RESEARCH-ARTICLE

VRCaptions: Design Captions for DHH Users in Multiplayer Communication in VR

TIANZE XIE, Southern University of Science and Technology, Shenzhen, Guangdong, China

XUESONG ZHANG, Southern University of Science and Technology, Shenzhen, Guangdong, China

FEIYU HUANG, Southern University of Science and Technology, Shenzhen, Guangdong, China

DI LIU, Southern University of Science and Technology, Shenzhen, Guangdong, China

PENGCHENG AN, Southern University of Science and Technology, Shenzhen, Guangdong, China

SEUNGWOO JE, Southern University of Science and Technology, Shenzhen, Guangdong, China

Open Access Support provided by:

Southern University of Science and Technology

Published: 26 April 2025

CHI 2025: CHI Conference on Human Factors in Computing Systems
April 26 - May 1, 2025
Yokohama, Japan

Conference Sponsors:
SIGCHI

https://doi.org/10.1145/3706598.3714186
ISBN: 9798400713941

VRCaptions: Design Captions for DHH Users in Multiplayer Communication in VR

Tianze Xie*
Southern University of Science and Technology
Shenzhen, China
tiaraaaxie@gmail.com

Xuesong Zhang*
Southern University of Science and Technology
Shenzhen, China
hcisong@gmail.com

Feiyu Huang
Southern University of Science and Technology
Shenzhen, China
huang.feiyu@outlook.com

Di Liu
School of Design
Southern University of Science and Technology
Shenzhen, China
seucliudi@gmail.com

Pengcheng An
School of Design
Southern University of Science and Technology
Shenzhen, China
anpc@sustech.edu.cn

Seungwoo Je†
Southern University of Science and Technology
Shenzhen, China
seungwoo@sustech.edu.cn

Abstract

Accessing auditory information remains challenging for DHH individuals in real-world situations and multiplayer VR interactions. To improve this, we investigated caption designs that specialize in the needs of DHH users in multiplayer VR settings. First, we conducted three co-design workshops with DHH participants, social workers, and designers to gather insights into the specific needs of design directions for DHH users in the context of a room escape game in VR. We further refined our designs with 13 DHH users to determine the most preferred features. Based on this, we developed VRCaptions, a caption prototype for DHH users to better experience multiplayer conversations in VR. We lastly invited two mixed-hearing groups to participate in the VR room escape game with our VRCaptions to validate. The results demonstrate that VRCaptions can enhance the ability of DHH participants to access information and reduce the barrier to communication in VR.

CCS Concepts

• **Human-centered computing** → **Accessibility**.

Keywords

Accessibility, Communication, Virtual Reality, Deaf and Hard of Hearing, Caption Design

ACM Reference Format:
Tianze Xie, Xuesong Zhang, Feiyu Huang, Di Liu, Pengcheng An, and Seungwoo Je. 2025. VRCaptions: Design Captions for DHH Users in Multiplayer Communication in VR. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3714186

*Both authors contributed equally to this research.
†The corresponding author

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714186

1 Introduction

Deaf and Hard of Hearing (DHH) individuals often encounter obstacles in accessing crucial auditory information, such as spoken language, during group conversations in real-world settings. Real-time captioning serves as an essential communication aid, enabling DHH individuals to comprehend speech information better during group conversations [28, 50, 51]. In remote conversations, DHH individuals not only struggle to follow the content of the conversation through speech information accurately but also need to navigate the challenge of accessing non-speech information, such as emotion [18], emphasis [27], and sound awareness [35, 58] which convey critical conversation information and are essential for active communication in multiplayer conversations.

Technologies like video conferencing, Augmented Reality (AR), and Virtual Reality (VR) enhance the accessibility and communication of DHH individuals in remote multiplayer conversations, which has been explored in previous studies [40, 48, 50, 62, 73]. In the context of video conferencing, previous research has focused on the integration of non-speech elements, such as emotions [18, 28, 40], background noise [50], and speaker recognition [51], to rich the real-time captions with additional contextual information beyond speech for facilitating more comprehensive communication during remote conversation. AR has also demonstrated potential by combining caption design with speaking order prompts and speaker position display and other means [58, 63, 64], offering a more dynamic and interactive conversation.

In VR and 360° video, sound not only conveys essential information but also enhances communication through non-speech information [13, 30, 38, 50, 59]. Previous studies have explored ways to improve DHH users understanding of speech information in VR, such as using smart gloves [72] and live sign language interpretation [57]. Additionally, there have been efforts to help DHH users perceive various sounds by integrating visualizations of non-speech information in VR [37, 38, 46, 47]. However, the challenge of providing DHH users with captions that enable effective access to both the content of the conversation and non-speech

information during VR multiplayer conversations remains unexplored.

To support DHH users in accessible communication within the complex auditory environment of multiplayer VR conversations through captions, we proposed the VRCaptions system prototype and followed a user-centered design (UCD) principle.

We first conducted a literature review to identify DHH users' key challenges and needs when they were involved in remote or small group conversations. This was followed by three co-design workshops to verify design directions and generate ideas with the input of DHH participants, social workers, and designers. Throughout this process, seven design directions (*Caption Display Position, Speech Overlapping, Speaking Order, Caption Delay, Chat History, Speaker Identification,* and *Sound Location*) among three design dimensions (*Readability, Speech Information Transmission,* and *Non-speech Information Transmission*) were taken into consideration. We then summarized these design directions and conducted 13 semi-structured interviews with DHH participants to gain insights into their preferences. Based on these findings, we developed a caption system prototype, VRCaptions, and integrated it into a room-escape game for validation. Finally, two mixed-hearing groups were invited to experience the game and share their feedback on the VRCaptions system design, along with insights for future multiplayer conversations in VR.

The contributions of our work are: (1) We proposed design directions for the inclusive and accessible caption system for DHH individuals in VR multiplayer conversations based on three co-design workshops and 13 interviews; (2) We developed a caption system prototype, VRCaptions, which focuses on the experience of DHH individuals in VR multiplayer conversations and validates it in the game context; (3) By summarizing the accessibility needs and preferences of DHH users for the that we collected, we put forward suggestions for future caption design of DHH individuals in VR multiplayer conversation.

2 Related Work

In this section, we review prior work related to two main areas: (1) the design and analysis of communication for DHH individuals in both real and remote conversations and (2) the accessible design of better communication for DHH individuals in Extended Reality (XR).

2.1 Communication for DHH Individuals in Real-world and Remote Conversations

Participating in multiplayer conversations can be challenging for DHH individuals, whether in the real world or in remote collaboration. In real-world conversations, DHH individuals often rely on technologies such as real-time captions [8, 23, 50, 65], sign language translation [22], speechreading [19], or translation gloves [60, 72] to follow the conversation. In remote settings, understanding the context of the conversation through voice information is crucial for DHH individuals, who also have to contend with the absence of non-speech cues [2, 32, 40, 41, 66]. In remote multiplayer conversations, DHH individuals must simultaneously process visual information such as sign language translation, captions, and other users' body language, making it more likely for them to miss important details [43, 52, 68].

Previous studies have extensively examined the access needs of DHH individuals in remote conversations in order to address these challenges [9, 40, 41, 43–45, 50, 51, 62, 70]. For instance, Ang et al. conducted a study on the usage and design preferences of video conferencing tools by deaf signers and sign language interpreters [62]. Their findings demonstrated that existing video conferencing platforms do not sufficiently meet the DHH users' requirements, including information acquisition, communication, and collaboration. Lacerda et al. explored the perspectives of DHH users on incorporating in-speech emotional cues through caption text [18]. They concluded that a thoughtful caption design can effectively convey more information to users beyond just the text content. Additionally, Vogler et al. discovered that significant delays in online interpretation impede DHH users from engaging in remote conversations [70]. Furthermore, Kushalnagar and Vogler outlined the challenges that DHH users may face in video or telephone conferences [45]. They provided practical recommendations for making video conferencing more accessible for DHH users, including conversation guides, transcripts, and other necessary accommodations.

With the development of technology, extensive research has been dedicated to remote conversations. However, as XR technology continues to advance, remote collaborations using AR or VR are becoming increasingly prevalent. Therefore, there is a growing need to develop inclusive and accessible captions for XR environments to facilitate the participation of DHH individuals in multiplayer conversations.

2.2 Caption Design for DHH Individuals in XR

The integration of XR technologies, especially AR, with real-time speech and sound recognition technologies, has gained interest among researchers aiming to enhance accessibility for DHH individuals.

One research direction in AR applications is combining sound-related context with real-time captioning to support accessible information for DHH users [27, 34, 35, 48, 54, 58, 63, 64, 68, 73]. Olwal et al. demonstrates the feasibility of merging real-time sound transcription with AR in a lightweight design to offer all-day captioning, providing a continuous, accessible communication solution for DHH users in the daily life [54]. Additionally, AR-based caption systems can also be effectively applied in educational settings to enhance the learning experiences of DHH students, which could significantly improve their learning speed and effectiveness [63]. Meanwhile, Jain et al. focused on applications in moving contexts with AR headsets and introduced four design guidelines for HMD-based captioning, including aligning text with the speaker, adapting captions to dynamic contexts, disabling the wearer's voice, providing contextual information, and supporting user customization [36]. Moreover, Peng et al. proposed *SpeechBubbles*, a captioning system specifically designed for group conversations in AR,

which aids DHH users in identifying speakers through a "text bubbles" design [58]. Furthermore, in terms of sound source localization, Guo et al. developed *HoloSound*, a system that can identify sounds and visualize their locations using circular arcs, which enhance sound awareness for DHH users with HoloLens [27].

Unlike integrating AR techniques to enhance accessibility for DHH individuals in daily life, education context, or conversational scenarios, accessing sound information in a VR setting presents accessibility challenges for DHH individuals [37, 38, 46, 53, 69]. Since sound in immersive environments is different from the real world, researchers have come up with different solutions. For the 360° videos, Brown et al.'s research investigated the optimal way to display captions, which follows the user's head movements to enhance user experience while maintaining accessibility [12, 13]. In VR, Mirzaei et al. proposed *EarVR*, which analyses the 3D audio in a virtual environment and notifies DHH users of the sound location through haptic feedback on the ear with vibro-motors [53]. Furthermore, to better design VR sounds and their accessibility, researchers also develop a VR sound taxonomy that categorizes sounds into sound source and intent and also suggest mapping sound to visual and haptic feedback to improve its accessibitlity [37, 38]. In addition, Li et al. proposed *SoundVizVR* to improve VR experiences for DHH users by visually representing sound loudness, duration, and location through on-object indicators, full mini-maps, and partial mini-maps [46]. DHH users preferred combining full mini-map and object techniques together, which also helped them perform best in locating sound sources.

However, when multiplayer is involved in VR conversation scenarios, how to improve accessibility and enhance their communication remains underexplored. Our study extends the investigation to understand DHH users' needs and preferences for better comprehension of the context through caption design, which could facilitate more inclusive and accessible conversation for DHH individuals in VR multiplayer conversation.

