Alt Text and Alt Sense in VR: Engaging Screen Reader Users within the Metaverse Through Multisenses

Yu Sun
yu.sun@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

Carolin Stellmacher
cstellma@uni-bremen.de
University of Bremen
Bremen, Germany

Annika Kaltenhauser
annika.kaltenhauser@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

Nadine Wagener
nwagener@uni-bremen.de
University of Bremen
Bremen, Germany

Daniel Neumann
daniel.neumann@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

Johannes Schöning
johannes.schoening@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

Figure 1: Vision of an inclusive, social scene in the Metaverse

ABSTRACT

Alt text plays a crucial role in providing individuals who rely on screen readers with a textual representation of visual content. However, dynamic and computer-generated virtual worlds comprised of scenes, 3D objects, and avatars in virtual reality (VR) remain mostly inaccessible. Equally immersive experiences are harder to guarantee, as conventional alt text may not provide rich enough descriptions for screen reader users. Additionally, engaging with virtual worlds only through the audio channel poses risks of auditory overload. To overcome the limitations of conventional alt text and reduce the experience gap between screen reader users and non-screen reader users in VR, we propose multisensory alt attributes. In this position paper, we explore the novel design space of using alt text (auditory channel) and alt sense (haptic channel) to convey the rich visual appearance of the VR world and consider the communication of social cues, and the support of navigation.

CHI '23 Workshop: Towards an Inclusive and Accessible Metaverse, April 23–28, 2023, Hamburg, Germany

.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Accessibility**; **Virtual reality**.

KEYWORDS

Human Computer Interaction; Metaverse; Accessibility; VR, AR

ACM Reference Format:
Yu Sun, Carolin Stellmacher, Annika Kaltenhauser, Nadine Wagener, Daniel Neumann, and Johannes Schöning. Alt Text and Alt Sense in VR: Engaging Screen Reader Users within the Metaverse Through Multisenses. In *CHI '23 Workshop: Towards an Inclusive and Accessible Metaverse.* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION & RELATED WORK

While using web pages, users with limited or no vision often rely on screen readers to engage with the content via their auditory sense. The assistive technology renders visual elements of a web page, such as text, images, or buttons, into speech. Thereby, screen readers rely on additional information about graphic components,

so-called *alt text*, which is often limited to a few words or a short phrase. Although alt text can be automatically generated by some AI tools, such as ALTTEXT.AI[1], it is mostly manually provided by designers who include images in the prototype and then implemented by developers [2]. Extensive research has investigated the usability of alt text in 2D web pages, for example defining the design space of visual content for screen reader users [11], comparing alt text authoring tools [9], investigating users needs for image descriptions [13], exploring alt text for memes [4] and gifs [3].

Yet, while alt text is widely adopted in and researched for 2D web pages, HCI research regarding the 3D space of virtual reality (VR) is underrepresented. VR is often thought of as a primarily visual medium, which is increasingly becoming popular due to concepts such as the Metaverse, in which networked virtual worlds are populated by computer-generated objects as well as avatars of other users. First approaches for inclusivity heavily rely on audio to convey information about the graphic components [1, 14]. While auditory feedback for alt text is crucial for VR users with visual impairments, communicating suitable information through touch feedback could help distribute the information load onto two sensory channels, auditory and haptic. Previous work from Equal Entry incorporated vibrotactile feedback with sounds to render virtual objects accessible in an inclusive virtual environment (VE) [2]. To find and identify objects, users can scan the VE by moving the VR controller through space. Upon a collision of the controller's ray cast with a virtual object, users receive vibrotactile feedback and an audio alt description.

In this position paper, we aim to identify how different forms of multisensory alt attributes can improve the accessibility of the Metaverse for users who are visually challenged. This opens up a new design space to utilise not only the auditory channel based on the *alt text*, but also the haptic channel based on a novel attribute that we call *alt sense*. The goal of our alt text and alt sense is to i) provide access to visual appearances, behaviours, and interactions of 3D objects, avatars, and scenes, ii) convey social cues from avatars facilitating social interactions, and iii) disclose the spatial layout within the VE to support navigation. Inspired by the previously mentioned idea from Equal Entry allowing users to scan the VE and find elements within their surroundings, we propose various novel concepts on how alt text and alt sense could improve the experience of these elements. Thereby, we contribute by reflecting on how VR should be designed to support alt text and alt sense for Metaverse users who are visually challenged. Our proposal involves the use of comprehensive alt text descriptions to create virtual experiences, that are inclusive of users with visual impairments or no remaining vision. Alt text and alt sense immerse users through storytelling and explorative vibrotactile feedback to support the identification of objects and their locations. In the next sections, we first describe current alt text technologies. Then, we introduce the concept of multisensory alt attributes in the metaverse, conceptualising our ideas regarding i) 3D objects, ii) social cues, and iii) spatial layout.

