

SpeechCap: Leveraging Playful Impact Captions to Facilitate Interpersonal Communication in Social Virtual Reality

YU ZHANG, City University of Hong Kong, China

YI WEN, Texas A&M University, USA and City University of Hong Kong, China

SIYING HU, City University of Hong Kong, China

ZHICONG LU, George Mason University, USA



Fig. 1. SpeechCap showcases a novel way to mediate voice conversations in social VR by converting speech into playful impact captions. (A) From real-time speech, SpeechCap generates impact captions with customized visual design that conveys both verbal and non-verbal information. The caption “happy” is in a warm and bright color for the positive emotions, and a “smiling” emoji further highlight the pleasant and relaxing atmosphere. (B) Users can play with impact captions like real-world objects, e.g., throwing a “cake”. (C) Once the flying “cake” crashes on another user, it will explode like a firework.

Social Virtual Reality (VR) emerges as a promising platform bringing immersive, interactive, and engaging mechanisms for collaborative activities in virtual spaces. However, interpersonal communication in social VR is still limited with existing mediums and channels. To bridge the gap, we propose a novel method for mediating real-time conversation in social VR, which uses impact captions, a type of typographic visual effect widely used in videos, to convey both verbal and non-verbal information. We first investigated the design space of impact captions by content analysis and a co-design session with four experts. Next, we implemented SpeechCap as a proof-of-concept system, with which users can communicate with each other using speech-driven impact captions in VR. Through a user study ($n=14$), we evaluated the effectiveness of the visual and interaction design of impact captions, highlighting the interactivity and the integration of verbal and non-verbal information in communication mediums. Finally, we discussed topics of visual rhetoric, interactivity, and ambiguity as the main findings from the study, and further provided design implications for future work for facilitating interpersonal communication in social VR.

CCS Concepts: • **Human-centered computing** → **Virtual reality**; **Interactive systems and tools**; **Empirical studies in interaction design**.

Authors' Contact Information: Yu Zhang, yui.zhang@my.cityu.edu.hk, Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China; Yi Wen, cyberwenyi2357@tamu.edu, Texas A&M University, College Station, Texas, USA, School of Creative Media and City University of Hong Kong, Hong Kong SAR, China; Siying HU, sying.ch1026@gmail.com, Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China; Zhicong Lu, zlu6@gmu.edu, George Mason University, Fairfax, Virginia, USA.



This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2025 Copyright held by the owner/author(s).

ACM XXXX-XXXX/2025/2-ART

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

Additional Key Words and Phrases: Social Virtual Reality, Interpersonal Communication, Impact Captions

ACM Reference Format:

Yu Zhang, Yi Wen, Siying HU, and Zhicong Lu. 2025. SpeechCap: Leveraging Playful Impact Captions to Facilitate Interpersonal Communication in Social Virtual Reality. 1, 1 (February 2025), 34 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

With the increasing availability of consumer-level virtual reality head-mounted displays (HMDs) [7, 55], Social Virtual Reality (VR) emerges as a promising platform that brings immersive, interactive, and engaging experiences to users for a wide range of collaborative activities. Social VR now supports remote conferencing [1], immersive learning [78, 91], healthcare [53, 93], collaborative design [71], music composing [72] and daily activities such as sleeping [65] and even drinking [18]. Among these diverse application areas, social VR helps users overcome physical dispersion, enabling them to communicate and collaborate with each other in the same virtual environment [55, 77] by offering channels supporting both verbal and non-verbal communication [29, 68, 98]. Text messaging and voice chat are two channels commonly used for verbal communication. Pictures [54], avatars [12, 32, 33, 47], and embodied interactions [66, 99] are channels that can convey non-verbal information in social VR.

Yet, these existing communication channels in social VR are preliminary, as they are either borrowed from traditional communication media (e.g., texts, voices, and pictures) lacking adaptations to the nature of social VR [29, 38, 74, 81], or difficult to be reproduced very well by current VR technologies (e.g., gestures, facial expressions, embodied actions) [88, 90, 98]. Specifically, texts as a communication medium in VR still suffer from difficulties in user inputting [74] and readability issues in presentations [38, 81]. Voice-based conversation in VR can easily be disturbed by sound intrusions from the environment [4]. Furthermore, non-verbal communication mediums, such as gestures and facial expressions, cannot be well-simulated with the avatars in current VR applications [10, 12, 32], and users may find it hard to achieve specific actions or movements using controllers with joysticks and buttons [54, 88, 90].

Except for social VR, previous research in HCI and CSCW has explored ways to enhance interpersonal communication in different scenarios. Those scenarios include video conferencing [60, 102], augmented reality (AR) [50, 52, 57], and text messaging interfaces [9]. These different scenarios demonstrate the power of the integration of both verbal and non-verbal cues in the mediums to support interpersonal communication. On the other hand, researchers have also envisioned the potential of social VR in providing situational awareness and enabling new social interactions with novel communication mediums that are unattainable in traditional computer-mediated communication (CMC) platforms [68, 69].

On top of the prior work, we proposed a new approach to support interpersonal communication in social VR. Our approach is inspired by *Impact Captions*, a type of expressive typographic 2D visual effect with engaging visual elements and animated motions. Impact captions are traditionally used in entertaining television shows and online videos [76, 82] to attract viewers' attention and create engaging watching experiences [83]. Particularly, we introduced impact captions as a communication medium to facilitate interpersonal communication and interactions in social VR. In our design, impact captions encode both verbal and non-verbal information extracted from real-time speech; they also offer interactivity that allows users to intentionally play with the captions for communicative purposes.

To implement and evaluate our impact-caption-inspired approach, we first explored the design space of impact captions regarding the visual appearances and interactivity through a content analysis of a collection of TV show videos. Based on the design space, we held a co-design session

with 4 experts (i.e., HCI researchers) to further derive specific design goals for building a system to demonstrate our idea. Following the findings so far, we developed SpeechCap, a proof-of-concept system that takes real-time speech as input to generate interactive impact captions with customized visual appearances in a multi-user VR environment. Using SpeechCap, we conducted an in-lab study with 14 participants to assess impact captions as a medium for supporting interpersonal communication in social VR.

The user study results highlighted the integration of verbal and non-verbal information offered by impact captions, indicating the design can make social VR communication experiences not only clear but also engaging, while the risks of miscommunication caused by ambiguity inherent from the complex visual design were noticed. Moreover, we found that impact captions can also facilitate interactions between users beyond enhancing speech conversations in social VR. Furthermore, we illustrated three application scenarios to demonstrate the generalizability of our impact-caption-inspired approach. Finally, this research ends up with discussions on the findings from the user study, in which we provided insights and implications for future work and explained the limitations and possible improvements on the scalability of the current SpeechCap system.

In summary, this research contributes: (i) A comprehensive design space of impact captions in terms of visual and interaction design for supporting interpersonal communication; (ii) A proof-of-concept system, SpeechCap, that enables users to conduct real-time conversations in social VR with the help of interactive impact captions; (iii) An evaluation study of SpeechCap that demonstrates the effectiveness of our impact-caption-inspired approach for communication, highlighting the value of employing creative mediums for supporting interpersonal communication in social VR.

2 Related Work

Starting from the background of interpersonal communication in social virtual reality (VR), we highlight the needs for better ways to support communication with both verbal and non-verbal information, and then introduce previous research systems to augment computer-mediated communication (CMC), which inspired our work. Finally, we introduce impact captions and captioning systems for communication purposes in HCI, showing our motivation for this research.

2.1 Interpersonal Communication in Social Virtual Reality

Social Virtual Reality (VR) refers to immersive virtual spaces where multiple users can synchronously interact with each other through VR head-mounted displays (HMDs) [32, 70]. It provides high-fidelity physical presence to facilitate various forms of interpersonal communication with both verbal and non-verbal cues over other computer-mediated communication channels [68, 104]. With an immersive and synchronous manner, social VR can effectively support collaborative tasks and promote social interactions in a wide range of application scenarios, including collaborative prototyping [71], healthcare and treatment [53, 93], immersive learning [42, 78, 91], intimate relationship building [31, 96], and inter-generational communication [11, 85, 97].

Although social VR has been increasingly applied in various domains, it remains far from being a socially and emotionally fulfilling digital space for users' socialization needs [66, 90, 98]. Previous research identified several limitations regarding interpersonal communication in social VR. A known issue is that current social VR applications do not allow users to easily express emotions (i.e., moods and excitements) and intentions through non-verbal cues that are common in real-world conversations [66, 88, 101], because virtual avatars reduce or even filter out several important non-verbal signals such as facial expressions and body language [12, 32, 33, 51, 105]. As an extreme case, imperfect avatars with realism will lead to the uncanny valley effect, bringing uncomfortable social experiences to users [48, 49]. Besides the lack of expressiveness, the voice-based communication,

as a common way to communicate in social VR, encounters the difficulty of coordinating speakers in multi-user virtual spaces [103].

Recent research proposed several computational methods to improve interpersonal communication experiences in social VR, such as generating realistic avatars that could capture users' facial expressions and body movements using deep learning models [94], adding haptic technologies to facilitate co-presence to allow affective communication [3, 30], and simulating spatial sound effects to enhance the convenience and flexibility of voice-based communication [103]. In summary, all of these efforts aim to make the interpersonal communication in social VR close to the situations of the realistic world.

However, people appreciate social VR not because it can simulate the real world. Instead, a significant reason is that the unique affordances of social VR can offer users immersive and unrealistic social experiences [32, 66]. In other words, social VR can augment social signaling and unlock new social interactions that are unattainable in the realistic world [68, 69]. At present, the potential of virtual reality to create unrealistic experiences for enhancing interpersonal communication for socialization needs is still underexplored. This leaves us opportunities to dive deep into the potentials of virtual reality and build novel tools with the idea of "superpowers" [68, 69] to support interpersonal communication in social VR.

2.2 Combining Verbal and Non-verbal Information to Augment Communication

Both verbal and non-verbal information is crucial in interpersonal communication in social VR [77]. Besides verbal cues that mainly convey semantic meanings, non-verbal cues convey emotions [58, 63], enhance understanding [2], regulate interactions [66], reflect cultural norms [31], and overall enrich the communication experiences no matter in offline face-to-face scenarios or online digital spaces.

Recent research in the HCI and CSCW fields has explored ways to involve non-verbal signals into verbal-centric communications (e.g., live presentations) to facilitate interpersonal communication and build social connections through computational approaches [6, 9, 16, 19, 21, 57, 60]. A thread of research aims to provide visual aids for computer-mediated spoken presentations. "RealityTalk" [57] explores a speech-driven approach that allows users to give a speech with predefined visual elements while mentioning particular key words. "Visual Captions" [60] and "CrossTalk" [102] further automate the process by introducing machine-learning-based models to predict users' intentions and automatically provide potential visual elements while processing speech input. "Elastica" [16] further enables adaptive animations on the associated visuals to effectively enhance the expressiveness of presentations.

Beyond visual aids, affective enhancement is another important dimension explored by previous research. For voice messaging interfaces, "Emowear" [6] invents a concept named "Emotional Teasers", which refers to a collection of animated emoticons, to explicitly show the emotional tone of an upcoming voice message for smartwatch users. For text-centered communication mediums such as text messaging interfaces, previous research has explored how the color and shape of speech bubbles [9, 19], and the typeface of textual content [21, 27] can be purposefully designed and computationally generated to convey the speaker's moods in conversations.

However, still less research explores how communication mediums in social VR can borrow the ideas from other CMC forms to enhance the interpersonal communication experiences, especially with the considerations of integrating both verbal and non-verbal information. So far, previous research revealed that non-verbal cues play an important and unique role in interpersonal communication in social VR. In unrealistic virtual spaces, users tend to engage in social interactions with more active and bold non-verbal behaviors [18, 65]. This phenomenon reflects a "proteus effect" in

which users intentionally manipulate virtual avatars to express themselves beyond pure verbal-language-based communication [66]. With more application scenarios and novel user activities emerging in social VR, strong demands for tools for facilitating interpersonal communication and interactions arise [90].

2.3 Impact Captions: A Typographic-Centered Visual Design Beyond Pure Verbal Communication

Impact captions refer to a type of typographic-centered visual effect that is prevalent in TV shows and online videos to engage viewers [82]. Traditionally, impact captions are used as a supplementary channel that provides information outside of the screen (e.g., a commentary message from the TV show's editor [76]) or visualizes the implicit non-verbal information in frames (e.g., the characters' moods [76] or environmental sound effects [95]). Unlike visual cues in previous related works for improving interpersonal communication [6, 9, 16, 19, 21, 57, 60], impact captions always use text as the main visual component that conveys verbal information and apply artistic modifications with additional visual decorations to textual elements to convey non-verbal cues simultaneously. The integration of both verbal and non-verbal information in the visual design of impact captions makes the captions a powerful medium that can provide rich semantics, afford viewers' affective reactions, and provoke thoughts [22, 76, 83].

HCI researchers have noticed the potential of captioning mechanisms in supporting communication and have built captioning systems to help people with deaf and hard-of-hearing (DHH) situations perceive in-depth information in videos [15, 27, 45, 84, 95]. In these cases, modifications on textual elements (e.g., irregular typefaces, coloring, and introducing dynamic visual effects) of subtitles have been proved to be effective in conveying semantic information of both verbal words and implicit non-verbal content of videos [27, 45, 84]. Yet, their explorations have not reached the scope of immersive media, such as VR. When considering textual captions in VR, another line of research indicates that the reading experience of texts in VR varies depending on the displaying conditions [81], and prolonged reading would reduce the comfort of using VR [92]. Yet, it is also unknown whether typographic-centered impact captions can effectively mediate communication in social VR.

To fill the research gap on tools for interpersonal communication in social VR and to respond to the needs of social VR users, this work explores how impact captions, as a type of elegant typographic design, can be applied to mediate and facilitate interpersonal communication in social VR with the inspirations from previous research on augmenting computer-mediated communication using visual cues and captioning mechanisms. By investigating a design space for using impact captions as a communication tool and conducting a study with the proof-of-concept system, we demonstrate a concrete example envisioning the potentials of social VR in interpersonal communication and socialization.

