Im Schlamm ruht sich ein großes Nilpferd aus

Virtuelle Videosafaris für blinde und sehbehinderte Personen

Oliver Bendel und Doris Jovic

Abstract *Inclusive AI ist ein junges Forschungsfeld und zugleich ein Anwendungsgebiet, wo KI-Systeme aller Art verwendet werden, um behinderte und beeinträchtigte Personen zu unterstützen und ihnen Teilhabe zu ermöglichen. Dabei spielt u.a. generative KI eine Rolle, etwa in Form von Multimodal Large Language Models (MLLM). Diese sind auch in die Animal-Computer Interaction eingezogen, etwa innerhalb von Apps, die Verhaltenstipps für Benutzer geben, die auf Tiere treffen. Im VISUAL-Projekt an der Hochschule für Wirtschaft FHNW wurden beide Disziplinen kombiniert, um neues Potenzial zu erschließen. Der Prototyp erlaubt es blinden und sehbehinderten Personen, auf virtuelle Videosafaris zu gehen. Es werden öffentlich verfügbare Wildtier-Webcams integriert und die Live-Bilder mit Hilfe eines MLLM analysiert und evaluiert. Für unterschiedliche Bedürfnisse sind unterschiedliche Profile und Modes verfügbar. Ein Text-to-Speech-System vermittelt die Beschreibungen und Erklärungen. Der Prototyp zeigt die technische Machbarkeit und das Potenzial für Inclusive AI und ACI. Es gilt allerdings auch Einschränkungen zu berücksichtigen, etwa mit Blick auf kommerzielle MLLMs und punktuell unzutreffende Beschreibungen, und es muss weitere Forschung und Entwicklung stattfinden, um das Erlebnis wirklich inklusiv und autonom zu machen.*

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

1 Einleitung

Inclusive AI (inklusive KI) will zum einen Phänomene der künstlichen Intelligenz (KI) mit exkludierendem Charakter wie Bias, Halluzination, Hate Speech und Deepfakes bekämpfen, zum anderen Anwendungen mit inkludierendem Charakter stärken und damit Betroffenen helfen (Bendel 2025). Eine große Rolle spielt die generative KI, die sich ab 2022 stark verbreitete. Die App Be My Eyes mit der Funktion Be My AI – die auf einem multimodalen großen Sprachmodell (Multimodal Large Language Model, MLLM) basiert – half ab 2023 blinden und sehbeeinträchtigten Personen dabei, ihre Umwelt wahrzunehmen und zu bewerten. Ihre Chancen und Risiken wurden von Forschern der Hochschule für Wirtschaft FHNW dargestellt (Bendel 2024b). Forscher der UC Davis haben 2024 mit Hilfe eines Brain-Computer Interface und eines Text-to-Speech-Systems die Stimme eines Patienten wiederhergestellt, der sie durch ALS verloren hatte. Sie war aus früheren Aufnahmen mit Hilfe von KI geklont worden (Yehya 2024). Gerade für Behinderte dürfte Inclusive AI einen wesentlichen Fortschritt darstellen. Weitere Möglichkeiten sieht man bei der Stärkung gefährdeter Sprachen und bedrohter Minderheiten.

Generative KI bietet Möglichkeiten in vielen Anwendungsbereichen, auch in der Animal-Computer Interaction (ACI) (Mancini 2011; Mancini und Nannoni 2023). In „The Animal Whisperer Project“ von 2024 an der Hochschule für Wirtschaft FHNW wurden drei Anwendungen geschaffen, mit denen man die Körpersprache von Tieren analysieren und evaluieren konnte, der Cow Whisperer, der Horse

Whisperer und der Dog Whisperer (Bendel und Zbinden 2024). Wie bei der Be-My-AI-Funktion kam hierbei ein MLLM zum Einsatz. Der Nutzerkreis der Apps war nicht spezifiziert. Neben Normalsichtigen konnten auch Menschen mit Sehbeeinträchtigung profitieren. Die Beschreibungen und Erklärungen wurden nicht nur mittels Textausgabe, sondern auch mittels Sprachausgabe vermittelt. Die Benutzer erhielten hilfreiche Empfehlungen für den Umgang mit dem Tier, dem sie – etwa auf einer Wanderung oder bei einem Spaziergang in der Stadt – begegneten.

Vor dem Hintergrund dieser Erfahrungen initiierte der Erstautor an der Hochschule für Wirtschaft FHNW im Februar 2025 ein Projekt mit dem Namen VISUAL. Das Akronym steht für „Virtual Inclusive Safaris for Unique Adventures and Learning". In der direkten Ansprache wird es in den Singular gesetzt: „Your Virtual Inclusive Safari …" Die Zweitautorin wurde als Mitarbeiterin gewonnen, die mit ihrer Arbeit am Projekt Anfang April startete und dazu auch ihre Abschlussarbeit schrieb. VISUAL verfolgt die Idee, mit Hilfe eines MLLM Inclusive AI und Animal-Computer Interaction zu verbinden, um einen Mehrwert zu schaffen. Genauer gesagt wurde nach einer Literaturanalyse und einer Onlineumfrage ein Prototyp entwickelt, mit dem blinde und sehbehinderte Personen virtuelle Videosafaris auf der ganzen Welt machen können, ohne ihr Zuhause oder ihre Umgebung verlassen zu müssen. Die Bilder werden von öffentlich verfügbaren Wildtier-Webcams bezogen und mit Hilfe eines MLLM analysiert und evaluiert. Ein Text-to-Speech-System vermittelt die Beschreibungen und Erklärungen – so wie es auch bei „The Animal Whisperer Project" möglich war. VISUAL wurde am 8. August 2025 abgeschlossen.

In Abschnitt 2 klären die beiden Autoren zunächst Grundbegriffe und legen Grundlagen. In Abschnitt 3 beschreiben sie die Umsetzung des Projekts. Dabei gehen sie insbesondere auf die Module des Systems und das Prompt Engineering ein, gefolgt von einer Beschreibung der Nutzung des Systems. Abschnitt 4 enthält eine kurze ethische Diskussion. Abschließend folgt in Abschnitt 5 eine Zusammenfassung mit Ausblick.

2 Grundlagen des Projekts

In diesem Abschnitt wird auf multimodale große Sprachmodelle eingegangen, zudem auf Wildtier-Webcams und auf den Zugang von blinden und sehbehinderten Personen zu Tieren. Zudem werden im Rahmen einer knappen Literaturrecherche verwandte Projekte vorgestellt. Obwohl Animal-Computer Interaction nicht im Detail untersucht wird, lässt sich das Projekt sowohl diesem Feld als auch dem Bereich der Inclusive AI zuordnen. Die Tiere interagieren nicht direkt mit dem VISUAL-System, doch das Projekt steht im Einklang mit den Zielen der ACI, wie sie in Mancinis Manifest formuliert sind, das die Förderung neuer Formen der Kommunikation und der Beziehungen zwischen Menschen und nichtmenschlichen Tieren betont (Mancini 2011). Indem VISUAL blinden und sehbehinderten Menschen den Zugang zu Wildtiererfahrungen ermöglicht, trägt das System zur Entwicklung von Empathie, Bewusstsein und einem vertieften Verständnis für Tiere und ihr Verhalten bei. Auf diese Weise leistet das Projekt einen Beitrag zur ACI, indem es das Spektrum der Mensch-Tier-Begegnungen erweitert und inklusive Formen interspezifischer Verbindungen fördert.

2.1 Multimodal Large Language Models

Ein großes Sprachmodell (Large Language Model, LLM) kann Sprache allgemein „verstehen" und erzeugen. Es nutzt große Datenmengen, um während des Trainings Milliarden oder sogar Billionen von Parametern zu erlernen. Multimodale große Sprachmodelle können neben Text auch Bilder, Audio und Video verarbeiten und ausgeben (Bendel 2024a, b). Sie alle gehören zur sogenannten generativen KI,

also zu einer Gruppe von Anwendungen der künstlichen Intelligenz (KI), die Inhalte erzeugen können. Ein bislang nicht zufriedenstellend gelöstes Problem, insbesondere bei Textgeneratoren, ist das sogenannte Halluzinieren, also die unbeabsichtigte Wiedergabe falscher Inhalte. Für das System im Projekt VISUAL ist es entscheidend, dass Bilder – auch Bildserien – verlässlich analysiert und evaluiert sowie Schlussfolgerungen daraus gezogen werden können. Im Hinblick auf blinde und sehbehinderte Menschen ist zudem eine Sprachausgabe von besonderer Bedeutung.

Um ein öffentlich verfügbares, bereits trainiertes MLLM anzupassen, werden im Wesentlichen drei Techniken verwendet. Die erste ist Prompt Engineering (Bendel 2024a). Dabei fügt der Entwickler Anweisungen an das System ein, an die sich dieses zu halten hat. Solche Anweisungen können sehr komplex sein. Sie können sich auf die Rolle und die Aufgaben des Systems beziehen. Die zweite Technik ist Retrieval-Augmented Generation (RAG). Dabei wird eine zusätzliche Wissensbasis aufgebaut oder genutzt, auf die sich das LLM oder MLLM beziehen kann. Man lädt einzelne Dokumente hoch oder baut eine Datenbank auf – oder bezieht wie im Projekt Wikipedia ein. Damit wird das System zum Spezialisten, und seine Halluzinationen werden vermindert. Eine dritte Technik ist das Finetuning, also das Trainieren des Sprachmodells mit geeigneten Daten. Es handelt sich um eine aufwändige und fehlerbehaftete Methode, die für das VISUAL-Projekt nicht infrage kam.

