There's a Large Hippo Resting in the Mud

Virtual Video Safaris for Blind and Visually Impaired People

Oliver Bendel and Doris Jovic

Abstract *Inclusive AI is a young field of research as well as an application domain where AI systems of various kinds are used to support individuals with disabilities and impairments, enabling their participation in different aspects of life. Generative AI plays an important role in this context, particularly in the form of multimodal large language models (MLLMs). These models have also found their way into Animal-Computer Interaction (ACI), for instance in apps that provide behavioral guidance to users encountering animals. In "The VISUAL Project" at the FHNW School of Business, both disciplines were combined to unlock new potential. The resulting prototype allows blind and visually impaired individuals to embark on virtual video safaris. It integrates publicly available wildlife webcams from various regions and uses an MLLM to analyze and evaluate the live footage. Different profiles and modes are available to meet various needs. A text-to-speech system delivers descriptions and explanations. The prototype demonstrates the technical feasibility and potential of Inclusive AI and ACI. However, limitations must be considered, such as those associated with commercial MLLMs and occasionally inaccurate descriptions. Further research and development are required to make the experience truly inclusive and autonomous.*

Keywords Generative AI, MLLM, Animal-Computer Interaction, Inclusive AI

1 Introduction

Inclusive AI aims, on the one hand, to address exclusionary phenomena associated with artificial intelligence (AI), such as bias, hallucinations, hate speech, and deepfakes, and, on the other hand, to promote inclusive applications that provide support for affected individuals (Bendel 2025). Generative AI, which saw widespread adoption beginning in 2022, plays a major role in this context. The app Be My Eyes, featuring the Be My AI function – based on a Multimodal Large Language Model (MLLM) – has supported blind and visually impaired individuals since 2023 in perceiving and assessing their environments. Its opportunities and risks were analyzed by researchers at the FHNW School of Business (Bendel 2024b). In 2024, researchers at UC Davis restored the voice of a patient who had lost it due to ALS by using a brain-computer interface and a text-to-speech system. The voice was cloned using AI based on earlier recordings (Yehya 2024). Inclusive AI may represent a significant advancement, especially for individuals with disabilities. Additional opportunities are seen in the preservation of endangered languages and the support of threatened minorities.

Generative AI offers potential in numerous application areas, including Animal-Computer Interaction (ACI) (Mancini 2011; Mancini and Nannoni 2023). In "The Animal Whisperer Project" from 2024 at the FHNW School of Business, three prototype applications were developed to analyze and evaluate

animal body language (Bendel and Zbinden 2024). Similar to the Be My AI function, these apps utilize an MLLM. The target user group was not explicitly defined. In addition to sighted users, individuals with visual impairments could potentially benefit as well. Descriptions and explanations were provided not only as text output but also via speech output. Users received helpful guidance for interacting with the animal they encountered – whether during a hike or a walk through the city.

Building on these experiences, the first author initiated a project at the FHNW School of Business in February 2025 titled VISUAL. The acronym stands for "Virtual Inclusive Safaris for Unique Adventures and Learning". In direct user communication, it is rendered in the singular: "Your Virtual Inclusive Safari …" The second author joined the project as a project employee, starting in early April, and also based a thesis on her work within the project. VISUAL aims to combine Inclusive AI and ACI through the use of an MLLM to create added value. Specifically, following a literature review and an online survey, a prototype was developed that enables blind and visually impaired individuals to go on virtual video safaris around the world without the need for physical travel. The images are sourced from publicly available wildlife webcams and are analyzed and evaluated with the help of an MLLM. A text-to-speech system delivers descriptions and explanations – just as it did in "The Animal Whisperer Project". The first development phase of VISUAL concluded on August 8, 2025, with a fully functional prototype.

In Section 2, the two authors define key terms and establish the conceptual foundation. Section 3 presents the project implementation, focusing in particular on the system's modules and the prompt engineering process, followed by a description of system usage. Section 4 provides a brief ethical and social discussion. The paper concludes with Section 5, which offers a summary and outlook.

2 Basics of the Project

This section focuses on multimodal large language models, wildlife webcams, and access to animals for blind and visually impaired individuals. Additionally, a brief literature review introduces related projects. Although Animal-Computer Interaction is not examined in detail, the project can be situated within this field as well as within the domain of Inclusive AI. The animals do not interact directly with the VISUAL system, but the project aligns with the goals of ACI as formulated in Mancini's manifesto, which emphasizes fostering new forms of communication and relations between humans and non-human animals (Mancini 2011). By enabling blind and visually impaired individuals to access wildlife experiences, VISUAL helps cultivate empathy, awareness, and a deeper understanding of animals and their behavior. In this way, the project contributes to ACI by broadening the range of human-animal encounters and promoting inclusive forms of interspecies connection.

2.1 Multimodal Large Language Models

A large language model (LLM) is capable of "understanding" and generating language in general. It uses vast amounts of data to learn billions or even trillions of parameters during training. Multimodal large language models can process and generate not only text, but also images and audio (Bendel 2024a, b). All of these models fall under the category of generative AI – a group of artificial intelligence (AI) applications designed to generate content. One persistent, unresolved issue, especially with text generators, is so-called hallucination: the unintended output of false information. For the system developed in the VISUAL project, it is essential that images – including image sequences – can be reliably analyzed, interpreted, and used to draw meaningful conclusions. For blind and visually impaired individuals, speech output is of particular importance.

To adapt a publicly available, pre-trained MLLM, three main techniques are typically used. The first is prompt engineering (Bendel 2024a), in which the developer inserts instructions that the system is expected to follow. These instructions can be highly complex and may relate to the system's role and tasks. The second technique is retrieval-augmented generation (RAG), which involves building or accessing an additional knowledge base that the LLM or MLLM can draw upon. This may involve uploading individual documents, creating a database, or integrating external sources such as Wikipedia, as was done in this project. This turns the system into a domain specialist and helps reduce hallucinations. A third technique is fine-tuning, which involves training the language model with domain-specific data. This method is resource-intensive and error-prone and was therefore not considered suitable for the low-budget VISUAL project.

In the context of VISUAL, MLLMs are expected to detect and describe animal presence and certain observable behaviors at a coarse level, such as feeding, grooming, or group movement. However, more subtle or complex ethological patterns are unlikely to be captured reliably. Rather than replacing expert observation, VISUAL leverages these capabilities as an accessible mediator that provides blind and visually impaired users with meaningful approximations of animal activity.

2.2 Wildlife Webcams

Webcams, as a form of telepresence media, have been a popular tool since the 1990s for gaining insight into remote regions or observing real-time conditions at specific locations. People often use them to check the weather in a certain area or to view exotic landscapes and cities. Thousands of publicly accessible webcams exist around the world (Jacobs et al. 2009). Older webcams from the 1990s typically offer low resolution, while newer ones from the 2020s often provide HD quality. A significant number of webcams have also been developed for wildlife observation. Platforms such as Explore.org (https://explore.org/livecams) stream webcam footage from a wide variety of global regions – also referred to here as environments or ecosystems. These appeared ideal for integration into the VISUAL prototype.

