Understanding Safety Risks and Safety Design in Social VR Environments

QINGXIAO ZHENG, School of Information Sciences, University of Illinois at Urbana-Champaign, USA
SHENGYANG XU*, College of Fine & Applied Arts, University of Illinois at Urbana-Champaign, USA
LINGQING WANG*, Department of Industrial Engineering, Tsinghua University, China
YILIU TANG, School of Informatics, University of Illinois at Urbana-Champaign, USA
ROHAN C. SALVI, School of Information Sciences, University of Illinois at Urbana-Champaign, USA
GUO FREEMAN, School of Computing, Clemson University, Clemson, USA
YUN HUANG, School of Information Sciences, University of Illinois at Urbana-Champaign, USA

Understanding emerging safety risks in nuanced social VR spaces and how existing safety features are used is crucial for the future development of safe and inclusive 3D social worlds. Prior research on safety risks in social VR is mainly based on interview or survey data about social VR users' experiences and opinions, which lacks "in-situ observations" of how individuals react to these risks. Using two empirical studies, this paper seeks to understand safety risks and safety design in social VR. In Study 1, we investigated 212 YouTube videos and their transcripts that document social VR users' immediate experiences of safety risks as victims, attackers, or bystanders. We also analyzed spectators' reactions to these risks shown in comments to the videos. In Study 2, we summarized 13 safety features across various social VR platforms and mapped how each existing safety feature in social VR can mitigate the risks identified in Study 1. Based on the uniqueness of social VR interaction dynamics and users' multi-modal simulated reactions, we call for further rethinking and reapproaching safety designs for future social VR environments and propose potential design implications for future safety protection mechanisms in social VR.

CCS Concepts: • **Human-centered computing → Collaborative and social computing design and evaluation methods**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: safety risks, safety design, social virtual reality, online harassment

ACM Reference Format:
Qingxiao Zheng, Shengyang Xu, Lingqing Wang, Yiliu Tang, Rohan c. Salvi, Guo Freeman, and Yun Huang. 2023. Understanding Safety Risks and Safety Design in Social VR Environments. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 154 (April 2023), 37 pages. https://doi.org/10.1145/3579630

*The second and third authors contributed equally to this research.

Authors' addresses: Qingxiao Zheng, School of Information Sciences, University of Illinois at Urbana-Champaign, USA, qzheng14@illinois.edu; Shengyang Xu, College of Fine & Applied Arts, University of Illinois at Urbana-Champaign, USA, sx21@illinois.edu; Lingqing Wang, Department of Industrial Engineering, Tsinghua University, China, wanglq19@mails.tsinghua.edu.cn; Yiliu Tang, School of Informatics, University of Illinois at Urbana-Champaign, USA, yiliut2@illinois.edu; Rohan c. Salvi, School of Information Sciences, University of Illinois at Urbana-Champaign, USA, rcsalvi2@illinois.edu; Guo Freeman, School of Computing, Clemson University, Clemson, South Carolina, USA, guof@clemson.edu; Yun Huang, School of Information Sciences, University of Illinois at Urbana-Champaign, USA, yunhuang@illinois.edu.

2573-0142/2023/4-ART154 $15.00
https://doi.org/10.1145/3579630

1 INTRODUCTION

Social Virtual Reality (VR) is a rapidly growing technology that allows multiple users to engage in immersive 360-degree virtual content through VR head-mounted displays [45, 59, 59, 60]. Unlike traditional video games that only allow users to pilot and watch avatars' movements on screen through keyboards and touch-pads, Social VR allows for full-body tracking, so that avatars move in the same way the user does in reality, and for users to make virtual physical connections with other users and grasp virtual objects [29, 30]. However, it is imperative to recognize that the embodiment and immersion characteristic of Social VR has resulted in an increasing number of reported cases of virtual harassment and abuse, particularly among underrepresented groups, including women and individuals who identify as LGBTQ [10, 13, 17, 26, 29, 30, 52].

Despite various commercial social VR platforms claiming to offer new safety tools and tightened safety safeguards to protect their users [72], it has been reported that social transgression and harassment occur every 7 minutes in Social VR [1], with 27% of women and 21% of men experiencing it, and over 40% observing it [70]. As a result, it is crucial to gain a comprehensive understanding of the emerging safety risks and the effectiveness of current safety features in social VR in order to design and create future social VR environments as safe spaces for all users.

Overall, we believe that three remaining research opportunities for further understanding social VR safety risks exist. First, past research has highlighted the importance of visibly labeling harassment as inappropriate in the context of social VR, which can serve to surface community norms and expectations for appropriate behavior [13]. Researchers cautioned that such classification can also marginalize users whose harassment experiences are not typical, or whose experiences are not accounted for in the system's development [13]. Despite the recognition of the necessity to understand different social norms in social VR [13, 29], there is a paucity of current research that has developed a classification of safety risks in social VR, likely due to the fact that this field is still in its nascent phase. Second, safety risks in social VR can be perceived as more severe and disruptive due to the realistic and immersive nature of the experience [30]. This may prompt more spontaneous and immediate actions, which could in turn lead to new safety risks and attacks. Previous research suggested that psychological attacks within immersive environments, such as social VR, may have a greater impact on an individual's mental health than traditional online attacks [10, 11, 57, 69]. Therefore, it is crucial to conduct a more comprehensive examination of specific behaviors or interactions within the context of social VR that may contribute to new safety risks. Lastly, there are a few solutions for social VR's novel safety risks. Due to the fact that many safety solutions are based on traditional online communities, they may not adequately handle social VR's unique challenges. Social VR allows for realistic and engaging social interactions, but it also exposes users to new threats beyond text- or voice-based online harassment [13, 29, 30]. Given this, it is crucial to critically analyze the extent to which current safety prevention techniques can provide secure social experiences in social VR environments.

To contribute towards addressing these limitations, in this paper we explore:

- **RQ1**: *What types of safety risks in social VR have been identified and documented?*
- **RQ2**: *How do people (e.g., as victims, attackers, bystanders, or spectators) react to these risks?*
- **RQ3**: *What risks do main social VR platforms address in their policies or features?*

We use the term "safety risk" in social VR to describe *various types of detrimental user behaviors involving abusive communications directed towards other users* (e.g., harassment, verbal abuse) and *disruptive behaviors that violate the rules and social norms of the platform* (e.g., griefing, spamming, and cheating). Social VR safety risks are often unpredictable, highly personal, and real-time, making documenting and data collecting difficult. To answer our questions, we conducted two studies. Study 1 addresses RQ1 and RQ2. We used YouTube videos as our dataset because users can freely

record and share their first-hand social VR experiences on Youtube. We then analyzed 212 YouTube videos (01/01/2022-06/01/2022) and their comments and transcripts to assess uploaders safety risks in social VR from their own perspectives. Built upon the findings of Study 1, study 2 examines RQ3 by evaluating how current social VR platforms address these safety risks. We reviewed safety feature descriptions on social VR platforms' community websites and directly tested them in social VR. We then mapped if the risks identified in Study 1 could be addressed by the existing safety features.

Our work contributes to the HCI and CSCW communities in the following ways. First, we conducted an in-depth investigation of the various types of safety risks in social VR, as well as how people react to these risks based on first-hand experiences captured in videos. Given the novelty and popularity of social VR, the findings can be used to better understand the new norms in the social virtual worlds. Our research demonstrates the feasibility of using user-generated videos as a new methodological approach to investigate safety risks in real-time, as opposed to previous approaches that frequently rely on interview and survey methods. Second, we took the first step toward evaluating existing safety features in social VR and determining their ability to address current safety risks. Our findings will help to shape the design and development of gesture-based and avatar-focused safety features in the future.

2 RELATED WORK

In this section, we explained the unique characteristics of social VR compared to other online social platforms, then summarized the challenges of managing safety risks in social VR.

2.1 Why is Social VR Unique?

Social VR is a novel digital technology with unique features of embodied avatars, simulated interactions, rich virtual content, and lifelike social environment [45, 59]. The use of embodied avatars allows social VR users to not only customize their avatar appearance but also control such avatars with real-time gestures and motions using full- or partial-body tracking [28, 29, 55]. Further, real-time interaction and rich virtual content offer immersive first-person perspectives [19], a myriad of activities and engaging experiences that may mimic offline activities, and many methods of expressing and understanding social cues through verbal or nonverbal behaviors (e.g., gestures and facial expressions) [55]. These features thus make social VR experientially different from traditional avatar-based virtual worlds settings (e.g., Second Life) that are played largely through a mouse, keyboards, and a flat screen [19, 57].

Moreover, with its lifelike social environment, social VR may help people foster substantial emotional bonds of friendship, intimacy, affection, and romance through simulated social activities than online social networking sites (e.g., Facebook) [25, 27, 55, 82]. Various social VR platforms tend to afford diverse activities and social atmospheres. For example, AltspaceVR is well-known for its unique combination of varied activities, such as interacting with people, attending events, and professional development; RecRoom is devoted to virtual reality gaming; High Fidelity VR is focuses on massive public performances and events; Meta Horizon Worlds is designed for virtual interactions with friends and family [29, 33]. As an online social space, social VR platforms are also free to enter and generally accessible. Some platforms even allow users without a VR headset to enter, bridging the gap between those who want to use a computer and those who prefer a virtual reality headset [55]. However, these unique technical and novel experiential features also introduced new safety risks, which we discussed in the next section.

2.2 Challenges of Addressing Safety Risks in Social VR

Existing research on social VR illustrated the complexity of managing safety risks in social VR, due to its embodiment nature, identity-related threats, presence of minors, and lack of consensus.

The *embodiment* of social VR can lead to a number of potential risks, including physical breakdowns, immersive harassment, and the compromise of user personal information. Physical breakdowns, such as failing, colliding with surroundings, and hitting spectators, can also occur in social VR environments[16]. Additionally, the use of voice function in social VR may lead to forced attention [30]. Most importantly, users usually reveal certain personally identifiable information, such as their voice in order to fully participate in social VR [30, 58, 74]. This can result in sensitive personal details about the individual's offline identity (e.g., in terms of height, ethnicity, looks, and gender) being revealed, making them more susceptible to stalking and other forms of harassment [58].

The compromise of user personal information can also lead to *identity-related threats*, because of the information revealed through user voice. For example, Maloney et al [57] have noted as voice is a primary mode of communication in social VR, this can lead to new forms of marginalization for certain groups, such as female users or non-native English speakers who may need to mute themselves to avoid unwanted attention or mockery of accent. Additionally, Freeman et al [29] have discovered the challenges faced by individuals who do not identify with the gender assigned to them at birth in social VR environments. These individuals often experience misgendering and stereotyping, which can make them more susceptible to harassment and discrimination in public areas of the virtual environment. These findings revealed different forms of safety risks due to the embodied nature of social VR, as well as underscore the need for further studies to further investigations [13, 30].

The *presence of minors* in social VR poses new and complicated safety risks due to the novel interactions that occur within these virtual environments. Research has shown that adults may have mixed sentiments towards interacting with minors in social VR, including irritation and annoyance [30], as well as enjoyment [54], highlighting the complexity of adult-minor co-existence in virtual environments. Additionally, teens have reported experiencing unique challenges and tensions in their daily use of social VR, such as different forms of harassment, disconnection from the offline world, addictive behavior, and fear of losing offline social skills [56]. Furthermore, studies have shown that minors tend to perceive their interactions with other minors in social VR as more realistic compared to other traditional virtual worlds [54], and as equal in value to offline interactions [79], further emphasizing the potential risks and negative impact of social VR on minors.

The *lack of consensus* on what constitutes a safety risk in social VR poses challenges in addressing safety concerns in these environments. Studies have shown that definitions of harassment vary among social VR users, with some viewing it as simply unpleasant experiences and others as more severe and damaging conduct [13, 30]. Furthermore, it is unclear if different social VR spaces may develop their own unique social norms pertaining to safety risks, depending on their specific focuses such as gaming, creativity, public events, and seminars, or professional advancement [30]. Additionally, the effectiveness of existing mechanisms for mitigating risks in social VR remains uncertain, as the instantaneous nature of communication in these environments may make it difficult to properly document and report instances of harassment [30].

Among all, these existing studies have employed a range of methodologies, including interviews [13, 27, 29, 30, 56, 58], surveys [74], observations [2, 16, 54, 57], and simulations [21], to investigate various aspects of social VR. However, there seems to be a significant research gap in understanding the complexity and nuance of *social cues*, or non-verbal behaviors, can be used and

interpreted in social VR and how they relate to safety risks. We found two studies that have used behavioral coding to understand user behaviors in social VR.

On the one side, Maloney et al. [57] have presented examples of observed nonverbal behaviors that primarily focused on the potential benefits of nonverbal behaviors in social VR. These nonverbal behaviors include indicating attention (nodding, head movement, gaze direction, hand gestures), expressing approval (emojis, applause), directing attention (pointing, patting chest), social grooming (waving, dancing, kissing), interpersonal provocation (poking, pushing, moving too close), social disruption (flying, excessive movement), and entertainment (dancing, emojis, playing with objects).

On the opposite, Fiani et al.[21] emphasized the harm the combination of nonverbal behaviors can cause in social VR. They created some 3D prototypes to use as simulations to explore how different combinations of nonverbal behaviors, such as proxemics, facial expressions, gaze, and voice, might influence the different levels of perceptions of bullying in social VR.

While both studies have provided insight into nonverbal behaviors in social VR, more research is needed to fully understand the patterns and impact of these behaviors on safety risks. This paper aims to address this research gap through two studies: one involving a comprehensive analysis of risk types and characteristics, and the other examining the effectiveness of existing safety features on social VR platforms. The aim of these studies is to understand what kind of safety risks are there, and how current features fail or can be effective to mitigate safety risks in social VR.

3 STUDY 1: YOUTUBE VIDEO ANALYSIS (RQ1 & RQ2)

In Study 1, we aimed to understand safety risks (RQ1)and users' reactions (RQ2) in social VR through a content analysis of YouTube videos posted by social VR users themselves. Previous studies using interview [2, 13, 30] and survey methods [70] to explore harassment in social VR may not fully capture the environmental factors and user experiences *in-the-moment*, instead, they mostly provided a record for users' experience *after* these risks. However, as previously noted, documenting and collecting data on safety risks in social VR can be challenging as they are often unexpected, personal, and occur in real time, making it difficult to gather empirical data.

Prior research in CHI and CSCW have used video content analysis of YouTube videos to gain insights into user interactions with novel technologies (e.g., [7, 35, 46]) or to gather empirical data on hard-to-find specific groups or events (e.g., [34, 63]). Therefore, we also chose to use YouTube as our data source for investigating how people interact and experience safety risks in social VR for the following reasons. First, the use of user-generated videos compensates for the lack of data on safety risks in social VR, as the unpredictability of these incidents makes it difficult to gather empirical data or recruit individuals who have experienced them. Second, video content analysis of incident recordings allowed us to observe more contexts of the incidents, the environment, and individual experiences than asking people to recall and reflect on what happened in the past by interview or survey when they could forget details or bring false information. Third, analyzing online user comments on these videos allows us to study people's perceptions and reactions to these situations without them having to be victims of these incidents.