Im Kontext von VISUAL sollen MLLMs die Präsenz von Tieren sowie bestimmte beobachtbare Verhaltensweisen in grober Weise erkennen und beschreiben, etwa Nahrungsaufnahme, Fellpflege oder Gruppenbewegungen. Feinere oder komplexere Verhaltensmuster lassen sich hingegen voraussichtlich nicht zuverlässig erfassen. Anstatt fachkundige Beobachtung zu ersetzen, nutzt VISUAL diese Fähigkeiten als zugänglicher Vermittler, der blinden und sehbehinderten Benutzern eine sinnvolle Annäherung an tierliche Aktivität ermöglicht.

2.2 Wildtier-Webcams

Webcams als eine Form von Telepräsenzmedien sind seit den 1990er-Jahren ein beliebtes Mittel, um Einblicke in fremde Regionen oder in die Situation vor Ort zu bekommen. Oft schaut man, wie das Wetter an einem Ort ist, oder blickt in exotische Landschaften und Städte. Es existieren Tausende öffentlich zugängliche Webcams auf der ganzen Welt (Jacobs et al. 2009). Die älteren aus den 1990er-Jahren bieten oft nur eine niedrige Auflösung, die neueren aus den 2020er-Jahren häufig HD-Qualität. Auch für die Beobachtung von Wildtieren hat sich ein beträchtliches Angebot entwickelt. Es gibt Plattformen wie Explore.org (https://explore.org/livecams), die Bilder von Webcams in vielen unterschiedlichen Regionen (hier auch Umwelt oder Ökosystem genannt) auf der Welt zeigen. Sie erschienen ideal, um sie in den VISUAL-Prototyp einzubinden.

Öffentliche Wildtier-Webcams ermöglichen es den Benutzern, entfernte Orte und das dort stattfindende Tierleben in Echtzeit zu beobachten. Kamphof (2011) erklärt, dass die Webcams trotz der tatsächlichen physischen Entfernung das Gefühl vermitteln, „vor Ort“ zu sein. Diese Echtzeitverbindung ermöglicht also nicht nur einen virtuellen Zugang, sondern schafft auch eine affektive Verbindung zwischen den Benutzern und den Ereignissen, die sie beobachten. Darüber hinaus stellen die Webkameras ein Gleichgewicht zwischen ihrer Nutzung zur Überwachung und ihrer Fähigkeit zur emotionalen Bindung und zum Engagement her (Jacobs et al. 2009).

Die Webcams von Explore.org zeigen Ökosysteme auf verschiedenen Kontinenten und sowohl nördliche als auch südliche bzw. niedrig und hoch gelegene Gegenden mit unterschiedlichen Klimaten. Sie richten sich auf Wildtiere in der Luft, auf dem Boden und im Wasser. Das Problem bei Vögeln ist,

dass sie in der Luft nur schwer zu erfassen sind. Manchmal sind Webcams auf Nester gerichtet, wo aber nur zu bestimmten Zeiten eine Aktivität zu verzeichnen ist. Auf dem Boden oder im Wasser hingegen hat man mehr oder weniger stabile Zustände, wenn der Standort richtig ausgewählt wurde. Solche Webcams erschienen für das VISUAL-Projekt optimal.

2.3 Der Zugang von blinden und sehbehinderten Personen zu Tieren

Blinde Personen können eine ganz unterschiedliche Lebensgeschichte haben. Manche sind von Geburt an blind, manche im Laufe ihres Lebens erblindet. In Bezug auf die Wahrnehmung von Tieren bedeutet dies einen erheblichen Unterschied. Weiter gibt es neben den blinden auch noch sehbehinderte Personen. Manche sind fast blind, andere haben erhebliche Einschränkungen, können aber z.B. noch Strukturen erkennen oder dunkle und helle Stellen unterscheiden. Die Weltgesundheitsorganisation (WHO) klassifiziert Sehbehinderungen in fünf Stufen (WHO 2022). Im VISUAL-Projekt wurden diese Kategorien nicht unterschieden, sondern der Extremfall von Menschen in den Fokus gerückt, die seit Geburt blind sind, auch wenn das System ebenso Personen mit anderen Formen von Sehbeeinträchtigung zugutekommen kann.

Kim et al. (2019) verglichen von Geburt an blinde oder früh erblindete mit sehenden Erwachsenen hinsichtlich ihrer Wahrnehmung von Tieren. Dabei zeigte sich, dass blinde Personen allgemein weniger mit Tieren vertraut waren. In Aufgaben zu Größe, Form, Hautstruktur und Farbe gab es viele Übereinstimmungen zwischen den Gruppen. So wurden etwa Elefanten von beiden Gruppen als größer als Nashörner und Giraffen als größer als Löwen eingeschätzt. Auch bei der Unterscheidung zwischen Wasser- und Landtieren waren sich beide Gruppen einig. Unterschiede zeigten sich jedoch bei Details: Sehende unterschieden Tiere z.B. nach Fellarten, während Blinde allgemeinere Merkmale und Kategorien wie Lebensraum oder Art heranzogen.

Ein System wie VISUAL, das sich an blinde und sehbehinderte Personen richtet, kann solchen Unterschieden über das Prompt Engineering und das Retrieval-Augmented Generation Rechnung tragen. Durch das Training über Internetdaten ist ein LLM oder MLLM eher auf die Gewohnheiten und Bedürfnisse der Mehrheit ausgerichtet. Das dürfte auch bedeuten, dass es eher die Sehgewohnheiten von Nichtblinden berücksichtigt als die Erfahrungen von Blinden. Über Prompt Engineering und RAG kann man zum einen die typischen Unterschiede in der Wahrnehmung berücksichtigen, zum anderen genau dort kompensieren, wo Lücken im Wissen und Fehler in der Einschätzung bestehen. Ein möglicher Weg ist, die Szenen in erzählerischer Weise mit einem hohen Grad an Genauigkeit darzustellen.

2.4 Verwandte Projekte

In einigen Projekten werden Möglichkeiten von Inclusive AI vorgestellt und diskutiert. Bereits erwähnt wurden die Tests von Be My Eyes mit der Funktion Be My AI (Bendel 2024b). Eine Studie von Forschern der Stanford University von 2025 arbeitet die Möglichkeiten heraus, bei der Entwicklung innovativer Systeme Lernunterschiede zu berücksichtigen. Dabei fokussiert man auf behinderte und beeinträchtigte Personen (McGee et al. 2025).

In anderen Projekten wird künstliche Intelligenz eingesetzt, um Wildtiere in Zoos und in freier Wildbahn zu überwachen – zum Beispiel in Bezug auf Position, Verhalten und Gesundheit (Congdon et al. 2022). In diesem Zusammenhang wurde der Einsatz von Gesichtserkennung vorgeschlagen, um einzelne Tiere verfolgen zu können (Mazurkiewicz 2024; Bendel und Yürekkirmaz 2023). Auch die Emotionserkennung bei Haustieren wird erforscht. Ein Forscherteam hat eine Deep-Learning-

Architektur vorgestellt, mit der individuelles und soziales Tierverhalten – selbst in komplexen Umgebungen – direkt aus Rohvideobildern klassifiziert werden kann (Tsuji et al. 2023). Projekte wie die BBC Springwatch nutzen Machine-Learning-Techniken zur Tiererkennung in Wildlife-Feeds (BBC 2020). Diese Systeme liefern aber keine narrative Sprachausgabe.

Zhang et al. (2025) haben in der Studie „I Can See Forever!“ ein System evaluiert, das Video-LLMs in Echtzeit verwendet, um blinden Personen bei Alltagsaufgaben zu helfen – etwa im Haushalt oder in sozialen Situationen. Allerdings wurde dies nicht speziell auf Tiervideos oder Webcam-Feeds angewandt. Chen et al. (2025) stellten ein multimodales Vision-Language-Modell vor, das durch Szenenbeschreibung blinden Menschen helfen soll, die Umgebung zu verstehen, etwa durch Wearables. Es bezieht sich auf reale Umgebungsszenen, nicht auf Tierwebcams oder Bündel von Einzelbildern. Weitere Arbeiten fokussieren auf Bild-zu-Text-Systeme, entwickelt für Alltagsszenen blinder Benutzer (Karamolegkou et al. 2025). Hier fehlt ebenfalls der Fokus auf Tier-Webcams oder Foto-Bundles. Das Accessible Aquarium nutzte Videostreams von Fischen und Aquarien, um eine Kombination aus Audiodeskription und Sonifikation beziehungsweise Musik zu erzeugen, die blinden Menschen ermöglichte, ein lebendiges Aquarium zu erleben (Pendse et al. 2008).