3 Design Space of Impact Captions

As a form of typographic design as well as visual rhetoric, impact captions can convey both verbal and non-verbal information on multiple dimensions. To introduce impact captions into social VR as a communication medium, we investigated the design space of impact captions with a concentration on the visual components and interactions.

3.1 Methods: Analyzing Videos with Impact Captions

To obtain a comprehensive understanding of impact captions currently used in videos (e.g., TV shows and online videos), we conducted a content analysis on a collection of videos of TV show recordings retrieved from popular online video-sharing platforms.

3.1.1 Data Collection. We collected relevant videos using a two-stage approach. In the first stage, we followed the top TV series rankings (i.e., IMDb¹, TV Time², and Douban³) to collect and review videos of three representative variety show series: “Arashi” (from Japan), “Running Man” (from South Korea), and “Who is the Murderer” (from China). These videos helped us get an initial sense and identify preliminary dimensions of the design space. In the second stage, we iteratively enhanced the design space by analyzing videos from an enlarged sample video set collected from online video-sharing platforms including YouTube and Bilibili. Specifically, we used the keywords “variety show with captions” and “popular variety shows” to search on these platforms with a newly registered account on YouTube and Bilibili, respectively. For each keyword, we focused on the top 30 videos of the raw results and filtered out irrelevant or redundant videos. As a result, a total of 46 sample videos were included in the dataset.

3.1.2 Content Analysis. Three authors collaboratively conducted content analysis on the video collection to build the design space. First, three authors randomly selected a subset of videos (N=10) to individually code the visual elements and behaviors of the impact captions in the videos, with the aim of determining initial dimensions. Next, three authors sat together to reflect and discuss their initial insights and ultimately generated a consensus codebook. Finally, three authors reviewed and analyzed the complete data set. Consequently, we categorized the impact caption design space into visual space and interaction space, in which visual space focuses on textual and non-textual elements (Figure 2), and interaction space (Figure 3) relates to motions, physicalization, and embodied interactive operations.

3.2 Visual Design Space of Textual Elements

Textual elements are normally the primary visual component in an impact caption, including words, phrases, sentences, and symbols with linguistic meaning. In TV shows and videos, the content of textual elements comes from different sources such as narrative voiceover, conversations, onomatopoeic words, and commentary with ironic or entertaining implications added during the post-editing process of TV show production [82].

Typeface (font type), size, and color are three primary visual dimensions for designing textual elements. Depending on the contexts and purposes, all of these dimensions can be used to encode the importance of information (e.g., highlighting keywords) or represent a character’s emotion (e.g., explicitly showing moods).

3.2.1 Typeface. Typeface denotes the fundamental visual style of text and affects the reader’s perceptions and emotional responses [5, 13]. Impact captions in videos widely utilize typeface to make impressions for audiences [83]. However, there is no strict standard that defines which typeface should be used under which circumstance. In our design space, we summarized the empirical rules based on content analysis on TV show videos.

A key perspective of the typeface used in the impact caption is the regularity of word characters’ strokes. Artistic typefaces with irregular strokes are usually used to exaggerate characters’ moods or to emphasize surprising situations. The rich shapes of the strokes allow the typeface to match changing emotions and vivacious performance. Relatively, formal and regular typefaces (e.g., serif fonts) commonly appear in impact captions that are used as annotations with neutral information. This kind of caption provides additional knowledge to the original video content and should be optimized for viewers to read. In addition, the visual style of the texts can also be designed to fit the

¹https://www.imdb.com/?ref_=nv_home

²<https://www.tvtime.com/>

³<https://m.douban.com/tv/tvshow>

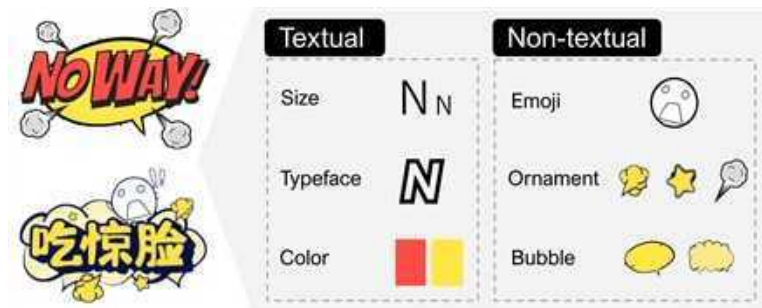


Fig. 2. Visual Design Space of Impact Captions. Impact captions consist of textual and non-textual visual elements. Textual elements include the typeface, color, and size of texts. Non-textual elements include emoji, ornament, and speech bubble.

overall visual theme of the video in TV shows. In other words, various deeply customized typefaces were created for impact captions in current TV shows beyond standard typefaces used in common computer interfaces such as desktop and mobile phones.

3.2.2 Size. The size of captions is another dimension commonly used to encode semantic meanings. Size can be used for several purposes, depending on the creator’s design intent and the overall atmosphere and theme of the context. Texts with changeable sizes attract attention more effectively, making the subtitles stand out on the screen and piquing viewers’ interest. The use of larger fonts helps to emphasize key information, punchlines, or significant events. Additionally, larger fonts and flashy designs enhance the show’s entertainment value, contributing to a lighthearted and enjoyable experience for the viewers. Text size can also help create a specific atmosphere, with larger fonts conveying liveliness and excitement, while smaller fonts may evoke warmth and intimacy. In terms of readability, larger text is easier to read on television screens, ensuring that the information is easily accessible to the audience. Additionally, highly abstract semantic meanings, such as strength, effectiveness, and severity, can be encoded by text size in impact captions.

3.2.3 Color. In TV shows and videos, the color design of impact captions aims to create the show’s atmosphere, emphasize key information, and attract the audience’s attention. Thus, the color of impact captions usually features vibrant and eye-catching styles. For example, bright colors are usually used to add liveliness and appeal to entertainment-oriented shows, making captions more noticeable on the screen.

Besides, the color of impact captions is also used to imply the moods of characters, which reflects the color-emotion associations in human perceptions and self-expressions [35, 79]. Empirical evidence shows that bright and highly-saturated colors tend to associate with positive feelings, while dark and low-saturated colors may associate with negative feelings [19, 35, 79, 100]. However, there are no standard rules to tell which mood should be represented by which color. Because the color-emotion associations are not strict one-to-one mapping rules. A single color may reflect more than one type of emotion and vice versa [35, 100]. Moreover, the perceptions of emotion conveyed by the same color also vary among different groups of people [19, 35, 100].

Nevertheless, the color-emotion associations still highlight the potential of using color to convey emotional information for interpersonal communication bi-directionally. Color can help users receive emotional information from others and express emotional feelings of themselves.

3.3 Visual Design Space for Non-verbal Elements

Beside textual elements, several *non-verbal elements*, such as typographic ornaments, speech balloons, emojis (emoticons), are involved in impact caption. Those elements provide visual cues to convey invisible information, enhance expressiveness, and leverage engaging communication experiences.

3.3.1 *Typographic ornaments.* Typographic ornaments refer to decorative elements or symbols used to embellish and enhance the appearance of impact captions. They are typically small, non-alphabetic characters that add artistic flair and visual interest to the text. Common examples of ornaments include swashes, which are elegant flourishes on letters often seen in decorative scripts; fleurons, floral or leaf-like designs used as separators or accents; bullets, various shapes used for lists and highlighting points; dingbats, pictorial symbols representing a wide range of subjects; and borders and rule lines for separating text. In current impact captions, the content of ornaments is flexible and can be deeply customized to fit text contexts.

3.3.2 *Speech bubbles.* A speech bubble (a.k.a. speech balloon, ornamental frame) is the decorative visual element around text as a container of impact caption. This concept comes from comics and graphic novels and is used to represent the speech or thoughts of characters [9]. It is a rounded or elliptical shape with a tail or pointer that connects to the character's mouth, indicating the source of the speech. Inside the bubble, the character's dialogue or thoughts are written in text to convey their words or inner monologue. In addition to speech, a speech bubble can also represent a character's inner thoughts, visualize loud or intense sounds, and provide narrative captions (additional narration or context to the story). Speech balloons play a crucial role in visual storytelling, as they allow audiences to understand and follow the characters' conversations and emotions. Different styles of speech balloons can be used to depict various emotions or tones of speech.

3.3.3 *Emojis and Emoticons.* Emojis are actual graphical images that depict a wide range of subjects, including facial expressions, animals, food, weather, etc. Similarly, an emoticon is a combination of characters, usually made from keyboard symbols, that represents a facial expression or an emotion [61]. Emoticons are commonly used in text-based communication, such as texting, messaging, social media posts, and online forums, to convey feelings or reactions that may not be easily expressed through text alone. Emojis and emoticons can be integrated into impact captions to enhance the emotional tone, clarity, and expressiveness.

3.4 Interactivity Design Space

When bringing impact captions from video-based media to VR, a significant change is that impact captions can be "alive" beyond unchangeable visual appearances and motions. In virtual reality, impact captions can be made interactive to respond to user input, including voice and embodied actions such as gestures and body movements. By playing with these virtual textual-visual elements, users can make fun, express themselves, and interact with others.

3.4.1 *Physicalization.* Using the physical engine provided by current VR technology, impact captions can be endowed with various physical effects. These physical effects allow impact captions to behave like objects in the real world, helping users establish perception beyond language. For impact captions with specific meanings, physical sensations can be employed to enhance the semantics of the text. Our design space includes two of the most common physical quantities: mass and velocity [Figure 3](#). Mass directly relates to people's perception of "light" and "heavy". When impact captions are assigned mass, gravity effects can be applied to them so that they would naturally convey the degree of abstraction in terms of weight. For example, assigning gravitational

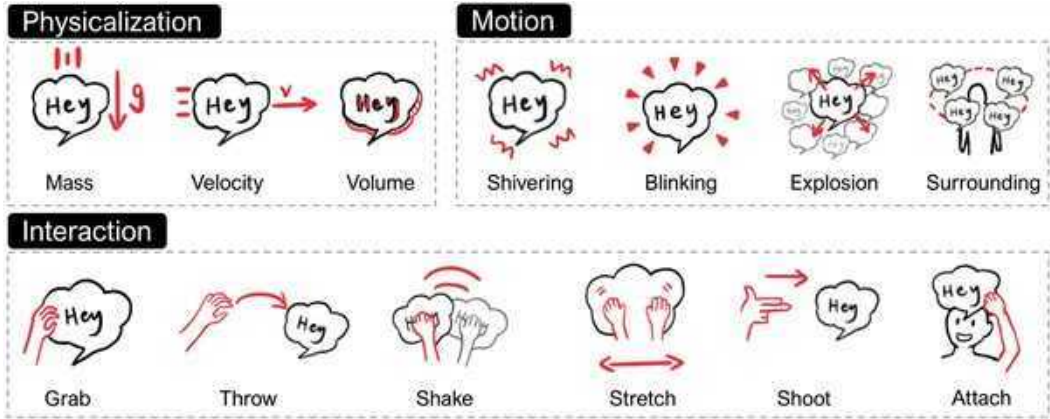


Fig. 3. Interactivity Design Space of Impact Captions. The design space include three dimensions: **Physicalization**, **Motion**, and **Interaction**. With **Physicalization**, impact captions can appear like physical objects to be affected by the gravity with mass, have velocity for movement, and take spaces with volume. With **Motion**, impact captions appear to be “alive” and be responsive to user actions. **Interaction** describes how users can play with impact captions using embodied interaction.

acceleration to an impact caption with the content “heavy-hearted” results in a dropping motion, physically conveying the feeling of “heaviness”. The concept of velocity is related to time. When adjusting the motion speed of impact captions, it can enhance the sense of urgency in expressing the textual content.

3.4.2 Motion. Motions make impact captions alive and attractive by endowing changing appearances. Motions can be independent (i.e., only affecting an impact caption itself) or dependent (i.e., relying on other objects to take effect). In our design space, we include three commonly observed independent animated motions: shivering, blinking, and explosion [Figure 3](#). When shivering, primary visual components (i.e., textual and non-textual) of an impact caption will swing periodically. When blinking, the color of visual components (mainly text) could change rapidly and keep glowing. Explosion is an effect where an original caption splits into several pieces of fragments or replicas spread around the whole space. As for dependent motions, character (or avatar in VR) is the most common object with which impact captions can interact. Surrounding is a useful motion where impact captions move following an orbit. Surrounding impact captions usually represent the unique personalities or properties of avatars.

3.4.3 Interaction. Embodied interaction emphasizes the integration of the human body and physical actions, such as gestures [\[63\]](#) and bodily movements [\[75\]](#), as integral components of interaction with digital technology [\[46, 86\]](#). Specifically, to support interpersonal communication using impact captions in VR, we recognized four common gestures (i.e., drag, drop, stretch, and throw) and one interaction with the virtual body (avatar) involved: attaching a caption with the avatar’s body. These embodied interactions can be used for either caption appearance customization or triggering particular actions. In this way, users can intuitively manipulate captions by their own hands rather than relying on a floating menu and clicking virtual buttons.

4 Co-designing an Impact-Caption-Inspired Communication Tool in Social VR

To concretize the design space to a feasible system design for our impact-caption-inspired approach, we conducted a co-design session with four HCI researchers and the four authors of this paper.

From the session, we obtained primary directions and design goals for building the proof-of-concept system.

4.1 Procedure

The co-design procedure contains two sub-sessions. The first sub-session focused on brainstorming and open-ended discussion, and the second sub-session aimed at paper prototyping.

4.1.1 Participants. Besides the four authors of this paper, we recruited four extra HCI researchers from our university to our co-design session. During the process, the first author played the role of instructor to host the sessions and also participated in the activities to contribute ideas. The eight contributors are either developers or designers of digital products. They all have rich experiences in VR gaming and social VR (e.g., VRChat⁴).

4.1.2 Sub-Session 1: Brainstorming and Discussion. In this sub-session, the main objective was to understand the scenarios, needs, and purposes of users' interpersonal communication behaviors in social VR settings. This session was followed by initial brainstorming and design discussions. Participants were first asked to share their experiences and pain points to communicate with other VR players and discuss possible ways to address the issues. The instructor then presented several examples of impact captions (sourced from example videos collected during the design space investigation stage) to illustrate the concept and potential use cases of these impact captions. Next, the participants were asked to discuss their views on applying impact captions to social VR scenarios and brainstorm any desired ways to apply impact caption's features to design communication tools in VR. The results of this session were eventually summarized by the instructor.