2 NON-VR ALT TEXT TECHNOLOGY

Typically, a screen reader checks an image on the web for alt attributes of the <img> tag and reads it aloud. However, ensuring accessibility becomes more challenging with more embedded rich media on the web, such as an interactive map or other third-party widgets. An inline frame (iframe) element is commonly used when embedding third-party widgets, but it does not support an alt attribute. Web accessibility guidelines[3] recommend including a title attribute to describe the purpose of the content. In some cases, such as an interactive map, an extra paragraph should be added on the web to describe the landmark, the location or directions. A-frame[4], is an open-source web framework based on HTML for building WebVR, which includes specific tags for 3D contents (scene and objects), such as <a-scene>, <a-sky>, <a-plane>, <a-box>, <a-cylinder>. While these markups have semantic meaning and structure the 3D world, these A-frame tags do not provide cues about which objects are important for storytelling, which ones are just for decoration, and if they can be interacted with. Therefore, we propose that these A-tags should be enhanced with alt attributes to introduce richer alt descriptions. Zooming out from web technologies, some other software systems do a great job in the auditory description, for example, VoiceOver[5] is used to describe people, objects, text, and graphs in great detail. It utilizes computer vision technology to recognize the image, generate descriptive text, and identify common objects like icons and buttons. The traditional idea of alt text as content-description should not be restricted to the context of web technologies. Instead, we envision that the meaning of the term alt text should be broadened and re-interpreted to also encompass VEs that are mainly built with game engines such as Unity and Unreal Engine.

3 CONCEPT FOR MULTISENSORY ALT ATTRIBUTES IN THE METAVERSE

Multisensory alt attributes, *alt text* and *alt sense*, consider the auditory and haptic sense in the design of assistive technology to convey visual appearances in VR. While alt descriptions should offer users with varying degrees of visual impairments immersive storytelling of the virtual world, vibrotactile feedback can free up auditory processing, and increase accessibility by communicating through the haptic channel (e.g., different intensities of vibration could support users in assessing relative distances and locations).

As current commercial applications such as Metaverse are highly social settings, we illustrate our ideas on a social scene (see Figure 1): Our scenery portrays social interactions within an envisioned inclusive Metaverse. A virtual world with trees is the backdrop for several groups of individuals with diverse avatar appearances and behaviours, such as people with visual impairments, prosthesis users, people playing the guitars, drinking coffee, talking, or taking selfies. In the subsequent sections, we will expound on our design concepts for alt text and alt sense about this scenario, encompassing alt attributes for objects and avatars, social cues, and navigation through this scene.

[1] https://alttext.ai/
[2] https://equalentry.com/interacting-with-objects-in-an-inclusive-virtual-environment/
[3] http://web-accessibility.carnegiemuseums.org/content/maps/
[4] https://aframe.io
[5] https://www.apple.com/accessibility/vision/

Table 1: Alt attributes to render visual presentations of objects, avatars, and scenes accessible to users with vision impairments in VR. Descriptions can be provided in advance by designers, developers, or users or automatically generated on the go through the use of recognition software.