Ein Projekt, das wie VISUAL Inclusive AI und ACI mit Hilfe der Möglichkeiten von generativer KI bzw. LLM und MLLM kombiniert, ist den Autoren dieses Beitrags nicht bekannt. Man muss dazu sagen, dass generative KI erst 2022 ihren Durchbruch hatte, auch wenn sie sich schon Jahre vorher angekündigt hatte, etwa als Erweiterung von sozialen Robotern (Coursey 2020). Inclusive AI ist ebenfalls ein junges Forschungs- und Entwicklungsfeld, und es gibt erst wenige Projekte, die blinde und sehbehinderte Personen berücksichtigen. ACI ist ein Bereich, der schon seit einigen Jahren erforscht wird, dessen Community aber sehr klein ist. Hier kam es erst ab ca. 2024 zu einer Hinwendung zu generativer KI.

3 Umsetzung des Projekts

Dieser Abschnitt beschreibt die Projektvorbereitung und die Entwicklung des VISUAL-Systems. Zudem wird die Onlineumfrage unter blinden und sehbehinderten Personen skizziert, einschließlich wichtiger Ergebnisse.

3.1 Ausgangsüberlegungen

Im Zentrum des VISUAL-Projekts steht die Implementierung eines innovativen Systems. Die Idee dazu hat sich durch die technologische Entwicklung in generativer KI und die Fortschritte in Inclusive AI und ACI ebenso ergeben wie durch die Beobachtung, dass beeinträchtigte Personen oft bestimmte Erfahrungen nicht machen können, die für viele selbstverständlich sind. Ziel des VISUAL-Prototyps ist es, blinden oder sehbehinderten Menschen virtuelle Videosafaris zu ermöglichen (Jovic 2025). Obwohl Live-Webcams faszinierende Bilder vom Aussehen und Leben von Tieren liefern, bleiben diese für Benutzer, die auf nichtvisuelle Zugangsformen angewiesen sind, weitgehend unzugänglich. Zudem können angebotene digitale Inhalte häufig nicht die Bedürfnisse blinder oder sehbehinderter Menschen erfüllen.

VISUAL möchte diese Barrieren überwinden, indem es Livestreams von Wildtier-Webcams nutzt und mit erzählerischen Audiobeschreibungen kombiniert (Jovic 2025). Diese sollen nicht nur das aktuelle Geschehen schildern, sondern auch durch kontextuelle und lehrreiche Informationen ergänzt werden, die je nach Altersgruppe und Nutzerpräferenz angepasst werden können. Damit werden blinde und

sehbehinderte Personen mit möglichst vielen Informationen versorgt, sodass sie sich ein Bild von der Szene und ihrem Hintergrund machen können. Sie werden aber auch gezielt dort unterstützt, wo sie aufgrund ihrer Einschränkung gewisse Lücken oder Verzerrungen haben.

Als Forschungsfrage wurde formuliert: „How can a multimodal large language model enable 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) Wie an der Hochschule für Wirtschaft FHNW und speziell im Forschungsbereich des Erstautors üblich, sollte die Forschungsfrage nicht nur auf theoretische, sondern auch auf praktische Weise beantwortet werden – eben mit Hilfe eines Prototyps. Dieser sollte grundsätzlich alle notwendigen Funktionalitäten aufweisen und technisch vollständig betriebsbereit sein. Eine öffentliche Bereitstellung war zu diesem Zeitpunkt nicht angedacht. Man kann aber sämtliche Dateien und Codes herunterladen und das System nachbauen (https://github.com/jovicyy/VISUAL_2025/).

3.2 Benutzeranalyse

Die im Projekt durch eine Literaturanalyse identifizierten Barrierefreiheitsbedürfnisse wurden in konkrete Gestaltungsprinzipien für den VISUAL-Prototyp übertragen (Jovic 2025). Die Struktur orientiert sich an den WCAG-POUR-Prinzipien: Wahrnehmbar (Perceivable), Bedienbar (Operable), Verständlich (Understandable) und Robust (Robust). Diese Prinzipien leiteten den Entwicklungsprozess:

- Wahrnehmbar (Perceivable): Blinde und sehbehinderte Benutzer sind stark auf reichhaltige und beschreibende Audiodeskriptionen angewiesen. Im VISUAL-Prototyp wird dies durch kontextuelle Audio-Nacherzählungen des Webcam-Livestreams umgesetzt, wobei der Fokus auf der tatsächlichen Beschreibung der Szene liegt, nicht auf interpretierendem Storytelling.
- Bedienbar (Operable): Visuell beeinträchtigte Benutzer erfahren häufig Frustrationen durch schwer zugängliche Navigation und mausabhängige Bedienung. VISUAL beugt dem vor, indem alle Funktionen vollständig per Tastatur steuerbar sind und alle Bedienelemente in logischer Lesereihenfolge angeordnet sind – optimal für automatische Screenreader.
- Verständlich (Understandable): Die Forschung ergab zudem, dass inkonsistente Layouts und unklare Strukturen die Nutzung digitaler Inhalte erschweren. Um dem entgegenzuwirken, verwendet VISUAL vorhersehbare Navigationsmuster, eine klare Struktur sowie einfache Sprache in Text und Audio. Zusätzlich gibt es einen speziellen Bereich, der den Benutzern erklärt, wie die Navigation funktioniert und was sie erwarten können.
- Robust (Robust): Wenn Inhalte nicht mit unterstützenden Technologien oder verschiedenen Geräten kompatibel sind, leidet die Barrierefreiheit. VISUAL stellt sicher, dass alle Benutzeroberflächenelemente korrekt beschriftet sind, sodass ihre Funktion von unterstützenden Technologien – insbesondere Screenreadern – richtig erkannt und genutzt werden kann.

Einzelne Funktionen wurden passend zum Gegenstand der Anwendung umgesetzt. So hört man nach erfolgreicher Navigation einen Löwen brüllen („Free Lion Roar Sound Effects Download“ von Pixabay, https://pixabay.com). Von visuellen Elementen wurde aus verschiedenen Gründen nicht abgesehen. Blinde und sehbeeinträchtigte Personen nutzen manchmal Anwendungen zusammen mit nicht beeinträchtigten Personen, etwa Assistenzpersonen, die so die Anwendung auf die für sie übliche Weise sehen und gegebenenfalls bedienen können.

3.3 Onlineumfrage

Die Onlineumfrage zum VISUAL-System wurde barrierefrei erstellt (Jovic 2025). Dabei half ein stark sehbehinderter Mitarbeiter der Hochschule für Wirtschaft FHNW, der in seinem Arbeitsalltag als Stundenplaner auf solche Systeme angewiesen ist. Ziel der Umfrage war es, Erkenntnisse von blinden und sehbehinderten Personen zu sammeln, um deren Bedürfnisse, Vorlieben und Erwartungen besser zu verstehen. Den Teilnehmern wurde mitgeteilt, dass ihre Unterstützung maßgeblich dazu beitrage, einen Prototyp zu entwickeln, der virtuelle Videosafaris zugänglicher und angenehmer gestalten soll. Es wurden 22 Fragen in sechs thematischen Abschnitten gestellt.

Über LinkedIn und Blogs wurde im Juni 2025 die auf Deutsch und Englisch verfügbare Onlineumfrage beworben. Die meisten Blindenorganisationen im deutschsprachigen Raum verweigerten eine Unterstützung, bis auf den Schweizerischen Blindenbund, der auf LinkedIn und auf seiner Website auf die Umfrage hinwies. Teilgenommen haben elf Personen. Dies scheint nicht viel zu sein, aber es ist schwierig, die Betroffenen zu erreichen, wenn man nicht selbst in den entsprechenden Netzwerken ist und es ansonsten an Unterstützung mangelt. Dennoch sind die Ergebnisse wertvoll für die explorative Forschung.

Im Folgenden werden die für das Systemdesign relevantesten Fragen und Antworten sowie zentrale Erkenntnisse zusammengefasst (Jovic 2025). Eine der zentralen Fragen, „Wenn Sie einer Audiodeskription einer Wildtierszene zuhören würden, welche Informationen wären für Sie am wichtigsten?“, zeigte, dass die Teilnehmer besonderen Wert darauf legten, zu erfahren, welche Tiere anwesend sind und welches Verhalten sie zeigen – beide Aspekte wurden von neun Befragten genannt. Die Umgebung wurde von acht Teilnehmern hervorgehoben, sechs betonten das Erscheinungsbild der Tiere, und fünf hoben die Atmosphäre oder Stimmung der Szene hervor. Daraus wurde geschlossen, dass Audiodeskriptionen Tierarten und Verhaltensweisen klar benennen und zusätzlich den Umweltkontext, physische Merkmale sowie die emotionale Atmosphäre der Szene einbeziehen sollten.

Eine weitere Frage, „Wie wichtig ist es, die Webcam-Region auswählen zu können?“, ergab, dass die meisten Befragten die regionale Auswahl als wichtig einschätzten: Zwei bewerteten sie als sehr wichtig, sechs als ziemlich wichtig und drei als neutral. In der Folge wurde die Regionsauswahl als zentrale, leicht zugängliche Funktion implementiert. Im Gegensatz dazu erhielt die Frage „Wie wichtig ist es, den Tier-Typ auswählen zu können?“ weniger starke Zustimmung: Nur eine Person bewertete sie als sehr wichtig, drei Personen bewerteten sie als ziemlich wichtig und sechs als neutral. Entsprechend wurde die Tierauswahl als sekundäre Option behandelt, während der Schwerpunkt auf der Auswahl des Lebensraums lag.

