"This is My Fault", Really? Understanding Blind and Low-Vision People's Perception of Hallucination in Large Vision Language Models

Yilin Tang
Zhejiang University
Hangzhou, Zhejiang, China
22251376@zju.edu.cn

Yuyang Fang
College of Computer Science and Technology
Zhejiang University
Hangzhou, Zhejiang, China
fangyuyang@zju.edu.cn

Tianle Wang
School of Digital Media and Design Arts
Beijing University of Posts and Telecommunications
Beijing, China
wtl02@bupt.edu.cn

Lingyun Sun
International Design Institute
Zhejiang University
Hangzhou, China
sunly@zju.edu.cn

Liuqing Chen*
College of Computer Science and Technology
Zhejiang University
Hangzhou, China
chenlq@zju.edu.cn

Abstract

Visual question-answering (VQA) tools powered by large visual language models (LVLMs) are used to assist blind and low-vision (BLV) individuals in overcoming visual challenges, raising concerns about hallucinations and associated risks. Existing literature overlooks the variations of hallucinations across distinct usage scenarios and types in the context of VQA for BLV people, resulting in limited understanding of their perceptions and insufficient guidance for targeted mitigation strategies. By analyzing 3,467 real-world VQA cases from BLV users, we developed a manifestation-scenario-based dual-dimensional hallucination typology, uncovering eight scenarios and five types of hallucinations. Through interviews with 16 BLV users, we examined their awareness levels, detection strategies, mental models of hallucinations, and their tolerance of associated risks, identifying key gaps between their perceptions and real situations. By designing with 12 BLV users, we uncovered their expectations for hallucination-mitigating solutions, including enhanced information provision, transparency in processing, verification strategies, and feedback mechanisms.

CCS Concepts

• **Human-centered computing** → **Accessibility**; *Empirical studies in accessibility*;

Keywords

Large visual language models (LVLMs), Artificial intelligence (AI), Hallucination, Blind and low vision, Human-centered AI

ACM Reference Format:
Yilin Tang, Yuyang Fang, Tianle Wang, Lingyun Sun, and Liuqing Chen. 2025. "This is My Fault", Really? Understanding Blind and Low-Vision People's Perception of Hallucination in Large Vision Language Models. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28–October 01, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3746059.3747597

*Corresponding author.

UIST '25, Busan, Republic of Korea

ACM ISBN 979-8-4007-2037-6/25/09
https://doi.org/10.1145/3746059.3747597

1 Introduction

Visual Question Answering (VQA) is an essential task that can support blind and low-vision (BLV) individuals by providing detailed interpretations of visual content in response to their queries, aiding them in tackling various challenges in their lives [18, 33]. The traditional solutions for assisting BLV users with visual interpretation involve human-powered VQA systems, such as VizWiz [12], Aira [3], and Be My Eyes [26], which transmit images and questions to remote sighted assistants for responses. However, human-dependent solutions come with various drawbacks, including high labor costs, lengthy response times, and privacy concerns [89]. The recent advancements in Large Visual Language Models (LVLMs) are considered to hold potential in addressing these drawbacks [28], as they can understand real-world images with human languages and generate rapid responses to assist BLV users in VQA tasks [58, 70, 96]. For instance, Be My AI utilizes GPT-4 to automatically answer visual questions or provide image descriptions from BLV users' uploaded images [27, 65].

Although LVLMs increase accessibility, they can generate hallucinations [58], which are defined as nonsensical, unfaithful, and undesirable textual output given the image inputs in the context of VQA [32, 50]. It is inevitable [58–60] and even the advanced LVLMs such as InstructBLIP still contain an astonishing 30% of hallucinated text [32]. This is risky for BLV users who tend to

trust the model's outputs [61] and rely on these to make important decisions in life[29]. For example, the LVLMs may misidentify a power adapter's voltage as 220 volts instead of the correct 120 volts, and incorrectly respond that a beverage is alcohol-free when it is not. These examples illustrate the unique risks that LVLMs may pose in VQA for BLV users, raising concerns about how to mitigate hallucinations or reduce their negative impact on users [93].

Previous studies have investigated LVLM-related hallucination issues in general context, focusing on the types[9, 59, 69, 87], measurements [43, 44, 59, 80, 88], and mitigation of hallucinations [30, 45, 57, 70, 83, 96]. The evaluations of the hallucinations among these studies typically employ predefined visual queries and general image datasets[80], such as MS-COCO [56]. However, the complexity of hallucination-related issues increases in VQA tasks for BLV individuals due to their visual queries with high lexical diversity, coupled with images that often have quality problems, such as blurriness, poor lighting, and sub-optimal framing [33, 60]. Such complexity potentially makes the general hallucination literature unsuitable to this specific context. Hallucination-related challenges have been sporadically mentioned in prior literature when researchers developed visual assistance tools for BLV people or broadly examined GenAI performance across various visual assistive tasks, such as information retrieval [1, 78], image captioning [4, 78], OCR [4], and VQA [1, 4, 78]. Among these visual tasks, VQA has the unique requirement that outputs need to align with both image and text inputs [18], thus, a specific investigation of hallucinations on VQA tasks for BLV people is necessary. Among existing studies [1, 4, 78] that partially involve VQA tasks, Huh et al. [40] found that BLV evaluators failed to detect inaccurate long-form answers generated by LVLMs for BLV people's visual questions. Alharbi et al. [4] and Adnin et al. [1] interviewed BLV individuals regarding their use of AI-driven visual assistance technologies and uncovered some hallucination-related findings, including how BLV people detect and verify potentially inaccurate outputs. However, these studies regard hallucinations in VQA for BLV people as a monolithic phenomenon, overlooking notable variations across distinct usage scenarios and types of hallucination. These variations among hallucinations have been found to lead to different user perceptions and necessitate targeted measurement and mitigation strategies in the general context [35, 63]. Therefore, it is essential to investigate the occurrence and potential causes of hallucinations from a more granular perspective, and on this basis, further explore BLV perceptions of hallucinations across various types and scenarios. In addition, current research has explored identifying the BLV community's needs for AI explainability and contestability [4], which may help mitigate hallucinations, but there is still a lack of practical design and technical recommendations tailored to mitigate hallucinations for BLV users.

To bridge the gaps, this study aims to provide insights into the following research questions:

RQ1: What types and scenarios of hallucinations arise in the context of VQA for BLV individuals?

RQ2: How do BLV individuals perceive hallucinations and the risks associated with them?

RQ3: What kind of solutions are desired by BLV individuals to help them mitigate the influence of hallucinations?

We conducted three complementary studies to address these research questions. To answer RQ1, we employed a mixed-methods approach[64] to analyze VQA cases (3467 cases, including 619 hallucination cases) from the VizWiz dataset and our self-built dataset. We developed a manifestation-scenario-based dual-dimensional hallucination typology, uncovering eight scenarios and five types of hallucinations, and the distribution of different hallucinations across these scenarios. To address RQ2, we conducted semi-structured interviews with 16 BLV individuals, where we investigated how and through what strategies they perceive hallucinations across different hallucination types, their tolerance of the risks associated with hallucinations across various usage scenarios, and their mental models regarding the formation of hallucinations. To answer RQ3, we held two co-design sessions with 12 BLV people to investigate the solutions desired by them to mitigate the impact of hallucinations. Overall, this research contributes an in-depth understanding of the occurrence, causes, and BLV people's perceptions of hallucinations from the perspective of manifestations and scenarios in the context of VQA for BLV community. Additionally, to our knowledge, this work is the first to co-design hallucination mitigation strategies with the BLV community, offering firsthand practical insights that highlight the importance of the collective efforts of various stakeholders, including AI practitioners, developers and designers.

2 Related Work

2.1 Artificial Intelligence(AI)-powered Visual Assistance Tools for BLV Individuals

AI-powered visual assistance tools for BLV users include captioning and VQA tools. Captioning tools describe visual elements (e.g., colors, text, objects) in the image [34, 62, 67, 86], offering general information [75] but cannot provide tailored responses to user queries.

In contrast, AI-powered VQA tools can take both visual content and a corresponding question as input and return an answer automatically [75]. For instance, Be My Eyes [27, 65] launched Be My AI, a GPT-4-based virtual assistant that is designed to replace human volunteers to answer visual questions based on the images uploaded by BLV users. These AI-powered VQA tools enhance accessibility[46], are favored in privacy-related scenarios [76], and strengthen the sense of independence of BLV users[24].

Despite the above advantages, Zhao et al. [93] proposed a benchmark for evaluating LVLMs' performance on VQA tasks and found that the outputs of all six LVLMs are not well-grounded in reality. Huh et al. [40] found that BLV individuals often failed to perceive the inaccuracies in long-form answers generated by LVLMs in response to their visual questions. Similarly, BLV user feedback from their use of visual assistive tools for image creation[39], real-time visual description [17], and image exploration [48] also echoes their difficulty in assessing the accuracy of AI-generated responses. Moreover, Gonzalez et al. [31] found that BLV users have lower satisfaction and trust in the responses provided by AI models. Hallucinations, as an important and emerging issue brought by LVLMs, have sparked widespread concern.

2.2 Hallucination Taxonomies in LVLMs

LVLMs, which evolved from previous Vision-Language Pretrained Models (VLPMs) by integrating the advanced capabilities of Large

Language Models (LLMs), are adept at tackling complex tasks involving both vision and natural language [9, 58, 80]. Despite integrating the strengths of LLMs and VLPMs, LVLMs also inherit the hallucination from both [80], resulting in unprecedentedly complex forms of hallucination.

AI hallucinations have yet to solidify into a universally agreed-upon definition and the boundary between AI hallucinations and AI errors or mistakes remains contested [77]. Research in computer vision strictly defines LVLMs' hallucinations as the misalignment between visual content and textual output[58, 81], while the HCI literature generally adopts a broader definition, defining them as nonsensical, unfaithful, or undesirable outputs [32, 50]. VQA tasks for BLV users require LVLMs to recognize visual content, understand the user's query, and reason over both. In this context, the strict definition, which focuses solely on mismatches between the output and visual content, overlooks cases where the output mismatches the user's textual query intention or the correct reasoning result[80]. This oversight is critical, as BLV people cannot accurately detect hallucinations, and hallucinations pose severe risks to them [40]. Thus, this paper builds upon the latter broader definition to cover a wider range of negative impacts on BLV users. To draw a clear line between hallucinations and stylistic or other subjective preferences, we constrain the notion of "nonsensical and undesirable outputs" in the broader definition by focusing on those outputs that fail to fulfill the VQA task's goal. For instance, an "undesirable" overly verbose response that helps a BLV user correctly understand the visual content, is not considered a hallucination.

A comprehensive taxonomy is vital for guiding AI development [47]. Existing studies typically categorize hallucinations in LVLMs either phenomenally or mechanically [87]. The former approach classifies hallucination based on the outcomes' manifestation, while the latter focuses on training and deployment methodologies[87]. Phenomenally, hallucinations can be classified as external (erroneous inferences based on image data) and internal (inconsistencies with image data), with internal hallucinations further divided into objects, attributes, multi-modal conflicts, and counter-commonsense hallucinations [59]. Additionally, based on visual semantics, hallucinations can be grouped into objects (nonexistent or incorrect object categories), attributes (incorrect color, shape etc.), relationships (such as incorrect human-object interactions), and events (fictional narratives about objects) [9, 42]. Rawte et al. further categorize hallucinations into eight types, including identity incongruity, geographic erratum, visual illusion, mild gender anomaly, VLM as classifier, wrong reading, and numeric discrepancy [69]. Mechanically, hallucinations can be attributed to components like the visual encoder, modality connection module, and LLM [58]. These taxonomies provide a structured foundation for analyzing and mitigating hallucinations in LVLMs.

However, existing taxonomies fall short in the context of VQA for BLV individuals, as they overlook the unique challenges in this context, where BLV people may struggle to articulate precise questions that accurately describe their needs and capture clear images[33]. In that case, LVLMs are required to provide open-ended responses to low-quality images and ambiguous visual questions, and these two factors may lead to a higher incidence of hallucinations [44, 60]. Thus, more in-depth research is needed to create a taxonomy specific to this context to fully understand the unique hallucinations encountered by BLV users when utilizing LVLM-based VQA solutions.

2.3 Mitigate the Negative Impact of Hallucination

Existing studies proposed various techniques to mitigate different types of LVLMs' hallucinations in general contexts. To mitigate "object hallucination" as identified in existing taxonomies, methods such as LVLM Hallucination Revisor (LURE) [96], Visual Contrastive Decoding (VCD) [51], Data Augmented Contrastive Tuning [74], Mitigating Hallucination via Classifier-Free Guidance (MARINE) [92], Hallucination Reduction via Adaptive Focal-Contrast Decoding (HALC) [19] were introduced. To evaluate "relationship hallucination", Wu et al. [85] proposed a benchmark called "R-Bench". In addition, a suite of strategies has also been employed, including formal methods guided iterative prompting [41], counterfactual inception[45], instruction tuning [25], interactive question-knowledge alignment [90], and "chain of knowledge" [54].