	Alt Text/Alt Sense	Information to Consider	Examples
Object	Visual properties	Size, shape, colour, etc.	"A tree with round orange and red leaves, 2m high."
	Behaviours	Animation, movement, effects, etc.	"The leaves are ruffling in the wind and some are dropping to the ground."
	Interactions	Grasping, lifting, opening or closing parts, haptics, etc.	"Upon touching the tree with your controller, it will further shed its leaves."
	Relationship w/ users	Ownership, meaningfulness, traces of use, etc.	"John Doe had previously carved a heart into the bark."
	Semantic relationship	Does it belong to a parent object or group of identical objects?	"The tree is part of a forest shaping a clearing."
	Spatial relationship	Distance, angle, height, etc.	"This tree is the closest to you, the others are in a circle around you, approximately 2m away from you."
	Accessible vibrotactile feedback	Type of object, type of interaction, relative distance, etc.	Vibration pattern indicates a portal interaction or relative distance
Scene	Setting	Surrounding, general description, etc.	"You are in a meadow."
	Weather	Weather conditions	"It is sunny with a light breeze."
	Season	Spring, summer, autumn, winter	"It is late summer."
	Light Conditions	Intensity, colour	"It is the golden hour for sunset."
	Mood	peaceful, threatening, exciting, etc.	"It is a cosy scene."
	Action	Main happening within the scene	"Many people are sitting around a campfire."
Avatar	Identity	Gender, cis or trans, age, disability or neurodivergence, race or ethnicity, skin colour	"I am a curious person in my thirties."
	Appearance	Hairstyle, height, clothes	"My black hair falls onto my shoulders and I am wearing sunglasses and a gray shirt"
	Belongings	Personal items, digital collections such as NFT	"My rainbow-coloured handbag lies next to the campfire."
	Social Cues	Attention, facial expression, hand gestures, etc.	"John Doe turns towards you and opens their arms in a welcoming gesture."
	Spatial relationship	Distance, angle, height, etc.	"An avatar is approaching diagonally from your right."
	Accessible vibrotactile feedback	Type of object, type of interaction, relative distance, etc.	Vibration pattern identifies the collision object as an avatar or indicates the distance

3.1 Multisensory Alt Attributes for Scenes, 3D Objects, and Avatars

Virtual objects, avatars, and scenes are crucial in creating immersive and convincing VEs [12]. They can direct the user's attention and impact users' feelings [6]. While size, shape, or colour convey the visual look, animations and movements mediate how objects or avatars behave, embedding them into the living virtual scene and contributing to a convincing VR simulation. In the context of social interactions, personal objects of avatars, such as a guitar, may disclose a person's personality trait and enhance social connectedness. Recent work even showed that previous interactions with objects, so-called *traces of use*, can enhance the social presence of users in VR [5]. Therefore, knowing the history of objects and how they were used by others can be crucial for social interactions. A comprehensive alt description should, therefore, rather tell a story and consider diverse aspects to provide a rich and convincing experience and immerse users with limited or no eyesight in VR. An overview of the attributes that we envision should have an alt text or alt sense description is provided in Table 1 and is explained in more detail in the following sections.

Setting the Scene. Engaging in storytelling in a 3D world is like watching a theatre show. An audio description should tell the audience the feel of the stage (e.g. the tone of the location as threatening or joyful) or the action of what is happening in the scene, which is essential to understand the story, visual look or design of the production. Explaining the scene upon entering the Metaverse can be crucial for rendering it equally realistic and setting the mood and emotions, despite previous findings for conventional image alt text to avoid decorative elements [16]. The weather and light information could be cued and even influence the storytelling. To avoid being overwhelmed, not all visible objects need to be described, but rather specific scene-related attributes. Besides, mentioning things that are already recognisable based on the sounds they are producing should also be avoided, such as birds twittering or the ripple of a nearby water stream.

Enriching the Experience of 3D Objects. To render virtual objects accessible through multisensory alt attributes, we propose alt texts to describe 1) the visual properties of an object (e.g. size or colour), 2) the behaviour, animations, movement, or other effects (e.g. the steaming effect of coffee), 3) the available interaction (e.g. opening

or closing parts of the object), and 4) the relationship with other users such as ownership or traces of use. Additionally, we propose to add the 5) alt sense attribute "accessible vibrotactile feedback" that conveys visual information through touch. Here, we extend on the previously mentioned idea of receiving vibrations upon a ray-cast collision when scanning the VE with the controller by different types of haptic feedback. Particular vibration patterns could highlight specific types of interactions (e.g. portals), or specific types of elements (e.g. avatars or nature assets).

As comprehensive descriptions and haptic feedback can be overwhelming in a complex VE, we envision different strategies to filter and tailor the information to the user's needs. Objects forming a parent object (e.g., stones, flames, and wood logs depicting a campfire) can be grouped and described by one standard alt text or alt sense. An additional alt attribute could be, thus, 6) the semantic relationship with other objects and scenes. Further, prioritisation can highlight only crucial objects that are at the centre of the experience. We envision that a book held by the conversation partner that symbolises the hobby of that person could be prioritised to provide a starting point for the conversation. We propose to apply the same to objects that are in the vicinity of avatars, such as flowers, that could be used to enhance the social connection by being given them to the conversation partner.