Schließlich zeigte die Frage „Wie hilfreich wäre es, die Option zu haben, Ton oder Stil der Audiodeskriptionen an Ihre Vorlieben anzupassen?“, dass die Mehrheit diese Funktion als nützlich erachtete: Fünf stuften sie als äußerst hilfreich ein, vier als sehr hilfreich, eine Person als neutral und eine als nicht hilfreich. Als Reaktion darauf integrierte das VISUAL-System eine Tonwahlfunktion in Form von drei unterschiedlichen Modes, die es den Benutzern ermöglicht, den Erzählstil – etwa ruhig, sachlich oder erzählerisch – an die persönlichen Präferenzen anzupassen.

3.4 Systemanforderungen

Die Systemanforderungen zum VISUAL-System sind in zwei Kategorien unterteilt: funktionale und nichtfunktionale Anforderungen (Jovic 2025). Funktionale Anforderungen beschreiben, welche

Aktionen das System ausführen muss, um die Bedürfnisse der Benutzer zu erfüllen. Sie wurden auf der Basis von Literaturanalyse und Teammeetings entwickelt.

Das System verarbeitet Live-Webcam-Streams mit Tieren in unterschiedlichen Regionen und Umgebungen, darunter Savannen, Dschungel, Wälder, Gebirge, Polarregionen und Ozeane. Es verwendet das Multimodal Large Language Model GPT-4o zur Generierung von Szenenbeschreibungen, die anschließend durch ein Text-to-Speech-System in gesprochene Sprache umgewandelt werden. Die Benutzer haben die Kontrolle über die Audiowiedergabe und können diese starten, stoppen, wiederholen und die Abspielgeschwindigkeit anpassen. Das System ermöglicht den Wechsel zwischen verschiedenen Regionen und Kameras. Die Inhalte können an das Alter des Benutzers angepasst werden, mit getrennten Einstellungen für Kinder und Erwachsene. Zusätzlich kann zwischen drei Erzählmodi gewählt werden: Adventurer (z.B. Safari Adventurer), Field Scientist oder Calm Observer. Das System bietet zudem Bildungsinhalte, indem es mittels RAG relevantes Faktenwissen aus Wikipedia bezieht. Die Zugänglichkeit wird durch vollständige Tastaturbedienbarkeit, Screenreader-Kompatibilität und eine intuitive Navigation sichergestellt.

Nichtfunktionale Anforderungen beschreiben, wie das System arbeiten soll. VISUAL hält die WCAG-Richtlinien vollständig ein, verwendet semantisch korrektes HTML mit passenden Labels und ARIA-Rollen und stellt eine kontrastreiche Benutzeroberfläche sicher. Externe APIs werden kontrolliert und sicher eingesetzt, um Datenschutz und Zuverlässigkeit zu gewährleisten.

Auch wenn keine formale Evaluation verschiedener MLLMs durchgeführt wurde, zeigten Tests mit der Google Cloud Vision API und GPT-4o, dass GPT-4o in diesem Kontext deutlich besser abschnitt. Es erkannte Tiere selbst dann, wenn sie nur schwer zu sehen waren, und schätzte ihre Anzahl in dichten Gruppen oder Herden präziser. Zudem war GPT-4o bereits in früheren Projekten an der Hochschule erfolgreich eingesetzt worden, was es zu einer logischen und effektiven Wahl für diese Implementierung machte. GPT-5 wurde erst nach Abschluss des Projekts veröffentlicht. Die Entwicklung eines domänenspezifischen MLLM von Grund auf war aufgrund des erforderlichen Aufwands nicht realisierbar.

3.5 Konzeptionelle Architektur

Der VISUAL-Prototyp basiert auf einer mehrschichtigen Architektur, die festlegt, wie die Systemkomponenten zusammenarbeiten, um Webcam-Livestreams in barrierefreie Audiobeschreibungen umzuwandeln (Jovic 2025).

- Der Prozess beginnt auf der Frontend-Ebene, wo die Benutzer ihre Erfahrung individuell anpassen: Sie wählen das bevorzugte Ökosystem, worauf automatisch die Wildtier-Webcam ausgewählt wird, dann ihr Altersprofil (Kind/Erwachsener) und den gewünschten Abenteuer-Modus (Adventurer, Field Scientist oder Calm Observer). Diese Einstellungen werden als strukturierte JSON-Daten an das Backend weitergeleitet.
- Im Backend wird anschließend eine Headless-Browser-Seite über Puppeteer (eine JavaScript Library) gestartet. Dort werden drei aufeinanderfolgende Screenshots vom gewählten Livestream gemacht, jeweils im Abstand von drei Sekunden. Diese Bilder werden zu einem temporären Paket gebündelt und an die nächste Ebene zur KI-Verarbeitung weitergeleitet.
- In der KI-Verarbeitungsschicht wird ein dynamischer System-Prompt erstellt, der die Benutzereinstellungen berücksichtigt (z.B. Rolle und Profil). Dieser Prompt wird an die GPT-4o-API von OpenAI gesendet. GPT-4o führt dann eine multimodale Analyse des Bildpakets durch und

generiert eine kontextuell passende und beschreibende Erzählung. Diese Rohbeschreibung wird anschließend an die nächste Schicht weitergeleitet.

- In der Nachbearbeitungsschicht wird ein Filter auf die erzeugte Beschreibung angewendet, um die Qualität sicherzustellen. Dabei werden unter anderem sogenannte „Anti-Halluzinations-Checks“ durchgeführt, um unrealistische Inhalte (z.B. Eisbären in der Savanne) zu entfernen. Dies wird allein durch das Sprachmodell selbst erledigt, da RAG hier keine Verbesserungen brachte. Zudem erfolgt eine sprachliche Überarbeitung, bei der visuelle Verben wie „sehen“ oder „anschauen“ durch räumliche Begriffe wie „vor dir“ oder „zu deiner Rechten“ ersetzt werden.
- Die bereinigte Erzählung wird schließlich an das Frontend zurückgesendet und dort mit Hilfe der Web Speech API in gesprochene Sprache umgewandelt, basierend auf dem Stimmprofil des gewählten Abenteuer-Modus und des Profils (weibliche und männliche Stimmen).
- Nach der Audiowiedergabe kann der Benutzer optional zusätzliche Bildungsinformationen über die erkannten Tiere anfordern. Diese werden über ein RAG-Verfahren bereitgestellt, das Wikipedia nach zuverlässigen Informationen durchsucht. Die Inhalte werden nicht 1:1 geliefert, sondern in aufbereiteter und verbesserter Form.

Diese Architektur legt besonderen Wert auf Barrierefreiheit und Modularität und stellt sicher, dass alle Komponenten – Benutzeroberfläche, Bildschirmaufnahme, KI-Analyse, Narrativgenerierung und Sprachausgabe – nahtlos zusammenarbeiten. Dadurch bietet das System eine intuitive, verlässliche und hochgradig anpassbare virtuelle Safarierfahrung für blinde und sehbehinderte Benutzer.

3.6 Modulübersicht

In diesem Abschnitt werden die Module des VISUAL-Systems beschrieben, zum einen die Explorer-Modes und die Profile, zum anderen die Regionen (Jovic 2025). Die Modes wurden von der Zweitautorin auf der Grundlage von Literaturanalyse und Analyse von Best Practices entwickelt. Ein Beispiel für das Prompt Engineering für Modes und Profile findet sich im nächsten Abschnitt.

3.6.1 Explorer-Modes

Der Safari Adventurer (im System als Adventurer geführt, der dann je nach Region zum Savanna Adventurer, Ocean Adventurer etc. wird) vermittelt das Gefühl, live auf Safari zu sein. Die Erzählung ist lebendig und schrittweise aufgebaut, und sie beschreibt konkret, was im Moment zu sehen ist. Ziel ist es, eine eindrucksvolle Naturbeobachtung zu ermöglichen – ohne spekulative oder erfundene Inhalte. Die Sprache ist anschaulich, aber natürlich, in direkter Ansprache gehalten („Vor dir …“) und legt den Fokus auf Bewegungen, Aussehen und Umgebung der Tiere. Diese Form eignet sich besonders für Benutzer, die ein intensives, aufregendes Naturerlebnis suchen.

Der Field Scientist legt den Schwerpunkt auf wissenschaftliche Genauigkeit und Bildung. Die Erzählweise ist sachlich und dokumentarisch, erklärt Tierverhalten sowie biologische Hintergründe klar und verständlich. Dabei bleibt der Ton ruhig und informativ, ohne spekulative oder erzählerische Ausschmückungen. Zielgruppe sind Benutzer, die sich für naturwissenschaftliche Zusammenhänge interessieren und beim Beobachten dazulernen möchten.

Der Calm Observer richtet sich an Benutzer, die Ruhe und Entspannung suchen. Die Erzählung ist sanft und meditativ und schafft ein Gefühl von Sicherheit und Zeitlosigkeit. Bewegungen werden nur dezent beschrieben, die Umgebung als beruhigend und schützend geschildert. Die Sprache ist bewusst weich

und atmosphärisch, mit einem Fokus auf Licht, Temperatur und Stille. Dieser Modus dient vor allem der emotionalen Entlastung und achtsamen Naturverbundenheit.

