

# "It's trained by non-disabled people": Evaluating How Image Quality Affects Product Captioning with Vision-Language Models

Kapil Garg

University of California, Irvine  
Irvine, California, USA  
kapilg@uci.edu

Dwayne R Morgan

University of California, Irvine  
Irvine, California, USA  
dwaynem@uci.edu

Xinru Tang

Department of Informatics  
University of California, Irvine  
Irvine, California, USA  
xinrut1@uci.edu

Darren Gergle

Northwestern University  
Evanston, Illinois, USA  
dgergle@northwestern.edu

Jimin Heo

Computer Science  
University of California, Irvine  
Irvine, California, USA  
heoj4@uci.edu

Erik B Suderth

Computer Science  
University of California, Irvine  
Irvine, California, USA  
sudderth@uci.edu

Anne Marie Piper

University of California, Irvine  
Irvine, California, USA  
ampiper@uci.edu



Figure 1: Example images taken by blind and low-vision (BLV) people featuring common household products. While all of the products in these images are visually recognizable by sighted people, common image quality issues, such as blur, framing, and rotation, make them difficult for vision-language models (VLMs) to recognize. None of the VLMs tested in this study (GPT-4.1, Gemini 2.5 Flash, Llama 3.2 90B, and Molmo 72B) fully and accurately recognized the products in these images.

## Abstract

Vision-Language Models (VLMs) are increasingly used by blind and low-vision (BLV) people to identify and understand products in their everyday lives, such as food, personal care items, and household goods. Despite their prevalence, we lack an empirical understanding of how common image quality issues—such as blur, misframing, and rotation—affect the accuracy of VLM-generated captions and whether the resulting captions meet BLV people's information needs. Based on a survey of 86 BLV participants, we

develop an annotated dataset of 1,859 product images from BLV people to systematically evaluate how image quality issues affect VLM-generated captions. While the best VLM achieves 98% accuracy on images with no quality issues, accuracy drops to 75% overall when quality issues are present, worsening considerably as issues compound. We discuss the need for model evaluations that center on disabled people's experiences throughout the process and offer concrete recommendations for HCI and ML researchers to make VLMs more reliable for BLV people.



This work is licensed under a Creative Commons Attribution 4.0 International License.  
CHI '26, Barcelona, Spain

© 2026 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-2278-3/26/04  
<https://doi.org/10.1145/3772318.3791309>

## CCS Concepts

• Human-centered computing → Empirical studies in HCI;  
Empirical studies in accessibility.

## Keywords

blind and low-vision (BLV) people, image captioning, product identification, hallucinations, image quality, disability-centric evaluation, vision-language model (VLM), large-language model (LLM)

### ACM Reference Format:

Kapil Garg, Xinru Tang, Jimin Heo, Dwayne R Morgan, Darren Gergle, Erik B Sudderth, and Anne Marie Piper. 2026. "It's trained by non-disabled people": Evaluating How Image Quality Affects Product Captioning with Vision-Language Models. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain*. ACM, New York, NY, USA, 27 pages. <https://doi.org/10.1145/3772318.3791309>

## 1 Introduction

Blind and low-vision (BLV) people regularly use automated (e.g., Microsoft Seeing AI, Be My AI, TapTapSee) and human-powered (e.g., Aira, Be My Eyes) tools to understand visual information [9, 77, 120, 132]. While Vision-Language Model (VLM)-based<sup>1</sup> image captioning research has focused on many types of content (e.g., social media photos, scenes, objects) [24, 50, 90, 96], one widely-studied use case is to support BLV people in identifying products, such as packaged foods and household goods. As such, AI tools and their underlying VLMs are becoming more integral to how BLV people perform a variety of everyday tasks, including grocery shopping, cooking, cleaning, and personal care [9, 120, 132]. Yet, we know little about the real-world experiences of BLV people using these tools for product identification or how well VLMs accurately identify products in naturalistic images, where objects of interest may be blurry, out of frame, or rotated.

Despite enthusiasm for VLM-based captioning tools in identifying and understanding products, three challenges complicate their real-world use and evaluation. First, extensive prior work has studied and introduced VLM-based captioning tools to help BLV people understand objects and products in their environment [9, 120, 132]. However, we know less about the factors (e.g., privacy, accuracy, safety) that shape their decision to turn to automated systems rather than humans, and about their experience with captioning errors using existing tools (e.g., Be My AI, Seeing AI). Second, VLM-based image captioning tools perform best when BLV users take and upload high-quality photos, a known challenge for BLV people [30, 33, 55]. Prior work identifies various image quality issues (e.g., blur, rotation, framing, lighting) [33, 55], automatically detects such distortions [30], and introduces techniques to help BLV people take better photos [6, 60, 64, 74, 97, 111, 124, 127]. However, limited prior work has examined how BLV people assess image quality issues with existing VLM tools and perceive their impact on captions [61]. Third, interview studies with BLV people indicate that pervasive image quality issues affect whether images are captioned accurately [9, 152]; however, the relationship between image quality factors and the accuracy of resulting product captions has yet to be systematically analyzed. Prior datasets examine the prevalence of image quality issues, but evaluation of these issues remains coarse (e.g., determining whether an image is captionable or not) [30, 33, 55]. Moreover, existing evaluation approaches for image captions focus on how well a generated caption aligns with a reference text. This

<sup>1</sup>We use "VLMs" to refer to tools like ChatGPT or Gemini that integrate vision-language models and are colloquially known as "AI". We use "AI" in the survey study (Section 3) since participants may be unfamiliar with "VLMs" versus "AI".

can result in false positives, where a caption appears reasonable even when it contains serious errors or omits critical information. Understanding how pervasive image quality issues affect the captions generated by state-of-the-art VLMs is critical, given that these models are used in a wide range of assistive technologies and research prototypes [24, 25, 58, 62, 96, 126].

To help bridge these gaps in the literature, this paper examines challenges in using VLM tools to identify and understand products through two complementary efforts. First, we report results from a survey of 86 BLV participants that detail their experiences and perspectives on captioning product images with existing VLM-based tools. More than half of survey respondents emphasized using only AI tools (over human assistance) when personal privacy matters most, and roughly two-thirds said they would most often use AI when reading a food label, identifying personal care products or toiletries, and identifying an unknown item in their home. Taking a good photo remains the hardest part of the process for many participants (echoing [61, 74]). Even with current tools that provide photo-taking guidance (e.g., SeeingAI, Be My AI), detecting and resolving image quality issues remains challenging. Moreover, the most frequently encountered error in product image captions is missing critical information, such as product brand names and ingredients, which can be obscured when images are of poor quality.

Building on our survey findings, we then develop a structured, annotated dataset of 1,859 naturalistic product images (based on the VizWiz dataset [30, 56]) and use it to evaluate how robust four top-performing VLMs—GPT-4.1, Gemini 2.5 Flash, Llama 3.2 90B, and Molmo 72B—are to common image quality issues. All VLMs were proficient at product identification for high-quality images (i.e., without blur, framing, rotation, or other issues) taken by BLV people, with accuracy rates of 95% or better for GPT and Gemini. Performance across all VLMs drops substantially for low-quality images, with the best model, GPT, achieving only 75% accuracy. Accuracy is even lower when images have multiple image-quality issues, with GPT dropping to 69% accuracy; see Figure 1. Our regression analysis confirms that all models are sensitive to image quality issues and specific content (e.g., cans with rounded labels, nutritional facts text panel) that reduce performance, and it also identifies which image quality issues specific models are more susceptible to and should be a focus for improving their performance.

This paper makes three primary contributions to the accessibility and HCI literature. First, we provide further empirical evidence of BLV people's preferences and experiences with VLM-based tools for product image captioning, underscoring the continued need for improvements in real-world product captioning applications. Second, we discuss the complexities of disability-centered approaches to model evaluation, including task and data selection, annotation procedures, and determining which models and metrics to use. Our work not only benchmarks the performance of four widely-used VLMs, which underlie many modern-day accessibility tools, but it also provides an example of how to approach the evaluation of VLMs that center on BLV people's information needs, answering prior calls to understand and address disability bias in AI models and systems (e.g., [47, 103, 112, 124]). Third, we provide concrete recommendations for making VLMs more reliable for BLV people at all stages of the development pipeline, including data curation, improving model performance, and addressing captioning errors.

## 2 Background

Describing images for BLV people has been a long-standing research area in HCI. Historically, on-demand human assistance was the primary means by which BLV people accessed visual information about their environment. These include remote interpretation services, such as Aira [7], that connect the caller with a trained visual interpreter; crowdsourcing-based systems, including VizWiz [18] and Be My Eyes [14], which ask a paid worker or volunteer to describe an image or video; and friends, colleagues, or family members. In the last decade, advances in computer vision have enabled machines to provide such descriptions (e.g., [129]). For example, early versions of Seeing AI from Microsoft combined various deep learning techniques for computer vision and natural language processing to describe images [83].<sup>2</sup> More recently, vision-language models (VLMs) have exploded in prevalence and capability, with many tools that support image description, like ChatGPT, Gemini, and Be My AI, all using variants of these models. Given their ubiquity, our work focuses on understanding these technologies in the context of BLV people's need for product identification, and their limitations when describing degraded images.