3 General Approach

During the design process, we adhered to the UCD principle [1], with a specific focus on participatory design. The design procedure consisted of four phases:

(1) **Understanding User Needs with Co-design Workshops**: We conducted a literature review and three co-design workshops involving DHH users, social workers, and designers to ensure that outcomes align with the needs and contexts of DHH users.
(2) **Refining Design Directions**: From insights gathered in the initial co-design workshops, we refined design directions that reflect DHH user needs and preferences.
(3) **Selecting Design with Semi-structured Interviews**: We adopted participatory design by actively engaging users not only as sources of information but also as decision-makers [67]. Thirteen semi-structured interviews were conducted to collect DHH users' preferences on these designs and to confirm the design choices.
(4) **Validating Design with User Study**: The final phase involved a user study to validate the feasibility of the selected design. Feedback from this phase can be used to refine the design further.

4 Co-design Workshops for Caption Designs

In this study, we aimed to explore the caption design for DHH users to enhance their communication experiences for multiplayer conversation in VR. Firstly, we identified seven design directions based on a literature review to address the unique needs of DHH users. Then, we conducted three co-design workshops with DHH users, social workers, and designers to gather empirical insights from diverse perspectives. During the workshop, participants engaged with a VR game that lacked accommodations for DHH users. After interacting, the seven proposed design directions were discussed, and participants provided valuable design ideas and feedback.

4.1 Literature Review to Propose Design Directions

We proposed seven design directions for a collaborative VR. These directions were derived using a deductive approach, focusing on a comprehensive literature review to identify the challenges faced by DHH participants in this specific context. We did not directly gather participants' needs through system interaction and interviews due to their lack of VR experience and the absence of an existing accommodations system. These limitations hinder participants' ability to fully recognize and express their needs in the VR environment.

The literature review was conducted through the ACM Digital Library, targeting publications from January 2018 to July 2024. Using the keywords *"DHH" AND "caption"*, we identified 236 results. Following a comprehensive review, studies not explicitly aimed at the design of captions for DHH users were excluded. This includes research focusing on individuals who stutter, Automatic Speech Recognition (ASR), the conversion of sign language into captions, studies proposing captions as a potential solution without detailed analysis, and those exploring caption styles primarily for emotional expression. A subset of 20 papers([2, 6–8, 21, 27, 32, 34, 36–38, 40, 42, 45–47, 50, 55, 58, 63]) were further analyzed to explore barriers encountered by the DHH community in daily life, online interactions, and AR/VR environments within small group conversations or gaming contexts. From this analysis, we extracted key challenges and proposed the following design directions:

Challenge 1: Inappropriate caption placement.

The placement and style of captions impacts the user's reading experience [8, 32, 34, 50, 63]. Positioning captions within the field of view can obscure crucial information [6].

Whether the captions are aligned with the camera influences whether users miss information or need to repeatedly return to specific positions. Such disturbances can hinder user interaction within the game environment.

- **Caption Display Position**: The placement of captions requires careful consideration to strike a balance between visibility and usability, ensuring they enhance the readability for DHH users.

Challenge 2: Difficulties in localizing sounds.

The absence of sound localization might limit players' comprehension of the context and environment [27, 37, 38, 46, 47, 50, 58]. In daily life, DHH users often cannot retrieve sound information from behind due to constraints of modern hearing aids [58]. This limitation does not occur in VR scenarios where users wear headsets, contrasting with their everyday experiences. This inconsistency can cause confusion for DHH users in VR as they try to identify sounds. Additionally, when two VR users are not co-located in the same virtual space, it can be challenging for them to understand each other's positions.

- **Sound Locations**: The source of sounds should be visualized for DHH users to enhance their spatial awareness and interaction within the environment.

Challenge 3: Specific needs in small group conversations.

When more than two participants are involved, DHH users commonly encounter difficulties in identifying the speaker. In everyday contexts, they may rely on observing lip movements or others gaze direction to discern who is speaking [2, 36]. Observing others' faces and body languages was crucial for DHH users during a discussion [40, 42, 55]. However, in VR environments, those elements are generally not captured or visualized, posing significant challenges for speaker identification. Additionally, understanding the content and sequential order of captions becomes challenging when multiple individuals speak continuously or simultaneously, as the rapid exchange complicates the tracking of dialogue progression [2, 7, 21, 45, 50].

- **Adaptive Caption Design**: Caption layouts should dynamically adjust based on the number of speakers to optimize readability and context comprehension for DHH users.
- **Speaker Identification**: The captions should be matched to the specific speaker or, when avatars are absent, include contextual information to clarify the source of dialogue.
- **Speaker Order**: The system should display utterances in a manner that mimics natural speech patterns, enhancing comprehension and user experience.
- **Speech Overlapping**: The captions should be clearly differentiated and synchronized with their corresponding speakers to ensure DHH users can follow the dialogue without confusion.

Challenge 4: Technical concerns.

Due to limitations in the effectiveness of real-time captioning and potential internet failures, the timeliness of information delivery cannot always be guaranteed, and thus the user experience may suffer [2, 27, 50].

- **Caption Delay**: Information about the level of lag should be conveyed to DHH users to ensure they can fully understand the context and appropriately adjust their behavior or conversation with other users.

4.2 Co-design Workshops

We designed and conducted three co-design workshops. Participants were invited to carefully assess and discuss the design directions identified through the literature review and to propose their own innovative design concepts.

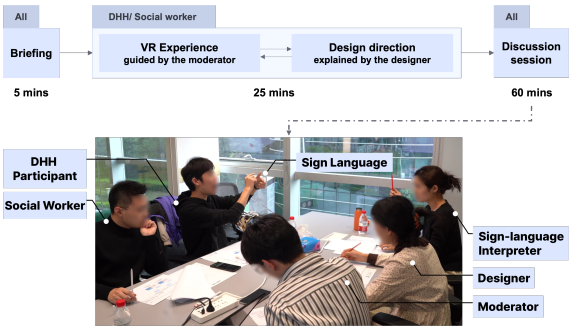

Figure 1: The upper part illustrates the procedure of the co-design workshop, while the lower part shows the setup for the discussion session, which included one DHH participant, one social worker, and one designer joined. Since the DHH participants prefer to use sign language for communication, we invite one additional sign-language interpreter to translate. One of the authors facilitated the session as the moderator.

4.2.1 Participants. We recruited nine participants (aged between 22 and 40, $Mean = 29.11, SD = 7.03$; five self-identified males and four self-identified females), comprising three DHH participants, three social workers who were familiar with DHH individuals (a sign language interpreter, a volunteer from a deaf association, and an expert from the local accessibility society), and three designers (two with six years of experience as industrial designers and one specializing in XR design; all had completed HCI-related coursework). In the subsequent co-design process, social workers serve as proxies for DHH participants, representing the needs and perspectives of vulnerable or hard-to-reach populations (i.e., DHH users) [33]. The three DHH participants were classified with grade 1 hearing loss, indicating extremely severe impairment in both the structural and functional aspects of the auditory system, characterized by an average hearing loss of 91 dB or greater in the better ear. All DHH participants used hearing aids and sign language; one preferred sign language as the first communication method. All participants were fluent in Chinese Mandarin. Table 1 shows the participants' demographic.

4.2.2 Procedure. Figure 1 illustrates the procedure of the co-design workshop and shows the co-design discussion session setup. Each co-design session lasted 1.5 hours, beginning with a brief five-minute introduction to general needs in VR. Following the 25-minute phase, the moderator guided

Table 1: Participant Information of Co-design Workshops

Group No.	Participant No.	Age	Gender	Description
Group 1	S1P1	27	Male	DHH Participant
Group 1	S1S1	39	Female	Sign Language Interpreter
Group 1	S1D1	23	Male	Designer
Group 2	S1P2	26	Female	DHH Participant
Group 2	S1S2	40	Female	Volunteer from the DHH Association
Group 2	S1D2	22	Male	Designer
Group 3	S1P3	27	Male	DHH Participant
Group 3	S1S3	35	Male	Expert from Accessibility Society
Group 3	S1D3	23	Female	Designer

one participant through a VR collaborative game demo in the context of an escape room game while not yet adapted for DHH users on the HTC Vive Pro Eye headset, ensuring the participant's safety throughout the experience. Concurrently, the designer explained the seven proposed design directions to the other participants. Once both non-designer participants were familiar with the VR environment and design directions, all three engaged in a discussion about these directions, proposed new ones, and brainstormed possible design solutions. The moderator encouraged participants to articulate their design rationale throughout the session, which concluded with a reflective discussion on shared ideas and priorities. All participants got 20 USD for their time. This study was approved by our Institutional Review Board.

4.3 Refine Design Directions

Throughout the co-design workshops, we compared the needs of DHH users and proposed design directions. Participants were asked to rate the necessity of these design directions on a 5-point Likert scale, where 1 indicates strong disagreement with the necessity and 5 indicates strong agreement. The results are shown in Table 2. While most design directions accurately reflected the needs of DHH users, there was still mismatch.

Table 2: Necessity Score from Co-design Workshop

Design Directions	Score
Caption Display Position	5.00
Speaker Identification	4.89
Sound Location	4.33
Adaptive Caption Design	2.67
Speech Overlapping	3.33
Speaker Order	4.56
Caption Delay	3.78

Regarding the *Adaptive Caption Design*, which was intended to tailor caption layouts based on the number of speakers, participants found it redundant for two-person conversations. They noted that a design suitable for small group conversations could naturally accommodate dialogues between two people, rendering a specific design for such scenarios unnecessary.

Additionally, participants raised new concerns about overall context comprehension, highlighting the need for a *History* feature. This feature would allow users to review all previous conversation records when they lose track of the discussion, such as when they are interacting with virtual objects and may ignore the captions. It is necessary to allow captions to be maintained temporarily [21] or permanently. Furthermore, the importance of a mechanism to record their own words was highlighted as essential, enabling users to reflect on and adjust their communication in real-time.

Following the design workshops, we confirmed and refined these design directions, categorizing them into the following three dimensions based on the benefits they offer to the user:

- **Readability**: Caption Display Position; Speech Overlapping;
- **Speech Information Transmission**: Speaking Order; Caption Delay; Chat History;
- **Non-speech Information**: Speaker Identification; Sound Location.

4.4 Findings of Caption Design from Co-design Workshop

We thoroughly reviewed caption designs provided by all participants in the co-design workshop for collecting the DHH users' needs and preferences for various design directions. Intending to refine the detailed caption designs in VR, we analyzed and summarized 15 critical design options across seven design directions, considering the requirements within the three main design dimensions. Table 3 presents the 15 options of the caption design.

Caption Display Position. All participants agreed that the positioning of the captions should follow the movement of the users' heads. Specifically, eight participants suggested placing the captions in the lower-middle position, which aligns better with everyday usage habits. S1S1, S1D3, and S1S3 stated that captions for narrators' or hosts' voices should be distinguished from player speech captions and displayed in a set position, like the upper-middle of the view. Additionally, S1P2 and S1S2 suggested that different functions should be distinguished; for example, prompts or warnings should separately appear in the center of the screen to make them noticeable.

After summarizing the design preferences of all participants regarding caption display positions, we created two design options to display player speech captions: (1) place the captions at the lower-middle (Figure 2-a), and (2) place them at the upper-middle (Figure 2-b) of the participants' view in VR.