3.6.2 Altersprofile

Die Erwachsenenprofile passen die Sprache der gewählten Explorer-Modi leicht an, indem sie ein etwas reicheres Vokabular und zurückhaltende Kontextinformationen verwenden, etwa zur Umgebung oder Atmosphäre. Der Ton bleibt sachlich und natürlich – jeweils im Stil des gewählten Modus – und vermittelt eine ausgewogene, reife Perspektive, ohne zu vereinfachen oder überzuerklären.

Die Kinderprofile gestalten die Erzählungen so, dass sie leicht verständlich, sicher und anschaulich sind. Die Sprache ist einfach, rhythmisch und bildhaft, mit betont freundlichem und beruhigendem Ton. Auch potenziell angespannte Szenen werden neutral beschrieben, ohne belastende Inhalte. Ziel ist eine sichere, positive Naturerfahrung, die leicht nachvollziehbar bleibt und kindgerechte Vergleiche einsetzt.

3.6.3 Regionen

Jede Region verfügt über ein eigenes Set an Kameras, die sorgfältig ausgewählt wurden, um typische Tiere des jeweiligen Ökosystems abzubilden. Aus Gründen der Einfachheit wurde mehrheitlich Explore.org verwendet, zudem eine Webcam von Africam (https://africam.com). In der Regel zeigen die Kameras die genannten Tierarten, jedoch sind auch andere Sichtungen möglich. Zu sehen sind Land-, Wasser- und Lufttiere in ihrem natürlichen Lebensraum sowie in Schutzgebieten. Haustiere oder Nutztiere werden nicht gezeigt. Wegen der geringen Auswahl in der Polarregion wurde dort nur eine Kamera verwendet. Tabelle 1 zeigt eine Übersicht über alle integrierten Kameras, geordnet nach Regionen, Kameraquelle, Tierart und Beispielen für die Beschreibung. Es wurden von der Zweitautorin alle Tiere gesichtet, bis auf den Löwen – stattdessen waren dort Hyänen und Vervet Monkeys zu sehen.

Tabelle 1: Übersicht über die Webcams (nach Jovic 2025)

Region	Kameraquelle	Tierart	Verhalten/Merkmale
Afrikanische Savanne	Mpala Live Camera (Kenia)	Afrikanischer Elefant	Große Herden, komplexes Sozialverhalten, Rüsselgebrauch
Afrikanische Savanne	Mpala Live Camera (Kenia)	Löwe	Rudelverhalten, Jagdverhalten, territoriales Verhalten
Afrikanische Savanne	Botswana Wildlife Safari Cam by Africam	Zebra	Migrationsmuster, Herdenbildung, Weideverhalten
Afrikanische Savanne	Mpala Live Camera (Kenia)	Nilpferd	Halbaquatischer Lebensstil, territoriales Verhalten, Familiengruppen
Afrikanische Savanne	Mpala Live Camera (Kenia)	Giraffe	Fressverhalten, soziale Interaktion, Bewegungsmuster
Afrikanische Savanne	Mpala Live Camera (Kenia)	Diverse Antilopenarten	Interaktionen verschiedener Arten (Impala, Gazellen, Wasserbock)

Region	Kameraquelle	Tierart	Verhalten/Merkmale
Tropischer Dschungel	GRACE Gorilla Forest Corridor	Gorilla	Familienstrukturen, Fellpflege, Futtersuche
Tropischer Dschungel	Toucan TV	Tukan	Fressverhalten, territoriales Verhalten, Lautäußerungen
Tropischer Dschungel	GRACE Gorilla Forest Corridor	Verschiedene Primaten	Soziale Vielfalt und Verhalten
Gemäßigter Wald	Brooks Falls, Katmai National Park	Braunbär	Lachsfischen, saisonales Verhalten, Mutter-Kind-Interaktionen
Gemäßigter Wald	International Wolf Center	Grauwolf	Rudelverhalten, territoriales Verhalten, Jagdstrategien
Gemäßigter Wald	Brooks Falls, Katmai National Park	Waldvögel	Greifvögel, Singvögel, Wasservögel
Gemäßigter Wald	Brooks Falls, Katmai National Park	Lachswanderungen	Fischmigration und Einfluss auf das Ökosystem
Polarregion	Cape East Camera, Churchill Cam	Eisbär	Anpassung an Kälte, Jagdverhalten, Spielkämpfe, Mutter-Kind-Verhalten
Ozean	Utopia Village Reef Cam	Haie (verschiedene Arten)	Fress- und Territorialverhalten
Ozean	Shark Cam	Haie (verschiedene Arten)	Fress- und Territorialverhalten
Ozean	Utopia Village Reef Cam	Tropische Fische	Schwarm- und Einzelverhalten
Ozean	Utopia Village Reef Cam	Manta-Rochen	Schwimmmuster, Fressverhalten

3.7 Prototypentwicklung und Gestaltungsprozess

Dieser Abschnitt erklärt die Methoden, Werkzeuge und Entscheidungen, mit denen das zuvor beschriebene Rahmenwerk in einen barrierefreien und funktionalen Prototyp umgesetzt wurde (Jovic 2025). Dann zeigt er das Prompt Engineering am Beispiel des Safari Adventurer auf. Zudem enthält er eine Schritt-für-Schritt-Darstellung der Nutzung des Prototyps, den „Walkthrough“.

3.7.1 Gestaltungsansatz und verwendete Werkzeuge

VISUAL wurde mit mehreren Tools entwickelt, nämlich Figma für das UI-Design (mit einem designorientierten Low-Code-Ansatz) und Cursor für die KI-gestützte Code-Erzeugung (Jovic 2025). Diese Kombination ermöglichte die Erstellung eines funktionalen Proof-of-Concept-Prototyps, ohne Kompromisse bei Benutzerfreundlichkeit oder Barrierefreiheit eingehen zu müssen.

Zur Überprüfung der Barrierefreiheit und des Nutzerflusses kam das Figma-Plugin „Stark – Contrast & Accessibility Checker“ zum Einsatz. Nachdem die Zugänglichkeit und Navigationslogik gewährleistet war, wurde der Dev Mode MCP Server von Figma verwendet. Dieses Tool war entscheidend für die Umsetzung des Designs in Code, da es KI-basierten Entwicklungsumgebungen – wie Cursor – Zugriff

auf strukturierte Designinformationen gibt (z.B. Farben, Abstände, Komponenten, Hierarchien). Diese Daten sind weitaus präziser als Screenshots oder statische Bilder.

In der Praxis bedeutete dies: Sobald der MCP-Server aktiviert und mit Cursor verbunden war, konnten ausgewählte Design-Frames direkt in Code übersetzt werden. Ein erster Prompt an Cursor lautete: „Hey, bitte erstelle mir einen klickbaren Prototyp einer ‚visuellen Safari' für blinde und sehbehinderte Benutzer mit semantischem HTML und Tailwind CSS. Halte dich strikt an das Layout der Startseite. Die Seite soll nicht scrollbar sein, alles muss auf den Bildschirm passen – exakt wie in meiner Figma-Datei. Weitere Seiten folgen später, bitte erstelle nur die Startseite. Vielen Dank!" Durch die Bereitstellung strukturierter Designdaten und präziser Anweisungen konnte Cursor funktionale und barrierefreie Komponenten generieren.

Das Backend von VISUAL wurde mit Node.js umgesetzt, das als zentrales Steuerungssystem für alle Kernprozesse dient. Hier laufen alle erforderlichen APIs zusammen – z. B. Puppeteer, GPT-4o, Wikipedia und die Web Speech API. Die Entscheidung für die Nutzung von Cursor beruhte auf drei Hauptgründen:

- Effizienz: Die Entwicklungszeit wurde stark verkürzt, da Figma-Designs schnell und zuverlässig in ein funktionierendes Frontend und Backend übertragen werden konnten.
- Barrierefreiheitstests: Die kontinuierliche Prüfung und Optimierung barrierefreier Funktionen waren direkt im Prozess integriert.
- Fokus auf Funktionalität: Da die Infrastruktur automatisiert generiert wurde, konnte sich die Entwicklung auf zentrale Aspekte wie Erzählqualität, API-Kompatibilität und Benutzerfreundlichkeit konzentrieren.

Allerdings zeigte sich auch, dass KI-gestütztes Coden nicht automatisches Coden bedeutet. Um gute Ergebnisse zu erzielen, waren präzise Prompts, technische Kenntnisse und gezielte Recherchen notwendig. So war etwa ein vager Prompt wie „Bitte erstelle eine Funktion, um Screenshots aus der eingebetteten YouTube-API zu machen" nicht ausreichend. Stattdessen war ein präziser Prompt erforderlich, z.B.: „Schreibe eine Funktion mit puppeteer-core, die eine YouTube-Video-URL im Headless-Modus (Microsoft Edge) öffnet, das Video startet und drei Screenshots im Abstand von 3 Sekunden aufnimmt. Speichere die Screenshots als .png-Dateien im Ordner ‚safari-screenshots'." Diese Herangehensweise erforderte technisches Verständnis und die Fähigkeit, verschiedene Quellen zu kombinieren und die KI präzise anzuleiten.

