gaze cursors or static FoV frustums treat attention as a fixed spatial coordinate rather than a dynamic cognitive process.

Failure to Handle Noise (Sensory Level): Raw gaze cues transmit high-frequency sensory

noise (e.g., reflexive saccades to motion or brightness), which does not necessarily reflect the user’s collaborative intent. Existing cues lack the filtering mechanisms required to distinguish between meaningful fixations and bottom-up distractions.

Static Scale (Perceptual Level): Cognitive attention naturally fluctuates between a ‘global’ scanning scope and a ‘narrow’ specific focus (Section 2.2.5). However, current XR cues utilize static geometries (e.g., a fixed-size cursor or a rigid cone) that fail to adapt dynamically to these shifts in scale, often misrepresenting broad situational awareness as precise scrutiny.

Absence of Decoupling Support (Cognitive Level): Finally, visual gaze often decouples from mental focus during complex tasks. Current systems assume a 1:1 mapping between eye direction and attention, lacking the multimodal input (such as speech) required to maintain a valid cue when a user looks away to think or speak.

There is, therefore, a gap in the design of cognitive-aware attention cues—visualisations that can filter noise, dynamically scale from scene to object level, and persist or vanish based on task relevance rather than just head/eye mechanics.

2.5.3 Bi-directional FoV-style Visualisations in Collaborative VR

The existing literature on FoV-style visualisations in collaborative VR has primarily focused on mono-directional cues, where only the collaborator’s visual focus is visible to the user [40, 41]. While bi-directional eye-gaze visualisations have been studied (Section 2.2.3), their application to FoV-style visualisations remains largely unexplored. The potential of bi-directional FoV-style visualisations, which allow users to see both their own and their collaborator’s visual focus areas, remains under-explored. There is a gap in understanding the impact of self-awareness of FoV visualisation on communication and task performances. It is important to investigate this further, as gaze visualisations and head-based FoV-style visualisations may yield different results due to their inherent differences in accuracy and the level of control users have over them. Gaze visualisations provide precise information about eye movements, while FoV-style visualisations offer a broader estimation of the attentional area based on head orientation. Understanding these distinctions is crucial for developing effective visual attention

cues in collaborative VR. This gap limits our understanding of how such cues could further enhance mutual awareness and coordination in collaborative tasks [98, 41].

2.5.4 Field-of-View Visualisations Specificity

Field-of-View (FoV) visualisations have been widely used in VR and collaborative environments as an approximation of users' visual attention. However, a fundamental limitation of this method is that it only provides an estimate of where the user might be looking based on head orientation, rather than their exact point of focus. Previous research has highlighted that FoV-based approaches fail to capture the dynamic and nuanced nature of visual attention, particularly when distinguishing between focused and distributed attention [40, 42]. Furthermore, the reliance on head movement as a proxy for gaze is problematic because it does not account for finer shifts in eye movement, which are crucial for precise attention representation [174]. As a result, while FoV-style visualisations provide valuable cues, they lack the specificity required for more accurate representations of visual attention, especially in collaborative tasks [42].

2.5.5 Lack of Precision in Gaze Inference Models

Existing models that utilise head behaviour and visual saliency often achieve an angular accuracy of approximately 7° , which is insufficient in high-density informational environments where precision is essential [175]. This inaccuracy can lead to errors in collaborative settings, where accurate gaze direction is crucial for effective communication and coordination. Moreover, much of the existing research focuses on predicting the exact gaze location as a 2D point. However, this approach can be error-prone in collaborative settings with dense information, as even minor prediction errors can lead to significant misinterpretations [90].

2.5.6 Lack of Real-Time Gaze Inference Models for Untethered VR Devices

Existing gaze inference models are resource-intensive, rendering them unsuitable for untethered, low-cost head-mounted displays (HMDs) that are increasingly prevalent in consumer VR markets [168]. The high computational demands restrict their real-time application on devices with limited processing power. Current approaches often depend on eye-tracking sensors, which are not widely available in affordable VR devices, or on visual saliency, which is computationally expensive and impractical for real-time use in dynamic VR environments [176].

2.6 Conclusion

This chapter has examined the existing literature on the coordination of visual attention, verbal communication, and non-verbal gestures in collaborative environments, emphasising their independent functions as well as their interdependence in immersive settings. A review of foundational studies has underscored the critical role of visual attention (Section 2.2) and verbal cues (Section 2.2.6) in fostering mutual awareness and effective collaboration. Additionally, the complementary nature of gaze, referring expressions, and pointing gestures (Section 2.3) in facilitating shared understanding and coordination has been highlighted.

An examination of the evolution of XVA environments (Section 2.1) highlights the trend of adapting XR for 2D content and multi-purpose applications, emphasising the need for effective communication tools in these settings. While gaze is a powerful tool for visual coordination, the lack of eye-tracking technology in consumer VR headsets (Section 2.4.1) presents challenges. Alternative methods, such as head movements and FoV-style visualisations (Section 2.2.2), have emerged as viable solutions for maintaining collaborative awareness.

Verbal communication, particularly in the form of referring expressions (Section 2.2.7), plays a crucial role in guiding attention, especially when visual cues are unavailable or ambiguous. Speech-to-text technologies (Section 2.2.8) hold the potential to further enhance XR collabo-

ration by augmenting traditional communication with AI-driven solutions.

Despite these advancements, gaps remain, particularly in designing effective FoV-style visualisations for 2D content in VR (Section 2.5.1) and increasing the specificity of FoV cues (Section 2.5.4). Additionally, current gaze inference models struggle with accuracy in high-density environments and pose challenges for real-time use on untethered VR devices (Section 2.5.5, Section 2.5.6).

This thesis addresses these gaps by introducing a novel framework, known as "complementary reconstruction", which leverages the natural interplay between gaze, verbal communication, and gestures. This framework compensates for technological limitations, such as the lack of eye-tracking, by using naturally occurring collaborative behaviours as implicit inputs to XR systems, thereby improving both human-human collaboration and the machine comprehension of interactions in immersive environments. Building on insights from the literature, this research explores how gaze and communication cues can be integrated into XR environments to enhance collaboration, contributing to the development of systems that support more effective, seamless collaborative experiences.

The next chapter introduces the Cone of Vision, detailing its design and evaluation as a novel approach to enhancing collaboration in XR environments.

Chapter 3

Cone-of-Vision

Use of the head as a proxy of visual attention for VR screens

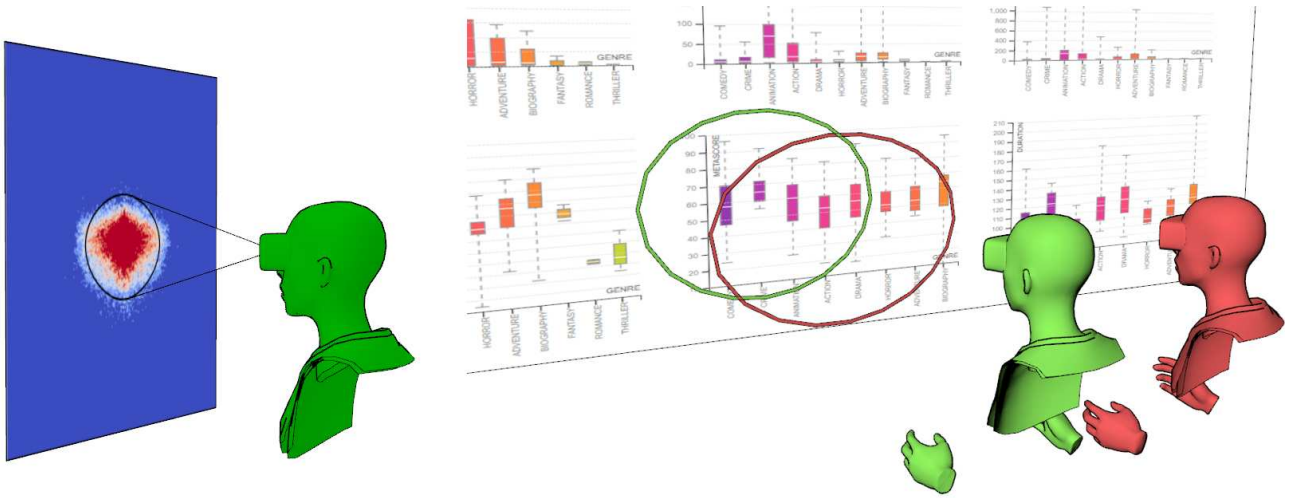


Figure 3.1: Left image: cone of vision modelled with average fixation map, an updated version of the FoV frustum visualisation. Right image: bi-directional attention cues between two users collaborating on an EDA task.

In this chapter, the concept of using head orientation as a proxy for visual attention is explored within VR environments, addressing the specific challenge of visualizing attention on 2D content displayed within VR. While previous research has demonstrated the potential of FoV-style visualisations for enhancing collaboration, these techniques have primarily focused on 3D environments and may not be optimal for emergent 2D-in-VR applications (Section 2.5.1). Moreover, existing FoV-style visualisations in collaborative VR settings have largely been mono-directional, meaning that they only display the collaborator’s visual focus without showing the

user’s own focus, thereby limiting mutual awareness (Section 2.5.3).

To address these gaps, the Cone of Vision (CoV) is introduced as a novel FoV visualisation technique tailored for 2D VR screens. The CoV estimates and displays the area where a user is most likely focusing their attention based on the direction of their head movements. By providing a shared understanding of visual focus, the CoV enhances collaborative experiences in VR in scenarios where eye-tracking technology is unavailable or impractical by fostering mutual awareness of visual attention (Section 3.3). Empirical studies highlight that CoV supports effective collaboration during specific phases (Section 3.3.1). Additionally, a comparison of bi-directional FoV-style visualisations to mono-directional ones is conducted to understand their respective impacts on communication efficiency, task performance, and user experience in collaborative VR settings. The findings offer insights into how self-awareness of FoV visualisation influences collaborative dynamics and informs the design of more effective attention-sharing mechanisms in immersive environments (Section 3.3.2).

The CoV’s capacity to leverage readily available head-tracking data aligns with the broader objective of exploring how naturally occurring user behaviours can be harnessed to support effective communication and collaboration in immersive XR environments, even in the absence of specialised hardware. This chapter contributes by introducing the CoV as a novel visual attention cue (Section 3.1.2.3), evaluating its effectiveness through empirical studies where participants perform an Exploratory Data Analysis (EDA) task, exploring its bi-directional and mono-directional representations (Section 3.3.2), and assessing its impact on collaboration (Section 3.3).

3.1 Study

The purpose of the study is to validate the proposed CoV design, understand which phases of collaborative EDA tasks benefit or are affected by the proposed CoV visualisation, and investigate how the self-awareness of CoV (Figure 3.2d) affects collaboration. Three conditions were defined: (1) a baseline condition where no CoV is visualised (Figure 3.2a); (2) a mono-

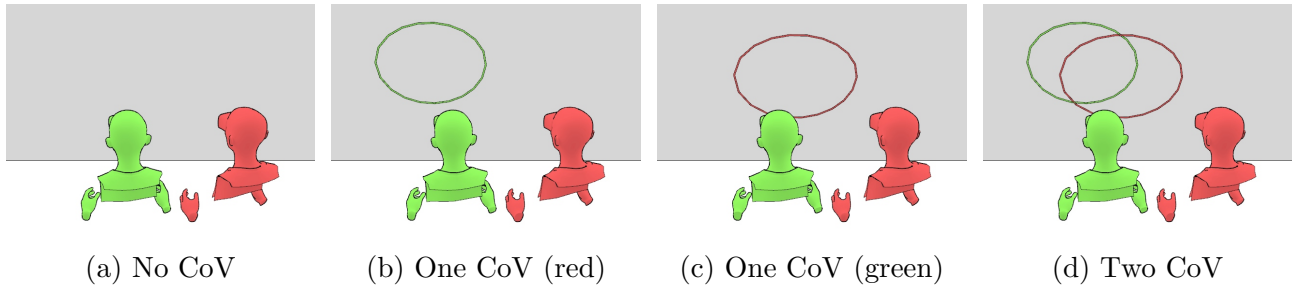


Figure 3.2: Experiment one Conditions. (a) No CoV: participants cannot see any CoV. (b)(c) One CoV: each participant can see the CoV of the collaborator but not their own. This visual cue is mono-directional as only the collaborator can see it. (d) Two CoV: participants can see their CoV and the collaborator’s CoV. This visual cue is bi-directional as it can be seen both by the user and the collaborator.

directional CoV where participants can see only their collaborator’s CoV (Figure 3.2b–3.2c); and (3) a bi-directional CoV where both participants can see their own and their collaborator’s CoV (Figure 3.2d).

3.1.1 Measures

To compare CoV conditions with the baseline, visual coordination and mutual awareness of visual attention were measured. Visual coordination was assessed by recording the behaviour of participants’ heads and calculating the rate of overlap of their field of view. This reflects how long participants spent pointing their heads concurrently toward the same target, as used in collaborative studies by Cordeil et al. [93]. To measure how FoV frustum visualisation increases mutual awareness during the collaborative stages, a self-reported score of mutual awareness and attention allocation was collected, [177]. In addition, the NASA-TLX questionnaire [178] was used to measure the load of the task changes across the CoV conditions.

To explore how the CoV impacts deictic pointing actions, the mutual understanding of spatial references was measured, along with the number of pointing actions performed by each participant. Mutual understanding of spatial references was assessed using an adapted subset of the co-presence questionnaire proposed by Biocca et al. [177]. The impact of the CoV on verbal communication was examined by measuring the amount of verbal communication [93] and the balance of verbal communication across participants [179].

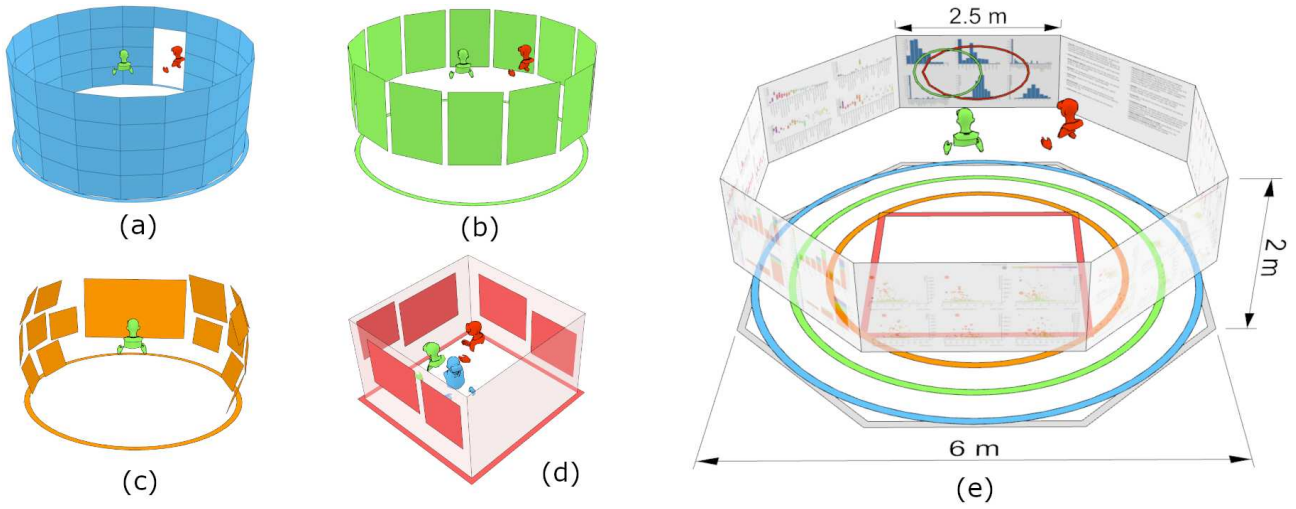


Figure 3.3: Inspired by Hybrid Reality Environment for Collaborative Information Analysis used by Cavallo et al. [84] image (b) and Febretti et al. [85] image (a) and in line with XVA studies Lee et al. [66] image (d) and Satriadi et al. [34] image (c) an ICVE was developed for the experiment (Figure 3.3) image (e) The ICVE is egocentric and consists of a series of 8 screens which are organised in a cylindrical layout around the participants defining the boundaries of the space. The image depicts the ICVE in which two participants can be seen collaborating while trying to complete the exploratory data analysis task. On the base of image (e) is possible to see a size comparison between previously explored layouts and our virtual space.

To understand how CoV conditions impact the performances of EDA tasks, the number of insights was recorded and classified by validity [180]. Participants were instructed to audio record insights by pressing a button on the controller before verbalising the insight and then pressing the same button to terminate the recording. All audio recordings were transcribed for further analysis. Whenever participants mentioned more than one fact extracted from the visualisation, insights were split unless they were comparing two facts or forming an interpretation. Only correct insights were counted for the analysis.

3.1.2 Apparatus

3.1.2.1 VR Environment

To keep the application development simple and to avoid noise caused by differences across VR HMDs, the study targeted only Oculus Quest and Quest 2 headsets. The selected VR headsets are six degrees of freedom (DoF) and untethered, chosen due to their popularity and low

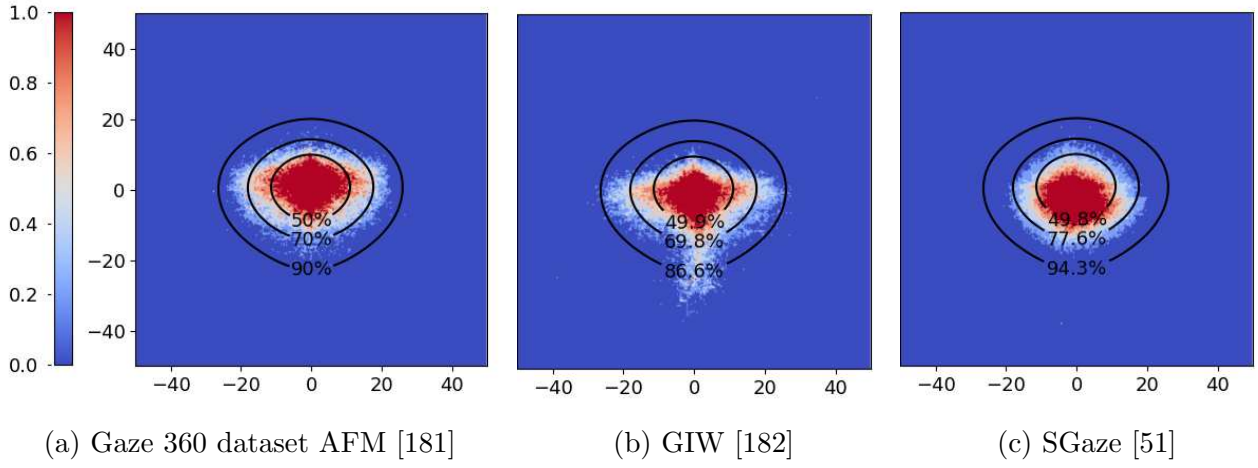


Figure 3.4: (a) (b) (c) shows the eye distribution in the longitudinal and latitude direction, for each dataset [181, 51, 182] and the amount of eye samples captured by the PBCs generated from [181]. On all three plots, the x-axes show the longitudinal eye angle and the y-axes show the latitudinal eye angle.

retail price. An application for collaborative visual analysis of 3D data was developed using Unity (version 2018.4.14f1) and the Oculus Unity SDK. The application enables displaying visualisations on web pages. Also, it allows each participant to join a real-time session in which the presence of other participants is represented by an avatar provided by the Oculus Avatar SDK (Figure 3.3e). Each participant in the VR space is free to move in any direction using the thumbstick controller or physically move using the six DoF of the VR HMD. Avatar movements are streamed via the network (using the Photon Unity Network API), so their behaviour (head and hand movements) and position in the virtual space are reproduced with low latency. The application also enables participants to talk to each other using the embedded microphone and speakers of the VR HMD. Additionally, the setup supports an observer/moderator to be present (i.e., being heard but not seen) during the VR session. The repository of the VR application can be found on a public GitHub repository.¹

3.1.2.2 Data Visualisations

Three visualisation views were developed using three different datasets: Hollywood movie gender bias (Theme A), the success of Hollywood movies (Theme B), and cars (Theme C). Theme

¹<https://github.com/Collaborative-Immersive-Visual-Toolkit/ConeOfVision>

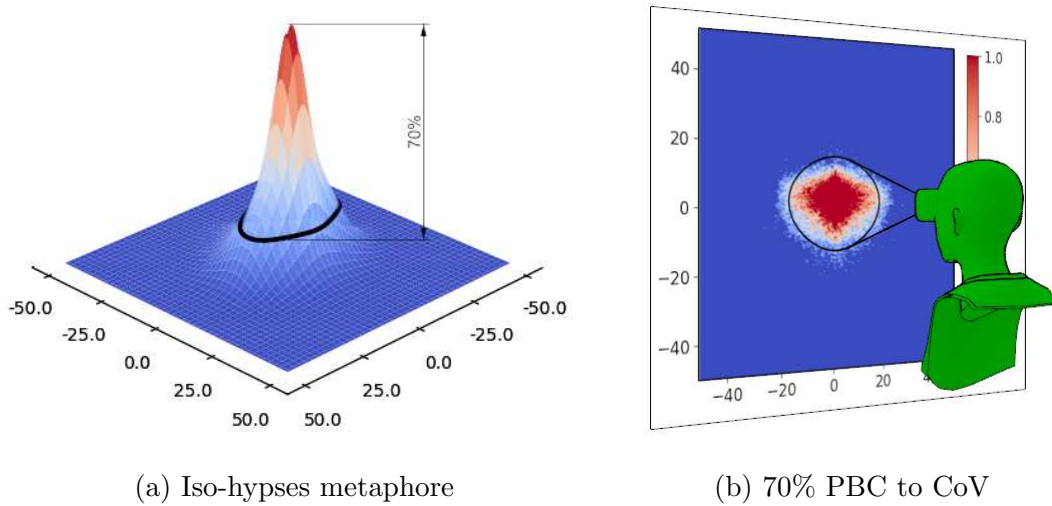


Figure 3.5: (a) shows the iso-hypsies metaphor where gaze probability is represented as the elevation, while the contour is extracted like an iso-hypsies. (b) From the identified percentile base contour (PBC) of 70% of the dataset [183], a 3D cone of view is generated.

A was inspired by The Bechdel Test [184]- a test that evaluates the presence of women in Hollywood movies by asking whether a movie features at least two named women characters who talk to each other about anything other than men. The data used for theme A was from [185]. Theme B was about the most successful movies of 2019, displaying attributes related to ratings, financial success, and available movie attributes (genre, language, etc.) taken from IMDB [186]. Lastly, Theme C displayed data about the risk rating calculated by insurance companies for cars, correlating with general attributes (engine size, fuel type, etc.), which was from the UCI machine learning repository [187]. Each theme contained 38 visualisations across the seven screens while one screen contained a glossary for the terms. Each theme also contained four chart types, namely scatter plots, box and whisker plots, histograms, and stacked bar plots. The scatter plot and box and whiskers plot had one interaction technique that allowed participants to click on items to get textual details (i.e., names of movies or details of the boxes). Each dashboard had a consistent number of visualisations and all visualisations were consistent in size. Captions were added under each visualisation to explain the visualisation and the labels used. The repository of the visualisations and dataset can be found on a public GitHub repository,¹ and the visualisations are publicly available on the Internet.²

²<https://graphs-for-collaborative-vr.web.app/>

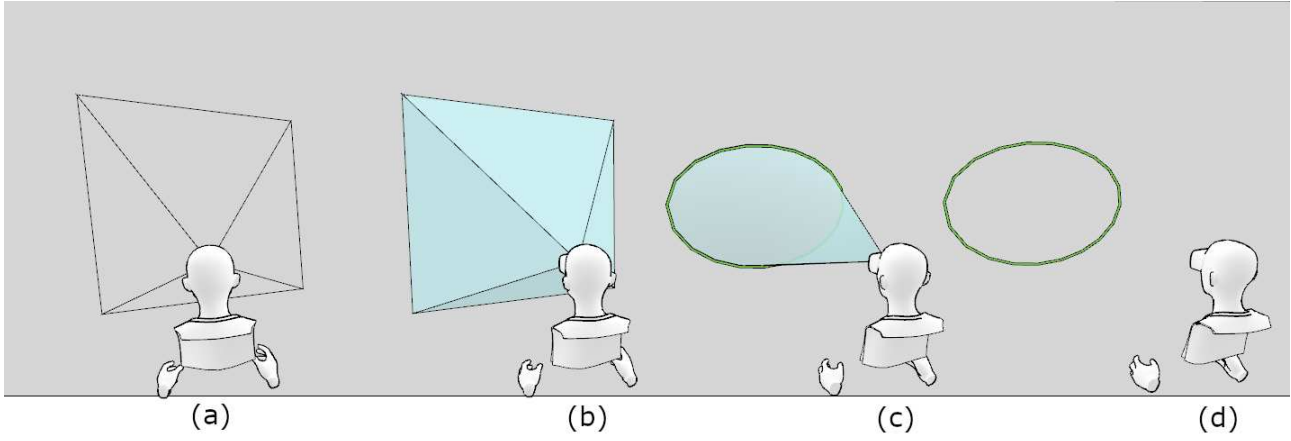


Figure 3.6: (a) Hindmarsh et al. [39] propose to visualise the FoV frustum as a wireframe (first on the left). (b) Piumsomboon et al. [41] propose to visualise the FoV frustum shaded (middle). (c) A shaded version of the CoV was also tested during the pilot phase. (d) Rather than displaying the FoV frustum, the intersection of a custom-calculated cone of vision (CoV) with the observed data is displayed.

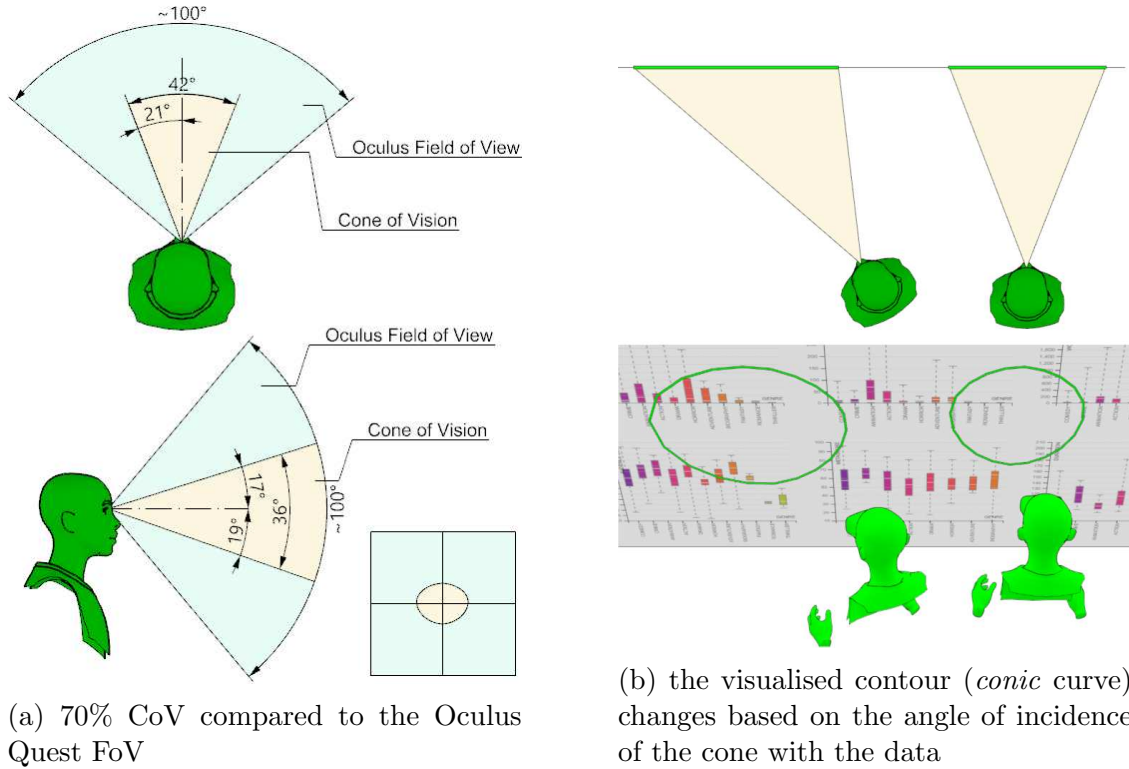


Figure 3.7: (a) For comparison, the Oculus Quest FoV is shown, which consists of approximately $\sim 100^\circ$ on the longitudinal and $\sim 100^\circ$ on the latitudinal. (b) Illustrates how the depicted intersection (green contour) changes according to the angle between the head and the 2D screen.