Speech Overlapping. Even though we think an overlap during the speech will cause problems for DHH users during the conversation, not all participants believe marking the speech overlaps during the workshop is necessary. S1S1, S1P3, and S1S3 felt that no additional design was needed, as users could rely on the captions with speaker identifiers to distinguish. On the contrary, S1D1, S1P2, S1S2, and S1D3 suggested adding subtle cues, such as changing the color of the avatars'

Table 3: Findings from the Co-design Workshop

Design Dimensions	Design Directions	Design Options
Readability	Caption Display Position	(1) Lower-middle; (2) Upper-middle.
	Speech Overlapping	(1) No Additional Design; (2) Small Cues.
Speech Information Transmission	Speaking Order	(1) Fixed-position Display; (2) Natural Bubble-based Display.
	Caption Delay	(1) Different Colors; (2) Delay Time When Threshold Is Exceeded.
	Chat History	(1) Separate Window; (2) Scrollbar.
Non-speech Information Transmission	Speaker Identification	(1) Avatar; (2) Nickname; (3) Color.
	Sound Location	(1) Mini-map; (2) Avatar-Related Cues.

outline for simultaneous speakers. They emphasized that these cues should not be too prominent to avoid overwhelming the user.

Based on their designs, we established two design options for handling speech overlapping: (1) no additional design (Figure 2-c), and (2) small cues that display small logo as cues (Figure 2-d).

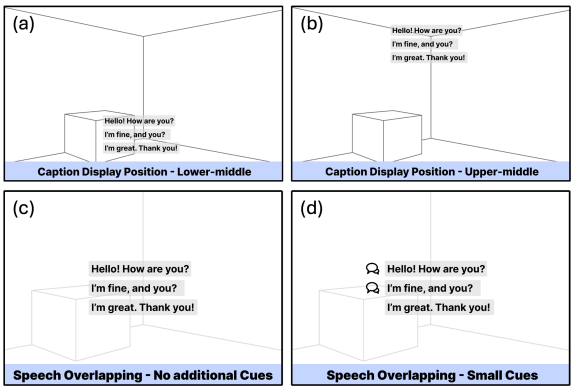

Figure 2: demonstrates the findings we collect from the co-design workshop about the caption display position (a and b) and speech overlapping (c and d). During the semi-structured interview, participants preferred to place the caption at the low-middle of the view and preferred no additional cues for marking the speech overlapping situation in multiplayer conversations in VR.

Speaking Order. During the discussion, two main methods were suggested for displaying speech orders in VR. Five participants (S1S1, S1S2, S1D2, S1S3, and S1D3) preferred a fixed position, meaning each speaker had a set place on the screen for their real-time captions, regardless of the order in which they spoke. Meanwhile, four participants (S1D1, S1P1, S1P2, and S1P3) favored a natural bubble-based ordering similar to the instant messaging software. In this way, each new caption appears below the previous one like a rising bubble, reflecting the chronological and sequential order of the conversation.

By analyzing the preferences, we developed two design options for displaying the speaking order: (1) a fixed position with each participant's speech updated in a single line (Figure 3-a) and (2) a natural bubble-based display (Figure 3-b).

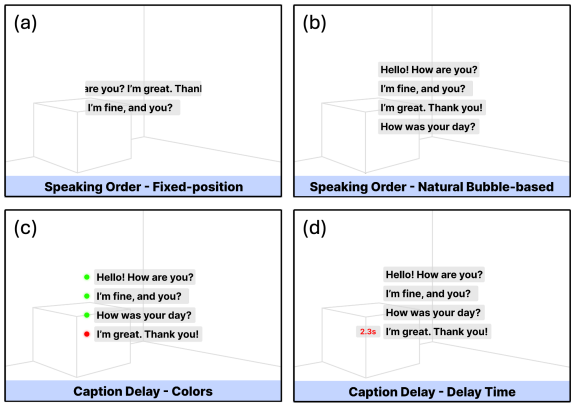

Figure 3: shows the examples of the display of speaking order (a and b) and caption delay (c and d). Participants expressed their preference for a natural bubble-based display to show the sequence of the sounds and a specific delay time when a threshold is exceeded to describe the transmission delay between speech and text.

Caption Delay. To hint at the delay between speech and caption caused by network connection issues, participants suggested using color or showing detailed delay seconds to highlight the unsynchronized generation. S1D1 and S1S3 recommended using different colors to indicate delays, such as changing the outline color of the avatar or bubble to make it stand out. Another two participants (S1S1 and S1P3) suggested displaying the delay time throughout the conversation to keep users informed. Additionally, three participants (S1D2, S1D3, and S1S3) preferred revealing the delay time only when it exceeds a specified time gap limit, as they mentioned that "*This way of expression does not impact communication under normal circumstances, and when unusual*

conversations suddenly occur, the reason can be immediately known through the prompts."

Through our analysis, we made two design options for alerting caption delay: (1) using different colors to highlight delays (Figure 3-c) and (2) showing the delay time in seconds only when a threshold is exceeded (Figure 3-d).

Chat History. Various suggestions on accessing and reviewing previous conversations were proposed through the co-design workshop, particularly when nearby captions are missed due to fast-paced conversations. S1S1 and S1P3 suggested using a separate window to view chat history, which can be hidden when unnecessary. Meanwhile, four participants (S1P1, S1D1, S1P2, and S1D2) preferred scrollable captions near the real-time. All participants suggested having a scrollbar on the side of the captions to allow users to scroll back and view past conversations.

Based on the participants' preferences, we created two design options that allow users to navigate past conversations using a scrollbar alongside the captions: (1) a separate window for viewing caption history (Figure 4-a) and (2) a scrollable caption (Figure 4-b).

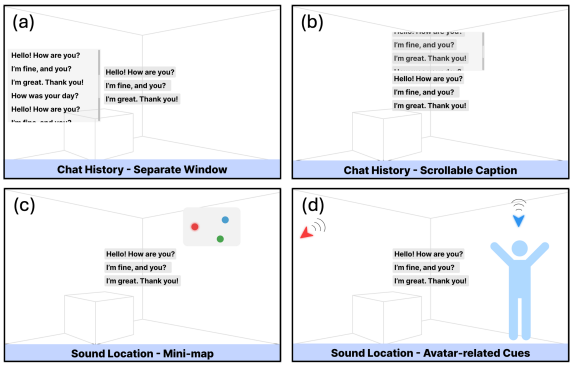

Figure 4: presents different ways to display the chat history (a and b) and identify the sound location (c and d). As a result, participants tended to use a scrollable caption to show the chat history and an avatar-related cue to recognize the sound location.

Speaker Identification. Participants proposed various methods for effectively identifying speakers in VR. Three participants (S1P2, S1D2, and S1D3) recommended using avatars to represent the current speaker as they emphasized the need for quick information retrieval within the game and pointed out that avatars distinguish the speaker and provide directional sound information. In contrast, the other three participants (S1S1, S1D1, and S1P3) preferred using nicknames for speaker identification. They suggested that as each participant has their own account and nickname during the start setting, it could serve as direct identification. Additionally, S1S2 and S1S3 recommended color coding for speaker distinction, as it can be seamlessly integrated into caption text without occupying additional space.

We identified three main design options by analyzing the identification preferences: (1) avatar (Figure 5-a), (2) nickname (Figure 5-b), and (3) color (Figure 5-c) to indicate the speakers.

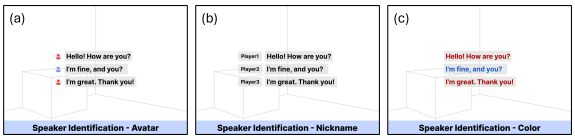

Figure 5: illustrates the methods for identifying speakers. Among the three methods, participants favored chose avatars to present and identify the speakers in multiplayer conversations in VR.

Sound Location. One method to indicate the location of the sound is to utilize a mini-map, as five participants (S1P1, S1D1, S1P2, S1S2, and S1S3) recommended using mini-map pins at the top of the view to indicate the location of the sound, similar to the display of player positions in popular games [39, 49]. Meanwhile, participants also contributed some valuable design ideas for clarifying where the sound came from. Three participants (S1D2, S1P3, and S1D3) preferred adding cues within the field of view, which can provide a more precise indication, especially since the view edges of VR can be challenging to see.

Aiming to indicate the location of the sound, we developed two design options: (1) a mini-map at the top right of the view, but away from the edge (Figure 4-c), and (2) avatar-related cues to clarify the sound's location (Figure 4-d).

5 Interviews for Design Selection

To collect the preferences of DHH users regarding caption design, we conducted semi-structured interviews with 13 DHH participants. We then developed a caption prototype for the multiplayer conversation in VR based on the interview results. In this section, we first introduce the process of semi-structured interviews and then describe the prototype design of our proposed VRCaptions.

5.1 Semi-strucrtued Interview

5.1.1 Demographics. Thirteen participants (aged between 21 to 28, *Mean* = 24.23, *SD* = 2.13; seven self-identified males and six self-identified females, all native Mandarin Chinese speakers) were involved in the interview; two previously attend our co-design workshop. Participants varied in the degree of hearing loss, the demographic as shown in Table 4; all reported using hearing aids.

5.1.2 Methods. We prepared a Caption System Demo that includes all the caption designs based on the findings from co-design workshops. Three hearing moderators repeatedly played the Caption System Demo to create several video clips using different caption designs, recorded in the Unity engine.

Our interviews lasted around one hour, beginning with a 10-minute introduction to caption designs and design directions. Participants were then asked to watch 15 video clips recorded from the Caption System Demo with different designs and experience the prototype implementation on the HTC Vive Pro Eye VR headset for 15 minutes. Afterward, participants were invited to rank all design options in each direction. Based on the results, we followed the previous

Table 4: Participant Information of Design Selection

Participant No.	Age	Gender	Description
S2P1	25	Female	First Level of hearing loss (> 91dB)
S2P2	26	Female	First Level of hearing loss (> 91dB)
S2P3	27	Male	First Level of hearing loss (> 91dB)
S2P4	26	Female	Second Level of hearing loss (81 - 90dB)
S2P5	28	Female	Third Level of hearing loss (61 - 80dB)
S2P6	22	Female	Forth Level of hearing loss (41 - 60dB)
S2P7	25	Male	Forth Level of hearing loss (41 - 60dB)
S2P8	22	Male	Forth Level of hearing loss (41 - 60dB)
S2P9	23	Male	Forth Level of hearing loss (41 - 60dB)
S2P10	23	Male	21 - 41dB hearing loss
S2P11	23	Male	21 - 41dB hearing loss
S2P12	24	Male	21 - 41dB hearing loss
S2P13	22	Female	21 - 41dB hearing loss

work [29, 58] and conducted 35-minute semi-structured interviews to collect the participants' feelings and reasons for their selections. To ensure full involvement, real-time captioning was activated during the semi-structured interviews. Each participant received 15 USD for their time. This study was approved by our Institutional Review Board.

5.1.3 Results. Based on the semi-structured interview, we explored DHH participants' preferences for various caption design directions in VR, focusing on caption display position, speech overlapping, speaking order, caption delay, chat history, speaker identification, and sound location.