3.1 Data Collection

We utilized YouTube API [1] in-title search feature to locate relevant videos. We queried a combination of broader VR terms (i.e., VR, virtual reality, metaverse, social VR), names of mainstream social VR applications (i.e., VRchat, Rec room, Altspace vr, and Horizon Worlds) [28, 30], and synonyms of safety risk (i.e., danger, hatred, issue, risk, poor conduct, safety, harm, privacy, sick, harassment, toxic), which were developed and refined via multiple rounds of search exploration. We

[1]https://developers.google.com/youtube/v3

automatically crawled the following metadata for each video: video title, video link, video length, video publish date, view counts, video transcripts, comment counts, comment content, like counts, channel title, channel topics, and channel subscriber counts.

To ensure the relevance and quality of the data, we cleaned the initial dataset by removing broken links, non-English videos, videos that were less than 2 seconds in duration, as it is difficult to understand the issue within that short time frame, and videos that had "news" in the title as these were highly related to news reports rather than social VR users' personal and first-hand experiences. Three researchers screened the videos by month and included the videos that met the following criteria: (1) the video is filmed in first-person view in a social VR app; (2) there are multiple users socializing in the video; (3) their interactions are not only about game playing (e.g., excluding videos with the focus of shooting games); and (4) the video demonstrates some behaviors that influence the social VR users' experience. We also included videos that recorded highlight moments and compilations.

3.2 Dataset

We collected a total of 4,748 videos using the YouTube API's in-title search and after cleaning the data to remove noise and irrelevant videos, a sample of 212 videos was used for further analysis. This sample size is comparable to that of prior HCI research using YouTube analysis [16, 35, 63]. The mean length of the sample is 14 minutes and 26 seconds (SD = 23 minutes and 1 second). The mean video *view* count was 98,670 (SD = 195,189), the mean video *like* count was 3,777 (SD = 6,637), and the mean video comment count was 116 (SD = 183). The videos were created by 135 unique Youtubers, with a mean channel subscriber count of 78,701 (SD = 212,460). These Youtubers' channel topics revolved around video game culture, role-playing video games, and action-adventure games, and their channel descriptions focused on VR content, featuring popular VR apps such as VRChat and Rec Room. The transcripts of the videos have an average length of 880 words.

3.3 Codebook and Coding

To answer RQ1, we developed our coding scheme by drawing on prior frameworks in the field: Freeman et al. [30] utilized user interviews to identify several types of emerging harassment experiences in social VR; Blackwell et al. [13] classified harassment on Facebook into physical, verbal, and spatial categories; and Thomas et al. [77] proposed a taxonomy for online hate and harassment identifying seven types of attacks, such as toxic content and surveillance, stemming from different attacker motives and abilities. These results serve as baseline knowledge when we coded safety risks in our data.

To answer RQ2, we coded four perspectives - victims, attackers, bystanders, and spectators - inspired by prior studies in cybercrime [5, 8, 9, 44, 81]. Cues, such as expressions, gender, gestures, and voice variations, play a significant role in interpersonal interactions [18]. They are also crucial in assessing interpersonal interactions in social VR, as users pilot their avatars' physical behaviors in real-time. We drew inspiration from Feine et al [20] who previously differentiated between cues, social signals, social reactions, and social cues in the context of conversational agents. We defined cue-related terminology in the social VR context using their work, as seen in Table 1. We coded social cues that elicit safety risk-related social reactions using previously developed categories of social cues [20, 68].

- **Paralinguistic cues** are verbal cues embedded in voice communication but independent from language or semantic content (e.g., pitch, tone, volume, emoticons, and inflection);
- **Linguistic cues** are verbal cues referring to word choice and sentence structure (e.g., dialect and syntax);

Table 1. Definitions of Social Cues in VR

Term	Definition *(examples)*
Cue	• A cue is any behavior that presents a source of information [73, 80]. *E.g., User A gives a hand-shaking gesture delivered through the avatar captured by VR controllers and motion detectors.*
Social Signal	• A social signal is the way that cues are interpreted, either consciously or unconsciously, in the form of thoughts or feelings [61]. *E.g., The handshaking gesture can be perceived by user B as a signal of showing friendliness.*
Social Reaction	• A social reaction is how a person feels (emotional), thinks (cognitive), or acts (behavioral) in response to the social signal he or she produced in VR [47]. *E.g., User B reacts to user A's hand shaking gesture.*
Social Cue	• A social cue is what makes the user act in a social way to the people who send out a cue [23, 61]. *E.g., User B reacts to user A's hand shaking gesture by holding hands with user A.*

- **Kinectis** relate to body posture and movements, head and hand gestures (e.g., nodding, pointing, waving);
- **Proxemic cues** pertain to how we perceive our surroundings and are influenced by factors such as societal norms, situational factors, individual traits, and familiarity (e.g., personal space, spatialize audio);
- **Eye movements** are nonverbal cues that have evolved to improve gaze perception by displaying a dark iris within a large, white sclera, creating a high contrast to allow people to follow each other's gaze (e.g., looking, staring, and blinking);
- **Facial expressions** help to convey emotions like happiness, sadness, anger, and fear (e.g., eyebrow movement, blushing).

Three authors evaluated a subset of videos and refined the coding scheme across eight rounds to resolve disagreements. For each round, five to ten videos were randomly selected and researchers proposed several doubtful videos to refine the codebook. We analyzed 20 video clips and checked inter-coder reliability among three coders [22] using ReCal3 calculator [24], and the result for both types of risks and social cues indicated a satisfactory level, proving that the coders are able to code the cases objectively. Finally, two authors coded each video in our dataset. See sample code below 2.

Table 2. Sample Code

Risk Type: Scaring others
Context: Kids presented

Perspective	Social Cues	Social Reaction
Attacker	Kinetics; Proxemic	Multiple identical avatars approach one
Victim	Paralinguistic; Linguistic; Kinetics	Stepping back; screaming; scolding back
Bystander	Paralinguistic; Linguistic	Screaming and yelling F words

*ID:V22'1'14

3.4 Comments and Transcripts Mining

To further address RQ1, we analyzed video transcripts to gain a deeper understanding of language patterns; to further address RQ2, we analyzed comments from viewers to gain the perspectives of spectators, who are people who watch the videos.

We employed *transcript analysis* to understand the language used in safety risk situations and how it varied among different types of risks. YouTube automatically censors potentially inappropriate words in the transcripts and replaces them with "[__]". We analyzed the frequency and density of YouTube-censored words in the transcripts, as well as negative opinion words from a dictionary compiled by Hu et al. [36], which was previously used to mine YouTube video comments to understand the context and negative opinion words used in the transcripts [66, 67]. We also created word clouds and mined for different parts of speech to identify unique speaking patterns among each type of safety risk video.

We used *comment analysis* to understand spectators' perceptions of different types of risks in social VR. Comment analysis can be used to see if the risk was either perceived as not severe enough to report or as acceptable from spectators' perspectives [36, 75]. We classified comments into positive, neutral, and negative categories using VADER, a lexicon, and rule-based sentiment analysis tool, and filtered out short, meaningless comments (i.e., less than 3 words) [38]. We also analyzed the top 10 most frequent words in each video to obtain the highlighted subtopics and subjects of the comments. This approach was inspired by the *"banter"* phenomenon reported by Beres et al. [12] where some interactions that may be perceived as toxic or insulting by some can also be seen as playful and positive by others.

4 STUDY 1 FINDINGS

4.1 Types and Characteristics of Safety Risks In Social VR (RQ1)

We demonstrated the identified 5 types of emerging virtual risks (9 videos) that are unique in the social VR setting, and 7 categories of severer safety risks as *virtual violence* (15 videos), *virtual scaring* (44 videos), *virtual abuse* (44 videos), *virtual sexual harassment* (9 videos), *virtual crashing* (7 videos), *virtual voice trolling* (40 videos), and *virtual trash actions* (44 videos), see Table 3. We also presented the language characteristics of some of these safety risks.

4.1.1 Emerging Safety Risks in Social VR. The following safety risks are unique to social VR settings and stem from varying social norms and cultural values.

Role-playing. *Role-playing* is the act of make-believe and pretending in which one assumes another role. When roleplaying, users may take a new personality, values, and goals, or be an alternative version of themselves. Some role-players also imitate famous celebrities. For example, video [R22'3'7] was shot from the uploader's view. As he entered a room, where users were posing as children and daycare workers, one child used a toy to hit the other while bystanders yelled in fear or begged them to stop. In this case, bystanders could not differentiate whether the kids were really having unpleasant experiences or just playing around. Fraud-impersonation is also role-playing but is done with the specific intent to deceive, e.g., purposefully identifying oneself as another individual or group, such as a social VR employee or an existing social VR user [V18'10'1].

Immersive dwellers. Mirror dwellers are E-daters or individuals in social VR who often wear costly virtual outfits and stand in front of mirrors in public rooms for hours. In the video [R22'2'3], a mirror couple (mirror dwellers as couples who wear matching outfits) stood in front of a public mirror to attract attention. Sometimes when other users suggested that they should use mirrors in private spaces, they replied *"mind your own business"* in an unfriendly manner. Similarly, there are some dwellers who slept in social VR or randomly run away from the keyboard for hours, and their avatars were displayed as sleeping [R22'2'1]. These inactive inhabitants sometimes enticed

Table 3. Definitions of Safety Risks in Social VR

Risks	Definitions
Emerging safety risks	Emerging unique risks in social VR, such as role-playing, fraud-impersonation, immersive dweller, misuse safety features, misread cues, minors picking on adults.
Virtual violence	When an attacker or group of attackers engages in physical contact with the avatars of their victims, such as striking or slapping them, violence occurs.
Virtual crashing	Crashers use tactics or bugs to ruin others' experience, e.g., adding particle effects like fire spawn animation, electric bundle animation in avatars to cause damages.
Virtual scaring	Users employ scary-looking avatars to scare other users, and either rush towards them or appear in front of them out of nowhere.
Virtual abuse	Verbal insults and hate speech aim against a person or group based on gender, ethnicity, sexual orientation, sexual identity, disabilities, or others.
Virtual sexual harassment	Sexual harassment includes making sexually provocative gestures in front of others and sexually assaulting others explicitly and implicitly.
Virtual voice-trolling	Voice trollers employ a gender-mismatch voice or contrasting voice like a lovely avatar with a frightening voice to amuse themselves or scare other users.
Virtual trash actions	An umbrella term for activities that are typically intended to spoil the experiences of others but cannot be categorized in the previous categories.

other users to take part in criminal behavior. In the video [R22'2'3], for instance, a user was shown sleeping on a couch in a public lobby. When other users saw this, they began to physically harass the sleeping avatar by stepping on the dweller's face. Soon, more users joined in. Although these immersing dwellers probably may not have meant to harm others, they nevertheless constituted a threat by disturbing the peace in the public space or inciting unpleasant activities.

Misread cues. Miscommunication among users may lead to potential safety issues. For example, in the video [VR22'5'9], a female avatar user got very close to a male avatar user, which was shown by the change in the avatar's visual effect: when an avatar gets too close to a user, it becomes transparent with a white outline in the user's view. The male avatar thought that the female avatar would get close to him so she could kiss him. So, he told her to stop and let her know that he was worried. The female avatar seemed confused and then left. This showed that an enthusiastic way of greeting could be misunderstood as an insult.

Misused safety features. While safety features are to protect users, the malicious use of these features can also generate new problems that disrupt the community. For example, in Rec Room, a vote kick is used to remove a user from a room if the majority of users in that room vote yes[2]. This feature can sometimes be maliciously exploited to abuse other uses. For example, in the video [VR22'4'4], one user was kicked out of the platform because this user did not give presents to other users. The user had the impression that other users' responses (i.e., kicking this user out) were excessive. Also, in the video [VR22'5'4], a group of attackers voted to kick one furry user out (the furries are a subgroup in VRchat who like to wear furry costumes), after they gave hate speech against furry users and showed offensive gestures towards this furry user. These incidents reveal that, while some safety features help protect the social VR community, they also allow for new harmful behaviors if not used responsibly.

Minors picking on adults. In social VR, the tension of adult-minor co-existence is prevalent. In our dataset, 9.4% of the videos contain minors or interaction with minors. Among these videos, we discovered that minors were more likely to be attackers than adults. The harms adults brought to kids included displaying inappropriate content when minors were present [V22'3'8] and abusing

[2]https://recroom.com/comfortandsafety

minors [VR22'3'16]. Surprisingly, many videos also depicted minors assaulting behaviors, such as excessive cursing [VR22'3'4, VR22'3'10], which were quite similar to that of adult users [V22'5'17, VR22'2'1]. These videos usually were titled with keywords such as "toxic kids". This implied that in the adult-minor co-existence environment of social VR, there seems to be an intense conflict between adults and minors, and minors may learn from inappropriate content posed by adult users, and behave accordingly to imitate.

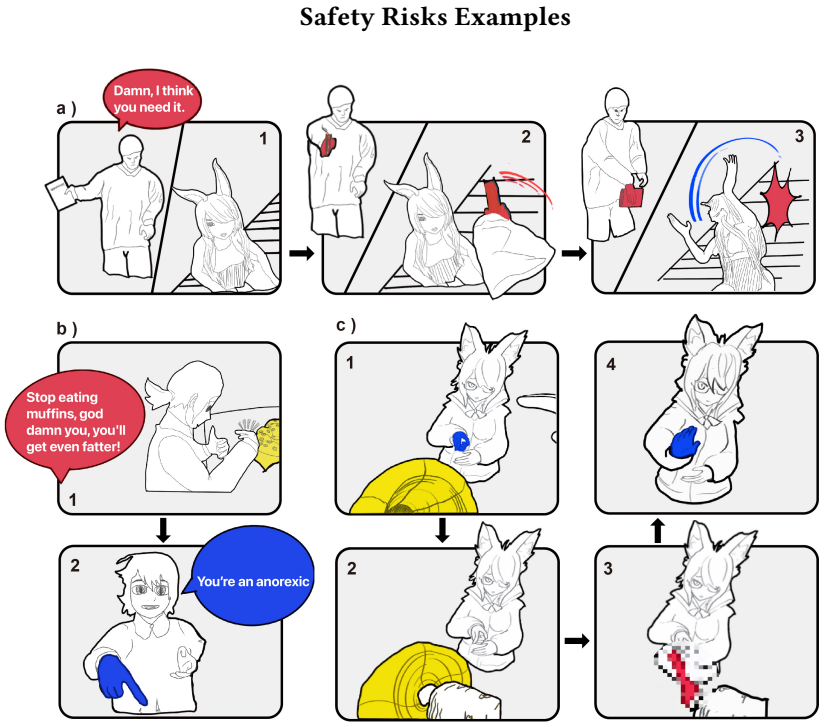

Fig. 1. a) ***Virtual Violence***: An attacker approached a victim while carrying a book and stroked the victim with it. Even though the attack occurred in a virtual environment, the victim responded as if she had been physically smacked in the head and fell down on the ground in reality; b) ***Virtual Abuse***: The attacker used profanity, begged the victim to stop eating, and engaged in body shaming by telling her, *"You'll grow bigger than you are now"* as the victim was snacking on muffins in Social VR. *"You're anorexic!"*, the victim yelled in a counterattack; c) ***Virtual Sexual Harassment***: The attacker approached the victim with his hand holding a virtual object (a hat). The victim was intrigued by the mysterious object. The attacker then grabbed inside his hat and suddenly pulled out an object with obvious sexual indication. The victim was terrified.