Meeting Different Avatars and Their Identity. In a highly social virtual world like the Metaverse, users who are visually challenged should be able to recognise other avatars to engage in social activities. Alt text descriptions of avatars could be provided by Metaverse users themselves, who then can decide on how they want to express themselves to other users through this medium. Research from Zhang et al. [17] investigated the preferences of showing disabilities in Avatars and found some users want to disclose their disabilities to increase their presence in the virtual world. Therefore, describing various avatars can not only show diverse self-presentations but also help screen reader users recognize people and provide starting points for a conversation. As a notable example, Vocaleyes developed guidance of self-description for inclusive meetings [15] and outlined possible visual information that could be included in a self-description for people with visual impairments: 1) the identity (e.g. gender, cis/trans, age, disability/neurodivergence, race/ethnicity, skin colour), 2) appearance (e.g. height, hairstyle), 3) belongings (e.g. clothes, accessories). Selective description and prioritization should be applied here as well. For instance, while the description of a person's shirt may not be inherently interesting, it could be a unique or eye-catching garment that could potentially become a source of humour during social interactions. Building upon our previous idea that a particular type of vibrotactile pattern could highlight specific interactions, we envision other avatars being equipped with an alt sense attribute defining a pattern that identifies them as other users. To extend the avatar alt attributes with visual dynamic social cues, we discuss possible approaches in section 3.2.

3.2 Multisensory Alt Attributes for Social Cues

To engage in social interactions, people with visual impairments need information [7], such as 1) knowing how many people are present and where they are located relative to oneself, 2) where a person's attention is directed, 3) identities of people, 4) appearances of people, 5) changes of appearance, 6) facial expressions, or 7) hand gestures and body motions. Hence, we think it is crucial to include social cues in the alt attributes, such as if people are approaching, if they are making eye contact, and if they want to initiate a conversation. In this regard, we were inspired by Mai et al. [10], who designed animal-sense inspired systems where users can sense the world via vibration on their feet or EMS on the skin, which could be incorporated into the alt sense. For example, users can feel the magnet field bending if one is approaching, mimicking the electroreception of eels. We can integrate this design into wearables, controllers, or canes (see [18]) as well. Especially during navigation or social interaction, it will reduce the overload of audio feedback and increase the social presence by using haptic sensation to receive information on people that are within one's contact range and if their body language reveals that they are approachable.

3.3 Multisensory Alt Attributes for Spatial Layout and Navigation

Understanding the own position and one's relation to other objects is crucial for the orientation within the VE. Therefore, alt attributes should also be available for the dimensions and relative positions within the landscape of the scene. Previous work from Zhao et al. [18] developed a cane that provides haptic feedback for people with visual impairment to navigate in a VE, which could render the alt sense. In the context of currently available consumer VR controllers, we envision that increasing vibration intensity indicates that the user gets closer to an object. Hence, objects or avatars that are far off are then only noticeable by a subtle signal, closer objects create a more substantial effect. Based on the proximity, we could also set filters to exclude objects triggering vibrations that are far away, so that users can only sense their immediate surroundings.

To offer alt text descriptions of obstacles or nearby objects during navigation, an assistant such as DeepNAVI [8] could be used. The system could actively assist the user by presenting the current position, distance to the point of interest, and movement guidance (turn left/right/back, stop, move forward). Additionally, the system could harness alt texts and alt senses of obstacles or nearby objects to inform the users of their immediate surroundings.

4 CONCLUSION & FUTURE VISION

It is challenging for users to rely on screen readers to actively engage in social VEs. Alt text descriptions of scenes, objects and avatars are necessary to build an inclusive Metaverse. However, given the richness VR, it is crucial to filter the information stream and reduce the cognitive load. We propose that a new alt attribute *alt sense* should be added to shift information to the haptic sense and empower users in identifying avatars through a specific vibration pattern or support navigation by different intensities of feedback. To mitigate the extensive work load for creating multisensory alt attributes, and to be tailored to individual's preference, AI-based systems can create a personalised metaverse based on user's social behaviour and interests. Through a combination of alt text and alt sense, we can design engaging and joyful Metaverse which is accessible for all kinds of users, specifically including users with limited or no eyesight.