4.1.3 Sub-Session 2: Prototyping and Presentation. After brainstorming, the main objective was to continue the co-design process, concretizing ideas in the creation of low-fidelity prototypes. At the beginning, the instructor led a review of the results of the previous session and then informed the participants that they had 30 minutes to create their own prototype using white paper by drawing or handcrafting. Participants were encouraged to discuss their ideas with others and to think aloud during the session. However, participants still needed to create their own prototypes individually. When the time was up, each participant presented his/her prototype to the group, explaining the design concept, and discussing the prototypes.

We recorded the entire process of the co-design session with permissions from the contributors, transcribed it into text, and collected the prototypes they designed. Using the method proposed by Chen and Zhang [17], three of our authors conducted open coding and categorization of the design prototypes of the participants. In particular, components involved in the design of each prototype (such as input format, output results), information flow, context, and other visual elements (such as present style, text content, and colors) were coded by referring to existing work regarding Visual Captions [60] and Augmented Presentation [57]. Finally, we revealed qualitative findings regarding how impact captions can support interpersonal communication and derived design goals for implementing a prototype system.

4.2 Findings

From the co-design activities, we identified three common patterns of using impact captions to facilitate interpersonal communication.

4.2.1 Enhancing Emotional Expressing. Emotions are abstract and inconspicuous. Impact captions could be an effective way to visualize and concretize human emotions as visible and touchable

⁴<https://hello.vrchat.com/>

shapes in VR. This allows information senders in communication to share their emotional feelings in perceptible forms, so that the receivers could also easily understand the sender's moods and intentions in the VR space.

The textual elements in the design space can encode various perspectives of emotional expressions. Text color is often used to represent positive or negative emotions in avatar dialogues. Generally, cool colors are used to convey more negative emotions, whereas warm colors represent more positive and enthusiastic emotions. For example, the word "indifference" could be rendered in deep blue, while "happy" is often depicted in bright orange or yellow colors. In addition to color, the typeface of text is also associated with emotions. More regular fonts are typically linked to serious and earnest emotions, whereas more artistic and irregular fonts convey a sense of openness and freedom. As for non-textual visual elements, emojis and emoticons are already commonly used elements for conveying people's emotions in social media and instant messaging. Previous research has also found that different shapes of bubbles can carry different emotional information to express calm or intense expressions.

4.2.2 Highlighting Key Information in Speech. In conversations, impact captions can be used to highlight key information and, with their positioning or bubble identifiers, attribute them to the speakers. Typeface, color, and text size are useful dimensions for highlighting key information in a lengthy speech. Texts in distinctive colors or larger sizes are more likely to be read first by the recipient. Additionally, simpler and more regular fonts are more readable compared to overly complex artistic typefaces. In practical scenarios, multiple visual dimensions can be used individually or simultaneously to emphasize key information while aligning with the artistic style of the context.

4.2.3 Identifying Speakers. Although 3D audio and spatial sound effects allow users in social VR to trace the source of the sound and identify who is speaking, not everyone can accurately discern direction through simulated sound. Moreover, when multiple people speak simultaneously, the situation becomes more challenging. Impact captions, as visible virtual objects, are generated near the speaking avatars with a bubble tail pointing to the corresponding character in social VR. This allows participants to easily identify the current speaker through visual perception. Furthermore, when impact captions generated during conversations are allowed to persist in the environment for a prolonged period, newcomers can also see previous conversation content and gauge the level of activity of other participants based on the number of impact captions.

4.3 Design Goals

Based on the results of the co-design, we propose the following design goals to implement a proof-of-concept tool to support interpersonal communication in social VR through impact captions.

4.3.1 G1: Generating impact caption in time. It was agreed by the participants that the impact captions should be generated when users were speaking, as voice-based "face-to-face" conversation is the most common way to interpersonal communication in social VR. Ideally, impact captions should appear simultaneously with the speech.

4.3.2 G2: Maintaining readability and clarity. Once impact captions were generated, a key aspect in making them useful for communication is to maintain readability and clarity so that other users in the VR space can clearly perceive the information carried out by impact captions. This requires each impact caption to appear in a proper location with a correct facing direction. In addition, multiple impact captions should not overlap each other.

4.3.3 G3: Enabling interpersonal interactions with playful impact captions. Interactive impact captions could enhance or introduce new forms of interaction between users, making these captions

a type of new medium that facilitates connections among multiple users in social VR. Taking advantage of the interactivity capabilities provided by VR techniques, users should be able to intuitively handle and play with the impact captions generated by themselves.

5 SpeechCap: System Design and Implementation

We implemented SpeechCap as a proof-of-concept system to demonstrate and evaluate our impact-caption-inspired approach to support interpersonal communication in social VR. According to the design goals (subsection 4.3), SpeechCap takes the user's real-time speech to drive the generation of impact captions in VR (G1), uses rule-based methods to filter out non-important words to be rendered as impact captions in VR (G2), and provides multiple impact-caption-mediated interactions for communication (G2, G3).

5.1 Technical Pipeline: Algorithms, Software, and Hardware Implementations

The core technical pipeline of SpeechCap (Figure 4) consists of three main software modules: (1) a *Voice Interface* that transcribes speech to text in real time and stores the transcribed text for subsequent processing; (2) a *Text Processor* that decides which words should be made as impact captions and the appearance of each impact caption to be rendered in VR based on the analysis of user speech. (3) a *VR Application* that provides a multi-user VR environment for users to realistically enjoy communication using interactive impact captions.

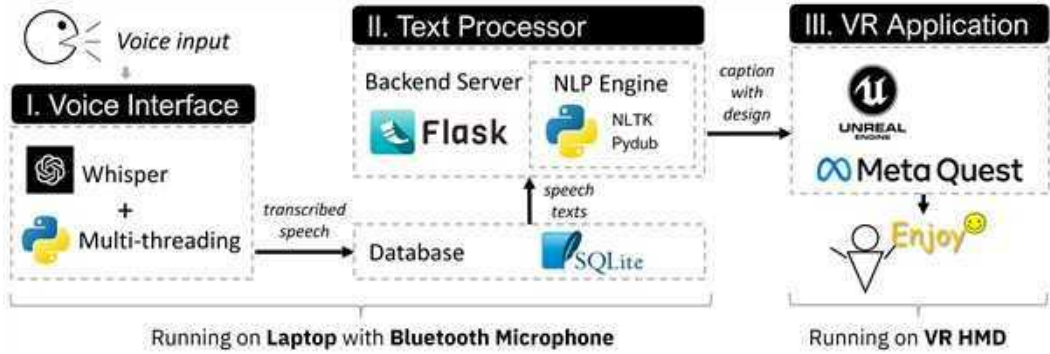


Fig. 4. Technical Pipeline of SpeechCap. The system consists of Voice Interface, Text Processor, and VR Application. The Voice Interface processes users' real-time voice input and stores the transcribed texts. The VR Application keeps polling the Text Processor for fetching upcoming impact captions, driving the Text Processor to convert transcribed texts into impact captions with customized visual design.

5.1.1 Voice Interface. We implemented a Python module with basic real-time speech-to-text capability using OpenAI's Whisper ⁵ [80] as the core model for speech recognition. As a state-of-the-art AI model designed for speech recognition tasks, Whisper's accuracy is close to that of professional human transcribers [80], and its speed is above 10 words per second when using the minimized pruned model (i.e., "tiny.en") for transcribing English on an ordinary hardware platform [36].

Since Whisper was not originally designed for real-time transcribing, we employed a self-developed multi-thread algorithm to approximately process voice input in a real-time manner. When starting, the algorithm recurrently chunks the input voice stream into fragments with a fixed duration (e.g. 1 second by default). Once a fragment is created, the algorithm would allocate

⁵<https://github.com/openai/whisper>

a thread to call Whisper for transcribing, and then save the text result with a sequence number. When running the algorithm, the sequence number starts from 0 and increases by 1 each time for a newly processed fragment. The sequence number mechanism ensures that the downstream Text Processor and VR Application can always keep alignment with the original speech when they process texts and render impact captions in VR.

In terms of accuracy, our method could be slightly lower than the core model because we rely on the multi-threading strategy to transcribe the chunks of speech input one-by-one. Thus, a single word could possibly be separated into two successive chunks and then fail to be recognized, even if the chance is small. For speeding up, Whisper provides a sort of pruned model with smaller sizes to reduce computational consumptions. And in our algorithm, the maximum number of simultaneous working threads is configurable. More available threads can accelerate transcribing while requiring more computational power. Ideally, the delay between a word being recorded and transcribed can be calculated by summing up the chunk duration (e.g. 1 second), Whisper processing time, and the time of waiting for an available thread. Particularly, in our user study, we adopted the “tiny.en” model of Whisper (with the minimum number of parameters and specified for English) and a maximum of 4 threads to maintain an acceptable transcribing speed for running the Voice Interface on Apple MacBook Pro with the M1 Pro chipset.

Our real-time speech recognition algorithm is scalable in terms of both accuracy and efficiency. By employing Whisper models with different sizes or other SOTA speech recognition models, both dimensions could possibly be improved, although the trade-off between accuracy and efficiency always holds. Except for the core model, by configuring different chunk durations and the number of available threads (i.e. thread pool size), the efficiency is optimizable. In addition, advanced hardware resources other than laptops can bring significant improvements to the performance of the Voice Interface module.

5.1.2 Text Processor. To decide which words should be made as impact captions and the specific visual design of each individual caption, we developed the Text Processor module to analyze transcribed speech using NLTK ⁶ and Pydub ⁷. Text Processor and upper-stream Voice Interface are connected by a shared database (SQLite) in which transcribed texts are stored by Voice Interface. And to respond to the needs of the downstream VR Application, we encapsulate the Text Processor in a back-end web server using the Flask framework ⁸. The server is responsible for providing processed data (i.e., the content and visual design of impact captions) for the VR Application through HTTP APIs.

For keyword extraction in Text Processor, we followed a strategy using the part-of-speech (POS) tag to filter out function words (e.g., *conjunction*, *particle*, and *adposition*) as these categories of words mainly convey grammar functionalities with less semantics, and to keep words from the tags *noun*, *verb*, *adjective*, and *adverb*. In addition, we also keep the words of *interjections* as they indicate emotional expressions in conversations, and emotional expression is one of the primary usages of impact captions based on our previous findings.

For the visual design of impact captions in VR, Text Processor applies a rule-based approach to generate a customized visual appearance regarding the design space (section 3) for each individual impact caption. The rule-based approach utilizes semantic and acoustic features extracted from the user’s speech to determine the appearances on each dimension of the design space for an impact caption instance. Details about the rules are discussed in the next section subsection 5.2.

⁶<https://www.nltk.org/>

⁷<https://pydub.com/>

⁸<https://flask.palletsprojects.com/>

5.1.3 VR Application. We implemented a VR application to render interactive impact captions and provide multi-user virtual reality environments to simulate social VR conversation scenarios. Technically, the VR application is built with Unreal Engine 5.1⁹ and Meta XR Tools¹⁰, as we adopted Meta Quest 2 as the target VR hardware device. Besides the overall settings, we utilized the built-in 3D Text Actor of Unreal Engine [39] to implement the impact captions in VR.

To achieve a multi-user VR environment for at least two players (i.e., running VR simulation in the same virtual session on multiple HMDs), we developed our VR application based on the “Listen Server” mode offered by Unreal Engine [40]. In this mode, the first player who starts the application will take the responsibility as a “server” to create and host a session over a local area network (LAN), while the “server” itself is also a “client”. Once a session has started, other players in the same LAN can join the session and play as a pure “client”. Technically, whether or not to play the role of “server”, each individual VR HMD in the multi-player session needs to simulate the entire virtual reality environment independently. To make the independently simulated virtual worlds have the same appearances for all the players, the “clients” have to share the same parameters for rendering objects and simulating interactions through a replication mechanism. The replication mechanism continuously synchronizes the parameters from the “server” to the “clients” and reports changes made by the “clients” to the “server”. In this way, multiple players can eventually experience the same VR environment using their own devices.

Nevertheless, the replication mechanism does not force every “client” to strictly have the same appearance all the time, since each individual “client” always reserves full control of the objects in its own simulation. This allows us to make local modifications to the objects before they are finally rendered on each individual “client”. In this way, we adjust the rotation parameters of the impact captions on each “client” HMD, making the impact captions always face towards the HMD’s owner player in a readable angle (i.e., the replicants of the “same” impact caption are rendered to face different directions in different “clients”).

5.1.4 Hardware Settings. Hardware settings include a laptop, a wireless microphone, and a VR head-mounted device with controllers for a single user. Specifically, we employed an Apple MacBook Pro (with an M1 Pro chipset) laptop for running the Voice Interface and Text Processor modules with the associated database and web server, and a Bluetooth wireless microphone connected to the laptop for collecting the user’s voice input. Meta Quest 2 with controllers is used to run the SpeechCap VR Application.

To create a stable network environment for serving the multi-player session among the devices, we configured a router to provide a local wireless network with fixed IP addresses for each quest device and the laptops. Additionally, during a session, we use a remote meeting software (e.g., ZOOM) on the laptops to transmit the voices of the users so that they can hear each other in the VR environment.

5.2 Speech-driven Impact Caption Generation

In SpeechCap, we adopted a rule-based method to determine the specific visual appearance for impact captions on each dimension of the design space. Our method takes verbal content, affective factors (e.g., valence) and prosodic factors (e.g., loudness) extracted from the speech to compute the specific visual design of each impact caption (Figure 5), making impact captions appropriate for the corresponding conversations.