3.1.2.3 Cone of Vision

To generate the CoV, an average fixation map (AFM) [188] was created using users' gaze samples from a VR dataset recorded by Agtzidis et al. [181]. The dataset contains 14 partic-

ipants watching 360 videos in a VR headset. To generate the average fixation map, all gaze coordinates were converted from a global coordinate system to a head-based coordinate system, representing the latitude and longitude angles of eye direction in head coordinates (Figure 3.4a). The average fixation map serves as a good approximation of the probability density function (PDF) due to the high number of fixations contained in the dataset. The PDF approximation was used to generate percentile-based contours (PBC). Therefore, the procedure to extract the PBC at the selected percentile uses the PDF gradient to prioritise the areas that have a higher probability of fixations and stop when the count of gaze samples reaches the right percentile. This procedure can be compared to the extraction of iso-hypes from ground elevation information where the iso-hypes are the PBCs, and the elevation information is the gaze probability (Figure 3.5a). The contour shape is therefore generated from the data distribution rather than arbitrarily chosen as in previous FoV visualisation studies [40]. PBC extraction is performed using the library "matplotlib"³. After a geometric down-sample, each contour is reduced to a sequence of 20 points with two coordinates. Such percentage-based contours effectively represent the probability of the gaze location in a movement-agnostic and context-agnostic way. Based on the extracted PBC, the 3D cone was generated (Figure 3.7a). The 3D cone is static in shape but follows the head direction of the user. The 3D cone is not visualised in the ICVE; instead, when the generated cone intersects any of the 2D surfaces present in the scene, the cone-plane intersection is visualised, forming a *conic* curve. Only closed conics are allowed, and any open conic is automatically closed by following the edges of the 2D screen. The shape of the conic depends on the angle between the head (cone) direction and the plane (Figure 3.7b). The intersection is visualised as an elliptically shaped polyline where the colour depends on the user casting the cone. The size of the CoV is $\frac{1}{9}$ of the Oculus FoV (Figure 3.7a).

As the CoV is generated from datasets where participants were watching panoramic videos in VR [181], there might be important deviations from contours generated from user tests with different tasks and contexts. To understand how well the extracted contours capture visual attention from other scenarios, an evaluation was conducted across two other datasets: one from Hu et al. [51] and the second from Kothari et al. [182]. The SGaze dataset contains

³<https://matplotlib.org/>

the visual attention data of participants engaged in free exploration of a virtual scene with a 6-degree-of-freedom (DoF) VR HMD [51]. The Gaze in the Wild dataset (GIW) instead contains visual attention data of participants performing everyday tasks while wearing a mobile eye tracker [182]. Three PBCs (90%, 70%, 50%) were extracted from the probability density function (PDF) of the 360 video dataset [181], and their effectiveness in capturing data from GIW and SGaze was evaluated. Results show that the extracted PBCs capture a very similar amount of data in SGaze (91%, 75%, 51% as shown in Figure 3.4c) and in GIW (85%, 69%, 51% as shown in Figure 3.4b). The repository containing the code used to extract the PCB and evaluate it across the datasets is available on a public GitHub repository.⁴

3.1.2.4 Pilot

While the range of motion of the eyes is approximately 50° in any direction for a healthy adult [189], it has been found that they rarely rotate beyond 30° relative to the head [165, 190], and a more comfortable amplitude for observing targets is within 15° [191]. Eye amplitudes exceed 15° during head shifts, whereas lower amplitudes correspond to static target observations [192]. Based on this, the contour that best represents visual attention aligned with a comfortable eye position while observing static targets was identified. A pilot study was conducted with three pairs of participants from the department. Participants were given the option to choose the PBC (90%, 70%, 50%) that best depicted the focus of their visual attention. The 90% and 50% PBCs were excluded based on participant feedback. The 90% PBC was considered too broad to accurately reflect visual focus, whereas the 50% PBC was perceived as being too restrictive, frequently excluding those areas where participants felt their visual focus was directed. The selected PBC corresponds to 70% of the data, aligning with previous findings, indicating that a 15° horizontal eye amplitude is a comfortable eye position. Additionally, a 3D representation of the CoV was tested (Figure 3.6 (c)), with the lateral surface of the cone assigned a transparency of 10%. However, this semi-transparent structure made it more difficult for participants to see the data. The solutions suggested by Hindmarsh et al. [39], Piumsomboon et al. [41], as shown in Figure 3.6 (a-b), were not tested, as the size of the FoV was greater

⁴<https://graphs-for-collaborative-vr.web.app/>

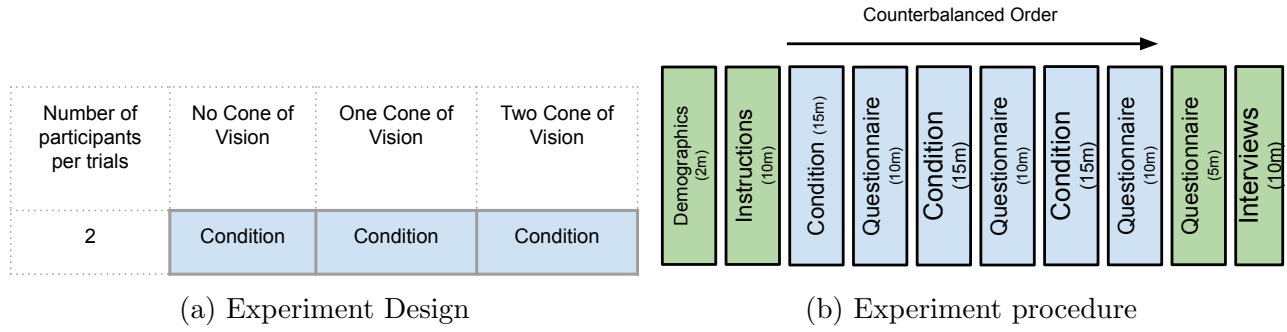


Figure 3.8: (a) In the first experiment 12 pairs of participants experience both experimental conditions: Absence of CoV/presence of collaborator CoV/presence of CoV as well as the presence of own CoV. (b) Experimental procedure.

than the already excluded 90% PBC shown in Figure 3.7.

3.1.3 Participants

Across a period of 90 days, 20 participants were recruited (6 women, $M_{\text{Age}} = 30.5$, $SD_{\text{Age}} = 5.9$) through online platforms. The study was advertised on the Oculus Quest 2 public Facebook group,⁵ Oculus subreddit community pages,⁶ the XR Distributed Research Network (XRDRN),⁷ and Prolific.⁸ These platforms were selected to target a population with an interest in VR who could potentially benefit from the CoV feature. A screening process was conducted with the following inclusion criteria: owning or having access to an Oculus Quest 1 or 2 (as the study was conducted remotely), full proficiency in spoken English, and normal or corrected-to-normal vision.

Nine participants were identified as heavy VR users, five as moderate users, and six as occasional users. Seventeen participants used Oculus Quest 2, while three participants used Oculus Quest 1. An additional requirement was familiarity with the chart types used for the EDA task, which consisted of histograms, bar plots, stacked bar plots, and scatter plots. Chart familiarity was measured through a preliminary screening questionnaire containing binary responses (yes or no). All participants had at least a bachelor's degree. Participants received a voucher equivalent

⁵<https://www.facebook.com/groups/oculusquest2>

⁶<https://www.reddit.com/r/oculus/>

⁷<https://www.xrdrn.org/>

⁸<https://www.prolific.co/>

Between conditions questionnaire		
Mutual awareness of visual focus	Q1: I was often aware of the other's visual focus in the environment Q2: The other was often aware of my visual focus in the environment	Likert scale agreement 7
Mutual understanding of spatial references	Q3: I understood the exact location of the other's spatial references Q4: The other often understood the exact location of my spatial references	Likert scale agreement 7
Embodiment	Q5: I watched the other participant's avatar to understand what they were doing Q6: The other person watched my avatar to understand what I was doing Q7: I spoke face-to-face with the other's avatar	Likert scale frequency 7
NASA-TLX	Q7-Q13 [178]	
System Usability Scale	Q13-Q23 [193]	

Table 3.1: Questionnaires

to \$20 for a 90-minute study. As an incentive, participants were encouraged to perform at their best to receive a bonus voucher equivalent to \$30, awarded to the pair of participants who reported the highest number of valid insights. Collaboration was further incentivised by informing participants that the chance of winning the bonus was shared between the pair and, therefore, each reported insight should be double-checked with the other participant.

3.1.4 Study Design

The study follows a within-group design, with three conditions (Figure 3.8). Participants were embodied in an avatar and used a hand pointer to reference the observed dataset. The order of the conditions was counterbalanced to prevent participant learning effects from biasing the study results. Additionally, the order of the visualisation themes (Section 3.1.2.2) used in each condition was counterbalanced to prevent any differences in data complexity from biasing the results.

3.1.5 Procedure

Participants had to fill out a screening questionnaire before joining the study. Participants were contacted via email to obtain consent and arrange a time for the study. They were then contacted via email to obtain consent and arrange a time for the study. The study began with a web conference to provide an overview of the study procedure and the three visualisation themes. Next, participants were asked to perform an EDA task, which involved extracting insights from the displayed visualisations. Examples of valid insights were provided. Instructions on navigating the virtual environment, recording insights, and using the hand pointer were then given. Once the instructions were clear, participants wore the Oculus Quest connected to the virtual environment and started the collaborative task. Between each VR trial (i.e., experimental condition), participants completed the mutual awareness questionnaire, task load, and usability questionnaire (Table 3.1). At the end of the experiment, semi-structured interviews were conducted with each pair of participants. During the interviews, participants were asked to report cases in which the CoV was helpful for the assigned task and cases where it was not (Section 3.2.2). The total time spent on the experiment ranged between 1 hr and 7.3 min and 1 hr and 32.7 min ($M = 1 \text{ h } 20 \text{ m}$, $SD = 12.7$), and the trial duration lasted between 10 min and 16 min ($M = 13 \text{ m}$, $SD = 3\text{m}$). The stop condition was reached when participants declared that the task was complete (i.e., saturation of insights) or when the 13-minute time limit was reached, at which point they were urged to complete the final insights. This study was approved by the UCL Interaction Centre (UCLIC) Research Department’s Ethics Chair.

3.2 Results

3.2.1 Statistical Analysis

To analyse the collected data, Friedman tests were performed across the three experimental conditions. Significance was set at 0.05, and the Nemenyi test was used for post-hoc analysis, where required. In summary, significant differences were found in the average overlap of CoV

(Figure 3.9a), workload (NASA-TLX) (Figure 3.9b), self-reported mutual awareness of visual focus (Figure 3.10b, Figure 3.10a), self-reported mutual understanding of spatial references (Figure 3.10c, Figure 3.10d), and embodiment usage (Figure 3.10f, Figure 3.10e).

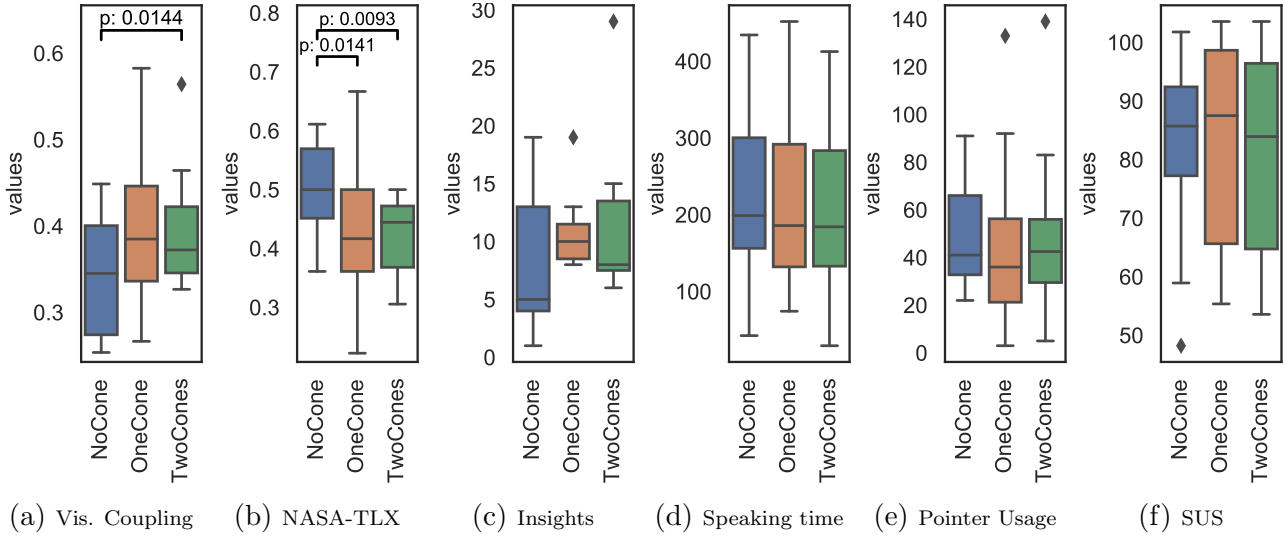


Figure 3.9: Behavioural Questionnaire Results

	Friedman test		Nemenyi test	
	Chi-Square	p-value	condition	p-value
(a) Visual coupling	14.4645	0.0285*	No CoV, One CoV	0.0781
			No CoV, Two CoV	0.0144*
			One CoV, Two CoV	0.9974
(b) NASA-TLX overall score	9.0	0.0111*	No CoV, One CoV	0.0141*
			No CoV, Two CoV	0.0093**
			One CoV, Two CoV	1.0
(c) Number of insight reported	4.16	0.124	-	-
(d) Speaking time	6.9905	0.033	-	-
(e) Pointer Usage	26.6	0.774	-	-
(f) System Usability Scale Overall score	15.0052	0.0257	-	-

* $p < .05$, ** $p \leq .01$, *** $p \leq .001$

Table 3.2: Statistical Analysis

3.2.1.1 Visual Coupling

Visual coupling (Figure 3.9a and Table 3.2) is a measure of visual coordination, giving us an idea of how visually synchronised the two participants were throughout the experimental condition. The measure is expressed in a unit scale where 1 means that their head direction was perfectly

synchronised and 0 means they never pointed their head towards the same target. As shown in Figure 3.9a, results show a significant difference in visual coupling between No CoV ($M_{\text{NoCoV}} = 0.34$, $SD_{\text{NoCoV}} = 0.07$) and Two CoV ($M_{\text{TwoCoV}} = 0.4$, $SD_{\text{TwoCoV}} = 0.08$). Therefore, the presence of the two CoVs leads to a significant 4% increase in visual synchronisation in the field of view overlap compared to no CoV.

3.2.1.2 NASA-TLX Overall Score

The NASA-TLX overall score (unweighted version [178]) was analysed by performing a Friedman test $\chi^2 = 9.0$, $p = .0111$. Since the resulting p-value is significant, each condition was compared using a post-hoc test (*Nemenyi*). The test shows a significant difference between No CoV ($M_{\text{NoCoV}} = 0.5$, $SD_{\text{NoCoV}} = 0.07$) and One CoV ($M_{\text{OneCoV}} = 0.42$, $SD_{\text{OneCoV}} = 0.07$) and between No CoV and Two CoVs ($M_{\text{TwoCoV}} = 0.42$, $SD_{\text{TwoCoV}} = 0.07$), but no significant difference between One CoV and Two CoVs. Therefore, the overall NASA-TLX value of the No CoV condition is statistically significantly higher than the One CoV condition and the Two CoVs condition. The presence of the CoV leads to a 5% reduction in the reported task load (Figure 3.9b and Table 3.2).

3.2.1.3 Number of Insights Reported

A total of 164 insights were reported by pairs of participants and statistically compared by condition. The analysis of the results did not highlight any significant difference across the experimental conditions, as shown in Figure 3.9c and Table 3.2.

3.2.1.4 Speaking Time

Verbal communication is an essential part of collaborative dynamics [93]. Using the HMD's microphones, the amount of time participants spent speaking was measured to determine whether the experimental conditions influenced the extent of participant communication. The results indicate no significant differences across conditions (Figure 3.9d and Table 3.2).

3.2.1.5 Pointer Usage

The number of pointing actions performed by participants during the trial was measured. The analysis did not reveal any significant differences across the experimental conditions (Figure 3.9e and Table 3.2).

3.2.1.6 System Usability Scale

The System Usability Scale overall score was analysed [194]. The analysis did not reveal any significant differences across the experimental conditions (Figure 3.9f and Table 3.2).

3.2.1.7 Mutual Awareness of Visual Focus (Q1 and Q2)

As part of the questionnaires, participants rated the following statements on a 7-point Likert scale (0 to 6): *Q1 I was often aware of the other visual focus in the environment* (Table 3.1 and Figure 3.10a), and *Q2 The other was often aware of my visual focus in the room* (Table 3.1 and Figure 3.10b). Variance was analysed using a Friedman test, and since the resulting p-value was significant, post-hoc comparisons were conducted using a Nemenyi test (Table 3.3). The presence of the CoV resulted in an increase in the perceived awareness of visual focus, shifting responses from undecided/slightly agree to agree/strongly agree. Notably, the increase was more pronounced in the two CoVs (i.e., self-aware CoV) condition for both Q1 and Q2.

3.2.1.8 Mutual Understanding of Spatial References (Q3 and Q4)

As part of the questionnaires, participants rated the following statements: *Q3 The other often understood the exact location of my spatial references* (Table 3.1 and Figure 3.10d) and *Q4 I understood the exact location of the other's spatial references* (Table 3.1 and Figure 3.10c). Variance was analysed using a Friedman test for Q3 and Q4 and, since the resulting p-value was statistically significant, post-hoc comparisons were conducted using a Nemenyi test (Table 3.3).

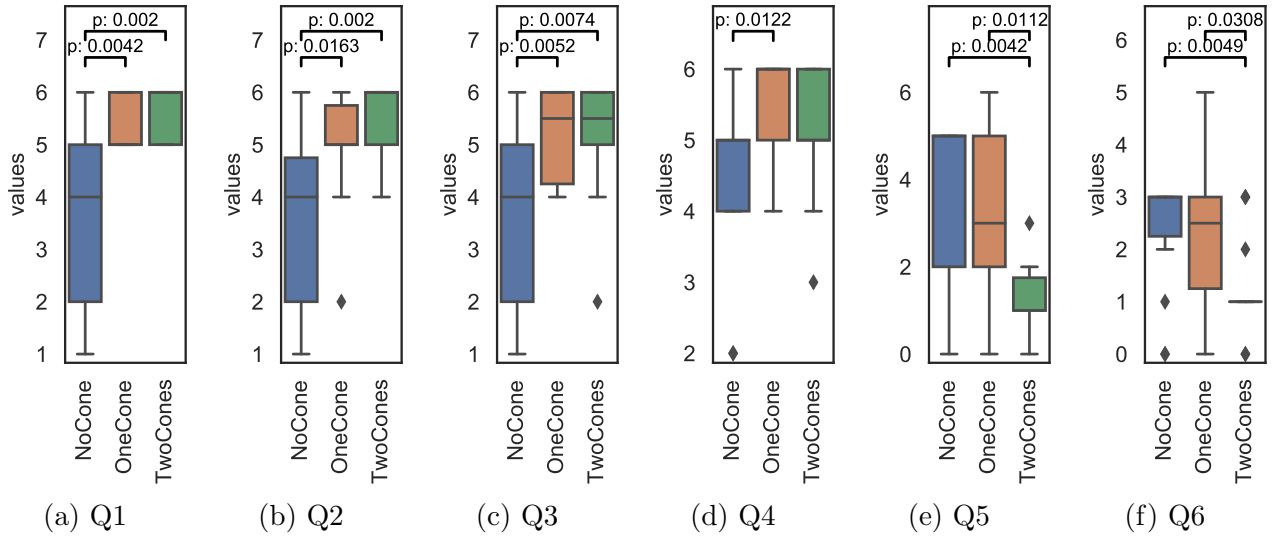


Figure 3.10: Questionnaires Answers

The perception of being understood showed an increase only in the One CoV condition, whereas the perception of understanding the other increased equally whenever the CoV was used.

	Friedman test		Nemenyi test	
	Chi-Square	p-value	condition	p-value
Q1: I was often aware of the other visual focus in the environment	18.0476	0.0001***	No CoV, One CoV	0.0042**
			No CoV, Two CoV	0.0020**
			One CoV, Two CoV	0.1797
Q2: The other was often aware of my visual focus in the room	15.6	0.0004***	No CoV, One CoV	0.0163**
			No CoV, Two CoV	0.002**
			One CoV, Two CoV	0.1605
Q3: I understood the exact location of the other's spatial references	12.4615	0.0019**	No CoV, One CoV	0.0052**
			No CoV, Two CoV	0.0074**
			One CoV, Two CoV	0.9141
Q4: The other often understood the exact location of my spatial references	6.8205	0.033*	No CoV, One CoV	0.0122*
			No CoV, Two CoV	0.1006
			One CoV, Two CoV	0.3804
Q5: I watched the other participant's avatar to understand what they were doing	12.1333	0.0023**	No CoV, One CoV	0.3822
			No CoV, Two CoV	0.0042**
			One CoV, Two CoV	0.0112*
Q6: The other person watched my avatar to understand what I was doing	10.1052	0.0257*	No CoV, One CoV	0.3822
			No CoV, Two CoV	0.0049**
			One CoV, Two CoV	0.0308*

* $p < .05$, ** $p \leq .01$, *** $p \leq .001$

Table 3.3: Statistical Analysis of Questionnaires

3.2.1.9 Embodiment (Q5 and Q6)

As part of the questionnaires, participants rated the following statements: *Q5 The other person watched my avatar to understand what I was doing* (Table 3.1 and Figure 3.10d) and *Q6 I watched the other participant's avatar to understand what they were doing* (Table 3.1 and Figure 3.10c). Variance was analysed using a Friedman test for Q5 and Q6 and, since the resulting p-value was significant for both, post-hoc comparisons were conducted using a Nemenyi test (Table 3.3). When Two CoVs were present (i.e., self-aware CoV), a decrease was observed in the extent to which participants reported observing the other avatar to understand their collaborator's actions. Simultaneously, a decrease was registered in how often participants perceived the collaborator observing their avatar.

3.2.2 Qualitative Analysis

The post-experiment semi-structured interviews were audio-recorded and fully transcribed. All transcriptions were coded, with initial codes derived from research questions and subsequently supplemented with those emerging from the interviews. The 42 codes identified were grouped into four themes which are reported in the following subsections.

3.2.2.1 The Cone of Vision Increases Mutual Awareness

The ability to see each other's CoV increases participants' mutual awareness of visual attention compared to the condition in which the CoV is absent. This means that they can follow each other better when moving their attention from one visualisation task to another. They were also able to better switch between collaborative and individual work.

Helped pacing collaborative analysis: Participants found that seeing the other's CoV was beneficial to pacing the progression of their collaborative analysis. Since participants can independently choose to focus on the same or different areas, they may split up without both being fully aware. For example, one of the two participants could move to another area without

notifying the other participant, which led to one of them being left behind without being aware of the split. On the contrary, one might move to another area, expecting the other to follow. Participants reported that, when the CoV was present, they would not unknowingly split: *“It was useful to see where the other would progress... because I could sort of see when P9 vision started to move on to the next series of graphs [P10]”, “If we were doing the same thing and he’d moved on, I don’t know if I would have caught it straight away without CoV [P15]”*.

The CoV plays an important role in enabling participants to keep their attention coupled. In other words, CoV’s presence helped to avoid splitting attention because of the lack of communication. The CoV allows participants to be aware of the other person’s head movements. Therefore, participants are able to know whether the other person was following their progression to the next visualisation. In fact, participants overwhelmingly reported how the benefit of the CoV consisted of being able to follow each other during collaboration: *“I was following his CoV more than I was following his pointer [P8]”, “I found the CoV useful in terms of continuing to make sure that we were focusing on the same visualisation [P9]”*.

Beyond showing vision spatial attention, CoV notifies when visual attention is shifting: *“It was useful for two things: be on the same page ... but also to pace us in moving so not just the position but also the movements of the cone vision was a useful signal [P12]”, “... if I started talking about something I could tell her CoV moved towards mine [P14]”*. Such comments highlight how the CoV differs from the pointer because it is implicit and therefore shows behavioural clues that would not normally be deliberate. In particular, some participants reported that such progressions to the next charts were happening without verbally discussing them: *“On some occasions at least I felt compelled to look at the same group of graphs that P2 was looking at even if we weren’t already talking about them or we did not already discuss to look at those [P3]”*.

Participants further reported that such synchronous pacing and progression of their visual attention across the visualised data happened without having to rely on the user embodiment: *“I didn’t understand if I didn’t look at his avatar unless I had the cone [P10]”, “seeing his CoV moving without having to turn and interpret the direction of their avatar made it easier*

to keep the same pace [P6]”. To observe the other’s CoV depiction, a user keeps facing the data, although to observe the collaborator’s embodiment, a user needs to look away from the data. Therefore, the CoV allows one to be aware of the other without turning away from the observed data.

Facilitate switching from independent to collaborative work: Participants found that seeing the other’s CoV was helpful when switching from independent to collaborative work. During the collaborative session, participants sometimes divided their attention to focus on different goals and then returned to a shared visual context and collaborated towards a common goal. When switching back to collaborative work, the CoV allowed the participant to understand where the other’s visual attention was without interpreting the other’s embodiment (i.e., interpreting the other head direction). *“when I was looking at something where P1 wasn’t looking then I would know that the CoV would help him to actually get on my field of view” [P2], “I think it was useful when I was looking at the glossary and P8 was looking at a visualisation then I could jump exactly to that visualisation that he was on” [P9], “like I said to be able to turn around and go like, oh that’s the direction that he’s looking at” [P13], “I found the cone was helpful to realign our focus” [P6].*

The visual cone depiction, therefore, removed the necessity of looking at the other person’s avatar as well as removed any ambiguity about the collaborator’s head direction: *“Instead of looking at him and trying to triangulate you know I was looking only the cone”[P9], “I didn’t have to like look at his avatar and then figure out I could just turn and then I’d see it already and my attention would then gravitate towards that” [P7].*

3.2.2.2 CoV specificity and hand pointer

Participants found that sometimes the CoV was not specific enough: the CoV provided the general direction of the visual attention but not its exact location. In other words, they highlighted how the CoV shows the general area: *“the cone told me what graph” [P1], “was useful to highlight the graph I was on” [P6], “I knew which window and graph the other was” [P8].* While

the pointer would highlight a specific feature: *"... while the pointer told me which column of the plot" [P1], "... but then again, I would use the pointer to increase the focus" [P6], "...but to point out the particular result I wanted to talk about I had to be more specific" [P8].*

While elaborating on the problem of specificity, participants highlighted how the eyes move independently from the head, thereby reinforcing how the CoV may only partially represent the location of the gaze, *"I move my eyes without moving my head so sometimes there was quite a disconnect between the CoV and my gaze" [P7], "it obviously was following tracking of my head rather than my eye " [P3];* some participants even mentioned eye-tracking as a feature to increase the specificity of the visual clue: *"if you had eye-tracking and you could pinpoint clearer where exactly in the graph I was looking that would be a bit better" [P8].*

Participants also mentioned how the size of the cone (and therefore its specificity) changes depending on their distance from the observed data. When far away from the data visualisations, the cone does feel not informative *"... the cone wasn't always representative of the graph that I was looking at because I was so far back and the cone was so big [P8]".* However, when moving closer, the size of the CoV felt more specific/accurate: *"To show that I was looking at a particular boxing whisker or a particular histogram bar I moved forward to narrow the radius [P10]".* Sometimes participants used locomotion just to make the cone smaller by moving closer to the data. However, when they felt the CoV was not precise enough, most participants used the hand pointer, e.g.: *"I would use the pointer rather because that was more focused" [P1], or "I would use the pointer to increase the focus" [P6].* Participants also highlighted that their familiarity with the hand pointer made it easy to use: *"Everybody knows the pointer, so it was in general fantastic [P5]", "when like in meetings I'm often using a laser pointer [P8]".*

While participants reported using the hand pointer when the cone was not specific enough, they also reported that they used the hand pointer less than in the case in which CoV was not visible. *"We used the pointer more in the first one because we didn't have the CoV [P12]", "we still used it to point out the details absolutely but we used it much much less because we didn't need to [P6]", "we didn't have to use the laser pointer every two seconds [P4]", "we would not use the hand pointer a lot [P15]", "I used the pointer less or at least like the impression [P7]".*

In addition, they reported using the pointer more frequently when their own cone was not visible to them: *“I used the pointer more for that last one [no CoV condition] to give more of an indication of where I’m looking at even though P4 could see where I’m looking at [P8]”, “When I could not see my cone, I wasn’t sure what the other could see so I used the pointer more to make sure he would understand me [P7].”*

3.2.2.3 The self-awareness of CoV offers a consistent shared representation

In their self-awareness of the CoV condition, participants reported that they had a better perception of what the other collaborator would see and this allowed them to rely more on the CoV as a behavioural cue. In contrast, in the other conditions, they were unable to clearly depict what the other participant could see: for example, two participants highlighted that, because they could not see their own CoV, they felt they used the pointer more often to ensure that the other participant understood what they were talking about: *“.. I knew the other could see my CoV, however, because I could not see it, I would use the pointer more... [P13]”, “..because we could not see our own cone, we could not rely upon it that much [P15]”*. Another example consisted of two participants being confused by the experimental condition in which their own CoV was not visible. Despite having explained the different experimental conditions to participants in the instructions, two participants believed that, because they could not see their own CoV, their collaborator could not see it either: *“In the first condition I thought only my collaborator’s CoV was visible, it took me a while to realise the other could see mine [P2] ”, “..yes at the beginning I did not realise my cone was visible, for me it’s easier to have a shared reality so that if he sees my cone I can see it as well so that both cones of vision can be seen [P7]”*. Essentially, a consistent representation of the shared reality leaves no doubts.