Despite these technological efforts, hallucinations cannot be entirely eliminated and continue to threaten users [32]. Thus, other studies focus on addressing hallucinations' adverse impact on non-BLV users from a human-centered perspective[38]. Educating users to enhance their awareness of hallucinations has been found to contribute to increasing their trust in AI systems [14]. Leiser et al. [49, 50] suggested displaying confidence thresholds, sources, and ethical considerations to help users identify LLMs' hallucinations and reduce over-reliance. Cheng et al. [20] designed RELIC, an interactive system that helps users identify LLMs' hallucinations in generated content by investigating factual consistency across multiple responses.

Beyond the technical and non-technical mitigations developed for general contexts, recent work initially explored how BLV users perceive hallucinations. Adnin et al.[1] investigated BLV users' overall adoption and comprehension of generative-AI tools and identified BLV users' unique verification strategies for hallucinations, including leveraging prior knowledge, answer consistency, and cross-checking with other information sources. Additionally, Alharbi et al. [4] further supplemented that BLV individuals also use non-visual senses to validate hallucinations and highlighted their needs for AI explainability and contestability. Despite these insights, there remains a lack of practical design or technical recommendations specifically tailored to help BLV individuals mitigate the negative impacts of hallucinations. Therefore, further exploring BLV users' perceptions of different types of hallucinations across various scenarios and developing targeted mitigation strategies are crucial for enhancing BLV users' trust in and experience with AI systems.

3 Study 1: Understanding Hallucinations in the Context of VQA for BLV Individuals: Types, Scenarios, and Proportions

3.1 Method

To comprehend the potential hallucinations that may arise when BLV people use LVLMs for VQA, we employed GPT-4o, the current state-of-the-art LVLM[21] (with prompts refer to the Supplementary materials), to generate answers for real-world VQA cases from BLV people and then analyzed the hallucinations present in the answers. All VQA cases analyzed in this study originated from VizWiz and our self-built dataset. VizWiz is a popular dataset that

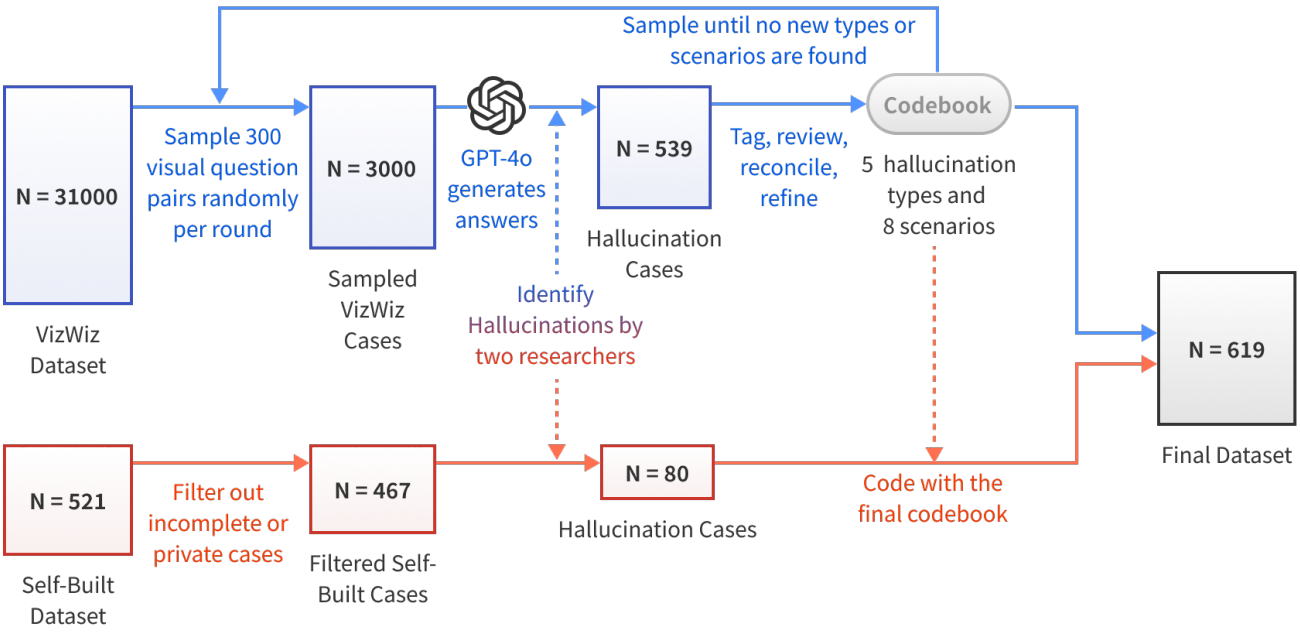


Figure 1: We randomly sampled 3,000 cases from the VizWiz dataset, which includes visual questions with images sent by BLV individuals to human volunteers. We used GPT-4o to regenerate the answers of these VQA cases. Two researchers reviewed these responses, identified 539 hallucination cases, and developed a codebook to classify hallucination types and scenarios. For our self-built dataset (which includes visual questions with images from BLV individuals and GPT-4o-generated answers), incomplete and private cases were filtered out, leaving 467 valid cases. Of these, 80 hallucination cases were identified and categorized using the same codebook developed from the VizWiz dataset. The two datasets were then combined, resulting in a final dataset of 619 hallucination cases.

includes over 31,000 real-life VQA cases sourced from BLV people worldwide [33]. Each VQA case in VizWiz includes a visual question with an image posed by BLV people to remote human volunteers, and 10 human-generated answers. For these cases, we discarded all human-generated answers and regenerated answers using GPT-4o based on the visual questions and images, focusing on investigating potential hallucinations within these GPT-generated answers. However, visual questions in VizWiz still reflect the questioning approaches and usage scenarios of BLV individuals when seeking visual assistance from humans, which may differ from their approaches when requesting help from LVLMs [31]. Such differences in visual questions could potentially impact the performance of LVLMs in answering. Thus, we supplemented a self-built dataset that represents the VQA cases posed by BLV people to LVLMs. We employed inductive qualitative content analysis[7] for these VQA cases, as depicted in Figure 1.

3.1.1 VizWiz Dataset. To balance the diversity of cases and the practicality of manual analysis, we used multiple rounds of random sampling, sending 300 pairs of visual questions with photos from the VizWiz dataset to GPT-4o for answers each round [91]. For each round of sampling, two researchers independently reviewed and identified whether the answers contained hallucinations according to hallucination's definition [80]. Afterwards, they were instructed to code the hallucinations based on the external phenomenon rather than their underlying mechanisms (as detailed in Section 2.2) [87] to ensure a user-centered perspective. Specifically, they inductively assigned a word or short phrase to the external phenomena of each hallucination case and wrote analytic memos to capture initial concepts. They then collectively discussed these initial concepts, reconciled coding discrepancies, and reorganized them through an iterative process to form initial codes [84]. The above sampling and encoding process is repeated until no new types of hallucinations appear and saturation is reached[91]. We sampled a total of 3,000 cases from the VizWiz dataset over 10 rounds, and identified five types of hallucinations in the final codebook (see Table 2). Based on the final codebook, the two researchers independently re-coded all cases and achieved high consistency, with Cohen's Kappa greater than 0.8 for each type of hallucination (the Cohen's Kappa values are provided in Supplementary materials).

Given the importance of creating a typology by synthesizing real-world scenarios in AI ethics [68], we used the affinity diagramming method [11] to summarize the scenarios where hallucinations occur. The researchers placed the cases on sticky notes, grouped them, and iteratively labeled each group with descriptors to capture their theme. Ultimately, eight scenarios were identified (the codebook is shown in Supplementary materials), encompassing the majority of contexts where BLV individuals utilize assistive AI techniques [29].

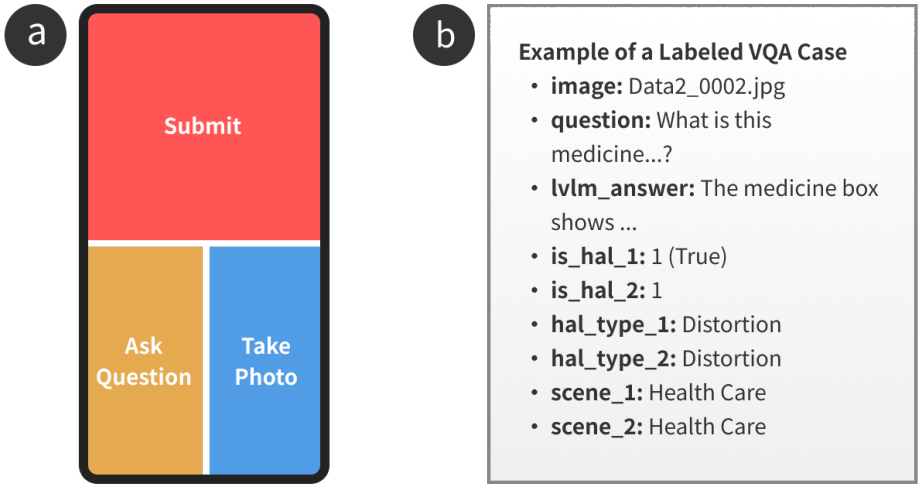


Figure 2: (a) AskVision user interface, where users can record a spoken question by tapping the "Ask Question" button, capture an image with the "Take Photo" button, and submit both by pressing the "Submit" button. The system ensures that both the audio and the image are provided before allowing submission. (b) An example of a labeled VQA case includes an image from BLV users, the corresponding question, the answer generated by the LVLM, and annotations from two independent researchers. These annotations include whether the answer contains hallucinations (is_hal), the type of hallucination (hal_type), and scene classification (scene), with each label provided separately by the two researchers. For example, scene_1 represents the annotation of researcher 1 for the scenario identification.

3.1.2 Self-built Dataset. To gain a more comprehensive understanding, we invited 16 BLV participants to engage with "AskVision", our developed LVLM-based VQA tool to build a dataset.

Participant Recruitment: We recruited participants through snowball sampling from BLV forums and local BLV community centers with the following criteria: (1) aged 18 or above, (2) experiencing blindness or low vision. As shown in Table 1, 16 participants (7 men, 9 women), aged from 22 to 71 with various professions were included in this study. All participants were familiar with visual access technologies such as Be My Eyes [26] and iOS VoiceOver [6], and six of them had experience with LVLMs for visual assistance, including Dou Bao [16] and ChatGLM [94]. All participants are native speakers of Mandarin.

AskVision: Considering the data collection challenges and usability issues stemming from complex functionalities in existing LVLM-based assistants[16, 94], we developed AskVision, specifically focused on VQA tasks for BLV users, featuring only three buttons on the interface for easy operation, as shown in Figure 2 (a). It enables users to ask questions through voice input in Chinese on their phones when taking photos, leveraging GPT-4o's API [66] to generate responses, and utilizing Baidu Smart Cloud's text-to-speech services for audio feedback [23]. Additionally, AskVision is developed following W3C guidelines to ensure its compatibility with mainstream screen readers.

Procedure: This study is conducted in a Disabled Persons' Federation (DPF) facility in a city in southern China. We first explained the study's purposes and procedures. Participants were then guided through downloading AskVision onto their phones and received step-by-step instructions on its utilization. Subsequently, participants had 15 minutes to explore AskVision, with any issues addressed in real-time. Once ready, participants commenced a one-week experience with AskVision, tasked with solving various visual problems they encountered in real-life scenarios. The conversations between participants and AskVision were all in Chinese. During this, they were asked to use their phone's voice memo feature to record as many instances as possible where they suspected the LVLM's responses to deviate from reality, indicating potential hallucinations. They were informed that all conversation logs would be recorded in the backend and were advised to avoid capturing personal privacy information. Our study was approved by the Institutional Review Board (IRB) and obtained informed consent from all participants.

Data analysis: The VQA cases recorded in conversational logs were incorporated into our self-built dataset. A total of 521 cases were collected, each comprising an image, a question, and a response from GPT-4o. Three researchers screened the data, excluding cases that were submitted incompletely and those containing private information, leaving 467 cases. Two researchers encoded these cases using the coding scheme employed for the VizWiz dataset, achieving Cohen's Kappa greater than 0.8. As a result, no new scenarios or hallucination types were identified, such commonality between the two datasets will be discussed in Section 3.4. Additionally, the hallucination cases identified and recorded by the participants will be analyzed in Study 2.