Public wildlife webcams allow users to observe distant locations and the animal life occurring there in real time. Kamphof (2011) explains that, despite the actual physical distance, webcams create a sense of "being there". This real-time connection enables not only virtual access but also an emotional connection between users and the events they observe. Moreover, the webcams strike a balance between their function as surveillance tools and their capacity to foster emotional engagement and attachment (Jacobs et al. 2009).

The webcams of Explore.org showcase ecosystems on various continents, including both northern and southern regions, as well as lowland and highland areas with diverse climates. They focus on wildlife in the air, on land, and in water. A key challenge with birds is that they are difficult to capture while in flight. Sometimes, webcams are aimed at nests, where activity is limited to specific times of year or day. On land and in water, however, more stable conditions can be achieved – provided the location is chosen carefully. Such webcams were deemed well-suited for integration into the VISUAL prototype.

2.3 Access to Animals for Blind and Visually Impaired People

Blind individuals can have widely varying life experiences. Some are blind from birth, while others lose their vision later in life. This results in significant differences in how animals are perceived. In addition to blind individuals, there are also people with visual impairments. Some are nearly blind, while others

have substantial limitations but can still, for example, recognize shapes or distinguish between dark and light areas. The World Health Organization (WHO) classifies visual impairments into five levels (WHO 2022). The VISUAL project did not differentiate among these categories but focused on the extreme case of individuals who are blind from birth, while the system may also benefit people with other forms of visual impairment.

Kim et al. (2019) compared adults who were congenitally blind or had lost their sight early in life with sighted adults in terms of their perception of animals. The study found that blind individuals were generally less familiar with animals. In tasks related to size, shape, skin texture, and color, there were many similarities between the two groups. For example, both groups judged elephants to be larger than rhinoceroses and giraffes to be taller than lions. They also agreed in distinguishing between aquatic and terrestrial animals. However, differences emerged in the level of detail: sighted individuals tended to differentiate animals by types of fur, whereas blind individuals used broader features and categories such as habitat or species.

A system like VISUAL, which is designed for blind and visually impaired users, can take such differences into account through prompt engineering and retrieval-augmented generation. Since most LLMs and MLLMs are trained primarily on internet-based data, they tend to reflect the habits and needs of the majority. This likely includes a bias toward visual conventions familiar to sighted individuals rather than the lived experiences of blind users. Prompt engineering and RAG can help address typical perceptual differences and compensate where knowledge gaps or misinterpretations exist. One promising approach is to describe scenes in a narrative style with a high level of accuracy.

2.4 Related Projects

Several projects have introduced and discussed the possibilities of Inclusive AI. The tests conducted with Be My Eyes and its Be My AI feature have already been mentioned (Bendel 2024b). A 2025 study by researchers at Stanford University explores ways to consider learning differences in the development of innovative systems, with a particular focus on individuals with disabilities and impairments (McGee et al. 2025).

In other projects, artificial intelligence has been used to monitor wildlife in zoos and in the wild – tracking position, behavior, and health (Congdon et al. 2022). In this context, facial recognition has been proposed as a method to track individual animals (Mazurkiewicz 2024; Bendel and Yürekkirmaz 2023). Emotion recognition in pets is also an area of ongoing research. A research team developed a system that simulates animal breathing to convey a sense of life during online safari tours – an aspect often missing in digital experiences (Tsuji et al. 2023). Projects such as “BBC Springwatch” utilize machine learning techniques for animal detection in wildlife feeds (BBC 2020). However, these systems do not provide narrative speech output.

Zhang et al. (2025), in their study “I Can See Forever!”, evaluated a system that uses video-based LLMs in real time to assist blind individuals with everyday tasks – for example, in domestic settings or social interactions. However, the system was not specifically applied to animal videos or webcam feeds. Chen et al. (2025) introduced a multimodal vision-language model designed to help blind users understand their environments through scene descriptions, often via wearable devices. This system focuses on real-world environments rather than animal webcams or bundles of still images. Other studies have concentrated on image-to-text systems developed for everyday scenes encountered by blind users (Karamolegkou et al. 2025), but these also lack a focus on animal webcams or curated image sets. The

Accessible Aquarium used video feeds of fish and aquariums to generate a mix of narration and sonification or music to allow blind individuals to experience a live aquarium (Pendse et al. 2008).

To the authors' knowledge, no existing project combines Inclusive AI and Animal-Computer Interaction through the capabilities of generative AI – specifically LLMs or MLLMs – in the way that VISUAL does. It should be noted that generative AI saw its major public breakthrough in 2022, although it had been anticipated for years, for instance as an extension of social robotics (Coursey 2020). Inclusive AI is likewise a young field of research and development, and there are still only a few projects that specifically address the needs of blind and visually impaired individuals. ACI has been under investigation for several years, but its research community remains small. A noticeable shift toward generative AI in ACI has only occurred since around 2024.

3 Implementation of the Project

This section describes the project preparation and the development of the VISUAL system. It also outlines the online survey conducted with blind and visually impaired individuals, including key findings.

3.1 Initial Considerations

At the heart of the VISUAL project is the implementation of an innovative system. The idea emerged from technological advancements in generative AI, progress in Inclusive AI and ACI, and the recognition that individuals with impairments are often excluded from experiences that are taken for granted by others. The goal of the VISUAL prototype is to enable blind or visually impaired individuals to go on virtual video safaris (Jovic 2025). Although live webcams offer fascinating visual insights into the appearance and behavior of animals, these experiences remain largely inaccessible to users who rely on non-visual forms of access. In addition, many existing digital offerings fail to meet the specific needs of blind or visually impaired individuals.

VISUAL aims to overcome these barriers by combining live streams from wildlife webcams with narrative audio descriptions (Jovic 2025). These descriptions are designed not only to convey the current scene but also to provide contextual and educational information that can be adapted to the user's age and preferences. In this way, blind and visually impaired individuals receive as much information as possible to help them build a mental picture of the scene and its background. At the same time, the system is designed to support users where gaps or distortions in perception exist due to their visual impairment.

The research question was formulated as follows: "How can a multimodal large language model enable blind and visually impaired individuals to experience a virtual safari utilizing public wildlife webcams, interpreting their pictures, and translating the content into descriptive audio narratives?" (Jovic 2025). As is customary at the FHNW School of Business, and particularly within the research area of the first author, the research question was intended to be addressed not only theoretically but also through a practical implementation – in this case, a prototype. The prototype was expected to include all essential functionalities and be fully operational from a technical standpoint. A public release as an online version was not planned at this stage; however, all files and code are available for download and reproduction (https://github.com/jovicyy/VISUAL_2025/).

3.2 User Analysis

The accessibility needs identified through a literature review during the project were translated into concrete design principles for the VISUAL prototype (Jovic 2025). The structure follows the WCAG POUR principles: Perceivable, Operable, Understandable, and Robust. These principles guided the development process:

- Perceivable: Blind and visually impaired users rely heavily on rich, descriptive audio descriptions. In the VISUAL prototype, this is implemented through contextual audio narrations of the webcam livestreams, with a focus on describing the actual scene rather than using interpretive storytelling.
- Operable: Visually impaired users often face frustration with inaccessible navigation and mouse-dependent controls. VISUAL addresses this by ensuring that all functions are fully keyboard-accessible and that all interface elements are arranged in a logical reading order – optimized for screen readers.
- Understandable: Research has shown that inconsistent layouts and unclear structures complicate the use of digital content. To counter this, VISUAL employs predictable navigation patterns, a clear and consistent structure, and plain language in both text and audio. Additionally, a dedicated section explains how to navigate the system and what users can expect.
- Robust: Accessibility suffers when content is not compatible with assistive technologies or various devices. VISUAL ensures that all user interface elements are properly labeled so that their functions can be accurately recognized and used by assistive technologies – especially screen readers.