3.2.2.4 Confusion and distraction caused by the CoV

If both CoVs are visible, participants can be confused about the ownership of the CoV. To clear their own doubt, participants moved their heads or memorised the colour of their CoV. Participants reported: *“I was a little bit confused which one was mine and which one was hers*

and I had to move my head [P10]”, “I was relying on the colour of the CoV I didn’t have any problem distinguishing the two of them [P8]”, “... got a bit confused because I forgot which colour I was [P5]”, “mine always moved with me so it was difficult to get confused [P9]”.

One participant found that the collaborator’s CoV may distract the collaborator when they are explicitly engaged with the same dataset: *“In that case, I already knew where we were. When we were already looking at the same data and we’re talking explicitly about it, so it was for me sometimes distracting [P8]”*. Participants also reported that CoV overlapping occurred even if this was negligible: *“The only time that it would overlap is it just when I wasn’t paying attention or something like that but it never was the case [P3]”, “I mean when we had both our CoV things showing up even in the surveys when I mentioned if it overlapped certain data points it never happened really [P6]”, “I don’t think it ever got in the way for me... maybe there was a couple of points where the outline of it was overlap a part of the graph I wanted to look at, in which case I just moved my head slightly and so it was it wasn’t like practically an issue at all [P12]”, “The cone never got in the way of what I was doing [P10]”.*

3.3 Discussion

When the CoV was present, an increment in the awareness of the collaborator’s visual attention was measured. This increment was more pronounced in the two CoV conditions tested (Figure 3.10a). While the number of reported insights remained consistent across conditions (Figure 3.9c), the conditions in which the CoV was present required a lower cognitive load (Figure 3.9b, Section 3.2.1.2).

The qualitative analysis provides insights into the observed reduction in cognitive load (Section 3.2.2.1). Increased awareness of visual focus was particularly beneficial when participants transitioned from independent to collaborative work and when coordinating progression across visualisation. CoV visualisations facilitated these tasks by eliminating the need to check and interpret the collaborator’s avatar head direction. The effort saved in these scenarios accounts for the decrease in overall cognitive task load.

However, the qualitative analysis also indicates that this increase in awareness pertains only to the general direction of visual attention rather than its precise focus (Section 3.2.2.2). To compensate for this lack of specificity, participants either used the hand pointer or, in rare cases, moved closer to the displayed data to narrow the CoV. Notably, the CoV did not influence laser-pointing behaviour during collaboration, although participants perceived a reduced need to use the pointer.

3.3.1 Cone of Vision and Collaboration Phases

Piumsomboon et al. [41] suggest that an intelligent system could autonomously choose when to show or hide FoV frustum visualisations based on the detection of collaborative context. However, the previous literature does not explore which phases of collaboration the FoV visualisation should display while suggesting such an idea. Our study identifies three principal collaboration phases during EDA that appear to benefit from CoV visualisations. The first phase consists of switching from independent to collaborative work. A context-aware system could, therefore, temporarily display the CoVs when it detects one participant joining the other for collaborative work. The second phase consists of pacing progression across the visualisations and essentially means the negotiation phase in which participants choose to move to the subsequent visualisation on the displayed dataset. In this case, again, an intelligent system could temporarily display the CoVs when it detects one participant moving to the next visualisation/laser-pointing screen. The third phase consists of displaying the CoV when a laser-pointing action is performed by one of the participants. In this case, again, an intelligent system could temporarily display the CoVs when a laser-pointing action is performed. In this case, the producer of the pointing gesture will be able to understand whether the laser pointer is seen or otherwise by the collaborator.

3.3.2 Cone of Vision Self-Awareness (bi-directional cue)

Results also highlight certain benefits concerning the CoV's self-awareness (i.e., two CoV conditions, bi-directional method Figure 3.2d). The statistical analysis of the questionnaire responses shows that, when participants could see their CoV, they reported watching the other avatar less than in the other two conditions (Figure 3.10e). These results show that CoV self-awareness makes participants feel that their collaborator understands what they are doing without watching their avatar. Statistical analysis of questionnaires also shows that participants reported being watched less by their collaborator (Figure 3.10f) in the self-awareness of CoV condition (i.e., two CoV conditions). These second results show that they watch the other avatar less in the two CoVs conditions than in the other experimental conditions. Qualitative analysis gives us an insight into why self-awareness of CoV impacts how participants perceive the other one, even if his/her own CoV movements are untangled from the other. One possible explanation is that the CoV perceived reliability is correlated with CoV self-awareness (Section 3.2.2.3). Qualitative analysis confirms our hypothesis that the bi-directional cue ended up creating a more consistent representation of the shared reality. Such a representation led to greater reliance on the CoV visual feedback in general. Our results highlight a second benefit concerning the Two CoVs (i.e., CoV's self-awareness), for which a statistically significant 4% increase in visual coupling was measured (Figure 3.9a, Section 3.2.1.1). Participants spent more time pointing their heads toward the same data when they could see their own CoVs (i.e., approximately 15s on a 6 min task). Both the qualitative analysis and questionnaire analysis give us ideas about how the CoV's self-awareness caused an increment in visual coupling. The questionnaire's statistical analysis shows us that, in the two cones of vision, there is a need to check the collaborator's avatar to understand the focus of their attention (Figure 3.10e). The time saved by avoiding checking and interpreting the collaborator avatar during EDA progression could be the reason for the increment in head direction overlap. By comparing mono-directional (i.e., One CoV) and bi-directional (i.e., Two CoVs) visual attention cues, the benefits of the latter over the former are outlined, thereby extending the work of Jing et al. [98].

3.3.3 Deictic Pointing

An important implication for design is that, when integrating the CoV into applications, it should not be considered a replacement for pointing. Our results highlight that CoV and laser pointing complement each other, especially during deictic pointing. Indeed, when CoV was present, participants reported that their understanding of spatial references, either verbal or pointing actions, improved with the use of CoV, particularly under conditions of self-awareness of CoV (Figure 3.10d, Section 3.2.1.8). Such increments can be explained by the fact that the CoV provides a continuous spatial reference (i.e., continuously depicting the general area of visual attention), which allows interlocutors to anticipate and verify referential intentions more efficiently. Unlike discrete cues such as pointing or gaze shifts, the CoV persists over time, reducing ambiguity in pointing references. Indeed, Wong and Gutwin [195] described how awareness of visual focus plays a role in the different phases of deictic pointing actions. The CoV supports deictic pointing actions by enabling the producer of the gesture to evaluate whether the observer will be able to see it (i.e., collaborator's CoV over the gesture's intended target). Additionally, the CoV can support the holding phase of deictic pointing: consisting of holding a pointing gesture until the intended observer has seen it (i.e., laser pointer directed within the observer's CoV). The benefit, in this case, is for the producer of a gesture to be able to understand whether a gesture will be seen, or has been seen, without looking at the observer's avatar or hearing a verbal confirmation from the observer.

Chapter 4

Model for gaze probability regions

Use of head movements to improve the probability of gaze

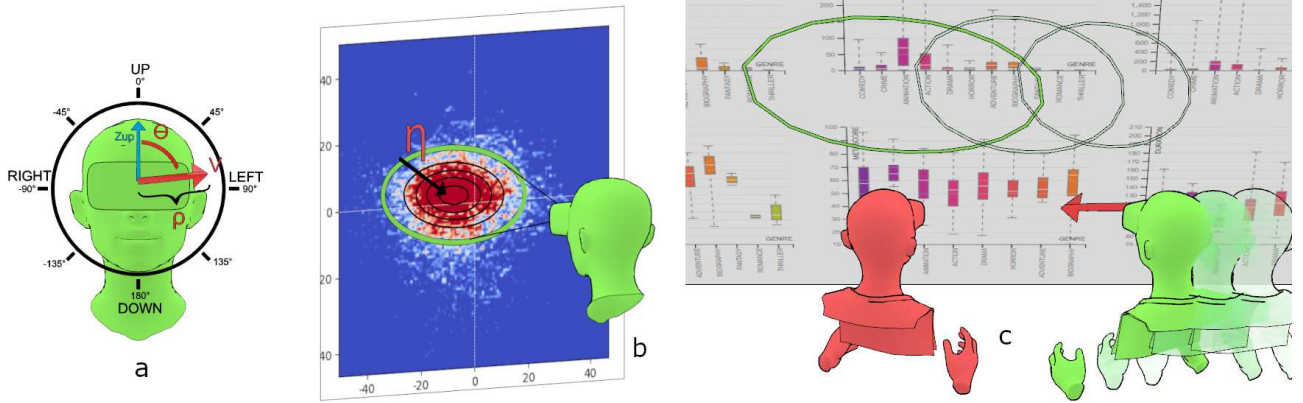


Figure 4.1: a) The head rotational velocity direction (θ) and magnitude (ρ) are extracted during a VR session. b) Probability density functions are extracted from eye-gaze distributions that correspond to the head rotational velocity and are converted into a series of percentile-based contours (η). c) Our real-time model uses the three parameters (θ, ρ, η) to provide a novel representation of visual attention for VR collaboration or interaction.

In Chapter 3, we introduced the CoV as a technique through which to enhance visual attention inference in VR. While the CoV significantly improved mutual awareness of visual focus and collaboration, our findings also revealed a key limitation in that the inferred gaze region was still too wide and lacked the necessary specificity (Section 3.2.2.2). Participants often compensated for this limitation by relying on explicit gestures or refining their verbal communication. This limitation underscores the need for a more accurate, real-time gaze inference approach that reduces ambiguity.

This chapter explores the enhancement of visual attention inference by leveraging head movements through a real-time deep-learning model. This model is designed to infer gaze probability regions in VR environments, particularly in cases where eye-tracking technology is unavailable or impractical. Current methods for inferring and displaying visual attention often fail to address the ambiguity of gaze regions (Section 2.5.5), which can impact the accuracy and efficiency of communication and collaboration in VR [36]. Additionally, existing gaze inference methods are resource-intensive, making them unsuitable for untethered, low-cost HMDs (Section 2.5.6).

To address these two gaps, this chapter proposes a model that enhances visual attention inference by utilising head movements through a real-time deep learning approach. By analysing head rotational velocity, expressed as magnitude (ρ) and direction (θ) (Figure 5.1a), the model generates a PDF that identifies likely gaze regions as percentile contours. A multi-stage pipeline is employed (Figure 4.3), involving an autoencoder for PDF generation (Section 4.1.4.1) and a Contour Encoder for real-time inference (Section 4.1.4.4). Unlike traditional Gaussian-based gaze estimation methods [172, 169], the proposed model generates gaze probability contours using an autoencoder-based approach, allowing for a more nuanced representation of visual attention that adapts to head movement dynamics. Validation against a dataset of gaze and head movements in VR demonstrates that the model, particularly during head movements (Section 4.1.4.2), surpasses head-based baselines in accurately predicting gaze probability regions, offering a refined understanding of visual attention that can significantly improve user interaction and collaboration in VR.

This chapter presents a real-time deep learning model designed to estimate gaze probability in VR, advancing the goal of leveraging natural user behaviours to infer gaze without the need for specialised hardware. By offering an accessible solution, this work enhances the precision of FoV visualisation in VR, particularly in contexts where cost or technological limitations preclude the use of eye-tracking systems.

4.1 Model

Previous works focused on predicting the exact gaze location as a 2D point. This approach can be error-prone in collaborative settings when the scene contains high-density areas of information since even a few degrees of prediction error can point towards entirely different pieces of information. To address this issue, I proposed capturing the shape of the gaze distribution instead of focusing solely on the most informative point. This approach exploits the trade-off between the ambiguity of a region’s information and the error that a point can introduce. Our model’s inputs consist of the head’s rotational vector (θ, ρ) and contour percentile (η) ; the latter represents the trade-off between point/region and error/ambiguity (Figure 5.1b, Figure 4.2b and Figure 4.2c).

4.1.1 Task-independent Nature of Eye-Head Coordination

The proposed model posits that eye-head coordination is governed by “hard-wired” physiological constraints—specifically ballistic [100, 164, 163] and anticipatory mechanisms [165, 166, 99]—that remain consistent across visual tasks (see Section 2.4.2). Whether a user is passively watching a video in VR or actively searching for data on a VR screen, the biomechanics of head rotation are a stable, bottom-up process [47, 51]. Consequently, a model trained on panoramic video data [181] captures a robust gaze distribution that should generalize effectively to other contexts. This premise is supported by our evaluation in Chapter 3 (see Figure 3.4a), where static contours derived from passive video viewing [181] successfully generalized to active task datasets (SGaze [51], Gaze-in-Wild [182]). Since the proposed model extends this logic by conditioning on dynamic head velocity rather than static orientation, we expect this cross-task robustness to persist.

4.1.2 Visualisation Dimensionality

PDFs are commonly visualised as a gradient to show the model’s continuous probability output in space. Alternative visualisations of PDFs comprise percentile-based contour lines or a closed line that divides the space into two regions, inside and outside, offering a binary classification. These visualisations only partially capture the PDF unless multiple lines are visualised (e.g., isohypse geological altitude representations or isobars in meteorology maps). As the model’s purpose is to visualise where a user’s visual attention is likely to be without occluding the virtual scene, a line-type visualisation is sufficient to convey the information and, unlike a gradient-type visualisation, does not occlude the scene. Moreover, a gradient in 2D space requires a bi-dimensional coordinate plus the associated value, while a line requires a bi-dimensional coordinate but no associated value.

4.1.3 Head Rotational Velocity Direction and Magnitude

Head rotational velocity is parametrized as magnitude (ρ) and the angle with the z-axis of the space (θ) (Figure 5.1). Such parameterization was chosen to enhance interpretability by decoupling velocity direction and magnitude, unlike previous research where these components are embedded within two Cartesian coordinates. This parameterization allows us to create a simple analysis tool to navigate the dataset distributions and inform our machine-learning method.

4.1.4 Deep Learning Pipeline

A dataset created by Agtzidis et al. [181], which includes both eye-tracking and head-tracking data in a virtual environment, was selected. They recruited 13 participants and played 15 360-degree panoramic videos for each participant for a total of 195 video sessions. The sessions were randomly split, with a ratio of 2:3 (130 videos, 936000 data samples) for training and 1:3 (65 videos, 468000 data samples) for evaluation. A series of gaze distributions were generated

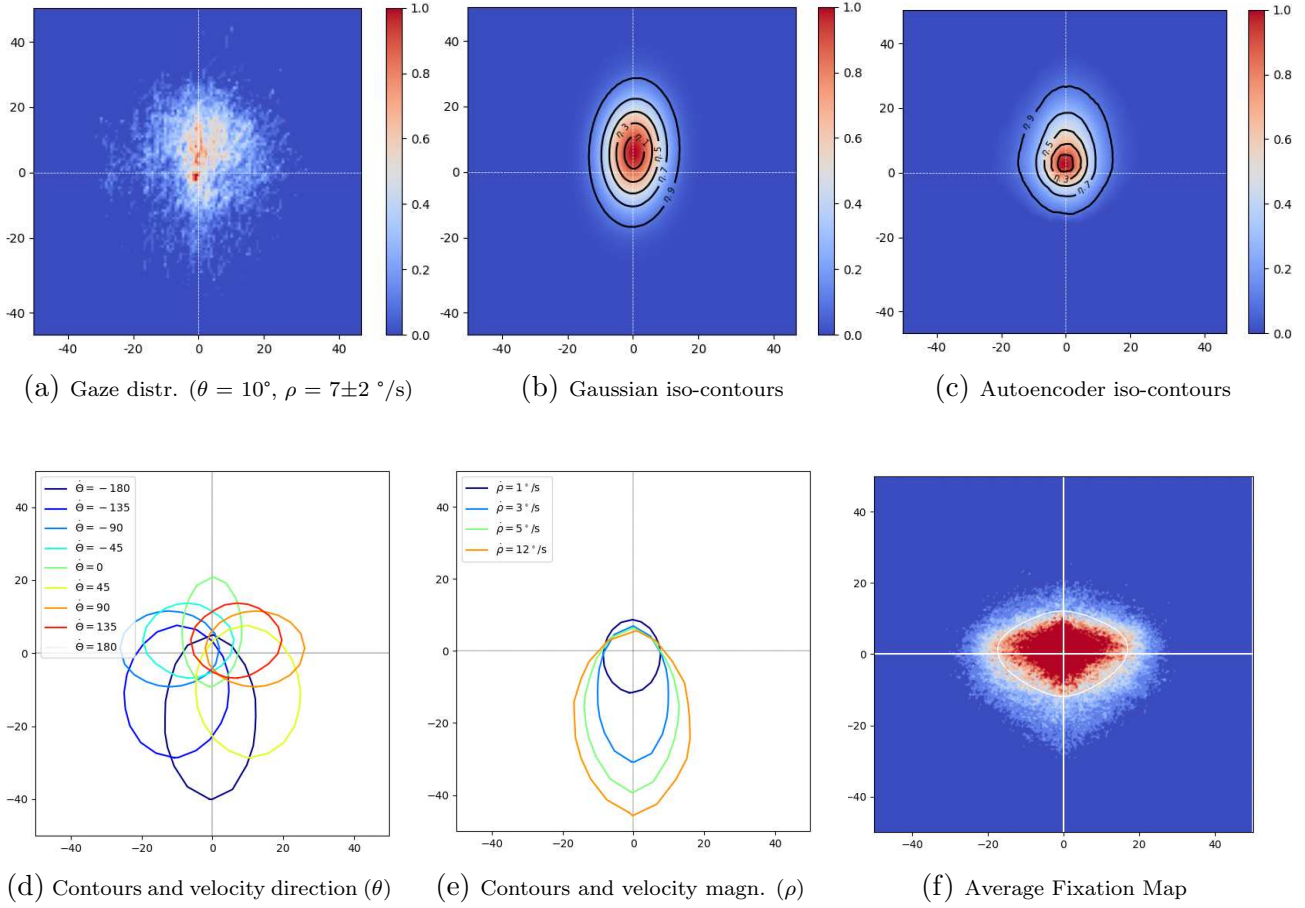


Figure 4.2: a) A distribution of eye gaze samples generated from the two parameters of head angular velocity direction $\theta = 10^\circ$ and magnitude $\rho = 7^\circ/\text{s}$ b) The PDF extracted via the MGF (Section 4.1.4.2), and the percentile-based contours extracted from the PDF. c) The PDF extracted via the Autoencoder (Section 4.1.4.1), and the percentile-based contours extracted from the PDF. d) A depiction of how contours change depending on the value of head angular velocity direction (θ) while the percentile ρ is fixed. e) A depiction of how contours change depending on the value of head rotational velocity magnitude (ρ) while the percentile parameter and θ are fixed. f) Average Fixation Map (AFM): All gaze data samples from the dataset [181] are plotted, and an AFM is generated as proposed by Radkowski [196]. From the AFM, the contour representing %70 of the data is extracted. This image clearly shows that the gaze distribution does not follow a Gaussian shape.

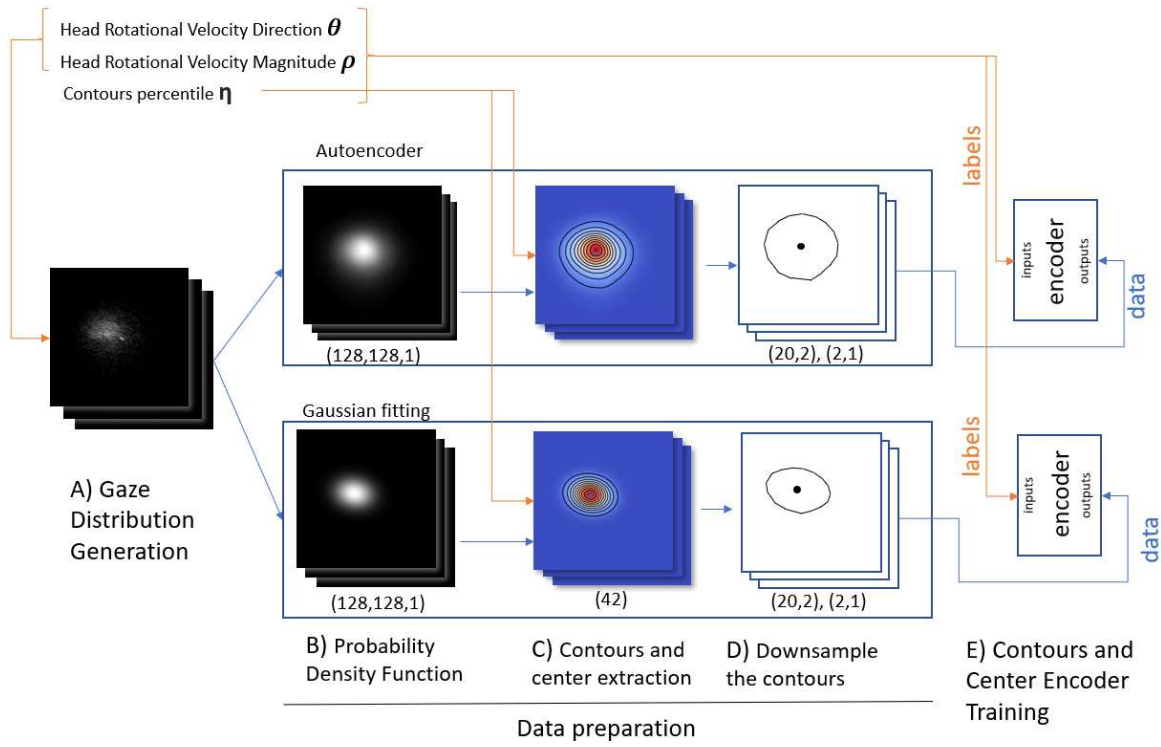


Figure 4.3: Our model’s three-stage training pipeline involves two deep learning models and a contour extraction phase to process velocity angle and magnitude (θ, ρ) to generate the contour that contains the requested probability of gaze area. The contour encloses the visual interest area and includes the predicted gaze location. The first pipeline training stage consists of an autoencoder reconstructing the 2D PDF from gaze location distribution samples, or an MGF approach. The second stage is a multi-perceptron architecture that encodes contours from the angles and magnitudes of head-shift velocity.

from the dataset by filtering the data using different head rotational velocity directions (θ) and magnitudes (ρ). These distributions were then saved as grayscale images with a resolution of 128×128 . Observation highlighted that the gaze samples did not always exhibit a Gaussian shape, necessitating the use of two different methods for generating PDFs. The first method employed a trained autoencoder (Section 4.1.4.1) that captures the shape of the distribution, while the second method involved fitting a multivariate Gaussian distribution (Section 4.1.4.2). Both models output gray-scaled images of the same size (128×128). The centre and percentile contours were extracted from each PDF, paired with the corresponding head angular velocity angles and magnitudes. A Contour Encoder was then trained to learn the various centres and contours from the PDFs, grouped by head angular velocity (θ, ρ) and the contours percentile (η , Figure 5.1b). A shallow model was used, avoiding complex neural network architectures, such as those proposed by Hu et al. [51, 168].

4.1.4.1 PDF from Autoencoder

The dataset's eye gaze distribution does not exhibit a Gaussian shape, as shown in Figure 4.2f, where a diamond-like shape is visible in the AFM. Therefore, the autoencoder aims to provide an alternative approach to the standard Gaussian model. The autoencoder was developed using Keras/TensorFlow 2.7.0 and follows a standard autoencoder architecture. It was trained with a set of multivariate distributions, including Gaussian, Poisson, and Skewed Gaussian. To generate the training set, the PDFs of these distributions were used as labels, while the input data were generated by applying uniform noise to the PDFs. This method produces distributions that visually resemble real-world gaze-tracking data. This variant is referred to as the *PDF Autoencoder*, which accepts grayscale 128×128 images as input. The output of the PDF Autoencoder is the estimated PDF, given the heatmap of the gaze distributions obtained after filtering the data based on the direction and amplitude of the head's rotational velocity.

4.1.4.2 PDF from Multivariate Gaussian Fitting (MGF)

Previous work applied Gaussian fitting to describe the distribution of gaze locations [172, 169]. However, these implementations split the X and Y components of the distribution, resulting in a loss of correlation and the removal of the Gaussian shape's tilting orientation. In contrast, the proposed MGF method is designed to allow the Gaussian shape to incorporate a tilt value, enabling a better fit to the distribution shape. Scipy 1.7.2 was used to compute MGF. Both the training set for this method and the dataset for the PDF Autoencoder were generated using the same initial samples, with random values assigned to θ and ρ .

4.1.4.3 Contour extraction

The PDF images generated by the auto-encoder and the MGF are processed to extract contours. Every percentile-based contour is associated with the two parameters of head velocity (θ, ρ) and the percentile value (η). The three parameters (θ, ρ, η) represent the training set for the next stage of the pipeline (Figure 4.3). The prediction stage does not require the autoencoder, and

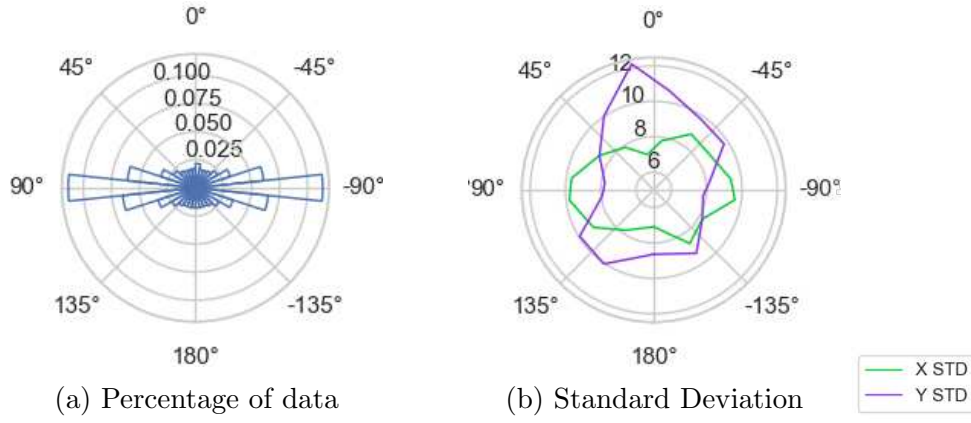


Figure 4.4: A) Percentage of data samples in the dataset for each head angular direction. B) X and Y standard deviation of the eye-gaze samples' distribution for each head angular direction.

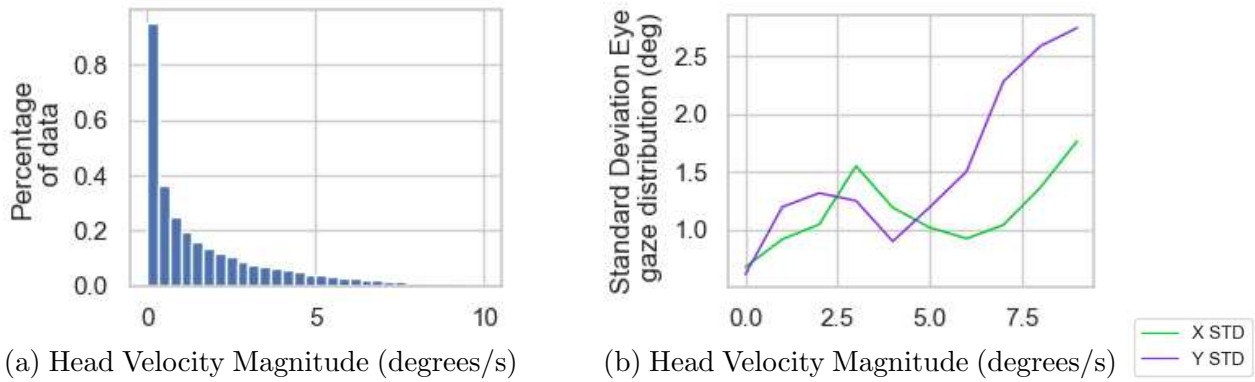


Figure 4.5: A) Percentage of data present in the dataset for each Head Velocity Magnitude (degrees/s). B) X and Y Standard deviation of the eye-gaze samples' distribution for each value of Head Velocity Magnitude (degrees/s).

such knowledge is embedded directly into the next model (Contour Encoder Section 4.1.4.4). After generating the resulting PDF images from the autoencoder and the MGF, contour extraction was performed on the following percentiles ($\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_7, \eta_8, \eta_9, \eta_{11}$) of the samples of the gaze distributions. Contour extraction is performed via a Python script and the library "Shapely". After a geometric downsample, each contour is reduced to a sequence of 20 points with two coordinates. The centre of the contour was also included as an additional dataset point to represent the estimated gaze location. The grouping of numerical input data and these contours represents the training dataset used in the final training phase.

4.1.4.4 Contour Encoder

The second part of the training pipeline consisted of the Contour Encoder, which is utilised for predicting the contour and the gaze location. The PDF is visualized as an iso-contour, adopting a lower dimensionality than a PDF gradient. Consequently, the neural network does not require 2D convolution (Section 4.1.2). Furthermore, Recurrence Neural Networks (RNNs) were not employed for two reasons. First, the resource consumption associated with recurrence may not meet the real-time performance requirements. Second, the training of RNNs can be affected by a vanishing gradient problem, which is exacerbated by the high number of eye fixations with a head rotational velocity of zero in the dataset (see the histogram in Figure 4.5a). To achieve fast inference and portability to mobile headsets, the Neural Network (NN) architecture was kept as simple as possible. Therefore, an MLP architecture was selected. The NN was designed as a simple fully connected model with four layers of 2048 hidden units. The model was trained using the contours generated from both the PDF Autoencoder and MGF for comparison. The input to the Contour Encoder consists of three parameters (θ, ρ, η) , while the output consists of 42 parameters: 20 2D points for the contour and one 2D point for the gaze.