4.1.2 Severer Safety Risks in Social VR. The following safety risks are similar to those that occur in other online communities, but the immersive experiences and enhanced avatar control in VR can amplify their severity and cause greater harm.

Virtual violence. Violence in the virtual world may result in bodily injury in the real world. For example, in the video [V22'2'5], as demonstrated in Fig. 1(a), the victim fell into the ground after being virtually slapped by the attacker. This implies that virtual violence in VR might trigger synchronous realistic sensations in users, thus eliciting immediate real-world impact.

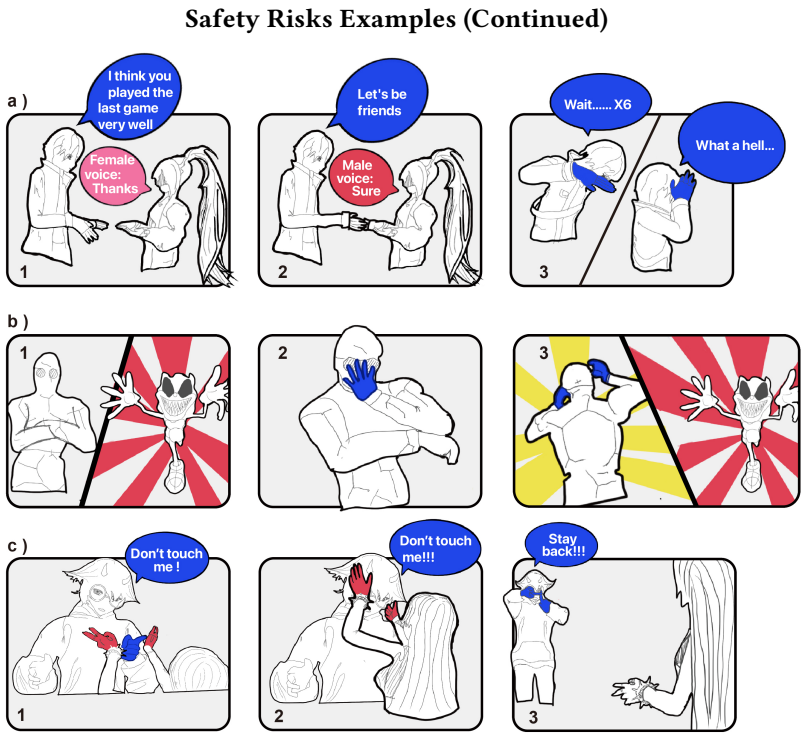

Fig. 2. a) ***Virtual Voice Trolling***: A female user got praised by a male player in a virtual reality game, and the female user responded verbally. The interaction was positive for them, and the male user even extended an invitation to become friends with the female user. They chatted but then the female user suddenly started talking in a masculine voice, startling the male user; b) ***Virtual Scaring***: The attacker made use of a terrifying avatar and sped towards the direction of the victim. The victim fled as a result of hearing that. The victim continued to use his hands to shield his head as the attacker chased after him as if this can keep him "protected"; and c) ***Virtual Trash Actions***: The victim shouted a warning to the attacker who approached him. While the victim tried to warn the attacker away from his face, the attacker continued to touch him. The victim moved back and made a "cross" gesture to warn the attacker not to approach.

Virtual crashing. Crashing often happens due to technological reasons[3] and can be perceived differently. Crashing in social VR is complicated both due to how such activity violates community norms and the need to uphold community privacy. On the one hand, crashers can be perceived as "terrorists." Some of them made crashes in public spaces to show off their skills and demonstrated the whole operation process in the videos, such as [V22'2'16] and [R22'3'4]. Also, some crashers formed underground groups to plot attacks, and other VR users referred to them as "crasher gangs". For example, one video's title and description contained *"The Massacres of VRChat Streamers"* [V22'5'28] and hate speech. On the other hand, crashers uploaded a series of crash compilations and saw themselves as "chivalrously" upholding social order in the social VR community – because they took revenge on social VR streamers who broadcasted in public VR space and therefore breached others' privacy, such as the series of *"Vrchat: We hate Streamers"* [VR22'5'6].

[3]https://hello.vrchat.com/community-guidelines

Virtual abuse. Compared to other online platforms, the simulated avatar and voice mode in social VR may expose more personal traits about a user, leading to personal abuse. Besides common abuse targets similar to traditional social platforms, such as race [R22'2'8] and age [VR22'3'16], attackers in social VR also made negative comments on other users' aesthetic taste in virtual clothing [R22'2'1], bullied a minor based on their high-pitch voice [VR22'3'16], and body shamed other users' behaviors, for example, Fig. 1(b) [V22'5'24]. Moreover, some groups became common targets for virtual abuse. For instance, it is rather easy to spot and insult furry users, who are interested in using anthropomorphic animal furry avatars, in social VR. Such users often suffer from attackers' excessive curses and blame for brutal sexual activities with animals based on the zoophilia orientation stereotype of furry users [V22'3'20, VR22'5'4]. Overall, virtual abuse often arose when social VR users just met. The attacker did not need to check other users' profiles before they could decide on their attacks as in many other social platforms.

Virtual sexual harassment. The *virtual sexual harassment* manifests in multiple ways in social VR, including visual, auditory, and physical sexual harassment. In our data, visual sexual harassment included making sexually harassing gestures [V22'2'2], or displaying sexual content [V22'1'19], e.g., Fig.1(c). Auditory sexual harassment included verbally harassing others, such as saying "hot" and "cute" in a sexual way [R22'2'5]. Physical sexuality included direct physical touch on others' avatars, such as touching the victim's breast [VR22'3'21]. All incidents of virtual sexual harassment occurred in public lobbies in social VR.

Virtual voice trolling. Voice communication is widely used in social VR and can be used to intentionally trigger a wide range of emotional responses. "Female voice trolling" was often used in video titles to describe male users starting out with a female avatar and a feminine voice, then switching to their male voices for trolling. For example, an attacker tricked the victim to fall in love with him by faking a girl's voice and using a cute feminine avatar as shown in Fig.2(a), and then uploaded the video to make fun of the victim [V22'3'28]. Also, deep, hoarse "corpse-like" voice was commonly employed to frighten other users, sometimes with terrifying avatars [V22'3'18, V22'3'26] or with cutesy or diminutive avatars [V22'3'19, V22'3'22] to contrast with the eerie voice.

Virtual scaring. In most cases of *virtual scaring*, attackers suddenly appeared in front of the victim with scaring avatars to frighten them, as illustrated in Fig. 2(b). These incidents of *virtual scaring* often occurred by intentionally using (1) unrealistic physical properties that conflicted with users' common sense. For example, attackers could immerse themselves in the floor and jump out to scare users nearby [V22'5'1]; and (2) virtual representations of physical settings that make people uncomfortable, such as a dark corridor completely blocking users' sight view [V22'4'12].

Virtual trash actions. *Virtual trash actions* included (1) disrupting physical acts, such as laughing at a user while they were describing a negative experience they had encountered [V22'1'10], dancing in front of other users and making farting noises (V22'4'19), peeing at other users [V22'4'6], and reading other users' bios aloud in public [R22'2'1]; (2) unwanted acts, such as repeatedly following other users against their will [V22'1'5, V22'5'20], or intentionally touching other users avatars in a non-sexual manner but making them feel uncomfortable [V22'2'6, V22'2'14], as shown in Fig. 2(c); and (3) verbal acts, such as "mic spamming" that produced abruptly loud sound or playing loud music to force others' attention [VR22'4'6, V22'1'21].

4.1.3 Language Characteristics of Safety Risks. The following showed the language patterns for different safety risks.

Varying frequent words and the use of distinct nouns, verbs, and adjectives shape a unique set of social vocabularies in social VR culture. In *virtual abuse* videos, common verbs included *"accuse"*, as victims' attempted to defend themselves against attackers' *verbal virtual abuse*,

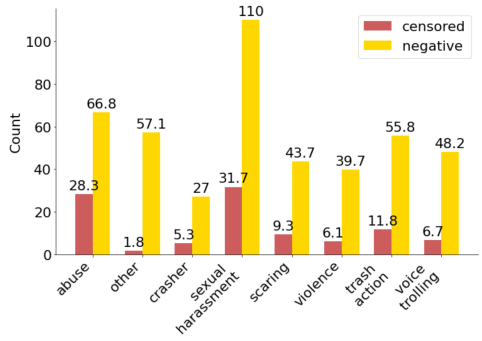

Fig. 3. Average number of censored and negative words per video transcript

Fig. 4. Word cloud based on *trash action* video transcripts

e.g., *"that's not me hey man don't you ever [__] accuse me of some [__]"* [VR22'4'2]. Common nouns were *"furry"* and *"bunny"*, associated with users dressed as animals in social VR, who are often mocked or attacked. These users were often made fun of or assaulted for their appearances, e.g., *"I'm not joking stop talking you're a [__] furry your parents hate you"* and *"most furries commit zoophilia"* [V22'3'21, VR22'5'4]. Common adjectives such as *"stupid"*, *"nuts"*, and *"homo"* were used in attackers' insults and are often offensive and discriminatory. For example, *"where's that homo oh god don't do that cringe [__] Jesus christ you can't be homo"* [V22'3'17].

In *virtual trash action* videos, common verbs included *"smack"*, *"kill"*, *"hate"*, *"shut"*, and *"suck"*, used by attackers to threaten or insult, e.g., *"i don't like that bro shut your [__] axs up i'll smack your dumb axs"* [R22'5'4], or directly used these words as insults by saying *"suck my bxxxs"* [R22'4'1]. Common nouns included *"trash"*, *"hell"*, *"dxxk"*, *"axs"*, e.g., *"axs [__] [__] garbage [__] [__] trash"* and were used in sentences like *"i will [__] you in your axs dude"* [V22'2'13, VR22'4'2]. Common adjectives were *"toxic"*, *"crazy"*, and *"dead"*, which were used by victims to describe the environment, e.g., *"my god bro every time it's so toxic"* and *"there's some crazy dudes out here like spreading [__] it's been it's been a tough night"* [V22'2'13, VR22'5'1].

In *virtual crashing* videos: Common nouns were *"lobby"*, *"anti-cheat"*, *"hackers"*, *"cheat bypass"*, *"problem"*, *"inconvenience"*, and *"avatar"*. For example, *"so yeah basically somebody started selling an easy anti-cheat bypass which basically means the easy anti-cheat isn't really doing anything at that point"* [R22'2'7]; or expressed disapproval or resentment towards cheating, e.g., *"it is basically a hacking lobby you're always finding hackers in there people always doing dumb stuff toxic people it's just not a fun code"* and *"now excuse me i'm gonna hunt some hackers"* [VR22'2'4, R22'2'7].

In *virtual scaring* videos: Common verbs were *"please"*, *"stop"*, *"stuck"*, and *"scared"*, which were utterances of victims in the form of pleas when they experienced *virtual scaring*. For example, *"i thought you was eating my [__] ass that scared the [__] out of me ... can you go first please"* [V22'2'15]. In *virtual violence* videos: Common nouns were *"pxxxs"*, *"kid"*, *"dad"*, *"father"*, which were often used by attackers as insults towards victims, accompanying virtual physical violence. For example, the attacker said *"i am your father you have to listen to what i say i created you with my pxxxs and i can destroy you with my pxxxs i will destroy you with my pxxxs"* to the victim [V22'2'5] and slapped the victim at the same time.

***Virtual sexual harassment* had the highest densities of negative words used and *virtual crashing* the lowest.** See Fig. 3. *Virtual sexual harassment* and *virtual abuse* videos contained significantly more censored words (31.7 and 28.3 each) than other risk types. *Virtual sexual harassment* videos had the highest average number of negative words (110 counts) and the highest censored

word density (2.2%). *virtual violence* videos had the lowest density of 0.8%. It is likely that *virtual sexual harassment* was mainly verbal, while *virtual violence* was mainly physical, explaining the higher density of negative words in the former. Many inappropriate expressions in *virtual sexual harassment* are related to sexual language and gestures, e.g., *"shut yo [__] ass up bro [__] [__] out of here"* [V22'2'2], *"i'm gonna grab your dxxk and i'm gonna twist i'ma bite that [__] off"*, and *"he's giving me kisses i'm loving his bxxt get my [__] and i'm kissing his mouth"* [R22'2'5, V22'3'21].

The most frequently used negative words are "sorry", "bad", and "crash." The most frequently used negative word in *virtual scaring*, *virtual violence*, *virtual trash actions*, and *virtua-l voice trolling* videos was *"sorry"*, which was often used by attackers to apologize to victims after the attacks [V22'2'7, R22'4'2], as shown in Fig. 4. For example, *"i'm sorry that we bullied you so much"*, and *"i'm sorry for harassing you i will never do that again."* Victims sometimes said *"sorry"* when they demanded apologies, e.g., *"you deserve that. Say sorry. You have to stay like that."* Meanwhile, *"bad"* was the most commonly used negative word in *virtual abuse* videos. For example, attackers said *"are you going mad because you're bad"* [VR22'3'2] and *"bro you're actually you're really bad player in this game i'm not kidding"* [V22'5'17]. Victims sometimes also used this word as a counter-attack, e.g., *"come on now he's making me the bad guy"* [VR22'4'2]. *"Crash"* is the most frequently used negative word in *virtual crashing* videos. Attackers used it when plotting and boasting about their attacks, such as *"we're gonna start crashing people right [__] now"* [V22'2'16] and *"this is like the second lobby i've crashed in the video"* [VR22'3'18].

In summary, our findings highlight that social VR poses new safety risks and exacerbates existing online risks. The primary cause of these safety issues is the ongoing negotiation of social norms in this developing community. These risks are ***severe*** (e.g. *virtual violence* can result in offline injuries), ***ambiguous*** (e.g. unclear limits in *role-playing*), and **unexpected** (e.g. users hiding under penetrable floors to scare others).