Certain functions were tailored to the nature of the application. For instance, after successful navigation, users hear a lion's roar (*Free Lion Roar Sound Effects Download* from Pixabay, https://pixabay.com). Visual elements were intentionally retained to support mixed-use scenarios, allowing blind or visually impaired individuals to use the application together with sighted assistants who benefit from conventional visual interfaces.

3.3 Online Survey

The online survey for the VISUAL system was designed to be fully accessible (Jovic 2025). A staff member at the FHNW School of Business who is severely visually impaired and relies on such systems in his daily work as a scheduling coordinator provided valuable assistance in this regard. The goal of the survey was to gather insights from blind and visually impaired individuals in order to better understand their needs, preferences, and expectations. Participants were informed that their input would significantly contribute to the development of a prototype aimed at making virtual video safaris more accessible and enjoyable. A total of 22 questions were asked, divided into six thematic sections.

The survey, available in both German and English, was promoted in June 2025 via LinkedIn and on various blogs. Most blindness organizations in German-speaking countries declined to support the initiative, with the exception of the Schweizerischer Blindenbund (Swiss Federation of the Blind), which shared the survey link on LinkedIn and its website. Eleven individuals participated. While this number may appear modest, it reflects the challenges of reaching the target group without access to relevant networks or institutional support and is nevertheless valuable for exploratory research.

The following summarizes the questions and responses most relevant to the system's design, along with key takeaways (Jovic 2025). One of the central questions, "If you were listening to an audio description of a wildlife scene, what information would be most important to you?", revealed that participants placed

particular value on knowing which animals were present and what behaviors they were exhibiting – both were mentioned by nine respondents. The environment was noted by eight participants, while six emphasized the appearance of the animals, and five highlighted the atmosphere or mood of the scene. The conclusion was that audio descriptions should name animal species and behaviors clearly and also include environmental context, physical features, and the emotional tone of the scene.

Another question, “How important is it to choose the webcam region?”, showed that most respondents considered regional selection to be important, with two rating it as very important, six as fairly important, and three as neutral. As a result, the ability to choose the region was implemented as a central, easily accessible feature. By contrast, the question “How important is it to choose the animal type?” received fewer strong endorsements, with only one participant rating it as very important, three as fairly important, and six as neutral. Accordingly, animal selection was treated as a secondary option, while the main focus was placed on habitat selection.

Finally, the question “How beneficial would it be to have the option to choose the tone or style of the audio descriptions to suit your preferences?” revealed that this feature was considered useful by the majority of participants, with five rating it as extremely beneficial, four as very beneficial, one as neutral, and one as not beneficial. In response to this, the VISUAL system incorporated a tone selection feature through the introduction of three distinct modes, allowing users to adapt the narrative style – such as calm, factual, or storytelling – according to their personal preferences.

3.4 System Requirements

The system requirements for the VISUAL prototype were divided into two categories: functional and non-functional (Jovic 2025). Functional requirements define the actions the system must perform to meet users’ needs. These were derived from a literature review and team meetings.

The system processes live webcam streams featuring animals in various regions and environments, including savannas, jungles, forests, mountains, polar regions, and oceans. It uses the multimodal large language model GPT-4o to generate scene descriptions, which are then converted into spoken language by a text-to-speech system. Users have control over the audio playback, with the ability to start, stop, repeat, and adjust the playback speed. The system allows users to switch between different regions and cameras. Content can be adapted to the user’s age, offering separate settings for children and adults. In addition, one can choose between three narrative modes: *Adventurer* (e.g., *Safari Adventurer*), *Field Scientist*, or *Calm Observer*. The system also provides educational content by retrieving relevant factual information from Wikipedia using RAG. Accessibility is ensured through full keyboard operability, screen reader compatibility, and intuitive navigation.

Non-functional requirements describe how the system should operate. VISUAL adheres fully to WCAG guidelines, using semantically correct HTML with appropriate labels and ARIA roles, and ensures a high-contrast interface. External APIs are used in a controlled and secure manner to guarantee data protection and reliability.

Although no formal evaluation of different MLLMs was conducted, tests with the Google Cloud Vision API and GPT-4o showed that GPT-4o performed significantly better in this context, recognizing animals even when barely visible and more accurately estimating their numbers in tight groups or herds. Furthermore, GPT-4o had already been used successfully in previous projects at the university, which made it a logical and effective choice for this implementation. GPT-5 was not released until the project

was completed. Developing a domain-specific MLLM from scratch was not feasible due to the resources required.

3.5 Conceptual Architecture

The VISUAL prototype is based on a multi-layered architecture that defines how the system components interact to transform webcam livestreams into accessible audio descriptions (Jovic 2025).

- The process begins at the frontend level, where users personalize their experience by selecting a preferred ecosystem. Based on this selection, a corresponding wildlife webcam is automatically chosen. Users then specify their *Age* profile (*Child* or *Adult*) and select a desired *Explorer* mode: *Adventurer*, *Field Scientist*, or *Calm Observer*. These preferences are sent to the backend as structured JSON data.
- On the backend, a headless browser session is launched using Puppeteer, a JavaScript library. From the selected livestream, three consecutive screenshots are taken at three-second intervals. These images are bundled into a temporary package and passed on to the next layer for AI processing.
- In the AI processing layer, a dynamic system prompt is generated based on the user's settings, including role and profile. This prompt is sent to OpenAI's GPT-4o API, which conducts a multimodal analysis of the image bundle and generates a contextually appropriate and descriptive narrative. The raw description is then forwarded to the next processing stage.
- In the post-processing layer, a quality filter is applied to the generated text. This includes so-called "anti-hallucination checks" to eliminate unrealistic content – such as polar bears appearing in savannas. These checks are performed entirely by the large language model itself, as retrieval-augmented generation did not yield improvements in this context. In addition, the language is refined to replace visually biased verbs such as "see" or "watch" with spatial expressions like "in front of you" or "to your right".
- The finalized narrative is returned to the frontend and converted into speech using the Web Speech API. The voice used depends on the selected *Explorer* mode and Age profile, including both female and male voice options.
- After playback, users have the option to request additional educational information about the animals detected. This is provided using a RAG mechanism, which queries Wikipedia for reliable content. The resulting information is not delivered verbatim but rather reformulated and enhanced to suit the narrative context and the target group.

This architecture places particular emphasis on accessibility and modularity, ensuring that all components – user interface, screen capture, AI analysis, narrative generation, and speech output – work together seamlessly. As a result, the system offers an intuitive, reliable, and highly customizable virtual safari experience for blind and visually impaired users.

3.6 Module Overview

This section describes the modules of the VISUAL system, specifically the *Explorer* modes and *Age* profiles, as well as the available regions (Jovic 2025). The modes were developed by the second author based on a literature review and an analysis of best practices. An example of the prompt engineering used for implementing the modes and profiles is provided in the following section.