### 2.1 How BLV People Use VLMs for Image Understanding

Image captioning is a well-studied task in computer vision that aims to generate descriptive text for images and has led to extensive work within accessible computing [56, 73, 90, 96, 116]. It is often studied alongside other visual tasks such as visual question answering [18, 21], object recognition [66, 97, 124], and image obfuscation [8]. With the introduction of VLMs, researchers are exploring many new applications of image captioning for BLV people, such as context-aware captions for web images [96], assisting with image editing [25], and real-time scene interpretation of live environments [24, 26, 50, 153]. Among these applications, object recognition is a core aspect of visual access tasks [143] and represents a critical need among BLV people [19]. Significant efforts have been dedicated to helping BLV people identify objects [48], including personal belongings [97, 124] and specific products [60].

More broadly, a substantial body of work has investigated how VLMs perform in object recognition. Modern VLMs are highly performant on zero-shot image identification benchmarks, such as ImageNet [37] and MS COCO [28, 82], which cover a broad range of objects [85]. When VLMs fail, recent work suggests that failures are not due to inference-time (e.g., prompts; decoding strategies) or training-time issues (e.g., learning objective) but rather to limited data frequency for the objects the model is trying to identify [148]. Besides lacking knowledge of image content, VLMs can also fail when the input image is distorted. While significant work has studied how to *measure image quality issues* in photographs (e.g., [5, 44, 49, 89, 137]), relatively little has focused on the *impact of quality issues on captioning*. Initial studies have examined the negative impact of visual variations [43] and the effect of synthetic image degradation [57, 107] on captioning output, but the literature on systematically understanding the impact of real-world image distortions on captioning accuracy is limited.

<sup>2</sup>Architecture details for Seeing AI are sparse, but the original system's release date suggests it lacked attention-based mechanisms found in modern VLMs.

### 2.2 Understanding and Addressing Image Quality Issues

A key issue in using VLMs for BLV people's visual needs lies in the photos they take. From analyzing VizWiz images, Gurari et al. [55] found that blind users often struggle to take high-quality photographs, and many visual questions go unanswered because images fail to capture the relevant objects [55]. While these "low-quality" images are often treated as edge cases (labeled as "other" [19], excluded in analysis [56], or treated as a direction for future work [24]), they make up a significant portion of the photos taken by blind individuals [33]. Image quality has been identified as a major challenge in both model development [56] and user interactions [152], leading to issues with annotation [17, 54, 113, 135] and poor model performance [152]. For example, Davis et al. [33] analyzed 265 medication package images from the VizWiz dataset and found that only 46% were legible. The prevalence of low-quality images has made image quality assessment a stand-alone task in developing image captioning tools for BLV individuals [30].

Recognizing the importance of image quality, tool designers have made considerable efforts to support BLV people in taking photos that both VLMs and humans can caption, with training and instruction playing a crucial role in data collection for model development [66, 97, 111, 124]. Various techniques have been explored to improve data collection, such as using video feeds to capture objects [97, 124], taking sequential photos of objects [66], and sending notifications when objects are out of frame [97]. While training may help, BLV users still find it hard to properly orient objects or avoid unintentionally capturing private content in the background [111, 124]. They may also be uncertain about how to fix photos, even when they know objects are poorly framed [60]. Across this literature, the emphasis is on having BLV people produce "high-quality" images for recognition, rather than systematically understanding how image-quality issues affect their experiences with VLMs accuracy when high-quality photos are not possible.

### 2.3 BLV People's Perspectives on AI Errors

There is growing awareness among BLV people regarding AI tools and errors, leading to many creative and adaptive strategies to identify them [3, 9, 50, 120, 121]. Yet identifying errors can still be difficult for BLV people. For example, when using a prototype object recognizer to identify common food items (e.g., soda, bags of chips, canned foods), BLV participants were only able to identify half of the object recognition errors, even with successive attempts, potentially due to objects' similarity in shape and size [61]. Moreover, most platforms provide little support for helping BLV users understand errors, such as confidence rates or multiple likely image descriptions [3, 9]. In addition, external factors, such as low-quality images or unreliable internet connectivity, often exacerbate perceived inaccuracies in image captioning [152]. When users encounter delays or fail to receive meaningful responses, they may view the system as inaccurate or untrustworthy, even if the underlying model is functioning properly [152]. As more products integrate VLMs into accessibility applications for BLV people (e.g., [24, 62]), it is critical to understand how robust they are to issues of accuracy in everyday tasks—such as identifying household products or goods—where details matter and inaccuracies can affect one's health and safety.

### 3 Study 1: Understanding BLV People's Preferences, Experiences, and Challenges with AI-based Captioning of Product Images

To understand how image quality issues relate to errors during captioning, we first study BLV people's experiences using VLM-based tools to identify and understand products, such as household goods and foods. We extend prior work on how BLV people use AI tools for object recognition [60, 61] by including the specific kinds of products, what information they are seeking, and errors that occur; the tradeoffs between using AI and human assistance based on privacy risks [114, 115], social norms [76, 120], speed, and other factors, as related to product identification; and the impact of image quality on their trust and confidence in the AI tool's output.

#### 3.1 Method

We conducted an online survey with 86 BLV people who use AI tools for image captioning. To clarify the distinction between varying kinds of captioning support, we first asked about the general use of (1) *human-assistance* through remote sighted interpreting services that provide crowdsourced support (e.g., Be My Eyes) or a trained visual interpreter (e.g., Aira); (2) *accessibility-specific AI tools* (e.g., Microsoft Seeing AI, Be My AI, TapTapSee, Access AI); and (3) *general-purpose AI tools* (e.g., OpenAI's ChatGPT, Google's Gemini, Anthropic's Claude). Then, we focused on their use of AI tools to identify and understand products, which we defined as, “packaged items and objects, such as foods, toiletries, cleaning supplies, and other household goods.” Finally, the survey covered their preferences for using AI tools versus human alternatives, and their experiences using AI to understand products (e.g., taking photos, image quality issues, captioning errors).

We revised the survey over three iterations. First, two researchers took the entire survey multiple times to check for language and length. This led to revisions to the survey structure and question wording. Then, we deployed the survey to 10 participants, including an open-ended question at the end that allowed participants to share any confusion or suggestions for improving the survey. This round resulted in two questions being removed and the rewording of others. Following these corrections, we distributed the survey to another 10 participants. No major issues were noted at this stage, and we proceeded with the final deployment. The final survey took approximately 10 minutes to complete. Complete survey questions are provided in the supplementary material.

The survey was hosted via Google Forms, which is accessible to screen reader users, and was open in March 2025. Participants were recruited through email lists maintained by the research team, as well as those of the National Federation of the Blind (NFB) [2] and the American Foundation for the Blind (AFB) [1]. Interested participants signed up through a pre-survey screener. Eligible participants must identify as blind or low-vision, be age 18 or older, use a screen reader to access digital content, speak English, and reside in the United States. Given the focus of our study, participants must have regularly used at least one AI tool (e.g., Be My AI, Seeing AI, ChatGPT, Gemini, Claude) for image captioning. Upon confirming eligibility and excluding any bot-like responses, participants were invited to take the survey using their unique email address. Participants provided consent before beginning. Each participant received

a \$20 Amazon gift card after completing the survey. The survey study was approved by our university IRB.

We received 97 survey responses, which the research team reviewed for duplicates and quality issues (e.g., spam-like responses or those lacking variation). To mitigate bot responses, we required participants to enter the email address to which the survey invitation was sent; responses with invalid email addresses were removed. In total, eleven responses were removed, resulting in a final sample of 86. More respondents in our sample identified as women ( $n = 58$ , 67.4%) than men ( $n = 24$ , 27.9%) or non-binary ( $n = 4$ , 4.7%). Most participants were aged 39–49 ( $n = 44$ , 51.2%) or 50–64 ( $n = 25$ , 29.1%), with smaller groups reporting age 18–29 ( $n = 9$ , 10.5%) and 65 or older ( $n = 8$ , 9.3%). Roughly 67.4% ( $n = 58$ ) of participants identified as white, with some identifying as Asian ( $n = 16$ , 18.6%), Black or African American ( $n = 9$ , 10.5%), and/or Native American or Alaska Native or Native Hawaiian ( $n = 3$ , 3.5%). About 7% ( $n = 6$ ) identified as Hispanic, Latino, or Spanish. More than 80% ( $n = 72$ ) of our sample had earned a bachelor's degree or higher.

#### 3.2 Analysis

We present descriptive statistics of our survey below. Where appropriate, we compare the experiences of BLV users and the effects of different tools or image quality issues on image captioning. We use a Mann-Whitney U Test for inferential statistics because the data being compared are ordinal (Likert scale) [93]. Finally, we present excerpts of quotes from open-ended responses that provide additional context for our interpretations.