4.2 Reactions from Victims, Attackers, Bystanders, and Spectators to Safety Risks (RQ2)

This section presented our findings on the association between social cues and safety risks in social VR, shown in Fig. 5, and illustrated the reactions of different users based on YouTube video analysis.

- Attackers: users that attack other users in social VR;
- Victims: users who are harmed as a result of the safety risks in social VR;
- Bystanders: users who are present when attackers attack victims in social VR;
- Spectators: viewers who watch the Youtube videos and post comments.

4.2.1 Reactions from Attackers. ***Attackers used multi-modal social cues, with customized avatars effects and virtual objects, to make the most threatening attacks.*** In 60.4% of the videos in our dataset, attackers' behavior was the result of a synthesis of social cues that, taken together, generated seemingly realistic and threatening effects.

This was especially the case in the risk of *virtual violence*. As shown in Fig. 5, during *virtual violence*, attackers mainly used *kinetics cues*, such as poking others' head [VR22'5'7] and trampling on others' face [R22'2'2]. Meanwhile, these attackers usually displayed other social cues to make their attacks more fierce, such as using *linguistic cues* of verbal threatening [VR22'5'7] or *proximity cues* of rushing back and forth to crash the victims [R22'5'1]. For instance, the attacker in [R22'5'1] performed a malevolent car crush with great realism by driving his car into another object, forcing the object back, and then rushing back and forth to flip the object into the air.

A similar pattern was found in *virtual abuse* and *virtual scaring*. For *virtual abuse* videos, besides using *linguistic cues* such as abusive words, attackers also utilized *proximity cues* such as repeatedly

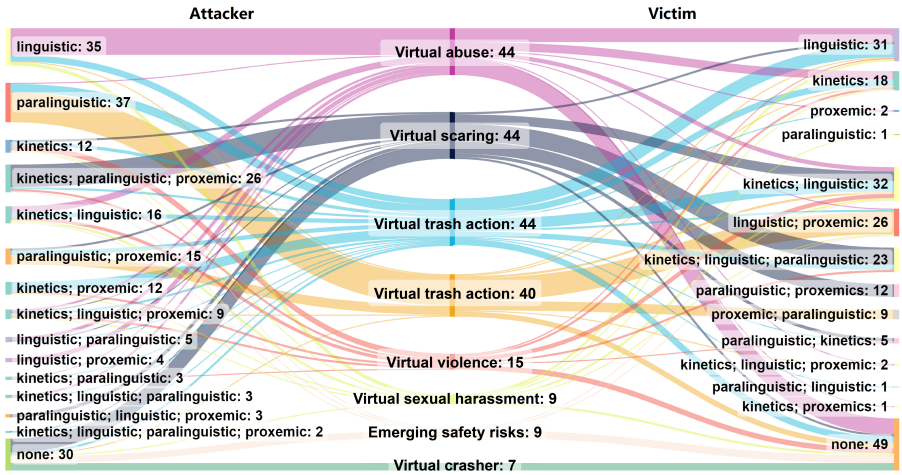

Fig. 5. Social Cues of Attackers and Victims in Each Type of Risks (2022/01/01 - 2022/06/01). The column in the middle indicates the risk type and the corresponding counts of the videos. Columns on the sides represent the combinations of different social cues for attackers (Left) and victims (Right) respectively. "None" category includes cases below: (1) Videos that focus mostly on victims' reactions but do not show the attackers' actions; (2) videos that were cut off just after the attack happened which lack the behaviors of victims, (3) for crashing, the victims were not able to react, because the system was crashed and the avatar could not be manipulated.

following others [VR22'3'12], and *kinetics cues*, such as posing offensive gestures [VR22'5'14]. In *virtual scaring* videos, 95.5% of videos demonstrate a combination of these social cues.

Furthermore, attackers also took advantage of avatar customization and virtual objects in social VR to magnify their threats. For avatars, the majority of *virtual scaring* videos featured customized scary avatars, some of which were self-created by users, for example, zombie-like avatars [V22'2'11]. Others were based on characters in pop culture. For example, in the video [V22'2'1], the scary avatar was based on "Demogorgon" in the television series "Stranger Things", a predatory humanoid creature with a head that could spread like flower petals to reveal sharp teeth and a large mouth inside. Notably, a group of users used identical avatars (e.g., [V22'1'14, V22'1'13]) to chase, block, and scare other users in the public lobby. For objects, weapons were commonly used to point at others (e.g., [V22'3'15, R22'3'6]). In one video [V22'2'8], the attacker used a small avatar with ice axes to climb on others' bodies. With each movement, he pretended to swing the ice axes and stuck them into the victims' bodies, as if their bodies were the mountains he climbed.

4.2.2 Reactions from Victims. ***Victims instinctively covered their heads and ears, shielding themselves with their arms, as if in the physical world, until they realized their efforts were useless in VR.*** Most victims attempted to defend themselves with *kinetics cues* to avoid attacks, such as covering their ears to block out unpleasant sound effects (e.g. [R22'5'3, R22'2'1]), covering eyes to block vision for themselves (e.g. [V22'4'1, V22'1'13]), covering the head (e.g. [V22'2'4] and [V22'1'10]) to protect themselves from being injured by flying virtual objects, or even covering nose when the attacker pretended to fart though they would not smell it in VR [V22'4'19].

Meanwhile, victims sought to physically distance themselves from these safety risks by stepping back [R22'2'4, V22'2'7] or running away [V22'5'20, VR22'3'16, V22'3'23], often accompanied by *para-linguistic cues* of screaming and crying [V22'3'9, V22'2'7]. These two types of reactions were often found in *virtual scaring*, *virtual trash actions*, and *virtual violence* when the attack suddenly and unexpectedly happened.

Unfortunately, several victims only realized afterward that their reactions had been pointless. For example, in the video [V22'2'5], when the attacker scared the victim with a scary avatar and a creepy voice from the ceiling, the victim cried out and swung his arms frantically, trying to drive the attacker away despite being in a virtual world where the attacker could not really touch them. The victim's behavior showcased how he tried his best to exert himself to push the attacker away as if the scary act indeed happened in the physical world.

However, only 13.6% of the videos showed that victims fought back in similar ways, such as starting to swear [VR22'5'7]. Such proactive reactions were mainly identified in *virtual trash action* videos, indicating that relatively few victims attempted to counter-attack or use any safety features in virtual spaces to protect themselves. Sometimes victims exhibited strong *linguistic cues* asking the attacker to cease harmful and damaging behaviors (e.g. [R22'2'2, V22'2'4]). For example, in *virtual trash action* and *virtual abuse*, victims' reactions focused on *linguistic cues*, such as swearing to express resentment [R22'3'1] or complaining *"you're so annoying"* [VR22'5'2], as shown in Fig. 1(a). To our surprise, while most victims mentioned safety features provided by the platform, such as *"ban this guy"* [VR22'5'3] or *"I'm gonna mute him"* [R22'2'6], most of them hardly actually took action to self-protect even though they mentioned these features, except in a video where a victim used a "stop" gesture while other victims did nothing [V22'1'19].

In a few cases, victims were left wondering what had happened. For instance, *"you can't have that voice while you in that avatar"* [V22'3'19] is a common line of victims' reactions in *virtual voice trolling* videos, e.g., [V22'5'3, V22'5'3]. Victims often just walked away, gasped in horror, and then walked back and forth toward the attacker and vocally expressed their surprise when they heard sounds that did not match the avatars or voices that were particularly frightening.

4.2.3 Reactions from Bystanders. ***Bystanders typically either observed, ignored the attack, left the area, laughed, or screamed and cried, as if such safety incidents were accepted in social VR; but sometimes their presence did pose pressure to the attackers.*** Bystanders were found in only about a quarter of the videos in our dataset, and there were a few instances of them trying to interfere with the attack [V22'4'15, VR22'5'5]. In fact, most bystanders acted as if nothing was wrong or even laughed [VR22'3'21, V22'5'11]. For example, in the video [VR22'3'21], an attacker sprayed shiny stuff on a female user and wiped her breast with a rag. Despite the victim's obvious discomfort and panic, the attacker continued to chase and sexually touched her while bystanders watched or laughed instead of showing concern or offering help.

However, bystanders did demonstrate the potential to stop harm in some scenarios. In one case, the bystander witnessed the attacker touching the victim's back, which the victim would not be able to see or feel [VR22'5'1]. Though the bystander did not take any action to stop the attack, the attacker stopped once he noticed that the bystander was watching. The attacker then immediately pretended to start a conversation and asked the bystander to join them. Additionally, the presence of bystanders sometimes may make certain sexual content less personal. For instance, when many bystanders were presented, users tended to be at ease or less bothered by immersive sexual harassment materials, such as when one user spoke filthy jokes [V22'3'8] or exhibited sexual content [V22'3'21].

4.2.4 Reactions from Spectators. We also analyzed spectators' reactions to social VR safety risks based on their comments on these videos.

Table 4. Percentage of spectator reactions per safety risk

Category (num of comments)	Positive	Neutral	Negative
Virtual abuse (29)	20.15%	69%	10.85%
Virtual violence (161)	17.3%	77%	5.7%
Virtual scaring (216)	24.13%	69.14%	6.73%
Virtual crashing (29)	10.75%	82.08%	7.17%
Virtual sexual harassment (177)	11.9%	74.4%	13.7%
Virtual voice trolling (125)	27.36%	68.04%	4.6%
Virtual trash actions (83)	22.35%	72.48%	5.17%

Debates centered on the toxicity of social VR and minor users. In *virtual abuse* videos, 20.15% of comments depicted toxicity in social VR as entertaining and normal, e.g., *"The perfect amount of voice crack, screaming, and explicit cursing. Beautiful!"* [V22'4'15] and *"Toxicity is something needed in vrchat. Honestly, there has to be balanced."* [V22'3'2]. 10.85% of comments reflected spectators' own experiences of being attacked, such as *"vrchat is prettty toxic... I am mute and there are always people that are adults making funs of me bcs of my disability"* [V22'3'2] and *"I got insulted once because I have a French accent."* [V22'3'2]. In 34% of the comments, spectators expressed concerns about toxicity and recommended blocking, reporting, and voting to kick out attackers,e.g., *"If you ain't reporting people, the VRC staff won't know a thing about them or their record. The more reports and kicks a user gets, the more likely they are to drop in rank as a user" [V22'3'2].*

Similarly, in *virtual scaring* videos, 24.13% of comments described such behaviors as entertaining. Spectators praised the scary avatars, e.g., *"I like seeing the spider avatar more because I like spiders every time i WATCH you I feel honoured hearing screams"* [V22'5'1]. Spectators even enjoyed the victim's reaction to these attack and some wished to have that experience themselves, e.g., *"It would be an honour to be jumpscared by you in VR Chat"* [V22'5'1] and *"Love it, hope i will be a victim of this someday"* [V22'5'1]. In contrast, other spectators were scared and terrified, e.g., *"I feel bad for those guys getting scared. Careful with this someone with heart problems might have a heart attack"*[V22'5'1]. A small portion called it "mental abuse"[V22'5'9] in which spectators sympathized the victims, e.g., *"It breaks my heart a little that so many of the victims cry afterwards"* [V22'5'9].

Moreover, spectators expressed mixed opinions about minor social VR users, with "kids" frequently used in the comments. In *virtual abuse* videos, some comments viewed minors as the problem, e.g. *"I have found a bunch of toxic people on vrchat, and most of the time it is literally little kids who just their grubby hands on a vr headset and they just harass everyone."* [V22'3'2]. Some did not mind attacking minors, e.g. *"These kids need to be put in their place"* [V22'4'15]. Others felt bad for minor users, e.g. *"Bro bro the little kid had a mental breakdown"* [V22'1'7]. These reactions align with the findings in 4.1.2 about adult-minor interactions in social VR leading to conflicts or abuse.

Spectators asked "How-To" questions while watching the videos. Spectators were curious about the special effects of social VR. *Virtual crashing* videos received the highest percentage of neutral comments (82.08%), with questions about how to get crashing avatars [V22'2'16] and using "Easy-Anti-Cheat (EAC)" bypass to hide from anti-cheat detection [R22'3'4]. The EAC is a tool to detect malicious behaviors and avoid crashers. However, with this bypass, people can easily cheat and hide from EAC.

Videos of *virtual voice trolling* and *virtual trash actions* received the most positive comments (27.36% and 22.35%, respectively), as shown in Table 4, with spectators commenting on the voice

effects, avatar design, and victims' reactions. For example, spectators were amazed by a boy faking a girl's voice, e.g., "amazingly generated a girl's voice" [V22'3'13]; abrupt change into a deep voice, e.g., *"They're in for another surprise after you reveal your real deep voice hehe"* [V22'3'35]); and novel avatar features, e.g.,*"What the hell is this why would the stupid puffer fish just pee on people"* [V22'4'6].

Being serious, showing empathy, and feeling hurt after watching the safety risks videos. Among all, *virtual sexual harassment* videos received the highest negative comments (13.7%). Many spectators empathized with the victims and shared their personal experiences of *virtual sexual harassment*. For example, some self-disclosed how they were also harassed after revealing their offline gender in social VR, e.g., *"Everytime i say im a girl (which is true) i get sexual harrassed so i change into a boy avatar and dont say anything about my gender i feel safe saying it here because nothing bad can happen i could remember one time i just said i was actually a girl and people just started to im just gonna say slap me on a not so good place lol i didnt play VRchat for a long time after that…"* [V22'3'2].

In *virtual crashing* videos, 7.17% of comments were negative, ranking third among all categories, and with negative words such as "cheat" and "engine" frequently appearing. Spectators expressed concerns about cheating with cheat engines, which allow users to bypass the block and kick features, infinitely jump, block global events, and access infinite avatars and crashing avatars. Spectators also worried about the impact of crashers on hardware and social VR experiences, as seen in the comment *"I don't know if y'all even trying to fix it or already fixed, but there are people selling the script and updating it, the scripts are getting worse and worse with people figuring out how to do all kinds of things, pull out maker pen, get unreleased items and even knowing how to kick people from games and I feel like it just getting worse form here"* [R22'2'7].

In summary, our findings showed that **attackers** frequently gave sensational reactions by combining multi-modal social cues, customized avatars, and virtual objects to make their attacks more threatening. **Victims** usually responded with self-defense gestures similar to those seen in the physical world, along with various emotional reactions (such as crying, screaming, or counter-attacking). **Bystanders** tended to have minimal reactions, such as observing, ignoring, or leaving the scene when a safety risk arises. **Spectators** viewed *virtual voice trolling* and *virtual trash actions* as less severe compared to *virtual crashing* and *virtual sexual harassment*.