Die Zusammenarbeit von Figma (inkl. MCP-Server) und Cursor ermöglichte es, die Lücke zwischen Konzept und funktionalem Prototyp zu schließen. Dieser Ansatz beschleunigte nicht nur die Entwicklung, sondern gewährleistete auch durchgehend die Einhaltung inklusiver Designprinzipien.

Tabelle 2: Beispiel für Prompts (Safari Adventurer) (nach Jovic 2025)

Kategorie	**Instruktion**
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

Kategorie	Instruktion
	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. 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 Beispiel für Prompt Engineering

Es wurde für die Rollen und Profile umfangreiches Prompt Engineering vorgenommen. Aus Platzgründen können nicht alle Prompts gezeigt werden. Es wird aber ein Beispiel für den Safari Adventurer gezeigt, in der Kombination mit dem Profil (hier Erwachsener). Um den Aufbau deutlich zu machen, wurde in diesem Beitrag die Tabellenform gewählt (Tabelle 2).

Das Output-Beispiel von einer Gorilla-Szene lautet in einem Ausschnitt: „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." Die Beschreibung eines Nilpferds ist Abbildung 4 zu entnehmen. Sie beginnt mit den Worten „In front of you, there's a large hippo resting in the mud …".

3.7.3 Prototyp-Durchgang („Walkthrough")

Der VISUAL-Prototyp besteht aus mehreren miteinander verbundenen Seiten, die jeweils im Hinblick auf einfache Navigation und Zugänglichkeit gestaltet wurden (Jovic 2025). Der folgende Abschnitt stellt jede Bildschirmseite vor, zeigt deren Darstellung und erläutert Zweck, Funktionen und gestalterische Überlegungen. Wie bereits angemerkt, erzeugt jede erfolgreiche Navigation ein Löwengeräusch, um zu bestätigen, dass es jetzt weitergeht und es geklappt hat.

Willkommensseite (Welcome Page)

Die Willkommensseite dient als Einstiegspunkt zu VISUAL und führt den Benutzer in das Konzept einer inklusiven virtuellen Safari ein. Sie verwendet ein klares und minimalistisches Layout mit dem Bild eines brüllenden Löwen. Dieses Bild wurde von der Zweitautorin mit Hilfe von ChatGPT-4o erstellt und wird von einem beschreibenden Alternativtext begleitet: „Ein majestätischer Löwe mit goldener Mähne, der sein Maul weit zu einem kraftvollen Brüllen geöffnet hat".

Die Seite zeichnet sich durch folgende Merkmale aus: 1. Eine prägnante Überschrift begrüßt den Benutzer und wird von einer kurzen Beschreibung gefolgt. Die Überschriften sind hierarchisch strukturiert und nutzen unterschiedliche Ebenen, damit automatische Bildschirmleser den Text in logischer Reihenfolge vorlesen können. Dies wird im Code umgesetzt, indem „Welcome to VISUAL" als „H1", die Tagline als „H2" und die Beschreibung als „H3" ausgezeichnet sind. Diese semantische Gliederung ermöglicht dem Benutzer eine hierarchische Navigation und erlaubt es ihm, gezielt zwischen den Überschriften zu wechseln. Zudem wird der Text dadurch klar von interaktiven Elementen abgegrenzt. 2. Diese Schaltflächen befinden sich im unteren Bereich der Seite und sind – u.a. als Hilfe für farbenblinde Benutzer – mit kontrastreichen Farben gestaltet: „How it works" (this button opens a quick start guide, which is intended for first time users), „Explore by Animal" (this button enables the direct selection of a desired animal species) und „Start Exploring" (this button starts the actual virtual safari and leads to the different regions).

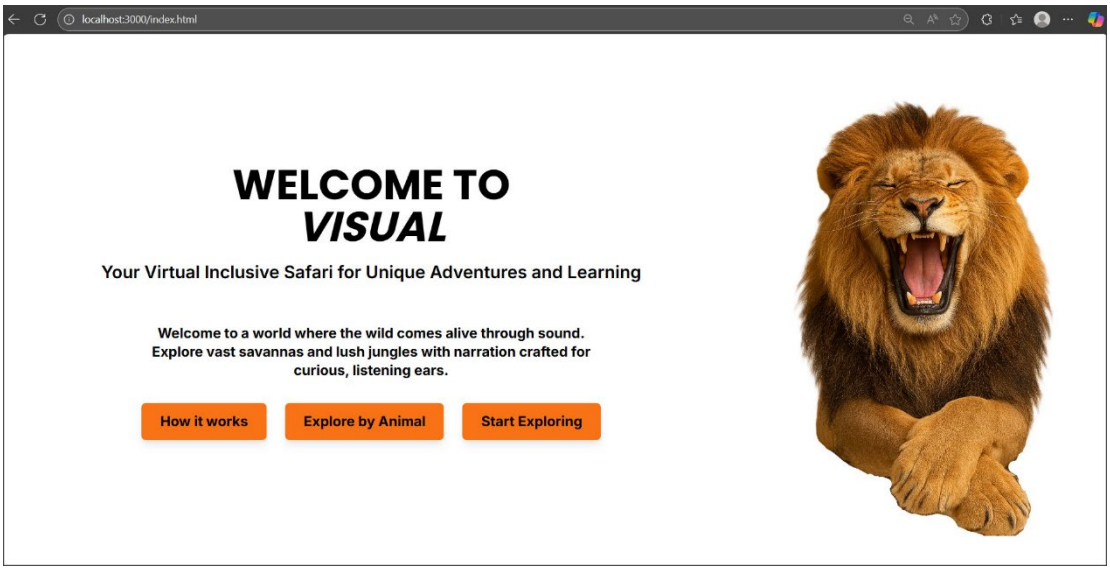

Abbildung 1: Welcome Page

Seite „How it works“ – Quick Start Guide

Die Seite „How it works“ dient als Schnellstart-Anleitung für neue Benutzer, um ihnen die grundlegenden Funktionen zu erklären und eine Beschreibung dessen zu geben, was sie erwartet und wie sie sich auf der Seite zurechtfinden können. Sie bietet eine schrittweise Einführung in die Nutzung von VISUAL, einschließlich der Navigation auf der Startseite, der Auswahl einer Region oder einer bestimmten Tierart, der Personalisierung der Erzählweise anhand von Profilen (Erwachsener/Kind) sowie drei verschiedenen Erzählstilen oder Modi (Safari Adventurer, Field Scientist, Calm Observer).

Darüber hinaus werden die möglichen Steuerungsmöglichkeiten der Erzählung erläutert, wie das Anpassen der Wiedergabegeschwindigkeit, das Pausieren und das erneute Abspielen der Erzählung. Zusätzlich beschreibt die Seite weitere Safari-Funktionen, wie die Möglichkeit, sachliche Informationen über erkannte Tiere abzurufen, die Kamera auszutauschen oder in eine andere Region zu wechseln.

Aus struktureller Sicht folgt die Seite ebenfalls den Richtlinien zur Barrierefreiheit, mit korrekter Auszeichnung von Texten und Elementen. Im Gegensatz zu den anderen Seiten wurde hier ein scrollbares Layout verwendet, um eine umfassende Anleitung mit allen notwendigen Informationen bereitzustellen, ohne die Seite zu überladen. Ein Home-Button wurde in das Design integriert, um dem Benutzer eine einfache Rückkehr zur Willkommensseite zu ermöglichen, wo die Safari beginnen kann. Die Platzierung dieses Buttons bleibt auf allen weiteren Seiten (außer der Willkommensseite) einheitlich und trägt damit zur Konsistenz des Gesamtlayouts bei. Diese Entscheidung wurde getroffen, um eine klare und benutzerfreundliche Navigation zu gewährleisten.