⁹<https://www.unrealengine.com/en-US/unreal-engine-5>

¹⁰<https://developers.meta.com/horizon/develop/>

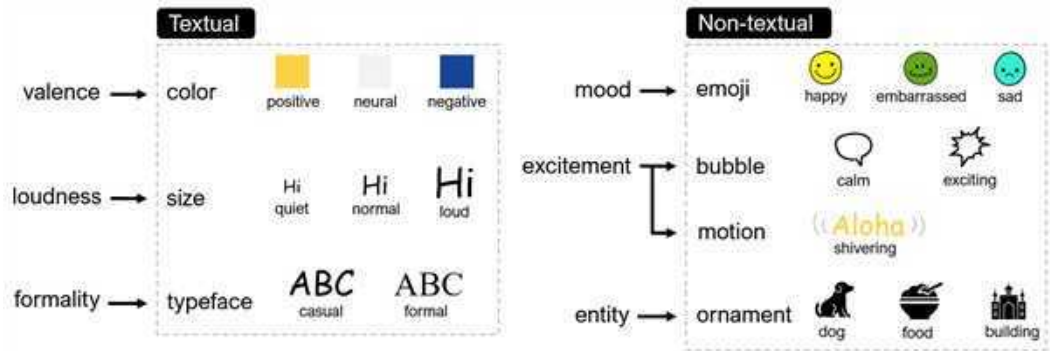


Fig. 5. Mappings Semantics to Impact Caption Design for Proof of the Concept. Valence links to text color where a warm color for positive moods and a cold color for negative moods. Loudness links to the size of captions. The larger the louder. Formality links to typeface. “Time New Roman” is used for formal and “Comic Sans” is used for casual words. Emoji is applied for words regarding happy, embarrassed, and sad feelings. Speech Bubble and a “shivering” motion is applied for words with excitement. Ornaments is applied for words representing specific entities.

5.2.1 *Text Color.* Based on our design space (subsubsection 3.2.3) and the empirical knowledge of color-emotion associations [35, 79, 100], we selected a bright and warm color (■, close to “light orange”) for representing captions with positive emotional expressions, a dark and cold color (■, close to “dark blue”) for negative captions, and white for neutral captions. Technically, the valence of an impact caption is determined by NLTK’s “opinion_lexicon”, which refers to lists of positive and negative words for sentiment analysis.

5.2.2 *Caption Size.* The size of an impact caption is determined by the loudness (volume) of its belonging speech recording fragment. We estimate the loudness by decibels relative to full scale (dBFS) because dBFS measures volume on the same scale so that comparable results can be generated from different audio recording fragments. Using dBFS, the value 0.0 represents the maximum loudness, and negative numbers are used for lower loudness. For example, -20 in dBFS means that the volume is 20 dBs less than the maximum. In SpeechCap, we implemented a rule that allows impact captions to be rendered in three sizes: “small”, “medium”, and “large”. Small size is applied when the dBFS value is under -40. Medium size is applied when the dBFS value is between -40 and -20. Large size applies when the dBFS value is above -20. The louder the speech is, the larger the impact captions will be.

5.2.3 *Typeface.* Referring to the design space and previous studies on human perceptions of typeface [5, 13], we use typeface to encode and represent the formality of speech. As pointed out by previous research, the formality of texts is not a dimension that can be directly calculated but inferred by the words [37]. Thus, we decided to classify words in speech as either “formal” or “casual” in terms of their formality and apply a typeface for the two groups, respectively.

To identify the formality of words, we prepared a pre-defined word list of formal words so that we can determine an incoming word by checking if it is in the list. As for the typefaces, we referred to the literature to guide our design choices. For formal words, we applied “Times New Roman”¹¹, which is a serif typeface originally designed for serious publications. It is also known to make

¹¹https://en.wikipedia.org/wiki/Times_New_Roman

people feel formal and serious [64]. For casual words, we adopted “Comic Sans”¹², as it is also studied by previous research as a casual, sans-serif typeface with a playful and informal look [5]. Its visual characteristics feature rounded edges and irregular strokes.

5.2.4 Emoji and Emoticon. As a non-textual element that directly links to emotions, we selected three common moods to be explicitly displayed by emojis in our impact captions: a “smiling face”, a “sad face”, and an “embarrassed face”, as faces with common expressions belong to a category of emojis that are known to be commonly used across different cultural backgrounds [25]. Similarly to the rules for deciding typefaces, we employed three special word lists for the three moods, respectively. Once a word in the lists is mentioned, the corresponding emoji will be attached aside from the textual part of an impact caption. As the most straightforward example, the word “happy” in impact captions would be accompanied by the “smiling face” emoji, while “crying” will have a “sad face” emoji.

5.2.5 Ornament. According to the design space, typographic ornaments include various types of non-alphabetic characters or icons that are relevant to the words within the impact caption. In SpeechCap, we focus on one of the main purposes of using the ornament, which is to visually highlight objects or entities (e.g., “cake”) that were mentioned in speech. Particularly, we specified a list of keywords of common concepts in daily life for ornaments. The graphic icons for these words are retrieved from Noun Project¹³, a free open-source icon library. The assets of the icons were pre-installed in the VR Application module of SpeechCap, and the back-end Text Processor will make decisions for each word based on the word list.

5.2.6 Speech Bubble. Speech bubble in SpeechCap is used for two types of words: greetings (e.g., “hello”) and interjections (e.g., “HHHHHH” for laughing), since they indicate attention requirements or speech with excitement. For impact captions with greeting words, the balloon is rounded-edged to make the captions different from others and attractive. For interjections, a spiky-edged balloon is used to simulate surprising sounds or speech.

5.2.7 Motion Effects. A “shivering” motion effect will be added to the words with excitement, such as “shocking” and “surprised”, when spawned. Rather than adding more motions to the impact captions while they are generated, we leave the potential of triggering motion effects to the users, since VR naturally provides such interactivity over videos where impact captions originate from. In SpeechCap, the impact captions can always be alive and interactive.

5.3 Communication with Interactive Impact Captions in VR

SpeechCap allows users to have conversations in a shared VR space with the support of interactive impact captions. In SpeechCap, impact captions can only be triggered by speech input. When a particular word is mentioned in the speech, SpeechCap will create a corresponding impact caption at a position roughly in front of the speaker’s avatar at the height of the chest so that users can easily grab the captions. To avoid overlapping, the position is slightly randomized each time for a newly incoming caption (subsubsection 5.3.4).

The visual appearance of each caption instance is determined by the semantics and affective factors of the corresponding speech fragment (subsection 5.2). During a conversation, both the speaker and the listener can see and play with the captions, making their communication process visible, interactive, and playful beyond the traditional voice-only experience.

¹²https://en.wikipedia.org/wiki/Comic_Sans

¹³<https://thenounproject.com/>

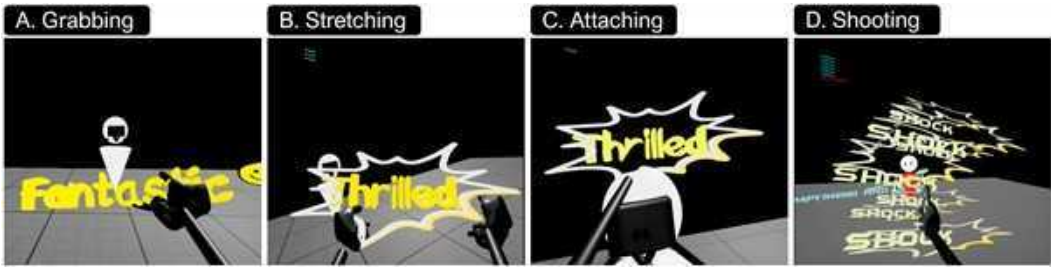


Fig. 6. Interactions with Impact Caption in SpeechCap. (A) **Grab** to hold and place a caption to a new position; (B) **Stretch** with two hands to resize an caption; (C) **Attach** an caption to (another player's) head; (D) **Shoot** an caption to trigger the **Explosion** effect.

5.3.1 Fundamental Interactions. SpeechCap enables users to play with impact captions in VR intuitively using a set of embodied actions (G3), which achieves fundamental interactions between user and the captions. **Grab** (Figure 6 A.) is an action that allows users to hold a caption with their hands using controllers, change its orientation, and place it elsewhere just like taking a physical item in the real world. When grabbing an impact caption, users can further **Shake** the caption to make it start shivering and blinking. When both hands grab the same impact caption simultaneously, the caption can be **Stretched** (Figure 6 B.) for resizing. By resizing, users may create huge captions larger than their avatars, or tiny captions like a bullet. The scalability of size extends the usage of impact captions beyond speech-driven subtitles. Users may combine captions of different sizes with other actions to achieve interactions with others. For example, making a huge caption and shaking it to trigger the blinking effects to enhance the visual impact to emphasize the key points of conversation.

Furthermore, when holding an impact caption, users can **Throw** it away by releasing the grab button while moving their arms, just like the natural “throw” action. As an enhanced version of “throwing”, users can **Shoot** a caption, letting it emit towards a direction in a straight line. The shooting action is bonded with the front trigger button on the controllers. Users can hold the button for up to 3 seconds to charge the action. Longer charging time rewards a stronger initial force for emission, resulting in a high initial velocity of the “bullet” caption.

5.3.2 Decorate Avatars for Self Expression. Avatar plays an important role in social VR that supports self-presentation and reflects users’ emotions and personalities [31, 89]. In SpeechCap, users can **Attach** (Figure 6 C.) an impact caption to the top of their virtual head by tapping or lightly touching the head while holding the caption. Once attached, the caption could be fixed continuously like a “tag” or “hat”. This feature is inspired by the typical usage of impact captions in TV shows where impact captions are used to tag the characters, revealing their emotional status. Besides the head, SpeechCap also allows users to attach an impact caption to the body part of avatars, in which the attached caption will start to periodically revolve around the avatar, achieving the surrounding motion effect in the design space.

5.3.3 Mediating Interpersonal Interactions. In SpeechCap, impact captions can react to the actions and avatars of multiple users so that they can be used as a medium to facilitate the interactions between the users. When thrown or shot by users, impact captions start flying in the virtual space. Once the flying impact caption crashes with another impact caption or avatar, an explosion effect will be triggered (Figure 6 D.), in which several duplications of the flying caption will be generated

at the crashing point and emit outwards fast, like the explosion of fireworks. Such a dynamic effect may be used as a non-verbal cue to attract attention and add more fun to the social VR experience.

5.3.4 Reducing Visual Clutter and Overlapping. To maintain the readability and clarity of floating impact captions in conversations (G2), SpeechCap introduces a “time-to-live (TTL)” mechanism to softly control the number of captions. The “TTL” mechanism allows each caption to live for 5 seconds once generated. If users do not interact with a newly generated impact caption within the time period, the caption will automatically disappear afterwards. Otherwise, if the user wants to keep the impact caption for any purpose, the simplest action to do so is to touch the caption or to grab it at least once before it disappears. Once the caption has interactions with a user, it will no longer be constrained by the living time limitation until users intentionally delete it by holding it and clicking a button on the VR controller (i.e., “X” button for the left-hand controller and “A” button for the right-hand controller).

Additionally, to avoid multiple impact captions overlapping with each other when a user keeps speaking, SpeechCap randomizes the positions for spawning. Specifically, the exact spawn position of a caption is calculated by a fixed position plus random offsets on x , y , and z dimensions. As a result, this strategy makes multiple impact captions look like a “word cloud” emerging in front of the speaker’s avatar.

6 Application Scenarios

Besides the scenarios and potential creative use cases committed in the user study, SpeechCap further supports a variety of communicative needs under different contexts. We briefly introduce the following additional application scenarios to illustrate the generalizability of SpeechCap.

6.1 Making Social VR Accessible to Deaf and Hard-of-Hearing (DHH) People

Recent research revealed that current VR applications fail to provide sufficient accessibility support for deaf and hard-of-hearing (DHH) people to neither experience immersive digital content nor conduct remote communication and socialization [14, 41]. Captioning systems, known as useful tools for making content accessible in other digital media such as videos, are still under-explored in VR contexts [26, 27, 45]. SpeechCap not only have similar functionalities as previous systems that turn speech into visible captions for DHH people [27, 45, 56], but also extend the idea by incorporating multiple visual cues and adding interactivity to convey rich non-verbal information (i.e., valence, excitement, etc.) in virtual reality.

For example, when participating in a conversation with someone who is talking about a vacation travel plan in social VR with SpeechCap, a DHH user can perceive the speaker’s positive mood through the captions in a bright color with a laughing emoji. An additional shivering motion can notify the user that the speaker is also really excited about the trip. By dragging and placing impact captions of the names of the must-visit spots mentioned in speech, the speaker can demonstrate the plan with clarity. Further, when the speaker picks up a caption containing “volcano” and talks about her passion for hiking and exploration, the impact caption with a corresponding volcano icon can give clear reference to the ongoing speech. Overall, informative captions, engaging visuals, and interactive playfulness provided by SpeechCap can enlarge the space for DHH individuals to communicate with other people and the virtual world in social VR.

6.2 Enhancing Interactions for Teaching and Learning in Social VR

VR has shown a great potential in education for the increased social presence it provides in remote learning, and the access to a wide range of learning contexts that may not be accessible in reality [42, 78, 91]. However, it is still far from being a commonly used tool due to the challenges of

creating educational content that is easy for reviewing, providing inclusive functionalities, and supporting collaborative learning [43]. In a virtual lecture, SpeechCap allows instructors to create instructional cues (e.g., notes) by placing impact captions of keywords in the space. The shared interactive captions can further act as a medium for the interactions between the instructor and the students, facilitating learning activities like question answering or hands-on demonstration.

For example, when teaching the structures of plants in a botany class, the instructor would usually introduce the roots, stems, leaves, and flowers in sequence. With SpeechCap, once these words were mentioned, the relevant impact captions would be generated with texts accompanied by symbolic icons. Then, the instructor can grab and stretch to enlarge the caption for emphasizing, and further captures the students' attention by shaking the caption to trigger the blinking effect. If a student has questions about a specific concept mentioned before, he could also make an impact caption of words for that concept and shoot the words, creating an explosion of the words so that the instructor can easily know which part the student is confused about. In addition, further use cases can be explored to adapt to different contexts in teaching and learning. SpeechCap in the current stage provides a starting point to a way to achieve engaging and playful interpersonal interactions for social VR users in the future.

6.3 Facilitating Engaging Live Streaming Experiences in Social VR

Recent research suggests that VR streamers face challenges in building emotional connections since their facial expressions can be obscured by VR headsets, making it difficult for viewers to perceive the emotions of the streamer directly [101], while emotional connection is a key factor in live streaming [62]. SpeechCap can provide visualized emotions for streamers by detecting the expressed moods within the speech and adding appropriate emoticons, motions to relevant captions. SpeechCap can also enhance the communication between the streamer and the viewers by rendering the viewers' reactions as interactive captions in the streamer's virtual space, mediating the interactions between the streamer and the viewers. This could be a way to satisfy the demands of visible and spatialized objects in engaging live streaming experiences [62, 101].