5 STUDY 2: EXISTING SAFETY FEATURES IN SOCIAL VR (RQ3)

In Study 2, we evaluated the effectiveness of current safety features offered by social VR platforms in addressing the safety risks identified in Study 1. Given the pervasiveness of the safety risks in social VR, it's crucial to assess the adequacy and effectiveness of these safety features.

5.1 Methods

5.1.1 Research Sites. Recent studies have identified several popular social VR platforms [42, 83]. Among them, Friston et al. compared VRChat, RecRoom, AltSpaceVR, BigScreen, Spatial, and Mozalla Hubs in terms of their system architectures [31]. Freeman et al. explored user experiences of VRChat, RecRoom, AltSpaceVR, BigScreen, Facebook Spaces, High Fidelity, Mozilla Hubs, and so forth [30, 57]; and Blackwell et al. [13] interviewed users of VRChat, RecRoom, AltSpaceVR, Oculus Rooms, Facebook spaces, vTime, and Oculus Venues.

Based on numbers of downloads across popular VR app stores including Oculus Store[4], Steam VR[5], and Sidequest[6], we focused on four popular social VR platforms: (1) *Rec Room*[7], which allows users to manipulate a variety of objects and engage with mini-games with other users; (2) *VRChat*[8], which is known for novel social interactions and performative memes; (3) *Meta Horizon Worlds*[9], which allows users to explore diverse virtual spaces and engage in content consumption and creation; and (4) *AltspaceVR*[10], which is used for live and virtual events, empowering artists, brands, and businesses.

5.1.2 Data Analysis. We conducted a review of existing safety features provided by four social VR platforms. We reviewed the platform's community safety websites to identify the types of safety risks, features, and coping strategies mentioned. The websites were identified by searching for the platform's name and "safety feature." These websites included:

- VRChat (Safety and Trust System): *https://docs.vrchat.com/docs/vrchat-safety-and-trust-system*
- RecRoom (Comfort and Safety): *https://recroom.com/safety*
- HorizonWorlds (Meta Quest) :*https://about.fb.com/news/category/technologies/oculus/*
- AltSpaceVR (User Safety and Moderation): *https://docs.microsoft.com/en-us/windows/mixed-reality/altspace-vr/user-safety*

After identifying the community websites, a researcher thoroughly read the content, noted the safety features and coping strategies mentioned, and summarized and synthesized them across the four platforms. It is important to note that we only summarized the safety features that were explicitly addressed on the platforms' official websites. To verify these features, three researchers used Oculus Quest 2 headsets to trigger the features and identify any new ones. Findings from Study 1 were then combined to determine which of the identified safety risks were and were not addressed, and why some features were not user-friendly.

6 STUDY 2 FINDINGS

6.1 Major Safety Features (RQ3)

We identified 13 safety features provided by four social VR platforms, grouped into three categories, as displayed in Table 5. Specifically, the "*Social Boundaries Settings*" (features 1-7) aim to maintain a comfortable social distance for users in social VR before safety risks arise. "*Quick Reactions*" (features 8-11) are meant to address safety risks once they occur, offering users ways to respond to boundary violations in social VR, including safe zones and safety reports. "*User Agreements*" (features 12-13) inform users of potential dangers and community norms in social VR experiences. For further details, see Appendix A which provides explanations, examples, and screenshots of each safety feature.

Intimacy Proxemic. This feature echoes Hall's personal space theory, which describes that there are four zones in personal space including intimate zone, personal space, social zone, and public distance [32]. See an example illustrated in Fig. 9.

Trust Reputation. This feature usually assesses users' virtual trustworthiness through activities such as uploading content, engaging with different users, and others interacting with your content, which demonstrates how a user wants to keep the platform a safe space. When displayed, the rank

[4]https://www.oculus.com/experiences/quest
[5]https://store.steampowered.com/vr
[6]https://sidequestvr.com/all-apps
[7]https://recroom.com
[8]https://hello.vrchat.com/
[9]https://www.oculus.com/horizon-worlds
[10]https://altvr.com

Table 5. High-level Analysis of Safety Features Across Social VR Platforms

	Safety Features	Cues Involved	Design Contexts	VRChat	RecR	Horizon	AltSpace
	(1) Intimacy Proxemics	Proxemics	Virtual bodily harassment [11]	✓	✓	✓	✓
Boundary	(2) Trust Reputation	Proxemics	Idling, bad quality content, mass-friending [12]	✓			
Settings	(3) Social Spaces	Proxemics	Unwanted social connections [13]	✓	✓		
(Before)	(4) Shield Levels	Proxemics	Unwanted social connections [14]	✓	✓	✓	
	(5) Voice Control	Para-linguistics	Garble voices [15]			✓	
	(6) Avatar Control	Eyes and Faces	Privacy, unwanted touch [16]	✓	✓		
	(7) UserDemographics	NA	Parental supervision over teens [17]	✓	✓	✓	
Quick-	(8) Safety Gestures	Kinetics	Disconnect with individual [18]		✓		
Reactions	(9) Safety Zone	Proxemics	Unwanted touch and talk [19]	✓	✓		✓
(After)	(10) Vote Kick	Proxemics	Kick out unwanted users in private space				
	(11) Safety Reports	NA	Last mins activities will be recorded [20]	✓	✓	✓	✓
Agreements	(12) Codes of Conduct	NA	Harass, abuse, offend, impersonate[21]	✓	✓	✓	✓
(All-time)	(13) Consent	NA	Terms of services [22]	✓	✓	✓	✓

signals to users that they have positively contributed to the platform and have earned the trust of other users. See an example illustrated in Fig. 10.

Social Spaces. *Private space* (e.g., dormitory) is a controlled space where all individuals are known to each other. *Semi-public space* is a semi-private space where all individuals are associated with each other. In contrast, *public space* is open to different, separate groups of people who might not have established relationships or connections. Room owners are fully responsible for ensuring their rooms not devolving into a toxic mess, even if they are going on vacations in the offline world. We found that local expectations also exist for conducts in *public spaces*. For example, it must be private if a club involves sexual themes. See an example illustrated in Fig. 11.

Interaction Shields. Users can adjust control settings to maintain desirable social boundaries with others (based on different users) [23]. We describe these adjustable settings as shields. There are multiple aspects that users can shield, e.g., voice, avatar, audio, animation, particles and lights, and so forth. See an example illustrated in Fig. 12.

Voice gibberish. Horizon Worlds has a feature of "Garble Voices," which is a voice mode that turns voice chat from any nearby non-friends into "unintelligible, friendly sounds" that will make users feel safer. An icon will appear above the display names to indicate to others that "garble" is being used. See an example illustrated in Fig. 13.

Avatar Shield. All platforms studied in this paper provide embodied avatars. However, how they support different avatars vary. Several platforms support avatar customization, and others (e.g. VRChat and Mozilla Hubs) allow importing customized avatar models from third-party applications. Users can both adjust their own avatars' identities, such as gender, race, and voice, and select to hide or show specific users' avatars. For example, in VRchat, when a user encounters a user that, despite their higher Trust Rank, is wearing an avatar they feel uncomfortable, they can choose "hide avatar" in their social panel. See an example illustrated in Fig. 14.

User Demographics. Age limits are set in social VR. For example, Rec Room enables a junior account managed by an account owned by the parent or guardian for users under 13 years of age, which comes with safety features to prevent children from disclosing their identifiable information in social VR. It also allows age-based matchmaking to recommend players in VR games close to their age range. See an example illustrated in Fig. 15.

Safety Reactions. These are communication gestures and shortcuts that allow quick-action remediation in challenging situations. For example, Rec Room's "talk to the hand" and pointing gestures can instantly trigger users' comfort and moderation menu, allowing users to mute, block,

[23]https://docs.vrchat.com/docs/vrchat-safety-and-trust-system

Table 6. Mapping Safety Risks with Safety Features

Safety Risks (RQ1)	Attacker Cues (RQ2)	Victim Cues (RQ2)	The Gap of Safety Features (Mapping)
Virtual violence	Mostly kinetics; Some with (para-)linguistics	Most kinetics and linguistics Some kinetics	(1)-(4) Proxemics designs No kinetics, (para-)linguistics designs
Virtual crashing	Malicious	Most kinetics and linguistics Some (para-)linguistic	(8-11) Quick reactions No regulation
Virtual scaring	Mostly kinetics, paralinguistic, proxemics Some paralinguistic, linguistic, proxemics	Most kinetics, (para-)linguistics Some para-linguistics, proxemics	(1)-(5) Proxemics and papra-linguistics No kinetics designs
Virtual abuse	Mostly linguistics; Some with kinetics or para-linguistics	Most linguistics or kinetics Some kinetics and linguistics	(5) voices settings (para-linguistics) No linguistics and kinetics design
Virtual sexual harassment	Mostly linguistics; Some linguistics with kinetics	Most kinetics or/and linguistics	(1)-(4) Proxemics designs No linguistics and kinetics design
Virtual voice trolling	Mostly paralinguistic; Some pare-linguistic and proxemics	Most linguistics and proxemics Some proxemics and paralinguistic	(1)-(3) Proxemics , (5)Para-linguistic No linguistics designs
Virtual trash actions	Malicious	Most linguistics or/and kinetics Some kinetics, (para-)linguistics	Multiple features Should focus on victim's cues
Emerging virtual risks	NA	NA	Notice technical issues like crashing

and report other users. These mechanisms enable users to report a hurting experience instantly without interrupting or degrading their experience. See an example illustrated in Fig. 16.

Safety Zone. VRchat has a shortcut called "safe mode," which can immediately disable all features on all users around. Horizon worlds offer a one-touch button called "Safe Zone" that can quickly remove users from a situation. Similarly, AltspaceVR has a "title screen" feature, which immediately drops users back into the command center, removing their avatars from the previous space and scene. See an example illustrated in Fig. 17.

Vote Kick. Users may issue a vote kick against another user by using the stop gesture or by navigating to the user's profile via the people menu. All other users in the room will be alerted to the vote kick initiation and asked to vote yes or no. This action requires collective reactions, so we separate it from the feature (8). See an example illustrated in Fig. 18.

Safety Reports. Users can report other users, rooms, or event places that violate codes of conduct or ask for extra support. See an example illustrated in Fig. 19.

Codes of Conducts. Some platforms, for example, Rec Room, require users to finish a tutorial on codes of conduct for the platform. These norms are introduced to newcomers and repeatedly displayed to members of the communities. See an example illustrated in Fig. 20.

Informed Consent. All platforms have informed Consent before users enter the platform. Informed consent often centers on health and safety warnings ranging from the immersive effects on users' psychological status, a notice of risky content during social interactions, and user privacy. See an example illustrated in Fig. 21.

6.2 Mapping the Limitations of Safety Features (RQ3)

The results of Study 1 indicated that only a few of the videos in the sample demonstrate people utilizing safety features effectively to address safety risks in social VR, e.g., [VR22'5'11, V22'3'30]. This suggests that the current safety features may not be adequate for ensuring user safety. To better understand the limitations of current safety features and users' practices, we further analyzed the relationship between safety risks (RQ1), social cues (RQ2), and the current social VR safety features (RQ3).

Overall, the social cues of attackers and victims primarily involve proxemics, linguistics, para-linguistics, and kinetics, but these cues are not adequately addressed by the existing safety features. We presented our findings in Table 6. Below, we summarized three limitations that were identified in the existing safety features of social VR platforms:

Modifying the proximity-centered safety design: Proxemics has been widely used as the primary cue in designing safety features across social VR platforms. However, our findings suggested that different types of safety risks may have distinct social cue characteristics. Using a proxemics-only approach is insufficient in addressing all types of safety risks. For instance, linguistic cues are more useful in addressing *virtual sexual harassment* and *voice trolling*, while the cue of kinetics is crucial in addressing *virtual violence* and *virtual scaring*.

Lacking natural reactive gestures: In social VR, gestures like high-fiving and clapping play an important role in social interactions. However, only the "stop gesture," or "talking to the hand gesture," has been designed for banning users. These shortcuts were barely found in existing safety features. Utilizing gestures as safety features offers a significant advantage, as it allows victims to quickly block offenders without having to navigate through a menu [59]. This can be particularly useful when the victim is flustered during a safety risk. Our research shows that kinetics is linked to almost all types of victim risks, and some users demonstrated intuitive responses to specific types of risks. For instance, when users wanted to block out the sounds or felt threatened by other users, they instinctively covered their ears and heads. Designing hotkeys and shortcuts for safety features based on these natural reactive gestures could make it easier for users to manage safety risks.

Transforming reactive solutions into preventive protections: Most existing safety features in social VR are reactive in nature, meaning that they only address safety risks after they have occurred. This is reflected in their design, which typically involves setting options or limiting certain actions by users. However, this approach may not be enough to fully address the growing safety concerns in virtual reality. Future designs should place more emphasis on preventative safety features. This could involve incorporating social cues from attackers and their behaviors to identify potential risks before they occur. For example, monitoring linguistic cues such as certain words or phrases that may indicate risk, or using kinetics to detect if a user is performing gestures that may be harmful to others.

7 DISCUSSION

Our research has investigated current safety risks presented in social VR platforms, taking into account the unique dynamics of social VR interactions and users' multi-modal reactions to such risks. We have also highlighted the limitations of existing safety designs. In this section, we discuss how these findings inform HCI and CSCW researchers to re-evaluate and re-approach novel safety risks in social VR. We also propose three design implications based on our findings to enhance user safety in social VR.

7.1 Re-Approaching Safety Risks in Social VR

A growing number of HCI studies have focused on the emerging risks in social VR, yet they mainly relied on interview [13, 30] and survey methods [62, 70]. While participants in these studies have often described behaviors or encounter that they perceive as harassment [30], there currently exists a lack of a shared vocabulary across various social VR platforms to describe safety risks in a consistent manner [13]. That is, the *definitions* of what constitutes safety risks also remain vague. Our research contributes to the development of a shared community vocabulary for describing how specific interaction behaviors may be viewed as safety risks in social VR.

We reported the characteristics of these safety risks as severe, ambiguous, and unpredictable, through content analysis of videos, transcripts, and comments, and also explicated the actual incidents of these safety risks with contextual details (e.g., environment, object, and bystander reactions). This approach offers valuable insights into the first-hand experiences of both the victims

and attackers and the reactions of bystanders within social VR and spectators outside social VR watching these videos. This is a unique contribution compared to studies that focused on inquiring about victims' [13, 30] and bystanders' [16] prior experiences with safety risks.

These insights motivate us to re-evaluate and re-approach these novel safety risks, thereby fostering a deeper understanding of how to mitigate these risks in the future. Below are several highlights.

The immersive nature of social VR can result in realistic, simulated bodily injury due to safety risks. Our findings demonstrate that despite occurring in virtual spaces, these safety risks can have substantial and even physical consequences in the offline world. One instance is the case of virtual violence where the victim fell to the ground. In such scenarios, users may experience "*phantom limb pain*", which occurs when people treat fake limbs as if they were real and feel pain and physically react when they are hurt. Therefore, when players feel that their virtual characters are at risk in social VR, similar reactions can occur. This highlights why social VR may pose a greater risk to users compared to traditional social platforms [13].