For the **Caption Display Position**, ten participants chose the "lower-middle" position to place the caption. Most participants who preferred the lower-middle position mentioned that it aligns more with their habit of reading captions in daily life. For example, S2P10 and S2P11 stated, "*Captions in TV shows or movies are usually at the bottom of the screen, so having them at the top would feel strange.*" S2P12 considered that when using VR equipment, "*[...] currently the HMDs are relatively heavy, so even at rest, I also tend to lower my head naturally or shift the gaze downward unconsciously.*" On the other hand, the three participants who favored the upper-middle position mentioned that this was more related to "*personal habits or intuitive ways of looking at things.*"

Regarding **Speech Overlapping**, 11 participants expressed that no particular indicator is needed, which aligns with the feedback from our previous workshop. They indicated that it's possible to tell others who are speaking right now through the captions. "*If several captions appear simultaneously, it indicates that many people are talking at the same time,*" said S2P2. Additionally, four of the participants (S2P6, S2P7, S2P12, and S2P13) mentioned that even though they are DHH users, if they can still hear to some extent, they can realize the sound when others speak. In contrast, the two participants who preferred small cues explained that "*[...] also it can be overwhelming when multiple people are talking simultaneously, and having a small cue could inform others and remind the current speakers.*"

In selecting **Speaking Order**, eight participants chose the natural bubble-based ordering, while the remaining five chose a fixed position. The participants who preferred the bubble ordering believed that the conversation sequence was important and the bubble-based captions are commonly used and can naturally display the conversation order. S2P5 asserted that the order is essential for the context, and S2P6 emphasized, "*It's essential to know the order in which things are said, especially when trying to understand the conversations.*" Conversely, the participants who favored the fixed position felt that having a static position for each speaker's captions made it easier to identify who was speaking. As S2P7 mentioned, "*[...] and if you miss the point, you can ask the speaker to explain again.*" Meanwhile, S2P12 expressed concern that "*if the conversation is fast-paced, bubble-based captions might make it difficult to catch and read the captions.*"

To display the **Caption Delay**, nine participants preferred to show in seconds only when a threshold is exceeded rather than using different colored lights to indicate caption delay in VR. They believed the delay "*should only be shown when necessary.*" S2P10 remarked that "*it's important to minimize the elements on the screen as much as possible,*" and S2P12 agreed that "*hiding it when not needed allows for a more immersive experience.*" Some participants also felt that using different colored lights could be confusing. S2P1 pointed out, "*I might need to remember the meaning of each color indicating since there will be many elements shown with the caption [...], which may distract attention, [...] and make it harder to read captions.*" In comparison, the four participants who favored using various colored lights pointed out that it could simplify the interface. According to S2P11, "*A small light offers an intuitive means of determining the network state of the other participants.*" Yet, they also expressed similar concerns regarding potential distraction, with S2P8 observing, "*Even though the small light is easy to check, it is important to explain the function of this light in advance.*"

We found that participants had varying preferences regarding the **Chat History**. Nine participants believed that a separate history window took up too much space while using scrollable captions would be more straightforward. S2P6 pointed out scrolling directly through the captions "*is more convenient to operate,*" and S2P13 highlighted that "*scrollable captions would not unnecessarily block the view compared to a separate window.*" Additionally, S2P9 remarked that the separate window occupies too much space and limited multitasking, "*even with the option to hide it, I feel like I can't do anything else while watching it.*" While four other participants preferred the separate window, allowing them to view more complete conversations, as S2P3 stated, "*to check the history using bubble through messages one by one is inefficient, and having a separate history window allows me to retrieve information when necessary and can also be hidden when not needed.*"

To consider the current **Speaker Identification** in VR, eight participants favored using avatars for clarity. At the same time, three preferred using colors that were easier to distinguish, and only two chose to use nicknames that would make individuals more straightforward to identify. Of the eight participants who preferred avatars, four expressed concern that "*using colors or nicknames would lead*

to more confusion." As S2P1 stated, "*[...] avatars are more noticeable than nicknames and less distracting than colors.*" Similarly, S2P7 expressed that " *If it is a conversation between many people, colors will be easily confused, and a lot of colors will easily interrupt my train of thought, while nicknames normally with varying lengths could disrupt readability.*" S2P9 pointed out that "*it's difficult to define colors, as there are already various colors in our VR scenes, and if the chosen color overlaps with the scene, it will become hard to see.*" Three Participants who preferred using colors noted that avatars could affect the overall layout, with both S2P10 and S2P12 mentioning that "*avatars are opaque and might block part of the view.*" Participants who chose nicknames explained that using nicknames makes it easier to distinguish between people, with S2P2 also indicating that "*Especially if everyone is a stranger, [...] I prefer to use nicknames.*"

Surprisingly, all participants preferred to use avatar-related cues to determine the **Sound Location**, such as small arrows or highlighted avatar edges. Compared with the design of the mini-map, participants believed that arrows or avatars-related cues could provide a more intuitive way to indicate who was speaking and which direction the sound was from. S2P7 further suggested that utilizing a small chat bubble as the arrow design would enhance the clarity, "*I prefer a what-you-see-is-what-you-get design,*" as also mentioned by S2P9 and S2P12. "*Using small arrows feels more fitting for VR.*", as S2P5 noted. Meanwhile, S2P8, S2P10, S2P11, and S2P13 raised concerns that "*the mini-map would occupy a certain field of view.*" Some participants also thought that a mini-map would bring unnecessary complexity. As S2P8 stated, "*a mini-map is more useful when you already know the overall layout of the current space, but it can be confusing for newcomers to understand what the mini-map is pointing to.*" S2P1 further emphasized that "*sound location cues should only be displayed when the sound appears; if no one is speaking, there's no need to occupy visual space.*" Additionally, S2P10 underlined that "*Mini-maps are usually placed in the corner, so you have to divert your attention to figure out where the others are, [...] while arrows just use a single icon can make it obvious at a glance.*"

Our findings revealed some consensus in our previous workshop, such as the presenting of speaker identification techniques and the preference for lower-middle caption positioning, which aligns with daily viewing habits. Meanwhile, participants also expressed differing opinions on design directions, such as displaying the speaking order and indicating the sound location, reflecting specific conditions and usage contexts. We summarized these findings and prepared VRCaptions, a prototype for a multiplayer VR game.

5.2 Proposed VRCaptions System

Based on the interview results, we summarized the preferences of all the design directions and developed a prototype of a caption system, VRCaptions, for validate the caption designs in VR multiplayer conversation with the following setups. Each decision was supported by the preferred results agreed by more than eight DHH participants during the semi-structured interview.

- **Caption Display Position — Lower-middle.** We designed the captions positioned in the lower-middle part of the users' view in VR, placed at approximately 30% of the vertical screen height to ensure readability without obstructing the main interaction.
- **Speech Overlapping — No Additional Design.** We did not implement additional speech overlapping designs, as most participants preferred to rely on the existing speaker identification and natural flow of the conversation.
- **Speaking Order — Natural Bubble-based Display.** We implemented a natural bubble-based display system where new captions appear in the sequence below previous ones, presenting a conversation flow in the time order of speech.
- **Caption Delay — Show Delay Time.** We designed the captions to display the delay time in seconds only when the delay exceeds a set threshold, ensuring users are aware of significant lag without cluttering the interface and interrupting attention.
- **Chat History — Scrollable Caption.** We created a scrollable chat history under the fixed position of the caption, allowing users to scroll back through the previous context without overwhelming the visual space during real-time conversations. By double-clicking the trigger, participants could turn on and turn off the chat history.
- **Speaker Identification — Avatar.** We utilized avatars for speaker identification, helping users quickly and clearly identify who is speaking based on visual representations.
- **Sound Location — Avatar-related Cues.** We used avatars-related cues that highlight the sound location in VR, displaying visual indicators around the avatar to provide an intuitive way to locate speakers.

Figure 6 shows the VRCaptions system. Following the previous design, we designed the captions in the game to be Noto Sans Mono CJK TC [58] based on the Mandarin language the participants prefer to use, with a maximum of 12 characters per line [71] and a caption retention time of 3-5 seconds [20]. The VRCaptions system prototype enables DHH users to clearly and intuitively understand conversations and receive non-speech information in multiplayer conversation in VR.

6 User Study for Design Validation

To validate the design of the VRCaptions, we integrated the caption prototype into a multiplayer VR game and conducted an validation study with both DHH participants and hearing users as mix-hearing groups. Following that, we interviewed the DHH participants and collected their suggestions for caption design in multiplayer games and future multiplayer conversation in VR for the DHH community. This section introduces the game experience first and then summarizes the interview results.

6.1 Game Experience with Mix-hearing Groups

To better understand the caption design with DHH users for VR multiplayer conversation, we developed a VR game *Room*

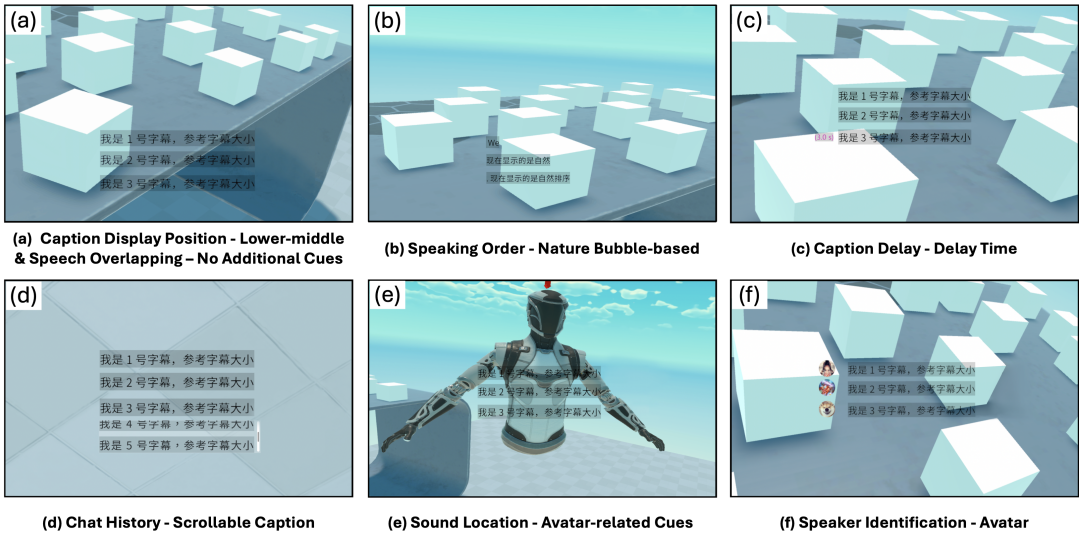

Figure 6: shows the prototype of VRCaptions. Based on the participants' preferences, we proposed the demo to present the caption design for DHH users in VR multiplayer conversations. Figures (a) to (f) illustrate the detailed design directions separately.

Escape with the caption prototype VRCaptions and invited two mix-hearing groups to participate. Figure 7 shows the in-game screenshots and participants in the game.

6.1.1 Game Design. *Room Escape* is a multiplayer collaborative escape room game we developed in VR. Escape room games are thought to improve communication and teamwork [16, 56]. Utilizing VR escape room games as the context can reflect the access needs of DHH users for speech communication and multiplayer collaboration, offering interactive environments and immersive experiences. In *Room Escape*, we designed two game scenes based on verbal discussion representing localized and non-localized speech [38] respectively, which could allow us to assess how DHH users interact with VR environments and validate accessibility in various interaction contexts in VR multiplayer conversations. In the first scene, all participants were present as cartoon-like avatars and interacted in the same virtual room, as shown in Figure 7-d. They needed to collaborate with others to find the corresponding keys based on clues, and only after collecting all the keys could they complete the scene. Throughout the game, participants could see each other's avatars and identify the speaking participant's location based on the source of the sound. In the second one, participants were placed in the same but independent virtual rooms in the second scene and could connect with the others only through speech exchange, as shown in Figure 7-e. They were recommended to collaborate with the others by sharing the clues they found and could complete the game when all three participants had located their respective keys. Since each participant was alone in their virtual room, it was impossible to determine the other participants' locations based on sound, necessitating information exchange through speech-only communication for successful collaboration. To ensure that participants could explore all design directions of the VRCaptions, both scenes require participants to stay engaged with the conversation in real-time (to validate the design for caption display position, speaking order, caption delay, and chat history), distinguish who is currently speaking (to validate the design for speaker identification), and exchange their location information (to validate the design for sound location).