Table 1: Demographics of participants

ID	G.	Age	Edu.	Occupation	Diagnosis	Onset	Vis.Exp.	Vis. Acc.(Tools)	AI. Acc. (Tools)
P01	W	51	H.S.	Masseur	Retinitis Pigmentosa	31	ATB	3 (V. O., Be My Eyes)	1
P02	M	60	J.H.S.	Community Coordinator	Trauma	21	ATB	3 (V. O., Be My Eyes)	1
P03	W	42	H.S.	Customer Service	Optic Neuropathy	23	AB-LC	3 (V. O.)	1
P04	W	71	P.S.	Retired	Glaucoma	0	CTB	3 (TBack)	1
P05*	M	61	H.S	Retired	Retinal Detachment	16	ALV	3 (TBack)	3 (DouBao, SeeingAI)
P06*	W	28	B.E	Liberal Professions	Hereditary Retinal Diseases	10	AB-LC	3 (V. O., NVDA)	3 (Be my AI, ChatGPT-4V)
P07*	M	22	H.S.	Masseur	Glaucoma	0	CTB	2 (V. O.)	2 (ChatGLM)
P08*	W	49	J.H.S.	Masseur	Trauma	20	ATB	1	1
P09*	M	32	B.A.	Piano tuner	Retinoblastoma	0	CB-LC	3 (V. O.)	1
P10*	M	40	B.E.	Programmer	Optic Atrophy	0	CTB	3 (V. O., NVDA)	2 (ChatGLM)
P11*	W	53	M.A.	Professor in art	Neuromyelitis Optica	33	ALV	3 (V. O.)	2 (DouBao)
P12*	M	37	B.E.	Online sales	Retinitis Pigmentosa	0	CB-LC	3 (V. O., Be My Eyes)	2 (DouBao)
P13*	W	70	J.H.S.	Retired	Diabetic Retinopathy	45	ATB	2 (TBack)	1
P14*	W	69	H.S.	Retired	Glaucoma	58	AB-LC	2 (TBack)	1
P15*	W	53	B.A.	Soprano of a choir	Cataract	0	CTB	2 (V. O.)	1
P16*	M	70	P.S.	Fortune teller	Trauma	3	ATB	3 (V. O.)	1

Note: G=Gender (M=Man; W=Woman). Edu=Education (P.S.= primary school; J.H.S.= Junior High School; H.S.=High School; B.E.= Bachelor in Engineering; B.A.=Bachelor in Arts; M.A.=Masters in Arts). Vis Exp=Visual Experience (CTB = Congenital Total Blindness: No visual cues or direct visual experience with images; CB-LC = Congenital Blindness with some light/color perception: No direct experience with images; ATB = Acquired Total Blindness: Prior experience with images; AB-LC = Acquired Blindness with some light/color perception: Prior direct experience with images; ALV = Acquired Low-Vision: Prior experience with images). Vis./AI. Acc.= Visual/AI Access Technologies. Vis. Acc. Tools (V.O.=iOS Voice Over; TBack=Android TBack; NVDA=Windows NonVisual Desktop Access;). Experience (1-basic knowledge; 2-limited experience; 3-practical application; 4-applied theory; 5-recognized authority). All participants took part in study 1 and 2, those marked with * were also involved in study 3.

3.2 Hallucination Types Identified in VQA cases

We combined 3000 cases extracted from VizWiz with 467 cases collected from our self-built dataset, forming a final dataset consisting of 3,467 VQA cases (detailed in supplementary materials) and Figure 2 (b) shows an example of a labeled VQA case in it. In the final dataset, we identified 619 cases involving hallucination cases, 539 from VizWiz and 80 from our self-built dataset, resulting in occurrence probabilities of 17.97% and 17.13%, respectively. This highlights that even advanced LVLMs still have a notable probability of generating hallucinations.

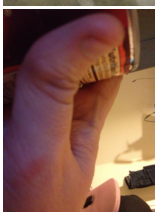
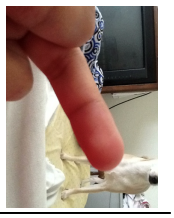
According to our coding results, we introduced a taxonomy of hallucinations based on their observable manifestations, categorizing them into five types: Distortion, Fabrication, Irrelevance, Fallacy and Misfocus, as shown in Table 2. Distortion manifests as LVLMs provide answers that are inconsistent with the verifiable information in images for questions that could be answerable based on the images. Fabrication is characterized by LVLMs making up answers when the image information is insufficient for the questions. Irrelevance is marked by LVLMs' answers that do not conflict with the image but are off-topic. Fallacy manifests as LVLMs answering incorrectly when they need to reason based on information in the image. Misfocus is characterized by LVLMs seemingly attempting to answer the input question but erroneously focusing on parts of the image that are unrelated to the intent of the question.

Among them, Distortion and Fabrication were found to cover existing visual-semantic-based hallucination taxonomies [9], which include object, attribute, relationship, and event hallucinations (as described in Section 2.2). In addition, we identified three new categories of hallucinations that are seldom mentioned in the existing taxonomies, including Irrelevance, Fallacy, and Misfocus. Through case review, we found that Irrelevance arises when the questions posed are less relevant to the images provided, and Misfocus is often linked to vague questions (as shown in Table2). These two types of hallucinations reflect the visual challenges faced by BLV individuals during VQA, as they are unable to see the items they are interested in and therefore cannot accurately capture and ask questions. Fallacy emerges when BLV people pose a range of image-based reasoning questions to LVLMs that exceed LVLMs' reasoning capabilities [36]. It reflects that BLV people's reasoning needs in real-world scenarios using VQA go beyond mere image recognition in traditional hallucination assessment tasks[69].

3.3 The Potential Causes of Hallucinations

Considering that visual questions and images are the only two inputs in the VQA task, we investigated the impact of content word numbers in visual questions and the quality of images on different types of hallucinations. We employed the method used by Bird et al.[13] to calculate the number of content words-referring to nouns, verbs, adjectives, and adverbs-in the visual questions. For our self-built dataset, the original Chinese questions were first translated into English using GPT-4o-mini (with prompts see Supplementary materials) and then manually reviewed to ensure the accuracy of the translation. A Kruskal-Wallis test revealed significant differences in the number of content words in visual questions across different types of hallucination ($H=47.61$, $p < 0.01$). Further Dunn's post hoc tests showed that the number of content words in Fabrication cases (Median=3.00, Mean=3.62) was significantly higher (Cliff's Delta = -0.18, $p < 0.01$) than those in no-hallucination cases (Median=2.00, Mean=3.10) and also significantly higher (Cliff's Delta= -0.37, $p < 0.01$) than those in Misfocus cases (Median=2.00, Mean=2.27). This suggests that more informative visual questions are prone to result

Table 2: Taxonomy of LVLMs' hallucination in the context of VQA for BLV people.

Hallucination	Definition	Case	
Distortion	When presented with an image that **CONTAINS** the fact capable of answering the question, LVLMs provide answers that formally adhere to the structure of the question but are **INCONSISTENT** with the facts.	Q: What color is this? A: Blue and white. *Although this response adheres to the structure of the question by addressing color, it is inconsistent with the fact in the image that "light purple with dark purple patterns".*	
Fabrication	When providing an image that does **NOT CONTAIN** the fact capable of answering the question, LVLMs provides answers that formally adhere to the structure of the question, but **FABRICATE** the fact.	Q: What's the sodium content? A: 70 mg. *The image does not contain the sodium content. This response adheres to the question' structure but fabricates the fact instead of truthfully stating "Unknown".*	
Irrelevance	LVLMs provide answers that do **NOT ADHERE to** the structure of the question.	Q: What is the layout of this fridge? A: There are various colorful magnets on the fridge door. *This response does not conform to the structure of the question.*	
Fallacy	When **REASONING** is required based on the facts in the image, the answer provided by LVLMs formally adheres to the question' structure but is **INCORRECT**.	Q: What region does this license plate from? A: The license plate displays "浙A," indicating that it is from Ningbo, Zhejiang Province, China. *Although this response structurally fits the question and correctly identifies the content on the license plate, it incorrectly infers that "浙A" represents Ningbo while it represents Hangzhou.*	
Misfocus	When presented with an image that **CONTAINS** facts to answer an **AMBIGUOUS** question, LVLMs provide answers that adhere to the question' structure but Misfocus on the fact, failing to align with the **INTENTION** of the question.	Q: What is this? A: The finger. *Although this response structurally fits the question and has no inconsistencies with the image, it mistakenly focuses on the finger when the expected answer is the dog (all human volunteers answered correctly).*	

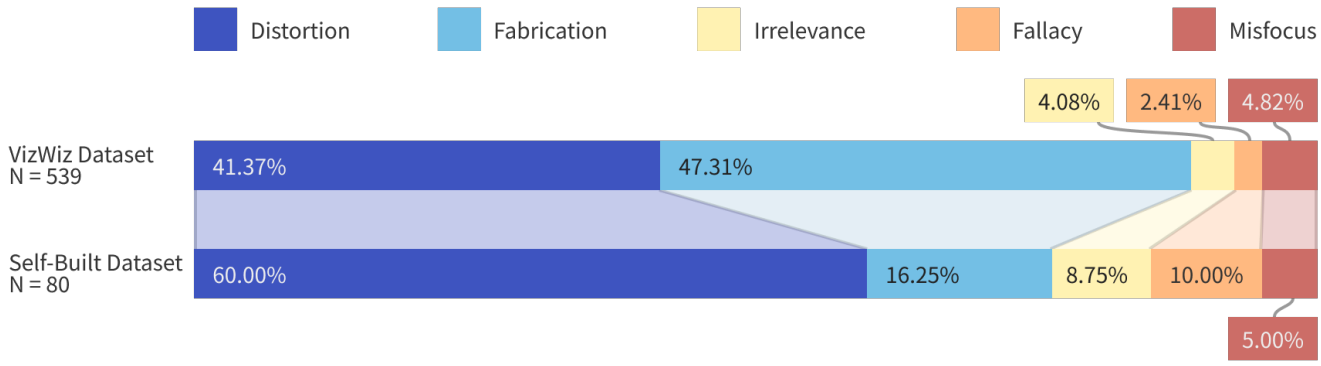


Figure 3: Proportions of Hallucination Types Across VizWiz and Self-Built Datasets

in Fabrication, while less informative questions are more likely to result in Misfocus.

To evaluate image quality, two researchers used the criteria from VizWiz[22] to label each image in the self-built datasets with perceived quality flaws, including no-flaws, blur, overexposure (bright), underexposure (dark), improper framing, obstructions, rotated views and unrecognizable. We further illustrated Figure 4 to demonstrate the relationship between image quality flaws and types of hallucinations. It shows that images with framing and unrecognizable issues predominantly led to Fabrication, while those with blur, bright and dark issues mostly resulted in Distortion. Notably, even images without quality issues have a 10.36% chance of causing hallucinations.

Based on the above results and manual case review, we observed that when facing images that are blurry or dark, LVLMs are prone to misinterpret or incorrectly identify parts of the images that could answer the visual questions, leading to Distortions. When confronted with informative visual questions that require extensive detail or facing images with framing or unrecognizable issues, LVLMs are unable to extract the necessary detail from the images and have to fabricate information to fill these gaps, leading to Fabrication. Low informative questions that are likely to be ambiguous, can misguide LVLMs to concentrate on the wrong objects within the image, leading to Misfocus.

3.4 Comparing Results from VizWiz and Self-Built Datasets

The types of hallucinations in VizWiz and self-built datasets were found to be the same, which may indicate that the overall demand for visual assistance from BLV individuals remains constant, regardless of whether they are interacting with humans or LVLMs, as well as the robustness of our taxonomy. We further calculated each hallucination type's probability (P_T), by determining the ratio of cases of this type to the total number of hallucination cases in both datasets, as shown in Figure 3. However, a Chi-square test showed a significant difference (χ^2=36.03, p<0.01) in the distribution of hallucination types between the two datasets. Post hoc tests with Bonferroni's corrections indicated that there were significant differences in the probabilities of Distortion, Fabrication, and Fallacy (all $p < 0.01$).

To elucidate these differences, we employed the methods from Section 3.3 to assess image quality and the number of content words of visual questions in both datasets. A Mann-Whitney U test showed a significant difference in the number of content words (U=498125.00, p<0.01) between VizWiz (median=3.00, IQR=2.00) and self-built datasets (median=1.00, IQR=2.00), indicating that BLV individuals pose more concise and less informative questions to LVLMs.