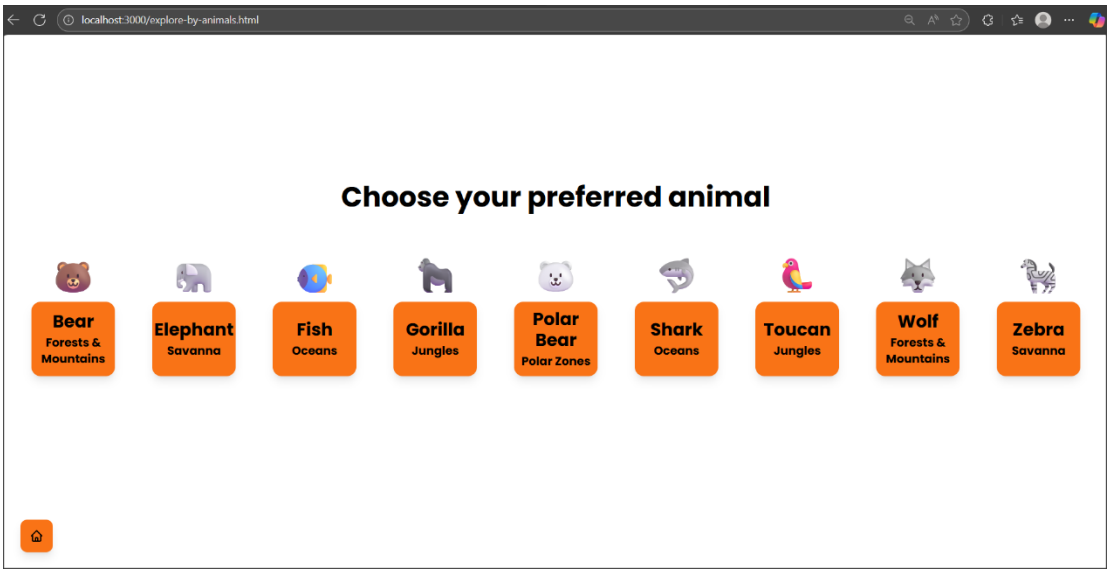

Abbildung 2: Explore by Animal

Seite „Explore by Animal" – Species-Focused Exploration

Die Seite „Explore by Animal" bietet die Möglichkeit, direkt zum Lebensraum der gewünschten Tierart zu navigieren (Abbildung 2). Sowohl der Name der Tierart als auch die zugehörige Region bzw. das zugehörige Ökosystem – etwa „Zebra" in der „Afrikanischen Savanne" – werden angezeigt und geben dem Benutzer somit bereits vor der Auswahl einen ersten Kontext. Diese Seite wurde für alle entwickelt, die eine tierartspezifische Erfahrung gegenüber einer regionalen Erkundung bevorzugen. Die eingebetteten Livestreams wurden so ausgewählt, dass eine hohe Wahrscheinlichkeit besteht, die gewünschte Tierart zu sehen. Ihre tatsächliche Anwesenheit kann jedoch nicht zu jedem Zeitpunkt garantiert werden.

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

Die Seite „Start Exploring (Choose by Region)" bildet den Einstieg in das eigentliche virtuelle Safari-Erlebnis (Abbildung 3). Sie präsentiert fünf unterschiedliche Regionen oder Ökosysteme: Savanne, Dschungel, Wälder und Gebirge, Polarzonen und Ozeane. Jede davon wird durch ein dekoratives, großes Emoji-Symbol dargestellt, das jedoch bewusst für automatische Bildschirmleser ausgeblendet ist, um redundante Informationen zu vermeiden und sicherzustellen, dass nur relevante Inhalte vorgelesen werden. Dies unterstützt zusätzlich eine reibungslose Navigation. Die Schaltflächen selbst sind bildschirmlesegerätfreundlich gestaltet und leiten den Benutzer nach der Auswahl direkt zur gewählten Region weiter.

Seite „Guided Safari" – The Core Experience

Die Seite „Guided Safari" bildet das Herzstück des VISUAL-Prototyps (Abbildung 4). Für jedes Ökosystem existiert eine eigene Unterseite, beginnend mit dem Präfix „guided-ecosystem", beispielsweise „guided-jungle" für den Dschungel. Die Aufteilung in separate Seiten ermöglicht eine gezielte Anpassung und Integration ökosystemspezifischer Informationen sowie der jeweiligen eingebetteten Livestreams. Darüber hinaus ist jede dieser Seiten farblich passend zum jeweiligen

Lebensraum gestaltet, wobei gleichzeitig auf hohen Kontrast geachtet wurde, um Barrierefreiheit sicherzustellen.

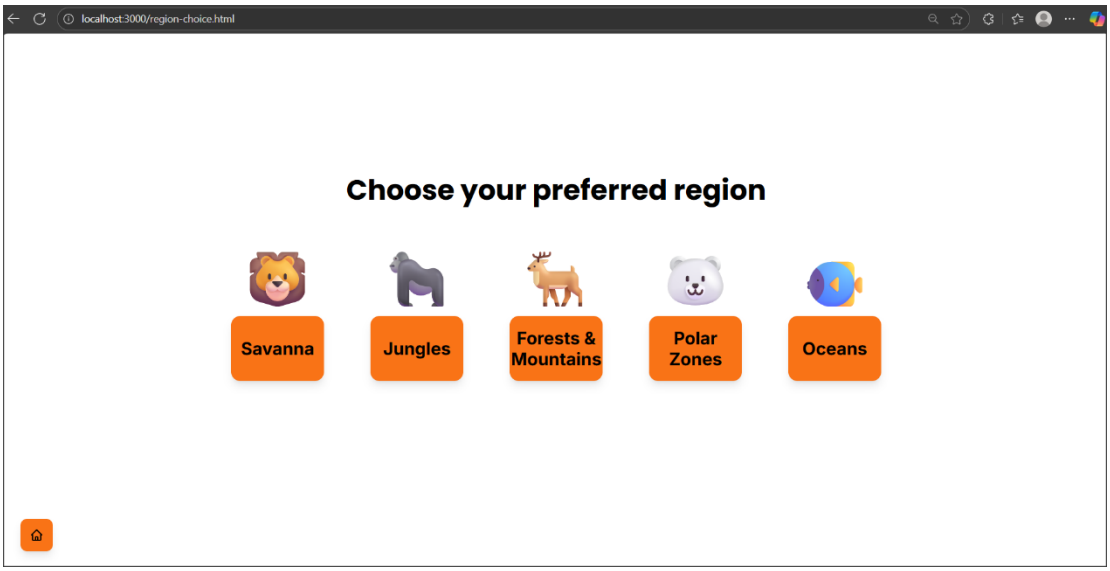

Abbildung 3: Choose by Region

Die Seite integriert den Livestream einer Wildtier-Webcam von Explore.org aus dem ausgewählten Ökosystem. Zudem steht dem Benutzer ein Anpassungspanel zur Verfügung, in dem das Profil sowie der gewünschte Modus gewählt werden können. Standardmäßig ist der Safari Adventurer vorausgewählt. Die Seite bietet mehrere Optionen zur Steuerung der Erzählung, darunter das Anpassen der Wiedergabegeschwindigkeit, Starten, Pausieren und Wiederholen. Zusätzlich kann der Benutzer optionale Fakten anfordern sowie zwischen Kameras oder Regionen wechseln – und das, ohne die Seite verlassen zu müssen. Alle Funktionen lassen sich über klar beschriftete Schaltflächen intuitiv steuern.

Auf der rechten Seite befindet sich ein Platzhalter für das Transkript der Erzählung („Safari Narrative Transcript“). Sobald die Erzählung erfolgreich generiert wurde, wird der Text dort angezeigt. Da die Sprachausgabe über die Web Speech API erfolgt, ist dieses Feld bewusst für automatische Bildschirmleser ausgeblendet, um ein doppeltes Vorlesen zu vermeiden.

Da das Erstellen von Screenshots und deren Analyse eine gewisse Zeit in Anspruch nehmen, wird der Benutzer über gesprochene Hinweise über den aktuellen Status informiert. Die erste Audioansage nach dem Klick auf „Erzählung starten“ lautet: „Safari wird vorbereitet“, nach erfolgreicher Aufnahme der Screenshots folgt „Tiere werden analysiert“. Das Ergebnis wird im erwähnten Fenster präsentiert.

Am oberen Seitenrand befindet sich zudem ein Indikator zur Serververbindung, der – ebenso wie das Analysepanel – ursprünglich zu Kontrollzwecken während der Entwicklung eingebaut wurde. Bei erfolgreicher Verbindung wird dort – grün hinterlegt und mit einem grünen, leuchtenden Punkt versehen – „Server: Online“ gemeldet.

Ergänzend zur Erzählung kann der Benutzer über den Button „Get Animal Facts“ zusätzliche wissensbasierte Informationen abrufen. Dabei wird ein RAG-Prozess ausgelöst, bei dem systemgestützt

faktenbasierte Inhalte aus Wikipedia ausgegeben werden. Diese Informationen sind auf das zuvor gewählte Altersprofil abgestimmt.

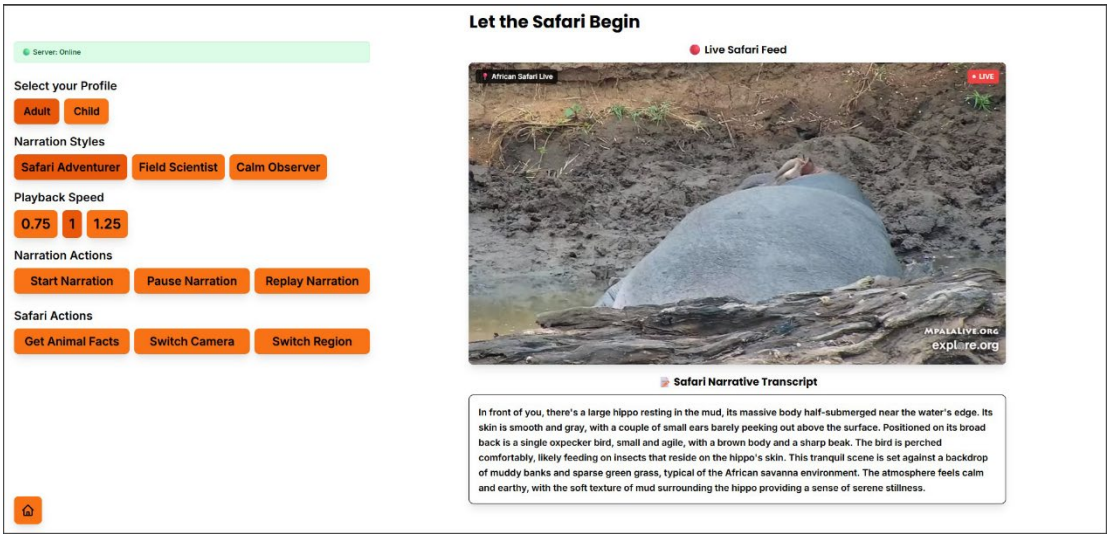

Abbildung 4: Guided Safari

3.7.4 Externe Tests

Anfang August 2025 fanden zwei externe Tests mit sehbehinderten Personen statt, die zusätzlich zu den internen Tests die volle Funktionsfähigkeit des VISUAL-Systems zeigen sollten (Jovic 2025). Aus Zeit- und Ressourcengründen war mehr nicht möglich. Es ist nicht das Ziel des Teams, ein Produkt zu entwickeln. Die Forschungsarbeit des Erstautors beinhaltet stets nur Prototypen, deren technische Grundlagen und Codes dann kostenlos zur Verfügung gestellt werden. Alle Tests verliefen unauffällig.