The amount of personal information disclosed through embodiment and flexible avatar control in social VR can lead to more targeted attacks. It is easier to target potential victims for *virtual abuse* and *virtual sexual harassment* (e.g., women) due to the primary use of voice communication and enhanced avatar control (e.g., full body tracking). We found that many *virtual sexual harassment* actually took place in public lobbies where attackers could easily access and identify their victims. Similar to what has been shown in earlier research [13], *virtual sexual harassment* may take many forms in social VR due to the multi-sensory experience. Also, prior studies have shown that harassment in social VR is a simulated immersive experience (in the environment, activities, and conducts) compared with traditional online gaming and virtual worlds, which can increase the possibilities of harassment [30].

Confusing design features lead to various safety risks that are against common sense. This means, how a social VR platform is designed, rather than individual users alone, may also pose safety risks. For example, the ability to hide under the floor [V22'5'1] or pass through walls [V22'2'1] can be used to intimidate users. Attackers may exploit design features that go against users' expectations, leading to severe psychological trauma if platforms do not provide adequate warnings or guidelines. Additionally, realistic environmental settings, like a vivid cave, can also trigger the *"uncanny valley"* effects [40] that cause fear or discomfort. Additionally, platform designers should consider potential novel safety risks of *"misread cue"* [VR22'5'9] and implement measures to prevent such misuse. A safe and secure virtual environment is crucial for a positive user experience and to prevent potential harm to users.

Many safety risks in social VR were influenced not only by design issues but also by community standards and the lack of clear guidelines in the grey areas. We found that crashers in *virtual crashing* felt that they were preserving social order in VR being *"chivalrously"*, particularly against VR streamers, and believed that streaming in public VR spaces violated privacy. This highlights the need for clear policies on streaming in VR as the lack of regulation led to such incidents. Our research also sheds light on the growing subcultures and user groups in social VR, such as *furry fandom*, *role-playing* users, and *cross-dressing*, which current rules and regulations fail to address. For example, the VRChat community has noted that *role-playing* can be used as a guise for malicious behavior, but there is currently no clear definition of the boundaries in this regard [24]. Although the performing nature of social interactions in VR adds adventurous experiences to social VR interactions [60], this also creates a grey area for potential risks because it can be difficult to distinguish between performing and actual harmful behavior. Further research is necessary

[24]https://hello.vrchat.com/community-guidelines

to understand the nuances of role-playing in social VR and how to differentiate it from harmful conduct.

The subjectivity of users' perceptions of social norms, which can differ from offline interactions, exacerbates safety risks in social VR. Our findings echo previous research stating that the definition of online harassment in social VR is personal and highly subjective, making it challenging to regulate[13, 30]. For instance, it is difficult to determine when the boundary is crossed in incidents involving *immersive dwellers*. Similarly, attackers often view *virtual harassment* as "just a game" where offline social norms do not apply, making it hard to distinguish it from actual harassment. The co-existence of adults and minors in social VR can create tension, with minors potentially harassing adults due to differences in interaction dynamics across age groups [54]. In our work, we further found that this behavior may result from minors imitating inappropriate behaviors displayed by adults displaying inappropriate content. Also, we did find that risks that adults posed to children were consistent with those reported in prior work [53, 54], such as displaying inappropriate content when minors might be present [V22'3'8] and abusing minors [VR22'3'16]. Some players reported feeling uncomfortable in environments with minors, as they felt the need to be mindful of their words and actions, just as they would in the offline world. Thus, future research is needed to explore how and to what extent offline social norms can be strengthened, implemented, or relaxed in the context of social VR.

The abuse or misuse of safety features in social VR can also pose significant challenges to managing safety risks. Our research has revealed instances where users in a private room abused the voting safety feature to collectively force a victim to leave the room. This misuse of safety features highlights the potential for abuse within the group dynamic in social VR environments and presents a pressing issue for social VR safety in the future. To mitigate such challenges, previous studies suggest incorporating ethical considerations into the design process, such as considering the ethical dimensions of technology use and researching the use and misuse of similar technologies [49]. This approach can help ensure that social VR platforms are designed and managed in a socially responsible manner.

Therefore, we urge the HCI community to re-examine the ways in which safety in social VR can be effectively researched and developed. It is imperative to address the pressing need to continually revise and make clear the community guidelines, so as to align with the continuously changing social VR culture, thereby reducing ambiguity and limiting the potential for harmful and hazardous actions to occur within these virtual environments.

7.2 Designing Reactive and Preventive Safety Features with the Aid of Cues

Prior research has primarily focused on individual cues, such as hand movements, body movements, and kinetics in social VR [37, 64], without conducting a holistic examination of social cues. Studies have revealed that social VR users tend to communicate social cues through non-verbal behaviors to strengthen relationships in online communities [57, 76]. Our findings contribute to the current understanding by demonstrating that social cues can also effectively convey important information in addressing safety concerns in social VR – Different cues can have varying significance in different safety risk situations. For instance, *kinematic cues* are often relied upon heavily by both attackers and victims of *virtual violence*, while *linguistic cues* are relied upon more heavily by those experiencing *virtual abuse*. Also, attackers who aim to scare others often use a combination of paralinguistic and kinematic cues. This suggests that it may be possible to categorize the use of different cues and create standardized safety triggers across various social VR platforms. Moreover, victims consistently exhibit similar kinematic responses to various instances of *virtual violence*, such as covering their eyes or ears when feeling threatened or during mic spam. These patterns indicate that victims often respond in a similar manner to particular safety risks. Based on our

research findings, we suggest the implementation of several design features to improve the safety of social VR environments.

7.2.1 Non-player Characters (NPCs) as Safety Companions. Our observations from the videos revealed that the majority of users learned about the safety features either through accidental triggers. However, we also noted instances where bystanders stepped in and instructed at-risk users on how to protect themselves. This highlights the need for user education on recognizing and managing safety risks in social VR. To address this issue, we propose the design of a Non-Playable Character (NPC) safety risk educator.

In the world of video games and virtual environments, NPCs are computer-controlled characters that serve various purposes. They have been found to enhance players' gaming experiences, such as improving immersion [39] and creativity [65], as well as provide guidance and customization options for players [15]. Additionally, research suggests that users respond to NPCs as if they were real people [41]. Given the importance of NPCs in virtual environments, it is surprising that they are not widely used in social VR platforms (except it is used to simply direct users' attention in a multiple-user scenario [59]). We propose the opportunity for using NPCs as a tool to educate users about safety risks in social VR.

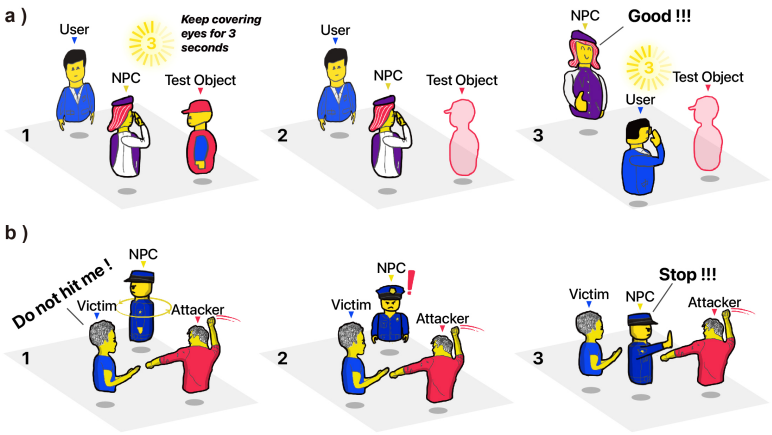

Fig. 6. a) *NPCs act as an educator*: In personal learning mode, an NPC is teaching the user to make the test object in front of him transparent and contoured using an eye-covering gesture. b) *NPCs act as mediators*: In a public area, a "mediator" NPC is patrolling, suddenly it finds a user is about to attack another user, the NPC immediately goes up and stops it.

Scenario 1: As shown in Fig. 6(a), in the video [VR22'2'6], a user is shocked when they are subjected to verbal abuse by a group of attackers and is kicked out of public space. To address this type of safety risk, an NPC can be designed to teach the user different safety gestures. The NPC would simulate different safety risk scenarios, showing how an attacker may try to harm them, and then demonstrate different gestures to trigger specific safety protection functions. This design recommendation is based on the concept of *social learning*, which involves observing or interacting with others to acquire new knowledge or skills. Social learning is particularly effective because it allows individuals to learn from the experiences of others, making changes in behavior, cognition, or emotional state as a result. Research has also shown that humans are more likely

to adopt behaviors that they observe in others, especially when those behaviors provide valuable information for survival or well-being [48, 50].

Scenario 2: As shown in Fig. 6(b), the attacker stopped their harmful actions upon realizing that other users were watching [VR22'5'1]. This evidence can inform the design of NPCs as mediators or virtual police officers in public areas. For example, some NPCs dressed as virtual police officers could patrol the space and the more safety reports received, the more virtual police NPCs would be present in the area. This would allow users to assess the safety of an area based on the number of patrolling NPCs and avoid potentially dangerous areas. This design could be made possible through crowd-sourced reporting, similar to reporting traffic situations through data collection, integration, analysis, prediction, and mining[78].

To address the issue of adult-minor co-existence in social VR, we recommend designing a system similar to movie ratings for auto-detecting inappropriate content [4, 14]. An NPC can remind minor users of reported inappropriate content in a certain area and suggest alternative spaces for them to play in.

7.2.2 Personalized and Natural Reactive Safety Measures with Easy Triggers. Gestures are a key way for users to express their intentions and activate system operations in VR environments [76]. However, the current "talking to hand" gesture in Recroom is not enough to effectively trigger safety features and protect users from attackers who can escape [43]. Harassment actions in social VR occur quickly, giving victims no time to respond or trigger safety features [V22'2'1]. In some platforms, victims may need to go through historical records to ban attackers who have already left. Proximity-based safety designs, such as space bubbles or 4-foot boundaries, are ineffective as attackers can constantly hover and use the boundaries to prevent intervention. Developing gesture-based safety features is necessary to better protect victims.

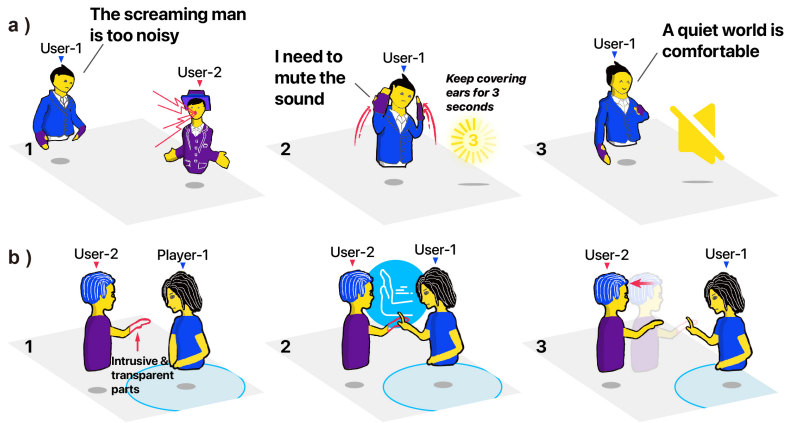

Fig. 7. a) **Covering ears**: The victim encountered screaming with another user. The victim found it too harsh and covered his ears with his hands for 3 seconds. This triggered the system's mute mechanism, which helped the victim turn off the ambient sound. b) **Push-away action**: When a user invades inside another user's boundary, the invaded user can choose whether or not to push him/her out of the way. If a push-away action is made, a push-away sign appears and pushes the invader away.

Our findings of the victims' behaviors revealed common, unconscious gestures made during initial attacks, such as covering their eyes or ears, screaming, backing up, and falling to the ground.

These gestures can inform the design of future safety features, with the goal that users can quickly protect themselves from harm using unconscious actions and overcome the difficulty in triggering protection mechanisms in complex systems.

Scenario 3: Based on observations, e.g.,[R22'5'3, V22'1'13], as illustrated in Fig.7(a), in a situation of abuse or mic-spamming, the victim can cover their ears for three seconds and the system will automatically eliminate the surrounding noise. Similarly, if the victim covers their eyes due to fear or encountering offensive behavior, the attacker will be highlighted in front of the victim temporarily.

Scenario 4: Based on the scenario, e.g., [V22'2'6], as shown in Fig. 7(b), When an intruder appears, the victim can push them away with a push-away gesture. This gesture can be set individually by the victim or automatically detected by the system, with the number of times the intruder is pushed away determining the size of the boundary distance. This allows the victim to create a personalized, protective boundary range.

7.2.3 Enhanced Controls for Avatar and Voice Shields. Our research found that attackers exhibit unique traits and behaviors during the attack preparation phase. For instance, they often establish a terrifying avatar and use a corpse-like voice. We can draw inspiration from existing privacy preservation measures, such as platform-embedded voice modulators [28] and the *"proteus effect"* where people's behavior aligns with their digital representatives [6]. This leads us to recommend restrictions on avatar appearance and vocal qualities to prevent harm. For example, users should be alerted of a potentially risky avatar or voice and given the option to see or hear it.

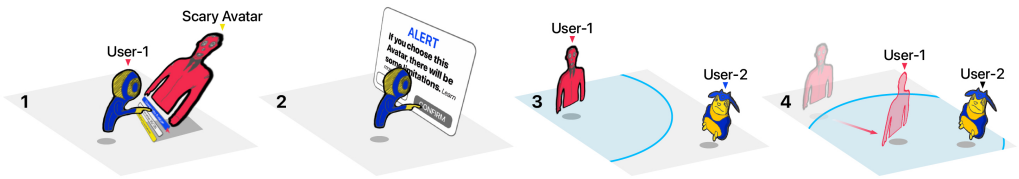

Fig. 8. **Enhanced Controls for Avatar and Voice Shields**: A user is trying to use the dreaded Avatar and still uses it after being prompted by the system. When he is at another user's distance, he can be seen by this user. But when he approaches, the system makes the scary avatar transparent and outlined, making it impossible for the user to see the scary avatar.

Scenario 5: As shown in Fig.8, a user sets up a scary avatar as shown in Fig.8. The system informs the user that restrictions will take effect once the setup is complete. Despite the warning, the user persists with the scary avatar. In public spaces, the avatar is automatically reduced to a translucent outline for the safety of other users. Using corpse-like or uncomfortable deep voices is restricted and only allowed in specific VR games. This helps prevent potential attacks and ensures the safety of all users.