We developed the game in Unity3D Engine and packaged it to install on three laptops in three separate rooms. We utilized the real-time speech transcription service from iFlytek [26] for speech-text recognition. All laptops were run on the same operating system with the same configuration.

6.1.2 Procedure. All participants were asked to complete a demographic questionnaire before the game started and were guided by the moderator to three separate rooms. They were first assisted in putting on the HTC Vive Pro Eye VR headset and then conducted visual and audio calibration. Once all the calibration was complete, participants started the *Room Escape* game.

Our *Room Escape* has two scenes, participants were asked to finish the tasks in each scene. As a result, participants spent 20 minutes on the first scene and 10 minutes on the second. After they finished the game, DHH participants were invited to complete a questionnaire to assess their satisfaction with the captions with a 5-point Likert scale (1 indicates very unsatisfied and 5 indicates very satisfied). The questionnaire contains seven questions that cover the design directions we chose for the VRCaptions. We then invited participants to join a 20 mintes interview to discuss the overall gaming experience and whether VRCaptions improved the readability of captions, speech, and non-speech information accessibility for DHH participants in-game. The procedure lasted around one hour, and all participants got 15 USD for their time. This study was approved by our Institutional Review Board.

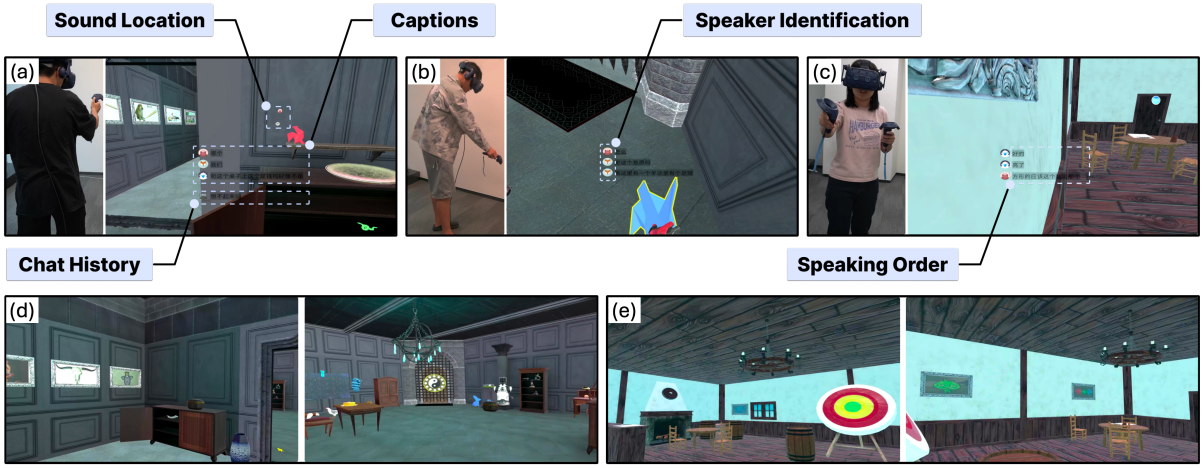

Figure 7: illustrates the in-game scene of *Room Scape* we developed to verify our VRCaptions. Three participants verbally collaborated in the game. (a) and (b) show the participants and the VRCaptions in the first scene, and (c) present the second one. We found no caption delay caused by the internet issue during the study. (d) and (e) demonstrate the design of each game scene.

Table 5: Participant Information of Validation Study

Group No.	Participant No.	Age	Gender	Description
Group 1	S3P1	26	Male	DHH Participant
Group 1	S3P2	27	Male	DHH Participant
Group 1	S3H1	23	Female	Hearing Participant
Group 2	S3P3	23	Female	DHH Participant
Group 2	S3P4	23	Male	DHH Participant
Group 2	S3H2	25	Male	Hearing Participant

6.1.3 Demographics. Inspired by previous resarch [58], we recruited six participants (Age: $Mean = 24.67$, $SD = 1.63$; four self-identified males and two self-identified females) as two groups to attend our validation study. Every three participants form a mixed hearing small group (two DHH participants and one hearing participant), enabling active interaction with both the system and each other. Table 5 shows the demographic of the participants. Two DHH participants (S3P1 and S3P2) have 21-41dB degrees of hearing loss, and another two (S3P3 and S3P4) have more than 91 dB. All the participants reported they had VR game experience before.

6.2 Results

The questionnaire results in Table 6 indicate that DHH participants are generally satisfied with the display position of the captions. They are also content with the performance regarding speech order, chat history, and speaker identification. This suggests that the captions effectively convey caption readability, speech, and some non-speech information transmission. There is a lower rating for caption delay since there's no delay happening during the game.

Following the questionnaire, we summarized the interview results across three dimensions.

Table 6: Satisfaction Score for VRCaptions in VR Game

Design Directions	Score
Caption Display Position	5.00
Speech Overlapping	5.00
Speaker Order	4.50
Caption Delay	3.75
Chat History	4.50
Speaker Identification	4.00
Sound Locatation	4.50

Overall Gaming Experience. All participants agreed that the captions were easy to understand and greatly facilitated in-game communication, effectively reducing communication barriers. As S3P1 mentioned, "*I got a lot of help from the captions, like [...] which assist me in communication.*" S3P4 also expressed that "*With this caption system (VRCaptions), communication becomes less of a problem, especially for a DHH like me who can't speak clearly. It makes me feel more involved in the game.*"

Caption Readability. We found that all participants appreciate fixing the captions in the lower-middle part of the field of view as it will make them easy to follow. At the same time, some participants shared concerns that the readability of captions can be challenging when the caption color is too similar to the background. S3P3 found that the fixed caption layout sometimes could block the view in VR, as stated, "*especially when captions overlap with in-game objects, [...] which impact the visual experience.*" Related with these, S3P1 mentioned that "*for some complex and high interactive scenes, feels like it will be flexible if we could customize the caption position.*" Concerning speech overlapping, no participant reported overlapping issues affecting the gaming experience

or readability of captions in the game. S3P2 suggested that "*if there are many people in the scene, I might expect that there could be a small mark in front of the avatar or some small cues to indicate who is speaking at the same time, but it shouldn't be too prominent.*" Yet, S3P1 noted that the number of captions displayed is limited when multiplayer speak simultaneously, "*[...] which may become difficult to follow the conversation through the rapid update captions when a speech overlapping happens.*" For this, participants suggested increasing the number of captions to improve comprehension.

Speech Information Transmission. All of the participants could follow the conversation through the captions, which accurately displayed the speech order. One participant (S3P4) mentioned that some information might be missed during fast-paced conversations. All participants appreciated the scrollable history, noting that the double-click function could solve the issue of history captions potentially blocking their view in certain game scenes. S3P2 emphasized, "*I noticed that I could double-click to turn it off, which is helpful for me as the history frame does obscure my view a bit when it's on, [...] especially when I'm looking at information in the history while searching [...] in the scene.*" S3P1 also remarked, "*The history feature is extremely helpful during the frequent conversations, with the speaking order providing valuable context, that's really useful.*" S3P4 further pointed out that "*it does block part of the field of view, but it could be hidden by double-clicking, which provides a positive experience.*" Since there was no caption delay during the game, participants could not experience and share their comments.

Non-speech Information Transmission. All participants mentioned that they could easily identify the current speaker through the avatars, with S3P3 and S3P4 stated, "*I can easily distinguish who is speaking through the avatar.*" Additionally, S3P1 mentioned that "*in future designs if we can upload our customized images or icons as an avatar, that would be better for identification.*" For the sound location, all participants indicated that avatar-related cues could provide brief information about sound locations in specific situations. S3P3 pointed out that explaining these visual cues in advance is necessary, as different systems represent sound location in various ways. "*If DHH users are unfamiliar with games or similar caption systems, they might find these cues confusing when they first start participating in the conversation.*" S3P2 further shared that while using the avatar-related cues helped in identifying the direction of the sound, it might be helpful to provide them more information to identify the detailed location directly through the captions, "*I can tell that the speaker is in front of me in this direction, but I can't judge the specific location because I don't know how far he is from me.*"

7 Discussion

In the real world, DHH individuals usually require additional speech-to-text devices or sign-language interpreters for multiplayer conversations. VR with integrated real-time captions or sign language [63, 72] improves the accessibility of speech information for DHH individuals in remote collaboration. But with multiplayer, conversations in VR contain more information, such as non-speech information that identifies the current speaker and the location of the sound, which brings access barriers for DHH individuals [50]. Although researchers have started providing more information through caption design for DHH users in multiplayer conversations, the needs in VR multiplayer conversations remain unexplored. In this paper, we explored the caption design of DHH users in VR multiplayer conversations for the first time. Through co-design workshops and semi-structured interviews, we reported the design needs and preferences of DHH users in seven design directions across three design dimensions. We then developed a VR caption prototype based on the above design preferences and implemented the prototype into a VR multiplayer game for verification. Based on our insights into VR caption design for DHH individuals, we discussed the following reflections.

7.1 Reflection on VRCaptions for VR Multiplayer Conversation Designs

Especially in the growing number of remote conversations such as videoconferencing and collaborations in AR and VR, DHH individuals increasingly demand the accessibility of multiplayer conversations. For DHH individuals, caption design should accurately convery conversation content with low delay, ensure high readability for large text volumes, and effectively represent non-speech information [58]. Based on previous research, this work further explored the needs of DHH users in VR multiplayer conversations in the seven design directions, specifically lying on **caption display position, speech overlapping, speaking order, caption delay, chat history, speaker identification, and sound location.**

Caption Display Position. Immersive environments often provide additional spatial information that significantly differs from 2D media, such as those in VR or 360° videos [12, 23, 31]. Hence, it is essential to strategically position captions to maintain the immersive experience while accessing critical spatial information within these environments [61]. In 360° videos, previous studies have found that users prefer captions to follow their head orientation for an enhanced experience, allowing each new caption to appear directly in front of them [12, 13]. Similarly, in our work, all co-design workshop participants agreed that captions for DHH users in VR multiplayer conversations should follow head movements, which aligns with their VR experience. Furthermore, previous studies have demonstrated that DHH users prefer familiar positions for captions, such as those found in TV shows and movies, even when alternative caption designs are considered [51]. Research conducted by Berke et al. also showed that DHH users prefer captions placed at the bottom of the video in group meetings [8]. Consistent with previous studies, our research shows that DHH users prefer positioning captions in the lower-middle part of the field of view during VR multiplayer conversations, similar to the positioning commonly used in videos or TV shows. As mentioned by participants that this design aligns with daily usage experience, we proposed that this familiar

placement allows DHH users to focus on the conversation without additional effort to adjust to new caption positions.