For example, in a typical VR live streaming scenario, a streamer was self-evaluating his playing performance on Beat Saber, a popular VR rhythm game. He didn't do well in the game but would like to tell the audience that he had a lot of fun playing it and felt like a champion. This strong emotion can be communicated and augmented by SpeechCap with colored captions of words like "Fun" and "Exciting" accompanied by laughing emojis. He then made a "Champion" caption with a trophy icon and attached it to his avatar's head like a funny crown to show self-mockery. By shooting the caption "Exciting" outwards, the explosion motion effect can be triggered once the caption crashes with an object in the VR scene, making the atmosphere more interesting and engaging. From the viewers' perspective, they can also send their feedback and comments as impact captions, using the visible and interactive captions to make humorous responses to the streamer's poor performance.

7 User Study

To understand the effectiveness of the design space and SpeechCap system in supporting interpersonal communication in social VR, we conducted a user study, inviting 14 participants of various backgrounds from local universities to experience simulated conversations in VR with interactive impact captions using SpeechCap. User feedback was evaluated from both quantitative and qualitative perspectives through a post-study survey and semi-structured interviews.

7.1 Participants

We recruited 14 participants (6 female and 8 male, aged between 18 and 35) from local universities by advertising on social media. There is no overlap between this group of participants and the four experts who contributed to the previous co-design activity. In terms of social VR experiences, five participants (annotated P1 to P5) were novice users who had never used VR HMDs before, while the other nine participants (P6 to P14) self-reported that they have rich experiences in VR games, social VR features, and VR application developments.

7.2 Procedure

The user study was conducted individually for each participant. The study session for a participant was run in a one-on-one manner in which an experimenter (either the first or second author of this paper) serves the participant to go through the entire procedure. The study sessions lasted 75 to 90 minutes and were held in a lab space in our university. After the study, the participant received a gift card valued \$50 in local currency. For analysis, the study processes were fully audio-taped and the activities in VR were screen-recorded. Regarding hardware settings, two sets of Meta Quest 2 devices and two laptop computers (paired with Bluetooth microphones) were used for the experimenter and the participant in a session. An additional router was configured to host a local area network to serve the multi-user session of the SpeechCap's VR application (subsection 5.1.4).

The core design of this user study is to evaluate SpeechCap through two typical conversation scenarios in social VR: (1) a private conversation between two close friends and (2) a public discussion on a debatable topic with multiple participants. First, the dyad conversation scenario simulates a private conversation in which rich emotional expressions occur. In this scenario, the experimenter and the participant play as close friends to talk about their daily life and personal feelings with the help of impact captions. In the multi-person conversation scenario, the participant will join an ongoing discussion to listen to others and then provide his/her own opinion. In the two scenarios, participants will have opportunities to play the role of both speaker (i.e., information sender) and listener (i.e., information receiver) in turns, allowing them to play with various impact caption instances in very different visual and interactive appearances. Specifically, a full user study session consists of four sections as follows:

7.2.1 Introduction and System Walk-through (~20 min). At the beginning, the experimenter briefly introduced the background and main purpose of SpeechCap (i.e., facilitating interpersonal communication in social VR). Next, the experimenter guided the participant to get familiar with basic usages of the hardware devices if they were not, and then walked through the SpeechCap system. During the walk-through, the experimenter provided verbal and hand-in-hand guidance, ensuring that the participant could see and generate impact captions with their own speech, and know how to interact with the impact captions in VR. After the walk-through, the experimenter would inform the participants about the next dyad and multi-person scenarios.

7.2.2 Dyad Conversation Scenario (~15-20 min). Once the participant and the hardware settings were ready, the experimenter would take another set of devices, go to a separate room (avoiding conflict with the speech-to-text program running on the laptop), and initiate the conversation in VR by telling a pre-defined sad story about losing her cat recently. During the storytelling, the experimenter employed several impact captions to enhance the speech and seek emotional support from the participant. The participant was then asked to comfort the experimenter by telling some happy moments from recent daily life, using impact captions generated along with the conversation's development.

7.2.3 *Public Discussion Scenario (~15-20 min)*. In this scenario, the participant was asked to join a discussion of the topic “*what privacy and security issues will arise around AI?*” The discussion was started between the experimenter and a “NPC” character who was configured to speak based on a pre-defined script and timeline, due to the limitation of hardware resources. Once joining the session, the participant was asked to first listen to the ongoing conversation for a while and then present her/his own opinions with the help of impact captions.

7.2.4 *Survey and Semi-structured Interview (~25-30 min)*. After the two conversation scenarios, the participant was asked to complete a survey about their experience with SpeechCap. The participant was also encouraged to think aloud while completing the questionnaire. When the survey was done, the experimenter would conduct a semi-structured interview with the participant, diving into their experiences and thoughts in depth.

7.3 Results and Findings of Post-study Survey

The post-study survey questionnaires consist of two sections. The first section employs four questions that ask whether the impact caption design space is understandable, interesting, and meaningful. The second section contains six questions about the user experience with SpeechCap, considering enjoyment, emotional expression, clarity of presentation, and overall experience of having conversations with impact captions in VR. The questions are designed in a 7-point Likert scale format [44], requiring the participants to respond with their level of agreement on the statements.

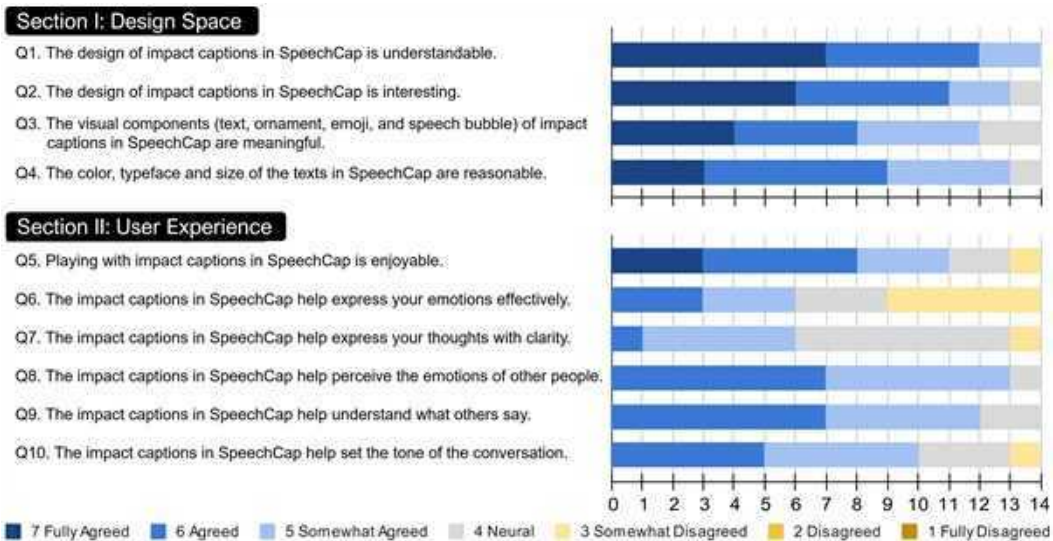


Fig. 7. Post-study Survey with Results. The survey consists of two sections in which the first section contains four questions (Q1-Q4) about the design space of impact captions and the second section contains six questions (Q6-Q10) about the experience with SpeechCap. Overall, participants believed the design space of impact captions was meaningful and the experiences of using SpeechCap is enjoyable and helpful.

In the first section (Figure 7 upper), the participants agreed that the overall design of impact captions to be used for communication in VR was understandable (Q1, M=6.36, SD=0.74) and interesting (Q2, M=6.14, SD=0.95). For the visual appearances and design of individual impact captions, 12 participants agreed that the integration of non-textual elements (ornaments, emojis,

speech bubbles) with texts was meaningful (Q3, $M=5.71$, $SD=1.07$). And 13 participants believed that the visual design (i.e., color, typeface, and size) of textual elements was reasonable (Q4, $M=5.79$, $SD=0.89$). No participants responded with disagreement on the design of impact captions.

In the second section (Figure 7 lower), most of the participants (11/14) agreed that their experiences with SpeechCap were somewhat enjoyable (Q5, $M=5.50$, $SD=1.22$). From an information sender's perspective, participants diverged on the effectiveness of using impact captions for emotional expression in social VR (Q6, $M=4.29$, $SD=1.20$). Six participants agreed that impact captions can help express their emotions effectively, while the other five committed objections. Two participants responded with a neutral opinion. This may relate to the ambiguity caused by the mapping rules for non-verbal visual elements. For the next question about whether impact captions were helpful for presentation clarity (Q7, $M=4.43$, $SD=0.76$), six participants responded with agreement, while five participants stayed neutral and one responded "somewhat disagreed". From a receiver's perspective, the majority of the participants (13/14) agreed that they could perceive the emotions of others by seeing and playing with the impact captions in VR (Q8, $M=5.43$, $SD=0.65$). Most of the participants (12/14) also agreed that impact captions could improve presentation clarity and help them understand other people in VR (Q9, $M=5.36$, $SD=0.74$). In addition, ten participants agreed that impact captions were useful in setting the tone of conversation (Q10, $M=5.00$, $SD=0.96$).

7.4 Results and Findings of Semi-structured Interviews

During the user study, all participants fully experienced the two scenarios and played the roles of both information sender and receiver in conversations. Creative usages of impact captions in VR emerged from their play-through. In the semi-structured interviews, we discussed questions regarding the impact captions' design space and user experiences of SpeechCap with the participants based on their individual performance. Overall, the participants responded positively, indicating that the impact captions were "interesting and enjoyable to play with." (P1, P2, P4, P10, P13). Two authors analyzed the transcribed scripts of the interviews by iterative open coding [23] and finally revealed the following findings showing the meaningfulness of our impact captions' design space, and the usefulness of SpeechCap system.

7.4.1 F1: Textual Elements Can Effectively Support Conversations in Social VR. Participants generally believed that visible texts in impact captions enhance their conversational experience in VR (P2, P3, P4, P6, P7, P11, P12, P13, P14). On one hand, text elements help users accurately express complex or profound meanings. P12 emphasized the expressive capability of impact captions: "The texts within impact captions have a high potential in expressiveness, because texts can convey complex, abstract concepts as well as accurately represent ironic or literary content, which non-textual elements cannot achieve."

On the other hand, some participants (P2, P7, P8, P13, P14) believed impact captions help extract and emphasize key information from dialogues. In SpeechCap, impact captions automatically disappear if not touched by users in several seconds (subsubsection 5.3.4). Therefore, besides filtering trivial words in the back-end processing (subsubsection 5.1.2), SpeechCap also allows users to actively select impact captions with keywords they consider important to remain in VR. "...if I was sharing some knowledge with others with SpeechCap, I could easily make the keywords of my speech persistent and let them floating around my avatar so that my audiences won't lose focus...", said P7. And when talking about the multi-person conversation scenario, P14 said "I'd appreciate SpeechCap if it can always correctly recognize and show the meaningful words by impact captions. This prevents me from listening to lengthy speeches in meetings."

Additionally, one participant (P7), a non-native English speaker, felt that impact captions lower the barrier for her to engage in English conversations. P7 said, "English is not my native language,

and when speaking with native speakers, I often struggle to understand due to different accents and fast speed of speech... but impact captions allow me to clearly and accurately see what others are saying, and they can remain in the VR space as a conversation record for later review."

7.4.2 F2: Non-textual Elements are Engaging but Possibly Ambiguous. Overall, participants agreed that the non-textual visual elements in impact captions are engaging and making the speech-driven captions really "impactful" (P1, P2, P3, P4, P5, P6, P8, P10, P11, P12, P13). P1 said *"I love the idea you combine those cute emojis with the colorful texts. It makes those captions not only interesting but also meaningful."* Among the multiple non-textual dimensions in the design space, the colors, ornaments, emojis, and motion effects are perceived by participants as the most noticeable features (P1, P3, P4, P7, P10, P11, P12).

Nevertheless, the non-textual elements in impact captions may not be able to convey meanings as clearly as texts can do. First, the current colors of impact captions in SpeechCap were reported to be confusing and ambiguous by some participants (P1, P8, P9). P1 asked *"I noticed that those impact captions had different colors, but actually I didn't figure out why a word should be rendered in blue while others were not?"*, and P9 made another similar comment saying that *"I don't quite understand why you choose orange color for the word 'happy' while other words are in white."* Similar issues arise for emojis. P4 commented, *"Expressing with visual elements alone can lead to misunderstandings; some people like to use a smiling face for implicit sarcasm or be ironic'."* These cases may relate to findings of previous research that different people may have different perceptions of the same visual elements when conveying non-verbal meanings [25, 35, 73, 100], revealing the existence of ambiguity in our design of impact captions.

However, on the other hand, three participants (P3, P10, P12) believed that the integration of textual and non-textual elements in impact captions can also mitigate the potential ambiguity. P3 highlighted, *"I might misunderstand the meaning when only seeing the smiling face, but if there was a word 'happy' nearby, it would be much clearer."* In other words, the meanings of different components of a single impact caption can mutually reinforce each other, reducing the chances of misinterpretation by the recipient. When discussing the color-mapping rules in SpeechCap, P10 commented, *"I think using different colors to show the emotions in a conversation is a good idea, and I don't worry about misunderstandings because when the overall tone is positive, I can see most words are in orange. The number of colored captions not only can avoid the ambiguity produced by a single word, but also can reflect how strong the emotional expression is."*

7.4.3 F3: Interactivity is a Key to Engaging Communication Experience in Social VR. Most participants (P1, P2, P4, P5, P6, P7, P8, P10, P11, P13, P14) agreed that the interactivity of impact captions in SpeechCap could bring novel collaborative social experiences and help develop a sense of intimacy between the users in social VR. *"Throwing a caption to people in social VR makes me think of the 'squat and strait' movement before starting a game when playing 'Fall Guys', in which everybody feels fun together."*, said P6. She believed the interpersonal actions mediated by impact captions in SpeechCap were meaningful for building connections with others in social VR, even such interactive capabilities were somehow "unnecessary" for a speech-driven captioning system.