#### 3.3 Results

Most participants (76.7%,  $n = 66$ ) reported using AI tools to identify and understand products at least weekly, and half (50.0%,  $n = 43$ ) used remote, sighted visual interpreting applications for these purposes at least weekly. The top accessibility-focused tools used by our respondents to identify and understand products included Be My AI (76.7%,  $n = 66$ ), Microsoft Seeing AI (69.8%,  $n = 60$ ), and AI captioning built into screen readers (51.2%,  $n = 44$ ); see Table 1. Among general-purpose AI tools, participants reported using ChatGPT (38.4%,  $n = 33$ ), Ray-Ban Meta Glasses (29.1%,  $n = 25$ ), and other tools. While we expected most users to regularly use AI tools (given our recruitment criteria), a majority of participants continue to rely on human assistance for product identification. We detail their preferences and challenges with AI tools below.

**3.3.1 AI Captioning Is Preferred for Identifying Food, Personal Products, and Items in the Home.** Building on previous research showing that BLV people move across human and AI assistance for access [3, 9, 120], our findings reveal their preferences and the trade-offs they consider when choosing between the two sources; see Figure 2, left. When considering common scenarios for product identification, roughly two-thirds of the participants said they would almost always or most often only use AI when reading a label on a food item (68.6%,  $n = 59$ ), identifying personal care products or toiletries (67.4%,  $n = 58$ ), and identifying an unknown item in their home (64%,  $n = 55$ ). Surprisingly, more than 45% of participants ( $n = 39$ ) said they would almost always or most often rely on AI to read a medication label, despite multiple AI tools issuing warnings about such use. Fewer participants said they would mainly rely on AI when

**Table 1: Total number of survey respondents who used various accessibility-specific and general-purpose VLM-based tools for identifying products in photographs they took. Participants often used multiple tools for their visual information needs.**

Type	Tool	Percent of Respondents (Count)
Accessibility-focused	Be My AI	76.7% (66)
	Microsoft Seeing AI	69.8% (60)
	AI captioning in screen reader	51.2% (44)
	Access AI	30.2% (26)
	TapTapSee	26.7% (23)
	Google Lookout	11.6% (10)
General-purpose	WayAround	7.0% (6)
	ChatGPT	38.4% (33)
	Ray-Ban Meta Glasses	29.1% (25)
	Google Gemini	15.1% (13)
	Microsoft Copilot	12.8% (11)
	Claude AI	4.7% (4)
	Clarifai	1.2% (1)

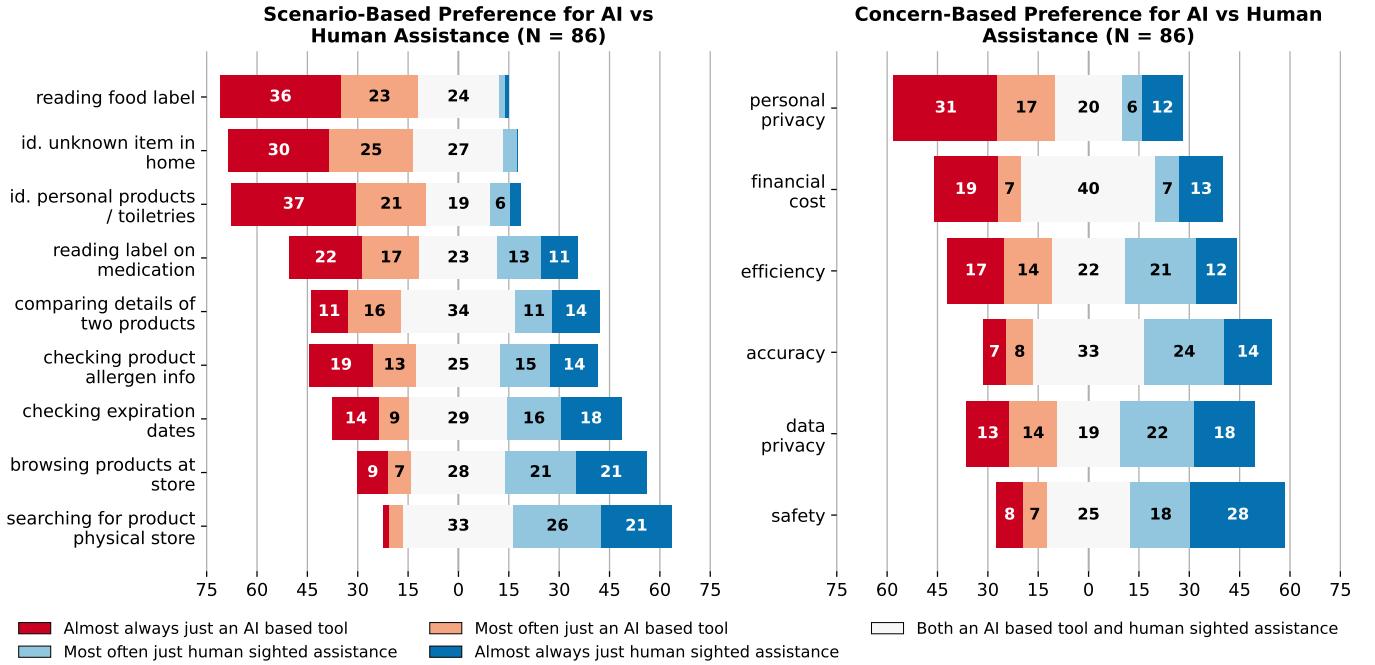
checking allergen information on products (37.2%,  $n = 32$ ), comparing the details of two products side by side (31.4%,  $n = 27$ ), or checking product expiration dates (27.6%,  $n = 23$ ). Although there has been prior work on object recognition when grocery shopping [71, 130, 151], half of the participants said they leaned towards just relying on human-sighted assistance when searching for a specific product at a physical store (54.7%,  $n = 47$ ) or browsing in a physical store (48.8%,  $n = 42$ ). The cost of searching in a large space was a key reason for this preference, with participants explaining that in a grocery store, “a human can often infer or already know where to go. Would take longer with just AI.”

Echoing prior work that highlights concerns about social norms [114, 115], more than half of the participants (55.8%,  $n = 48$ ) said they would most often or almost always only use AI to caption products when personal privacy matters most; see Figure 2, right. The most cited concerns include feeling embarrassed discussing personal matters with real people and the potential misuse of their personal information. In contrast, more participants indicated they would most often rely on human assistance when data privacy was most important (46.5%,  $n = 40$ ). Several people noted the dilemma between personal and data privacy, saying, “it’s a catch-22: go with AI-generated [services] where they store a photo, or a person who could be copying down my information,” which led to varying priorities regarding the associated risks. People were seen as a greater direct risk due to the potential for bad human actors (e.g., “If I ask a human, someone will know”), while AI tools presented a broader indirect risk (e.g., “A human has a limited number of people they could potentially share the information with, but AI means more companies can access your data”). Like data privacy, most participants leaned towards using human-sighted assistance when safety (53.5%,  $n = 46$ ) and accuracy (44.2%,  $n = 38$ ) mattered most, as humans were perceived as more reliable, especially when there was a clear and specific need, such as counting or reading text.

**3.3.2 Taking Photos Remains Time-Consuming and Challenging.** Despite research on supporting BLV people to take photos [6, 64, 74, 127], it remains a key challenge for image captioning. With the

tool they used most, nearly half of the participants (47.7%,  $n = 41$ ) said it took 2–4 minutes to get the desired information, followed by 0–1 minutes (27.9%,  $n = 24$ ) or 5–9 minutes (19.8%,  $n = 17$ ). Two-thirds of participants (67.1%,  $n = 49$ ) said taking a good photo was the hardest part of the captioning process, and roughly half (45.9%,  $n = 34$ ) said it took the longest. Multiple photos were often needed, with most participants (62.8%,  $n = 54$ ) saying 2–3 photos; fewer needed just one photo (23.3%,  $n = 20$ ) or more than four photos (10.5%,  $n = 9$ ). For some participants, taking photos was difficult due to physical disabilities that made it hard to hold the camera steady. Participants described learning to take photos over time, including learning from how Aira interpreters guide them to angle their camera and adjust the environment for visual captioning.