3.6.1 Explorer Modes

The *Adventurer* mode – adapted according to the selected region (e.g., *Safari Adventurer* – "Safari" was used here instead of "Savanna" –, *Polar Adventurer, Ocean Adventurer*) – is designed to convey the feeling of being on a live safari. The narration is vivid and sequential, providing a concrete description of what is currently visible. The goal is to enable an immersive wildlife observation experience without speculative or fictional elements. The language is expressive yet natural, delivered in direct address (e.g., "In front of you …"), with a focus on the animals' movements, appearance, and environments. This mode is especially well suited for users seeking a rich and exciting nature experience.

The *Field Scientist* mode emphasizes scientific accuracy and education. The narrative style is objective and documentary, offering clear and accessible explanations of animal behavior and biological context. The tone remains calm and informative, avoiding speculative or narrative embellishment. This mode is intended for users who are interested in natural science and wish to gain knowledge while observing.

The *Calm Observer* mode is aimed at users who seek peace and relaxation. The narration is gentle and meditative, creating a sense of safety and timelessness. Movements are described only subtly, while the environment is depicted as calming and protective. The language is intentionally soft and atmospheric, focusing on light, temperature, and silence. This mode is primarily designed to provide emotional relief and promote mindful connection with nature.

3.6.2 Age Profiles

The adult profiles make slight adjustments to the language used in the selected *Explorer* modes by incorporating a somewhat richer vocabulary and restrained contextual information, such as details about the environment or atmosphere. The tone remains factual and natural – always in line with the chosen mode – and conveys a balanced, mature perspective without oversimplifying or overexplaining.

The children's profiles are designed to ensure that the narratives are easy to understand, safe, and engaging. The language is simple, rhythmic, and vivid, with a deliberately friendly and calming tone. Even potentially tense scenes are described in a neutral manner, avoiding distressing content. The goal is to provide a safe and positive nature experience that is easy to follow and supported by age-appropriate comparisons.

3.6.3 Regions

Each region includes its own set of carefully selected cameras intended to capture representative animals of the respective ecosystem. For reasons of simplicity, the majority of streams were sourced from Explore.org, with one additional camera from Africam (https://africam.com). While the cameras generally display the animal species associated with each region, other sightings are also possible. Viewers may observe terrestrial, aquatic, and airborne animals in their natural habitats or within wildlife reserves. Domesticated or farm animals are not included. Due to the limited availability of suitable options in polar regions, only a single camera was used for that area. Table 1 provides an overview of all integrated webcams, organized by region, camera source, animal species, and sample descriptions. All animal footage was reviewed by the second author, with the exception of the lion stream, where instead hyenas and vervet monkeys were observed.

Table 1: Overview of the webcams (following Jovic 2025)

Region	Camera Source	Animal Species	Behavior/Characteristics
African Savanna	Mpala Live Camera (Kenya)	African Elephant	Large herds, complex social behavior, trunk usage
African Savanna	Mpala Live Camera (Kenya)	Lion	Group behavior, hunting behavior, territorial behavior
African Savanna	Botswana Wildlife Safari Cam by Africam	Zebra	Migration patterns, herd formation, grazing behavior
African Savanna	Mpala Live Camera (Kenya)	Hippopotamus	Semi-aquatic lifestyle, territorial behavior, family groups
African Savanna	Mpala Live Camera (Kenya)	Giraffe	Feeding behavior, social interaction, movement patterns
African Savanna	Mpala Live Camera (Kenya)	Various Antelope Species	Interactions among different species (impala, gazelles, waterbuck)
Tropical Jungle	GRACE Gorilla Forest Corridor	Gorilla	Family structures, grooming, foraging
Tropical Jungle	Toucan TV	Toucan	Feeding behavior, territorial behavior, vocalizations
Tropical Jungle	GRACE Gorilla Forest Corridor	Various Primate Species	Social diversity and behavior
Temperate Forest	Brooks Falls, Katmai National Park	Brown Bear	Salmon fishing, seasonal behavior, mother-offspring interactions
Temperate Forest	International Wolf Center	Gray Wolf	Pack behavior, territorial behavior, hunting strategies
Temperate Forest	Brooks Falls, Katmai National Park	Forest Birds	Behavior of birds of prey, songbirds, waterbirds
Temperate Forest	Brooks Falls, Katmai National Park	Salmon Migration	Fish migration and its impact on the ecosystem
Polar Region	Cape East Camera, Churchill Cam	Polar Bear	Adaptation to cold, hunting behavior, play-fighting, mother-offspring behavior
Ocean	Utopia Village Reef Cam	Sharks (various species)	Feeding and territorial behavior
Ocean	Shark Cam	Sharks (various species)	Feeding and territorial behavior
Ocean	Utopia Village Reef Cam	Tropical Fish	Schooling and solitary behavior
Ocean	Utopia Village Reef Cam	Manta Rays	Swimming patterns, feeding behavior

3.7 Prototype Development and Design Process

This section explains the methods, tools, and design decisions used to implement the previously described framework into an accessible and functional prototype (Jovic 2025). It then illustrates the prompt engineering process using the *Safari Adventurer* mode as an example. Additionally, it includes a step-by-step walkthrough of how the prototype is used.

3.7.1 Design Approach and Tools Used

VISUAL was developed using several tools, including Figma for UI design (following a design-oriented low-code approach) and Cursor for AI-assisted code generation (Jovic 2025). This combination enabled the creation of a functional proof-of-concept prototype without compromising usability or accessibility.

To ensure accessibility and user flow, the Figma plugin Stark – Contrast & Accessibility Checker was employed. Once accessibility and navigation logic were verified, Figma's Dev Mode MCP Server was used. This tool was crucial for implementing the design in code, as it grants AI-based development environments such as Cursor access to structured design data – such as colors, spacing, components, and hierarchies – which are far more precise than screenshots or static images.

In practice, this meant that once the MCP Server was activated and connected to Cursor, selected design frames could be translated directly into code. An initial prompt to Cursor read: "Hey, please create a clickable prototype of a 'visual safari' for blind and visually impaired users using semantic HTML and Tailwind CSS. Strictly follow the layout of the homepage. The page should not be scrollable – everything must fit on the screen exactly as in my Figma file. Other pages will follow later, so please create only the homepage for now. Thank you!" By combining structured design data with precise instructions, Cursor was able to generate functional and accessible components.

The backend of VISUAL was implemented using Node.js, which serves as the central control system for all core processes. All essential APIs – such as Puppeteer, GPT-4o, Wikipedia, and the Web Speech API – are integrated into this environment. The decision to use Cursor was based on three main reasons:

- Efficiency: Development time was significantly reduced, as Figma designs could be quickly and reliably transferred into a functioning frontend and backend.
- Accessibility testing: Continuous evaluation and refinement of accessibility features were integrated directly into the workflow.
- Focus on functionality: Since much of the infrastructure could be generated automatically, development could focus on critical aspects such as narrative quality, API compatibility, and overall usability.