8 LIMITATIONS AND FUTURE WORK

The limitations of using YouTube videos as the main data source for our research must be acknowledged. Firstly, some of the videos in our dataset may be performative in nature, created with the intention of drawing public attention, and may not reflect the actual situation of safety risks in social VR. To gain a comprehensive understanding of users' experiences, it is essential to conduct further research utilizing various data sources such as participatory observations within social VR

and data collected from platforms like Twitch. Secondly, our results based on YouTube videos may not be comprehensive, as our dataset is limited to English-based videos posted in the recent year. However, our approach aligns with the scope of other YouTube-based analysis[3, 51, 71], and our findings may still provide valuable insights into the safety risks in social VR. Lastly, the variability of the topics and qualities of the videos presents challenges in the coding process. For example, it was challenging to identify videos that contain actual harmful impacts and to distinguish between the languages of attackers, victims, and bystanders in the transcripts. Our coding process relied partly on our expertise as CSCW/HCI researchers and designers.

Despite the limitations and challenges, our approach of video content analysis offers valuable insight into the safety risks in social VR, based on first-hand evidence and real-time observations of people's actions and reactions. As such, YouTube videos provide a rich source of data that can be useful in advancing the field of CSCW and HCI research, without the need for researcher presence and on-site observations [63]. Future research may focus on examining the extent to which first-person behaviors reflect their perceptions of risks, as well as exploring the unique reactions of special groups, such as LGBTQ, minority, and disabled users when they experience safety risks in social VR. In addition, while the primary context in our dataset is social interaction in public areas, it would be beneficial to explore risks in other settings, such as game environments and private rooms.

9 CONCLUSION

In conclusion, this study sheds light on the various safety risks present in social VR, as documented through analysis of Youtube videos. The risks are found to be severe, ambiguous, and often unexpected. Through analysis of the videos and Youtube users' comments, the study highlights the varying reactions of attackers, victims, bystanders, and spectators to such risks, which are influenced by their differing understandings of social VR community norms and the application of offline cultural norms in virtual spaces. By mapping the gap of these safety risks and existing safety features, our findings also point to the need for improved safety designs in social VR systems to create safer virtual environments for users.

REFERENCES

[1] 2021. *FACEBOOK'S METAVERSE One incident of abuse and harassment every 7 minutes.* https://counterhate.com/research/facebooks-metaverse,urldate={2021-12-30}

[2] Dane Acena and Guo Freeman. 2021. "In My Safe Space": Social Support for LGBTQ Users in Social Virtual Reality. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.

[3] Maria Vraka Mikael B. Skov Aida Komkaite, Liga Lavrinovica. 2019. Underneath the Skin: An Analysis of YouTube Videos to Understand Insertable Device Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[4] Sharifa Alghowinem. 2018. A Safer YouTube Kids: An Extra Layer of Content Filtering Using Automated Multimodal Analysis. In *Proceedings of SAI Intelligent Systems Conference*. 294–308.

[5] Kimberley R Allison and Kay Bussey. 2016. Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying. *Children and Youth Services Review* 65 (2016), 183–194.

[6] Daniel Görlich Anna Samira Praetorius. 2020. How Avatars Influence User Behavior: A Review on the Proteus Effect in Virtual Environments and Video Games. In *International Conference on the Foundations of Digital Games*. 1–9.

[7] Lisa Anthony, YooJin Kim, and Leah Findlater. 2013. Analyzing user-generated youtube videos to understand touchscreen use by people with motor impairments. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1223–1232.

[8] Budi Arief and Mohd Azeem Bin Adzmi. 2015. Understanding cybercrime from its stakeholders' perspectives: Part 2–defenders and victims. *IEEE Security & Privacy* 13, 2 (2015), 84–88.

[9] Budi Arief, Mohd Azeem Bin Adzmi, and Thomas Gross. 2015. Understanding cybercrime from its stakeholders' perspectives: Part 1–attackers. *IEEE Security & Privacy* 13, 1 (2015), 71–76.

[10] Maura Barrett and Forte Douglas. 2019. *Metaverse virtual worlds lack adequate safety precautions.* https://www.nbcnews.com/tech/internet/metaverse-virtual-worlds-lack-adequate-safety-precautions-critics-say-rcna15418

[11] Tanya Basuarchive. 2021. *The metaverse has a groping problem already.* https://www.technologyreview.com/2021/12/16/1042516/the-metaverse-has-a-groping-problem

[12] Nicole A Beres, Julian Frommel, Elizabeth Reid, Regan L Mandryk, and Madison Klarkowski. 2021. Don't you know that you're toxic: Normalization of toxicity in online gaming. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[13] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in social virtual reality: Challenges for platform governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[14] Zanoni Dias Anderson Rocha Bruno Malveira Peixoto, Sandra Avila. 2018. Breaking down violence: A deep-learning strategy to model and classify violence in videos. In *In Proceedings of the 13th International Conference on Availability, Reliability and Security (ARES 2018)*. 1–7.

[15] Sri Kalyanaraman Daniel Pimentel. 2020. Your Own Worst Enemy: Implications of the Customization, and Destruction, of Non-Player Characters. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 93–106.

[16] Emily Dao, Andreea Muresan, Kasper Hornbæk, and Jarrod Knibbe. 2021. Bad breakdowns, useful seams, and face slapping: Analysis of vr fails on youtube. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[17] Ellysse Dick. 2021. *Balancing User Privacy and Innovation in Augmented and Virtual Reality.* Technical Report. Information Technology and Innovation Foundation.

[18] J Donath. 2011. Signals, cues and meaning (February draft for Signals, Truth and Design. MIT Press).

[19] Nicolas Ducheneaut, Ming-Hui Wen, Nicholas Yee, and Greg Wadley. 2009. Body and mind: a study of avatar personalization in three virtual worlds. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1151–1160.

[20] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies* 132 (2019), 138–161.

[21] Cristina Fiani and Stacy Marsella. 2022. Investigating the Non-Verbal Behavior Features of Bullying for the Development of an Automatic Recognition System in Social Virtual Reality. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces*. 1–3.

[22] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[23] Brian J Fogg. 2002. Persuasive technology: using computers to change what we think and do. *Ubiquity* 2002, December (2002), 2.

[24] Deen G Freelon. 2010. ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science* 5, 1 (2010), 20–33.

[25] Guo Freeman and Dane Acena. 2021. Hugging from A Distance: Building Interpersonal Relationships in Social Virtual Reality. In *ACM International Conference on Interactive Media Experiences*. 84–95.

[26] Guo Freeman and Dane Acena. 2022. " Acting Out" Queer Identity: The Embodied Visibility in Social Virtual Reality. *Proceedings of the ACM on humancomputer interaction* 6, CSCW2 (2022).

[27] Guo Freeman, Dane Acena, Nathan J McNeese, and Kelsea Schulenberg. 2022. Working Together Apart through Embodiment: Engaging in Everyday Collaborative Activities in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–25.

[28] Guo Freeman and Divine Maloney. 2021. Body, avatar, and me: The presentation and perception of self in social virtual reality. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–27.

[29] Guo Freeman, Divine Maloney, Dane Acena, and Catherine Barwulor. 2022. (Re) discovering the Physical Body Online: Strategies and Challenges to Approach Non-Cisgender Identity in Social Virtual Reality. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

[30] Guo Freeman, Samaneh Zamanifard, Divine Maloney, and Dane Acena. 2022. Disturbing the Peace: Experiencing and Mitigating Emerging Harassment in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–30.

[31] Sebastian J Friston, Ben J Congdon, David Swapp, Lisa Izzouzi, Klara Brandstätter, Daniel Archer, Otto Olkkonen, Felix Johannes Thiel, and Anthony Steed. 2021. Ubiq: A system to build flexible social virtual reality experiences. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*. 1–11.

[32] Edward T Hall, Ray L Birdwhistell, Bernhard Bock, Paul Bohannan, A Richard Diebold Jr, Marshall Durbin, Munro S Edmonson, JL Fischer, Dell Hymes, Solon T Kimball, et al. 1968. Proxemics [and comments and replies]. *Current anthropology* 9, 2/3 (1968), 83–108.

[33] RYAN HANDLEY, BERT GUERRA, RUKKMINI GOLI, and DOUGLAS ZYTKO. 2022. Designing Social VR: A Collection of Design Choices Across Commercial and Research Applications. (2022).

[34] Dave Harley and Geraldine Fitzpatrick. 2009. Creating a conversational context through video blogging: A case study of Geriatric1927. *Computers in Human Behavior* 25, 3 (2009), 679–689.
[35] Juan Pablo Hourcade, Sarah L Mascher, David Wu, and Luiza Pantoja. 2015. Look, my baby is using an iPad! An analysis of YouTube videos of infants and toddlers using tablets. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1915–1924.
[36] Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Seattle, WA, USA) *(KDD '04)*. Association for Computing Machinery, New York, NY, USA, 168–177. https://doi.org/10.1145/1014052.1014073
[37] Cloe Huesser, Simon Schubiger, and Arzu Çöltekin. 2021. Gesture interaction in virtual reality. In *IFIP Conference on Human-Computer Interaction*. Springer, 151–160.
[38] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 216–225.
[39] Ryan Ng Jeffrey C. F. Ho. 2020. Perspective-Taking of Non-Player Characters in Prosocial Virtual Reality Games: Effects on Closeness, Empathy, and Game Immersion. In *Behaviour Information Technology*. 1185–1198.
[40] Ernst Jentsch. 1997. On the Psychology of the Uncanny (1906). *Angelaki: Journal of the Theoretical Humanities* 2, 1 (1997), 7–16.
[41] Dylan Arena Jesse Fox and Jeremy N. Bailenson. 2009. Virtual Reality: A Survival Guide for the Social Scientist. In *Journal of Media Psychology Theories Methods and Applications*. 95–113.
[42] Marcel Jonas, Steven Said, Daniel Yu, Chris Aiello, Nicholas Furlo, and Douglas Zytko. 2019. Towards a taxonomy of social vr application design. In *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. 437–444.
[43] Katherine Isbister Joshua McVeigh-Schultz, Anya Kolesnichenko. 2019. Shaping Pro-Social Interaction in VR: An Emerging Design Framework. In *Proceedings of CHI Conference on Human Factors in Computing Systems*. 1–12.
[44] Franccesca Kazerooni, Samuel Hardman Taylor, Natalya N Bazarova, and Janis Whitlock. 2018. Cyberbullying bystander intervention: The number of offenders and retweeting predict likelihood of helping a cyberbullying victim. *Journal of Computer-Mediated Communication* 23, 3 (2018), 146–162.
[45] 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*. 241–252.
[46] Aida Komkaite, Liga Lavrinovica, Maria Vraka, and Mikael B Skov. 2019. Underneath the Skin: An Analysis of YouTube Videos to Understand Insertable Device Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[47] Nicole C Krämer. 2005. Social communicative effects of a virtual program guide. In *International Workshop on Intelligent Virtual Agents*. Springer, 442–453.
[48] D. COWNDEN M. ENQUIST K. ERIKSSON M. W. FELDMAN L. FOGARTY S. GHIRLANDA T. LILLICRAP L. RENDELL, R. BOYD and K. N. LALAND. 2010. Why Copy Others? Insights from the Social Learning Strategies Tournament.. In *SCIENCE*. 208–213.
[49] Qinglan Li and Ioana Literat. 2017. Misuse or misdesign? Yik Yak on college campuses and the moral dimensions of technology design. *First Monday* (2017).
[50] Olsson Andreas Lindström, Björn. 2015. Mechanisms of social avoidance learning can explain the emergence of adaptive and arbitrary behavioral traditions in humans.. In *Journal of Experimental Psychology: General*. 688–703.
[51] Leah Findlater Lisa Anthony, YooJin Kim. 2013. Analyzing user-generated youtube videos to understand touchscreen use by people with motor impairments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1223––1232.
[52] Fade Lorne. 2021. *What Is Virtual Reality, And How Can It Be Used In The Workplace?* https://www.forbes.com/sites/forbesbusinesscouncil/2021/04/05/what-is-virtual-reality-and-how-can-it-be-used-in-the-workplace/?sh=1905b7b27a1e
[53] Divine Maloney, Guo Freeman, and Andrew Robb. 2020. It is complicated: Interacting with children in social virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 343–347.
[54] Divine Maloney, Guo Freeman, and Andrew Robb. 2020. A Virtual Space for All: Exploring Children's Experience in Social Virtual Reality. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 472–483.
[55] Divine Maloney, Guo Freeman, and Andrew Robb. 2021. Social virtual reality: ethical considerations and future directions for an emerging research space. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 271–277.
[56] Divine Maloney, Guo Freeman, and Andrew Robb. 2021. Stay Connected in An Immersive World: Why Teenagers Engage in Social Virtual Reality. In *Interaction Design and Children*. 69–79.