4.1.4.5 VR-NN model implementation

TensorFlow Lite was utilised to port the model to an untethered VR headset by converting it into an Android-compatible version that can be loaded and executed via the Unity framework. Additionally, model stability was improved when velocity approached zero by linearly combining the contour estimation of the NN with the AFM, as defined in Formula 4.1. In this formulation, V represents the head angular velocity, and V_{thres} is a velocity threshold above which no linear combination is performed. A threshold value of $V_{thres} = 0.05$ was selected based on visual estimation. The parameter t ranges from 0.0 to 1.0, where $t = 0.0$ corresponds to displaying the AFM contour, and $t = 1.0$ corresponds to selecting the pure NN contour.

$$Contour_{V < V_{thres}} = (1 - t) * Contour_{AFM} + t * Contour_{Neural} \quad (4.1)$$

4.2 Results

4.2.1 Data Exploration

To contextualise the evaluation of the autoencoder and Gaussian models, the overall dataset distribution of head angular velocity magnitude and direction was analysed. The histogram of data samples relative to velocity magnitude indicates that most of the data are concentrated at low-velocity magnitudes, where the head remains stationary (Figure 4.5a). Additionally, the data reveals that gaze samples corresponding to head movement directions are predominantly found in horizontal directions (Figure 4.4a). The standard deviation (STD) of the eye-gaze distributions provides a measure of sample spread. By calculating the STD on data filtered by head angular velocity direction and magnitude, variations in the size and shape of PDFs and contours as a function of head rotational velocity can be observed. When analysing STD variations in relation to head direction (Figure 4.4b), it is evident that the X-axis STD reaches its maximum when the head is rotating towards the left ($\theta = 90^\circ$) or right ($\theta = -90^\circ$), while the Y-axis STD reaches its maximum when the head rotates upward ($\theta = 0^\circ$) or downward ($\theta = 180^\circ$ or -180°). These results indicate a correlation between STD and head rotation direction as, when the head rotation is upward or downward, the Y-axis STD increases while the X-axis STD decreases and, conversely, when the head rotation is left or right, the X-axis STD increases while the Y-axis STD decreases. Such a change in STD is confirmed by the shape of the contours generated by the models (Figure 4.2d). The contour size changes on the X and Y axes according to X STD and Y STD. Furthermore, the STD increases as the head velocity magnitude increases. Such a result is consistent with the size of contours, which increase with increments in head velocity magnitude (ρ), as shown in Figure 4.2e.

4.2.2 Performance

The temporal performance of the model was tested on two untethered devices. The average inference time was measured to be 8 ms on the Oculus Quest and 3 ms on the Oculus Quest

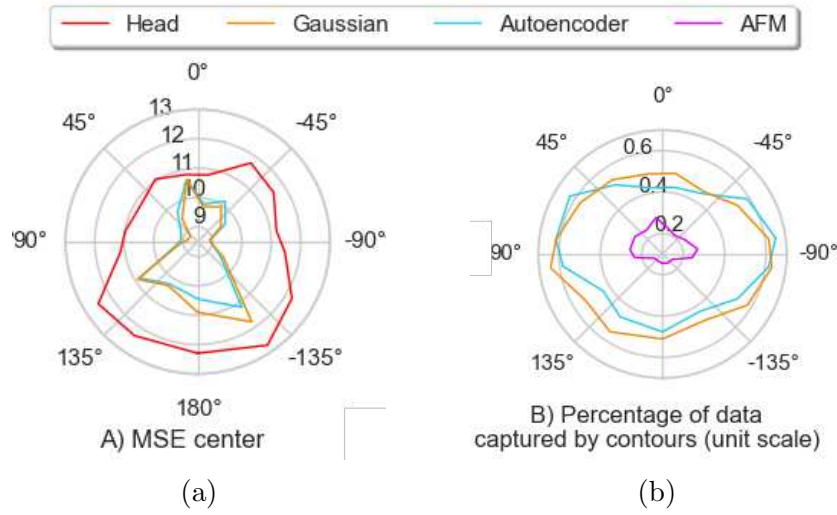


Figure 4.6: Results of the Autoencoder, Gaussian, and head direction baseline. A) RMSE of the predicted centre relative to the distributions of eye-gaze samples for each head direction. B) Percentage of data captured by the model contours for each head rotation direction.

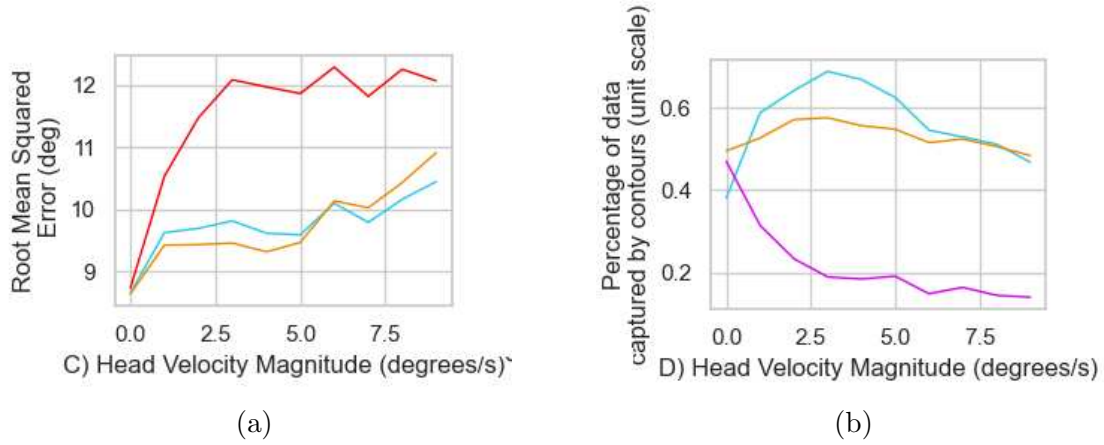


Figure 4.7: Results of the Autoencoder, Gaussian, and head direction baseline. A) RMSE of the models relative to Head Angular Velocity (degrees/s). B) The percentage of the data captured by the models' 70th percentile contours relative to the head's angular velocity magnitude (degrees/s).

2. These results align with the Oculus refresh rate, indicating that the model operates in real time. To evaluate the model, a comparison was conducted against the identified baselines: head direction and the Average Fixation Map, using the recorded eye-tracking data as ground truth. We compare the model's inferred gaze direction to this ground-truth data using two metrics. First, the Root Mean Squared Error (RMSE) is used to quantify the error in the predicted centre relative to the eye-gaze distributions. Second, we measure the proportion of eye-tracking samples contained within the predicted probability contours on a unit scale.

4.2.2.1 RMSE

The performance of the models related to inferred gaze position was evaluated by calculating the RMSE between the inferred gaze and the real gaze (Figure 4.6a and Figure 4.7a). A comparison of the RMSE results indicates that the autoencoder model outperforms the Gaussian model for $\theta = [-90^\circ, 90^\circ]$ (Figure 4.6a) and for $\rho = [0^\circ/s, 6^\circ/s]$ (Figure 4.7a). Both models consistently demonstrate superior performance compared to the baseline (i.e., head direction). However, all models, including the baseline, exhibit the lowest accuracy during downward head movements (Figure 4.6a). This result aligns with previous studies that describe how the eyes contribute more significantly than the head during downward gaze shifts [47]. As eye contribution increases, so the accuracy of both models and the baseline decreases during downward movements. Furthermore, the models achieve their lowest accuracy during head movements to the left or right, indicating that they are most effective during lateral head movements. This observation aligns with findings from Hu et al. [51], which explored the Pearson's Correlation Coefficient (PCC) of the X and Y components of head rotational velocity and reported a stronger correlation between the X component and gaze location as compared to the Y component. Model performance to the head velocity magnitude (ρ) (Figure 4.7a) shows that, when ρ is close to zero, the accuracy of both models is equivalent to the baseline. On the other hand, the models perform significantly better than the baseline across most of the range of ρ . Due to the limited number of samples with a head velocity above $10^\circ/s$, it is not possible to reliably estimate how RMSE evolves beyond these velocities.

4.2.2.2 Percentage of data captured

The performance of the contours in capturing the eye-gaze samples was also evaluated. For this analysis, the 70% percentile contour was used (Figure 4.6b and Figure 4.7b). This metric allows for a comparison between the models and the baseline, the AFM (Figure 4.2f). The results indicate that both models outperform the AFM, with the best performance corresponding to the left/right directions, similar to the MSE of the centre. A comparison with the AFM further reveals that, when head rotational velocity is close to zero, the AFM performs equivalently to

the models (Figure 4.7b), similar to the MSE centre baseline. However, when head movement occurs, both models demonstrate a significantly better performance when compared to the baseline. This characteristic of the AFM is leveraged to stabilise the model’s output when zero velocity magnitude is detected. A comparison of the percentage of data captured by the models shows that the autoencoder model outperforms the Gaussian model for $\theta = [-90^\circ, -45^\circ]$, $[45^\circ, 80^\circ]$ (Figure 4.6b) and for $\rho = [1^\circ/s, 6.5^\circ/s]$ (Figure 4.7b). These results suggest that the proposed approach to generating PDFs using an autoencoder better captures the shape of gaze distribution in relevant scenarios, such as left/right head movements.

4.3 Discussion

During collaboration, visual attention cues contribute to clarifying the shared visual context, enabling participants to perceive each other’s focus. Mutual awareness of visual attention facilitates communication by reducing the necessity for verbally explicit references or hand-pointing gestures to identify and negotiate the current direction of the collaborative effort. This chapter introduces a lightweight deep-learning model capable of predicting the PDF of gaze in VR based on head motion. The model is founded on a novel method for generating PDFs, which enhances interaction in collaborative VR by explicitly depicting the collaborator’s likely area of visual focus in real time. Two models for inferring gaze and its PDF, as a percentile contour, were developed and compared. The results of the RMSE of the predicted eye gaze indicate that both models outperform the AFM baseline. When compared to each other, the models demonstrate similar accuracy. The comparable performance of both models suggests that the autoencoder can serve as an effective alternative to the standard Gaussian 2D approach. Furthermore, the evaluation demonstrates that the multi-perception model can be utilised in real-time applications on untethered devices. The model and source code have been made available to the research and VR developer communities via GitHub¹. The implementation achieves real-time performance on a low-cost Oculus Quest with an inference time of 7 ms, indicating that this approach can serve as an alternative to head orientation for approximating

¹https://github.com/Collaborative-Immersive-Visual-Toolkit/VR_Iso_Gaze_Contours

gaze in collaborative VR applications. Moreover, the same technique could be used in post-process analysis scenarios as a fast method for estimating gaze from 3 DoF head data. In addition, this paper introduces a novel type of coordinates for head rotational speed with the characteristics of being human-readable and that can be used to analyse gaze distributions in the function of the head rotational velocity. Further work could be carried out to compare our model with the AFM baseline via psychophysical experimentation and extend the model with different dataset types. For interactive applications, the model could be tested in a pointing task [197]. The same comparison, instead of using the models' contours, could be carried out for collaborative applications performing mixed-methods studies such as [41, 97] comprehensive of quantitative and qualitative measures.

Chapter 5

Speech-Augmented Cone-of-Vision

Use of Collaborative Verbal Communication to Augment Visual Attention

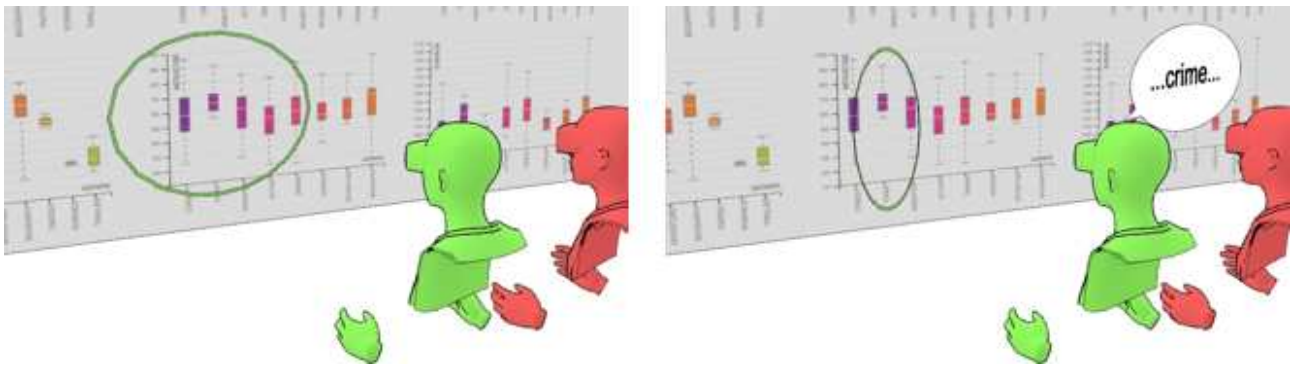


Figure 5.1: CoV+Speech is a multi-modal visual attention cue that narrows the cone of visual attention around keywords uttered by participants during collaborative verbal communication.

The Cone of Vision (CoV) was introduced in Chapter 3 as a technique to enhance visual attention inference in VR. While CoV significantly improved mutual awareness of visual focus and collaboration, our findings revealed a key limitation as, in certain cases, the inferred gaze region lacked specificity (Section 3.2.2.2). This limitation highlights the need for a more precise, real-time gaze inference approach – one that reduces ambiguity. Notably, when this limitation arises, verbal communication often serves as a compensatory mechanism. This observation aligns with previous research, which indicates that collaborative verbal communication frequently serves as a fallback when visual cues are otherwise inadequate or when calibration

errors arise [37]. Indeed, prior studies emphasise the crucial role of verbal communication in aligning and coordinating visual attention during collaboration, both in-person and in remote immersive environments (Section 2.2.7). While the interplay between verbal communication and visual attention has been extensively studied, it has yet to be actively leveraged to enhance collaborative communication.

This chapter explores the use of head orientation and speech to enhance visual attention representations, aiming to improve the accuracy of head-based visualisations in collaborative VR settings (Section 2.5.4). A novel attention visualisation method, the Speech-Augmented Cone-of-Vision (CoV+Speech), is introduced, refining the representation of visual focus by integrating head orientation with speech-based references that naturally occur during collaborative analysis (Section 2.2.7). As the system is aware of the context (i.e., has access to the underlying HTML of the data visualisation), it can adjust its focus in real time to highlight referents mentioned by the user, thereby refining the representation of visual attention.

The empirical study evaluating the Speech-Augmented Cone-of-Vision demonstrated that integrating speech with head direction significantly improves eye-gaze inference (Figure 5.8a). These findings provide valuable insights into how Speech-Augmented Cone-of-Vision enhances collaborative dynamics, offering guidance for designing more effective attention-sharing mechanisms in immersive environments (Section 3.3.2). This chapter introduces the fundamental concept of leveraging naturally occurring verbal communication as an input modality for XR systems, moving beyond explicit commands to a more seamless, context-aware interaction paradigm.

5.1 Study

A within-group study was designed to compare three conditions: CoV, CoV+Speech, and Eye-Gaze Cursor (Figure 5.2). Participants were embodied in an avatar using ReadyPlayerMe¹ and could also use a hand pointer to reference the observed dataset. A Latin Square approach was

¹<https://readyplayer.me/>

used to counterbalance the three conditions. However, due to the number of participants, one ordering had an additional pair of participants².

5.1.1 Conditions

5.1.1.1 Cone of Vision (CoV)

The CoV was introduced in Section 3.1.2.3.

5.1.1.2 Speech-Augmented Cone-of-Vision (CoV+Speech)

The second visual cue is a combination of CoV and the effects of the user's verbal interaction with the system. Although the same CoV calculation as in the CoV condition is utilised, a novel algorithm is designed to modify the CoV after processing the user's speech. As described in Section 5.1.2.1, the dataset is contained in the HTML pages rendered by the virtual screens in the 3D office. To extract the semantics of the speech, the user's audio is first captured during communication with their collaborator. This audio is then streamed to an online speech service that transcribes the speech and returns a string to parse and process with NLP algorithms. The extracted information is subsequently searched in the HTML context to identify those elements containing keywords isolated by NLP. The bounding box coordinates of these elements within the browser page are retrieved, converted into world coordinates, and added to a list. The next phase involves modifying the current CoV. To reduce the CoV size, the principal component of the coordinates is determined through linear regression. The standard deviation of the points is then calculated along the principal component and its orthogonal direction. An ellipse is drawn using the coordinates of the centre of mass, with the two standard deviations serving as its radii. Finally, interpolation between the CoV points and the ellipse points is performed with a factor of 0.5 (Figure 5.4 (d)). The visual cue is displayed as shown in Figure 5.4 (e). This condition integrates two different input channels: head-based position and orientation derived

²Additional statistical analyses were performed to ensure that no ordering or learning effects influenced the results, as detailed in Section 5.2.

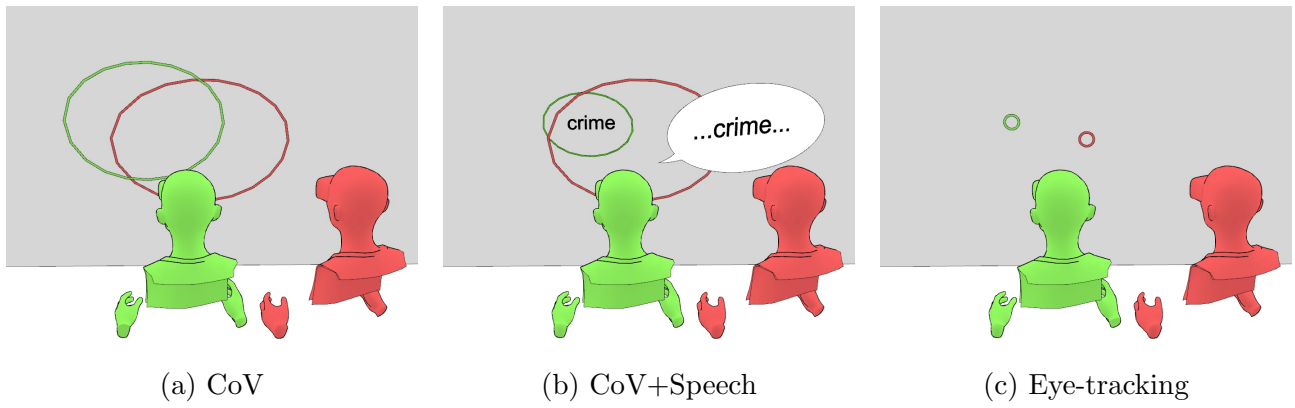


Figure 5.2: Experiment One Conditions: (a) CoV introduced in Chapter 3; (b) CoV+Speech: when the participant enunciates a word which is displayed on the VR screen, the application search within the HTML context and if the word is within the CoV, it shrinks the cone to around the bounding box of the matching word; (c) Eye-tracking: a visual cursor is displayed for each participant in the raw gaze location of the eye-gaze.

from the HMD hardware, and an analysis that converts speech into a morphing function of the CoV. While the CoV calculation is executed through internal code, a portion of the speech processing relies on an external service.

5.1.1.3 Eye-Gaze Cursor

The Eye-Gaze Cursor condition displays a graphical cursor at the gaze location on the VR screen. The position is calculated by calculating the intersection between the gaze direction and the VR screen. The cursor is visualised as a ring (Figure 5.2(b)) and has a radius of $\frac{1}{3}$ of the radius of the foveal region as determined by pilot testing to ensure that the cursor is noticeable. The Eye-Gaze Cursor is subjected to noise from the eye-tracker.

5.1.2 Apparatus

For this study, each participant was provided with a PicoNeo 2 HMD featuring a 4K resolution (3840×2160) at a refresh rate of 75Hz. The HMD includes an embedded Tobii eye-tracker operating at 90Hz with a declared accuracy of 0.5 degrees. For verbal communication, a Microsoft Teams connection was established using Bluetooth headphones and a microphone for participant interaction, while the PicoNeo 2 microphone captured the audio stream for transcription.

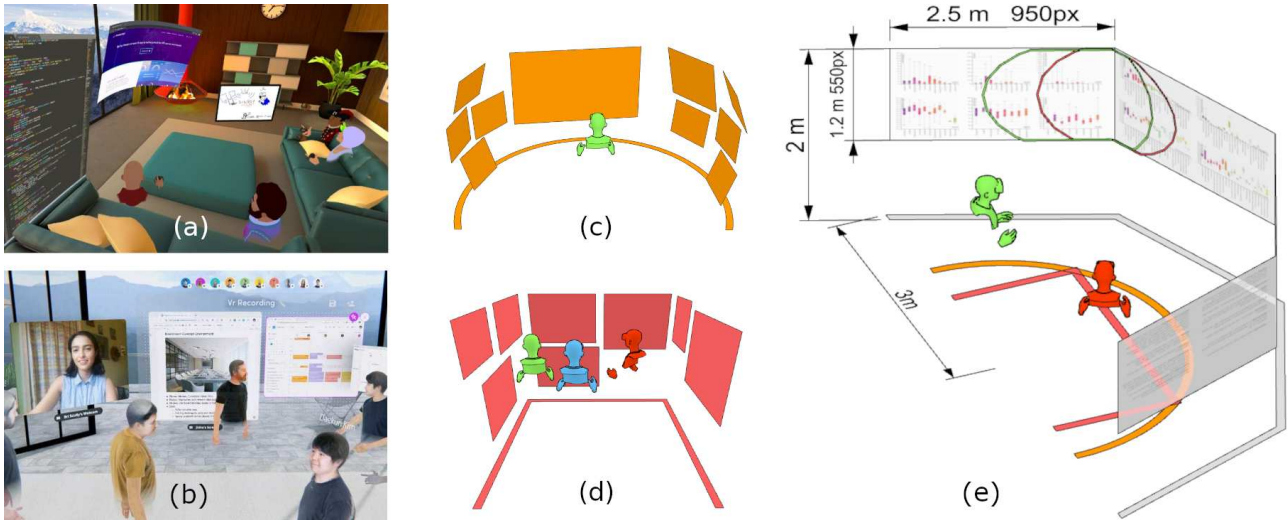


Figure 5.3: This Figure shows on the left two collaborative environments: (a) Immersed [198], and (b) Spatial.io [76], which enable users to generate their configurations of 2D VR screens layouts. Figure (c) highlights a circular VR scenario in cross-virtual analytics studies of Satriadi et al. [34] and (d) a squared configuration from the work of Lee et al. [66]. Our immersive collaborative virtual environment is depicted in (e). It consists of 4 VR screens with HTML web pages displayed in a cylindrical layout.

A collaborative VR application was designed and implemented using Unity 2020.3.34.f1, enabling two users to share the same digital space while remaining physically separate. The application supports three different methods for exchanging visual cues during the exploratory data analysis task. Each user's visual cues were displayed using distinct colours: red for the local visual cue and green for the remote visual cue. All cues were refreshed in each Unity loop at a fixed rate of 50Hz.

5.1.2.1 VR Environment

The same VR environment described in Chapter 3 is used. However, modifications were made to the layout to better suit the current study. Specifically, instead of a cylindrical arrangement, the environment was updated to a semi-cylindrical configuration to optimise participant visibility and align with the commercial version of VR offices, as described in Section 2.1.2. This adjustment maintains the egocentric convex layout inspired by prior work [34, 66], while accommodating the specific requirements of the present study. Unlike Chapter 3's study, participants remained in fixed locations to prevent obstruction of each other's views. ReadyPlayerMe avatars were used, retaining torso representation along with lip and eye synchronisation. The

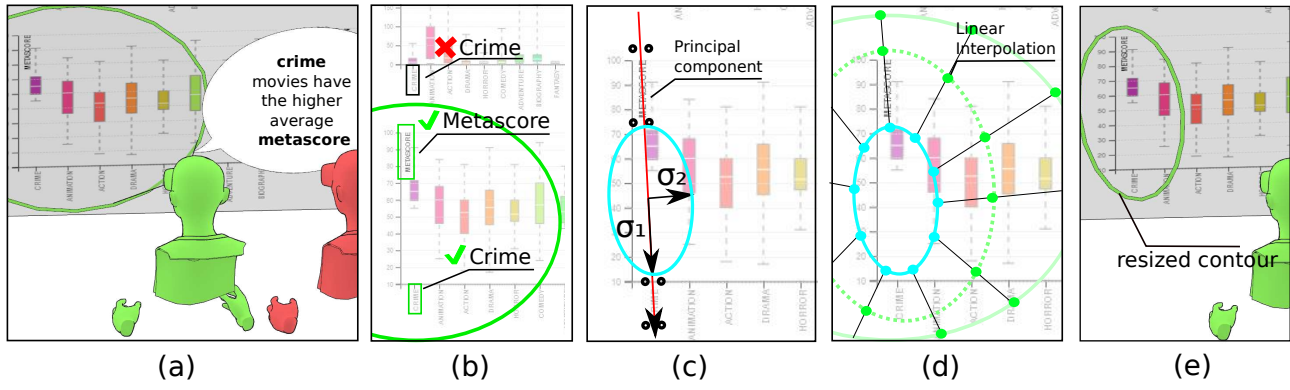


Figure 5.4: This sequence of figures illustrates the algorithm’s steps for generating the visual cue from the combined action of CoV and speech inputs. Initially, (a) the CoV is displayed on the VR screen while the user speaks to the collaborator. After speech recognition, (b) keywords present on the VR screen are identified and filtered based on their positions. Only keywords within the CoV are retained, while others are discarded. An ellipse is then fitted to the distribution of the selected keywords, with its axes corresponding to the standard deviation of their X and Y components (c). In (d), the CoV is interpolated with the fitted ellipse using a balanced ratio of 0.5. Finally, (e) the resulting visual cue is presented to the users.

four VR screens continued to display charts rendered via an embedded internal browser, decoding datasets from HTML/JavaScript files. The datasets and visualisations remained the same as those described in Section 3.1.2.2. Additionally, keyword positions within the HTML were extracted and utilised in the CoV+Speech condition.

Speech-to-Text Real-time captioning services provide transcriptions of spoken information from audio streams. Google Speech-To-Text[137], Microsoft Cognitive Services[138], Dialogflow [139], IBM Watson[140], and Amazon Transcribe [141] are among the most widely used services that allow real-time API transcription integration into a system. Google Speech-to-Text was selected due to its compatibility with the Android platform and the Unity framework. The Google Speech-to-Text service can be configured with various parameters, including language, sample rate, automatic punctuation, and context adaptation. Several pilot tests were conducted to optimise accuracy and reduce latency. The audio sample rate was set to 16,000 Hz, the language was configured as British English, and only native British speakers were included to ensure optimal transcription quality. Punctuation was filtered out to streamline the transcription process.

5.1.3 Participants

A total of 20 participants were recruited over two weeks (13 women, 7 men, $M_{\text{Age}} = 29.4$, $SD_{\text{Age}} = 9.1$) through an online platform managed by University College London³. Several inclusion criteria were applied during screening: participants had to be native English speakers, have normal vision, and possess at least a high school education. The latter criterion ensured sufficient proficiency in interpreting the graphs used in the visualisation. Additionally, participants were required to be confident in interpreting the charts included in the study, which consisted of bar or candlestick plots, histograms, and scatter plots. A self-assessment questionnaire was used to verify participants' familiarity with these visualisations, and a summary of their characteristics was provided during the task presentation. One participant identified as an expert VR user, five reported average experience, six as occasional users, eleven with low experience, and four had no prior experience with VR. Compensation of £15 was provided for each participant for a 90-minute study. To incentivise optimal performance, an additional reward of £15 was introduced for the pair that reported the highest number of valid insights. Participants were encouraged to collaborate rather than split their attention across different visualisation areas.

5.1.4 Procedure

Upon arrival, participants read the information sheet and signed the consent form. The experiment was conducted in a laboratory setting, with each participant assigned to a separate office. The task duration and the three experimental conditions were introduced, and participants were given one minute to test each condition. Following this, participants performed an exploratory data analysis task, extracting insights from the displayed visualisations. Examples of valid insights were provided. A valid insight was defined as a recorded speech conveying a precise and deep understanding of two or more measures displayed on a graph or a series of graphs [180]. Instructions were given on how to record insights and use the hand pointer. Once instructions were understood, participants wore the Pico Neo Eye 2, completed the eye-tracking calibration process, connected to the virtual environment, and started the collaborative task.

³<https://uclpsychology.sona-systems.com/>

Category	Description
(1) Keyword	When a participant explicitly references an element by naming a specific word or name displayed in the data visualisation. For example, “I would say that a movie’s budget is ...”
(2) Sequential	When a participant references an element by naming a number representing the element, for example, “shall we move to the third panel” or “look at the second element”. An alternative expression might consist of the participant suggesting moving the focus of the collaboration to the next or previous graph/element/page, for example: “the one next to it”.
(3) Color	When a participant references an element by naming its colour: “...the green one..”
(4) Context reference	When a participant references an element based on its location within the page, for example, saying “...top left corner...” or “...the graph below...”. Such references can also be related to data on data, for example: “but the actual value is much lower”.
(5) Pointer	When a participant performs an implicit verbal reference by using the laser pointer to highlight an element directly and utter words like “...this...”, “...over here...”, “... the graph we are looking at...” or directly mentioning “...where I’m pointing...”
(6) User relative	When a participant references an element in relation to the frame of reference of the other user, such as: on your right/left/side, close to/far from you, above/below you ... for example, “...the graph on your right...”
(7) Temporal	When a participant references an element previously envisioned, such as before, after, or earlier... for example, “Let me check if I can see something else from the previous one.”
(8) Visual cue	When a participant uses a deictic reference such as this, that or here “...this graph over here...” or when a participant directly refers to the gaze visualisation, for example, “Right where I am looking.”