However, the process also demonstrated that AI-assisted coding does not equal automated coding. Achieving good results required precise prompts, technical expertise, and targeted research. For example, a vague prompt such as "Please create a function to take screenshots from the embedded YouTube API" was insufficient. Instead, a far more specific prompt was needed, such as: "Write a function using puppeteer-core that opens a YouTube video URL in headless mode (Microsoft Edge), starts the video, and takes three screenshots at three-second intervals. Save the screenshots as .png files in the folder 'safari-screenshots'." This approach demanded both technical understanding and the ability to integrate various sources while guiding the AI with precision.

The integration of Figma (including the MCP Server) and Cursor made it possible to bridge the gap between conceptual design and functional prototype. This approach not only accelerated the development process but also ensured consistent adherence to inclusive design principles throughout.

Table 2: Example Prompt (*Safari Adventurer*) (following Jovic 2025)

Category	**Instruction**
Task ADVENTURER MODE	Provide an engaging and adventurous narration of a wildlife scene for blind or visually impaired visitors.
Action ADVENTURER MODE	You are an experienced safari guide narrating wildlife scenes for blind visitors. Your top priority: Describe the animals first – count them, identify their species, and describe their posture, physical features, and visible actions in concrete detail. Avoid vague or speculative language. After that, briefly describe the environment. Use accessible language (never "see/look/watch/observe"). Do not invent or narrate sounds. You may include light sensory impressions like light, texture, or atmosphere.
Goal ADVENTURER MODE	Write 7–9 short, natural sentences, focusing on the animals' actions as if explaining to a friend on safari.
Accessibility	Never use these words: see, look, watch, observe, notice, appear, visible, sight, view, glance, gaze, peer, glimpse, spot, witness. Use instead: present, there is/are, positioned, located, in front of you, nearby, to your left/right, resting, moving, situated, found, detected.
Detection protocol	Examine the image carefully – are any animals clearly identifiable? Animals present: If YES: Count them, identify the species, and describe their physical features, posture, movements, and interactions in detail. No animals: If NO: Provide a simple, brief description of the environment (plants, ground, light). Only describe what is visually certain. Never guess or invent details.
Task ADULT OVERLAY	Deliver wildlife narration that feels natural, conversational, and mature for an adult listener.
Action ADULT OVERLAY	Use clear, precise animal details with direct and factual language. Avoid unnecessary embellishment or overly emotional tones. Maintain a natural flow, as if explaining the scene to a friend.
Goal ADULT OVERLAY	Produce an accessible, well-structured narration that informs and engages an adult audience without oversimplifying content.
Region	You're narrating wildlife from the African savanna with grasslands and acacia trees.
Critical accuracy enhancement	NEVER assume animals are present just because the habitat suggests they should be there.

Category	Instruction
	Look for specific animal indicators: eyes, ears, tails, legs, fur patterns, movement. If an object is motionless and unclear, treat it as landscape/vegetation. Count only animals you can distinguish as separate individuals with certainty. NEVER identify vegetation, rocks, shadows, or tree formations as animals. When in doubt, describe the environment rather than guessing about animals.
Additional instructions	Don't say "in the first image" or "in the second image". Treat multiple images as one scene.

3.7.2 Example of Prompt Engineering

Extensive prompt engineering was carried out to define the system's roles and profiles. Due to space limitations, not all prompts can be presented in this paper. However, one representative example is provided for the *Safari Adventurer* mode in combination with the *Adult* profile. To clearly illustrate the structure, this contribution uses a tabular format (Table 2). An excerpt from the output for a gorilla scene reads: "In front of you, a small group of gorillas sits closely together on the jungle floor. Two of them are pressed side by side. One leans forward, using its fingers to part and pick carefully through the thick fur on the other's head and shoulders. The other gorilla stays still, letting the grooming continue. Around them, the ground is littered with scattered leaves and dense jungle vegetation." The description of a hippopotamus can be found in Figure 4. It begins with the words "In front of you, there's a large hippo resting in the mud ...".

3.7.3 Prototype Walkthrough

The VISUAL prototype consists of several interconnected pages, each designed with a focus on simple navigation and accessibility (Jovic 2025). The following section presents each screen, explaining its layout, purpose, functions, and design considerations. As previously mentioned, each successful navigation action is accompanied by a lion's roar, confirming that the process has succeeded and that the user is moving forward.

Welcome Page

The welcome page serves as the entry point to VISUAL and introduces users to the concept of an inclusive virtual safari. It features a clear and minimalist layout, accompanied by an image of a roaring lion. This image was created by the second author using ChatGPT/4o Image and is accompanied by a descriptive alt text: "A majestic lion with a golden mane, mouth wide open in a powerful roar."

The page includes the following key elements: 1. A concise headline welcomes the user and is followed by a brief description. The headings are structured hierarchically using different levels so that screen readers can read them in a logical order. In the code, "Welcome to VISUAL" is marked as an H1, the tagline as H2, and the description as H3. This semantic hierarchy enables users to navigate between headings more easily and helps distinguish text content from interactive elements. 2. Buttons are located

in the lower part of the page and are styled with high-contrast colors – among other things, to support colorblind users. These include: “How it works” (this button opens a quick start guide intended for first-time users), “Explore by Animal” (this button enables direct selection of a desired animal species), and “Start Exploring” (this button begins the actual virtual safari and leads to the regional selection page).

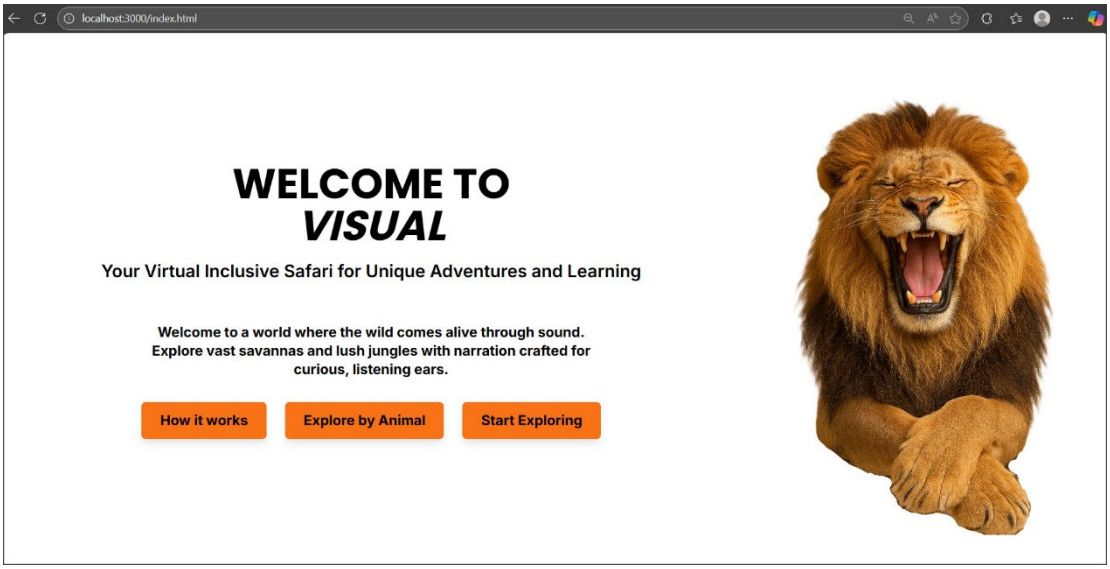

Figure 1: Welcome page

“How it works” Page – Quick Start Guide

The “How it works” page serves as a quick-start guide for new users, explaining the core functionalities of VISUAL and offering an overview of what to expect and how to navigate the system. It provides a step-by-step introduction to using VISUAL, including navigating the homepage, selecting a region or a specific animal species, and personalizing the narration through *Age* profiles (*Adult* or *Child*) and three different narrative styles or modes (*Adventurer*, *Field Scientist*, *Calm Observer*).