Besides the fundamental interactions, more creative usages of impact captions were explored by the participants. Note-taking is one method commonly proposed. P7 recalled his experience of the multi-person discussion scenario in our study and said, *"During the group discussion, I would like to grab the keywords mentioned by other members, enlarge them, and place them in the virtual space, just like using a whiteboard."* P4 also talked about similar use cases, saying that *"I used to capture the key points mentioned by others and record them with notes in meetings. In SpeechCap, impact captions can be the notes, I can just take the captions aside ... this could also be used in multiplayer reasoning games, right? You can note what others have said (as a basis for reasoning)"*. Note-taking

allows impact captions to persistently present the conversation in the virtual space like dialogue records, achieving a means of asynchronous communication. *“Impact captions can be like email ... other users can read the meeting record whenever they join the session, as long as the impact captions remain there.”* said P12.

Another creative use case was conducted by P10. When conducting the dyad conversation scenario in which there was a sad story about a missing cat, he purposefully generated several impact captions with negative words and used these captions as blocks to construct a small shelter in VR by resizing and placing. P10 explained, *“When I felt sad and uncomfortable, I always find somewhere to hide myself avoiding meeting with other people. So I take these many “sad” words to build a blue room for taking a break.”* He also invited the researcher to join his room made by impact captions to relax in VR.

7.4.4 F4: Needs for Identifying Keywords in Conversations. Although the participants mostly provided us with positive feedback to SpeechCap, they also pointed out some drawbacks. One of the most commonly asked questions by the participants when they started using SpeechCap was about how we decided the words to be rendered as impact captions, as they usually found there were too many impact captions emerging when speaking continuously. After knowing our simple POS-tag-based filtering method to remove functional words, most participants suggested an improvement on it. P2 suggested, *“Don not turn every word into an impact caption. Please extract the key information (of the speech) and only show impact captions of the relevant keywords.”* Other participants, like P11, thought too much text content increased the reading and comprehension costs and then reduced the overall experience of having a conversation with SpeechCap. P11 complained, *“No one wants to read that much text when using VR.”* Particularly, in the multi-person conversation scenario, as more impact captions were spawned and accumulated in limited space, some participants (P1, P2) reported that they felt the field of vision was partially obstructed.

8 Discussion

The user study findings demonstrate the effectiveness of our impact caption design space and the usefulness of our SpeechCap system. Related to findings and the literature, we first discuss the potentials of using impact captions as a form of visual rhetoric with interactivity to enhance communication and interactions in social VR. Next, we present an in-depth understanding of the ambiguity in SpeechCap and proposed future research directions regarding it for designing communication tools in social VR. Then, we provide three design implications for future tools for interpersonal communication in social VR based on the study results. Finally, we explain the scalability of SpeechCap, showing the limitations and corresponding future directions to improve our SpeechCap.

8.1 Designing Visual Rhetoric with Interactivity for Facilitating Communication in Social VR

As suggested by previous research, mediating interpersonal communication in virtual spaces is actually about supporting people in constructing cognitive and affective connections [68, 77] and discovering self-identity beyond the real world [32, 66, 89]. Although most of the known communication methods in social VR are still imperfect [4, 12, 29, 88, 90, 98].

8.1.1 Highlighting the Visual Rhetoric and Interactivity. To fill this gap, our work incorporates speech-driven captioning mechanisms into social VR with improvements on captions' visual design and interactivity. Inspired by the original impact captions used in entertaining TV shows [82], we design and implement a form of visual rhetoric with an integration of textual and non-textual

elements to convey verbal and non-verbal information simultaneously and provide several human-caption interactions to support speech conversations in social VR. The user study results confirmed that our impact-caption-mediated method for communication can effectively enhance clarity (subsubsection 7.4.1) and reinforce non-verbal cues (subsubsection 7.4.2) in social VR conversations. Moreover, we also found that the rich interactivity offered by impact captions can promote interpersonal interactions between users (subsubsection 7.4.3). Particularly, SpeechCap presents a concrete example of how expressive visual objects with interactivity can mediate communication and further support social interactions in social VR.

8.1.2 Considering Generalizability. As for the generalizability of our impact-caption-mediated method, the aforementioned application scenarios (section 6) demonstrate that SpeechCap can be widely applied to different cases to support user needs in social VR. It can enhance emotional expressions to create engaging and intimate atmospheres, support live presentations for teaching scientific knowledge in educational scenarios, and further provide a promising captioning mechanism that can bring immersive social VR experiences to deaf and hard-of-hearing (DHH) individuals.

8.1.3 Research Directions for Future Work. In summary, our work emphasizes the value of visual rhetoric integrating with interactivity for facilitating communication in social VR, showing how invisible verbal and non-verbal information can be made perceptible and playable. SpeechCap expands on the “superpowers” of social VR [68, 69], leading the way towards unique immersive communication experiences and fostering novel interactions not only between two human users, but also between humans and the digital media.

As the design space of impact captions in SpeechCap is generalizable, future work can borrow from or build on our impact captions to create novel digital mediums to support communication needs in a wide range of social VR applications, including facilitating distinctive activities [18, 65], enhancing self-presentations [32, 66, 89], and constructing social connections [54]. Besides, further studies are required to investigate the long-term influence of using SpeechCap for communication on users and explore whether such a method can interfere with the “proteus effect” in social VR [66] to reveal opportunities for expanding and enriching the application scenarios.

8.2 Understanding the Ambiguity in SpeechCap

In computer-mediated communication (CMC) systems, ambiguity is inevitable due to multiple socio-technical factors, such as cultural differences, personal preferences, and technical limitations [8, 87]. In SpeechCap, impact captions also face the risks of miscommunication caused by the ambiguity inherent in the complex design space. Through the user study, a concern raised by participants is that the current heuristic rule-based method for deciding the visual appearances may fail to accommodate the diverse needs of users with different preferences, since they may have different perceptions when seeing the same impact captions (subsubsection 7.4.2). Previous research pointed out that visual elements in the design space, such as color-emotion association [35, 100] and emoji [25, 73], could imply ambiguous interpretations by different people, which also supports our study findings.

However, the inevitable ambiguity is a double-edged sword in digital media design. While bringing risks of miscommunication, it can also facilitate creativity [87], support negotiation [34], and even enhance the human-likeness of AI agents for human-AI communication [59]. Therefore, a trade-off arises: for CMC tools, designers must balance the clarity of communication while maintaining certain ambiguity to meet other design goals [8, 87]. This trade-off leads to two directions of dealing with the ambiguity: mitigating and utilizing.

Yet, the way to mitigate ambiguity for communication in social VR has not been fully explored. Though previous work revealed the potential of using biosignal visualizations [51]. As a complement, our impact caption design space and SpeechCap illustrate another way to reduce the risks of miscommunication caused by the ambiguity in social VR conversations. Particularly, our impact captions can reduce ambiguity by providing contextual information, which is a known effective method for communication in contexts other than social VR [24, 28, 73].

For a single impact caption in SpeechCap, the integration of textual and non-textual visual elements allows the elements to complement each other and provide mutual confirmation in terms of their meanings. This can reduce ambiguity conveyed by a single element (i.e., a single emoji) (subsubsection 7.4.2). At the speech level, the speech-driven approach produces impact captions sequentially aligning with the voice input, in which the conversation contexts could be naturally visualized and presented. Moreover, users can persistently present an impact caption by simply dragging and placing it at an arbitrary position in the virtual space (subsubsection 5.3.4) to intentionally create environmental cues [24] for providing contextual information with flexibility. Overall, these features embedded in SpeechCap pave the way towards designing communication tools in social VR with the consideration of mitigating ambiguity. Future work can refer to the self-explanatory visual design of impact captions to construct visual cues that can convey meanings with clarity and can also leave space for users to actively create environmental cues in the virtual space to support their communication activities.

As for utilizing the ambiguity for particular design aims, which is not covered in the current stage of SpeechCap and the design space of impact captions, we believe it would be a promising research direction for future work that aims to support interpersonal communication and social interactions in social VR. With the unique affordances [31, 32, 69, 98] offered by social VR and growing application scenarios with communication needs [18, 65, 71], ambiguity should have great potential to foster valuable design and mediums to satisfy specific user needs [34, 87].

8.3 Design Implications

Based on SpeechCap and the findings, we identify the following potential design implications for designing digital mediums and tools to support interpersonal communication in social VR. With the fast evolution of VR and related computing technologies, the contexts of conducting interpersonal communication in VR may change rapidly. However, we hope these implications can highlight the invariant knowledge learned from SpeechCap to inspire future work.

8.3.1 Presenting Verbal Information Selectively. Verbal information is fundamental in interpersonal communication [60]. However, when enhancing verbal information for social VR conversations, we should consider selective presentations rather than making all the words in speech visible. In our user study, participants suggested avoiding presenting every word in speech, since too many insignificant words could be distracting and cause unexpected visual clutter (subsubsection 7.4.4). Although we have already designed a mechanism to softly control the number of impact captions, trying to minimize the possibility of visual clutter (subsubsection 5.3.4), the POS-tag-based filtering mechanism in the text processor of SpeechCap is too simple to be an effective keyword extraction method. With the fact evolution of AI and NLP technologies, future work can explore the ways to involve advanced NLP technologies such as Large Language Models (LLMs) [67] to achieve keyword extraction with accurate understandings of the speech content and the user's intentions.

8.3.2 Using Visual Rhetoric to Present Non-verbal Information. Non-verbal information plays an important role in communication, social interactions [2, 58, 66, 69], and self-presentations [32, 89, 105] in social VR. Our work demonstrates that rich visual elements can effectively convey non-verbal information in social VR, showcasing the potential of visual rhetoric design for communication

tools. However, we currently only use simple mapping rules to decide how captions correspond to specific visual elements in SpeechCap. These rules are inadequate for sufficiently and accurately expressing the various types of non-verbal information, as pointed out by participants in the user study. Therefore, future research can start from seeking in-depth understandings of the associations between specific visual elements and non-verbal information (such as color-emotion association [35, 100]), and create novel visual rhetoric designs with effectiveness and elegance for supporting various applications of social VR.

8.3.3 Enhancing the Interactivity of Communication Medium. The impact captions in SpeechCap embed rich interactivity, enabling users to create multiple new forms of interpersonal interactive actions (subsection 5.3). This again demonstrates the unique interactive and immersive nature of social VR [66], highlighting the value of the interactivity for supporting creative, engaging, and emotional communication experiences. On top of SpeechCap and previous work, future research can explore more on the interactivity afforded by the digital mediums in social VR, investigating their influences and creating novel designs and technologies to enhance the existing communication channels or even introduce new channels to support interpersonal communication and interactions in social VR.

8.4 Technical Scalability and Future Improvements

SpeechCap is theoretically scalable in terms of architecture and algorithms (subsection 5.1). However, the current implementation of SpeechCap as a proof-of-concept system has limited scalability when considering real-world deployment. We identified three main aspects of the limitations and discuss possible solutions for future improvements.

8.4.1 Seeking for Better Speech-to-text Solution. First, our custom speech transcribing algorithm introduces limitations. The algorithm integrates the Whisper model [80] with multi-thread processing for approximate real-time processing, which is inferior to the cloud-based end-to-end speech-to-text services provided by business companies in terms of performance (i.e., speed and accuracy). Running on laptops further restricts its performance due to the limited computational resources. For future improvement, we can replace the voice interface module with a well-developed real-time speech transcribing solution, such as a cloud-based business solution or a locally-deployable SDK.

8.4.2 Reducing Network Load by Simplifying Architecture. Second, the architecture of SpeechCap heavily relies on network communication. The back-end web server works as a centric hub that generates parameters of impact captions and exposes APIs to feed VR applications. The VR application, as the user end, keeps polling the back-end server to continuously fetch newly generated impact captions through the network. In the meantime, multiple VR application instances also continuously sync with each other to achieve the multi-user session across the network. As pointed out by recent research, the growing number of users will significantly increase the network load for social VR applications [20]. This challenge also applies to SpeechCap. For improvement, through the perspective of architecture, the back-end server could be reduced, because ideally, all the computations, including speech transcribing and caption generation, can be processed in place in a single computational node (i.e., the VR HMDs). Though this relies on the growth of computational power provided by advanced chips in the future. Without the back-end server, only the synchronization mechanism among VR devices requires network resources.

8.4.3 Simplifying the Hardware Requirements. Third, the hardware setting is too complex. Currently, we employ a laptop, a microphone, and a VR HMD to serve a single user. This setting requires coordination among the devices to smoothly run SpeechCap. Even though the setting is manageable under lab settings for user study, it is far from being used by ordinary people out of the lab.

This limitation can also be addressed with the simplification of architecture, as the laptop and external microphone are used for running the back end for speech input and processing. Once those computational tasks are embedded into VR HMDs, no extra hardware will be needed for SpeechCap.

9 Conclusion

This research demonstrates the potential of impact captions as a novel tool to facilitate interpersonal communication in social VR. By integrating both verbal and non-verbal cues through dynamic and playful typographic elements, impact captions enrich the interpersonal interactions in real-time conversations in virtual spaces. Our proof-of-concept system, SpeechCap, showcases the feasibility and benefits of this approach, while also identifying areas for improvement. Future work could focus on refining the design of impact captions to minimize ambiguity and exploring customization features to better meet diverse user needs and enhance inclusiveness. Overall, our impact-caption-inspired approach offers new avenues for engaging and meaningful interactions among social VR users, paving the way for more immersive and effective communication experiences.