For image quality, the Chi-square test showed significant differences (χ^2=254.54, p<0.01) in the distribution of image quality flaws between the two datasets. Further post hoc tests with Bonferroni correction revealed the proportion of "no-flaws" images in VizWiz (15.63%) was significantly lower (χ^2=110.13, p<0.01) than those in self-built dataset (35.97%), suggesting that BLV individuals are more likely to pose images without quality flaws to LVLMs. Additionally, the results indicated that the self-built dataset contains a significantly higher proportion of images with darkness (VizWiz: 11.97%, self-built: 17.99%; χ^2=12.61, p<0.01) issues and a significantly lower proportion of images with framing (VizWiz: 7.07%, self-built: 2.78%; χ^2=11.52, p<0.01) issues compared to the VizWiz dataset.

Building on the findings from Section 3.3, the significantly lower probability of Fabrication in the self-built dataset compared to VizWiz can be explained by the fact that BLV individuals, when engaging with LVLMs, ask questions that are less demanding of intricate details and provide photos with better framing. These behaviors may decrease the number of instances where LVLMs need to resort to fabricating details to respond to queries. While the significantly higher probability of Distortion in the self-built dataset could be attributed to the increased presence of images with darkness issues in it, which predominantly contribute to Distortion, as shown in Figure 4.

3.5 Scenarios of Hallucination

We identified eight scenarios in which hallucinations occur: House Maintenance, Cooking and Dining, Healthcare, Finance, Entertainment, Work and Education, Outdoor and Dressing. We further determined the probability distribution of five types of hallucinations across these scenarios, as detailed in Table 3. The probability of each scenario in a dataset(P_S) is calculated as the ratio of the number of cases in that scenario to the total number of cases in the dataset. The probability of hallucination in a specific scenario (P_{HS}) is calculated as the ratio of hallucination cases in that scenario to the number of cases in that scenario. The probability of a specific type of hallucination in a given scenario (P_{HT}) is calculated as the ratio of that type's cases in the specific scenario to the total number of hallucination cases in the same scenario. It is worth noting that, due to the limited number of hallucination cases in our self-built dataset, calculating the P_{HT} may be susceptible to the influence of outliers or random events. Thus, only the results from the VizWiz dataset are presented in Table 3.

Table 3 shows that the higher probability of a scenario does not necessarily correlate with a higher probability of hallucinations in that scenario. We further calculated the five most frequent words in visual questions for each scenario, as shown in Table 4, which implies that the visual tasks vary across scenarios and this may account for the varying distribution of hallucinations in different scenarios. Finance, and Cooking and Dining are the two scenarios where LVLMs are most prone to hallucinations. Examination of cases reveals that tasks in these scenarios require the extraction of details from intricate images, such as discerning the denomination on crumpled currency and reading ingredients from reflective food packaging. Meanwhile, the diversity of food and currency across different regions might make it difficult for the LVLMs to recognize items according to their outdated and insufficient training data [80]. For example, the Chinese pancake was mistakenly identified as the French crepe. These complexities can lead to difficulties for LVLMs to accurately interpret and process the information, increasing the likelihood of hallucinations.

Across all scenarios, Distortions and Fabrications are the most common types of hallucinations. A higher probability of Fabrications is observed in Cooking and Dining, Healthcare, and Entertainment scenarios, where LVLMs need to analyze detailed information such as flavors and names, as shown in Table 4. On the other hand,

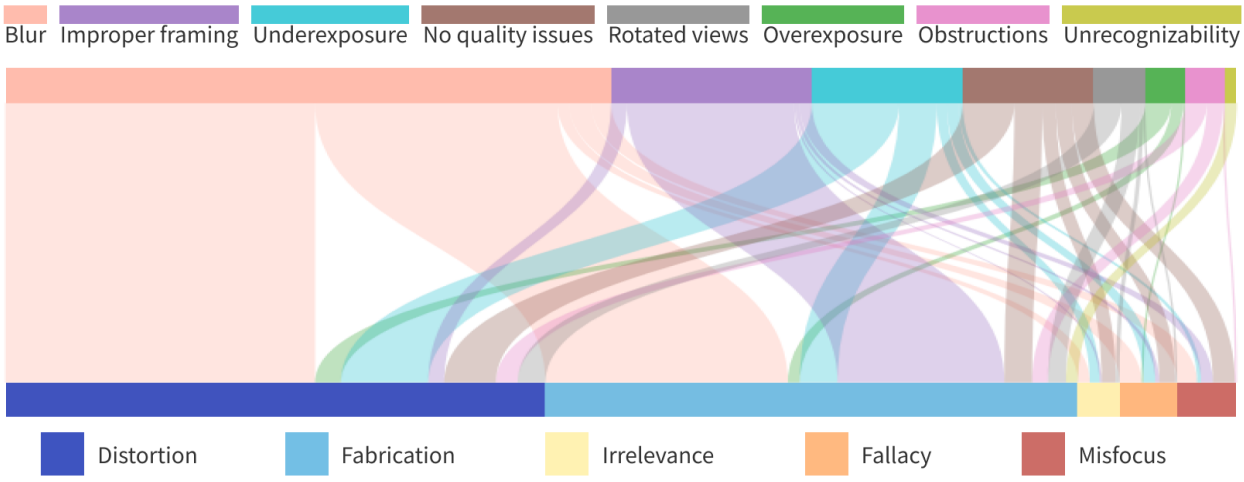


Figure 4: The Sankey diagram visualizes the contribution of image quality issues to different hallucination types.

in House Maintenance, Finance, Work and Education, and Dressing scenarios, where LVLMs are required to describe the basic characteristics of objects including their kinds and colors, Distortions are more likely to arise. Fallacies are most concentrated in Finance scenarios, followed by Work and Education scenarios. This implies that complex tasks requiring multi-step reasoning, such as calculating bills, interpreting documents, and providing medication guidance, pose challenges to LVLMs. It is worth emphasizing that the 0.00% in Table 3 indicates this hallucination type within the given scenario with zero occurrences across the 3000 cases examined. Nonetheless, this does not eliminate the potential for its occurrence.

In summary, the uneven distribution of hallucinations suggests that the performance of LVLMs depends on the usage scenarios and may be related to the characteristics of images and task requirements within those scenarios. Thus, it is imperative to devise tailored strategies for various usage scenarios to mitigate the adverse effects of hallucinations on BLV users, and these will be discussed in Section 6.

4 Study 2: Understanding BLV Individuals' Perceptions of Hallucinations and the Risks Associated with Them

4.1 Method

To understand BLV individuals' perceptions of hallucinations and associated risks, we conducted semi-structured interviews with 16 participants who experienced AskVision for a week in Study 1. Approved by the IRB and with informed consent from participants, study 2 was conducted at the DPF facility mentioned earlier.

4.1.1 Interviews. We checked the conversation logs between each participant and AskVision and conducted one-hour interviews based on the logs, using questions presented in the supplementary materials.

We first asked them to share the hallucination cases they detected and recorded during the experience process and their detection strategies. Afterwards, we inquired about their concerns regarding the risks associated with hallucinations, and then collaboratively summarized the risks in eight scenarios by sharing typical hallucination cases from the final dataset with the participants. Finally, participants were prompted to verbally describe their mental models regarding how hallucinations occur. With consent, over 18 hours of interviews were audio-recorded. Additionally, by presenting typical hallucination cases of different types and scenarios summarized in Study 1, we utilized two 5-point Likert scales to gauge BLV users' awareness levels of five hallucination types ("I am aware of this type of hallucination") and their tolerance levels of risks in eight scenarios ("I can tolerate the risk of hallucinations in this scenario"). Responses for both scales ranged from "Strongly Disagree" to "Strongly Agree."

4.1.2 Data Analysis. The audio recordings from the interviews were transcribed and divided into two approximately equal-sized sets for theme analysis. Three researchers independently analyzed the first set to generate codes, which were furthure refined into a codebook during collaborative sessions. These codes were applied to the second set, achieving high reliability (Cohen's Kappa > 0.84). The 5-point Likert Scale Ratings were used for statistical analysis to assess participants' awareness and tolerance for hallucinations and associated risks.

4.2 BLV Individuals' Awareness of Hallucinations

4.2.1 Overview of the Awareness of Hallucinations. Comparing hallucination cases from logs and participants' reports, less than one-third of the hallucination cases were detected by the participants, highlighting their insufficient awareness. For instance, P08 firmly believed that she only encountered two hallucination cases, but in reality she experienced seven. Even when we mentioned the LVLMs' mistake of describing her white jade pendant as green, she was surprised, "*Oh my god, I really thought it was green.*"

Deceived by LVLMs' rhetoric: 11 of 16 participants were attracted by the LVLM's "*well-organized* (P13)" and "*richly detailed*

Table 3: Probability distribution of five types of hallucinations across eight scenarios

Scenarios	P_S		P_{HS}		Hallucination (P_{HT})				
	in VizWiz	in self-built	in VizWiz	in self-built	Distortion	Fabrication	Irrelevance	Fallacy	Misfocus
House Maintenance	21.60%	32.12%	15.99%	14.77%	**45.63%**	37.86%	6.80%	0.00%	9.71%
Cooking and Dining	34.07%	14.13%	19.49%	30.30%	27.14%	**64.82%**	2.01%	2.01%	4.02%
Healthcare	7.90%	14.35%	18.49%	14.71%	40.91%	**47.73%**	4.55%	4.55%	2.27%
Finance	1.63%	2.36%	30.00%	54.55%	**46.67%**	26.67%	6.67%	20.00%	0.00%
Entertainment	10.07%	3.00%	20.46%	14.29%	38.71%	**50.00%**	4.84%	1.61%	4.84%
Work and Education	10.43%	9.85%	16.50%	10.87%	**59.62%**	21.15%	7.69%	5.77%	5.77%
Outdoor	3.43%	15.63%	11.54%	19.18%	41.67%	**50.00%**	8.33%	0.00%	0.00%
Dressing	10.87%	8.57%	15.95%	2.50%	**71.15%**	26.92%	0.00%	0.00%	1.92%

Note: P_S is calculated as the ratio of the number of cases in a specific scenario to the total number of cases in the dataset. P_{HS} is calculated as the ratio of the number of hallucinations in a specific scenario to the total number of cases in that scenario. P_{HT} is calculated as the ratio of the number of a specific hallucination type in a scenario to the total number of hallucinations in that scenario.

Table 4: High-Frequency Words and Their Frequencies in Eight Scenarios

House Maintenance	Cooking and Dining	Healthcare	Finance	Entertainment	Work and Education	Outdoor	Dressing
color (.10)	kind (.14)	bottle (.13)	much (.37)	cd (.08)	screen (.10)	like (.09)	color (.57)
say (.04)	flavor (.06)	medicine (.06)	money (.32)	picture (.07)	say (.09)	front (.07)	shirt (.22)
see (.04)	box (.05)	kind (.06)	bill (.08)	kind (.07)	captcha (.04)	look (.07)	pants (.04)
kind (.03)	product (.03)	name (.05)	coin (.07)	book (.05)	color (.04)	color (.06)	say (.04)
picture (.03)	coffee (.03)	know (.05)	picture (.05)	name (.04)	read (.03)	picture (.06)	description (.02)

Note: Frequencies in parentheses represent each word's proportion in its scenario, calculated as the ratio of the Number of Questions include the Specific Word to the Total Questions in the Scenario, rounded to two decimals. The words "please", "tell" and "thank" were excluded as they do not help differentiate between tasks in different scenarios and are present in every scenario.

(P01)" expressions, thereby overlooking the hallucination information contained within them. As P05 remarked, "*I was immersed in its vivid portrayal of the street scene... the market in front... I had no idea that it made mistakes with the stores' names and the products they sell.*" Nine participants mentioned that they were convinced by the LVLM's authoritative tone and the multitude of details in the outputs, as P06 stated, "*If it could provide so many details, it should be true.*" Even when participants sometimes recognize that hallucinations may have occurred in the outputs, they still tend to rely on these outputs to interpret the actual situation, making it difficult for them to grasp the truth. For example, when P08 heard the output described as "*a framed map of Long Island, New York,*" she doubted, saying, "*I don't think a hotel room should have this; maybe it is a world map*" but it was just a patterned tile.

Blame themselves instead of the LVLMs: When encountering hallucinations, a considerable portion of participants (7 out of 16) blamed themselves for hallucinations instead of the LVLM. P05 defended the model's improper output by saying, "*This is my fault...it must have been blurry in my photo,*" though logs showed the photos were clear, proving it was the model's error. P02 offered another explanation for the unsatisfactory response,"*The photo processing has a delay, and the environment might change during this time, which the model isn't aware of.*"

In summary, BLV participants are easily deceived by the LVLM's rhetoric and tend to blame themselves rather than the model, demonstrating a lack of awareness of hallucinations.