In addition, the page outlines the available controls for the narration, such as adjusting playback speed, pausing, and replaying the audio. It also explains further safari-related features, including the option to retrieve factual information about detected animals, switch between cameras, or move to a different region.

From a structural perspective, the page follows accessibility guidelines, with correct labeling of all texts and interface elements. Unlike the other pages, this screen uses a scrollable layout in order to provide a comprehensive guide without overcrowding the interface. A “Home” button is included in the design, allowing users to easily return to the welcome page, where the safari experience begins. The placement of this button remains consistent across all subsequent pages (except the welcome page itself), contributing to a coherent and user-friendly navigation system. This design decision was made to support clarity and consistency throughout the user experience.

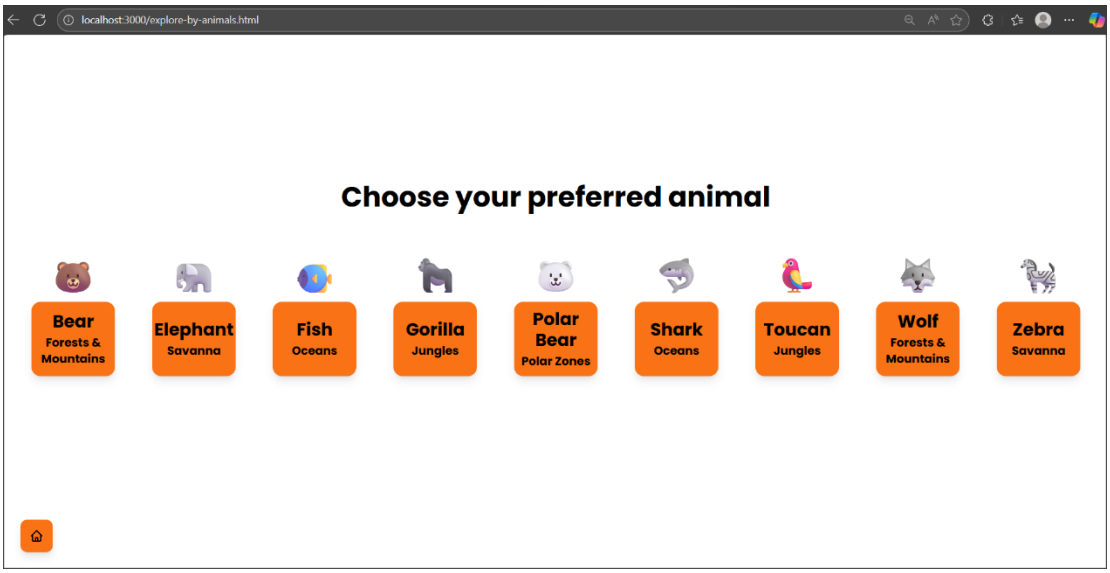

Figure 2: Explore by Animal

"Explore by Animal" Page – Species-Focused Exploration

The "Explore by Animal" page allows users to navigate directly to the habitat of a selected animal species (Figure 2). Both the name of the species and the associated region or ecosystem – such as "Zebra" in the "Savanna" – are displayed, providing users with contextual information even before making a selection. This page is designed for those who prefer a species-specific experience over regional exploration. The embedded livestreams were chosen to ensure a high likelihood of observing the selected species; however, their actual presence at any given time cannot be guaranteed.

"Start Exploring" Page (Choose by Region) – Ecosystem-Based Exploration

The "Start Exploring" page (Choose by Region) marks the entry point into the core virtual safari experience (Figure 3). It presents five distinct regions or ecosystems: savanna, jungle, forests and mountains, polar zones, and oceans. Each region is visually represented by a large decorative emoji symbol, which is intentionally hidden from screen readers to avoid redundant information and to ensure that only relevant content is read aloud. This measure also contributes to smoother and more efficient navigation. The buttons themselves are designed to be fully screen reader–friendly and lead users directly to the selected region upon activation.

"Guided Safari" Page – The Core Experience

The "Guided Safari" page forms the core of the VISUAL prototype (Figure 4). Each ecosystem is represented by its own dedicated subpage, prefixed with "guided-ecosystem", such as "guided-jungle" for the jungle. This separation into individual pages allows for the targeted integration of ecosystem-specific information and corresponding embedded livestreams. Additionally, each page is color-coded to match the visual character of its habitat, while maintaining high contrast to ensure accessibility.

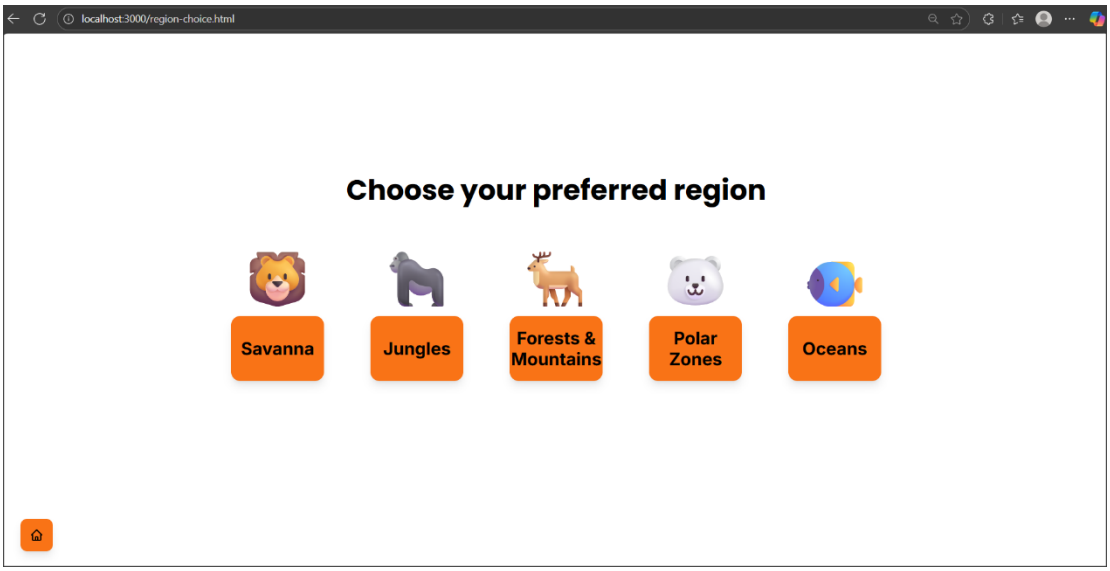

Figure 3: Choose by Region

The page integrates a wildlife webcam livestream from Explore.org specific to the selected ecosystem. A customization panel is available, allowing users to choose their preferred profile and narration mode. By default, the *Safari Adventurer* mode is preselected. The page offers various options for controlling the narration, including adjusting playback speed, starting, pausing, and replaying the audio. Users can also request optional factual information, switch between cameras, or change regions – all without having to leave the page. All functions are accessible via clearly labeled and intuitive buttons.