Additionally, we found that DHH users have different design considerations for various types of captions. During our semi-structured interviews, some participants suggested that prompt-related captions like narrators or system alerts should be placed separately. This would help DHH users better understand conversations while ensuring they do not miss critical flow-related cues. Previous research has explored using pop-up windows for important prompts in videoconferencing, such as when captions are mistranslated [51]. While it was also noted that these pop-up windows still require further refinement to reduce the disruption to the ongoing conversation. Therefore, positioning captions in VR aligning with users' familiar visual habits and conversation needs could offer DHH users a more intuitive and coherent multiplayer conversation experience.

Speech Overlapping. Speech overlapping or rapidly displaying conversation content becomes difficult to follow, which affects the readability of information and causes access barriers for DHH individuals [7, 29, 50, 51]. Even though the presenting of overlapping conversations has been investigated, the ambiguity caused by multiple speakers in VR remains underexplored [35, 45]. In multiplayer video conferences, McDonnell et al. implemented pop-up prompts to notify the conversation overlaps [51], while some participants reported that this method could disrupt the flow of communication and create fairness issues. In line with this, our work found that additional cues to indicate overlapping conversations were generally unpreferred among DHH participants since the speaker avatar has already shown out when they were speaking. Unlike online meetings constrained by 2D screens, users in VR could interact with their surroundings, making them more sensitive to visual cues and symbols in their field of vision [31]. Meanwhile, in our semi-structured interviews, most participants noted that additional cues for conversation overlap are unnecessary when the speaker could already be identified.

Speaker Order. Captions are designed to present content and order in various ways, such as through scrolling, pop-up, or cinematic modes in 2D videos [29]. In this work, we designed natural bubble-based ordering captions that update speech in a rising order within a fixed area. Similarly, we found that in AR, Peng et al. designed captions for DHH using enumerated bubbles that corresponded to the time order of the conversation [58]. With the conversation content presented to DHH users in chronological order through the rising of the bubbles, this design was intuitive and easy to understand as the participants noted it was commonly used in online text-based conversation in the semi-structured interview. Additionally, for the single-line setting, we found that users always need to switch between other users' captions to follow the conversation while simultaneously exploring the environment. In the semi-structured interviews, participants who preferred the bubble-based design instead of the single-line setting mentioned that a coherent presentation of conversation content is essential for better understanding and obtaining context information.

Caption Delay. In the study by McDonnell et al., it was observed while partial DHH users were interested in communicating their delay to others, certain participants believed that captions would always be perceived with delay, rendering feedback on this issue unnecessary [50]. Similarly, our work found that most DHH users disliked the continuous display of caption delays, often perceiving it as redundant. During semi-structured interviews, some participants suggested using colored lights to indicate caption delays. Nevertheless, they also expressed concerns about potential distractions and increased complexity, particularly for DHH users unfamiliar with the system. Most participants emphasized the need to minimize on-screen elements to maintain an optimal user experience, aligning with findings from previous studies [17, 50]. This reminds us that when designing accessible VR environments, overly explicit or constant cues, even if well-intentioned, can possibly disrupt the user experience in an immersive environment. In our work, we displayed the delay in seconds only when it exceeds a threshold, aiming to strike a balance between providing valuable information and preserving the user's sense of immersion. Although there was no significant delay during our escape room game, which prevented us from collecting feedback on this aspect, we believe implementing corresponding responses when necessary could improve communication for DHH users without compromising the immersive experience.

Chat History. The ability to review the chat history is an important feature due to the limited space available for displaying captions [24, 25], as it provides a possibility to revisit dialogue and better understand the context. Previous research highlights the importance of allowing users to manage their attention between real-time captions and supplementary information, which pointed out that although chat history could help better understand and integrate information from the video, it could increase the cognitive load when users are required to switch between the video and the history window [44]. Our findings align with these prior insights, revealing a better following and understanding of the conversation's progression through chat history. While participants also shared their concern that displaying the chat history could occupy more screen space and make it harder for users to engage in real-time conversation. In light of previous research, we introduced the chat history as a single line that is scrollable below real-time captions and can be hidden by double-clicking. DHH participants in the validation study noted that this design neither obstructs their view nor disrupts the immersive experience, which indicates that it could allow DHH users to review conversation content smoothly.

While a separate history window can offer a more structured presentation of past conversations, it also presents challenges by occupying more screen space and making it harder for users to engage with real-time conversations. As noted by previous studies, captions that occlude onscreen content may limit the amount of visual information that DHH users can perceive [5]. In semi-structured interviews, participants shared that they still felt unable to do anything else while watching it, even with the option to hide it. Hence, dynamically managing information based on immediate needs can help DHH users balance content accessibility and

immersion in VR, especially as the interaction complexity increases in multiplayer conversation.

Speaker Identification. Speaker identification should be carefully considered for DHH individuals in multiplayer conversations. Previous research indicated that in immersive environments, visual elements can draw users' attention away from the surrounding environment or the main content, with excessive visual cues potentially leading to confusion and distraction [31]. Our work found that most DHH participants preferred using avatars to identify the current speaker since they can provide a clear visual representation of the speakers and help reduce prominent visual elements in the user's field of view. Some participants in semi-structured interviews expressed concerns about avatars affecting the overall layout, particularly opaque avatars that might block part of the view.

Moreover, previous research suggested that color-coding speakers could be an effective method to identify speakers in group conversations [50]. Nonetheless, during semi-structured interviews, our DHH participants noted that relying solely on color to identify each speaker could be challenging. They emphasized that users unfamiliar with the system might need additional cognitive effort to associate specific colors with corresponding speakers. Additionally, some DHH participants suggested using nicknames to identify speakers, especially in scenarios involving strangers. At the same time, others were concerned that the varying lengths of nicknames posed a potential risk to the readability. When incorporating visual elements for speaker identification, the overall layout and the user's familiarity with the system should be considered.

Sound Location. Sound Location should be effectively identified for supporting DHH individuals in more accessible group communication [4, 35, 38, 50]. DHH users face challenges in distinguishing audio source locations, even with hearing aids [58]. In previous research, Li et al. suggested combining a mini-map with object indicators to help DHH users recognize the location of the sound source in VR [46, 75]. Surprisingly, our work shows that DHH participants prefer avatar-related cues rather than mini-maps during VR multiplayer conversations, despite the widespread use of mini-maps in video games to track environmental updates [39, 49]. This preference likely stems from the difference between conventional 2D video game environments and immersive VR experiences, where users need to engage more actively with their surroundings in VR [14]. DHH participants in semi-structured interviews pointed out that mini-maps in VR can occlude their field of view and create unnecessary visual clutter. Some participants suggested that if mini-maps are used to identify sound locations, they should not occupy visual space when no sound is present. Additionally, for users unfamiliar with the virtual environment, using mini-maps may lead them to struggle to orient themselves and locate others. This aligns with previous research, which suggests that users need a clear understanding of the representation of the current surroundings to comprehend the position and scale of the virtual environment [49].

Therefore, we proposed avatar-related cues to indicate sound location, which could engage DHH users to participate in conversations more effectively. As noted by DHH participants in semi-structured interviews, visual indicators, such as arrows or highlighted avatar edges, could enhance the VR experience by maintaining visual clarity while providing crucial information about sound sources and active speakers. This finding aligns with Jain et al.'s work, which demonstrated the effectiveness of directional cues in enhancing spatial awareness in multiplayer games [35]. Nevertheless, participants also expressed concerns that these cues might initially introduce cognitive confusion, particularly without proper instructions, as DHH users transition between systems that represent sound locations differently. This underscores the importance of providing clear guidance on these cues' functionalities, especially for DHH users unfamiliar with current captioning systems.

7.2 Design Consideration for future VRCaptions design

A primary area of this paper is to investigate the caption needs and preferences of DHH users for multiplayer conversations in VR since the rich diversity of sound in VR enhances the immersive and realistic experience and provides many important notifications [10, 15]. Previous research has shown that low accessibility to multiple sounds in VR can limit the experience for DHH users, such as missing important cues during the conversation [37]. Researchers pointed out that caption design should address trade-offs between various factors to improve accessibility in 360° videos [12, 31]. Similarly, in VR multiplayer conversations, it is crucial to balance multiple design directions to enhance caption accessibility while preserving the immersive experience for DHH users. Building on these insights, we propose the following design considerations for future development.

Readability. In our work, DHH participants expressed a preference for positioning captions in the lower-middle part of the field of view, as this placement aligns with their familiar experiences of watching TV shows or movies. This preference suggests that incorporating familiar visual arrangements into caption design could enhance accessibility and reduce the learning curve for DHH users [8, 51]. Furthermore, as previous research on 360° videos highlighted, background quality can affect caption readability when selecting positions, and a typical reason for missed content is the captions obstructing the view [11, 13]. Similarly, in VR, the impact of the VR environmental background and potential occlusion of content within the DHH user's field of view should be carefully considered, as participants in our validation study suggested applying adaptive designs, such as auto-adjusting caption color or transparency based on current background, to minimize these issues and support better engagement with conversations. Additionally, we found that DHH participants preferred not to have extra cues indicating overlapping speech in VR. While previous studies have shown that speech overlap can affect the readability of fast-paced conversations in video conferencing scenarios [45, 50], DHH participants in our semi-structured interview expressed more concern about the value of adding visual cues for speech overlap within their field of view. According to the insights from our participants, when speaker

identification is clear, we propose carefully considering having additional cues for speech overlap to help maintain readability and conversation flow while minimizing visual clutter.

Building on our findings, we suggest the caption design for DHH users should align with their natural viewing habits and learning curves, especially when adapting to new caption systems in VR. At the same time, the potential readability challenges should also be considered, which could posed by VR environmental backgrounds and the additional visual elements for prompting speech status. Furthermore, given that users have different preferences for interacting with the same content, the previous study suggested that allowing customization of caption appearance can reduce the cognitive load associated with reading captions in 2D videos [51]. Customization may also help minimize content occlusion when positioning captions in multiplayer conversations, as participants in the validation study mentioned that VR environments are highly dynamic and ever-changing. Allowing users to adjust caption positions or adaptive background transparency could further enhance caption accessibility for DHH users [8, 63], all while preserving the immersive experience in VR multiplayer conversations.

Speech Information Transmission. Our work found that the natural bubble-based display, which integrates enumerated bubbles [58] and fixed-position captions [4, 18, 63], could order captions chronologically and align well with DHH users' needs for understanding conversation sequences. We suggest that when designing captions for multiplayer conversations, it is necessary to prioritize designs that facilitate DHH users understanding of conversation content. Furthermore, we propose a scrollable chat history in a single line below the real-time captions, which can be hidden by double-clicking. This approach helps users stay connected to the ongoing conversation and avoids disrupting the immersive experience with unnecessary additional visual displays. Additionally, we indicate displaying caption delay in seconds only when it exceeded the threshold, underscoring the importance of minimizing the visual prompts [17]. This design could help avoid overwhelming the user with excessive feedback that might cause distraction or additional learning load.

Therefore, ensuring DHH users can follow the flow of conversation is essential for effective multiplayer conversations in VR. Caption design should provide coherent conversation content for DHH users while minimizing visual disruption and cognitive load, which could help create a more inclusive VR multiplayer conversation experience for a wider range of users.