4 Ethische und soziale Diskussion

In diesem Abschnitt wird eine kurze ethische und soziale Diskussion zum VISUAL-System geführt. Aus Platzgründen können nicht alle Aspekte aufgenommen werden.

- Inklusion und Zugänglichkeit: VISUAL kann die Inklusion von blinden und sehbehinderten Menschen verbessern, indem es ihnen mehr Unabhängigkeit und Zugang zu Informationen bietet. Dies fördert eine gleichberechtigtere Teilnahme am gesellschaftlichen Leben. Dabei eröffnen sich Möglichkeiten, die es bisher nicht gab, nämlich eine individuelle, professionelle Tour, wie sie allenfalls speziell angeheuerte, fachkundige Assistenzkräfte hätten bieten können, die die Wildtier-Webcams aufrufen und die Szenen beschreiben.
- Privat- und Intimsphäre: Obwohl Wildtier-Webcams vor allem auf die Lebensbereiche von Wildtieren gerichtet sind, können sie in Einzelfällen auch Menschen erfassen und deren Privat- und Intimsphäre verletzen (Bendel und Zbinden 2024). Dabei wäre zu unterscheiden zwischen befugten Personen (wie Wildhütern) und unbefugten (wie Wilderern). Wildtier-Webcams und MLLMs könnten unter Umständen dabei helfen, Fälle von Wilderei aufzuklären, was dem befürchteten Schaden beim Datenschutz einen möglichen Nutzen bei der Kriminalitätsbekämpfung hinzufügen

würde. Ob die Tiere selbst in ihrer Privat- und Intimsphäre beeinträchtigt werden, wird in der ACI-Community diskutiert (Paci et al. 2022).

- Genauigkeit und Zuverlässigkeit: Die Zuverlässigkeit der KI bei der Beschreibung, Einordnung und Bewertung von Bildern ist wichtig (Bendel und Zbinden 2024; Bendel 2024b). Fehlerhafte oder ungenaue Informationen könnten zu einem unbefriedigenden Erlebnis und zu einer falschen Einordnung des Aussehens und Verhaltens von Tieren führen und damit einem Zweck des Systems entgegenlaufen. Die internen Tests der Zweitautorin von Juni bis Juli 2025 haben keine gravierenden Mängel zutage gebracht. Allerdings ist bei ihr auch kein tiefes Wissen zu Tieren und Tierverhalten vorhanden.
- Manipulation und Steuerung: Die Abhängigkeit von einem Anbieter oder Entwickler kann in diesem Zusammenhang problematisch sein, hier vor allem von OpenAI (Bendel 2024b). Dieser kann durch seine Beschreibungen und Bewertungen der Umwelt den Benutzer manipulieren und in eine bestimmte Richtung steuern. Es ist zudem seine Weltanschauung und Moral, die die Beschreibungen prägt und überhaupt Beschreibungen erlaubt und verhindert. Amerikanische Konzerne fallen diesbezüglich immer wieder durch Restriktionen und Zensur auf, etwa in Bezug auf Sexualität. Ob davon auch Tiersexualität betroffen ist, konnte nicht eruiert werden.
- Digitaler Graben: Es besteht die Gefahr eines digitalen Grabens, da nicht alle blinden und sehbehinderten Personen gleichermaßen Zugang zu den neuesten Technologien und Internetverbindungen haben (Bendel 2024b). Zwar kann die Anwendung kostenlos bezogen werden. Aber es kann sich nicht jeder ein Tablet oder Laptop mit geeignetem Netzzugang leisten. Letztlich ist eine virtuelle Videosafari dieser Art viel günstiger als eine echte vor Ort.

Zusätzlich zu diesen ethischen Überlegungen – von denen die meisten Technik- und Informationsethik betreffen – stellen sich weitere Fragen. Zwar können blinde und sehbehinderte Menschen nun von ihrem eigenen Wohnzimmer aus per Video auf Safari gehen, sie könnten jedoch auch alleine oder mit Unterstützung in die betreffenden Länder reisen und Tieren in der Realität begegnen, nicht nur virtuell, sondern auch über Geruch und Berührung. Kamphof (2011) hat betont, dass man das Gefühl hat, vor Ort zu sein. Dennoch fehlen Aspekte der Wahrnehmung und des Erlebens, bei eingeschränkten und nicht eingeschränkten Personen. Obwohl Forscher ein innovatives System entwickelt haben, um das Erlebnis während Online-Safari-Parktouren zu verbessern, kann letztlich nichts das tatsächliche Vor-Ort-Sein ersetzen (Tsuji et al. 2023). Es erscheint daher wichtig, VISUAL nicht als Ersatz, sondern als Ergänzung zu realen Erfahrungen zu verstehen.

Eine weitere ethische Überlegung betrifft die Tatsache, dass das System in erster Linie von sehenden Entwicklern entworfen wurde. Dies führt zwangsläufig zu Verzerrungen, die auf visuellen Konventionen beruhen und möglicherweise nicht mit den Lebensrealitäten blinder Menschen übereinstimmen. Um dem entgegenzuwirken, holte das Team den Rat eines blinden Mitarbeiters ein und führte eine Onlineumfrage durch, um Perspektiven und Bedürfnisse blinder Personen zu erfassen. Während Prompt Engineering und RAG ebenfalls eingesetzt wurden, um potenzielle Verzerrungen zu mindern, würde eine stärkere und kontinuierliche Einbindung blinder Entwickler die Inklusivität weiter verbessern.

5 Zusammenfassung und Ausblick

Dieser Beitrag stellte das VISUAL-Projekt vor, in dessen Rahmen ein Prototyp entwickelt wurde, der blinden und sehbehinderten Menschen virtuelle Videosafaris ermöglicht. Dabei wurde folgende

Forschungsfrage behandelt: „How can a multimodal large language model enable visually impaired individuals to experience a virtual safari utilizing public wildlife webcams, interpreting their pictures, and translating the content into descriptive audio narratives?“ Das Projekt lieferte theoretische Erkenntnisse, vor allem aber einen vollständig funktionsfähigen Prototyp.

Der VISUAL-Prototyp zeigt die technische Machbarkeit und das Potenzial der Verbindung von Inclusive AI und Animal-Computer Interaction. Bestimmte Einschränkungen müssen jedoch berücksichtigt werden, etwa die Abhängigkeit von kommerziellen MLLMs oder gelegentliche Ungenauigkeiten in den Beschreibungen. Das System könnte von der Einbindung zoologischer und biologischer Fachkenntnis sowie von Beiträgen von Ethologen und weiteren Experten für Wildtierverhalten und Ökologie profitieren. Verfahren wie RAG, die in diesem Projekt nur teilweise eingesetzt wurden, könnten die faktische Genauigkeit weiter verbessern. Eine weitere mögliche Weiterentwicklung ist der Einsatz neuer MLLMs wie GPT-5, das Anfang August 2025 veröffentlicht wurde. Schließlich würde eine öffentliche Zugänglichmachung des Systems Tests mit einer größeren Nutzerschaft ermöglichen und die Zusammenarbeit mit potenziellen Partnerorganisationen erleichtern.

6 Danksagung

Die Autoren danken Artan Llugaxhija von der Hochschule für Wirtschaft FHNW, der als sehbehinderter Mensch wesentlich dazu beigetragen hat, die Onlineumfrage barrierefrei zu machen, beide in die Nutzung von Screenreadern einzuführen und ihnen zu zeigen, wie Notebooks von blinden oder sehbehinderten Benutzern verwendet werden.

Literatur

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

Bitte zitieren Sie wie folgt:

Bendel, Oliver; Jovic, Doris. Im Schlamm ruht sich ein großes Nilpferd aus: Virtuelle Videosafaris für blinde und sehbehinderte Personen. In: Wiley Industry News, 10. November 2025. https://wileyindustrynews.com/de/fachbeitraege/im-schlamm-ruht-sich-ein-grosses-nilpferd-aus.