On the right side, there is a placeholder labeled “Safari Narrative Transcript”, where the generated narration appears once it is successfully created. Since the audio playback is handled via the Web Speech API, this field is intentionally hidden from screen readers to prevent redundant reading of the same content.

Since capturing screenshots and analyzing them takes a certain amount of time, users are informed of the current system status through spoken prompts. The first audio message, triggered after clicking “Start Narration”, announces: “Preparing your safari”. Once the screenshots have been successfully captured, the system says: “Analyzing animals”. The final result is then displayed in the previously mentioned transcript window.

At the top of the page, a server connection indicator is displayed. Like the analysis panel, this element was initially implemented for internal monitoring during development. When the connection is active, it shows “Server: Online” with a green background and a glowing green dot.

In addition to the main narration, users can click the “Get Animal Facts” button to receive supplementary knowledge-based content. This triggers a retrieval-augmented generation process that automatically sources factual information from Wikipedia. The output is adapted to match the previously selected *Age* profile.

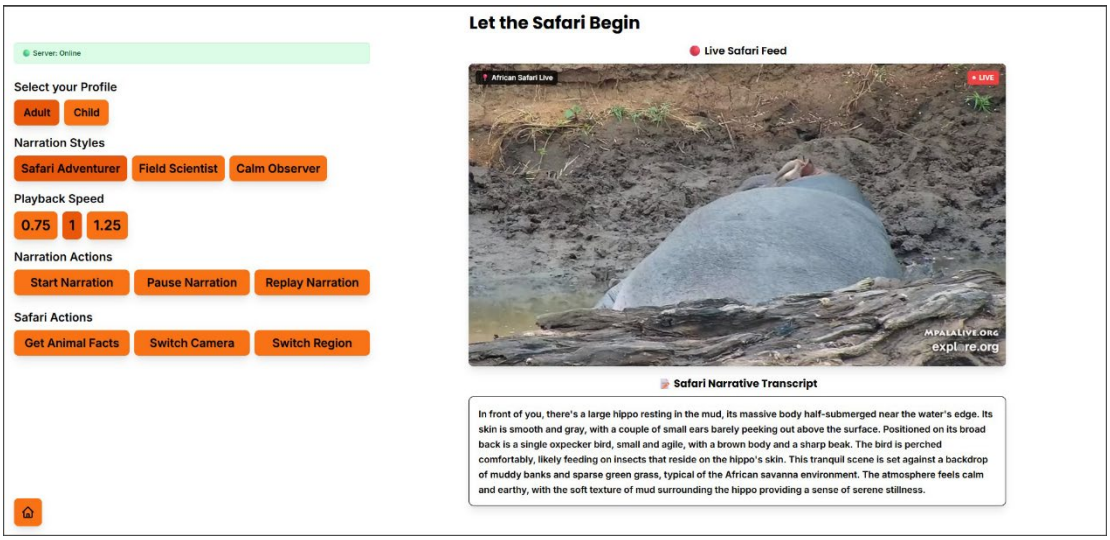

Figure 4: Guided Safari

3.7.4 External Tests

In early August 2025, two external tests were conducted with visually impaired individuals to complement the internal testing and demonstrate the full functionality of the VISUAL system (Jovic 2025). Due to time and resource constraints, no further testing was possible at this stage. The goal of the team is not to develop a commercial product. The first author's research consistently focuses on prototypes, with the technical foundations and source code made freely available. All tests proceeded without any notable issues.

4 Ethical and Social Discussion

This section offers a brief ethical and social reflection on the VISUAL system. Due to space constraints, not all relevant aspects can be included.

- Inclusion and Accessibility: VISUAL has the potential to enhance the inclusion of blind and visually impaired individuals by providing greater independence and access to information. This supports more equal participation in social life. It creates opportunities that previously did not exist – namely, individualized, professional tours that could otherwise only be offered by specially trained human assistants who would manually access wildlife webcams and describe the scenes in real time.
- Privacy and Intimacy: Although wildlife webcams are primarily directed at animal habitats, they may occasionally capture people, potentially violating their privacy or intimacy (Bendel and Zbinden 2024). A distinction must be made between authorized individuals (such as park rangers) and unauthorized ones (such as poachers). Wildlife webcams and MLLMs could, in principle, help uncover illegal activity like poaching, adding a potential societal benefit to what might otherwise be viewed purely as a privacy risk. Whether animals themselves are affected in their privacy or intimacy is a matter of ongoing debate within the ACI community (Paci et al. 2022).
- Accuracy and Reliability: The reliability of AI in describing, interpreting, and evaluating visual input is of critical importance (Bendel and Zbinden 2024; Bendel 2024b). Inaccurate or misleading

information could result in an unsatisfying user experience and foster misconceptions about the appearance or behavior of animals, thus undermining one of the system's core goals. Internal tests conducted by the second author between June and July 2025 revealed no major flaws, though it should be noted that she does not possess specialized knowledge in zoology or animal behavior.

- Manipulation and Control: Dependence on a single provider or developer – especially OpenAI in this case – raises concerns (Bendel 2024b). Through its descriptions and evaluations, the provider can influence or manipulate the user's perception of the environment. The worldview and moral framework embedded in the model shape what is described and what is excluded. American tech companies are repeatedly criticized for imposing restrictions or censorship, particularly concerning topics like sexuality. Whether this extends to depictions of animal sexuality could not be determined.
- Digital Divide: There is a risk of a digital divide, as not all blind and visually impaired individuals have equal access to the latest technologies or reliable internet connections (Bendel 2024b). Although the application itself is free, not everyone can afford a tablet or laptop with sufficient connectivity. Nonetheless, a virtual video safari of this kind is significantly less expensive than traveling to the location in person.

In addition to these ethical considerations – most of which relate to technology and information ethics – further questions arise. While blind and visually impaired individuals can now go on a video safari from their own living room, they could also travel alone or with assistance to the relevant countries and encounter animals in real life, not only virtually but also through smell and touch. Kamphof (2011) emphasized the feeling of "being there", yet aspects of perception and experience remain missing – both for impaired and non-impaired users. Although researchers have developed an innovative system to enhance the experience during online safari park tours, nothing truly compares to being there in person (Tsuji et al. 2023). It seems important, therefore, to understand VISUAL not as a replacement but as a supplement to real-world experiences.

An additional ethical consideration concerns the fact that the system was designed primarily by sighted developers. This inevitably introduces biases linked to visual conventions that may not align with the lived experiences of blind users. To counteract this, the team sought advice from a blind collaborator and conducted an online survey to gather the perspectives and needs of blind individuals. While prompt engineering and RAG were also used to mitigate potential biases, deeper and continuous involvement of blind developers would further enhance inclusivity.

5 Summary and Outlook

This paper presented the VISUAL project, in which a prototype was developed to enable blind and visually impaired individuals to embark on virtual video safaris. The following research question was addressed: "How can a multimodal large language model enable blind and visually impaired individuals to experience a virtual safari utilizing public wildlife webcams, interpreting their pictures, and translating the content into descriptive audio narratives?" The project delivered theoretical insights but, more importantly, a fully functioning prototype.