REFERENCES

[1] Eelke Folmer, Bei Yuan, Dave Carr, and Manjari Sapre. 2009. TextSL: a command-based virtual world interface for the visually impaired. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility (Assets '09)*. Association for Computing Machinery, New York, NY, USA, 59–66. https://doi.org/10.1145/1639642.1639654

[2] Caitlin Geier. 2018. How to Design Great Alt Text: An Introduction. https://www.deque.com/blog/great-alt-text-introduction/ Accessed on Feb 18, 2023.

[3] Cole Gleason, Amy Pavel, Himalini Gururaj, Kris Kitani, and Jeffrey Bigham. 2020. Making GIFs Accessible. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 24, 10 pages. https://doi.org/10.1145/3373625.3417027

[4] Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B. Chilton, and Jeffrey P. Bigham. 2019. Making Memes Accessible. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 367–376. https://doi.org/10.1145/3308561.3353792

[5] Linda Hirsch, Anna Haller, Andreas Butz, and Ceenu George. 2022. "What a Mess!": Traces of Use to Increase Asynchronous Social Presence in Shared Virtual Environments. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 598–599. https://doi.org/10.1109/VRW55335.2022.00150

[6] Hunter G Hoffman. 2021. Interacting with virtual objects via embodied avatar hands reduces pain intensity and diverts attention. *Scientific Reports* 11, 1 (2021), 1–13.

[7] Sreekar Krishna, Dirk Colbry, John Black, Vineeth Balasubramanian, and Sethuraman Panchanathan. 2008. A Systematic Requirements Analysis and Development of an Assistive Device to Enhance the Social Interaction of People Who are Blind or Visually Impaired. In *Workshop on Computer Vision Applications for the Visually Impaired*. James Coughlan and Roberto Manduchi, Marseille, France. https://hal.inria.fr/inria-00325432

[8] Bineeth Kuriakose, Raju Shrestha, and Frode Eika Sandnes. 2023. DeepNAVI: A deep learning based smartphone navigation assistant for people with visual impairments. *Expert Systems with Applications* 212 (2023), 118720. https://doi.org/10.1016/j.eswa.2022.118720

[9] Kelly Mack, Edward Cutrell, Bongshin Lee, and Meredith Ringel Morris. 2021. Designing Tools for High-Quality Alt Text Authoring. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 23, 14 pages. https://doi.org/10.1145/3441852.3471207

[10] Christian Mai, Mariam Hassib, and Ceenu George. 2017. Like Elephants Do: Sensing Bystanders During HMD Usage. In *Proceedings of the CHI 2017 Workshop on Amplification and Augmentation of Human Perception, May 07, 2017, Denver, CO, USA. Copyright is held by the owner/author(s).* (Denver, CO, USA) *(CHI '17)*. ACM.

[11] Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. 2018. Rich Representations of Visual Content for Screen Reader Users. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3173633

[12] Rebecca A Penn and Michael C Hout. 2018. Making reality virtual: How VR "Tricks" your brain. *Frontiers for Young Minds* 6 (2018).

[13] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. " Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[14] Shari Trewin, Vicki L. Hanson, Mark R. Laff, and Anna Cavender. 2008. PowerUp: An Accessible Virtual World. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility* (Halifax, Nova Scotia, Canada) *(Assets '08)*. Association for Computing Machinery, New York, NY, USA, 177–184. https://doi.org/10.1145/1414471.1414504

[15] VocalEyes. 2021. Self-description for inclusive meetings. https://vocaleyes.co.uk/services/resources/digital-accessibility-and-inclusion/self-description-for-inclusive-meetings/ Accessed on Feb 18, 2023.

[16] W3C. 2023. Decorative Images. https://www.w3.org/WAI/tutorials/images/decorative/ Accessed on Feb 18, 2023.

[17] Kexin Zhang, Elmira Deldari, Zhicong Lu, Yaxing Yao, and Yuhang Zhao. 2022. "It's Just Part of Me:" Understanding Avatar Diversity and Self-presentation of People with Disabilities in Social Virtual Reality. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–16.

[18] Yuhang Zhao, Cynthia L. Bennett, Hrvoje Benko, Edward Cutrell, Christian Holz, Meredith Ringel Morris, and Mike Sinclair. 2018. Enabling People with Visual Impairments to Navigate Virtual Reality with a Haptic and Auditory Cane Simulation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173690