4.2.2 *Factors Influencing Hallucination Awareness.*

The type of the hallucination: Participants were found to rely on the external manifestations of LVLMs' outputs to discern whether they contain hallucinations, which coincides with our manifestation-based hallucination taxonomy. As shown in Figure 5, a clear trend emerged from the participants' ratings: Fabrication, Fallacy and Distortion were harder to spot than Irrelevance and Misfocus supported by the results of the ANOVA ($F=32.426$, $p<0.001$) and post hoc tests with Bonferroni's corrections (see supplementary materials). Interviews revealed that since participants "*had an expectation* (P11)" when posing questions, responses that deviated from their intentions, such as Irrelevance and Misfocus, could be "*detected right away* (P15)." As for Fabrication and Fallacy, which need to "*dig into the logic* (P02)" or other cognitive processing, detecting them "*can be tricky* (P16)." Additionally, we further examined the actual awareness ratio of each hallucination type, calculated as the proportion of that type of hallucination detected by participants in practice relative to the total number of cases of that type, as shown in5. Overall, the actual and perceived awareness levels show consistency, except for distortions. Despite having the lowest actual awareness ratio, distortions were perceived at a

moderate level of awareness by BLV participants, suggesting that they overestimated their awareness of this type of hallucination.

Personal aspects: Hallucination awareness varied by individual (Figure 5). Experienced individuals with prior knowledge of a specific scenario typically detected hallucinations more easily within that context. As P09, who has cooked since childhood, shared, "*I knew something was wrong when the model stated that beef in this package needed 20 minutes to stew... it's far too short to tenderize properly.*" Additionally, participants (P9, P10, P14, P15) believed that individuals with critical thinking skills are more likely to detect hallucinations than an accepting thinker. All in all, the difficulty of BLV people detecting hallucinations is found to be related to the type of hallucination and users' personal experiences.

4.2.3 Strategies for Detecting Hallucinations. BLV participants reported employing four basic strategies to detect hallucinations and developed a mixed-methods approach, selectively applying or sequentially combining these strategies to address different types of hallucinations.

Four basic Strategies: (1) Contextual consistency[1] was checked by 5 out of 16 participants through rephrasing questions, inquiring about specific details, or uploading images from different angles to observe if the model's response remained consistent in multiple rounds of conversations. (2) Prior knowledge [1, 4], including memories of familiar objects, common sense, or logical reasoning, was used by all participants to detect hallucinations. (3) External knowledge[1], obtained by sharing images with family members, sighted volunteers, or other AI models that participants were more familiar with, was also utilized by all participants to detect hallucinations, with a clear preference for trusting human judgment. (4) Multi-sensory approaches [4], including sensitive touch (P01, P07, P14), olfaction (P08), hearing (P12), and even intuition (P16), were employed to discern hallucinations.

How BLV Users Select and Combine Strategies: Participants were found to flexibly select and combine the four basic detection strategies to deal with different types of hallucinations, achieving both efficiency and reliability in their detection. Irrelevance can be "*surefire detected* (P10)" through prior knowledge due to its disconnect with their expectations of the answers. Misfocus can usually be identified through contextual consistency and further corrected by rephrasing questions or re-uploading images to guide the model back to the user's intended focus. As P05 shared, "*It's probably because what I'm asking about takes up too little space or off-center in the image...So I just try a few more times.* " For Fallacy, participants were found to switch between prior and external knowledge based on the complexity of cases. Prior knowledge was used to detect simple fallacies, which directly contradict common sense or logical reasoning. For complex ones, such as calculation errors, participants tended to combine external knowledge as a "*double check (P07).*" When it comes to Distortion and Fabrication, participants did not adhere to specific strategies for detection. Instead, they had a prioritized sequence for using these strategies, demonstrating a preference for first using multi-sensory and prior knowledge, then checking contextual consistency, and finally seeking external knowledge. As P04 commented, "*I try to detect with my own first, then with other models, and as a last resort, I'll ask a sighted friend for help, but that's only when all else fails.*"

4.3 How BLV Individuals Concern about Risks Associated with Hallucination

Participants expressed concerns about the risks related to hallucinations but were willing to compromise on these risks in light of the potential benefits. They (P02, P05, P06, P07, P09, P11, P13) refrained from assigning tasks that *"may exceed the model's capabilities* (P01)" to mitigate uncontrollable risks. P10 and P14 were particularly wary of the security hazards posed by hallucinations, advocating, *"begin with photos of things you know in a familiar setting* (P10)." Despite their concerns, participants were willing to compromise on risk for the sake of the benefits, with sentiments like "*it's better than seeing nothing at all* (P12)," and "*just as long as it doesn't happen often...it's still just a machine* (P16)." Moreover, some participants (P5, P10, P11) viewed hallucinations as opportunities for model improvement, stating, "current risks are for future progress (P10)."

By analyzing typical hallucination cases in the eight scenarios outlined in Study 1 and combining their experience using AskVision, participants summarized the main risks of hallucinations in each scenario, as shown in Table 5. As shown in Figure 6, participants further rated these risks, and the results showed significant differences in tolerance thresholds across scenarios (ANOVA: $F=20.106$, $p<0.001$; post hoc tests with Bonferroni's corrections in Supplementary materials).

Intolerable: Healthcare risks were the least tolerable, with all participants giving negative ratings due to their concerns over misjudged health conditions or inappropriate treatment advice. P12 shared an instance where the LVLM mistakenly identified wood glue as eye drops, commenting, *"This is fatal. If you're not sure, say so. But when you give answer, it must be correct."* P08 added, "*If that happens even once, I'd stop using it right away.*"

Varies depending on the use case: We observed considerable individual differences in participants' risk tolerance across Finance, Cooking and Dining, Outdoor, House Maintenance, and Work and Education scenarios. Risks in the first three scenarios received negative ratings from more than half of the participants due to their direct threat to the participants' safety and property. In contrast, for the latter two scenarios, where risks might affect task efficiency and performance, participants demonstrated a higher degree of tolerance. Despite the overall trend, participants' ratings across these five scenarios varied depending on their specific use cases. For instance, in the work scenario, P07 (masseur), stated, "*I don't need LVLMs help when I'm giving a massage, so the risks are pretty low.*" On the other hand, P10, a programmer, warned, "*If it misreads the error on the screen and provides inappropriate guidance, that's when I'd be in deep trouble.*"

Tolerable: Risks in Entertainment and Dressing scenarios were deemed "*not a big deal* (P13)," as consequences such as "*unsatisfactory experiences* (P13)" or "*inappropriate attire* (P15)" were considered minor.

4.4 Mental Models of How Hallucinations Occur in LVLMs

Considering that understanding users' mental models is crucial for building responsible AI [55, 73] and optimizing human-AI collaboration [5, 10], we followed the methodology from previous literature

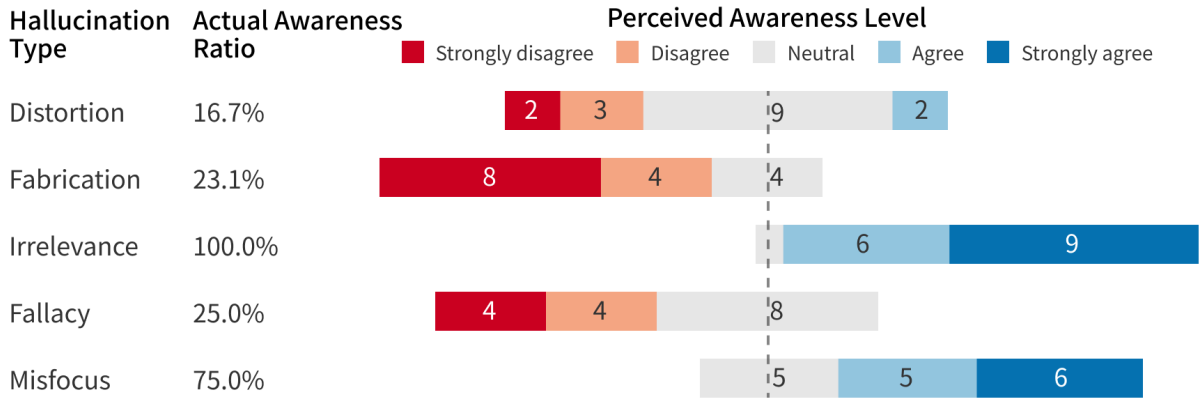


Figure 5: Actual Awareness Ratio and Perceived Awareness Level of Five Hallucination Types by BLV Individuals. The "Actual Awareness Ratio" column is calculated as the proportion of that type of hallucination detected by participants in practice relative to the total number of cases of that type. The "Perceived Awareness Level" is illustrated by responses to the statement, "I am aware of this type of hallucination," rated on a 5-point Likert scale. Numbers on the bars denote the count of responses for each rating.

Table 5: The Potential Risks of Hallucinations in Eight Scenarios

Scenarios	Tasks that hallucinations occur	Common risks
House Maintenance	Ascertain the type and location of daily necessities, as well as the operational status and usage methods of household appliances.	Damage to home items, inefficient chores, and even pose threat to personal safety when it comes to electrical operations.
Cooking and Dining	Inquire various aspects of food such as its type, flavor, cooking or storage methods, and expiration date.	Affect the eating experience and even pose risks of food allergies and poisoning.
Healthcare	Discern personal health metrics such as body weight, blood pressure, or pregnancy test results, as well as medication types or their ingredient concentrations.	Misjudge one's health condition and misuse the medication.
Finance	Identify the denomination of currency, calculate bills.	Financial loss.
Entertainment	Describe the titles and content of CDs, movies, and books, judge the situation of games.	Unsatisfactory entertainment experience.
Work and Education	Operate computers, complete homework problems, interpret files.	Poor learning and work performance.
Outdoor	Identify environmental features, such as plants, vehicles, and architecture .	Mishaps in navigation.
Dressing	Evaluate the color, style, and coordination of clothing.	Improper dressing and damage to personal image.

[91] to explore, for the first time, BLV individuals' mental models regarding the hallucination mechanisms of LVLMs. Four mental models were summarized: Model A: "Random Lottery", Model B: "Lost in Language", Model C: "Garbage In, Garbage Out", and Model D: "Cognitive Dissonance", which indicated that BLV users have insufficient understanding of how hallucinations are generated in LVLMs.

4.4.1 Mental Model A: "Random Lottery". Mental Model A represents the participants' (P02, P05, P13) superficial technical understanding of how LVLMs generate hallucinations. They believed that hallucinations occur purely by chance, as a random event. P13 stated, *"It's unpredictable, like playing the lottery...sometimes you hit a jackpot, sometimes not."* Similarly, P02 added, *"The more you use it, the higher the chance you'll hit a problem sooner or later."*

4.4.2 Mental Model B: "Lost in Language". Some participants (P01, P04, P07, P08, P14) were skeptical that LVLMs perform a "matching" task between users' queries and the existing database, and they believed that hallucinations occur when there is a mismatch. Specifically, they thought this mismatch arises from the model's inability to understand the diverse ways humans express themselves. For instance, P03 described the process as: *"It's like finding a needle in a haystack, but the same question can be phrased in many different ways. If the model fails to comprehend my queries and matches to the wrong answer, hallucinations occur."*

4.4.3 Mental Model C: "Garbage In, Garbage Out". Participants with this mental model (P03, P09, P11, P12, P15, P16) viewed the response process of LVLMs as akin to image recognition. Drawing from their usage experience, they concluded that hallucinations

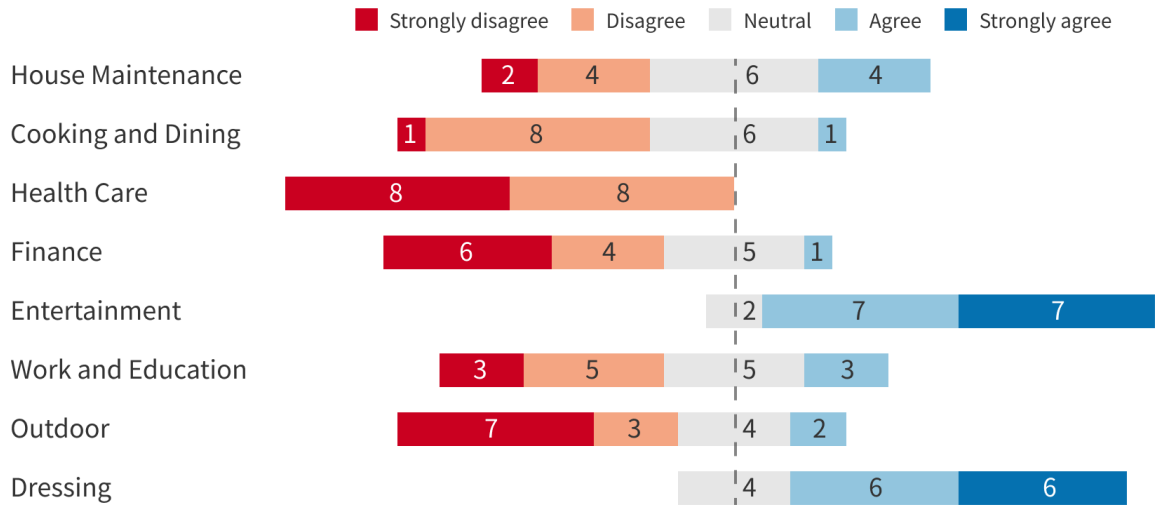


Figure 6: BLV Individuals' Tolerance Levels of Hallucination-Induced Risks in Eight Scenarios gauged by 5-point Likert scales with the statement "I can tolerate the risk of hallucinations in this scenario "

occur due to the poor quality of input images, which makes it difficult for the model to capture key features. As P03 commented, "*If it's not shot well, it (the LVLM) can't see clearly, spouting nonsense.*" Based on this understanding, they believed that "*automatically adjusting composition and lighting* (P12)" or "*uploading videos to provide more data to help models identify which features are key and which are just background noise* (P11)", can reduce hallucinations.