Non-speech Information Transmission. Our work found that using avatars to identify the current speaker allows DHH users to access critical conversational content while maintaining minimal visual distraction. Meanwhile, it should be carefully considered to avoid obstructing the view of important content by the opaque avatars. Additionally, applying avatar-related cues offers an effective strategy to represent sound locations. This approach provides an efficient way for DHH users to access others' locations without a prior comprehensive understanding of the VR environment. Also, customizable avatars mentioned by the DHH participants could gather information about the current speaker intuitively. Furthermore, one participant from our validation study suggested enhancing the sound location by incorporating distance information, which could make the experience more engaging. This suggestion is particularly relevant to the context of this study, as it was based on a VR escape room game. Inspired by this, we propose that optimizing the sound location based on specific topics might be a valuable approach to identifying the detailed location, thus enhancing the immersive experience for DHH users in VR multiplayer conversations.

Drawing from our findings, we suggest exploring the use of avatar-based or customizable identity-related designs for DHH users in VR, as these could effectively convey critical non-speech information while minimizing unnecessary distractions and maintaining the immersive experience.

7.3 Limitation and Future Work

This paper contributes to the caption design for DHH individuals in VR multiplayer conversations. Although our caption system prototype was designed and developed based on the identified needs of DHH users, several challenges remain.

First, most current caption generation methods rely on ASR [74] for transcription, in which unclear speech easily leads to transcription errors or failures. In the co-design workshop, three social workers raised such concerns, especially for DHH users with different degrees of hearing loss. Therefore, for future research, we recommend integrating additional technologies such as cameras that can capture lip reading [3, 19] to assist in transcribing speech or sign language input. Secondly, our caption design prototype has only been applied to multiplayer games in VR for verification. We will further explore diverse multiplayer conversations as remote collaboration or education settings [57, 63]. Third, since we failed to collect feedback about the design of caption delay, further study will explore more studies to specifically focus on the design to determine whether this design could help the DHH users to better communicate in VR multiplayer conversations. Lastly, the number of participants in these studies was limited, particularly in the user study for validation, which was conducted in a controlled lab setting with only two participant groups. To address this, we plan to release the VRCaptions system to the DHH community and invite more DHH individuals to participate in future studies, allowing us to gather additional data and further refine the caption design.

8 Conclusion

In this paper, we explored different caption design directions under three dimensions, *Readability*, *Speech Information Transmission*, and *Non-Speech Information Transmission* with DHH individuals to improve their experience in VR multiplayer conversations. We first conducted three co-design workshops to understand DHH individuals' needs for caption designs in VR multiplayer conversations. From this, we identified seven design directions. Subsequently, we invited 13 DHH participants to join semi-structured interviews to collect their design preferences. Using these insights, we developed a caption prototype, VRCaptions,

and integrated it into a VR multiplayer collaborative room escape game for validation. Two groups of mixed-hearing participants were invited to experience the game and evaluate our VRCaptions. By understanding the design needs of DHH participants in VR multiplayer conversations and integrating the design based on their preferences into VRCaptions, the designs we proposed could provide an inclusive experience for DHH participants. Through these studies, we have thoroughly explored the design of a VR caption system to better support DHH users, systematically collected and organized their design needs and preferences, and provided valuable suggestions for future design of caption systems for DHH individuals in VR.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(RS-2024-00337803).

References

[1] Chadia Abras, Diane Maloney-Krichmar, Jenny Preece, et al. 2004. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications* 37, 4 (2004), 445–456.

[2] Rahaf Alharbi, John Tang, and Karl Henderson. 2023. Accessibility Barriers, Conflicts, and Repairs: Understanding the Experience of Professionals with Disabilities in Hybrid Meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 605, 15 pages. https://doi.org/10.1145/3544548.3581541

[3] Shurug Alkhalifa and Muna Al-Razgan. 2018. Enssat: wearable technology application for the deaf and hard of hearing. *Multimedia Tools and Applications* 77 (2018), 22007–22031.

[4] Oliver Alonzo, Hijung Valentina Shin, and Dingzeyu Li. 2022. Beyond subtitles: captioning and visualizing non-speech sounds to improve accessibility of user-generated videos. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–12.

[5] Akhter Al Amin, Saad Hassan, and Matt Huenerfauth. 2021. Caption-occlusion severity judgments across live-television genres from deaf and hard-of-hearing viewers. In *Proceedings of the 18th International Web for All Conference*. 1–12.

[6] Akhter Al Amin, Saad Hassan, Sooyeon Lee, and Matt Huenerfauth. 2022. Watch It, Don't Imagine It: Creating a Better Caption-Occlusion Metric by Collecting More Ecologically Valid Judgments from DHH Viewers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 459, 14 pages. https://doi.org/10.1145/3491102.3517681

[7] Akhter Al Amin, Joseph Mendis, Raja Kushalnagar, Christian Vogler, and Matt Huenerfauth. 2023. Who is Speaking: Unpacking In-Text Speaker Identification Preference of Viewers Who Are Deaf and Hard of Hearing While Watching Live Captioned Television Program. In *Proceedings of the 20th International Web for All Conference* (Austin, TX, USA) *(W4A '23)*. Association for Computing Machinery, New York, NY, USA, 44–53. https://doi.org/10.1145/3587281.3587286

[8] Larwan Berke, Khaled Albusays, Matthew Seita, and Matt Huenerfauth. 2019. Preferred Appearance of Captions Generated by Automatic Speech Recognition for Deaf and Hard-of-Hearing Viewers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312921

[9] Larwan Berke, Christopher Caulfield, and Matt Huenerfauth. 2017. Deaf and hard-of-hearing perspectives on imperfect automatic speech recognition for captioning one-on-one meetings. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. 155–164.

[10] Danielle Bragg, Nicholas Huynh, and Richard E Ladner. 2016. A personalizable mobile sound detector app design for deaf and hard-of-hearing users. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*. 3–13.

[11] Andy Brown, Rhia Jones, Mike Crabb, James Sandford, Matthew Brooks, Mike Armstrong, and Caroline Jay. 2015. Dynamic subtitles: the user experience. In *Proceedings of the ACM international conference on interactive experiences for TV and online video*. 103–112.

[12] Andy Brown, Jayson Turner, Jake Patterson, Anastasia Schmitz, Mike Armstrong, and Maxine Glancy. 2017. Subtitles in 360-degree Video. In *Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*. 3–8.

[13] Andy Brown, Jayson Turner, Jake Patterson, Anastasia Schmitz, Mike Armstrong, and Maxine Glancy. 2018. Exploring subtitle behaviour for 360 video. *White Paper WHP* 330 (2018).

[14] Loïc Caroux, Katherine Isbister, Ludovic Le Bigot, and Nicolas Vibert. 2015. Player–video game interaction: A systematic review of current concepts. *Computers in human behavior* 48 (2015), 366–381.

[15] Anna Cavender and Richard E Ladner. 2008. Hearing impairments. *Web accessibility: A foundation for research* (2008), 25–35.

[16] David David, Edward Arman, Natalia Chandra, Nadia Nadia, et al. 2019. Development of escape room game using VR technology. *Procedia Computer Science* 157 (2019), 646–652.

[17] Caluã de Lacerda Pataca, Saad Hassan, Nathan Tinker, Roshan Lalintha Peiris, and Matt Huenerfauth. 2024. Caption Royale: Exploring the Design Space of Affective Captions from the Perspective of Deaf and Hard-of-Hearing Individuals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 899, 17 pages. https://doi.org/10.1145/3613904.3642258

[18] Caluã de Lacerda Pataca, Matthew Watkins, Roshan Peiris, Sooyeon Lee, and Matt Huenerfauth. 2023. Visualization of Speech Prosody and Emotion in Captions: Accessibility for Deaf and Hard-of-Hearing Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 831, 15 pages. https://doi.org/10.1145/3544548.3581511

[19] Aashaka Desai, Jennifer Mankoff, and Richard E Ladner. 2023. Understanding and Enhancing The Role of Speechreading in Online d/DHH Communication Accessibility. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[20] Gilbert CF Fong. 2009. Let the words do the talking: The nature and art of subtitling. *Dubbing and subtitling in a world context* (2009), 91–105.

[21] Abraham Glasser, Joseline Garcia, Chang Hwang, Christian Vogler, and Raja Kushalnagar. 2021. Effect of caption width on the TV user experience by deaf and hard of hearing viewers. In *Proceedings of the 18th International Web for All Conference* (Ljubljana, Slovenia) *(W4A '21)*. Association for Computing Machinery, New York, NY, USA, Article 27, 5 pages. https://doi.org/10.1145/3430263.3452435

[22] Abraham Glasser, Vaishnavi Mande, and Matt Huenerfauth. 2020. Accessibility for deaf and hard of hearing users: Sign language conversational user interfaces. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–3.

[23] Benjamin M Gorman, Michael Crabb, and Michael Armstrong. 2021. Adaptive subtitles: preferences and trade-offs in real-time media adaption. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[24] Stephen R Gulliver and Gheorghita Ghinea. 2002. Impact of captions on deaf and hearing perception of multimedia video clips. In *Proceedings. IEEE International Conference on Multimedia and Expo*, Vol. 1. IEEE, 753–756.

[25] Stephen R Gulliver and George Ghinea. 2003. How level and type of deafness affect user perception of multimedia video clips. *Universal Access in the Information Society* 2 (2003), 374–386.

[26] Meng Guo, Lili Han, and Marta Teixeira Anacleto. 2023. Computer-Assisted Interpreting Tools: Status Quo and Future Trends. *Theory and Practice in Language Studies* 13, 1 (2023), 89–99.

[27] Ru Guo, Yiru Yang, Johnson Kuang, Xue Bin, Dhruv Jain, Steven Goodman, Leah Findlater, and Jon Froehlich. 2020. HoloSound: Combining Speech and Sound Identification for Deaf or Hard of Hearing Users on a Head-Mounted Display. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 71, 4 pages. https://doi.org/10.1145/3373625.3418031

[28] Saad Hassan, Yao Ding, Agneya Abhimanyu Kerure, Christi Miller, John Burnett, Emily Biondo, and Brenden Gilbert. 2023. Exploring the Design Space of Automatically Generated Emotive Captions for Deaf or Hard of Hearing Users. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 125, 10 pages. https://doi.org/10.1145/3544549.3585880

[29] Richang Hong, Meng Wang, Mengdi Xu, Shuicheng Yan, and Tat-Seng Chua. 2010. Dynamic captioning: video accessibility enhancement for hearing impairment. In *Proceedings of the 18th ACM international conference on Multimedia*. 421–430.

[30] Matthias Hoppe, Jakob Karolus, Felix Dietz, Paweł W. Woźniak, Albrecht Schmidt, and Tonja-Katrin Machulla. 2019. VRsneaky: Increasing Presence in VR Through Gait-Aware Auditory Feedback. In *Proceedings of*

the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3290605.3300776

[31] Chris Hughes, Mario Montagud Climent, and Peter tho Pesch. 2019. Disruptive approaches for subtitling in immersive environments. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*. 216–229.

[32] Ryo Iijima, Akihisa Shitara, Sayan Sarcar, and Yoichi Ochiai. 2021. Word Cloud for Meeting: A Visualization System for DHH People in Online Meetings. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 99, 4 pages. https://doi.org/10.1145/3441852.3476547

[33] Anna Sigridur Islind, Johan Lundin, Katerina Cerna, Tomas Lindroth, Linda Åkeflo, and Gunnar Steineck. 2023. Proxy design: a method for involving proxy users to speak on behalf of vulnerable or unreachable users in co-design. *Information Technology & People* (2023).