Table 5.1: Verbal communication taxonomy.

After completing all VR trials (i.e., experimental conditions), participants filled out a questionnaire (Section 5.2.3). At the end of the experiment, semi-structured interviews were conducted individually. Participants reported cases where each experimental condition facilitated the task and cases where it did not. The study lasted between 75 and 90 minutes ($M = 80$, $SD = 12.7$), with each trial lasting approximately 10 minutes ($M = 13$, $SD = 3$). The stop condition was reached when the allocated time of 13 minutes expired.

5.1.5 Offline Analysis Methods

To examine the role of verbal communication in negotiating shared visual attention, recorded audio was transcribed to ensure high-accuracy transcriptions (Section 5.1.2.1). The transcribed

data was analysed to quantify the frequency with which participants referenced displayed keywords and the extent to which alternative referencing methods were used (Section 5.1.5.2). Additionally, an evaluation was conducted to determine whether participants' utterances could be leveraged alongside head direction to refine eye-gaze inference (Section 5.1.5.3).

5.1.5.1 Speech-to-Text

An offline speech recognition system was employed to analyse the audio recordings with higher accuracy than the real-time system used in the study. This framework, released in the second part of September 2022, is the open-source project Whisper [199], developed by OpenAI. The system is trained on extensive multilingual spoken language datasets and utilises an end-to-end encoder-decoder transformer architecture [200], producing highly accurate text captions. Whisper was implemented using Python 3.8.3 and PyTorch 1.12.3 [201]. The manual analysis described in Section 5.1.5.2 ensured the quality of the transcription.

5.1.5.2 Classification of Verbal References

To quantify the extent to which participants referenced displayed keywords verbally and how frequently alternative methods such as pointing gestures or implicit visual cue references were employed, a classification framework was established. Verbal reference taxonomies from D'Angelo and Begel [12] (i.e., remote pair programming; Table 5.1 (1,3,8)) and Pettersson et al. [13] (i.e., collaboration over tabletop map visualisations; Table 5.1 (3,5)) were merged to accommodate both textual and visual element references within the same context. The taxonomy was expanded to incorporate novel verbal references, including sequential statements (Table 5.1 (2)), which rely on implicit left-to-right (LTR) directional bias [202]. Additionally, user-relative references (Table 5.1 (4,6)) and temporal references (Table 5.1 (7)) were introduced to account for contextual variations in referencing behaviour. The transcripts were analysed alongside video and audio recordings to contextualise non-verbal communication (e.g., mutual alignment or opposite orientation of participants, pointing gestures with a laser pointer, etc.). Three coders conducted the analysis wherein each transcribed trial was initially reviewed by one coder, sub-

sequently validated by a second, and any discrepancies were resolved by a third coder. The roles of the coders rotated with each trial. For each transcribed sentence, verbal elements aimed at identifying or shifting the focus of the collaborative exploratory data analysis task in relation to the visualised data were identified. Sentences containing visual context negotiation were classified according to the predefined taxonomy, and the corresponding areas of interest (i.e., data, chart, page) were annotated. Following classification, the number of occurrences of each verbal reference type was quantified per experimental condition. A stark difference emerged between the "Keyword" class and all other categories, with keyword-based references occurring more frequently than the combined sum of all other reference types.

5.1.5.3 Head+Speech Gaze Inference

The refinement of eye-gaze inference using head direction and verbal utterances was evaluated. Data segments in which verbal communication was utilised were specifically considered to ensure a fair analysis. As illustrated in Figure 5.8b, the steps employed to compute the accuracy of the Head/Gaze and Head/Gaze+Speech methods in relation to the ground truth (i.e., eye-gaze data) are detailed. First, the model focuses on bi-grams, capturing the last two spoken words by the user at any given time. This number of words was determined to be optimal among other n-grams. Second, the Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) Lin [203], a well-established text similarity metric for longest common subsequences, was applied across all possible keywords located inside the CoV. The similarity scores range from 0 to 1, and only positive scores were retained. Bounding boxes of the accepted keywords were then identified, and the closest box was determined by calculating its Euclidean distance from the Head/Gaze position. Finally, the RMSE was computed for both Head-Gaze and Head/Gaze+Speech methods against the eye-gaze data.

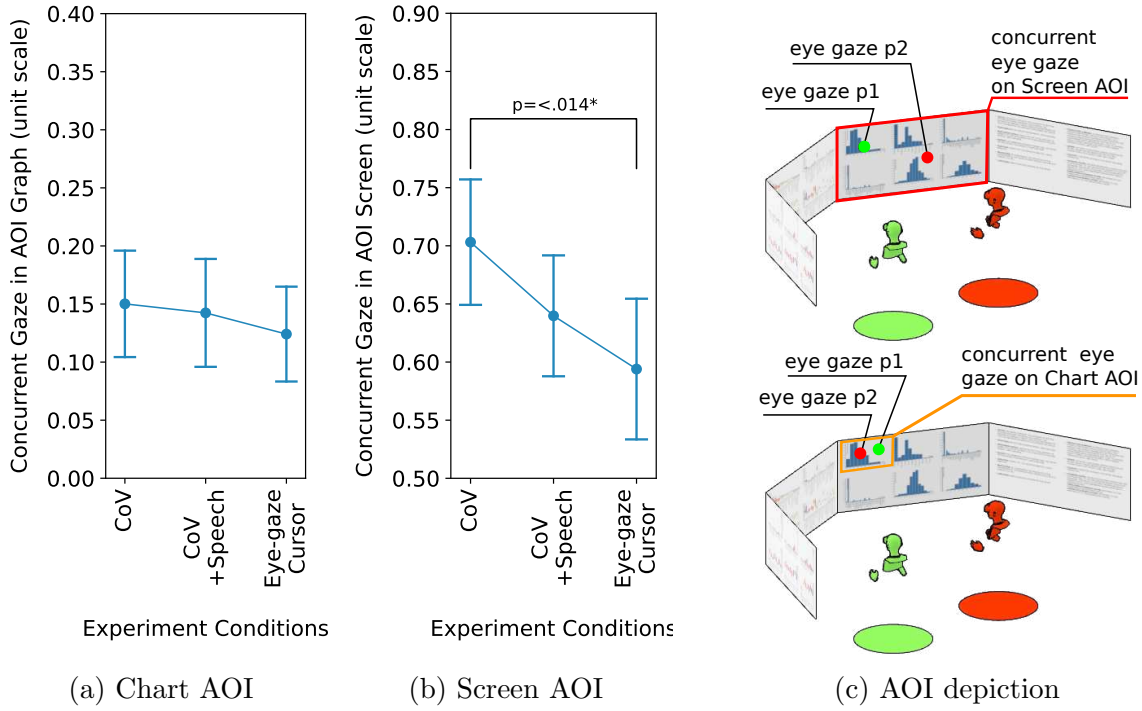


Figure 5.5: (a) The y-axis represents the percentage of time, in unit scale, that participants spent concurrently looking at the same graph. (b) The y-axis represents the percentage of time participants spent concurrently looking at the same page. (a)(b) The x-axis of both graphs denotes the experimental conditions. The error bars indicate a 95% confidence interval. (c) The granularity of joint attention measurement is illustrated. The top of (c) represents screen-level granularity, capturing instances where collaborators focus on the same screen. The bottom of (c) represents chart-level granularity, capturing instances where collaborators focus on the same chart.

5.2 Results

The results are structured into three sections. The first Section 5.2.1 presents the analysis of how different visual cues influenced concurrent (Section 5.2.1.1) and individual (Section 5.2.1.2) visual attention. The second Section 5.2.2 examines how participants employed verbal communication to negotiate the shared context of visual attention (Section 5.2.2.1) and evaluates the effectiveness of speech as an input to infer gaze during collaboration (Section 5.2.2.5). The third Section 5.2.3 assesses participants' experiences with different visual attention cues. Unless otherwise stated, the analysis was conducted using a one-way repeated measures (RM) ANOVA ($\alpha=.05$) with Condition (*COV*, *Eye-GazeCursor*, *CoV+Speech*) as an independent variable. When the assumption of sphericity was violated, as tested with Mauchly's test, Greenhouse-Geisser corrected values were used. QQ-plots were utilised to validate the assump-

tion of normality, and Holm-corrected post-hoc tests were applied when applicable.

As mentioned in Section 5.1, the Latin square was not fully balanced. Three participant pairs started with the CoV condition, three with the CoV+Speech condition, and four with the Eye-Gaze Cursor condition. To assess potential biases and confounding factors, a statistical analysis was conducted to examine ordering and learning effects. An ANOVA analysis was performed on joint attention ("concurrent AOI") and individual visual attention ("gaze on own cone") to detect any ordering effects, but no significant impact was found. Additionally, an ANOVA analysis of the number of insights generated across conditions showed no discernible learning effect. This indicates that the presentation order did not significantly influence participants' behaviour or performance, supporting the effectiveness of the approach in enhancing collaboration and communication in data analysis. All statistical analyses are included in the supplementary materials.

5.2.1 Visual Attention

The recorded visual attention data was segmented by dividing it for each reported insight. Participants spent, on average, 127.24s (SD=63.57s) on each reported insight, engaging in discussion, exploration, or examining relevant objects before articulating their observations. Concurrent (Section 5.2.1.1) and individual (Section 5.2.1.2) visual attention behaviours were averaged for each insight segment. Participants reported 179 insights. Each pair reported a mean number of 17.6 insights (SD=5.25). Statistical analysis did not reveal any significant differences in the number of insights. The duration of each insight was analysed by segmenting each trial into the time between consecutive insights. No statistical difference was found between the conditions.

5.2.1.1 Concurrent Visual Attention in Areas of Interest

The amount of concurrent visual attention within pairs was measured using a semantic segmentation of the visualisation. This was achieved by calculating the percentage of time participants

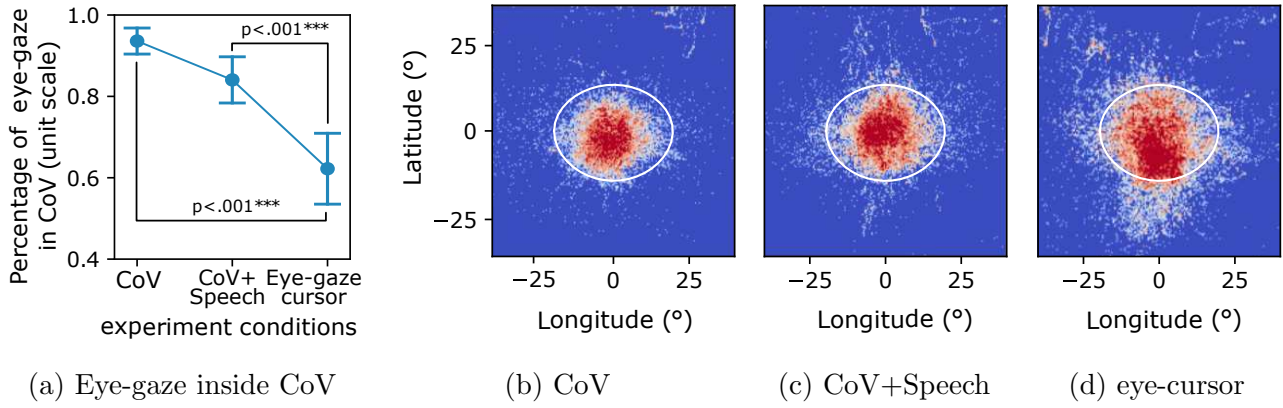


Figure 5.6: (a) Eye-gaze within the projected cone: The y-axis presents the percentage of time in unit scale that the eye-gaze was within the projected cone of vision. The x-axis presents the experimental conditions. (b), (c), and (d) Gaze sample distributions in head angular coordinates of the experimental condition of (b) CoV, (c) the CoV+speech, and (d) eye-cursor, respectively. On the x-axis, the longitude is presented in degrees while, on the y-axis, the latitude is presented in degrees. In (a),(b), and (c), the CoV is depicted in white.

spent concurrently looking at the same AOI for each insight. Two AOI levels were considered: *screens* and *charts*. The screen AOI was defined as a full page, 981px wide and 551px high (Figure 5.5c top). The chart AOI was defined as an individual chart, varying between 300px and 500px wide and 250px high (Figure 5.5c bottom). For the screen AOI level, the RM ANOVA analysis revealed a statistically significant difference ($F(2, 116)=4.191, p=.017$). Post-hoc comparisons identified a significant difference between CoV ($M = 0.703, SD = 0.211$) and Eye-tracking ($M = 0.583, SD = 0.238$) conditions ($t = 2.895, p = .014$). The effect size test ($Cohen'sd = 0.377$) indicated a small to medium-sized effect. The results suggest that the presence of the CoV led to an approximately 20% increase in the time participants spent concurrently looking at the same screen. For the chart AOI level, ANOVA analysis showed no significance. Participants spent, on average, 13% of each insight looking at the same graph, with no differences observed across experimental conditions.

5.2.1.2 Individual Visual Attention Behaviour

The effect of experimental conditions on participants' eye-gaze behaviour was measured by calculating the percentage of gaze samples within the visual cone (Figure 5.6a). This measure provides insight into the extent to which gaze diverges from the head's direction. RM

ANOVA revealed a significant difference ($F(1, 106)=29.007, p < .001$). Post-hoc comparisons showed significant differences between the CoV ($M = 0.929, SD = 0.151$) and Eye-tracking ($M = 0.541, SD = 0.383$) conditions ($t = 7.337, p < .001$), as well as between the CoV+Speech ($M = 0.829, SD = 0.232$) and Eye-tracking conditions ($t = 5.441, p < 0.001$). All significant pairwise differences exhibited Cohen's $d > 0.8$. These results indicate that, when the CoV's cone intersection is visualised, the participants' gaze is significantly less likely to fall outside the depicted region, suggesting that the visualisation acts as a container. This effect is also observed in the CoV+Speech condition, where the region dynamically adjusts based on voice input. To further examine the impact of these conditions on gaze behaviour, gaze distributions were plotted for each condition (Figure 5.6b, Figure 5.6c, Figure 5.6d). A comparison of these distributions reveals two key insights. First, gaze distribution in the Eye-tracking condition is more dispersed than in other conditions. Second, the gaze distribution in the Eye-tracking condition exhibits a greater downward spread, indicating that users tended to move their gaze further downward from the head direction relative to other directions, as previously reported [192]. This suggests that alternative conditions may encourage more head movement, as gaze remains within a more constrained area.

5.2.2 Speech and Visual Attention

5.2.2.1 Classification of Verbal References

A classification was conducted to quantify the types (Section 5.2.2.2) and targets (Section 5.2.2.3) of verbal references used in negotiating the shared context of visual attention. This classification facilitates the calculation of the frequency of different verbal references (Table 5.1) during collaboration. Additionally, it provides insight into which targets require more frequent referencing and whether the frequency of verbal reference types varies based on the target (Section 5.2.2.3).

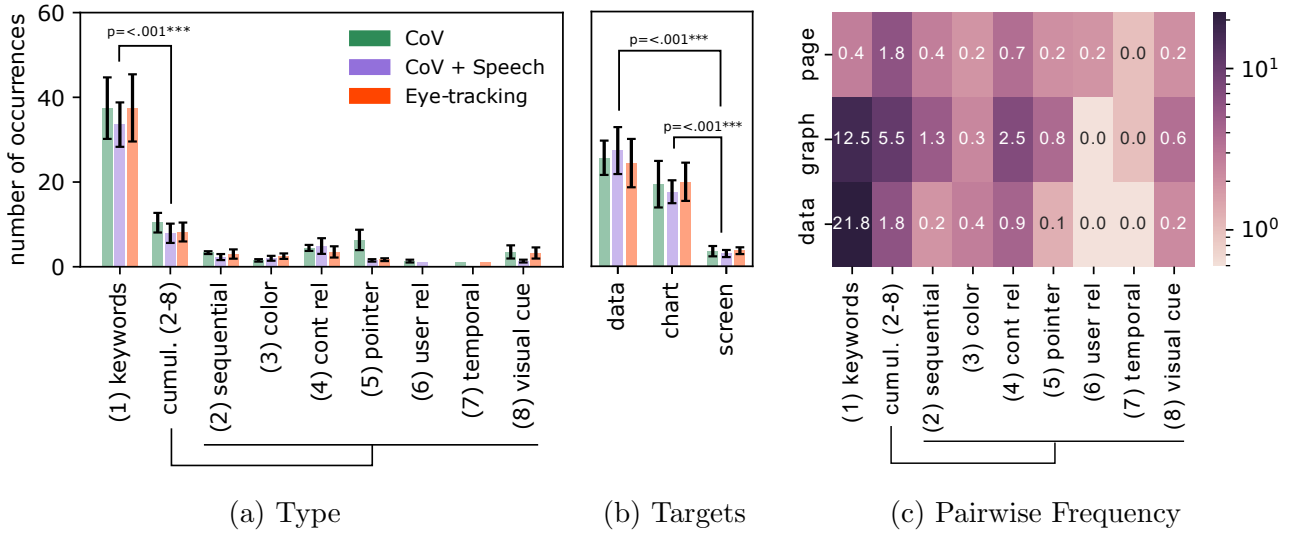


Figure 5.7: Analysis of verbal references to the dataset. (a) Bar charts representing the mean count of verbal reference occurrence per trial grouped by type (types description can be seen in Table 5.1). On the x-axis, the classes representing the different types of verbal communication reference the visualised data, while the y-axis charts the number of occurrences for each pair of participants during a single 15 min trial. Error bars represent unbiased standard error of the mean, normalised by $N-1$. (b) Bar charts representing the mean count of verbal reference occurrence per trial grouped by a target area. On the x-axis, the classes representing the different AOI targets of the verbal communication reference the visualised data, while the y-axis charts the number of occurrences for each pair of participants during a single 15 min trial. Error bars represent unbiased standard error of the mean, normalised by $N-1$. (c) Verbal communication pairwise (type/AOI) frequencies, with each cell showing the mean count per trial of each verbal reference combination type/AOI. The colour bar shows a logarithmic scale palette.

5.2.2.2 Verbal References Types

An RM ANOVA analysis was conducted to determine whether there were significant differences in the type of verbal reference ("(1) keywords" and "cumulative (2-8)") under various experimental conditions (Figure 5.7a). There was a significant difference observed in the number of verbal references consisting of participants enunciating visualisation keywords relative to the cumulative sum of all other types of verbal references ($F(1, 8)=50.190, p_i.001$). These results highlight that our proposed approach, which intersects meanings extracted from verbal communication with keywords on the VR display, improves visual attention inference. RM ANOVA analysis showed no significant differences between the experimental conditions ($F(2, 16)=0.211, p=0.812$) nor any interactions between the reference type and the experimental conditions ($F(2, 16)=0.013, p=.987$). These results imply that participants did not change

their verbal communication depending on the type of visual attention cue, and further suggest that using verbal communication as an input does not impact the verbal behaviour of users.

5.2.2.3 Verbal Reference Targets

An RM ANOVA was conducted to examine the effect of the given experimental condition on the target AOI of verbal references (Figure 5.7b), revealing a significant difference ($F(2, 14)=30.368, p<.001$). Post hoc tests showed a significant difference ($p < .001$) between the data ($M = 25.333, SD = 11.313$) and screen ($M = 2.125, SD = 2.853$) targets, as well as a significant difference ($p < .001$) between chart ($M = 20.875, SD = 15.376$) and screen. These differences showed that participants equally negotiated visual attention via verbal communication for both the "chart" and "data" levels while, for the "screen" level, such negotiation is less necessary. The analysis did not show any main effect on experimental conditions ($F(2, 14)=0.211, p=.934$), nor any interaction between the target factors and the experimental conditions ($F(4, 14)=0.211, p=0.696$) meaning that participants did not change the way they verbally communicated depending on the type of visual attention cue displayed. These results imply that the experimental conditions did not alter verbal communication.

5.2.2.4 Pairwise Comparison of Types and Targets of Verbal References

The relationship between the target AOI and the verbal reference method was further examined by generating a pairwise frequency matrix (Figure 5.7c) to characterise the references in greater detail. For example, it is clear that, when participants refer to the chart AOI, they tend to use a larger array of methods which can be visualised by comparing the chart row and data row. This effect is even stronger for the page AOI where (1) keywords are no longer the most frequent method, as shown in data and chart AOIs.

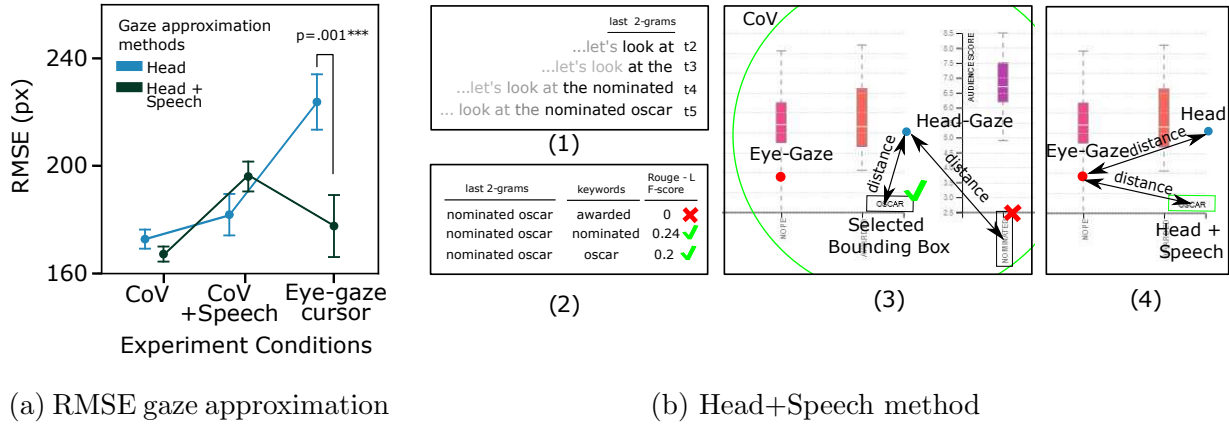


Figure 5.8: (a) Comparison between the RMSE of head-gaze and head-gaze + speech using eye-gaze as ground truth. The y-axis represents RMSE in pixels. (b) (1) A queue is created containing the last two spoken words (2-grams). (2) The ROUGE-L metric [203] is applied to each keyword within the CoV, yielding a similarity score of between 0 and 1, where only keywords scoring above 0 are retained. (3) The Euclidean distance between the head-gaze and each bounding box is computed, and the closest match is identified. (4) The RMSE is then measured.

5.2.2.5 Speech and Head-Gaze as Approximations of Eye Gaze

To assess whether speech improves eye gaze approximation, an offline simulation was conducted across the three experimental conditions using the method detailed in Figure 5.8b. Questionnaire Q10 results (Figure 5.9) indicated that the speech-to-text system used during the experiment exhibited accuracy and latency issues. Therefore, the recorded audio was transcribed using an improved speech-to-text algorithm (Section 5.1.2.1). As the focus was on evaluating verbal communication as a supplementary input for gaze inference, the analysis was restricted to time segments when participants were speaking. The Euclidean distance between head-gaze and the proposed method (Figure 5.8b) was calculated relative to the eye-gaze (ground truth), along with the RMSE of the distances. The RMSE was calculated in screen space and so, therefore, the results are given in pixels. The RM ANOVA results (Figure 5.8a) showed a main effect for the gaze approximation method ($F(1, 45)=7.065, p=.011$) and an interaction between the study condition and gaze approximation method ($F(1.501, 63.210)=8.420, p=.002$). Post hoc comparisons highlighted a difference between Head+Speech ($M = 178.327, SD = 142.517$) and Head ($M = 174.407, SD = 40.401$) in the Eye-tracking condition ($t = 4.703, p < .001$). This difference highlights how our method outperforms head gaze as an approximation of eye gaze in the eye-tracking condition. An interesting insight is that this difference is only present in the

eye-tracking condition where the CoV is not present. Under the CoV and CoV+Speech conditions, the presence of the contours keeps the vision closer to the head-gaze thereby hindering the eye-gaze from spreading wider.

5.2.3 Questionnaire Results

As part of the evaluation, a series of questionnaires with 11 questions answered on five-point Likert scales were administered (Figure 5.9). The first two questions (Q1 and Q2) come from the System Usability Scale (SUS) questionnaire [204], and the second pair of questions (Q3 and Q4) come from the NASA Task Load Index (NASA TLX) questionnaire [178]. Q5 and Q6 aim to understand how visual attention cues affect attention allocation. Q5 question comes from an attention allocation questionnaire and is intended to measure how visual cues attract attention, while Q6 is intended to measure how much visual cues act as a distraction [205].

We also introduced two questions (Q7 and Q8) related to communication to understand whether, across conditions, participants altered their verbal communication behaviour (Q8) or pointing-based communication behaviour (Q9). Q10 is specifically related to how accurately participants perceived the underlying technologies (i.e., eye-tracking, head-tracking, and speech-to-text). Finally, Q11 relates to how users actively alter their natural behaviour to improve feedback passed to collaborators via the visual attention cue.

To analyse Likert items, the Wilcoxon test was used, with a Kruskal–Wallis pairwise comparison for post-hoc testing, where relevant. Only results where statistically significant differences were found are reported. Tests revealed statistically significant differences for Q2 ($W = 5.0, p = 0.002$). A pairwise comparison revealed differences ($H = 9.26, p = .0097$) between CoV+Speech ($M = 2.40, SD = 1.02$) and Eye-tracking conditions ($M = 3.35, SD = 0.57$). A significant Wilcoxon test was found for Q10 ($W = 0.0, p < 0.001$). Pairwise comparisons revealed differences between the CoV+speech ($M = 1.65, SD = 1.11$) and Eye-tracking ($M = 3.15, SD = 0.73$) conditions ($H = 24.08, p < 0.001$), as well as between CoV+speech and CoV ($M = 3.30, SD = 0.78$) conditions ($H = 3.0, p < 0.001$). Responses to these questions highlight well-known problems, when dealing with speech recognition technologies, of

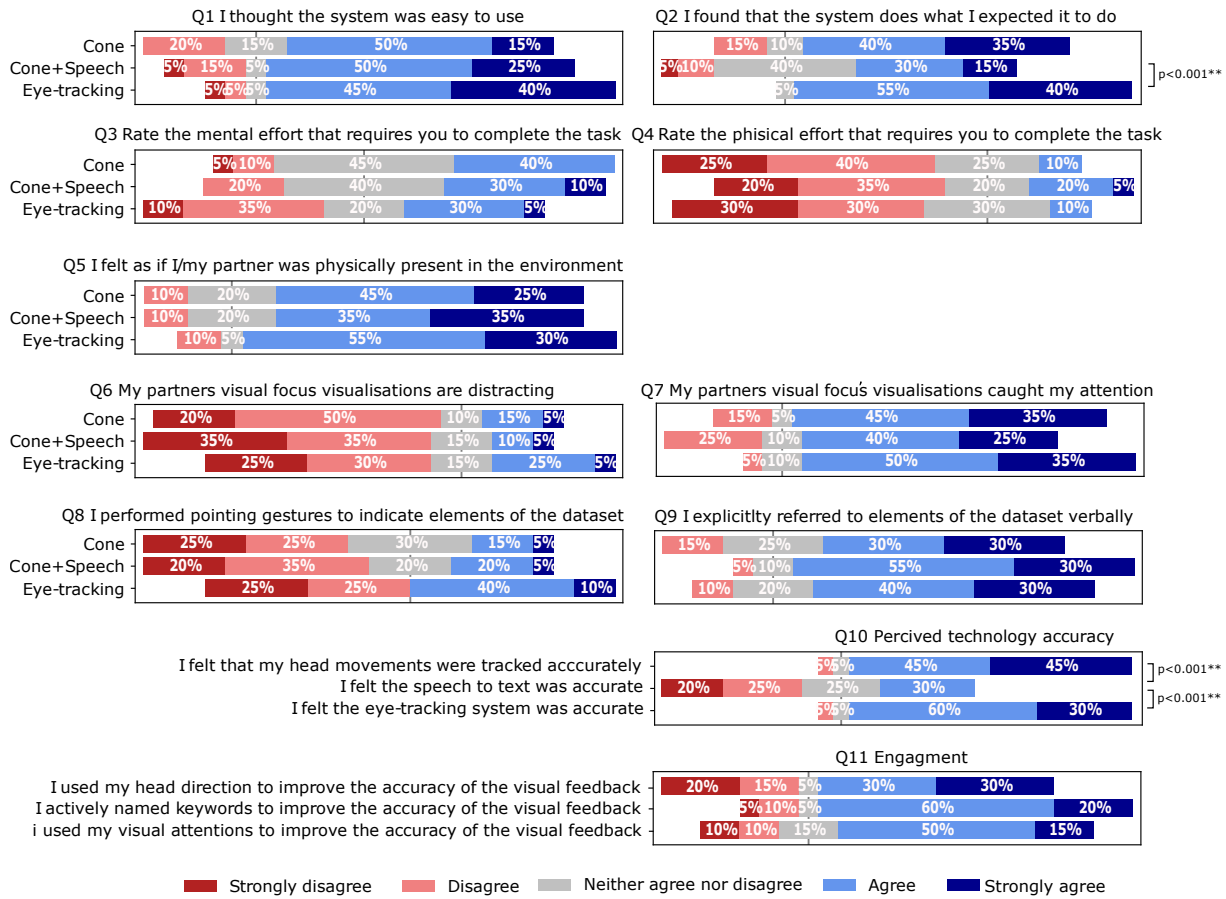


Figure 5.9: Questionnaire Results

latency during real-time use and difficulty in accent recognition [206]. No significant differences were found for NASA TLX, attention, allocation, distraction, verbal communication behaviour, pointing-based communication behaviour, or active engagement questions.