**3.3.3 Current Tools Make It Difficult to Assess and Resolve Image Quality Issues.** Difficulties during photo-taking can result in lower-quality photos, which then affect a VLM’s caption quality. We asked participants about their perceived impact of image quality issues on the quality of AI-generated captions for products, on a 4-point scale from 1: “not at all” to 4: “to a great extent” with the option of “I am not sure”; see Figure 3, left. Overall, BLV users perceived image quality issues of framing ( $m = 3.54$ ,  $s = 0.71$ ), blur ( $m = 3.5$ ,  $s = 0.69$ ), and distance to object ( $m = 3.45$ ,  $s = 0.6$ ) to affect caption quality the most, followed by hand placement and position ( $m = 3.35$ ,  $s = 0.74$ ), lighting ( $m = 3.15$ ,  $s = 0.71$ ), and rotation ( $m = 3.13$ ,  $s = 0.74$ ). A few respondents indicated “I am not sure”, most often for rotation ( $n = 15$ ), hand position ( $n = 11$ ), and distance ( $n = 9$ ), suggesting that the impact of these might be more subtle than that of other quality issues. We also examined the differences between Seeing AI and Be My AI (but not other tools, due to the limited sample size). We found that framing was the only image quality issue whose perceived impact on caption quality was different across tools, being more impactful for Seeing AI than Be My AI ( $m_{\text{Seeing AI}} = 3.79$  vs.  $m_{\text{Be My AI}} = 3.39$ ;  $p = 0.0076$ ,  $U = 529.5$ ;  $n_{\text{Seeing AI}} = n_{\text{Be My AI}} = 28$ ). This is not surprising given that Seeing AI has a feature specifically designed to support framing, which we discuss below.



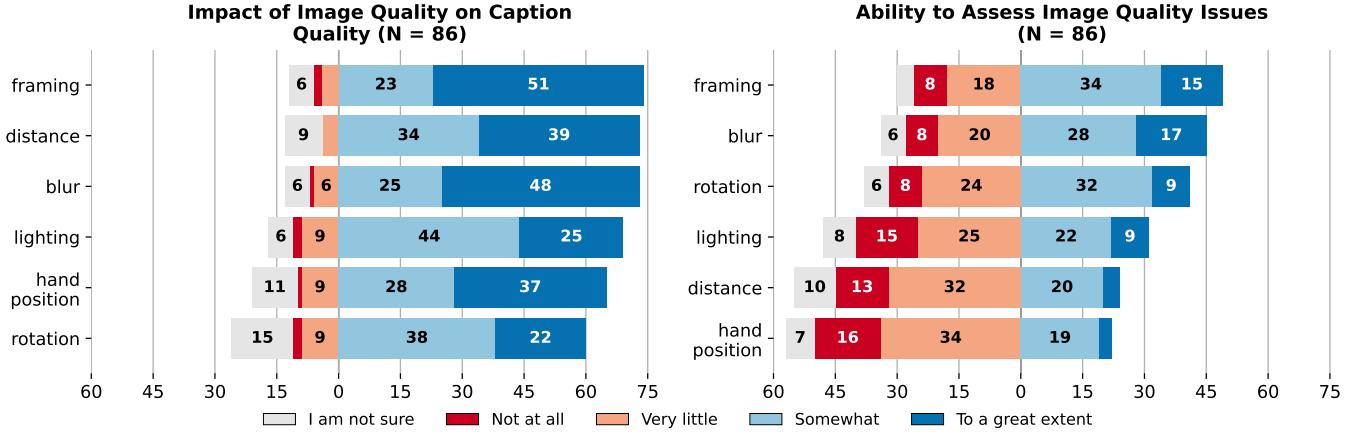
**Figure 2: Divergent stacked bar charts show the distribution of reported responses for scenario-based preferences for AI vs human assistance (left) and concern-based preferences for AI vs human assistance (right). The x-axis shows the number of participants indicating each response. Bar labels 5 and under are hidden due to bar size.**

Given the known challenges of taking good photos, multiple AI captioning tools include features to help BLV users understand image quality issues and adjust their camera position during captioning. We asked how well these tools helped participants assess image quality issues, using the same scale as the impact of image quality; see Figure 3, right. BLV people found framing ( $m = 2.75$ ,  $s = 0.90$ ), blur ( $m = 2.74$ ,  $s = 0.94$ ), and rotation ( $m = 2.58$ ,  $s = 0.85$ ) as the quality issues the tools helped them assess the best, followed by lighting ( $m = 2.35$ ,  $s = 0.96$ ), distance ( $m = 2.22$ ,  $s = 0.82$ ), and hand position ( $m = 2.12$ ,  $s = 0.80$ ). A few participants were unsure when asked whether the tools helped assess quality issues in their photographs, with the most common related to distance ( $n = 10$ ), lighting ( $n = 8$ ), and hand position ( $n = 7$ ), suggesting that the tools provide less support in addressing these issues when taking photos. We observed a significant difference between Seeing AI and Be My AI in how well they help assess whether an image is blurry (Be My AI more than Seeing AI;  $m_{\text{Seeing AI}} = 2.27$  vs.  $m_{\text{Be My AI}} = 2.92$ ;  $p = 0.0147$ ,  $U = 191.5$ ;  $n_{\text{Seeing AI}} = 26$ ,  $n_{\text{Be My AI}} = 24$ ), or if the product is obscured by hand positioning (Be My AI more than Seeing AI;  $m_{\text{Seeing AI}} = 1.81$  vs.  $m_{\text{Be My AI}} = 2.26$ ;  $p = 0.0297$ ,  $U = 198.5$ ;  $n_{\text{Seeing AI}} = 26$ ,  $n_{\text{Be My AI}} = 23$ ). While an in-depth analysis of why users perceive greater support for these two aspects is beyond the scope of the present paper, we present detailed user feedback below and note that Be My AI specifically instructs users to ask the system questions about whether an object is centered and focused [15]. Notably, the average scores for assessing each quality issue range from “Very Little” and “Somewhat”, suggesting

that both tools could do more to make quality issues apparent to BLV people.

BLV people’s open-ended responses suggest that the built-in features for assessing image quality issues are only partially effective. Seeing AI, for example, emits beeps to help users move an object or product barcode into the camera’s view. When asked about this feature, 14 people shared positive comments (e.g., “Does a great job of letting me know when the object is in full view” and “This really helps me when adjusting the angle of the camera and increases my confidence”). However, 27 people shared negative or mixed experiences with this feature, mentioning that it is not always accurate regarding alignment, it is hard to rotate objects to find the barcode, making slight adjustments and holding one’s hand steady is problematic, and the feedback can be misleading (e.g., forcing the camera to put a whole object in view when only a small portion is of interest). One person said, “It’s a game of hot and cold: it takes some trial and error every time to get it right, unless you have a good sense of where the barcode is.” The remainder stated that they had not used this feature ( $n = 21$ ) or did not answer the question.

While Be My AI offered more detailed feedback on photos, it also received mixed responses to its suggestions (e.g., asking users to take a new picture, contact a volunteer, or ask questions such as whether the photo is out of focus [15]). Of the 47 people who commented on the feedback feature, 26 had positive experiences. Others ( $n = 21$ ) provided mixed or negative comments, often noting that the feedback was limited, lacked clear guidance on resolving issues, and still required trial and error. One person said, “It’s good to know the photo is not clear enough, but tough to figure out sometimes



**Figure 3: Divergent stacked bar charts show the distribution of reported responses for perceived impact of image quality issues on caption quality (left) and the ability to assess a quality issue in their image (right). The x-axis shows the number of participants indicating each response. Bar labels 5 and under are hidden due to bar size.**

if it's a lighting, placement, or angle issue." Another commented, "It's really just overall not helpful...the devs still really don't get it. It's not enough to just say the photo's not of good quality; you have to tell someone how to fix it. Many of us have been blind since birth, and how to deal with photos completely escapes us." What's more, feedback was only given after the photo was taken, with multiple participants suggesting that the tool provide more detailed, real-time feedback on framing, lighting, and orientation.

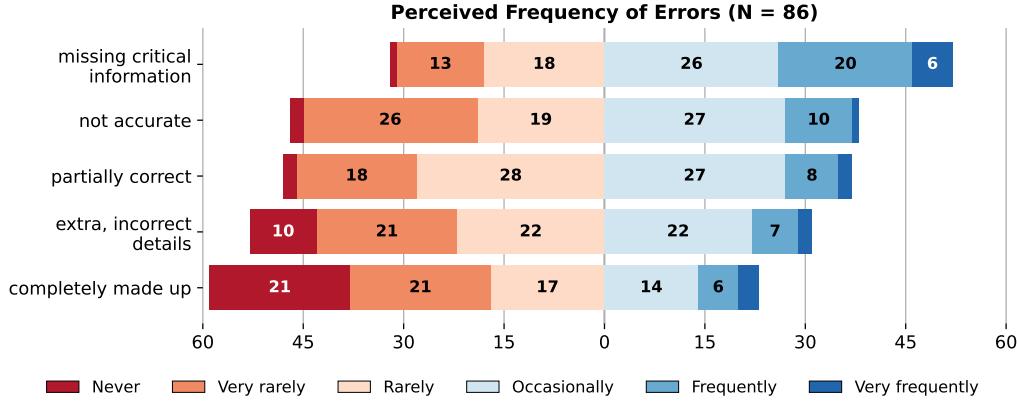
Finally, we asked participants to rate their confidence in knowing *why* a photo is not good enough, even when the tool says it is not good enough to caption or returns a similar error. On average, participants rated themselves as between "slightly confident" and "somewhat confident" ( $m = 2.39, s = 0.97$ ). Only six participants said they were "very confident" or "extremely confident" in knowing why an image was not good enough. There were no significant differences in confidence ratings between Be My AI and Seeing AI.