The VISUAL prototype demonstrates the technical feasibility and potential of combining Inclusive AI with Animal-Computer Interaction. However, certain limitations must be taken into account, such as the reliance on commercial MLLMs or occasional inaccuracies in descriptions. The system could benefit

from the inclusion of zoological and biological expertise, as well as contributions from ethologists and other specialists in wildlife behavior and ecology. Methods such as RAG, which were only partially applied in this project, could further enhance factual accuracy. Another possible improvement is the use of new MLLMs, such as GPT-5, released in early August 2025. Finally, making the system publicly accessible would allow testing with a larger user base and engaging potential partner organizations.

6 Acknowledgments

The authors would like to thank Artan Llugaxhija from the FHNW School of Business, who, as a visually impaired individual, contributed significantly by helping to make the online survey accessible, introducing them to screen reader usage, and demonstrating how laptops are used by blind or visually impaired users.

References

BBC (2020) Using AI to monitor wildlife cameras at Springwatch. https://www.bbc.com/rd/blog/2020-06-springwatch-artificial-intelligence-remote-camera

Bendel O (2025) Inclusive AI. In: Gabler Wirtschaftslexikon. Springer Gabler, Wiesbaden. https://wirtschaftslexikon.gabler.de/definition/inclusive-ai-171870

Bendel O, Zbinden N (2024) The Animal Whisperer Project: A GenAI App for Decoding Animal Body Language and Behavior. In: Proceedings of ACI2024, Glasgow University, Glasgow, UK. ACM, New York. https://dl.acm.org/doi/proceedings/10.1145/3702336

Bendel O. (2024a) 300 Keywords Generative KI. Springer Gabler, Wiesbaden

Bendel O (2024b) How Can Generative AI Enhance the Well-being of Blind? In: Proceedings of the AAAI 2024 Spring Symposium Series, Symposium "Impact of GenAI on Social and Individual Well-being", Stanford University, Stanford, California, March 25–27, 2024. The AAAI Press, Washington, DC. https://ojs.aaai.org/index.php/AAAI-SS/article/view/31232/33392

Bendel O, Yürekkirmaz A (2023) A Face Recognition System for Bears: Protection for Animals and Humans in the Alps. In: Proceedings of the Ninth International Conference on Animal-Computer Interaction (ACI'22), December 05–08, 2022, Newcastle-upon-Tyne, United Kingdom. ACM, New York. https://dl.acm.org/doi/proceedings/10.1145/3565995

Chen Z, Liu Z, Wang K et al (2025) A large vision-language model based environment perception system for visually impaired people. https://arxiv.org/abs/2504.18027

Congdon JV, Hosseini M, Gading EF, Masousi M, Franke M, MacDonald SE (2022) The future of artificial intelligence in monitoring animal identification, health, and behaviour. Animals 12(13):1711. https://doi.org/10.3390/ani12131711

Coursey K (2020) Speaking with Harmony. In: Bendel O (ed) Maschinenliebe. Springer Gabler, Wiesbaden

Jacobs N, Burgin W, Fridrich N et al (2009) The global network of outdoor webcams: Properties and applications. In: Proc ACM Int Symp Advances in Geographic Information Systems (GIS). https://doi.org/10.1145/1653771.1653789

Jovic D (2025) VISUAL: Virtual Inclusive Safaris for Unique Adventures and Learning. Bachelor Thesis. FHNW School of Business, Basel

Kamphof I (2011) Webcams to save nature: Online space as affective and ethical space. Found Sci 16:259–274. https://doi.org/10.1007/s10699-010-9194-7

Karamolegkou A, Nikandrou M, Pantazopoulos G et al (2025) Evaluating multimodal language models

as visual assistants for visually impaired users. https://arxiv.org/html/2503.22610v1

Kim JS, Elli GV, Bedny M (2019) Knowledge of animal appearance among sighted and blind adults. Proc Natl Acad Sci USA 116(23):11213–11222. https://doi.org/10.1073/pnas.1900952116

Mancini C (2011) Animal-computer interaction: a manifesto. Interactions 18(4):69–73. https://dl.acm.org/doi/10.1145/1978822.1978836

Mancini C, Nannoni E (2023) Editorial: Animal-computer interaction and beyond: The benefits of animal-centered research and design. Front Vet Sci 9:1109994. https://doi.org/10.3389/fvets.2022.1109994

Marino L, Allen K (2017) The psychology of cows. Anim Behav Cogn 4(4):474–498. https://doi.org/10.26451/abc.04.04.06.2017

Marks M, Jin Q, Sturman O et al (2022) Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments. Nat Mach Intell 4:331–340. https://doi.org/10.1038/s42256-022-00477-5

Mazurkiewicz J (2024) Artificial intelligence methods for pet emotions recognition. In: Zamojski W, Mazurkiewicz J, Sugier J, Walkowiak T, Kacprzyk J (eds) System Dependability – Theory and Applications. DepCoS-RELCOMEX 2024. Lecture Notes in Networks and Systems, vol 1026. Springer, Cham. https://doi.org/10.1007/978-3-031-61857-4_16

McGee NJ, Kozleski E, Lemons CJ, Hau IC (2025) AI + Learning Differences: Designing a future with no boundaries. Stanford Accelerator for Learning, Stanford

Paci P, Mancini C, Nuseibeh B (2022) The case for animal privacy in the design of technologically supported environments. Front Vet Sci 8:784794. https://doi.org/10.3389/fvets.2021.784794

Pendse A, Pate M, Walker BN (2008) The accessible aquarium: identifying and evaluating salient creature features for sonification. In: Proc 10th Int ACM SIGACCESS Conf Computers and Accessibility (Assets '08). ACM, New York, pp 297–298. https://doi.org/10.1145/1414471.1414546

Tsuji M, Takegawa Y, Matsumura K, Hirata K (2023) Investigation on enhancement of the sense of life in safari park online tours with animal breathing reproduction system. In: Proc 9th Int Conf Animal-Computer Interaction (ACI '22), Newcastle-upon-Tyne, UK, 5–8 Dec 2022. ACM, New York, Article 3, pp 1–5. https://doi.org/10.1145/3565995.3566024

WHO (2022) Blindness and vision impairment. https://who-dev5.prgsdev.com/m/news-room/fact-sheets/detail/blindness-and-visual-impairment

Yehya N (2024) New brain-computer interface allows man with ALS to 'speak' again. In: News of UC Davis Health, 14 August 2024. https://health.ucdavis.edu/news/headlines/new-brain-computer-interface-allows-man-with-als-to-speak-again/2024/08

Zhang Z, Sun Z, Zhang Z et al (2025) "I can see forever!": Evaluating real-time VideoLLMs for assisting individuals with visual impairments. https://arxiv.org/abs/2505.04488

Please cite as follows:

Bendel, Oliver; Jovic, Doris. There's a Large Hippo Resting in the Mud: Virtual Video Safaris for Blind and Visually Impaired People. In: Wiley Industry News, 10. November 2025. https://wileyindustrynews.com/en/contributions/theres-a-large-hippo-resting-in-the-mud.