5.3 Qualitative Analysis

Post-experiment semi-structured interviews were audio-recorded, fully transcribed, and analysed via thematic analysis [207]. The research questions focused on verbal communication as input for visual attention cues and, more broadly, the role of verbal communication in collaborative exploratory data analysis. The initial codes for the analysis were derived from these research questions, emphasising aspects related to the perception of the cues, comments on verbal communication, and the impact of the CoV. Additional codes emerged from the interviews, including lag and accuracy issues with speech-to-text technology, workarounds used when visual

attention cues lacked precision, and instances where the CoV facilitated focus. The resulting 30 codes were grouped into three themes, as detailed in the following subsections.

5.3.1 Comparing the Different Visual Cues

Although most participants reported the eye-tracking condition as their favourite due to its precision, some noted that it did not allow them to focus on the charts and, for this reason, preferred the larger contoured region of the cones. For example:

“Eye-tracking was helpful If I was trying to say something specific. But then if either of us were talking about something broader then it would not be helpful because you missed the bigger picture. [P12]”, or: “...so sometimes during a discussion you are not talking about the specific data point but more about the broader and the specific cursor led me to focus on one thing at the detriment of other facts. [P16]”

Some participants reported that gaze movements were hectic and distracting. For example, P13 mentioned: *“...it was like really distracting as I have ADHD, so it’s hard to focus its hard to concentrate on the task, so I could not focus on my collaborator’s visual cue because it was very confusing”*. Similarly P3 said: *“...the eye-tracking visual cue it felt like he’s pulling me away from the where I need to focus...”*

Most participants reported the CoV to be most useful during the initial phase of mutual alignment. For example: *“It was very helpful to get aligned initially and so just for the moment what he needs to align and maybe the moment when the other one goes away [P4]”* or: *“It was helpful when the person I was collaborating with was talking about something, but I didn’t know where she was looking for pictures so I could see the different colour area and turns towards it [P7]”*. P15 said that the CoV was useful in confirming that the two participants were looking at the same thing: *“I found it helpful because I knew where she was looking, so we were able to basically be on the same page”*.

Several participants reported that the CoV contours allowed them to focus on the encircled data: *“...the good part of the cone was that it helped me focus on where I was looking at, so I*

won't look in under directions. [P9]". P17 explicitly stated that they 'liked' the circle produced by the cones: *"We were both focusing on the same things. Like having a line around, you know what you are focusing your attention on, that kind of helps the kind of block out everything else"*. Similarly, P2 stated: *"The cone was helpful as it just helps me concentrate. I don't really feel like it was getting on the way."*

Participants commented that the visual attention cues made them feel more coupled and accountable rather than wandering around and helped them to stay on the same page. For example, P13 reported: *"The cone was helpful in that it kind of kept me in the room so I need to look at the same thing that she was describing, so she would see that I was looking at it was very helpful."*

Participants complained that in the CoV+Speech condition, the cue lagged considerably: *"I feel the speech was picking it up like in 10 seconds I did not find it to be reliable as when it was shirking down it would do so unreliable not in the specific area. [P13]"*. Similarly, P12 stated: *"It felt like it was slow"*.

5.3.2 Verbal Communication as a Fallback

Most participants reported that the CoV worked well for keeping them on the same page, although it was not very precise. So they reported using verbal communication as the default method to refine the accuracy of the CoV. For example: *"I think it was helpful seeing in general where the other person is looking at and then also aligning myself with that we did without yeah but it wasn't like with the specifics obviously it wasn't as helpful so I think we used more like verbal things to see like which chart each person's actual we read [P4]"*. P6 explicitly compared the CoV to the eye-tracking cues: *"With the fixed cone compared to the eye-tracker there is a lot more to verbalise so you had to find out like oh yeah, I have a look over there and get more details to say to the other collaborator"*.

Some participants reported feeling that the eye-tracking was sometimes not perfectly calibrated. Although, initially, they tried to compensate for the error by moving their eyes, they ended

up using verbal communication to specify the location of the data they were discussing: *“The eye-tracking did get in the way a little bit because I feel like it was sometimes it wasn’t calibrated that well, so I’m trying to fix my attention on the specific part of the chart and then the cursor was slightly off in another place so at first I was trying to compensate with my eyes but that wasn’t working so I just had to ignore it and communicate the region of interest verbally [P2].”*

5.3.3 CoV+Speech Condition

The CoV narrowing down on the region of interest was reported as a welcome confirmation of the shared visual attention: *“[it] was nice to have a confirmation of the cone shrinking as it increases the confidence that we were both looking at something [P5]”*. Participants commented that it rarely focused on the wrong area: *“I felt it rarely narrows down on the wrong area, but there was delay [P19]”*.

Sometimes participants went beyond explicitly looking for labels to refer to, and they attempted to direct the CoV, narrowing it with spatial voice commands (e.g., top left, bottom right, etc.). The positions were expected to be understood concerning the virtual screen at which the participant was looking. For example, P13 reported feeling disappointed that such a strategy did not work: *“I found also that was limited in the functionality as it would not recognise top left bottom right corners”*. P2 mentioned that the other participant instead quickly reacted to such spatial references: *“There were few charts in which some of the information on the y and x axes were the same; however, with them, I was mentioning top left or top right, and she would very quickly look there”*. Therefore, the future system that uses speech as an input for visual attention could integrate recognising these forms of verbal, spatial references to inform the visual cue contours without necessitating the semantic knowledge of the context.

5.4 Discussion

The results are discussed by subdividing the findings into Joint Attention (Section 5.4.1) and Individual Visual Attention (Section 5.4.2). The final section (Section 5.4.3) highlights the outcomes of the speech and semantic analysis.

5.4.1 Comparing Joint Attention

Quantitative results indicate that concurrent joint attention on VR screens (Figure 5.5b) was significantly better (increment of 20%) in the CoV condition than in the eye-tracking conditions of Section 5.2.1.1. Despite the fact that the qualitative results showed that the participants preferred the eye-tracking method, the interviews also revealed the reasons for the success of the CoV method in focusing joint attention on screen AOI. Participants mentioned that the cone was helpful in mutual orientation and they found the wider head-based cone contour much easier to find than the eye-tracking cursor. The size of the contour was not the only reason it was easier to find, as it also moved less. Mutual alignment (i.e., orienting themselves along the general direction of collaborators) has been defined by [32] as an essential phase in negotiating a shared visual attention context and is significant for joint attention. Such results extend previous work related to uni-directional visual attention cues of [41], showing that, in the context of 2D VR screens, bi-directional visual attention cues based on head direction outperform the overall mutual alignment when compared to eye-tracking cues.

Moreover, such a finding extends the work of [208], which focuses on bi-directional eye behaviour-based attention cues, as well as the work of [53], which focuses on bi-directional head behaviour-based attention cues, by comparing bidirectional eye-based to head-based visual attention cues. Moreover, quantitative results indicate that, for concurrent joint attention, for the chart area of interest (Figure 5.5a), all tree conditions perform equally well. This result is consistent with Q6 of the questionnaire, which indicates no statistical difference across the experimental conditions. Qualitative analysis of interviews suggests that this result could be due to the effectiveness of verbal communication in refining and specifying the location of the area of interest

(Section 5.3.2). While eye-tracking seems to be the favourite visual attention cue, not all VR headsets have eye-tracking capabilities, and perhaps future cheap VR headsets will never incorporate such capabilities. Our results show that, in the context of exploratory data analysis tasks on VR screens, cheap VR headsets using bi-directional CoV can still lead to the same amount of joint attention and are, therefore, effective.

5.4.2 How Visual Cues Contours Affect Individual Visual Attention

Quantitative results indicate that contour-based visual cues significantly alter individual visual attention by focusing it within the depicted visual cone, with an increase of 20% of gaze sample observed within the contour when the contour is displayed relative to a condition in which the contour is not displayed (Section 5.2.1.2). Furthermore, qualitative results highlight that people perceive such a change and focus more on the area within the depicted region (Section 5.3.1). These results are consistent with the data [181] used to generate the CoV (Figure 3.7) and validated by [53] with two separate datasets [51, 182]. While the eye-tracking condition gaze samples are comparable in percentage terms with the sample of the original dataset, the CoV conditions sample shows the reported 20% increase. This could be because the participants were aware (consciously or unconsciously) that the CoV was a signal of their visual attention to the other person. In other words, they kept their visual attention within the highlighted region where the other person would expect it to be. An alternative explanation is that the contour generates an attention tunnelling effect similar to that of small field of view AR/VR devices [209]. Such results could be interesting in relation to managing the attention of participants who have trouble concentrating (e.g., in ADHD) [210]. A key implication of such an effect is that displaying the CoV can be useful when eye-tracking is not available. For example, metrics about the success of collaboration based on eye-tracking [18, 17] could be more accurately approximated using head-tracking. Furthermore, it might be possible to more effectively adapt interaction techniques that support eye-tracking [42] which only relies on head-tracking.

5.4.3 Speech as Input for Collaboration Support

Our results indicate that participants utter keywords present during visualisation as the primary way to perform verbal references (Section 5.2.2.2, Figure 5.7a), independently of verbal communication being used as input for visual attention cues (Figure 5.7b). Such results indicate that there is potential for speech recognition to be used to refine gaze inference in two ways. First, searching the spoken keywords in the users' visual field (i.e. CoV) and exploiting their location to refine the region of the visual cue is the best strategy for interpreting verbal attention if compared to the other strategies aimed at interpreting different verbal references (Table 5.1) because of its frequency of usage (Table 5.1). Second, such verbal exchange occurs naturally, so our method does not require users to communicate differently. This suggests that, when correctly implemented, such a method could lead to no learning/usage costs for the users. Such results are also consistent with the outcomes of the qualitative analysis Section 5.3.2 in that participants reported using verbal communication during collaboration as a method to overcome the lack of precision of the CoV.

However, our results indicate that further technological advances in automatic speech recognition are needed for this approach to become viable, since real-time state-of-the-art speech-to-text services (Section 5.1.2.1) still suffer from accuracy and lag issues. Such problems emerged from our qualitative analysis (Section 5.3.1) and the questionnaire responses (Section 5.2.3, Figure 5.9). Nevertheless, a high-accuracy transcription was obtained using the audio recording from the experiment using an offline state-of-the-art speech-to-text model (Section 5.1.5.1). Such accurate transcription was used for additional analysis to test speech as an input for gaze inference across all conditions (Section 5.2.2.5, Figure 5.8a). These results illustrate that our method better approximates eye-gaze than head gaze alone when the CoV is not present (Section 5.2.2.5, Figure 5.8a). The results indicate that, in the Eye-Gaze Cursor condition, speech improves the accuracy of the head gaze by approximately 50px on average, and that this is statistically significant. Results from the same analysis also show that, under the conditions in which the CoV or CoV+Speech is used, our method does not show improvements, most probably because the gaze is constrained by the visual cues (as discussed in Section 5.4.2).

The dataset related to head, gaze, and accurate transcript speech behaviour has been released to encourage further research in this direction. Gaze inference models other than eye-tracking could be beneficial for those low-cost eye-tracker-less headsets or for offline analysis which lacks gaze data, but retains speech and head direction information.

Chapter 6

Conclusion

The primary aim of this thesis was to enhance collaboration and productivity in immersive XR office environments, in which 2D documents are visualised on VR screens within a 3D space. This required addressing a dual challenge: a hardware challenge, where low-cost, eye-tracker-less VR headsets lack the sensors needed to directly capture gaze, and a conceptual challenge, where even when gaze is available, it is an imperfect proxy for users' underlying higher order of attention.

To address this problem, the primary research objectives (Section 1.2) of this thesis were to:

- Develop novel methods to infer visual attention in eye-tracker-less XR headsets from accessible input data such as head movements and speech cues;
- Design adaptive attention cues that better approximate users' attention;
- Evaluate the impact of these methods on collaboration, user experience, and task efficiency through empirical studies.

This research successfully introduced novel approaches to estimate and visualise users' focus using head movement velocity and augment these estimates by interpreting users' collaborative speech. The proposed methods successfully improved focus estimation, allowing for real-time

adjustments of visual attention from a broad, scene-level awareness to a fine-grained, local focus in XR environments, effectively operationalizing a fundamental characteristic of the higher-order of attention. User studies demonstrated that these techniques significantly enhanced mutual awareness and task coordination, without eye-tracking technology, making them both effective and accessible for low-cost VR headsets.

These approaches help bridge the gap between high-cost VR headsets with eye tracking and low-cost eye-tracker-less headsets, while simultaneously laying the foundation to fully address the conceptual gap between raw visual sensors and the accurate reconstruction of cognitive intent. This goal aligns with the broader vision of fulfilling the promise of XR as a platform for spatial hyper-productivity.

6.1 Discussion

Chapter 3, Chapter 4, and Chapter 5 each contribute a piece of the puzzle, highlighting the modular nature of the contributions of this thesis. Chapter 3 introduced the Cone of Vision (*CoV*), an FoV-based visualisation cue designed to display visual attention for improved collaboration. Subsequently, Chapter 4–Chapter 5 introduced two complementary methods based on the CoV—a real-time deep-learning model for gaze probability and a Speech-Augmented CoV—each building on the original cue to meet increasingly refined requirements for mutual awareness in XR. The CoV establishes a fundamental head-based cue; the deep-learning model refines probabilistic gaze estimation; and the Speech-Augmented CoV allows for targeted, fine-grained attention when user communication clarifies ambiguous references. Collectively, these enhancements trace a clear evolution from a unimodal to a multi-modal approach, in which each added dimension—head direction, behaviour, and speech—further amplifies the prior technique’s effectiveness. This progression explores novel inputs and methods to both infer and visualise user attention in immersive collaborative environments.

From a technical perspective, Chapter 4 introduced two novel approaches to infer visual attention, focusing on optimising computational efficiency and predictive accuracy. For prediction

dimensionality (Section 4.1.2), a line-based PDF representation was designed to enhance CoV visualisation while minimising computational cost (Section 4.2.2). For accurate probability representation, an autoencoder-based approach (Section 4.1.4.4) outperformed bivariate Gaussians in certain cases, demonstrating its viability as an alternative (Section 4.2). Additionally, the model achieved its highest accuracy during left and right head movements (Section 4.2.1), suggesting that gaze estimation is most beneficial when clear directional head cues are present. The speech-based model (Chapter 5) introduced verbal input as a dynamic modifier, refining gaze inference and significantly improving head-gaze-based predictions (Section 5.2.2.5). However, its effectiveness remains contingent on verbal engagement during collaboration, making it more applicable to dialogue-rich tasks.

During evaluation, each method showcased distinct strengths. The *plain CoV* (Chapter 3) excels at providing broad mutual awareness when stable visual cues are needed (i.e. opposite to erratic like eye gaze pointers), for instance during transitions between different screens or when collaborators move from individual to joint work. By reducing the need to interpret each other's avatar orientation (Section 3.2.1.9), participants can remain focused on the data. However, while helpful for general alignment, plain CoV can lack the precision required in certain collaborative phases (Section 3.2.2.2), leading us to explore additional methods. The Deep-Learning Probabilistic Model (Chapter 4) offers more precise gaze estimation during head movement but was evaluated only on an existing dataset rather than through a dedicated user study. Although it performs well in tasks requiring frequent head rotation or real-time density mapping, communication in EDA tasks often happens when users hold their heads still, making its direct impact less evident. Lastly, the Speech-Augmented CoV (Chapter 5) is most effective in refining or disambiguating attention cues when precise references to displayed data arise or when conversations are dense and detail-oriented. This approach narrows the CoV in real time based on spoken keywords, thereby addressing precision needs that plain CoV alone cannot fulfil.

6.1.1 Region-Based Visualisations and Awareness

Our first study introduced the CoV in a collaborative 2D VR environment, demonstrating that head-based attention cues significantly improved mutual awareness and alignment during transitions between individual and joint tasks. Notably, participants reported relying less on avatar interpretation (Figure 3.10e) and qualitative feedback, highlighting the CoV’s role in preventing unintentional splits in focus (Section 3.2.2.1). In the second study, comparing CoV to eye-gaze cursors revealed that CoV attention cues supported joint attention equally or more effectively. Joint attention increased by approximately 20% when using head-based cues (Figure 5.5b, Section 5.2.1.1), underscoring the CoV’s ability to foster more stable and cohesive collaboration. Collectively, these two studies illustrate a clear progression from establishing the fundamental benefits of CoV for mutual awareness to confirming its superiority—or at least parity—when directly compared with eye-gaze methods.

6.1.2 Self-Awareness of Visual Cues

A recurring theme in our studies (Section 5.1, Section 3.1) was that users become more aware of their own attentional cues when these are visually augmented, whether through the CoV itself or in response to contextual speech references. This enhanced self-awareness amplifies both collaboration and individual behaviour. As participants monitor not only their collaborator’s attention but also the display of their own CoV, they naturally align or “contain” their gaze within that region (Chapter 3), thereby reducing misunderstandings and coordinating more fluidly in real time (Section 3.2.2.3). This behaviour suggests both conscious and less deliberate influences at play, highlighting the powerful impact of visual attention cues on individual gaze patterns. Meanwhile, even though our data show no significant statistical shift in the ratio of implicit vs. explicit speech references, participants reported feeling more intentional about their language when they knew it would reshape the CoV (Figure 5.7, Q9 in Figure 5.9, and Section 5.3.3). In this way, self-awareness and bidirectional feedback not only improve mutual understanding and task performance but also subtly influence how users direct their gaze and communicate verbally, underscoring the broader impact that visualising one’s own attention

can have on collaborative behaviour in XR.

6.2 Impact and Implications

This research challenges the assumption that raw eye-tracking is the gold standard for attention cues in XR. Literature warns of sensory noise and calibration drift [211, 161] (see Section 2.4.1)—flaws this thesis confirms can actively distract collaborators and force a reversion to verbal descriptions (see Section 5.3.2 and Section 5.3). In contrast, this work proves that visual focus can be robustly reconstructed by leveraging multimodal signals like head pose and speech. Offline analysis demonstrates that extracting semantic targets from speech significantly improves gaze inference accuracy over head orientation alone (see Section 5.1.5.3). By reducing spatial ambiguity and filtering sensory noise, this approach proposes a fundamental shift in design philosophy: moving from merely replicating gaze mechanics to utilizing higher-order of cognitive attention.

This shift has significant implications for the design of collaborative interfaces. It suggests that future systems should prioritize cognitive alignment—supporting the user’s fluctuating need for both ‘global’ and ‘local’ attention—over static sensor precision. As identified in the literature, attention is dynamic in scale, acting as a zoom lens that expands to scan scenes or contracts to scrutinize details [27, 28] (see Section 2.2.5). This thesis proves the importance of broad attention during collaborative alignment, demonstrating that head-based cues (CoV) significantly outperform precise gaze cursors during global phases such as screen transitions [143] (see Section 5.2.1.1). Furthermore, it establishes the importance of speech towards reducing ambiguity and visualizing ‘higher-order’ cognitive attention. By integrating verbal communication as a semantic anchor, the Speech-Augmented CoV (Chapter 5) offers a framework to bridge global awareness and local precision, using the user’s own descriptions to dynamically refine the visual cue. This implies that effective collaboration tools must be context-aware interpreters of human behaviour, capable of adapting to the changing scale of cognitive intent.

Finally, while these methods were initially motivated by the hardware constraints of low-cost headsets [16] (see Section 1.1.2), their utility extends far beyond cost-saving. In the emerging context of XVA environments and spatial knowledge work, users require interaction modalities that are seamless and intuitive. This work demonstrates that naturally occurring behaviours—looking and speaking—can be harnessed as powerful implicit inputs. Consequently, leveraging these natural signals is not merely a hardware workaround for today’s devices; it is a fundamental requirement for realising the seamless, hyper-productive augmented collaboration promised by the next generation of spatial computing.

6.3 Limitations and Future Work

6.3.1 Visual Attention Inference

6.3.1.1 Model Evaluation and Generalisability

At present, the Model presented in Chapter 4 is trained and validated offline, by comparing the inferred gaze distribution against ground-truth eye-tracking data using RMSE and the proportion of eye-gaze samples contained within the predicted probability contours (see Section 4.2). The proposed model relies on the premise that eye-head coordination is task-independent (see Section 4.1.1) because of its physiological nature (see Section 2.4.2). However, this offline validation does not yet confirm how precisely these results transfer to EDA task settings where cognitive intent drives attention differently than in the passive exploration of 360-degree video footage [181]. Future work should therefore extend this evaluation in two directions. First, the model should be tested on a dataset of EDA tasks performed within an XVA environment, to verify that gaze patterns learned from passive video consumption generalize to the dynamics of active EDA tasks, thereby mitigating potential biases introduced by the training data. Second, beyond offline error metrics, controlled user studies are needed to examine whether improvements in RMSE and coverage (i.e., the proportion of eye-gaze samples contained within the predicted probability contours) actually translate into better perceived visual cue specificity

and increased mutual awareness in immersive analytics and XR collaboration scenarios.

6.3.1.2 Accuracy of Speech-Based Input

The accuracy of speech-based input is contingent on the quality of the speech recognition system and the clarity of the user’s speech. Future work could investigate more robust speech recognition techniques or alternative methods for incorporating verbal communication, such as semantic analysis or understanding of natural language. As highlighted in our qualitative analysis (Section 5.3.3), participants occasionally made spatial references, such as ”on my left” or ”top right corner,” that were not fully utilised by our system. The system could be extended to better exploit such verbal communication. Future work should investigate how spatial references which do not include keywords, such as ”on my left” or ”in the top-right corner,” could be explicitly used to manipulate visual cues in real time. Additionally, exploring the integration of verbal inputs that reference specific colours, shapes, or even synonyms for visible keywords could further enhance mutual understanding during collaboration (Section 5.1.5.2). Moreover, expanding the system to function in more complex XR environments, such as those involving 3D data, would be an important next step. Leveraging 3D meta-information or real-time video segmentation could provide a richer dataset for verbal references and collaborative visualisations [212].

6.3.1.3 Co-design visual attention inference with UI Layout

Knowledge work interfaces are composed of discrete semantic elements—plots, buttons, headings—that are inherently high in saliency [52, 51]. Since gaze is strongly correlated with these informative regions [182, 213], the spatial layout itself acts as a strong prior for gaze prediction, modulating ambiguity even when head movements remain identical. While this thesis excluded generic visual saliency due to its prohibitive computational cost (Section 2.4.3), window-based VR environments might offer a more efficient alternative: UI layout metadata.

In our XVA environment Section 5.1.2.1, the 2D semantic layout is explicitly exposed, providing low-cost bounding boxes for all UI elements. Unlike pixel-based saliency, these metadata-

driven priors do not require per-frame re-computation, highlighting the significant merit in co-designing the attention inference system with the UI environment.

Concretely, these priors could be operationalised via two approaches: (i) probabilistic integration, where the semantic layout is output as a weighted probability texture and merged with the CoV’s underlying distribution on the GPU; or (ii) heuristic adjustments, drawing inspiration from eye driven interfaces where “magnetic” targets and dwell-based selection are common [214, 215, 216]. In the latter approach, the borders of the CoV would be attracted/repulsed to/by semantic UI boundaries, ensuring that region based contours reliably “locks on” to discrete objects disregarding empty background.

6.3.1.4 Reliance on Head Movement

The reliance on head movement for gaze estimation might be less effective for users with limited mobility or in scenarios where head movements are restricted. Although head-based methods proved effective in this context, they may not generalise well to all XR applications. Further work could explore the use of other behavioural cues, such as hand movements or body posture to further enhance gaze estimation accuracy, especially in cases where head movement is limited. For example, further development of the real-time deep learning model presented in Chapter 4 could include optimising the model for tasks such as interactive pointing (i.e. hand movements) or collaborative problem-solving (i.e. group body dynamics).

6.3.2 Attention Representation

While this thesis approximates the “Zoom Lens” and “Spotlight” models [27, 28] through complementary reconstruction, it does not fully reconstruct the underlying cognitive state. By shifting from raw sensory tracking to adaptive regions, sensory noise is filtered (Section 2.2.5), yet two significant limitations persist regarding the dependency on overt behaviour.

Reliance on Explicit Input (The Silent User): The Speech-Augmented CoV successfully realises a dynamic “Zoom Lens” during verbalisation, but lacks a mechanism for silent analysis.

When a user scrutinises a detail without speaking, the system defaults to the broad, static CoV, failing to capture the covert narrowing of cognitive focus. Consequently, the representation remains contingent on explicit, outward communication rather than truly reflecting the user’s internal state.

Persistence of Decoupling (Looking $\neq$ Seeing): While speech helps ground attention when users physically look away, the system cannot detect internal decoupling—such as mind-wandering or deep reasoning—where a user stares at a target but processes unrelated concepts. I have improved referential accuracy for overt behaviour, but the real-time distinction between merely “looking” and active “processing” remains unsolved.

Crucially, these gaps stem from fundamental disconnects between observable behaviour and internal cognition, affecting even high-fidelity eye-tracking systems. Bridging this final mile may require integrating physiological signals (e.g., EEG, BCI) or leveraging deep semantic context models (e.g., LLMs reasoning over interaction history) to infer intent even in the absence of explicit speech.

6.3.2.1 The “Horse Blinder” and “Midas” Effects

Externalising attention introduces two distinct behavioural effects: the exploration-limiting “Horse Blinder” effect and the performative “Midas” effect.

The “Horse Blinder” Effect: Our analysis (Section 5.2.1.2) shows that the CoV acts as an attention container, causing users to concentrate almost exclusively within the highlighted area. While this limits peripheral awareness, akin to “attentional narrowing” [217, 218]—it also offers specific benefits; for instance, a participant with ADHD reported that the containment helped reduce distractions (Section 5.3.1). Future work should investigate how CoV sizing can balance this trade-off between tunnel vision and useful guidance.

The “Midas Touch”: The “Midas effect” describes the dilemma where natural behaviours are interpreted as intentional commands [219, 220]. In our context, exposing gaze and speech as shared cues created a feedback loop: participants became aware their behaviours were driving the visualisation and adapted them to “game” the system. For instance, participants reported using more intentional language or issuing spatial commands specifically to reshape the CoV

(Figure 5.7, Figure 5.9, Section 5.3.3). While this makes communication less natural and more performative, it paradoxically acts as a mechanism for clarity, encouraging users to be more deliberate in their signalling.

6.3.2.2 Integration of Head and Eye Gaze for better Attention Cues

Participants in several experimental conditions reported that the CoV was easier to track than eye-tracking-based methods, primarily due to its stability. In contrast, the erratic movements of eye-tracking cursors proved distracting for some, highlighting the trade-off between stability and precision. While eye-tracking offers greater accuracy in pinpointing visual attention, its rapid and sometimes unpredictable movements can be overwhelming (Section 5.3.1). The CoV, based on head movements, provides a smoother and more stable representation, albeit at the expense of precision. Even in headsets equipped with eye-tracking, head direction could be concurrently utilised to enhance visual attention estimation. This suggests that a combination of head and eye tracking may offer the best of both approaches, balancing stability and precision to create a more effective and user-friendly visualisation of attention. Future research should explore the potential for multi-modal integration between head and eye tracking to develop a more comprehensive and functional representation of user behaviour. This approach could lead to a more stable and less distracting method of visualising attention, improving the efficiency of collaboration in XR environments. Additionally, expanding the use of speech as an input, coupled with more advanced natural language processing techniques, could further enhance the system's ability to interpret collaborative speech and improve the accuracy of visual attention estimation.

6.3.3 Asynchronous Collaboration

With the rise of global, time-distributed teams [221, 222], there is a significant opportunity to extend the Speech-Augmented CoV to support asynchronous collaboration. By transforming synchronous traces of visual attention and speech into persistent artefacts, future systems

could facilitate knowledge transfer across different time zones. This opens three key research directions.

First, future work should explore auto-generating annotation units. Unlike raw video segmentation [223, 224], XVA environments provide structured semantic data (speech, head pose, object labels) that allows for more robust segmentation. These units could be defined via simple spatial heuristics (e.g., joint dwell time) or derived from spatially grounded models, such as spatial LLMs [225], that reason over the shared 3D context.

Second, the visual representation and navigation of these units warrants investigation. Annotations could take the form of static summary cards anchored to specific semantic regions, or continuous representations like attention heatmaps, where intensity serves as a spatial metaphor for time—allowing users to click “hot” regions to jump to relevant moments in the recording. Finally, systems should investigate context-aware resurfacing, where past discussions are automatically presented to an asynchronous user based on their real-time visual focus or verbal queries, dynamically bridging the gap between past insights and current tasks.