[34] Dhruv Jain, Bonnie Chinh, Leah Findlater, Raja Kushalnagar, and Jon Froehlich. 2018. Exploring Augmented Reality Approaches to Real-Time Captioning: A Preliminary Autoethnographic Study. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems* (Hong Kong, China) *(DIS '18 Companion)*. Association for Computing Machinery, New York, NY, USA, 7–11. https://doi.org/10.1145/3197391.3205404

[35] Dhruv Jain, Leah Findlater, Jamie Gilkeson, Benjamin Holland, Ramani Duraiswami, Dmitry Zotkin, Christian Vogler, and Jon E Froehlich. 2015. Head-mounted display visualizations to support sound awareness for the deaf and hard of hearing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 241–250.

[36] Dhruv Jain, Rachel Franz, Leah Findlater, Jackson Cannon, Raja Kushalnagar, and Jon Froehlich. 2018. Towards Accessible Conversations in a Mobile Context for People Who Are Deaf and Hard of Hearing. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. Association for Computing Machinery, New York, NY, USA, 81–92. https://doi.org/10.1145/3234695.3236362

[37] Dhruv Jain, Sasa Junuzovic, Eyal Ofek, Mike Sinclair, John Porter, Chris Yoon, Swetha Machanavajhala, and Meredith Ringel Morris. 2021. A Taxonomy of Sounds in Virtual Reality. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 160–170. https://doi.org/10.1145/3461778.3462106

[38] Dhruv Jain, Sasa Junuzovic, Eyal Ofek, Mike Sinclair, John R. Porter, Chris Yoon, Swetha Machanavajhala, and Meredith Ringel Morris. 2021. Towards Sound Accessibility in Virtual Reality. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 80–91. https://doi.org/10.1145/3462244.3479946

[39] Numair Khan and Anis Ur Rahman. 2018. Rethinking the mini-map: A navigational aid to support spatial learning in urban game environments. *International Journal of Human–Computer Interaction* 34, 12 (2018), 1135–1147.

[40] Yeon Soo Kim, Hyeonjeong Im, Sunok Lee, Haena Cho, and Sangsu Lee. 2023. "We Speak Visually": User-Generated Icons for Better Video-Mediated Mixed-Group Communications Between Deaf and Hearing Participants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 610, 16 pages. https://doi.org/10.1145/3544548.3581151

[41] Yeon Soo Kim, Sunok Lee, and Sangsu Lee. 2022. A Participatory Design Approach to Explore Design Directions for Enhancing Videoconferencing Experience for Non-signing Deaf and Hard of Hearing Users. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 47, 4 pages. https://doi.org/10.1145/3517428.3550375

[42] Raja Kushalnagar. 2019. A Classroom Accessibility Analysis App for Deaf Students. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 569–571. https://doi.org/10.1145/3308561.3354640

[43] Raja Kushalnagar, Matthew Seita, and Abraham Glasser. 2017. Closed ASL interpreting for online videos. In *Proceedings of the 14th International Web for All Conference*. 1–4.

[44] Raja S Kushalnagar, Walter S Lasecki, and Jeffrey P Bigham. 2013. Captions versus transcripts for online video content. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. 1–4.

[45] Raja S. Kushalnagar and Christian Vogler. 2020. Teleconference Accessibility and Guidelines for Deaf and Hard of Hearing Users. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 9, 6 pages. https://doi.org/10.1145/3373625.3417299

[46] Ziming Li, Shannon Connell, Wendy Dannels, and Roshan Peiris. 2022. SoundVizVR: sound indicators for accessible sounds in virtual reality for deaf or hard-of-hearing users. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–13.

[47] Ziming Li, Kristen Shinohara, and Roshan L Peiris. 2023. Exploring the Use of the SoundVizVR Plugin with Game Developers in the Development of Sound-Accessible Virtual Reality Games. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.

[48] Sanzida Mojib Luna, Jiangnan Xu, Konstantinos Papangelis, Garreth W. Tigwell, Nicolas Lalone, Michael Saker, Alan Chamberlain, Samuli Laato, John Dunham, and Yihong Wang. 2024. Communication, Collaboration, and Coordination in a Co-located Shared Augmented Reality Game: Perspectives From Deaf and Hard of Hearing People. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 36, 14 pages. https://doi.org/10.1145/3613904.3642953

[49] Imran Mahalil, Azmi Mohd Yusof, Nazrita Ibrahim, Eze Manzura Mohd Mahidin, and Mohd Ezanee Rusli. 2019. Virtual reality mini map presentation techniques: lessons and experience learned. In *2019 IEEE Conference on Graphics and Media (GAME)*. IEEE, 26–31.

[50] Emma J. McDonnell, Ping Liu, Steven M. Goodman, Raja Kushalnagar, Jon E. Froehlich, and Leah Findlater. 2021. Social, Environmental, and Technical: Factors at Play in the Current Use and Future Design of Small-Group Captioning. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 434 (oct 2021), 25 pages. https://doi.org/10.1145/3479578

[51] Emma J McDonnell, Soo Hyun Moon, Lucy Jiang, Steven M. Goodman, Raja Kushalnagar, Jon E. Froehlich, and Leah Findlater. 2023. "Easier or Harder, Depending on Who the Hearing Person Is": Codesigning Videoconferencing Tools for Small Groups with Mixed Hearing Status. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 780, 15 pages. https://doi.org/10.1145/3544548.3580809

[52] Dorian Miller, Karl Gyllstrom, David Stotts, and James Culp. 2007. Semi-transparent video interfaces to assist deaf persons in meetings. In *Proceedings of the 45th annual southeast regional conference*. 501–506.

[53] Mohammadreza Mirzaei, Peter Kán, and Hannes Kaufmann. 2020. EarVR: Using Ear Haptics in Virtual Reality for Deaf and Hard-of-Hearing People. *IEEE Transactions on Visualization and Computer Graphics* 26, 5 (2020), 2084–2093. https://doi.org/10.1109/TVCG.2020.2973441

[54] Alex Olwal, Kevin Balke, Dmitrii Votintcev, Thad Starner, Paula Conn, Bonnie Chinh, and Benoit Corda. 2020. Wearable Subtitles: Augmenting Spoken Communication with Lightweight Eyewear for All-day Captioning. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 1108–1120. https://doi.org/10.1145/3379337.3415817

[55] Kotaro Oomori, Akihisa Shitara, Tatsuya Minagawa, Sayan Sarcar, and Yoichi Ochiai. 2020. A Preliminary Study on Understanding Voice-only Online Meetings Using Emoji-based Captioning for Deaf or Hard of Hearing Users. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 54, 4 pages. https://doi.org/10.1145/3373625.3418032

[56] Rui Pan, Henry Lo, and Carman Neustaedter. 2017. Collaboration, awareness, and communication in real-life escape rooms. In *Proceedings of the 2017 conference on designing interactive systems*. 1353–1364.

[57] Prajwal Paudyal, Ayan Banerjee, Yijian Hu, and Sandeep Gupta. 2019. Davee: A deaf accessible virtual environment for education. In *Proceedings of the 2019 Conference on Creativity and Cognition*. 522–526.

[58] Yi-Hao Peng, Ming-Wei Hsi, Paul Taele, Ting-Yu Lin, Po-En Lai, Leon Hsu, Tzu-chuan Chen, Te-Yen Wu, Yu-An Chen, Hsien-Hui Tang, and Mike Y. Chen. 2018. SpeechBubbles: Enhancing Captioning Experiences for Deaf and Hard-of-Hearing People in Group Conversations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3173867

[59] Sandra Poeschl, Konstantin Wall, and Nicola Doering. 2013. Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence. In *2013 IEEE Virtual Reality (VR)*. 129–130. https://doi.org/10.1109/VR.2013.6549396

[60] Nikhita Praveen, Naveen Karanth, and MS Megha. 2014. Sign language interpreter using a smart glove. In *2014 international conference on advances in electronics computers and communications*. IEEE, 1–5.

[61] Sylvia Rothe, Kim Tran, and Heinrich Hussmann. 2018. Positioning of Subtitles in Cinematic Virtual Reality.. In *ICAT-EGVE*. 1–8.

[62] Jazz Rui Xia Ang, Ping Liu, Emma McDonnell, and Sarah Coppola. 2022. "In this online environment, we're limited": Exploring Inclusive

Video Conferencing Design for Signers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 609, 16 pages. https://doi.org/10.1145/3491102.3517488

[63] Yasith Samaradivakara, Thavindu Ushan, Asela Pathirage, Prasanth Sasikumar, Kasun Karunanayaka, Chamath Keppitiyagama, and Suranga Nanayakkara. 2024. SeEar: Tailoring Real-time AR Caption Interfaces for Deaf and Hard-of-Hearing (DHH) Students in Specialized Educational Settings. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[64] Chris Schipper and Bo Brinkman. 2017. Caption Placement on an Augmented Reality Head Worn Device. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 365–366. https://doi.org/10.1145/3132525.3134786

[65] Matthew Seita. 2020. Designing automatic speech recognition technologies to improve accessibility for deaf and hard-of-hearing people in small group meetings. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[66] Matthew Seita, Sarah Andrew, and Matt Huenerfauth. 2021. Deaf and hard-of-hearing users' preferences for hearing speakers' behavior during technology-mediated in-person and remote conversations. In *Proceedings of the 18th International Web for All Conference*. 1–12.

[67] Clay Spinuzzi. 2005. The methodology of participatory design. *Technical communication* 52, 2 (2005), 163–174.

[68] Ippei Suzuki, Kenta Yamamoto, Akihisa Shitara, Ryosuke Hyakuta, Ryo Iijima, and Yoichi Ochiai. 2022. See-through captions in a museum guided tour: exploring museum guided tour for deaf and hard-of-hearing people with real-time captioning on transparent display. In *International Conference on Computers Helping People with Special Needs*. Springer, 542–552.

[69] Mauro Teófilo, Alvaro Lourenço, Juliana Postal, and Vicente F Lucena. 2018. Exploring virtual reality to enable deaf or hard of hearing accessibility in live theaters: A case study. In *Universal Access in Human-Computer Interaction. Virtual, Augmented, and Intelligent Environments: 12th International Conference, UAHCI 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings, Part II 12*. Springer, 132–148.

[70] Christian Vogler, Paula Tucker, and Norman Williams. 2013. Mixed local and remote participation in teleconferences from a deaf and hard of hearing perspective. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–5.

[71] Y Wang. 2006. Discussion on Technical Principle for Handling with Translation of Captions of Movies and Televisions. *Journal of Hebei Polytechnic College* 6, 1 (2006), 61–63.

[72] Feng Wen, Zixuan Zhang, Tianyiyi He, and Chengkuo Lee. 2021. AI enabled sign language recognition and VR space bidirectional communication using triboelectric smart glove. *Nature Communications* 12, 1 (10 Sep 2021), 5378. https://doi.org/10.1038/s41467-021-25637-w

[73] Kenta Yamamoto, Ippei Suzuki, Akihisa Shitara, and Yoichi Ochiai. 2021. See-through captions: real-time captioning on transparent display for deaf and hard-of-hearing people. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–4.

[74] Dong Yu and Lin Deng. 2016. *Automatic speech recognition*. Vol. 1. Springer.

[75] Krzysztof Zagata, Jacek Gulij, Łukasz Halik, and Beata Medyńska-Gulij. 2021. Mini-map for gamers who walk and teleport in a virtual stronghold. *ISPRS International Journal of Geo-Information* 10, 2 (2021), 96.