4.4.4 Mental Model D: "Cognitive Dissonance". Participants with Model D (P6, P10), both with technical backgrounds, viewed LVLMs as driven by vast data and algorithms, likening them to a "*not-so-bright brain*" (P6). They attributed hallucinations to "*cognitive dissonance*" (P6), caused by imbalances in input and training data. P10 illustrated this, saying, "*If training data always says 'this bag is green,' the model might insist it's green despite the actual image.*" P06 adds another perspective on the dissonance between scenarios, suggesting that *"when executing specific tasks, if the model does not have a profound understanding of the current scenario, hallucinations may occur. In such cases, designing targeted solutions for typical tasks might be more effective."*

5 Study 3: Understanding BLV Individuals' Expectations for Mitigating the Impact of Hallucinations

5.1 Method

Study 2 revealed that BLV individuals have a limited understanding of hallucinations, struggle to detect them, and are concerned about associated risks. Given that hallucinations cannot be completely eliminated technologically[32], Study 3 focuses on exploring strategies that assist BLV people in comprehending, detecting, and managing hallucinations, thereby mitigating their negative impact from a user perspective.

Considering that BLV people have fundamentally different experiences and capabilities compared to designers, engineers, and researchers who develop assistive tools [72], in study 3, we followed the participatory design method to include users in designing solutions that mitigate the negative impact of hallucinations.

5.1.1 Participant Recruitment. This study took place at a Disabled Persons' Federation facility with 12 participants (P05-P16) from earlier studies, as detailed in Table 1. Two 50-minute co-design sessions were held, each involving six BLV participants and two researchers. The sessions included: 1) Warm-up and introduction, 2) Research findings sharing, 3) Fictional Inquiry. Throughout the research process, researchers encouraged participants to express their ideas and discuss with each other, rather than attempting to steer the direction of the design. The study was IRB-approved, with informed consent obtained.

5.1.2 Procedure.

Warm-up and introduction: We began by introducing the purpose and procedures of the study. Then participants were invited to share their experiences and perspectives on hallucinations, to ease into the research context, and become acquainted with one another and the researchers.

Research findings sharing: Findings from Studies 1 and 2 were presented to enhance understanding and provoke discussion. This included the distribution of 5 types of hallucinations in 8 scenarios, participants' awareness, strategies of hallucinations, the potential scenario risks and participants' tolerance for these risks. Participants were then invited to share and discuss insights, surprises, and comments freely.

Fictional inquiry: Considering that BLV participants prefer discussing ideas over building low-fidelity prototypes, and scenario-based approaches can be effective for participatory design with them[15], we adopted the method of Fictional Inquiry from previous work[28]. Fictional inquiry is a participatory design technique that creates a fictional context to reduce the constraints and biases in current solutions and empower participants to explore innovative

ideas [28, 79]. We introduced a fictional hallucination expert capable of addressing various inquiries related to LVLMs' hallucinations. Then, we asked participants to envision this expert as a "digital assistant" in LVLM-based VQA solutions, inviting them to freely articulate their expectations of this assistant without providing any information about how the assistant might function. Finally, we delved into the detailed aspects of this assistant, including its information sources, interaction modalities, and operational patterns.

5.1.3 Data Analysis. Audio recordings from two co-design sessions, lasting 62 and 71 minutes respectively, were transcribed by the first and second authors. Subsequently, thematic analysis involving two rounds of coding was utilized to conduct a qualitative analysis of these transcripts. In the first round of coding, two researchers independently coded the same session. Through daily meetings, they discussed, reconciled coding differences, and iteratively merged their codes to obtain an initial codebook. Then, they collaboratively conducted axial coding to combine similar codes and form higher-level themes. In the second round of coding, one of the two researchers independently coded the other session using the codebook derived from the first round and made adjustments to the codes and themes with the agreement of the other researchers.

5.2 Findings

This section presents the strategies proposed by participants to mitigate the impact of hallucinations, along with their underlying reasons and thought processes. These strategies have been further compiled into specific guidelines, as listed in Table 6.

5.2.1 Enhancing Information Provision. After learning the findings from Studies 1 and 2, participants recognized gaps in their understanding and awareness of hallucinations and associated risks. Accordingly, they called for LVLM-based VQA solutions to provide more hallucinations-related information, outlining two potential directions:

Hallucinations overview disclosure: Participants commonly emphasized the importance of being informed about the probability of hallucinations. As P05 stated, "*I hope it clearly states from the beginning that the probability of experiencing hallucinations is higher than 17%, which means for every five questions I ask, one might be wrong. This would heighten my vigilance.*" They wished to learn about the types of hallucinations to "*avoid being caught off guard* (P11)", seeking knowledge of their characteristics and occurrence probabilities. They also noted that introducing the mechanism behind hallucinations is necessary to "*feeling confident about something* (P09)".

Real-time scenario-related hallucination alert: Participants called for real-time alerts that informed them about the most common type of hallucination and associated risks according to their current usage scenario. P15 envisioned, "*When I'm snapping a photo of food packaging, it could alert me that hallucinations in this (Cooking and Dining) scenarios typically involve fabricating non-existent details. That way, I'll double-check what it tells me, like the ingredients on the label.*" Especially, if the most common type of hallucination in this scenario is hard for BLV people to detect (refer to Table5), they should be specifically flagged. Some participants (P07, P08, P12) noted that being alerted to potential risks in the current scenario is necessary since it would help them take preventive measures more effectively. P10 suggested allowing users to choose only the alerts they deem necessary, to "*avoid lengthy voice broadcasts that could waste time.*"

5.2.2 Increasing the Transparency of Processing. Participants believed that their insufficient understanding of the system's inner workings hindered their ability to fully assess LVLM outputs. Thus, they hoped for increased transparency with which VQA systems process the uploaded images and questions:

Source: Participants were keen to know the specific sources of information that LVLMs utilize when answering questions to"*make sure the answers made sense (P12)*". This includes knowing "*which elements of the image were considered* (P12)", "*is there any additional references (P15)*" and "*the thought process if further reasoning is involved (P16).*"

Confidence level: Participants advocated for providing a confidence level in LVLMs' outputs. Given the diversity in their mental models of hallucinations, they collaboratively proposed a weighted confidence level calculation that incorporates various factors potentially associated with hallucinations, including the image quality, the clarity of the question, and the model's cognitive level in that scenario. They also pointed out that this confidence level should ideally be conveyed through tone such as "*if the confidence is high, say 'It is highly likely that...*(P11)'", instead of presenting it numerically, which sounds "*too rigid* (P13)" and "*incomprehensible* (P14)".

5.2.3 Introducing Hallucination Verification Strategies. After sharing their respective hallucination detection strategies, participants realized the limitations of their knowledge and advocated for the integration of a comprehensive hallucination verification strategy.

Automatic validation: Participants expressed a desire that LVLMs can conduct "*self-questioning* (P11)" on outputs with low confidence, and determine whether they contain hallucinations through consistency between multiple outputs. P10 commented, "*For hallucinations like Irrelevance, it (the LVLM) should realize on its own that something is amiss if it engages in more comprehensive thinking processes.*"

User-involved information supplement: Participants wished for LVLMs to feel free to ask them for confirmation or to provide additional information to address hallucinations arising from misunderstandings or incomplete data. Such additional information includes"*background information on the current task* (P09)," "*images from different perspectives* (P16)," and "*more specific questions* (P12)".

Third-party verification: Participants proposed that uncertain outputs could be sent to a third party for verification. They tend to trust humans, including family members and remote sighted volunteers, commenting, "*LVLMs can handle simple queries to reduce the frequent disturbance of humans, while complex ones should be handed over to humans for correction.*" On the other hand, participants with mental model D suggested that small-scale deterministic models, trained specifically for particular tasks or scenarios, could be employed for verification.

5.2.4 Incorporating Feedback Mechanism. P10 and P13 recommended introducing a feedback mechanism that allows users to tag and report hallucination cases. These cases could be incorporated into LVLM training datasets to improve the model. P13 believed

such user involvement in model improvement helps to build trust in the system.

6 Discussion

6.1 Analysing Hallucination-related Challenges through Probability- Hallucination-Risk Multidimensional Lens

During our interviews, we found that BLV users evaluate the challenges by weighing the probability of hallucinations occurring, their awareness levels of hallucinations, and their tolerance levels for associated risks, all of which are types and scenarios-dependent. This offers us a new perspective to comprehensively interpret the findings from this study. For instance, in the Healthcare scenario, the dataset analysis reveals a relatively low occurrence probability of hallucinations (see Table 3). However, the most common type of hallucination in this scenario is Distortion, which is the least detectable by BLV users (see Figure 5), and users exhibit the lowest tolerance for risks (see Figure 6). Therefore, addressing hallucination-related challenges in the Healthcare scenario remains a critical priority. Conversely, in the Dressing scenario, where the occurrence probability of hallucinations is low and users' tolerance for risks is also minimal, the challenges are not as pronounced.

Prior work has preliminarily noted that BLV people tend to overlook hallucinations [1, 17, 40, 48], and, through qualitative observations, suggested that usage scenarios may influence LVLM performance [17] and BLV people's concerns about hallucination risks[78], yet it has not examined the underlying factors that impact their awareness nor systematically quantified scenario-level variations. This research further uncovered that BLV awareness levels of hallucinations are tied to the manifestations of hallucination and quantitatively highlight the variation in BLV people's awareness and detection strategies across different hallucination types. Additionally, our findings quantitatively reveal how usage scenarios affect hallucination occurrence probabilities, types, and proportions and suggest that these differences may stem from varying image features and tasks within each scenario. We also found that BLV participants' concerns of risks are scenario-dependent and quantitatively assessed their risk tolerance across scenarios. The above findings indicate that, in the context of VQA for BLV people, hallucinations are not homogeneous. Thus, our developed manifestation-scenario-based dual-dimensional hallucination taxonomy could serve as a framework to scaffold in-depth discussions on BLV individuals' perceptions and targeted mitigation strategies of hallucinations from a more granular perspective. Our findings provide a multidimensional perspective for weighing the hallucination-related challenges, offering data-driven decision support for policymakers, technology developers, and BLV users.

6.2 Charting the Design Space for Hallucination-Mitigating VQA Solutions

Our findings, grounded in the real-life usage of LVLM-based VQA tools by BLV users, provide empirical evidence that the probability of hallucinations occurring is as high as above 17%. Correspondingly, BLV participants demonstrated incomplete mental models regarding hallucinations, identifying less than one-third of them, while the undetected ones can pose significant risks as mentioned in Table5. In addition, Irrelevance, Fallacy, and Misfocus, which account for over 13%, reflect the unique challenges faced by BLV individuals when using LVLMs for VQA, including accurately composing images, asking clear questions, and dealing with reasoning tasks.

Given the pressing and unique challenges of addressing hallucinations highlighted above, this study creates a new design space for hallucination-mitigating solutions tailored to the BLV community. Specifically, the HCI community should focus on introducing fundamental concepts of hallucinations to BLV users. Systems should adopt transparent designs that align with user expectations or proactively disclose unforeseen risks. Considering the limitations of BLV people in detecting hallucinations, there is an opportunity to integrate hallucination detection techniques to assist users in verifying LVLMs' outputs in an automated or semi-automated manner. Notably, hallucination mitigation approaches tailored for the BLV community should focus on leveraging non-visual feedback such as audio and haptic cues to alert users when the provided visual information is a hallucination. All the above strategies are expected to enhance BLV people's understanding of hallucinations, alleviate their self-blame, and reduce their over-reliance on LVLM-generated results.