[57] Divine Maloney, Guo Freeman, and Donghee Yvette Wohn. 2020. " Talking without a Voice" Understanding Non-verbal Communication in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.
[58] Divine Maloney, Samaneh Zamanifard, and Guo Freeman. 2020. Anonymity vs. familiarity: Self-disclosure and privacy in social virtual reality. In *26th ACM Symposium on Virtual Reality Software and Technology*. 1–9.
[59] 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*. 1–12.
[60] Joshua McVeigh-Schultz, Elena Márquez Segura, Nick Merrill, and Katherine Isbister. 2018. What's It Mean to" Be Social" in VR? Mapping the Social VR Design Ecology. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*. 289–294.
[61] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.
[62] Jessica Outlaw and Beth Duckles. 2018. Virtual harassment: The social experience of 600+ regular virtual reality (VR) users. *The Extended Mind Blog* 4 (2018).
[63] Jeni Paay, Jesper Kjeldskov, and Mikael B Skov. 2015. Connecting in the kitchen: an empirical study of physical interactions while cooking together at home. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 276–287.
[64] Shane L Rogers, Rebecca Broadbent, Jemma Brown, Allan Fraser, and Craig P Speelman. 2022. Realistic motion avatars are the future for social interaction in virtual reality. (2022).
[65] Cynthia Breazeal Safinah Ali, Hae Won Park. 2020. Can Children Emulate a Robotic Non-Player Character's Figural Creativity?. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 499–509.
[66] Aliaksei Severyn, Alessandro Moschitti, Olga Uryupina, Barbara Plank, and Katja Filippova. 2016. Multi-lingual opinion mining on YouTube. *Information Processing & Management* 52, 1 (2016), 46–60.
[67] Aliaksei Severyn, Olga Uryupina, Barbara Plank, Alessandro Moschitti, and Katja Filippova. 2014. Opinion mining on YouTube. (2014).
[68] Navya N Sharan, Alexander Toet, Tina Mioch, Omar Niamut, and Jan BF van Erp. 2021. The Relative Importance of Social Cues in Immersive Mediated Communication. In *International Conference on Human Interaction and Emerging Technologies*. Springer, 491–498.
[69] Frenkel Sheera and Browning Kellen. 2021. *The Metaverse's Dark Side: Here Come Harassment and Assaults.* https://www.nytimes.com/2021/12/30/technology/metaverse-harassment-assaults.html
[70] Ketaki Shriram and Raz Schwartz. 2017. All are welcome: Using VR ethnography to explore harassment behavior in immersive social virtual reality. In *2017 IEEE Virtual Reality (VR)*. IEEE, 225–226.
[71] Katherine G. McKim Scott McCrickard Shuo Niu, Cat Mai. 2021. TeamTrees: Investigating How YouTubers Participate in a Social Media Campaign. In *Proceedings of the ACM on Human-Computer Interaction*. 1–26.
[72] Skarredghost. 2022. *Altspace VR to introduce safety measures, Horizon audience is growing, and more!* https://skarredghost.com/2022/02/21/altspace-meta-horizon-safety
[73] John Maynard Smith, David Harper, et al. 2003. *Animal signals*. Oxford University Press.
[74] 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 *CHI Conference on Human Factors in Computing Systems*. 1–17.
[75] Mike Thelwall. 2018. Social media analytics for YouTube comments: Potential and limitations. *International Journal of Social Research Methodology* 21, 3 (2018), 303–316.
[76] Jeffrey Bryan Theresa Jean Tanenbaum, BNazely Hartoonian. 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*. 1–13.
[77] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 247–267.
[78] H. Mai Tan Tran Minh, H. N. Pham-Nguyen and N. Xuan Long. 2019. Traffic Congestion Estimation Based on Crowd-Sourced Data. In *2019 International Conference on Advanced Computing and Applications (ACOMP)*. 119–126.
[79] Gill Valentine and Sarah L Holloway. 2002. Cyberkids? Exploring Children's identities and social networks in On-line and Off-line worlds. *Annals of the association of American Geographers* 92, 2 (2002), 302–319.
[80] Travis J Wiltshire, Emilio J Lobato, Jonathan Velez, Florian Jentsch, and Stephen M Fiore. 2014. An interdisciplinary taxonomy of social cues and signals in the service of engineering robotic social intelligence. In *Unmanned Systems Technology XVI*, Vol. 9084. SPIE, 124–138.
[81] Rachel Young, Stephanie Miles, and Saleem Alhabash. 2018. Attacks by Anons: A content analysis of aggressive posts, victim responses, and bystander interventions on a social media site. *Social Media+ Society* 4, 1 (2018), 2056305118762444.

[82] Samaneh Zamanifard and Guo Freeman. 2019. " The Togetherness that We Crave" Experiencing Social VR in Long Distance Relationships. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 438–442.

[83] Douglas Zytko, Ryan Handley, Bert Guerra, and Rukkmini Goli. 2022. A Taxonomy of Social VR Design. *arXiv preprint arXiv:2201.02253* (2022).

A APPENDIX

The following figures explain the major safety features in section 6.1

Intimacy Proxemic

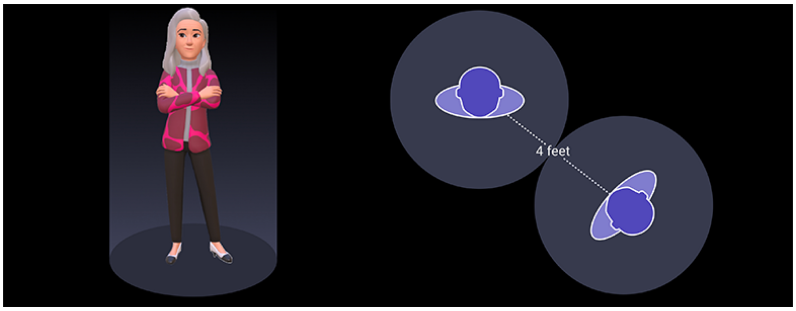

Fig. 9. Horizon Worlds released the personal boundary feature with a default setting making users feel like an almost 4-foot distance between their avatar and others. Similarly, Rec Room has an adjustable "personal space bubble," an invisible space that allows users to control how close other players interact with their avatars from close, medium, to large. (Retrieved from: https://arstechnica.com/gaming/2022/02/meta-establishes-four-foot-personal-boundary-to-deter-vr-groping/)

Trust Reputation

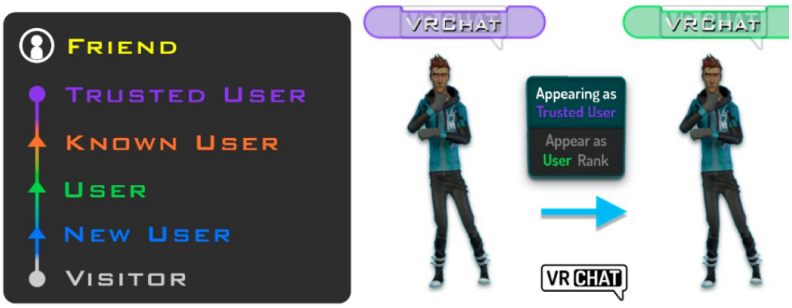

Fig. 10. When playing VRchat, users' trust feeds into a "trust rank" with several levels, i.e., trusted user, known user, user, new user, visitor, and nuisance.Retrieved from: https://docs.vrchat.com/docs/vrchat-safety-and-trust-system

Social Spaces

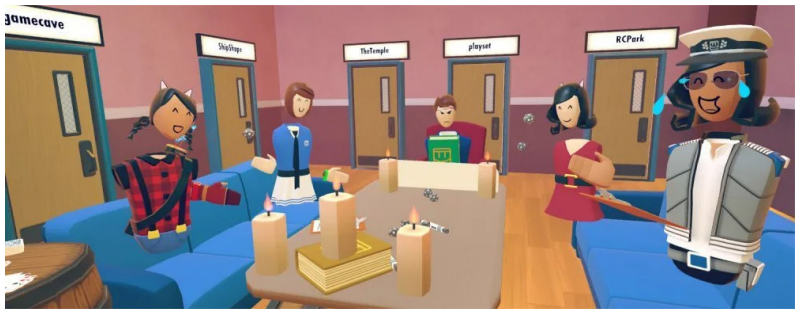

Fig. 11. Example is Rec Room's dorm room and user-created club. The dorm room is a completely private space. (Retrieved from: https://venturebeat.com/2017/12/16/rec-room-social-vr-app-update-adds-clubhouses/)

Interaction Shield

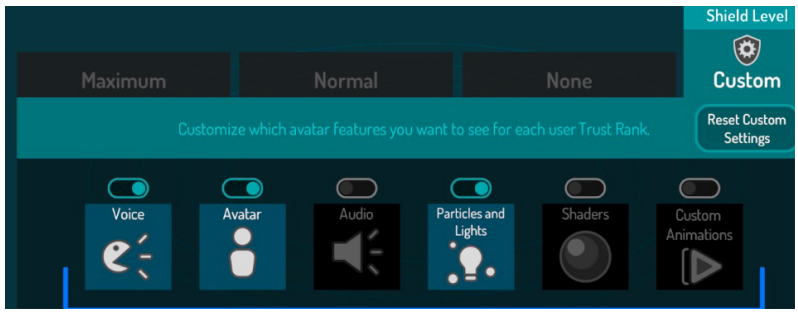

Fig. 12. Example is VRChat's trust and safety system, which can keep users safe from nuisance by adjusting microphones, screen-space shaders, loud sounds, or visually noisy or malicious particle effects. (Retrieved from: https://steamcommunity.com/sharedfiles/filedetails/?id=2421089516)

Voice gibberish

Fig. 13. When use Voice Mode to garble the speech of those around a user, an icon will appear above user's display name to indicate that he/she can't hear what strangers are saying. But if user like to temporarily unmute them they can do so by raising his/her in-game hand to their ear using a motion controller. Meta's Voice Mode feature shows it alongside a separate toggle that lets user mute non-friends entirely. (Retrieved from: https://www.theverge.com/2022/6/14/23167136/meta-horizon-worlds-garbled-voice-mode-harassment-vr-virtual-reality)

Avatar Setting

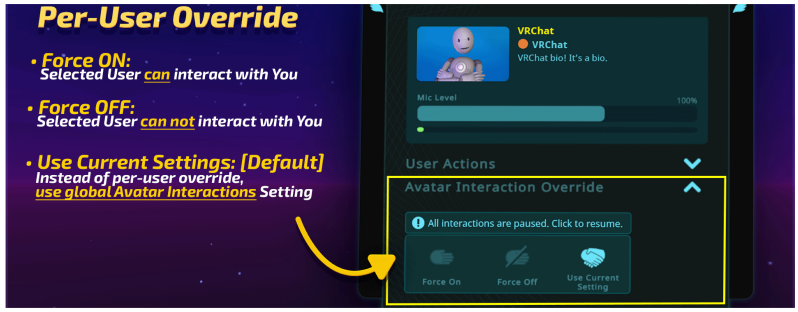

Fig. 14. Users can adjust who can and cannot interact with their avatars. For someone to be able to interact with an avatar, both parties must explicitly allow interactions with each other. (Retrieved from: https://hello.vrchat.com/blog/avatar-dynamics-live)

User Demographics

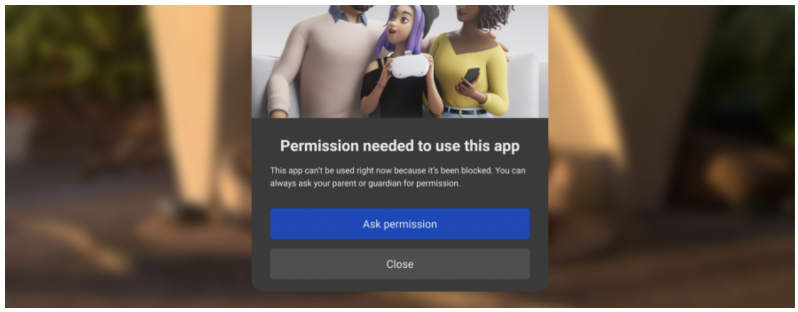

Fig. 15. Permission from parents is required for children to use the app. (Retrieved from: https://www.gamingdeputy.com/se/android/meta-to-arm-vr-headset-med-verktyg-for-foraldraovervakning/)

Safety Reactions

Fig. 16. The user utilizes the stop gesture to mute another user who is speaking. (Retrieved from: https://www.youtube.com/watch?v=T5lw58uruCwt=22s)

Safety Zone

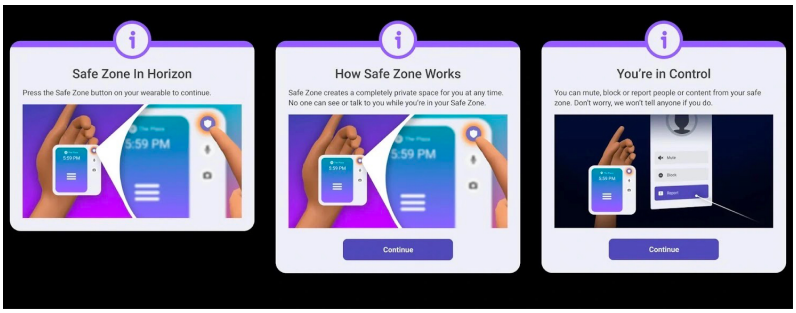

Fig. 17. Horizon World offers a one-touch button called "Safe Zone" that can quickly move users out of trouble. (Retrieved from: https://www.technologyreview.com/2021/12/16/1042516/the-metaverse-has-a-groping-problem/)

Vote Kick

Fig. 18. Users can navigate to the player's profile through the character menu and issue a vote to kick out this player.(Retrieved from: https://www.researchgate.net/figure/Safety-Options-Available-on-VRChat-Social-Quick-Menu)

Safety Report

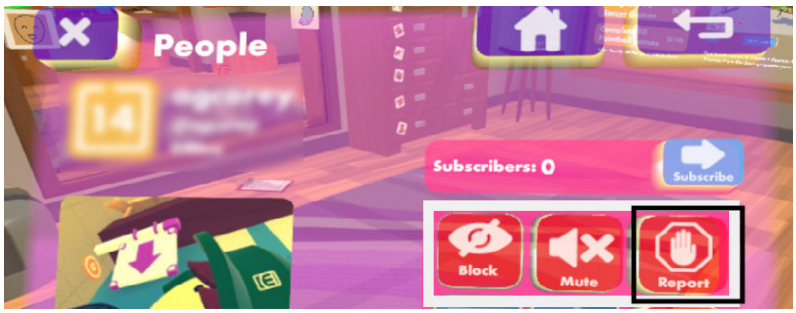

Fig. 19. Users can report users who violate the code of conduct.(Retrieved from: https://i.imgur.com/DL1PUn3.jpg)

Code of Conduct

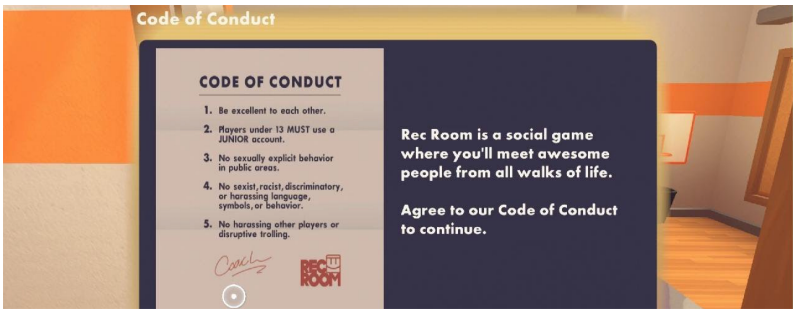

Fig. 20. Ask the user to complete a tutorial on the code of conduct in the virtual environment and re-present it to the user after he/she has violated it. (Retrieved from: https://www.youtube.com/watch?v=wYC45QyHyNU)

Informed Consent

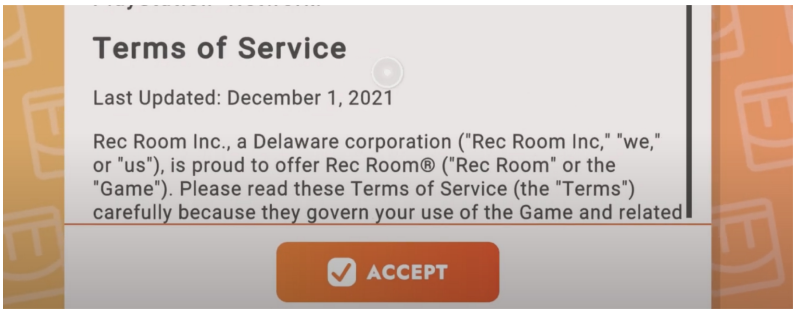

Fig. 21. Informed consent required to be accepted by the user prior to using access to the application. (Retrieved from https://www.youtube.com/watch?v=FRSccOVC09Y)

Received July 2022; revised October 2022; accepted January 2023