6.4 Closing Remarks

This thesis has explored innovative ways to close the gap between low-cost, eye-tracker-less XR headsets and high-end, eye-tracking-enabled models. By demonstrating that multimodal reconstruction can reduce ambiguity without relying on expensive hardware, this work validates a shift from merely replicating gaze mechanics to visualizing the higher-order of attention. The advancements presented here establish that leveraging natural collaborative behaviours is not just a workaround, but a fundamental requirement for intuitive, hyper-productive spatial computing. The journey towards an inclusive and accessible XR future is ongoing, and this work represents an important step towards realising that vision.

Bibliography

- [1] Mario A Rojas-Sánchez, Pedro R Palos-Sánchez, and José A Folgado-Fernández. Systematic literature review and bibliometric analysis on virtual reality and education. *Education and Information Technologies*, 28(1):155–192, 2023.
- [2] Caroline Fertleman, Phoebe Aubugeau-Williams, Carmel Sher, Ai-Nee Lim, Sophie Lumley, Sylvie Delacroix, and Xueni Pan. A discussion of virtual reality as a new tool for training healthcare professionals. *Frontiers in public health*, 6:44, 2018.
- [3] Antonio E. Uva, Michele Gattullo, Vito M. Manghisi, Daniele Spagnulo, Giuseppe L. Cascella, and Michele Fiorentino. Evaluating the effectiveness of spatial augmented reality in smart manufacturing: a solution for manual working stations. *The International Journal of Advanced Manufacturing Technology*, 94(1-4):509–521, 1 2018. ISSN 0268-3768. doi: 10.1007/s00170-017-0846-4. URL <http://link.springer.com/10.1007/s00170-017-0846-4>.
- [4] Markus Funk, Thomas Kosch, Scott W. Greenwald, and Albrecht Schmidt. A benchmark for interactive augmented reality instructions for assembly tasks. *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia - MUM '15*, (Mum):253–257, 2015. doi: 10.1145/2836041.2836067. URL <http://dl.acm.org/citation.cfm?doid=2836041.2836067>.
- [5] Apple. Apple vision pro. <https://www.apple.com/vision-pro/>, 2022.
- [6] Statista. Metaverse workplace in the united kingdom. <https://www.statista.com/>

- outlook/amo/metaverse/metaverse-workplace/united-kingdom, 2024. Accessed on 12 March 2024.
- [7] Verena Biener, Daniel Schneider, Travis Gesslein, Alexander Otte, Bastian Kuth, Per Ola Kristensson, Eyal Ofek, Michel Pahud, and Jens Grubert. Breaking the screen: Interaction across touchscreen boundaries in virtual reality for mobile knowledge workers. *IEEE transactions on visualization and computer graphics*, 26(12):3490–3502, 2020.
- [8] Leonardo Pavanatto and Doug A Bowman. Virtual displays for knowledge work: Extending or replacing physical monitors for more flexibility and screen space. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 667–669. IEEE, 2024.
- [9] Eyal Ofek, Jens Grubert, Michel Pahud, Mark Phillips, and Per Ola Kristensson. Towards a practical virtual office for mobile knowledge workers. *arXiv preprint arXiv:2009.02947*, 2020.
- [10] Nadia Fereydooni and Bruce N Walker. Virtual reality as a remote workspace platform: Opportunities and challenges. 2020.
- [11] Hiroshi Ishii and Minoru Kobayashi. Clearboard: A seamless medium for shared drawing and conversation with eye contact. *Human-Computer Interaction*, 7(3):341–347, 1992. doi: 10.1207/s15327051hci0703_3.
- [12] Sarah D’Angelo and Andrew Begel. Improving communication between pair programmers using shared gaze awareness. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI ’17, page 6245–6290, New York, NY, USA, 2017. ACM. ISBN 9781450346559. doi: 10.1145/3025453.3025573. URL <https://doi.org/10.1145/3025453.3025573>.
- [13] Lars Winkler Pettersson, Andreas Kjellin, Mats Lind, and Stefan Seipel. On the role of visual references in collaborative visualization. *Information Visualization*, 9(2):98–114, 2010. ISSN 14738716. doi: 10.1057/ivs.2009.2.

- [14] Amin Jourabloo, Fernando De la Torre, Jason Saragih, Shih-En Wei, Stephen Lombardi, Te-Li Wang, Danielle Belko, Autumn Trimble, and Hernan Badino. Robust egocentric photo-realistic facial expression transfer for virtual reality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20323–20332, 2022.
- [15] Young-Woon Cha, True Price, Zhen Wei, Xinran Lu, Nicholas Rewkowski, Rohan Chabra, Zihe Qin, Hyounghun Kim, Zhaoqi Su, Yebin Liu, Adrian Ilie, Andrei State, Zhenlin Xu, Jan-Michael Frahm, and Henry Fuchs. Towards fully mobile 3d face, body, and environment capture using only head-worn cameras. *IEEE Transactions on Visualization and Computer Graphics*, 24(11):2993–3004, 2018. doi: 10.1109/TVCG.2018.2868527.
- [16] Thomas Alsop. Virtual reality (vr) headset unit sales worldwide in 4th quarter 2019 and 4th quarter 2020, by device, February 2021. URL <https://www.statista.com/statistics/987701/vr-unit-sales-brand/>.
- [17] Hana Vrzakova, Mary Jean Amon, Angela E. B. Stewart, and Sidney K. D’Mello. Dynamics of visual attention in multiparty collaborative problem solving using multidimensional recurrence quantification analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1–14, New York, NY, USA, 2019. ACM. ISBN 9781450359702. doi: 10.1145/3290605.3300572. URL <https://doi.org/10.1145/3290605.3300572>.
- [18] Maureen Villamor and Ma. Mercedes Rodrigo. Predicting successful collaboration in a pair programming eye tracking experiment. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP ’18, page 263–268, New York, NY, USA, 2018. ACM. ISBN 9781450357845. doi: 10.1145/3213586.3225234. URL <https://doi.org/10.1145/3213586.3225234>.
- [19] Robert Moulder, Brandon Booth, Angelina Abitino, and Sidney D’Mello. Recurrence Quantification Analysis of Eye Gaze Dynamics during Team Collaboration. pages 430–440. Association for Computing Machinery, 3 2023. ISBN 9781450398657. doi: 10.1145/3576050.3576113.

- [20] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in psychology*, 6:1049, 2015.
- [21] Joseph H Goldberg and Jack C Schryver. Eye-gaze determination of user intent at the computer interface. In *Studies in Visual Information Processing*, volume 6, pages 491–502. Elsevier, 1995.
- [22] Carl Gutwin and Saul Greenberg. A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work (CSCW)*, 11(3):411–446, 2002.
- [23] Bertrand Schneider and Roy Pea. Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-Supported Collaborative Learning*, 8(4):375–397, 2013. ISSN 15561607. doi: 10.1007/s11412-013-9181-4.
- [24] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.
- [25] Jan Theeuwes. Top-down and bottom-up control of visual selection. *Acta psychologica*, 135(2):77–99, 2010.
- [26] Steven L Franconeri and Daniel J Simons. Moving and looming stimuli capture attention. *Perception & psychophysics*, 65(7):999–1010, 2003.
- [27] Charles W Eriksen and James D St. James. Visual attention within and around the field of focal attention: A zoom lens model. *Perception & psychophysics*, 40(4):225–240, 1986.
- [28] Kyle R Cave and Narcisse P Bichot. Visuospatial attention: Beyond a spotlight model. *Psychonomic bulletin & review*, 6(2):204–223, 1999.
- [29] Jonathan Smallwood and Jonathan W Schooler. The science of mind wandering: Empirically navigating the stream of consciousness. *Annual review of psychology*, 66(1):487–518, 2015.

- [30] Jonathan W Schooler, Jonathan Smallwood, Kalina Christoff, Todd C Handy, Erik D Reiche, and Michael A Sayette. Meta-awareness, perceptual decoupling and the wandering mind. *Trends in cognitive sciences*, 15(7):319–326, 2011.
- [31] Mathias Benedek, Robert Stoiser, Sonja Walcher, and Christof Körner. Eye behavior associated with internally versus externally directed cognition. *Frontiers in psychology*, 8:1092, 2017.
- [32] Nelson Wong and Carl Gutwin. 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, page 1377–1387, New York, NY, USA, 2014. ACM. ISBN 9781450325400. doi: 10.1145/2531602.2531691. URL <https://doi.org/10.1145/2531602.2531691>.
- [33] B. Fröhler, C. Anthes, F. Pointecker, J. Friedl, D. Schwajda, A. Riegler, S. Tripathi, C. Holzmann, M. Brunner, H. Jodlbauer, H. C. Jetter, and C. Heinzl. A Survey on Cross-Virtuality Analytics. *Computer Graphics Forum*, 41(1):465–494, 2022. ISSN 14678659. doi: 10.1111/cgf.14447.
- [34] Kadek Ananta Satriadi, Barrett Ens, Maxime Cordeil, Tobias Czauderna, and Bernhard Jenny. Maps around me: 3d multiview layouts in immersive spaces. *Proc. ACM Hum.-Comput. Interact.*, 4(ISS), nov 2020. doi: 10.1145/3427329. URL <https://doi.org/10.1145/3427329>.
- [35] Keita Higuch, Ryo Yonetani, and Yoichi Sato. Can eye help you? effects of visualizing eye fixations on remote collaboration scenarios for physical tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI ’16, page 5180–5190, New York, NY, USA, 2016. ACM. ISBN 9781450333627. doi: 10.1145/2858036.2858438. URL <https://doi.org/10.1145/2858036.2858438>.
- [36] Huidong Bai, Prasanth Sasikumar, Jing Yang, and Mark Billinghurst. A user study on mixed reality remote collaboration with eye gaze and hand gesture sharing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–13, New

- York, NY, USA, 2020. ACM. ISBN 9781450367080. URL <https://doi.org/10.1145/3313831.3376550>.
- [37] Nelson Wong and Carl Gutwin. Where are you pointing? The accuracy of deictic pointing in CVEs. In *Conference on Human Factors in Computing Systems - Proceedings*, volume 2, pages 1029–1038, New York, New York, USA, 2010. ACM Press. ISBN 9781605589299. doi: 10.1145/1753326.1753480. URL <http://portal.acm.org/citation.cfm?doid=1753326.1753480>.
- [38] Diederick C. Niehorster, Thiago Santini, Roy S. Hessels, Ignace T.C. Hooge, Enkelejda Kasneci, and Marcus Nyström. The impact of slippage on the data quality of head-worn eye trackers. *Behavior Research Methods*, 52(3):1140–1160, 2020. ISSN 15543528. doi: 10.3758/s13428-019-01307-0.
- [39] Jon Hindmarsh, Mike Fraser, Christian Heath, Steve Benford, and Chris Greenhalgh. Fragmented interaction: Establishing mutual orientation in virtual environments. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work, CSCW '98*, page 217–226, New York, NY, USA, 1998. ACM. ISBN 1581130090. doi: 10.1145/289444.289496. URL <https://doi.org/10.1145/289444.289496>.
- [40] Thammathip Piumsomboon, Arindam Dey, Barrett Ens, Gun Lee, and Mark Billinghurst. [poster] covar: Mixed-platform remote collaborative augmented and virtual realities system with shared collaboration cues. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pages 218–219, 2017. doi: 10.1109/ISMAR-Adjunct.2017.72.
- [41] Thammathip Piumsomboon, Arindam Dey, Barrett Ens, Gun Lee, and Mark Billinghurst. The effects of sharing awareness cues in collaborative mixed reality. *Frontiers Robotics AI*, 6(FEB), 2019. ISSN 22969144. doi: 10.3389/frobt.2019.00005.
- [42] William Steptoe, Oyewole Oyekoya, Alessio Murgia, Robin Wolff, John Rae, Estefania Guimaraes, David Roberts, and Anthony Steed. Eye tracking for avatar eye gaze control during Object-Focused multiparty interaction in immersive collaborative virtual

- environments. In *Proceedings - IEEE Virtual Reality*, pages 83–90. IEEE, 2009. ISBN 9781424439430. doi: 10.1109/VR.2009.4811003.
- [43] Ivan Moser, Martin Hlostá, Per Bergamin, Umesh Ramnarain, Christo Van der Westhuizen, Mafor Penn, Noluthando Mdlalose, Koketso Pila, and Ogegbo Ayodele. Approximating eye gaze with head pose in a virtual reality microteaching scenario for pre-service teachers. In *LAK Workshops*, pages 272–283, 2024.
- [44] Yirong Yang. Technical challenges affecting the popularization of vir. In *Proceedings of the 2024 International Conference on Mechanics, Electronics Engineering and Automation (ICMEEA 2024)*, volume 240, page 261. Springer Nature, 2024.
- [45] Robert Moulder, Brandon Booth, Angelina Abitino, and Sidney D’Mello. Recurrence Quantification Analysis of Eye Gaze Dynamics during Team Collaboration. In *ACM International Conference Proceeding Series*, pages 430–440. Association for Computing Machinery, 3 2023. ISBN 9781450398657. doi: 10.1145/3576050.3576113.
- [46] Seungwon Kim, Gun Lee, Mark Billinghurst, and Weidong Huang. The combination of visual communication cues in mixed reality remote collaboration. *Journal on Multimodal User Interfaces*, pages 1–15, 7 2020. ISSN 17838738. doi: 10.1007/s12193-020-00335-x. URL <https://doi.org/10.1007/s12193-020-00335-x>.
- [47] Ludwig Sidenmark and Hans Gellersen. Eye & Head: Synergetic eye and head movement for gaze pointing and selection. In *UIST 2019 - Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pages 1161–1174. Association for Computing Machinery, Inc, 10 2019. ISBN 9781450368162. doi: 10.1145/3332165.3347921.
- [48] Tobit Kollenberg, Alexander Neumann, Dorothe Schneider, Tessa Karina Tews, Thomas Hermann, Helge Ritter, Angelika Dierker, and Hendrik Koesling. *Visual search in the (Un)real world: How head-mounted displays affect eye movements, head movements and target detection*. 2010. ISBN 9781605589947. doi: 10.1145/1743666.1743696.

- [49] Peng Wang, Shusheng Zhang, Xiaoliang Bai, Mark Billinghurst, Weiping He, Shuxia Wang, Xiaokun Zhang, Jiaxiang Du, and Yongxing Chen. Head pointer or eye gaze: Which helps more in MR remote collaboration. In *26th IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2019 - Proceedings*, pages 1219–1220. Institute of Electrical and Electronics Engineers Inc., 3 2019. ISBN 9781728113777. doi: 10.1109/VR.2019.8798024.
- [50] Nelson Wong and Carl Gutwin. 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*, page 1377–1387, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450325400. doi: 10.1145/2531602.2531691. URL <https://doi.org/10.1145/2531602.2531691>.
- [51] Zhiming Hu, Congyi Zhang, Sheng Li, Guoping Wang, and Dinesh Manocha. SGaze: A Data-Driven Eye-Head Coordination Model for Realtime Gaze Prediction. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2002–2010, 2019. ISSN 19410506. doi: 10.1109/TVCG.2019.2899187.
- [52] Ali Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 438–445. IEEE, 2012.
- [53] Riccardo Bovo, Daniele Giunchi, Muna Alebri, Anthony Steed, Enrico Costanza, and Thomas Heinis. Cone of vision as a behavioural cue for vr collaboration. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 2022. doi: 10.1145/3555615.
- [54] Riccardo Bovo, Daniele Giunchi, Ludwig Sidenmark, Hans Gellersen, Enrico Costanza, and Thomas Heinis. Real-time head-based deep-learning model for gaze probability regions in collaborative vr. In *Proceedings of the Eye Tracking Research and Applications Symposium (ETRA)*, 2022. doi: 10.1145/3517031.3529642.
- [55] Riccardo Bovo, Daniele Giunchi, Ludwig Sidenmark, Joshua Newn, Hans Gellersen, Enrico Costanza, and Thomas Heinis. Speech-augmented cone-of-vision for exploratory data

- analysis. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2023. doi: 10.1145/3544548.3581283.
- [56] Riccardo Bovo, Daniele Giunchi, Enrico Costanza, Anthony Steed, and Thomas Heinis. Shall i describe it or shall i move closer? verbal references and locomotion in vr collaborative search tasks. In *Proceedings of 20th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET), 2022.
- [57] Riccardo Bovo, Steven Abreu, Karan Ahuja, Eric J. Gonzalez, Li-Te Cheng, and Mar Gonzalez-Franco. Embardiment: An embodied ai agent for productivity in xr. *arXiv preprint arXiv:2408.08158*, 2024. doi: 10.48550/arXiv.2408.08158.
- [58] Daniele Giunchi, Riccardo Bovo, Nels Numan, and Anthony Steed. Streamspace: A framework for window streaming in collaborative mixed reality environments. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (IEEEVR)*, 2024.
- [59] Riccardo Bovo, Eric J. Gonzalez, Li-Te Cheng, and Mar Gonzalez-Franco. Windowmirror: An open-source toolkit to bring interactive multi-window views into xr. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (IEEEVR)*, 2024.
- [60] Tiger Cross, Riccardo Bovo, and Thomas Heinis. Browsing internet content in multiple dimensions. In *Proceedings of the EDBT/ICDT Workshops*, 2021.
- [61] Daniele Giunchi, Riccardo Bovo, Panayiotis Charalambous, Fotis Liarokapis, Alastair Shipman, Stuart James, Anthony Steed, and Thomas Heinis. Perceived realism of pedestrian crowds trajectories in vr. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology (VRST)*, 2021.
- [62] Hugo Mayo, Alastair Shipman, Daniele Giunchi, Riccardo Bovo, Anthony Steed, and Thomas Heinis. Vr toolkit for identifying group characteristics. *Collective Dynamics*, 2021.

- [63] Wolfgang Büschel, Anke Lehmann, and Raimund Dachsel. Miria: A mixed reality toolkit for the in-situ visualization and analysis of spatio-temporal interaction data. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. ACM. ISBN 9781450380966. doi: 10.1145/3411764.3445651. URL <https://doi.org/10.1145/3411764.3445651>.
- [64] Patrick Reipschlag, Tamara Flemisch, and Raimund Dachsel. Personal augmented reality for information visualization on large interactive displays. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1182–1192, 2021. ISSN 19410506. doi: 10.1109/TVCG.2020.3030460.
- [65] Tahir Mahmood, Erik Butler, Nicholas Davis, Jian Huang, and Aidong Lu. Building multiple coordinated spaces for effective immersive analytics through distributed cognition. In *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)*, pages 1–11. IEEE, 2018. doi: 10.1109/BDVA.2018.8533893.
- [66] Benjamin Lee, Xiaoyun Hu, Maxime Cordeil, Arnaud Prouzeau, Bernhard Jenny, and Tim Dwyer. Shared surfaces and spaces: Collaborative data visualisation in a co-located immersive environment. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1171–1181, 2021. ISSN 19410506. doi: 10.1109/TVCG.2020.3030450.
- [67] Simon Butscher, Sebastian Hubenschmid, Jens Müller, Johannes Fuchs, and Harald Reiterer. Clusters, trends, and outliers: How immersive technologies can facilitate the collaborative analysis of multidimensional data. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA, 2018. ACM. ISBN 9781450356206. doi: 10.1145/3173574.3173664. URL <https://doi.org/10.1145/3173574.3173664>.
- [68] Arthur Nishimoto and Andrew E Johnson. Extending virtual reality display wall environments using augmented reality. In *Symposium on Spatial User Interaction*, SUI '19, New York, NY, USA, 2019. ACM. ISBN 9781450369756. doi: 10.1145/3357251.3357579. URL <https://doi.org/10.1145/3357251.3357579>.

- [69] Verena Biener, Daniel Schneider, Travis Gesslein, Alexander Otte, Bastian Kuth, Per Ola Kristensson, Eyal Ofek, Michel Pahud, and Jens Grubert. Breaking the Screen: Interaction across Touchscreen Boundaries in Virtual Reality for Mobile Knowledge Workers. Technical Report 12, 2020.
- [70] Nadia Fereydooni and Bruce N Walker. Virtual Reality as a Remote Workspace Platform : Opportunities and Challenges. Technical report, 2020.
- [71] Eyal Ofek, Jens Grubert, Michel Pahud, Mark Phillips, and Per Ola Kristensson. Towards a Practical Virtual Office for Mobile Knowledge Workers. Technical report, 2020. URL <http://arxiv.org/abs/2009.02947>.
- [72] Adrian H Hoppe, Florian van de Camp, and Rainer Stiefelhagen. Enabling interaction with arbitrary 2d applications in virtual environments. In *International Conference on Human-Computer Interaction*, pages 30–36. Springer, 2020.
- [73] Barrett Ens, Juan David Hincapié-Ramos, and Pourang Irani. Ethereal planes: a design framework for 2d information space in 3d mixed reality environments. In *Proceedings of the 2nd ACM symposium on Spatial user interaction*, pages 2–12, 2014.
- [74] Mark McGill, Aidan Kehoe, Euan Freeman, and Stephen Brewster. Expanding the Bounds of Seated Virtual Workspaces. *ACM Transactions on Computer-Human Interaction*, 27(3), 2020. ISSN 15577325. doi: 10.1145/3380959.
- [75] Leonardo Pavanatto, Chris North, Doug A. Bowman, Carmen Badea, and Richard Stoakley. Do we still need physical monitors? An evaluation of the usability of AR virtual monitors for productivity work. *Proceedings - 2021 IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2021*, pages 759–767, 2021. doi: 10.1109/VR50410.2021.00103.
- [76] Anand Agarwala; Jinha Lee. Spatial.io, 2021. URL <https://spatial.io/>.
- [77] Santtu Parikka and Juha Ruistola. Glue Collaboration, 2021. URL <https://glue.work/>.
- [78] Alex Howland. Virbela, 2021. URL <https://www.virbela.com/>.
- [79] David Whelan. engage vr, 2021. URL <https://engagevr.io/>.

- [80] Chris Anton and Rasmus Larsen. meetinvr, 2021. URL <https://www.meetinvr.com/>.
- [81] Meta. Infinite office, September 2022. URL <https://www.workplace.com/metaverse-work-infinite-office>.
- [82] Meta. Infinite office, 2020. URL <https://www.youtube.com/watch?v=ES9jArHRFHQ>.
- [83] Adam Glazier. Designing screen interfaces for vr (google i/o '17), 2017. URL <https://www.youtube.com/watch?v=ES9jArHRFHQ>.
- [84] Marco Cavallo, Mishal Dholakia, Matous Havlena, Kenneth Ocheltree, and Mark Podlaseck. Dataspace: A reconfigurable hybrid reality environment for collaborative information analysis. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 145–153. IEEE, 2019. doi: 10.1109/VR.2019.8797733.
- [85] Alessandro Febretti, Arthur Nishimoto, Terrance Thigpen, Jonas Talandis, Lance Long, J. D. Pirtle, Tom Peterka, Alan Verlo, Maxine Brown, Dana Plepys, Dan Sandin, Luc Renambot, Andrew Johnson, and Jason Leigh. CAVE2: a hybrid reality environment for immersive simulation and information analysis. *The Engineering Reality of Virtual Reality 2013*, 8649(March 2013):864903, 2013. ISSN 0277786X. doi: 10.1117/12.2005484.
- [86] Jay S. Efran. Looking for approval: Effects on visual behavior of approbation from persons differing in importance. *Journal of Personality and Social Psychology*, 10(1):21–25, 1968. doi: 10.1037/h0026383. URL <https://doi.org/10.1037/h0026383>.
- [87] Roel Vertegaal, Ivo Weevers, Changuk Sohn, and Chris Cheung. Gaze-2: Conveying eye contact in group video conferencing using eye-controlled camera direction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, page 521–528, New York, NY, USA, 2003. ACM. ISBN 1581136307. doi: 10.1145/642611.642702. URL <https://doi.org/10.1145/642611.642702>.
- [88] David Nguyen and John Canny. Multiview: Spatially faithful group video conferencing. page 799–808, 2005. doi: 10.1145/1054972.1055084. URL <https://doi.org/10.1145/1054972.1055084>.

- [89] Bertrand Schneider and Roy Pea. Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-Supported Collaborative Learning*, 8(4):375–397, 2013. ISSN 15561607. doi: 10.1007/s11412-013-9181-4.
- [90] Allison Jing, Kieran William May, Mahnoor Naeem, Gun Lee, and Mark Billinghurst. Eyemr-vis: Using bi-directional gaze behavioural cues to improve mixed reality remote collaboration. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2021. ACM. ISBN 9781450380959. URL <https://doi.org/10.1145/3411763.3451844>.
- [91] Gun A. Lee, Seungwon Kim, Youngho Lee, Arindam Dey, Thammathip Piumsomboon, Mitchell Norman, and Mark Billinghurst. Improving Collaboration in Augmented Video Conference using Mutually Shared Gaze. In Robert W. Lindeman, Gerd Bruder, and Daisuke Iwai, editors, *ICAT-EGVE 2017 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*. The Eurographics Association, 2017. ISBN 978-3-03868-038-3. doi: 10.2312/egve.20171359.
- [92] Yuan Li, Feiyu Lu, Wallace S Lages, and Doug Bowman. Gaze direction visualization techniques for collaborative wide-area model-free augmented reality. In *Symposium on Spatial User Interaction, SUI '19*, New York, NY, USA, 2019. ACM. ISBN 9781450369756. doi: 10.1145/3357251.3357583. URL <https://doi.org/10.1145/3357251.3357583>.
- [93] Maxime Cordeil, Tim Dwyer, Karsten Klein, Bireswar Laha, Kim Marriott, and Bruce H. Thomas. Immersive Collaborative Analysis of Network Connectivity: CAVE-style or Head-Mounted Display? *IEEE Transactions on Visualization and Computer Graphics*, 23(1):441–450, 2017. ISSN 10772626. doi: 10.1109/TVCG.2016.2599107.
- [94] Mike Fraser, Steve Benford, Jon Hindmarsh, and Christian Heath. Supporting awareness and interaction through collaborative virtual interfaces. In *Proceedings of the 12th Annual ACM Symposium on User Interface Software and Technology, UIST '99*, page 27–36, New York, NY, USA, 1999. ACM. ISBN 1581130759. doi: 10.1145/320719.322580. URL <https://doi.org/10.1145/320719.322580>.

- [95] Yin Li, Alireza Fathi, and James M. Rehg. Learning to predict gaze in egocentric video. In *2013 IEEE International Conference on Computer Vision*, pages 3216–3223, 2013. doi: 10.1109/ICCV.2013.399.
- [96] Rowel Atienza, Ryan Blonna, Maria Isabel Saldares, Joel Casimiro, and Vivencio Fuentes. Interaction techniques using head gaze for virtual reality. In *2016 IEEE Region 10 Symposium (TENSYP)*, pages 110–114. IEEE, 2016. doi: 10.1109/TENCONSpring.2016.7519387.
- [97] Michael Prilla. "i simply watched where she was looking at": Coordination in short-term synchronous cooperative mixed reality. *Proc. ACM Hum.-Comput. Interact.*, 3(GROUP), dec 2019. doi: 10.1145/3361127. URL <https://doi.org/10.1145/3361127>.
- [98] Allison Jing, Kieran William May, Mahnoor Naeem, Gun Lee, and Mark Billingham. Eyemr-vis: Using bi-directional gaze behavioural cues to improve mixed reality remote collaboration. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA, 2021. ACM. ISBN 9781450380959. doi: 10.1145/3411763.3451844. URL <https://doi.org/10.1145/3411763.3451844>.
- [99] Jeff Pelz, Mary Hayhoe, and Russ Loeber. The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research*, 139(3):266–277, 8 2001. ISSN 00144819. doi: 10.1007/s002210100745. URL <http://link.springer.com/10.1007/s002210100745>.
- [100] B. Biguer, M. Jeannerod, and C. Prablanc. The coordination of eye, head, and arm movements during reaching at a single visual target. *Experimental Brain Research*, 46(2):301–304, 5 1982. ISSN 0014-4819. doi: 10.1007/BF00237188. URL <http://link.springer.com/10.1007/BF00237188>.
- [101] Frederic Kaplan and Verena V Hafner. The challenges of joint attention. *Interaction Studies*, 7(2):135–169, 2006.
- [102] Michael Tomasello. Joint attention as social cognition. In *Joint attention*, pages 103–130. Psychology Press, 2014.