References

- [1] Ahsan Abdullah, Jan Kolkmeier, Vivian Lo, and Michael Neff. 2021. Videoconference and Embodied VR: Communication Patterns Across Task and Medium. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 453 (Oct. 2021), 29 pages. doi:10.1145/3479597
- [2] Nadine Aburumman, Marco Gillies, Jamie A. Ward, and Antonia F.de C. Hamilton. 2022. Nonverbal communication in virtual reality: Nodding as a social signal in virtual interactions. *International Journal of Human-Computer Studies* 164 (2022), 102819. doi:10.1016/j.ijhcs.2022.102819
- [3] Imtiaz Ahmed, Ville Harjunen, Giulio Jacucci, Eve Hoggan, Niklas Ravaja, and Michiel M. Spapé. 2016. Reach out and touch me: effects of four distinct haptic technologies on affective touch in virtual reality. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan) (ICMI '16). Association for Computing Machinery, New York, NY, USA, 341–348. doi:10.1145/2993148.2993171
- [4] Daniel Akselrad, Cyan DeVeaux, Eugy Han, Mark Roman Miller, and Jeremy N. Bailenson. 2023. Body Crumple, Sound Intrusion, and Embodiment Violation: Toward a Framework for Miscommunication in VR. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) (CSCW '23 Companion). Association for Computing Machinery, New York, NY, USA, 122–125. doi:10.1145/3584931.3606968
- [5] Nicole Amare and Alan Manning. 2012. Seeing typeface personality: Emotional responses to form as tone. In *2012 IEEE International Professional Communication Conference*. IEEE, 1–9. doi:10.1109/IPCC.2012.6408605
- [6] Pengcheng An, Jiawen Stefanie Zhu, Zibo Zhang, Yifei Yin, Qingyuan Ma, Che Yan, Linghao Du, and Jian Zhao. 2024. EmoWear: Exploring Emotional Teasers for Voice Message Interaction on Smartwatches. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 279, 16 pages. doi:10.1145/3613904.3642101
- [7] Christoph Anthes, Rubén Jesús García-Hernández, Markus Wiedemann, and Dieter Kranzlmüller. 2016. State of the art of virtual reality technology. In *2016 IEEE Aerospace Conference*. IEEE, 1–19. doi:10.1109/AERO.2016.7500674
- [8] Paul M. Aoki and Allison Woodruff. 2005. Making space for stories: ambiguity in the design of personal communication systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) (CHI '05). Association for Computing Machinery, New York, NY, USA, 181–190. doi:10.1145/1054972.1054998
- [9] Toshiki Aoki, Rintaro Chujo, Katsufumi Matsui, Saemi Choi, and Ari Hautasaari. 2022. EmoBalloon - Conveying Emotional Arousal in Text Chats with Speech Balloons. 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 527, 16 pages. doi:10.1145/3491102.3501920
- [10] Sahar Aseeri and Victoria Interrante. 2021. The influence of avatar representation on interpersonal communication in virtual social environments. *IEEE transactions on visualization and computer graphics* 27, 5 (2021), 2608–2617.
- [11] Steven Baker, Ryan M. Kelly, Jenny Waycott, Romina Carrasco, Thuong Hoang, Frances Batchelor, Elizabeth Ozanne, Briony Dow, Jeni Warburton, and Frank Vetere. 2019. Interrogating Social Virtual Reality as a Communication Medium for Older Adults. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 149 (Nov. 2019), 24 pages. doi:10.1145/3359251
- [12] Steven Baker, Jenny Waycott, Romina Carrasco, Ryan M. Kelly, Anthony John Jones, Jack Lilley, Briony Dow, Frances Batchelor, Thuong Hoang, and Frank Vetere. 2021. Avatar-Mediated Communication in Social VR: An In-depth Exploration of Older Adult Interaction in an Emerging Communication Platform. In *Proceedings of the 2021*

- CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 491, 13 pages. doi:10.1145/3411764.3445752
- [13] Renata Germano Bianchi, Kamila Rios da Hora Rodrigues, and Vânia Paula de Almeida Neris. 2021. Emotional Responses to Font Types and Sizes in Web Pages. In *Proceedings of the XX Brazilian Symposium on Human Factors in Computing Systems* (Virtual Event, Brazil) (IHC '21). Association for Computing Machinery, New York, NY, USA, Article 17, 11 pages. doi:10.1145/3472301.3484325
 - [14] Azadeh Borna, Seyede Zohre Mousavi, Farnaz Fathollahzadeh, Ahmadreza Nazari, and Rayan E Harari. 2024. Applications of Augmented and Virtual Reality in Enhancing Communication for Individuals Who Are Hard of Hearing: A Systematic Review. *American Journal of Audiology* 33, 4 (2024), 1378–1394. doi:10.1044/2024_AJA-24-00056
 - [15] Danielle Bragg, Shiri Azenkot, Kevin Larson, Ann Bessemans, and Adam Tauman Kalai. 2017. Designing and Evaluating Livefonts. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) (UIST '17). Association for Computing Machinery, New York, NY, USA, 481–492. doi:10.1145/3126594.3126660
 - [16] Yining Cao, Rubaiat Habib Kazi, Li-Yi Wei, Deepali Aneja, and Haijun Xia. 2024. Elastica: Adaptive Live Augmented Presentations with Elastic Mappings Across Modalities. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 599, 19 pages. doi:10.1145/3613904.3642725
 - [17] Kevin Chen and Haoqi Zhang. 2015. Remote Paper Prototype Testing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) (CHI '15). Association for Computing Machinery, New York, NY, USA, 77–80. doi:10.1145/2702123.2702423
 - [18] Qijia Chen, Andrea Bellucci, and Giulio Jacucci. 2024. “I’d rather drink in VRChat”: Understanding Drinking in Social Virtual Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 797, 16 pages. doi:10.1145/3613904.3642405
 - [19] Qinyue Chen, Yuchun Yan, and Hyeon-Jeong Suk. 2021. Bubble Coloring to Visualize the Speech Emotion. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI EA '21). Association for Computing Machinery, New York, NY, USA, Article 361, 6 pages. doi:10.1145/3411763.3451698
 - [20] Ruizhi Cheng, Nan Wu, Matteo Varvello, Songqing Chen, and Bo Han. 2022. Are we ready for metaverse? a measurement study of social virtual reality platforms. In *Proceedings of the 22nd ACM Internet Measurement Conference* (Nice, France) (IMC '22). Association for Computing Machinery, New York, NY, USA, 504–518. doi:10.1145/3517745.3561417
 - [21] Saemi Choi and Kiyoharu Aizawa. 2019. Emotype: Expressing emotions by changing typeface in mobile messenger texting. *Multimedia Tools and Applications* 78 (2019), 14155–14172.
 - [22] Yean Fun Chow. 2023. Impact caption translation on a streaming media platform: the case of a Chinese reality show. *Perspectives* (2023), 1–20.
 - [23] Juliet Corbin and Anselm Strauss. 2014. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.
 - [24] Paolo Cottone, Luca Pieti, Valentina Schiavinato, Dorian Soru, Massimiliano Martinelli, Diego Varotto, and Giuseppe Mantovani. 2009. “Solving” ambiguity in the virtual space: communication strategies in a collaborative virtual environment. *Cognition, Technology & Work* 11 (2009), 151–163.
 - [25] Justyna Czeżochowska, Kristina Gligorić, Maxime Peyrard, Yann Mentha, Michał Bień, Andrea Grütter, Anita Auer, Aris Xanthos, and Robert West. 2022. On the Context-Free Ambiguity of Emoji. *Proceedings of the International AAAI Conference on Web and Social Media* 16, 1 (May 2022), 1388–1392. doi:10.1609/icwsm.v16i1.19393
 - [26] 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 2024 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. doi:10.1145/3613904.3642258
 - [27] 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. doi:10.1145/3544548.3581511
 - [28] Anind K. Dey and Jennifer Mankoff. 2005. Designing mediation for context-aware applications. *ACM Trans. Comput.-Hum. Interact.* 12, 1 (March 2005), 53–80. doi:10.1145/1057237.1057241
 - [29] Elena Dzardanova, Vlasios Kasapakis, Damianos Gavalas, and Stella Sylaiou. 2022. Virtual reality as a communication medium: a comparative study of forced compliance in virtual reality versus physical world. *Virtual Reality* 26, 2 (2022), 737–757.

- [30] Leonor Fermoselle, Simon Gunkel, Frank ter Haar, Sylvie Dijkstra-Soudarissanane, Alexander Toet, Omar Niamut, and Nanda van der Stap. 2020. Let's Get in Touch! Adding Haptics to Social VR. In *Proceedings of the 2020 ACM International Conference on Interactive Media Experiences* (Cornella, Barcelona, Spain) (IMX '20). Association for Computing Machinery, New York, NY, USA, 174–179. doi:10.1145/3391614.3399396
- [31] Guo Freeman and Dane Acena. 2021. Hugging from A Distance: Building Interpersonal Relationships in Social Virtual Reality. In *Proceedings of the 2021 ACM International Conference on Interactive Media Experiences* (Virtual Event, USA) (IMX '21). Association for Computing Machinery, New York, NY, USA, 84–95. doi:10.1145/3452918.3458805
- [32] Guo Freeman and Divine Maloney. 2021. Body, Avatar, and Me: The Presentation and Perception of Self in Social Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 239 (Jan. 2021), 27 pages. doi:10.1145/3432938
- [33] Kexue Fu, Yixin Chen, Jiaxun Cao, Xin Tong, and RAY LC. 2023. "I Am a Mirror Dweller": Probing the Unique Strategies Users Take to Communicate in the Context of Mirrors in Social Virtual Reality. 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 385, 19 pages. doi:10.1145/3544548.3581464
- [34] William W. Gaver, Jacob Beaver, and Steve Benford. 2003. Ambiguity as a resource for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) (CHI '03). Association for Computing Machinery, New York, NY, USA, 233–240. doi:10.1145/642611.642653
- [35] Mitsuhiro Hanada. 2018. Correspondence Analysis of Color-emotion Associations. *Color Research & Application* 43, 2 (2018), 224–237. doi:10.1002/col.22171
- [36] Amma Liesvarastranta Haz, Evianita Dewi Fajrianti, Nobuo Funabiki, and Sritrasta Sukaridhoto. 2023. A Study of Audio-to-Text Conversion Software Using Whispers Model. In *2023 Sixth International Conference on Vocational Education and Electrical Engineering (ICVEE)*. 268–273. doi:10.1109/ICVEE59738.2023.10348186
- [37] Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Interneter Bericht, Center "Leo Apostel", Vrije Universiteit Brussel* 4, 1 (1999).
- [38] Ching-Yu Hsieh, Yi-Shyuan Chiang, Hung-Yu Chiu, and Yung-Ju Chang. 2020. Bridging the Virtual and Real Worlds: A Preliminary Study of Messaging Notifications in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376228
- [39] Epic Games Inc. 2024. *3D Text Actor in Unreal Engine*. Retrieved May 05, 2024 from https://dev.epicgames.com/documentation/en-us/unreal-engine/3d-text-actor-in-unreal-engine?application_version=5.1
- [40] Epic Games Inc. 2024. *Networking Overview*. Retrieved May 05, 2024 from https://dev.epicgames.com/documentation/en-us/unreal-engine/networking-overview-for-unreal-engine?application_version=5.1
- [41] 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. doi:10.1145/3462244.3479946
- [42] Lasse Jensen and Flemming Konradsen. 2018. A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies* 23 (2018), 1515–1529.
- [43] Qiao Jin, Yu Liu, Svetlana Yarosh, Bo Han, and Feng Qian. 2022. How Will VR Enter University Classrooms? Multi-stakeholders Investigation of VR in Higher Education. 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 563, 17 pages. doi:10.1145/3491102.3517542
- [44] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology* 7, 4 (2015), 396–403.
- [45] JooYeong Kim, SooYeon Ahn, and Jin-Hyuk Hong. 2023. Visible Nuances: A Caption System to Visualize Paralinguistic Speech Cues for Deaf and Hard-of-Hearing Individuals. 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 54, 15 pages. doi:10.1145/3544548.3581130
- [46] David Kirsh. 2013. Embodied cognition and the magical future of interaction design. *ACM Trans. Comput.-Hum. Interact.* 20, 1, Article 3 (April 2013), 30 pages. doi:10.1145/2442106.2442109
- [47] Anya Kolesnichenko, Joshua McVeigh-Schultz, and Katherine Isbister. 2019. Understanding Emerging Design Practices for Avatar Systems in the Commercial Social VR Ecology. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) (DIS '19). Association for Computing Machinery, New York, NY, USA, 241–252. doi:10.1145/3322276.3322352
- [48] Christos Kyriltsias and Despina Michael-Grigoriou. 2022. Social interaction with agents and avatars in immersive virtual environments: A survey. *Frontiers in Virtual Reality* 2 (2022), 168.
- [49] Marc Erich Latoschik, Daniel Roth, Dominik Gall, Jascha Achenbach, Thomas Waltemate, and Mario Botsch. 2017. The effect of avatar realism in immersive social virtual realities. In *Proceedings of the 23rd ACM Symposium on Virtual*