**3.3.4 Captions Frequently Lack Important Detail and Contain Inaccurate Information.** Although prior work indicates BLV people expect error-prone output from AI tools [3, 9, 120], we know less about the specific kinds of errors they experience when captioning products and their relative frequency. Given this, we asked participants how frequently they experienced various types of errors with the AI tool they used most, on a 6-point scale from 1: "Never" to 6: "Very Frequently"; see Figure 4. Overall, participants reported the highest frequency of errors involving accurate captions that are *missing critical information* ( $m = 3.82, s = 1.20$ ). Other frequently experienced errors were captions that are *not accurate* ( $m = 3.24, s = 1.12$ ) and product captions that are only *partially correct* ( $m = 3.32, s = 1.06$ ). They somewhat less frequently experienced captions that *include extra incorrect details* ( $m = 3.01, s = 1.25$ ) and captions that are *completely made up* ( $m = 2.66, s = 1.41$ ). There were no significant differences in perceived frequency of errors between users of Seeing AI and Be My AI. When asked how frequently they verify captioning output regarding products with a human visual interpreter or another sighted person, more than half of the participants said "rarely" or less ( $m = 3.44,$

$s = 1.27$ ). This aligns with Hong and Kacorri's findings on the overall verification frequency [61].

Many participants reported captions *missed critical information* and lacked details they were specifically seeking, especially regarding brand names, varieties, and ingredients. Respondents said, "I'm trying to find out the color of a lipstick I want to wear, it may capture every bit of info other than the color name, which is very frustrating," and "Many times, Seeing AI does not find the exact title of my yogurt." Others described receiving accurate, general information but lacking needed specificity, such as "AI just says that the product is 'beans' but doesn't specify what type of beans," and "I was trying to find out if I was holding a pack of pork chops or neck bones... It would only tell me it was a package of meat."

In addition to captions frequently missing critical information, participants described captions that were *not accurate* (e.g., recognizing a pregnancy test as a pen, a pair of boots as food item, protein bars as stuffing mix, green beans as spark plugs) as well as *partially correct*, such as getting the product type correct but the specific details wrong (e.g., garlic powder as turmeric spice, frozen shrimp as frozen chicken, agave nectar as maple syrup). Partially correct captions can be more difficult for BLV people to assess and cause potentially life-threatening issues. One person recalled that Be My Eyes AI correctly identified a lotion bottle but got the specific variety wrong, saying it, "left a horrible white cast on my skin, which I didn't notice until someone told me." Another explained that canned pears were identified as canned peaches and that, while both are canned fruits, such errors could be fatal: "My husband is very allergic to peaches, and this probably would have meant a Benadryl shot for him if he'd gotten the wrong product." Similarly, one person said, "It has told me completely different names of medicines than what is printed on the packet," highlighting yet another case where accurate details are essential. Respondents identified these errors based on their life experience (e.g., "it seemed like it was taking me way too long to finish the prescriptions and then I called the pharmacy to verify" and "It told me a name of medicine I knew I never had."), or by asking sighted people. Either way, people



**Figure 4: Distribution of perceived frequency of various types of errors in image captions when describing products. The x-axis shows the number of participants indicating each response. Bar labels 5 and under are hidden due to bar size.**

became cautious about using AI in life-threatening situations and turned to people or to more tested technologies (e.g., Script Talk for medications, which uses RFID technology [42]).

Survey respondents also provided insight into how product design and packaging affect caption accuracy, confirming prior work regarding medical product packaging [33]. Many of these errors stem from package designs that are difficult to photograph effectively, particularly those with rounded or reflective surfaces (e.g., “It won’t read...all instructions on rounded bottles like eye drop bottles,” and “It usually takes more time and lots of rotating the can to piece together the information I’m looking for”).

### 3.4 Final Reflections

Near the end of the survey, we asked BLV people what they would like to communicate to researchers and developers building these tools. Respondents across the board emphasized accuracy and precision, saying “I need accuracy and precise captions,” “Be more specific!” and “Please, please be sure your tools are accurate. Especially if people are using it for life-reliant things like medicine.” Others emphasized frustration (e.g., “It’s really frustrating that I have to go through so many hoops just to be able to find out what’s in a box or can”) and that there is more work to be done, saying developers need to “Take time to understand the specific use cases and needs that are unique to users who are blind or low vision.” Another suggested that many of the issues BLV people are facing with such tools are because underlying models are typically “trained by non-disabled people, [and] show implicit bias toward disabled people.” Their final reflections underscore the importance of evaluating how the image quality issues BLV people contend with daily—which are often set aside in research—affect whether VLMs can accurately identify products with the level of detail that BLV people need for safe and effective use.

## 4 Study 2: Evaluating VLM Caption Accuracy for Product Understanding

Given the pervasive challenges with using VLMs to identify products, we systematically examine how image quality issues affect a VLM’s ability to identify them correctly and in detail.

### 4.1 Challenges in Evaluating VLMs’ Product Captioning Performance

We initially conducted experiments using Gurari et al.’s VizWiz Image Captioning dataset [56], but encountered two challenges. First, captions from crowdworkers varied in whether they correctly identified products and the level of detail provided, making it difficult to assess whether a VLM was performing poorly or if we lacked accurate product identification data to benchmark the model against. Second, existing metrics for measuring caption quality—like BLEU [102], METEOR [12], ROUGE [81], CIDEr [128], SPICE [11], and BERTScore [147]—primarily measure *text alignment* and are unreliable for evaluating *correctness* of product information in captions. Adding images to such measures (e.g., Vilbertscore [72], TIGER [65], SCAN [75]) or using reference-free measures (e.g., CLIPScore [59]) can help, but degraded images can be a confound in their scores. In short, reliance on these metrics could lead to false positives, where a caption appears reasonable even when it contains serious errors or omits critical information. These challenges motivated us to develop a dataset with verified annotations to determine whether products were correctly identified in captions.

### 4.2 Method

**4.2.1 Data Selection.** To create a dataset focused on products, we start with Gurari et al.’s VizWiz Image Captioning dataset [56], which includes five crowdworker-provided image captions on photos taken by blind people, and Chiu et al.’s VizWiz Image Quality Assessment dataset [30], which includes annotations on image quality issues by five crowdworkers. Using their training dataset (23,431 images), we first filter for images for which humans can confidently provide a caption, indicating that image quality issues are not severe enough to prevent image description. We select these data by including images for which two or fewer crowdworkers indicated the image was unrecognizable (conversely, three to five crowdworkers provided a caption). Upon inspecting the dataset, we noticed that most product images included text; therefore, we only included images in which crowdworkers identified text as a heuristic for product identification. This resulted in a filtered dataset of 14,398 images (61.5% of the original dataset).

We then created two data subsets. First, we focused on *high-quality* product images without quality issues, serving as a benchmark for evaluating the performance of VLMs in product identification on natural images. We selected images for which 4 or 5 crowdworkers flagged no issues, and at most 1 person flagged each image quality issue, resulting in 2,599 images (11.1% of the original). Second, we created a dataset of *low-quality images*, where 4 or 5 crowdworkers flagged the image having an *image quality* issue (blur, rotation, framing, obstruction, being too bright, or being too dark), resulting in 5,432 images (23.2% of the original). This dataset corresponds to images for which human captioners felt confident providing a caption, despite identifying image quality issues that could potentially hinder their accuracy.

To assess how accurately VLMs identify products, we manually reviewed all images and identified those that appeared to contain products. Four researchers reviewed all images in each subset. They excluded images that did not include products, such as nondescript boxes or pictures of rooms in the home. We excluded images of computer screenshots, currency, printed papers, books, CDs, DVDs, clothing, and unpackaged electronic devices. These were, on the whole, difficult for annotators to verify objectively, such as identifying an article of clothing or the name of a book from a page of its text.<sup>3</sup> We also excluded any images where more than one product is pictured. For the high-quality images specifically, we also excluded any product images with even mild distortions (e.g., camera blur, lens flares) to ensure the subset was free of image-quality issues. This resulted in a high-quality subset of 729 images and a low-quality subset of 1,696 images.