From a technical perspective, the variances in the probability of hallucinations across different scenarios indicate that LVLMs have limitations when faced with diverse tasks. Thus, more image samples from scenes with high hallucination probabilities should be included in the training dataset, such as images of Cooking and Dining scenarios captured by BLV users, enabling LVLMs to adapt to various situations. For specific tasks such as medical image analysis, a hybrid model approach could be employed: LVLMs identify tasks and assign them to task-specific small-scale deterministic models to minimize the incidence of hallucinations. In addition, Fabrication and Misfocus may reflect an imbalance between the visual and linguistic modalities within LVLMs [2, 37], as the former manifests as LVLMs using prior textual knowledge to "infer" image content without understanding [71], while the latter manifests as LVLMs overly focusing on images without fully considering text queries. Thus, adjusting the model architecture to balance different modalities could be a promising direction. Our findings reveal the relationship between image quality and hallucinations, as depicted in Figure 4, suggesting the potential to detect different types of hallucinations through image quality. Especially, Irrelevance and Misfocus emphasize the challenges LVLMs face in understanding the intentions of BLV individuals through images and questions, and enhance the model's multimodal semantic alignment capabilities should be considered. When LVLMs encounter real-world scenarios requiring complex reasoning, self-play reasoning and retrieval-augmented generation may offer potential solutions to address Fallacy.

Implementing the guidelines in Table 6 requires a concerted, multi-stakeholder effort and presents notable technical and societal challenges. For example, to deliver a meaningful confidence level, AI practitioners need to consider factors outside the model's internal logits. This requires research into quantifying the relationship between input features, such as programmatically assessed image quality and question ambiguity, and the empirical probability of

Table 6: Guidelines to Mitigate the Negative Impact of Hallucinations for the BLV Community

Category	Subcategory	Guidelines
Enhancing Information Provision	**Overview Disclosure**	1. In the onboarding tutorial, incorporate voice narration to announce the overall probability of hallucinations (17%). 2. In the help documentation, provide examples of each type of hallucination, along with their characteristics and occurrence probabilities, in both textual and audio formats. 3. In the help documentation, elucidate the generation mechanism of hallucinations.
	Realtime Alert	1. Automatically recognize the current scenario and provide speech prompts on hallucination probability, the most common hallucination type, and potential risks in that scenario. 2. Use auditory cues or vibration to alert scenarios with low risk tolerance levels, high hallucination probability,, especially when the common type of hallucination in that scenario is one that BLV individuals have a low awareness of. 3. Allow users to customize the alerts they find necessary and the forms of alerting (voice, vibration, sound effects, etc.).
Increasing the Transparency of Processing	**Source**	1. Provide tactile feedback to let users understand the proportion and location of the image parts referenced in generating the answer. 2. Use different timbres to help users distinguish which parts of the answer come from the image, prior knowledge, external resources, etc. 3. Use speech to explain the thought process when reasoning is involved.
	Confidence Level	1. Assess answer confidence level based on image quality, question clarity, and the model's performance in that specific scenario, and express it through tone or wording. 2. Use speech prompts to explain the primary factors affecting the confidence level of the answer.
Introducing Hallucination Verification Strategies	**Automatic Validation**	1. Verify the answer's consistency with the structure or intent of the user's query. 2. Obtain multiple outputs for the same visual question using various models and check the consistency among the outputs. 3. Use other visual questions from the same time period as context to validate the reliability of the answers. 4. Verify the consistency of the answer with known facts or common sense.
	User-Involved Information Supplement	1. Use speech prompts to elicit additional task information or to request clarification of the intent behind the questions. 2. Prompt users to provide additional images or photos from different angles.
	Third-Party Verification	1. Help users connect with family, friends, or remote volunteers for verification. 2. Utilize small-scale deterministic models for verification of specific tasks.
Incorporating Feedback Mechanism	**Feedback-Driven Improvement**	1. Allow users to quickly tag and report cases they suspect contain hallucinations and the reasons for their suspicion through voice commands or specific gestures. 2. Incorporate user-reported hallucination cases into model training.

different hallucination types. Similarly, implementing third-party verification requires a sophisticated task-routing system to intelligently triage requests to different models, human volunteers, or family members based on content sensitivity, user-defined privacy settings, and the urgency of the task. This requires collaboration between HCI designers, backend engineers, and platforms hosting human volunteers. Furthermore, creating an effective feedback mechanism for model improvement extends beyond a user report button. Computer vision researchers need to develop continual learning frameworks that can integrate sparse user feedback, incorporating privacy-preserving techniques like on-device anonymization or federated learning to protect users while still allowing for model refinement. In short, these combined efforts are essential for creating more trustworthy and reliable AI systems.

6.3 BLV Individuals' Trust in AI-based Visual Assistance Tools

This study extends the current understanding of BLV users' trust in AI-based visual assistance Tools. MacLeod et al. [61] found that BLV individuals placed great trust in computer-generated captions while Gonzalez et al. [31] found that BLV users reported low trust in AI-generated scene descriptions. Our findings reconcile the conflict between the findings of above studies by revealing that while the BLV community lacks trust in the capabilities of AI-based visual

assistance tools, they have a strong belief in the results these tools generate. A recurring theme in the interview results is that, compared to human volunteers, BLV users have preconceived lower expectations of AI capabilities and tend to assign simpler tasks to AI to avoid exceeding the model's capabilities. This echoes the findings in Section 3.4 that BLV users were more likely to ask less informative visual questions and provide images without quality flaw when interacting with LVLMs. In that case, participants expressed thoughts like, "*How could such a simple task be wrong?* (P02)" and "*There shouldn't be much risk involved* (P13)," leading to an over-reliance on AI's outputs. When confronted with suspicious answers, they prefer to blame themselves rather than blaming the model, attributing the issue to factors such as poor photo quality. However, according to the findings in Section 3.3, over 10% of no-flaw images can also lead to hallucinations. Such paradoxical trust and self-attribution lead them to overlook many hallucination cases, highlighting the potential risks in this context.

6.4 Bridging BLV Individuals' Mental Models: From Generative AI to Hallucination

Our analysis reveals that BLV individuals often develop inaccurate, incomplete, or oversimplified mental models of hallucination, which partially mirror their mental models of Generative AI [1]. For instance, our participants with Model B "Lost in Language," believe that hallucinations arise because LVLMs fail to comprehend queries, leading to mismatches with the existing knowledge. This understanding reflects that their perceived operation mechanism of LVLMs aligns with the model "King of Knowledge"[1], where LVLMs are seen as conducting keyword-based searches directly within a massive database. Our participants with Model D "Cognitive Dissonance," attribute hallucinations to imbalances in input and training data, indicating that they perceive LVLMs as a form of "deeper artificial intelligence" [1]. Additionally, some participants recognized the probabilistic nature of AI behavior, viewing hallucinations as "unpredictable events" (Model A "Random Lottery"). In contrast, participants in prior research [1] see generative AI as "Word Generating Machine" that produces predictable outputs based on input sequences. The correspondence between BLV individuals' mental models of hallucinations and Generative AI stems from their practical experience with AI visual assistive tools, where they share common challenges. This further indicates that correcting misconceptions about generative AI and clarifying its mechanisms have potential to enhance BLV individuals' understanding of hallucinations.

6.5 Ethical Considerations

Participants in this research were compensated as follows: 30 RMB for Study 1 and 40 RMB for Studies 2 and 3, respectively. Acknowledging the potential privacy challenges in using LVLMs-based VQA tools, we informed our participants of the following and obtained their consent: (1) During the use of AskVision, endeavor not to capture personal privacy information, as all conversation logs will be recorded in the backend. (2) Researchers will manually review the conversation logs to remove or obfuscate cases containing private information. Cases with no privacy risks will be included in the publicly available dataset. (3) There is no disadvantage if participants withdraw midway, and any associated VQA cases will be deleted. Therefore, there is a likelihood that their behavior during the study might deviate from their typical usage patterns to circumvent the disclosure of sensitive information or for other considerations.

6.6 Limitation and Future work

Although we have initially explored the issues of LVLMs' hallucinations and associated risks in the context of VQA for BLV users, there are still some limitations. Firstly, to capture the latest advancements, GPT-4o was selected to generate answers for VQA in this study. This focus may limit our understanding of hallucination issues in other LVLM-based VQA tools with varying performance. Secondly, we refine and constrain the existing definition of hallucinations specifically within the context of VQA assistance for BLV users, aiming to more comprehensively understand the negative impacts of LVLMs on them, however, as the definition evolves and remains contentious, other contexts may use different definitions. In Study 1, the use of binary rating to identify hallucinations in VQA cases, despite achieving a high Kappa score, is limited in handling complex cases where even sighted individuals may have conflicts. Future research could address this by incorporating a "Not Sure" category. Furthermore, our motivation of creating self-built dataset was comparing how BLV individuals question LVLMs versus human volunteers. However, the differences between Chinese self-built dataset and the English VizWiz dataset introduce a known confounder, as model performance is sensitive to both language [95] and cultural context [8, 52, 82]. For instance, our dataset's lower number of content words may not only reflect these influences but also stem from core grammar. Chinese is a "topic-prominent" and "pro-drop" language [53], permitting subject and pronoun omission when the context, like an image, is clear. These variations, aligning with our hallucination findings, underscore the critical need for diverse datasets to ensure LVLM assistance is both effective and equitable worldwide. Additionally, the size of the self-built dataset is still small, and future research can expand upon this dataset for more comprehensive analysis. Lastly, although our analysis incorporated the VizWiz dataset, which comprises VQA cases from BLV individuals worldwide, participants in interviews and co-design sessions were from East Asian backgrounds, which may not fully represent perspectives on hallucinations and risks across other cultural contexts. Future research endeavors will focus on integrating the design implications gleaned from this study into LVLM-based VQA tools, aiming to explore their impact in mitigating hallucinations. Additionally, we are contemplating the integration of wearable devices to harvest a richer array of multimodal information, thereby bolstering efforts to alleviate hallucinations and forging a trustworthy LVLM-based VQA solution for BLV users.

7 Conclusion

This study contributes an understanding of the specific types of hallucinations that BLV users may encounter when interacting with LVLM-based VQA solutions, their perceptions of these hallucinations and the associated risks, as well as the support they seek to mitigate these issues. By analyzing 3467 real-world VQA cases, we developed a manifestation-scenario-based dual-dimensional hallucination typology, uncovering eight scenarios and five types of

hallucinations. Specifically, Irrelevance, Fallacy, and Misfocus highlight the unique challenges faced by BLV users in this context, including posing a range of complex questions to LVLMs, ambiguous queries that fail to accurately articulate their needs, and low-quality images. Through semi-structured interviews with 16 BLV individuals, we found that to deal with different hallucination types, they selectively applied or sequentially combined various detection strategies, including checking contextual consistency, seeking prior and external knowledge, and multi-sensory verification. However, deceived by the LVLMs' rhetoric and tending to self-blame, they were only aware of a small fraction of hallucination cases. Their awareness levels correlated with the manifestations of hallucinations, while their risk tolerance was scenario-dependent. Moreover, their mental models revealed their insufficient understanding of LVLMs' hallucinations. Co-design sessions with 12 BLV participants revealed their needs for hallucination-mitigating VQA solutions, emphasizing enhanced information provision, transparency in processing, verification strategies, and feedback mechanisms. Overall, our findings in-depth clarify hallucination-related challenges in the context of VQA for BLV individuals from the perspectives of manifestations and usage scenarios, and offer firsthand practical insights for developing trustworthy solutions.

Acknowledgments

We extend our warmest thanks to Yu Cai for the discussions during the project. We also thank the Blind and Low Vision participants and the anonymous reviewers for their valuable contributions.

References

[1] Rudaiba Adnin and Maitraye Das. 2024. "I look at it as the king of knowledge": How Blind People Use and Understand Generative AI Tools. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 64, 14 pages. doi:10.1145/3663548.3675631

[2] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the Behavior of Visual Question Answering Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 1955–1960.

[3] Aira. 2020. https://aira.io/

[4] Rahaf Alharbi, Pa Lor, Jaylin Herskovitz, Sarita Schoenebeck, and Robin N. Brewer. 2024. Misfitting With AI: How Blind People Verify and Contest AI Errors. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, NL, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, Article 61, 17 pages. doi:10.1145/3663548.3675659

[5] Robert W Andrews, J Mason Lilly, Divya Srivastava, and Karen M Feigh. 2023. The role of shared mental models in human-AI teams: a theoretical review. *Theoretical Issues in Ergonomics Science* 24, 2 (2023), 129–175.

[6] Apple. 2020. https://support.apple.com/en-hk/guide/iphone/iph3e2e415f/ios

[7] Mohammad Reza Armat, Abdolghader Assarroudi, Mostafa Rad, Hassan Sharifi, and Abbas Heydari. 2018. Inductive and deductive: Ambiguous labels in qualitative content analysis. *The Qualitative Report* 23, 1 (2018), 219–221.