- [103] Arthur M Glenberg, Jennifer L Schroeder, and David A Robertson. Averting the gaze disengages the environment and facilitates remembering. *Memory & cognition*, 26(4): 651–658, 1998.
- [104] Johannes Zagermann, Ulrike Pfeil, Roman Rädle, Hans-Christian Jetter, C. Klokmoose, and Harald Reiterer. When tablets meet tabletops: The effect of tabletop size on around-the-table collaboration with personal tablets. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5470–5481, 2016. doi: 10.1145/2858036.2858224. URL <https://dx.doi.org/10.1145/2858036.2858224>.
- [105] K. Stark, A. Shah, Jacob Borgman, Miko Somborac, J. Carson, Lauren Hauser, Krishna Kola, and Hermant Virkar. Data science, analytics and collaboration for a biosurveillance ecosystem. *Online Journal of Public Health Informatics*, 11(1):e127, 2019. doi: 10.5210/OJPHI.V11I1.9702. URL <https://dx.doi.org/10.5210/OJPHI.V11I1.9702>.
- [106] Andrew F. Monk, John F. Gale, and Lynn E. Dunlop. A look is worth a thousand words: Full gaze awareness in video-mediated conversation. *Discourse Processes*, 33(3):257–278, 2004. doi: 10.1207/s15326950dp3303_4.
- [107] Amrita S. Tulshan and Sudhir Namdeorao Dhage. Survey on virtual assistant: Google assistant, siri, cortana, alexa. In Sabu M. Thampi, Oge Marques, Sri Krishnan, Kuan-Ching Li, Domenico Ciuonzo, and Maheshkumar H. Kolekar, editors, *Advances in Signal Processing and Intelligent Recognition Systems*, pages 190–201, Singapore, 2019. Springer Singapore. ISBN 978-981-13-5758-9. doi: http://dx.doi.org/10.1007/978-981-13-5758-9_17.
- [108] Sean Li and Xiaojun (Jenny) Yuan. A review of the current intelligent personal agents. In Constantine Stephanidis, editor, *HCI International 2018 – Posters’ Extended Abstracts*, pages 253–257, Cham, 2018. Springer International Publishing. ISBN 978-3-319-92270-6. doi: 10.1007/978-3-319-92270-6_35.
- [109] Gustavo López, Luis Quesada, and Luis A. Guerrero. Alexa vs. siri vs. cortana vs. google assistant: A comparison of speech-based natural user interfaces. In Isabel L. Nunes,

- editor, *Advances in Human Factors and Systems Interaction*, pages 241–250, Cham, 2018. Springer International Publishing. ISBN 978-3-319-60366-7. doi: http://dx.doi.org/10.1007/978-3-319-60366-7_23.
- [110] Alexander Gruenstein, Bo-June Paul Hsu, James Glass, Stephanie Seneff, Lee Hetherington, Scott Cyphers, Ibrahim Badr, Chao Wang, and Sean Liu. A multimodal home entertainment interface via a mobile device. In *Proceedings of the ACL-08: HLT Workshop on Mobile Language Processing*, pages 1–9, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/W08-0801>.
- [111] Lina Zhou, Mohammedammar Shaikh, and Dongsong Zhang. Natural language interface to mobile devices. In Zhongzhi Shi and Qing He, editors, *Intelligent Information Processing II*, pages 283–286, Boston, MA, 2005. Springer US. ISBN 978-0-387-23152-5. doi: http://dx.doi.org/10.1007/0-387-23152-8_37.
- [112] David L. Waltz. An english language question answering system for a large relational database. *Commun. ACM*, 21(7):526–539, July 1978. ISSN 0001-0782. doi: 10.1145/359545.359550. URL <http://doi.acm.org/10.1145/359545.359550>.
- [113] Fei Li and Hosagrahar V Jagadish. Nalir: An interactive natural language interface for querying relational databases. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 709–712, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2376-5. doi: 10.1145/2588555.2594519. URL <http://doi.acm.org/10.1145/2588555.2594519>.
- [114] Jürgen M. Janas. The semantics-based natural language interface to relational databases. In Leonard Bolc and Matthias Jarke, editors, *Cooperative Interfaces to Information Systems*, pages 143–188, Berlin, Heidelberg, 1986. Springer Berlin Heidelberg. ISBN 978-3-642-82815-7. doi: http://dx.doi.org/10.1007/978-3-642-82815-7_6.
- [115] Shinya Kikuchi and Partha Chakroborty. Car-following model based on fuzzy inference system. *Transportation Research Record*, pages 82–91, 1992. URL <http://onlinepubs.trb.org/Onlinepubs/trr/1992/1365/1365-009.pdf>.

- [116] Niels Ole Bernsen and Laila Dybkjaer. Exploring natural interaction in the car. In *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue*, pages 75–79, Denmark, 2001. University of Southern Denmark.
- [117] J. H. L. Hansen, Rongqing Huang, P. Mangalath, Bowen Zhou, M. Seadle, and J. R. Deller. Speechfind: spoken document retrieval for a national gallery of the spoken word. In *Proceedings of the 6th Nordic Signal Processing Symposium, 2004. NORSIG 2004.*, pages 1–4, Espoo, Finland, 2004. IEEE. doi: <http://dx.doi.org/10.1109/TSA.2005.852088>.
- [118] Masataka Goto, Jun Ogata, and Kouichirou Eto. Podcastle: a web 2.0 approach to speech recognition research. In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, pages 2397–2400, Antwerp, Belgium, 2007. Curran Associates, Inc. URL http://www.isca-speech.org/archive/interspeech_2007/i07_2397.html.
- [119] James R. Glass, Timothy J. Hazen, D. Scott Cyphers, Igor Malioutov, David Huynh, and Regina Barzilay. Recent progress in the MIT spoken lecture processing project. In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, pages 2553–2556, 2007. URL http://www.isca-speech.org/archive/interspeech_2007/i07_2553.html.
- [120] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan. An audio indexing system for election video material. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4873–4876, Taipei, Taiwan, April 2009. IEEE. doi: 10.1109/ICASSP.2009.4960723.
- [121] Lin shan Lee Sheng-yi Kong, Miao ru Wu, Che kuang Lin, Yi sheng Fu, Yungyu Chung, Yu Huang, and Yun-Nung Chen. Ntu virtual instructor - a spoken language system offering services of learning on demand using video/audio/slides of course lectures. *ICASSP*, 2009.
- [122] Scott McGlashan. Speech interfaces to virtual reality, 1995.

- [123] Scott McGlashan and Tomas Axling. A speech interface to virtual environments, 1996.
- [124] J. Muller, C. Krapichler, Lam Son Nguyen, K. Hans Englmeier, and M. Lang. Speech interaction in virtual reality. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 6, pages 3757–3760 vol.6, 1998. doi: 10.1109/ICASSP.1998.679701.
- [125] Denis V. Dorozhkin and Judy M. Vance. Implementing Speech Recognition in Virtual Reality. volume Volume 1: 22nd Computers and Information in Engineering Conference of *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pages 61–65, 09 2002. doi: 10.1115/DETC2002/CIE-34390.
- [126] JM Noyes. Speech technology in the future. In *Interactive speech technology: Human factors issues in the application of speech input/output to computers*, pages 189–208. Taylor & Francis London, 1993.
- [127] C Baber and RB Stammers. Is it natural to talk to computers: an experiment using the wizard of oz technique. *ED Megaw,(Ed.), Contemporary Ergonomics 1989*, 1989.
- [128] Minkyung Lee and Mark Billinghurst. A wizard of oz study for an ar multimodal interface. In *Proceedings of the 10th International Conference on Multimodal Interfaces, ICMI '08*, page 249–256, New York, NY, USA, 2008. ACM. ISBN 9781605581989. doi: 10.1145/1452392.1452444. URL <https://doi.org/10.1145/1452392.1452444>.
- [129] Daniele Giunchi, Alejandro Sztrajman, Stuart James, and Anthony Steed. Mixing modalities of 3d sketching and speech for interactive model retrieval in virtual reality. In *ACM International Conference on Interactive Media Experiences, IMX '21*, page 144–155, New York, NY, USA, 2021. ACM. ISBN 9781450383899. doi: 10.1145/3452918.3458806. URL <https://doi.org/10.1145/3452918.3458806>.
- [130] Alexander Winstan Stedmon. *Putting Speech in, taking speech out: human factors in the use of speech interfaces*. PhD thesis, University of Nottingham, 2005.
- [131] C. Baber, B. Mellor, R. Graham, J.M. Noyes, and C. Tunley. Workload and the use of automatic speech recognition: The effects of time and resource demands. *Speech*

- Communication*, 20(1):37–53, 1996. ISSN 0167-6393. doi: [https://doi.org/10.1016/S0167-6393\(96\)00043-X](https://doi.org/10.1016/S0167-6393(96)00043-X). URL <https://www.sciencedirect.com/science/article/pii/S016763939600043X>.
- [132] Alex W Stedmon and Chris Baber. Evaluating stress in the development of speech interface technology. In *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces- Volume I-Volume I*, pages 545–549, 1999.
- [133] Alex W. Stedmon, Harshada Patel, Sarah C. Sharples, and John R. Wilson. Developing speech input for virtual reality applications: A reality based interaction approach. *International Journal of Human-Computer Studies*, 69(1):3–8, 2011. ISSN 1071-5819. doi: <https://doi.org/10.1016/j.ijhcs.2010.09.002>. URL <https://www.sciencedirect.com/science/article/pii/S1071581910001217>.
- [134] Niels ter Heijden and Willem-Paul Brinkman. Design and evaluation of a virtual reality exposure therapy system with automatic free speech interaction. *Journal of CyberTherapy and Rehabilitation*, 4(1):41–55, 2011.
- [135] Luís Coroado, Tiago Pedro, Jorge D’Alpuim, Sara Eloy, and MiguelSales Dias. Viar-modes: visualization and interaction in immersive virtual reality for architectural design process. 2015.
- [136] Aziz Siyaev and Geun-Sik Jo. Towards aircraft maintenance metaverse using speech interactions with virtual objects in mixed reality. *Sensors*, 21(6), 2021. ISSN 1424-8220. doi: 10.3390/s21062066. URL <https://www.mdpi.com/1424-8220/21/6/2066>.
- [137] Google. Speech to text, September 2022. URL <https://cloud.google.com/speech-to-text>.
- [138] Microsoft. Azure cognitive services, September 2022. URL <https://azure.microsoft.com/>.
- [139] Google. Dialogflow, September 2022. URL <https://dialogflow.cloud.google.com>.

- [140] IBM. Watson, September 2022. URL <https://www.ibm.com/watson>.
- [141] Amazon. Aws transcribe, September 2022. URL <https://aws.amazon.com/transcribe/>.
- [142] Thammathip Piumsomboon, Arindam Dey, Barrett Ens, Gun Lee, and Mark Billinghurst. The effects of sharing awareness cues in collaborative mixed reality. *Frontiers Robotics AI*, 6(FEB), 2019. ISSN 22969144. doi: 10.3389/frobt.2019.00005.
- [143] Riccardo Bovo, Daniele Giunchi, Enrico Costanza, and Anthony Steed. Shall I describe it or shall I move closer ? Verbal references and locomotion in VR collaborative search tasks . 2022. doi: 10.48340/ecscw2022.
- [144] Maurício Sousa, Rafael Kuffner Dos Anjos, Daniel Mendes, Mark Billinghurst, and Joaquim Jorge. Warping deixis: Distorting Gestures to Enhance Collaboration. In *Conference on Human Factors in Computing Systems - Proceedings*, volume 12, pages 1–12, New York, NY, USA, 5 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300838. URL <https://dl.acm.org/doi/10.1145/3290605.3300838>.
- [145] Sven Mayer, Valentin Schwind, Robin Schweigert, and Niels Henze. The effect of offset correction and cursor on mid-air Pointing in real and virtual environments. *Conference on Human Factors in Computing Systems - Proceedings*, 2018-April:1–13, 2018. doi: 10.1145/3173574.3174227.
- [146] Chris L. Schmidt. Adult Understanding of Spontaneous Attention-Directing Events: What Does Gesture Contribute? In *Ecological Psychology*, volume 11, pages 139–174, 1999. doi: 10.1207/s15326969eco1102{_}2.
- [147] Adrian Bangerter and Daniel M. Oppenheimer. Accuracy in detecting referents of pointing gestures unaccompanied by language. *Gesture*, 6(1):85–102, 2006. ISSN 1568-1475. doi: 10.1075/gest.6.1.05ban.
- [148] Anil Ufuk Batmaz and Wolfgang Stuerzlinger. Effects of 3d rotational jitter and selection methods on 3d pointing tasks. In *26th IEEE Conference on Virtual Reality and 3D User*

- Interfaces, VR 2019 - Proceedings*, pages 1687–1692, 2019. ISBN 9781728113777. doi: 10.1109/VR.2019.8798038.
- [149] Richard A. Bolt. “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, page 262–270, New York, NY, USA, 1980. Association for Computing Machinery. ISBN 0897910214. doi: 10.1145/800250.807503. URL <https://doi.org/10.1145/800250.807503>.
- [150] Darius Miniotas, Oleg Špakov, Ivan Tugoy, and I. Scott MacKenzie. Speech-Augmented Eye Gaze Interaction with Small Closely Spaced Targets. In *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications*, ETRA '06, page 67–72, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933050. doi: 10.1145/1117309.1117345. URL <https://doi.org/10.1145/1117309.1117345>.
- [151] Sven Mayer, Gierad Laput, and Chris Harrison. Enhancing Mobile Voice Assistants with WorldGaze. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 4 2020. ISBN 9781450367080. doi: 10.1145/3313831.3376479.
- [152] Karan Ahuja, Andy Kong, Mayank Goel, and Chris Harrison. Direction-of-voice (dov) estimation for intuitive speech interaction with smart devices ecosystems. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, page 1121–1131, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375146. doi: 10.1145/3379337.3415588. URL <https://doi.org/10.1145/3379337.3415588>.
- [153] Yevhen Romaniak, Anastasiia Smielova, Yevhenii Yakishyn, Valerii Dziubliuk, Mykhailo Zlotnyk, and Oleksandr Viatchaninov. 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, page 129–131, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375153. doi: 10.1145/3379350.3416153. URL <https://doi.org/10.1145/3379350.3416153>.

- [154] Sebastian Marwecki, Andrew D Wilson, Eyal Ofek, Mar Gonzalez Franco, and Christian Holz. Mise-unseen: Using eye tracking to hide virtual reality scene changes in plain sight. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology*, pages 777–789, 2019.
- [155] Mikko Kytö, Barrett Ens, Piumsomboon Thammathip, Lee Gun A., and Mark Billinghurst. Pinpointing. *Education + Training*, 23(3):88–89, 2018. ISSN 00400912. doi: 10.1108/eb016812.
- [156] Yuan Yuan Qian and Robert J. Teather. The eyes don’t have it. pages 91–98, 2017. doi: 10.1145/3131277.3132182.
- [157] Difeng Yu and Xueshi Lu. Gaze-supported 3d object manipulation in virtual reality. *Conference on Human Factors in Computing Systems - Proceedings*, 2021. doi: 10.1145/3411764.3445343.
- [158] Sunggeun Ahn, Jeongmin Son, Sangyoon Lee, and Geehyuk Lee. Verge-it: Gaze interaction for a binocular head-worn display using modulated disparity vergence eye movement. *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–7, 2020. doi: 10.1145/3334480.3382908.
- [159] Per Bækgaard, John Paulin Hansen, Katsumi Minakata, and I. Scott MacKenzie. A fitts’ law study of pupil dilations in a head-mounted display. *Eye Tracking Research and Applications Symposium (ETRA)*, 2019. doi: 10.1145/3314111.3319831.
- [160] Heiko Drewes, Mohamed Khamis, and Florian Alt. Smooth pursuit target speeds and trajectories. *ACM International Conference Proceeding Series*, pages 139–146, 2018. doi: 10.1145/3282894.3282913.
- [161] Pieter Blignaut and Daniël Wium. Eye-tracking data quality as affected by ethnicity and experimental design. *Behavior Research Methods*, 46(1):67–80, 2014. ISSN 15543528. doi: 10.3758/s13428-013-0343-0.
- [162] R John Leigh and David S Zee. *The neurology of eye movements*. Oxford university press, 2015.

- [163] Wolfgang H Zangemeister and Lawrence Stark. Types of gaze movement: variable interactions of eye and head movements. *Experimental neurology*, 77(3):563–577, 1982.
- [164] Emilio Bizzi, Ronald E Kalil, and Vincenzo Tagliasco. Eye-head coordination in monkeys: evidence for centrally patterned organization. *Science*, 173(3995):452–454, 1971.
- [165] Michael F. Land. Vision, eye movements, and natural behavior. *Visual Neuroscience*, 26(1):51–62, 2009. ISSN 09525238. doi: 10.1017/S0952523808080899.
- [166] Daniel Guitton and Michel Volle. Gaze control in humans: eye-head coordination during orienting movements to targets within and beyond the oculomotor range. *Journal of neurophysiology*, 58(3):427–459, 1987.
- [167] Kevin Pfeil, Eugene M. Taranta, Arun Kulshreshth, Pamela Wisniewski, and Joseph J. LaViola. A comparison of eye-head coordination between virtual and physical realities. *Proceedings - SAP 2018: ACM Symposium on Applied Perception*, 2018. doi: 10.1145/3225153.3225157. URL <https://doi.org/10.1145/3225153.3225157>.
- [168] Zhiming Hu, Sheng Li, Congyi Zhang, Kangrui Yi, Guoping Wang, and Dinesh Manocha. DGaze: CNN-Based Gaze Prediction in Dynamic Scenes. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1902–1911, 2020. ISSN 19410506. doi: 10.1109/TVCG.2020.2973473.
- [169] Kentaro Yamada, Yusuke Sugano, Takahiro Okabe, Yoichi Sato, Akihiro Sugimoto, and Kazuo Hiraki. Attention prediction in egocentric video using motion and visual saliency. In *Proceedings of the 5th Pacific Rim Conference on Advances in Image and Video Technology - Volume Part I*, PSIVT’11, page 277–288, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 9783642253669. doi: 10.1007/978-3-642-25367-6_25. URL https://doi.org/10.1007/978-3-642-25367-6_25.
- [170] Kenji Matsuo, Kentaro Yamada, Satoshi Ueno, and Sei Naito. An attention-based activity recognition for egocentric video. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 565–570, 2014. doi: 10.1109/CVPRW.2014.87.

- [171] Yin Li, Alireza Fathi, and James M Reh. Learning to Predict Gaze in Egocentric Video. pages 3216–3223, 2013. doi: 10.1109/ICCV.2013.399.
- [172] Ryoichi Nakashima, Yu Fang, Yasuhiro Hatori, Akinori Hiratani, Kazumichi Matsumiya, Ichiro Kuriki, and Satoshi Shioiri. Saliency-based gaze prediction based on head direction. *Vision Research*, 117:59–66, 2015. ISSN 0042-6989. doi: <https://doi.org/10.1016/j.visres.2015.10.001>. URL <https://www.sciencedirect.com/science/article/pii/S0042698915003338>.
- [173] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12919–12928, 2021.
- [174] Riccardo Bovo, Daniele Giunchi, Enrico Costanza, and Anthony Steed. Shall I describe it or shall I move closer ? Verbal references and locomotion in VR collaborative search tasks . 2022. doi: 10.48340/ecscw2022.
- [175] Kenneth Holmqvist, Marcus Nyström, and Fiona Mulvey. Eye tracker data quality: What it is and how to measure it. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, page 45–52, New York, NY, USA, 2012. ACM. ISBN 9781450312219. doi: 10.1145/2168556.2168563. URL <https://doi.org/10.1145/2168556.2168563>.
- [176] Rowel Atienza, Ryan Blonna, Maria Isabel Saludares, Joel Casimiro, and Vivencio Fuentes. Interaction techniques using head gaze for virtual reality. In *2016 IEEE Region 10 Symposium (TENSYP)*, pages 110–114, 2016. doi: 10.1109/TENCONSpring.2016.7519387.
- [177] Frank Biocca, Chad Harms, and Jenn Gregg. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. *4th Annual International Workshop on Presence*, (December):1–9, 2001. URL <http://astro.temple.edu/~lombard/ISPR/Proceedings/2001/Biocca2.pdf>.

- [178] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, 52(C):139–183, 1988. ISSN 01664115. doi: 10.1016/S0166-4115(08)62386-9.
- [179] Joan Morris Dimicco, Anna Pandolfo, and Walter Bender. Influencing Group Participation with a Shared Display. 2004.
- [180] Chris North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, 2006. ISSN 02721716. doi: 10.1109/MCG.2006.70.
- [181] Ioannis Agtzidis, Mikhail Startsev, and Michael Dorr. 360-degree video gaze behaviour: A ground-truth data set and a classification algorithm for eye movements. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM ’19, page 1007–1015, New York, NY, USA, 2019. ACM. ISBN 9781450368896. doi: 10.1145/3343031.3350947. URL <https://doi.org/10.1145/3343031.3350947>.
- [182] Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B. Pelz, and Gabriel J. Diaz. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific Reports*, 10(1):1–23, 2020. ISSN 20452322. doi: 10.1038/s41598-020-59251-5.
- [183] Ioannis Agtzidis, Mikhail Startsev, and Michael Dorr. 360-degree Video Gaze Behaviour: A ground-truth data set and a classification algorithm for eye movements. Technical report, 2019. URL <https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/>.
- [184] Alison Bechdel. Bechdel Test. https://en.wikipedia.org/wiki/Bechdel_test, 1985.
- [185] Sharon Brener. Bechdel test dataset. <https://data.world/sharon/bechdel-test/workspace/file?filename=movies.csv>, 2013.
- [186] IMDb. URL <https://www.imdb.com/interfaces/>.
- [187] Jeffrey Schlimmer. Automobile data set. <http://archive.ics.uci.edu/ml/datasets/Automobile>, 1985.

- [188] Hamed R. Tavakoli, Esa Rahtu, Juho Kannala, and Ali Borji. Digging deeper into ego-centric gaze prediction. In *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*, pages 273–282. Institute of Electrical and Electronics Engineers Inc., 3 2019. ISBN 9781728119755. doi: 10.1109/WACV.2019.00035.
- [189] Virgilio F Ferrario, Chiarella Sforza, Graziano Serrao, GianPiero Grassi, and Erio Mossi. Active range of motion of the head and cervical spine: a three-dimensional investigation in healthy young adults. *Journal of orthopaedic research*, 20(1):122–129, 2002.
- [190] Alvin R Tilley et al. Henry dreyfuss associates. *The measure of man and woman: human factors in design*, 2002.
- [191] Sohrab Saeb, Cornelius Weber, and Jochen Triesch. Learning the optimal control of coordinated eye and head movements. *PLoS Computational Biology*, 7(11), 2011. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002253.
- [192] Ludwig Sidenmark and Hans Gellersen. Eye, head and torso coordination during gaze shifts in virtual reality. *ACM Trans. Comput.-Hum. Interact.*, 27(1), dec 2019. ISSN 1073-0516. doi: 10.1145/3361218. URL <https://doi.org/10.1145/3361218>.
- [193] Marie Diener-west, Jeff Sauro, and John Brooke. SUS: A 'Quick and Dirty' Usability Scale. *Usability Evaluation In Industry*, pages 207–212, 2020. doi: 10.1201/9781498710411-35.
- [194] James R. Lewis. The System Usability Scale: Past, Present, and Future. *International Journal of Human-Computer Interaction*, 34(7):577–590, 2018. ISSN 15327590. doi: 10.1080/10447318.2018.1455307. URL <https://doi.org/10.1080/10447318.2018.1455307>.
- [195] Nelson Wong and Carl Gutwin. Where are you pointing? The accuracy of deictic pointing in CVEs. In *Conference on Human Factors in Computing Systems - Proceedings*, volume 2, pages 1029–1038, New York, New York, USA, 2010. ACM Press. ISBN 9781605589299. doi: 10.1145/1753326.1753480. URL <http://portal.acm.org/citation.cfm?doid=1753326.1753480>.

- [196] Rafael Radkowski. Investigation of visual features for augmented reality assembly assistance. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9179, pages 488–498. Springer, Cham, 2015. ISBN 9783319210667. doi: 10.1007/978-3-319-21067-4_{_}50. URL http://link.springer.com/10.1007/978-3-319-21067-4_50.
- [197] Robert J. Teather and Wolfgang Stuerzlinger. Pointing at 3D targets in a stereo head-tracked virtual environment. In *3DUI 2011 - IEEE Symposium on 3D User Interfaces 2011, Proceedings*, pages 87–94, 2011. ISBN 9781457700620. doi: 10.1109/3DUI.2011.5759222.
- [198] Immersed Company. Immersed, 2022. URL <https://immersed.com/>.
- [199] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *OpenAI Blog*, 2022.
- [200] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [201] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
- [202] Meghan E. Flath, Austen K. Smith, and Lorin J. Elias. Cultural differences in lateral biases on aesthetic judgments: The effect of native reading direction. *Culture and Brain*, 7(1):57–66, 2019. ISSN 2193-8652. doi: 10.1007/s40167-018-0062-6. URL <https://doi.org/10.1007/s40167-018-0062-6>.
- [203] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

- [204] Aaron Bangor, Philip T. Kortum, and James T. Miller. An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, 24(6): 574–594, 2008. doi: 10.1080/10447310802205776.
- [205] Allison Jing, Kieran William May, Mahnoor Naeem, Gun Lee, and Mark Billinghurst. Eyemr-vis: Using bi-directional gaze behavioural cues to improve mixed reality remote collaboration. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.
- [206] Leon Reicherts, Yvonne Rogers, Licia Capra, Ethan Wood, Tu Dinh Duong, and Neil Sebire. It’s good to talk: A comparison of using voice versus screen-based interactions for agent-assisted tasks. *ACM Trans. Comput.-Hum. Interact.*, 29(3), jan 2022. ISSN 1073-0516. doi: 10.1145/3484221. URL <https://doi.org/10.1145/3484221>.
- [207] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa.
- [208] Allison Jing, Kieran William May, Mahnoor Naeem, Gun Lee, and Mark Billinghurst. EyemR-Vis: Using Bi-Directional Gaze Behavioural Cues to Improve Mixed Reality Remote Collaboration. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM, may 2021. ISBN 9781450380959. doi: 10.1145/3411763.3451844. URL <https://doi.org/10.1145/3411763.3451844>.
- [209] Vinod Karar and Smarajit Ghosh. Attention tunneling: Effects of limiting field of view due to beam combiner frame of head-up display. *IEEE/OSA Journal of Display Technology*, 10(7):582–589, 2014. ISSN 1551319X. doi: 10.1109/JDT.2014.2311159.
- [210] J.M. Lee, B.H. Cho, J.H. Ku, J.S. Kim, J.H. Lee, I.Y. Kim, and S.I. Kim. A study on the system for treatment of adhd using virtual reality. In *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 4, pages 3754–3757 vol.4, 2001. doi: 10.1109/IEMBS.2001.1019654.
- [211] Diederick C. Niehorster, Thiago Santini, Roy S. Hessels, Ignace T.C. Hooge, Enkelejda Kasneci, and Marcus Nyström. The impact of slippage on the data quality of head-worn

- eye trackers. *Behavior Research Methods*, 52(3):1140–1160, 2020. ISSN 15543528. doi: 10.3758/s13428-019-01307-0.
- [212] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. doi: 10.1109/CVPR.2016.91.
- [213] Zhiming Hu, Congyi Zhang, Sheng Li, Guoping Wang, and DInesh Manocha. SGaze: A Data-Driven Eye-Head Coordination Model for Realtime Gaze Prediction. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2002–2010, 2019. ISSN 19410506. doi: 10.1109/TVCG.2019.2899187.
- [214] Jae-Young Lee, Hyung-Min Park, Seok-Han Lee, Soon-Ho Shin, Tae-Eun Kim, and Jong-Soo Choi. Design and implementation of an augmented reality system using gaze interaction. *Multimedia Tools and Applications*, 68(2):265–280, 2014.
- [215] Alexander Plopski, Teresa Hirzle, Nahal Norouzi, Long Qian, Gerd Bruder, and Tobias Langlotz. The eye in extended reality: A survey on gaze interaction and eye tracking in head-worn extended reality. *ACM Computing Surveys (CSUR)*, 55(3):1–39, 2022.
- [216] Baihan Cai, Shaoyao Zhang, Bo Wang, Xinlian Lu, Duming Wang, Chaoyue Qi, Xinqi Zhang, Meng Chen, Chunhui Wang, and Yu Tian. The magnetic cursor pointing technique (magcursor): A gaze-assisted manual pointing technique based on human eye-hand coordination patterns. *International Journal of Industrial Ergonomics*, 107:103760, 2025.
- [217] James A Easterbrook. The effect of emotion on cue utilization and the organization of behavior. *Psychological review*, 66(3):183, 1959.
- [218] Christopher D Wickens and Amy L Alexander. Attentional tunneling and task management in synthetic vision displays. *The international journal of aviation psychology*, 19(2):182–199, 2009.
- [219] Robert JK Jacob. What you look at is what you get: eye movement-based interaction techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 11–18, 1990.

- [220] Päivi Majaranta and Andreas Bulling. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*, pages 39–65. Springer, 2014.
- [221] Corinna Zeiser. *Processes in hybrid teams: Analyzing the relationship between availability and team performance*. PhD thesis, Technische Universität Wien, 2025.
- [222] Matheus Argôlo, Carlos Eduardo Barbosa, Lucas Nóbrega, Luiz Felipe Martinez, Yuri Oliveira, Jano de Souza, et al. Challenges and practices for knowledge transfer in global software development: An updated literature review. In *European Conference on Knowledge Management*, volume 1, pages 87–96. Academic Conferences International Limited, 2025.
- [223] Yiqing Shen and Dell Zhang. A survey of language-guided video object segmentation: from referring to reasoning. *Vicinagearth*, 2(1):1–20, 2025.
- [224] Danfeng Zhuang, Min Jiang, Hichem Arioui, and Hedi Tabia. Strengthening temporal action segmentation through diffusion models. *Engineering Applications of Artificial Intelligence*, 164:113163, 2026.
- [225] Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, et al. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8120–8132, 2025.