**4.2.2 Data Annotation.** Our survey results showed that BLV people want specific product information when selecting foods, medicines, and personal products. To capture how well VLMs meet these needs, we developed a three-part annotation scheme consisting of:

- Product: the generic term for the product (e.g., cereal, soup, meal, medication).
- Brand: any detectable brand information (e.g., Betty Crocker, Kraft, Great Value, Kellogg's).
- Variety: details about the type, flavor, or variety (e.g., peanut, low sodium)

A team of four researchers manually annotated each image using this structure. When annotating a product, researchers reviewed the image and crowdworkers. If unsure, researchers also searched online for product images or noted that they were unsure about the image, so another researcher could review it. The image was excluded from the dataset if researchers were uncertain about the pictured product. For example, we excluded images that showed only product barcodes or lacked the visible details required for product verification. To ensure the validity of product annotations, a second researcher then reviewed and confirmed agreement with each image and annotation. Any discrepancies were flagged for discussion, and if no agreement was reached, the image was removed from the dataset. To enable consistency in product naming, the research team aimed for the most specific name within the

<sup>3</sup>While these are real-world cases where objects are ambiguous and valuable to identify, we require images with clear, correct, and assessable annotations to understand how VLMs fail (our focus), where more ambiguous or hard to verify examples could create a confound in our analysis.

product (e.g., granola instead of cereal, Sprite instead of soda) and included both brand and sub-brand names when available. When possible, we included known flavor or ingredient details (e.g., vanilla soymilk; chicken with potatoes and green beans) and details related to dietary needs or potential allergies (e.g., zero-sugar Gatorade, peanut butter granola bars). This detailed, validated labeling is distinct from coarse object and product labels in prior work [55, 66]. This process yielded a final dataset of 1,859 images annotated with product details, comprising 729 high-quality and 1,130 low-quality images. See Appendix A, Table 8 for the number of images where 0–5 crowdworkers identified an image quality issue.

During data annotation, researchers also noted *product properties* that may affect caption quality, including when labels were rounded (e.g., on cans, bottles) or contained large panes of text (e.g., nutrition label, back of box recipes) by double-coding if a product had one or both properties. We identified 622 (33.5% of our dataset) products with rounded labels, 126 (6.8%) with large text panels, and 49 (2.6%) with both characteristics (e.g., back of a can).

Finally, two researchers observed that while the agreement of two crowdworkers on the presence of blur or framing issues effectively captured issue quality, it did not for rotation. Therefore, the researchers recoded rotation as an orientation beyond 45 degrees from the product's natural axis—depending on the product's top and bottom (such as a can) and text orientation—marking the image as rotated if both agreed. Tables 5 and 6 include selected examples of product images showcasing various image quality issues; additional examples can be found in Figure 1 and Appendix B.

**4.2.3 Generating Captions From VLMs.** We used four different VLMs to generate captions for our dataset. We include GPT-4.1 since the three most commonly used AI tools in our survey—Seeing AI [16], Be My AI [13], and OpenAI's ChatGPT [100]—all use a GPT-4 class model from OpenAI. We include Google's Gemini 2.5 Flash, another frequently used model. Finally, we include two recently released open-source models: Llama from Meta [53]<sup>4</sup> and Molmo from the Allen Institute for AI [35], which both exhibit comparable performance to closed-source industry models on benchmarks. We include open-source models since BLV people in our survey and prior work [114, 115] expressed concerns about data privacy when using LLMs, which open-source models can address when run locally. Moreover, open-source models provide access to the model architecture, training regime, and, in some cases, training data (e.g., Molmo), affording greater flexibility for improving performance than closed-source models. For GPT-4.1, we used OpenAI's API and selected the gpt-4.1-2025-04-14 [101] model checkpoint for reproducibility. For Gemini 2.5 Flash, we used Google's API for gemini-2.5-flash [52]; Google does not provide a more specific model checkpoint. For Llama and Molmo, we used Llama-3.290B-Vision-Instruct [94] and Molmo-7B-D-0924 [10] from Hugging Face, with 4-bit quantization.<sup>5</sup> Llama and Molmo were run locally on two NVIDIA RTX A6000 GPUs. For all models,

<sup>4</sup>The Ray-Ban Meta Glasses, the most used general-purpose AI tool after ChatGPT, are also powered by a version of Llama [95]

<sup>5</sup>We tested the smaller Llama-3.2-11B-Vision-Instruct and Molmo-7B-D-0924 with 16-bit precision, but found that the larger, quantized models performed better while fitting within our compute limitations. Prior work suggests that performance loss is marginal with 4-bit quantization [39, 46].

we set `temperature = 1.0` and `top_p = 0.95` to balance determinism and randomness of output generation<sup>6</sup>, and `max_new_tokens = 500` for generated tokens to allow for detailed captions. Before generating captions, images were converted to PNGs with the alpha channel removed—since some VLMs perform poorly with transparent images—but no additional processing was done (e.g., blur reduction; image super-resolution). For brevity, we refer to the VLMs as “GPT”, “Gemini”, “Llama”, and “Molmo” in the following.

We instructed each VLM to caption each image with the same prompt; see Appendix C. Our prompt was inspired by prior work using VLMs to describe images for BLV people [24, 63, 96], and developed following best practices [99]. We focused on prompting the VLM to identify key features, such as the object, product type, brand names, and variety details essential to understanding the product in the image while abstaining from vague language.

**4.2.4 Dataset Coding.** As the final step in our dataset creation process, we manually verified the correctness of each VLM-generated caption. We performed human coding due to the issues with existing captioning metrics (see Section 4.1) and because LLM-as-judges—while correlating well with human judgment for simple question-answer tasks [154]—may be falsely lenient on more open-ended tasks, like slightly incorrect product descriptions (e.g., Coke Zero versus Diet Coke) [122]. All VLM captions were anonymized to minimize potential bias during coding (i.e., Models A, B, C, D), and any image metadata (e.g., what quality issues were present) was concealed, except for the product annotations. The order of images was also randomized to reduce any ordering effects.

Four researchers coded the accuracy of the four VLM captions for each of the 1,859 images in our dataset (7,436 captions in total). Before coding, the research team developed a coding scheme that allowed for minor spelling mistakes and term variation (e.g., soda vs. soft drink; chips vs. crisps) but was strict on key details (e.g., brand and sub-brand; ingredients when describing food variety). Captions were marked as incorrect if there were major hallucinations (e.g., 12 ounces reported as 12-pack, for a soda can) or contained errors that changed their meaning (e.g., grilled chicken instead of fried chicken). In this way, our evaluation measures both recall (the model gets all details) and precision (what the model says is largely correct). Each researcher coded a sample of 50 randomly selected images with all VLM captions (a total of 200 captions). IRR was computed using Krippendorff’s alpha, with an agreement of 0.859. Following the training period, the four researchers independently coded the remaining images, marking ones they were unsure about as “maybe”. The team reviewed and discussed these and other challenging cases. Our final dataset is available in the supplementary materials.

**4.2.5 Analytical and Statistical Approach.** The first step in our analysis was to assess the overall accuracy of VLMs for identifying products across the range of image attributes and quality issues previously identified. We computed descriptive statistics to determine how often each VLM correctly identified products across different image quality types, image quality issues, and product properties.

<sup>6</sup>We tested various `temperature` (0–1.0) and `top_p` (0–1.0) settings. `temperature` had little effect on product identification quality. In contrast, `top_p` caused more noisy captions above 0.95 (the default for our models). These settings are similar to prior work that has used VLMs for image captioning [23, 98].

The next stage of our analysis applied inferential statistics to determine how different types of image degradations and product properties influence each VLM’s ability to accurately identify products. We modeled this relationship over a series of logistic regressions. We began with a single model that included all images and captions for each VLM. This allowed us to assess overall patterns in how the VLMs performed with degraded images, and to make direct statistical comparisons of performance across VLMs. The model predictors included binary variables capturing image quality dimensions (blur, framing, and rotation)<sup>7</sup>, product image properties (rounded label, text panel), and a categorical factor representing the VLMs (GPT, Gemini, Llama, Molmo). We binned each image quality variable as either true (if 2–5 crowdworkers reported the issue) or false (if 0–1 crowdworkers did). Because image quality issues co-occur [30], we included all two- and three-way interactions among image quality issues, two-way interactions between image quality issues and product properties, and two-way interactions between image quality issues and the VLM factor.

Our final analyses fit a set of independent logistic regression models for each VLM. This allows us to more clearly delineate and assess how a given VLM’s performance degrades across different image quality issues. The predictors in these models include image quality dimensions (blur, framing, and rotation) and all two- and three-way interactions among them; product properties were excluded because they led to poorer model fit.

We used the Akaike Information Criterion (AIC) metric during model development to compare candidate models and determine final model parameterizations. The AIC metric assesses the balance between model fit and complexity, penalizing models with excessive numbers of parameters to avoid overfitting. We observed no outliers in the dataset, nor evidence of multicollinearity in the final models (all variance inflation factor (VIF) scores were less than five). Statistical modeling was performed using R (v 4.5.2) [123].

Tables 4 and 7 present the logistic regression coefficients as logits (i.e., log-odds). In the findings below, we report these as the percentage change in the odds of correctly identifying products (i.e.,  $100 * (\exp(\beta) - 1)$ ) for interpretability.

## 4.3 Findings

**4.3.1 VLMs Struggle to Identify Products on Low-Quality Images.** All VLMs struggled to provide accurate captions for degraded images; see Table 2. For high-quality images, GPT and Gemini performed well, recognizing 98.5% and 95.7% of products, respectively. Accuracy for open-source models was slightly less, with Llama correctly recognizing 86.1% and Molmo 86.8%. Performance across all VLMs dropped substantially for low-quality images, with the best model, GPT, achieving only 74.9% accuracy. Gemini performed slightly worse than GPT (71.7% accuracy), but Llama and Molmo fared much worse, at 44.1% and 36.1% accuracy, respectively.