[8] Yujin Baek, ChaeHun Park, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-vlm collaboration. *arXiv preprint arXiv:2406.16469* (2024).

[9] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930* (2024). https://doi.org/10.48550/arXiv.2404.18930

[10] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[11] Hugh Beyer and Karen Holtzblatt. 1997. *Contextual Design: Defining Customer-Centered Systems*. Vol. 6. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 32–42 pages.

[12] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 333–342.

[13] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[14] Ali Borji. 2023. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494* (2023). https://doi.org/10.48550/arXiv.2302.03494

[15] Robin N Brewer. 2018. Facilitating discussion and shared meaning: Rethinking co-design sessions with people with vision impairments. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 258–262.

[16] ByteDance. 2024. https://www.doubao.com/

[17] Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. 2024. WorldScribe: Towards Context-Aware Live Visual Descriptions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 140, 18 pages. doi:10.1145/3654777.3676375

[18] Chongyan Chen, Samreen Anjum, and Danna Gurari. 2023. Vqa therapy: Exploring answer differences by visually grounding answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15315–15325.

[19] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 7824–7846. https://proceedings.mlr.press/v235/chen24bi.html

[20] Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. 2024. Relic: Investigating large language model responses using self-consistency. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[21] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 331, 30 pages.

[22] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. 2020. Assessing image quality issues for real-world problems. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3646–3656.

[23] Baidu Smart Cloud. 2022. https://cloud.baidu.com/

[24] Jazmin Collins, Crescentia Jung, Yeonju Jang, Danielle Montour, Andrea Stevenson Won, and Shiri Azenkot. 2023. "The Guide Has Your Back": Exploring How Sighted Guides Can Enhance Accessibility in Social Virtual Reality for Blind and Low Vision People. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–14.

[25] Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 49250–49267. https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6a435e75419a836fe47ab6793623e6-Paper-Conference.pdf

[26] Be My Eyes. 2015. https://www.bemyeyes.com/

[27] Be My Eyes. 2023. Introducing Be My AI(formerly virtual volunteer) for People who are blind or have low vision, powered by OpenAI's GPT-4. https://www.bemyeyes.com/blog/introducing-be-my-eyes-virtual-volunteer

[28] Yuanyuan Feng, Abhilasha Ravichander, Yaxing Yao, Shikun Zhang, and Rex Chen. 2024. Understanding How to Inform Blind and {Low-Vision} Users about Data Privacy through Privacy Question Answering Assistants. In *33rd USENIX Security Symposium (USENIX Security 24)*. 2065–2082.

[29] Bhanuka Gamage, Thanh-Toan Do, Nicholas Seow Chiang Price, Arthur Lowery, and Kim Marriott. 2023. What do Blind and Low-Vision People Really Want from Assistive Smart Devices? Comparison of the Literature with a Focus Study. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–21.

[30] Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. 2024. VDGD: Mitigating LVLM Hallucinations in Cognitive Prompts by Bridging the Visual Perception Gap. *arXiv preprint arXiv:2405.15683* (2024). https://doi.org/10.48550/arXiv.2405.15683

[31] Ricardo E Gonzalez Penuela, Jazmin Collins, Cynthia Bennett, and Shiri Azenkot. 2024. Investigating Use Cases of AI-Powered Scene Description Applications for Blind and Low Vision People. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.

[32] Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18135–18143.

[33] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3608–3617.

[34] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 417–434.

[35] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. doi:10.1145/3703155

[36] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55.

[37] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6700–6709.

[38] Shuxu Huffman, Si Chen, Kelly Avery Mack, Haotian Su, Qi Wang, and Raja Kushalnagar. 2025. "We do use it, but not how hearing people think": How the Deaf and Hard of Hearing Community Uses Large Language Model Tools. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 33, 9 pages. doi:10.1145/3706599.3719785

[39] Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. GenAssist: Making Image Generation Accessible. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 38, 17 pages. doi:10.1145/3586183.3606735

[40] Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Danna Gurari, Eunsol Choi, and Amy Pavel. 2024. Long-form answers to visual questions from blind and low vision people. In *Workshop on Demographic Diversity in Computer Vision@ CVPR 2025*.

[41] Susmit Jha, Sumit Kumar Jha, Patrick Lincoln, Nathaniel D Bastian, Alvaro Velasquez, and Sandeep Neema. 2023. Dehallucinating large language models using formal methods guided iterative prompting. In *2023 IEEE International Conference on Assured Autonomy (ICAA)*. IEEE, 149–152.

[42] Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. 2024. Hal-Eval: A Universal and Fine-grained Hallucination Evaluation Framework for Large Vision Language Models. (2024), 525‘Ä‘Ş534. doi:10.1145/3664647.3680576

[43] Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. FaithScore: Fine-grained Evaluations of Hallucinations in Large Vision-Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 5042–5063. doi:10.18653/v1/2024.findings-emnlp.290

[44] Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. 2024. THRONE: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27228–27238.

[45] Junho Kim, Kim Yeonju, and Yong Man Ro. 2024. What if...?: Thinking Counterfactual Keywords Helps to Mitigate Hallucination in Large Multi-modal Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 10672–10689. doi:10.18653/v1/2024.findings-emnlp.626

[46] Elizabeth Kupferstein, Yuhang Zhao, Shiri Azenkot, and Hathaitorn Rojnirun. 2020. Understanding the use of artificial intelligence based visual aids for people with visual impairments. *Investigative Ophthalmology & Visual Science* 61, 7 (2020), 932–932.

[47] Hao-Ping Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. 2024. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[48] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 2022. ImageExplorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 462, 15 pages. doi:10.1145/3491102.3501966

[49] Florian Leiser, Sven Eckhardt, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2023. From ChatGPT to FactGPT: A participatory design study to mitigate the effects of large language model hallucinations on users. In *Proceedings of Mensch und Computer 2023*. 81–90.

[50] Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2024. Hill: A hallucination identifier for large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–13.

[51] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13872–13882.

[52] Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems* 37 (2024), 84799–84838.

[53] Charles N Li and Sandra A Thompson. 1989. *Mandarin Chinese: A functional reference grammar*. Univ of California Press.

[54] Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq Joty, and Soujanya Poria. 2023. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269* 3 (2023). https://doi.org/10.48550/arXiv.2305.13269

[55] Q. Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. arXiv:2306.01941 [cs.HC] https://arxiv.org/abs/2306.01941

[56] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[57] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26296–26306.

[58] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253* (2024). https://doi.org/10.48550/arXiv.2402.00253

[59] Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2024. PhD: A prompted visual hallucination evaluation dataset. *arXiv preprint arXiv:2403.11116* (2024). https://doi.org/10.48550/arXiv.2403.11116

[60] Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2024. Negative Object Presence Evaluation (NOPE) to Measure Object Hallucination in Vision-Language Models. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, Jing Gu, Tsu-Jui (Ray) Fu, Drew Hudson, Asli Celikyilmaz, and William Wang (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 37–58. https://aclanthology.org/2024.alvr-1.4

[61] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *proceedings of the 2017 CHI conference on human factors in computing systems*. 5988–5999.

[62] Alexander Mathews, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[63] Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained Hallucination Detection and Editing for Language Models. arXiv:2401.06855 [cs.CL] https://arxiv.org/abs/2401.06855

[64] Alicia O'Cathain and Kate Thomas. 2006. Combining qualitative and quantitative methods. *Qualitative research in health care* (2006), 102–111.

[65] OpenAI. 2023. Be My Eyes uses GPT-4 to transform visual accessibility. https://openai.com/index/be-my-eyes/

[66] OpenAI. 2024. https://openai.com/index/hello-gpt-4o/

[67] JK Periasamy and IK Mukilan. 2022. Generating image description aiding blind people interpretation. In *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*. IEEE, 1–4.

[68] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972.

[69] Vipula Rawte, Anku Rani, Harshad Sharma, Neeraj Anand, Krishnav Rajbangshi, Amit Sheth, and Amitava Das. 2024. Visual hallucination: Definition, quantification, and prescriptive remediations. *arXiv preprint arXiv:2403.17306* (2024).

[70] Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922* (2023). https://doi.org/10.48550/arXiv.2309.05922

[71] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 4035–4045. doi:10.18653/v1/D18-1437

[72] Muhammad Rabani Mohd Romlay, Siti Fauziah Toha, Azhar Mohd Ibrahim, and Ibrahim Venkat. 2021. Methodologies and evaluation of electronic travel aids for the visually impaired people: a review. *Bulletin of Electrical Engineering and Informatics* 10, 3 (2021), 1747–1758.

[73] Heleen Rutjes, Martijn Willemsen, and Wijnand IJsselsteijn. 2019. Considerations on explainable AI and users' mental models. In *CHI 2019 Workshop: Where is the human? Bridging the gap between AI and HCI*. Association for Computing Machinery, Inc.

[74] Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2024. Mitigating Object Hallucination via Data Augmented Contrastive Tuning. *arXiv preprint arXiv:2405.18654* (2024). https://doi.org/10.48550/arXiv.2405.18654

[75] Abigale Stangl, Kristina Shiroma, Nathan Davis, Bo Xie, Kenneth R Fleischmann, Leah Findlater, and Danna Gurari. 2022. Privacy concerns for visual assistance technologies. *ACM Transactions on Accessible Computing (TACCESS)* 15, 2 (2022), 1–43.

[76] Abigale Stangl, Kristina Shiroma, Bo Xie, Kenneth R Fleischmann, and Danna Gurari. 2020. Visual content considered private by people who are blind. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–12.

[77] Yujie Sun, Dongfang Sheng, Zihan Zhou, and Yifei Wu. 2024. AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–14.

[78] Xinru Tang, Ali Abdolrahmani, Darren Gergle, and Anne Marie Piper. 2025. Everyday Uncertainty: How Blind People Use GenAI Tools for Information Access. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 63, 17 pages. doi:10.1145/3706598.3713433

[79] Ge Wang, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2023. 'Treat me as your friend, not a number in your database': Co-designing with Children to Cope with Datafication Online. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[80] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. 2023. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126* (2023). https://doi.org/10.48550/arXiv.2308.15126

[81] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. 2023. Evaluation and Analysis of Hallucination in Large Vision-Language Models. arXiv:2308.15126 [cs.LG] https://arxiv.org/abs/2308.15126

[82] Min Wang and Dan Wu. 2021. ICT-based assistive technology as the extension of human eyes: technological empowerment and social inclusion of visually impaired people in China. *Asian Journal of Communication* 31, 6 (2021), 470–484.

[83] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 15840–15853. https://aclanthology.org/2024.findings-acl.937

[84] Michael Williams and Tami Moser. 2019. The Art of Coding and Thematic Exploration in Qualitative Research. *International Management Review* 15, 1 (2019), 45–55. https://api.semanticscholar.org/CorpusID:198662452

[85] Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. 2024. Evaluating and Analyzing Relationship Hallucinations in Large Vision-Language Models. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 53553–53570. https://proceedings.mlr.press/v235/wu24l.html

[86] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1180–1192.

[87] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817* (2024). https://doi.org/10.48550/arXiv.2401.11817

[88] Bei Yan, Jie Zhang, Zheng Yuan, Shiguang Shan, and Xilin Chen. 2024. Evaluating the Quality of Hallucination Benchmarks for Large Vision-Language Models. *arXiv preprint arXiv:2406.17115* (2024). https://doi.org/10.48550/arXiv.2406.17115

[89] Xiaoyu Zeng, Yanan Wang, Tai-Yin Chiu, Nilavra Bhattacharya, and Danna Gurari. 2020. Vision skills needed to answer visual questions. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–31.

[90] Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2024. The Knowledge Alignment Problem: Bridging Human and External Knowledge for Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 2025–2038. https://aclanthology.org/2024.findings-acl.121

[91] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.

[92] Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2024. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv preprint arXiv:2402.08680* (2024). https://doi.org/10.48550/arXiv.2402.08680

[93] Yi Zhao, Yilin Zhang, Rong Xiang, Jing Li, and Hillming Li. 2024. VIALM: A survey and benchmark of visually impaired assistance with large models. *arXiv preprint arXiv:2402.01735* (2024). https://doi.org/10.48550/arXiv.2402.01735

[94] ZhipuAI. 2024. https://chatglm.cn/

[95] Qishuai Zhong, Yike Yun, and Aixin Sun. 2024. Cultural value differences of LLMs: Prompt, language, and model size. *arXiv preprint arXiv:2407.16891* (2024).

[96] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. Publisher Copyright: © 2024 12th International Conference on Learning Representations, ICLR 2024. All rights reserved.; 12th International Conference on Learning Representations, ICLR 2024 ; Conference date: 07-05-2024 Through 11-05-2024.