- Reality Software and Technology* (Gothenburg, Sweden) (VRST '17). Association for Computing Machinery, New York, NY, USA, Article 39, 10 pages. doi:10.1145/3139131.3139156
- [50] Kyungjun Lee, Hong Li, Muhammad Rizky Wellyanto, Yu Jiang Tham, Andrés Monroy-Hernández, Fannie Liu, Brian A. Smith, and Rajan Vaish. 2023. Exploring Immersive Interpersonal Communication via AR. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 50 (April 2023), 25 pages. doi:10.1145/3579483
- [51] Sueyoon Lee, Abdallah El Ali, Maarten Wijntjes, and Pablo Cesar. 2022. Understanding and Designing Avatar Biosignal Visualizations for Social Virtual Reality Entertainment. 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 425, 15 pages. doi:10.1145/3491102.3517451
- [52] Joanne Leong, Olivia Seow, Cathy Mengying Fang, Benny J. Tang, Rajan Vaish, and Pattie Maes. 2022. Wemoji: Towards Designing Complementary Communication Systems in Augmented Reality. In *Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) (UIST '22 Adjunct). Association for Computing Machinery, New York, NY, USA, Article 13, 3 pages. doi:10.1145/3526114.3558699
- [53] Jie Li, Guo Chen, Huib de Ridder, and Pablo Cesar. 2020. Designing a Social VR Clinic for Medical Consultations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI EA '20). Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3334480.3382836
- [54] Jie Li, Yiping Kong, Thomas Röggl, Francesca De Simone, Swamy Ananthanarayan, Huib de Ridder, Abdallah El Ali, and Pablo Cesar. 2019. Measuring and Understanding Photo Sharing Experiences in Social Virtual Reality. 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, 1–14. doi:10.1145/3290605.3300897
- [55] Jie Li, Vinoba Vinayagamoorthy, Julie Williamson, David A. Shamma, and Pablo Cesar. 2021. Social VR: A New Medium for Remote Communication and Collaboration. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI EA '21). Association for Computing Machinery, New York, NY, USA, Article 81, 6 pages. doi:10.1145/3411763.3441346
- [56] 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* (Athens, Greece) (ASSETS '22). Association for Computing Machinery, New York, NY, USA, Article 5, 13 pages. doi:10.1145/3517428.3544817
- [57] Jian Liao, Adnan Karim, Shivesh Singh Jadon, Rubaiat Habib Kazi, and Ryo Suzuki. 2022. RealityTalk: Real-Time Speech-Driven Augmented Presentation for AR Live Storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) (UIST '22). Association for Computing Machinery, New York, NY, USA, Article 17, 12 pages. doi:10.1145/3526113.3545702
- [58] Noah Liebman and Darren Gergle. 2016. It's (Not) Simply a Matter of Time: The Relationship Between CMC Cues and Interpersonal Affinity. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) (CSCW '16). Association for Computing Machinery, New York, NY, USA, 570–581. doi:10.1145/2818048.2819945
- [59] Lu Liu, Harm Van Essen, and Berry Eggen. 2024. Let Information Flow Into Awareness: A Design Space for Human-Like Experiences to Promote Informal Communication for Hybrid Work. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) (DIS '24). Association for Computing Machinery, New York, NY, USA, 1666–1680. doi:10.1145/3643834.3661549
- [60] Xingyu "Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, and Ruofei Du. 2023. Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 108, 20 pages. doi:10.1145/3544548.3581566
- [61] Shao-Kang Lo. 2008. The Nonverbal Communication Functions of Emoticons in Computer-Mediated Communication. *CyberPsychology & Behavior* 11, 5 (2008), 595–597. doi:10.1089/cpb.2007.0132
- [62] Zhicong Lu, Haijun Xia, Seongkook Heo, and Daniel Wigdor. 2018. You Watch, You Give, and You Engage: A Study of Live Streaming Practices in China. 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, 1–13. doi:10.1145/3173574.3174040
- [63] Yuhan Luo, Junnan Yu, Minhui Liang, Yichen Wan, Kening Zhu, and Shannon Sie Santosa. 2024. Emotion Embodied: Unveiling the Expressive Potential of Single-Hand Gestures. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 404, 17 pages. doi:10.1145/3613904.3642255
- [64] Jo Mackiewicz and Rachel Moeller. 2004. Why people perceive typefaces to have different personalities. In *International Professional Communication Conference, 2004. IPCC 2004. Proceedings*. IEEE, 304–313.

- [65] Divine Maloney and Guo Freeman. 2020. Falling Asleep Together: What Makes Activities in Social Virtual Reality Meaningful to Users. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Virtual Event, Canada) (*CHI PLAY '20*). Association for Computing Machinery, New York, NY, USA, 510–521. doi:10.1145/3410404.3414266
- [66] Divine Maloney, Guo Freeman, and Donghee Yvette Wohn. 2020. "Talking without a Voice": Understanding Non-verbal Communication in Social Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 175 (Oct. 2020), 25 pages. doi:10.1145/3415246
- [67] Reza Yousefi Maragheh, Chenhao Fang, Charan Chand Irugu, Parth Parikh, Jason Cho, Jianpeng Xu, Saranyan Sukumar, Malay Patel, Evren Korpeoglu, Sushant Kumar, et al. 2023. LLM-take: theme-aware keyword extraction using large language models. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 4318–4324.
- [68] Joshua McVeigh-Schultz and Katherine Isbister. 2021. The Case for "Weird Social" in VR/XR: A Vision of Social Superpowers Beyond Meatspace. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (*CHI EA '21*). Association for Computing Machinery, New York, NY, USA, Article 17, 10 pages. doi:10.1145/3411763.3450377
- [69] Joshua McVeigh-Schultz and Katherine Isbister. 2022. A "beyond being there" for VR meetings: envisioning the future of remote work. *Human-Computer Interaction* 37, 5 (2022), 433–453.
- [70] Joshua McVeigh-Schultz, Anya Kolesnichenko, and Katherine Isbister. 2019. Shaping Pro-Social Interaction in VR: An Emerging Design Framework. 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, 1–12. doi:10.1145/3290605.3300794
- [71] Yanni Mei, Jie Li, Huib de Ridder, and Pablo Cesar. 2021. CakeVR: A Social Virtual Reality (VR) Tool for Co-designing Cakes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (*CHI '21*). Association for Computing Machinery, New York, NY, USA, Article 572, 14 pages. doi:10.1145/3411764.3445503
- [72] Liang Men and Nick Bryan-Kinns. 2022. Supporting sonic interaction in creative, shared virtual environments. In *Sonic Interactions in Virtual Environments*. Springer International Publishing Cham, 237–267.
- [73] Hannah Miller, Daniel Kluver, Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Understanding Emoji Ambiguity in Context: The Role of Text in Emoji-Related Miscommunication. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (May 2017), 152–161. doi:10.1609/icwsm.v11i1.14901
- [74] Jesús Emmanuel Montoya Esquer and Graciela Lara López. 2023. Wordsphere: virtual reality text input interface. *Virtual Reality* 27, 3 (2023), 2769–2785.
- [75] Florian 'Floyd' Mueller, Richard Byrne, Josh Andres, and Rakesh Patibanda. 2018. Experiencing the Body as Play. 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, 1–13. doi:10.1145/3173574.3173784
- [76] Minako O'Hagan. 2010. Japanese TV entertainment: Framing humour with open caption telop. *Translation, humor and the media* 2 (2010), 70–88.
- [77] Mark T Palmer. 1995. Interpersonal communication and virtual reality: Mediating interpersonal relationships. *Communication in the age of virtual reality* (1995), 277–299.
- [78] Zhenhui Peng, Chuhan Shi, and Xiaojuan Ma. 2022. Exploring User Experience and Design Opportunities of Desktop Social Virtual Reality for Group Learning Activities. In *Proceedings of the Ninth International Symposium of Chinese CHI* (Online, Hong Kong) (*Chinese CHI '21*). Association for Computing Machinery, New York, NY, USA, 105–111. doi:10.1145/3490355.3490367
- [79] Robert Plutchik and Henry Kellerman. 2013. *Theories of emotion*. Vol. 1. Academic press.
- [80] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 28492–28518. https://proceedings.mlr.press/v202/radford23a.html
- [81] Rufat Rzayev, Polina Ugnivenko, Sarah Graf, Valentin Schwind, and Niels Henze. 2021. Reading in VR: The Effect of Text Presentation Type and Location. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (*CHI '21*). Association for Computing Machinery, New York, NY, USA, Article 531, 10 pages. doi:10.1145/3411764.3445606
- [82] Ryoko Sasamoto. 2014. Impact caption as a highlighting device: Attempts at viewer manipulation on TV. *Discourse, Context & Media* 6 (2014), 1–10.
- [83] Ryoko Sasamoto, Stephen Doherty, and Minako O'Hagan. 2021. The 'hookability' of multimodal impact captions: A mixed-methods exploratory study of Japanese TV viewers. *Translation, Cognition & Behavior* 4, 2 (2021), 253–280.
- [84] Shuichi Seto, Hiroshi Arai, Kimikazu Sugimori, Yuko Shimomura, and Hiroyuki Kawabe. 2010. Subtitle system visualizing non-verbal expressions in voice for hearing impaired-Ambient Font. *Proceeding of the 10th Asia-Pacific*

- Industrial Engineering and Management Systems* (2010).
- [85] Chenxinran Shen, Joanna Mcgrenere, and Dongwook Yoon. 2024. LegacySphere: Facilitating Intergenerational Communication Through Perspective-Taking and Storytelling in Embodied VR. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 119, 16 pages. doi:10.1145/3613904.3641923
 - [86] Harrison Jesse Smith and Michael Neff. 2018. Communication Behavior in Embodied Virtual Reality. 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, 1–12. doi:10.1145/3173574.3173863
 - [87] Martin Stacey and Claudia Eckert. 2003. Against Ambiguity. *Computer Supported Cooperative Work (CSCW)* 12 (2003), 153–183.
 - [88] Philipp Sykownik, Sukran Karaosmanoglu, Katharina Emmerich, Frank Steinicke, and Maic Masuch. 2023. VR Almost There: Simulating Co-located Multiplayer Experiences in Social Virtual Reality. 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 789, 19 pages. doi:10.1145/3544548.3581230
 - [89] Philipp Sykownik, Divine Maloney, Guo Freeman, and Maic Masuch. 2022. Something Personal from the Metaverse: Goals, Topics, and Contextual Factors of Self-Disclosure in Commercial Social VR. 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 632, 17 pages. doi:10.1145/3491102.3502008
 - [90] Theresa Jean Tanenbaum, Nazely Hartoonian, and Jeffrey Bryan. 2020. "How do I make this thing smile?": An Inventory of Expressive Nonverbal Communication in Commercial Social Virtual Reality Platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376606
 - [91] Santawat Thanyadit, Parinya Punpongsonan, Thammathip Piumsomboon, and Ting-Chuen Pong. 2022. Xr-live: Enhancing asynchronous shared-space demonstrations with spatial-temporal assistive toolsets for effective learning in immersive virtual laboratories. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–23.
 - [92] Sunday D Ubur, Naome A Etori, Shiva Ghasemi, Kenneth King, Denis Gračanin, and Maria Gini. 2024. EasyCaption: Investigating the Impact of Prolonged Exposure to Captioning on VR HMD on General Population. In *International Conference on Human-Computer Interaction*. Springer, Springer, Cham, Switzerland, 382–403.
 - [93] Udapola Balage Hansi Shashiprabha Udapola. 2022. Social VR for Socially Isolated Adolescents with Significant Illnesses. In *Companion Proceedings of the 2022 Conference on Interactive Surfaces and Spaces*. 50–53.
 - [94] Matthijs van der Boon, Leonor Fermoselle, Frank ter Haar, Sylvie Dijkstra-Soudarissanane, and Omar Niamut. 2022. Deep Learning Augmented Realistic Avatars for Social VR Human Representation. In *Proceedings of the 2022 ACM International Conference on Interactive Media Experiences* (Aveiro, JB, Portugal) (IMX '22). Association for Computing Machinery, New York, NY, USA, 311–318. doi:10.1145/3505284.3532976
 - [95] Fangzhou Wang, Hidehisa Nagano, Kunio Kashino, and Takeo Igarashi. 2016. Visualizing video sounds with sound word animation to enrich user experience. *IEEE Transactions on Multimedia* 19, 2 (2016), 418–429.
 - [96] Xian Wang, Xiaoyu Mo, Lik-Hang Lee, Xiaoying Wei, Xiaofu Jin, Mingming Fan, and Pan Hui. 2023. Designing Loving-Kindness Meditation in Virtual Reality for Long-Distance Romantic Relationships. In *Proceedings of the 31st ACM International Conference on Multimedia* (Ottawa ON, Canada) (MM '23). Association for Computing Machinery, New York, NY, USA, 7608–7617. doi:10.1145/3581783.3612438
 - [97] Xiaoying Wei, Yizheng Gu, Emily Kuang, Xian Wang, Beiyan Cao, Xiaofu Jin, and Mingming Fan. 2023. Bridging the Generational Gap: Exploring How Virtual Reality Supports Remote Communication Between Grandparents and Grandchildren. 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 444, 15 pages. doi:10.1145/3544548.3581405
 - [98] Xiaoying Wei, Xiaofu Jin, and Mingming Fan. 2024. Communication in Immersive Social Virtual Reality: A Systematic Review of 10 Years' Studies. In *Proceedings of the Tenth International Symposium of Chinese CHI* (Guangzhou, China and Online, China) (Chinese CHI '22). Association for Computing Machinery, New York, NY, USA, 27–37. doi:10.1145/3565698.3565767
 - [99] Markus Wieland, Lauren Thevin, Albrecht Schmidt, and Tonja Machulla. 2022. Non-verbal Communication and Joint Attention Between People with and Without Visual Impairments: Deriving Guidelines for Inclusive Conversations in Virtual Realities. In *International Conference on Computers Helping People with Special Needs*. Springer, 295–304.
 - [100] Lisa Wilms and Daniel Oberfeld. 2018. Color and emotion: effects of hue, saturation, and brightness. *Psychological research* 82, 5 (2018), 896–914.
 - [101] Liwei Wu, Qing Liu, Jian Zhao, and Edward Lank. 2023. Interactions across Displays and Space: A Study of Virtual Reality Streaming Practices on Twitch. *Proc. ACM Hum.-Comput. Interact.* 7, ISS, Article 437 (Nov. 2023), 24 pages. doi:10.1145/3626473

- [102] Haijun Xia, Tony Wang, Aditya Gunturu, Peiling Jiang, William Duan, and Xiaoshuo Yao. 2023. CrossTalk: Intelligent Substrates for Language-Oriented Interaction in Video-Based Communication and Collaboration. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) (*UIST '23*). Association for Computing Machinery, New York, NY, USA, Article 60, 16 pages. doi:10.1145/3586183.3606773
- [103] Yukang Yan, Haohua Liu, Yingtian Shi, Jingying Wang, Ruici Guo, Zisu Li, Xuhai Xu, Chun Yu, Yuntao Wang, and Yuanchun Shi. 2023. Conespeech: Exploring directional speech interaction for multi-person remote communication in virtual reality. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2647–2657.
- [104] Amal Yassien, Passant ElAgroudy, Elhassan Makled, and Slim Abdennadher. 2020. A Design Space for Social Presence in VR. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (Tallinn, Estonia) (*NordiCHI '20*). Association for Computing Machinery, New York, NY, USA, Article 8, 12 pages. doi:10.1145/3419249.3420112
- [105] Kexin Zhang, Elmira Deldari, Zhicong Lu, Yaxing Yao, and Yuhang Zhao. 2022. “It’s Just Part of Me:” Understanding Avatar Diversity and Self-presentation of People with Disabilities in Social Virtual Reality. 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 4, 16 pages. doi:10.1145/3517428.3544829