What’s more, accuracy is even worse when images have multiple distortions, with GPT dropping to 69.4%, Gemini to 66.9%, Llama to 36.1%, and Molmo to 27.8%. While identifying over two-thirds of products in images with image quality issues may not seem

<sup>7</sup>We excluded the variables for obstruction, too dark, and too bright from the analysis due to extreme class imbalance and an insufficient number of true cases, which violate the assumptions required for reliable model estimation. See Table 2, Low-Quality, Single Issue, Row “Other Quality Issues”.

**Table 2: VLM accuracy for identifying products, given different image quality issues present. All models perform well on high-quality images taken by BLV people. However, accuracy drops sharply as image quality issues compound.**

Image Type	Image Quality Issue	Num. Images	GPT	Gemini	Llama	Molmo
High-Quality	None	729 (100.0%)	718 (98.5%)	698 (95.7%)	628 (86.1%)	633 (86.8%)
Low-Quality Overall	All Issues	1130 (100.0%)	846 (74.9%)	810 (71.7%)	498 (44.1%)	408 (36.1%)
Low-Quality, Single Issue	Blur	143 (12.7%)	112 (78.3%)	113 (79.0%)	71 (49.7%)	74 (51.7%)
	Framing	250 (22.1%)	209 (83.6%)	191 (76.4%)	139 (55.6%)	124 (49.6%)
	Rotation	55 (4.9%)	49 (89.1%)	48 (87.3%)	39 (70.9%)	19 (34.5%)
	Other Quality Issues	12 (1.1%)	11 (91.7%)	10 (83.3%)	7 (58.3%)	5 (41.7%)
	Single Issue Total	460 (40.7%)	381 (82.8%)	362 (78.7%)	256 (55.7%)	222 (48.3%)
Low-Quality, Multiple Issues	Blur and Framing	242 (21.4%)	172 (71.1%)	164 (67.8%)	88 (36.4%)	86 (35.5%)
	Blur and Rotation	75 (6.6%)	43 (57.3%)	50 (66.7%)	27 (36.0%)	14 (18.7%)
	Framing and Rotation	146 (12.9%)	113 (77.4%)	103 (70.5%)	69 (47.3%)	41 (28.1%)
	Blur, Framing, and Rotation	132 (11.7%)	84 (63.6%)	86 (65.2%)	36 (27.3%)	21 (15.9%)
	Other Co-Occurring Issues	75 (6.6%)	53 (70.7%)	45 (60.0%)	22 (29.3%)	24 (32.0%)
Multiple Issues Total		670 (59.3%)	465 (69.4%)	448 (66.9%)	242 (36.1%)	186 (27.8%)

**Table 3: VLM product identification accuracy is not always affected by rounded labels, like canned foods, or text panels, like nutrition labels. Compared to images with no rounded label or text panel, GPT, Gemini, and Llama show little to no performance loss in the rounded-label-only and text-panel-only conditions across high- and low-quality images; across all images, Molmo's performance drops when only text panels are present. Gemini, Llama, and Molmo all experience performance drops when both a rounded label and a text label are present (e.g., a nutrition label on a can) across all images.**

Image Type	Product Image Property	Num. Images	GPT	Gemini	Llama	Molmo
High-Quality	Overall (All Images)	729 (100.0%)	718 (98.5%)	698 (95.7%)	628 (86.1%)	633 (86.8%)
	Without Rounded Label or Text Panel	356 (48.8%)	350 (98.3%)	338 (94.9%)	302 (84.8%)	311 (87.4%)
	Rounded Label Only	334 (45.8%)	330 (98.8%)	324 (97.0%)	293 (87.7%)	293 (87.7%)
	Text Panel Only	30 (4.1%)	29 (96.7%)	29 (96.7%)	26 (86.7%)	22 (73.3%)
	Rounded Label and Text Panel	9 (1.2%)	9 (100.0%)	7 (77.8%)	7 (77.8%)	7 (77.8%)
Low-Quality	Overall (All Images)	1130 (100.0%)	846 (74.9%)	810 (71.7%)	498 (44.1%)	408 (36.1%)
	Without Rounded Label or Text Panel	706 (62.5%)	522 (73.9%)	500 (70.8%)	303 (42.9%)	244 (34.6%)
	Rounded Label Only	288 (25.5%)	222 (77.1%)	211 (73.3%)	137 (47.6%)	123 (42.7%)
	Text Panel Only	96 (8.5%)	71 (74.0%)	72 (75.0%)	41 (42.7%)	29 (30.2%)
	Rounded Label and Text Panel	40 (3.5%)	31 (77.5%)	27 (67.5%)	17 (42.5%)	12 (30.0%)

problematic, the stakes for misidentifying products are higher for BLV people, especially for products with health or safety-related issues. For instance, Davis et al. showed how medical packaging presents a challenging task for VLMs and is a case where knowing the correct medicine and dosage is critical [33]. Moreover, images with multiple degradations are common in our dataset, comprising nearly 60% of all low-quality images and 36% of the entire dataset.

Recognizing products with rounded labels is generally challenging [33], as is identifying products from a large panel of text. However, we found that these product properties do not always affect the studied VLMs; see Table 3. For high-quality images with only rounded labels or only text panels, GPT, Gemini, and Llama had little to no performance loss compared to high-quality images with neither (maximum drop of 1.6%, for GPT on text panels only). For low-quality images, performance loss for these models was similar

(maximum drop of 0.2%, for Llama on text panels only). Molmo showed a larger drop in performance for text panels in high-quality images (87.4% to 73.3%) and low-quality images (34.6% to 30.2%). However, product images with *both* rounded labels and text panels had a greater impact on performance. While GPT remained unaffected, Gemini, Llama, and Molmo all dropped to 77.8% for high-quality images and similarly for low-quality images (67.5%, 42.5%, and 30.0%, respectively). We suspect that performance drops due to text panels occur because VLMs overfocus on visible text details, leading them to become conflicted between text and visual details, which in turn leads to incorrect inferences [36]. Molmo did this frequently, including one instance in which it labeled a carton of “O Organics almond milk” as “Horizon Organic” because it read “organics”, despite the carton’s completely different design; see Appendix B, Table 13.

**Table 4: Logistic regression model across all images and VLMs, which shows us general challenges VLMs face when describing degraded images. The model coefficients represent logits (i.e., log-odds). p-value significant at: \* 0.05; \*\* 0.01; \*\*\* 0.001.**

Independent Variable	Estimate
(Intercept)	3.6402***
Blur = True	-2.1414***
Framing = True	-1.8610***
Rotation = True	-1.5839***
Rounded Label = True	0.0938
Text Panel = True	-0.5707**
Model = Gemini	-0.5987**
Model = Llama	-1.7674***
Model = Molmo	-1.8242***
Blur and Framing = True	1.1892***
Blur and Rotation = True	0.5838*
Framing and Rotation = True	1.0371***
Blur, Framing, and Rotation = True	-0.5610*
Blur and Rounded Label = True	0.0202
Framing and Rounded Label = True	-0.1693
Rotation and Rounded Label = True	0.5980**
Blur and Text Panel = True	-0.1057
Framing and Text Panel = True	0.8561***
Rotation and Text Panel = True	-0.0923
Blur = True and Model = Gemini	0.4816*
Blur = True and Model = Llama	0.2167
Blur = True and Model = Molmo	0.3438
Framing = True and Model = Gemini	-0.0436
Framing = True and Model = Llama	0.0839
Framing = True and Model = Molmo	0.0260
Rotation = True and Model = Gemini	0.3053
Rotation = True and Model = Llama	0.2684
Rotation = True and Model = Molmo	-0.5686**
Null deviance (df = 7435)	9026.8
Residual deviance (df = 7408)	6966.6
AIC	7022.6

#### 4.3.2 Effects of Image Quality on Product Identification Accuracy

*Across VLMs.* Our regression results reveal that image quality issues impact all VLMs when identifying products; see Table 4. All image quality variables (blur, framing, and rotation) were statistically significant and negative, indicating that their presence increases the likelihood that the studied VLM would incorrectly identify a product. Blurred images were the most likely to be incorrect, reducing the odds of correct product identification by 88.3%. We hypothesize that all four VLMs are trained on high-quality (i.e., non-blurry) images and never learn to handle blurred images during inference. In examples of blurred images, we observe discrepancies in identifying the product generally versus providing necessary details for

BLV (see Table 5, row 1–2)<sup>8</sup> For example, GPT and Gemini correctly identify a box of Chewy Lemonhead & Friends candy, while Llama only identifies “Lemon Head” (missing “& Friends” sub-brand) and that it is candy (missing “chewy” variety). Molmo similarly misses sub-brand and variety details. This suggests that VLMs can capture large, easily readable text, such as brand labels, that is more resistant to distortion than fine-text details (e.g., food flavor). In another example, only GPT can correctly identify a can of Great Value light red Kidney Beans; Llama can identify the brand, but not “kidney beans”; Gemini and Molmo identify nothing correctly.

Framing was the second-most problematic image-quality issue across models, reducing the odds of correct product identification by 84.5%. Specific examples from show that framing issues even affect the identification of common U.S. brands (e.g., Tide detergent, Sprite Zero), which almost certainly occur frequently in the internet-scale training data for these models; see Table 5, row 3–4. What makes framing interesting is how well VLMs fill in or infer the rest of the content. Each VLM was varied in this regard. For example, GPT and Gemini could fill in “Tid” and “DS” for Tide Pods, while Llama could fill in “Tide” and Molmo filled in neither (despite recognizing it was laundry detergent). However, no models could fill in “Sprite”.

Finally, rotation was the least problematic image quality issue, reducing the odds of correct product identification by 79.5%. Qualitatively, we found that rotation makes it harder for VLMs to understand fine text details—which often includes key details about the product—compared to larger attributes, like brand text and logos, or well-known varieties (e.g., Diet for Coke); see Table 6. For product and variety details, we observed that GPT, Llama, and Molmo failed to identify the product (herb tea) and brand (Bigelow) of a K-Cup pod, whereas Gemini was correct. In the second example of Select Choice Chewy granola bars, all models identified the product (granola bars) and variety (chewy, with chocolate chips), but only GPT correctly recognizes the brand.

As shown earlier, co-occurring quality issues can negatively impact performance and are common in BLV people’s photos [30], complicating the challenge of using VLMs to identify products. The regression results reveal significant two-way interaction effects between blur and framing ( $p < 0.001$ ), blur and rotation ( $p < 0.05$ ), and framing and rotation ( $p < 0.001$ ). The interaction plots reveal that when two image quality issues co-occur (e.g., blur and misframing), the drop in performance is less steep than when only one issue is present. We also observe a significant three-way interaction among blur, framing, and rotation ( $p < 0.05$ ); inspection of this interaction plot reveals a similar pattern to the two-way interactions, where additional image quality issues reduce performance, but not to the same extent as a single issue. This suggests that once product images are sufficiently degraded, models struggle to identify them, regardless of further image degradation. Our qualitative observations echo these findings; see Appendix B, Table 12. For example, all four VLMs failed to identify a box of Mucinex Expectorant medication when the image is blurry, rotated 90 degrees, and half of the “M” in Mucinex is out of frame (despite the rest of the label being visible). Yet in a second image, moved ever so slightly so that

<sup>8</sup>Our examples focus on brands from English-speaking countries, primarily the U.S., which the studied models should perform the best on. While our dataset includes brands from other English-speaking countries (e.g., crisps in the U.K.), these examples are sparse and less likely to be in training data for models built by U.S. companies.

**Table 5: Examples of blurred (rows 1–2) and misframed (3–4) product images where VLMs struggle to correctly identify products. Captions had to include accurate product, brand, and variety information to be coded as correct. Captions were shortened for presentation purposes only, indicated by [...].**

Image	Annotation	GPT	Gemini	Llama	Molmo
	<p><b>Product:</b> candy</p> <p><b>Brand:</b> Lemonhead &amp; Friends</p> <p><b>Variety:</b> chewy</p>	<p>[Correct] Box of <b>Chewy Lemonhead &amp; Friends candy</b>, assorted fruit flavors, with a yellow background and cartoon fruit characters on the front.</p>	<p>[Correct] A rectangular yellow box of <b>Chewy Lemonhead &amp; Friends candy</b>. The box has a yellow background with a vertical rainbow stripe on the right side. The words “<b>Chewy Lemonhead &amp; Friends</b>” are printed in large, colorful letters on the front of the box. [...]</p>	<p>[Incorrect] Product Information:</p> <ul style="list-style-type: none"> <li>- Object Type: Box</li> <li>- Product Type: <b>Candy</b></li> <li>- Brand: Lemon Head</li> <li>- Variety: Lemon [...]</li> </ul> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- <b>brand: Lemonhead &amp; Friends</b></li> <li>- <b>variety: chewy</b></li> </ul>	<p>[Incorrect] Bag of Lemonhead <b>candy</b> featuring 4 fruity flavors, displayed against a wooden background.</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- <b>brand: Lemonhead &amp; Friends</b></li> <li>- <b>variety: chewy</b></li> </ul>
	<p><b>Product:</b> kidney beans</p> <p><b>Brand:</b> Great Value</p> <p><b>Variety:</b> light red</p>	<p>[Correct] Can of <b>Great Value light red kidney beans</b>, with a white label and image of kidney beans on the front.</p>	<p>[Incorrect] A metal can of Dole brand sliced strawberries in syrup. The can is white with red and blue lettering. The top of the can features the Dole logo in blue, with the words “sliced red sweet strawberries” beneath it in red. Below the text is an image of several red strawberries in syrup.</p>	<p>[Incorrect] The object is a can. The can is white with blue text and a red image on it. The can is a soup can, specifically a tomato soup can. The can is from the <b>Great Value</b> brand. The can is 14.5 ounces. The can is white, rectangular, and has a red image of tomatoes on it.</p>	<p>[Incorrect] Cylindrical white can containing Honeysuckle Hill Turkey Chili, featuring blue text and an image of cooked ground meat</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- <b>product: beans</b></li> <li>- <b>brand: Great Value</b></li> <li>- <b>variety: light red kidney</b></li> </ul>
	<p><b>Product:</b> laundry; detergent</p> <p><b>Brand:</b> Tide</p> <p><b>Variety:</b> PODS</p>	<p>[Correct] Plastic bag of <b>Tide Pods laundry detergent</b>, featuring the <b>Tide</b> brand logo with orange, yellow, and purple color patterns and the word “detergent” in multiple languages.</p>	<p>[Correct] A <b>Tide PODS laundry detergent</b> package. It is a large, rectangular, plastic pouch with a red background and orange, yellow, and blue stripes. The word “détérgeant” is printed in white letters on the front of the package.</p>	<p>[Incorrect] The product is a red plastic container with a yellow and orange label, featuring the brand name <b>Tide</b> in large white letters. [...] The label features the brand name prominently at the top, with the words “<b>detergent</b>” and “detergente” written in smaller text below it. [...]</p>	<p>[Incorrect] A round <b>laundry detergent</b> container with red coloring, featuring the letter ‘e’ on the left side and a bullseye-like circle with a blue wave in the center. The word ‘detergent’ is written in three languages.</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- <b>brand: Tide</b></li> <li>- <b>variety: PODS</b></li> </ul>
	<p><b>Product:</b> soda/soft drink</p> <p><b>Brand:</b> Sprite</p> <p><b>Variety:</b> zero; 12-pack</p>	<p>[Incorrect] <b>Twelve-pack</b> of Dr Pepper <b>Zero soda</b> in 12 fluid ounce cans, blue packaging with white and light blue text and bubble graphics.</p>	<p>[Incorrect] A <b>12-pack</b> of 12-ounce cans of <b>Pepsi Zero Sugar</b>. The packaging is a blue cardboard box with the word “zero” printed in white.</p>	<p>[Incorrect] Product Description</p> <ul style="list-style-type: none"> <li>- Object Type: Blue cardboard <b>12-pack</b></li> <li>- Product Type: <b>Soft drink</b></li> <li>- Brand: Pepsi</li> <li>- Variety: <b>Zero</b> Sugar</li> <li>- Visual Features: <ul style="list-style-type: none"> <li>+ Color: Blue</li> <li>+ Size: 12 cans, 12 fluid ounces each</li> </ul> </li> </ul>	<p>[Incorrect] Pepsi <b>Zero soda 12-pack</b>. Royal blue packaging with white text displaying brand name and product details. Case contains 12-12 fl oz cans (144 fl oz total). Zero sugar variety. Packaged in a cardboard box.</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- <b>brand: Sprite</b></li> </ul>

the “M” in Mucinex is fully in view but still blurred and rotated 90 degrees, three of the four VLMs correctly identify it. Further disentangling how co-occurring image quality issues affect product identification is an important area for future work.

As we saw earlier, rounded labels and text panels had varying effects on model performance; our regression results provide a clearer illustration. Only text panels caused a significant drop in performance, reducing the odds of correct product identification by 43.5%. An interaction effect for framing by text panel was also significant ( $p < 0.001$ ), with the interaction plot showing that framing generally reduces performance, but no text panel results

in *poorer* performance when misframed. This suggests that text panels can provide the VLM with clues about the product (e.g., from a longer description of a frozen meal), even if other identifying features are not in clear view (e.g., the brand logo or meal title). While having a rounded label had an insignificant effect on product identification odds, the interaction effect for rotation by rounded labels was significant ( $p < 0.01$ ), with the drop in performance being less steep than when only one variable is true (similar to image quality interactions). Appendix B, Table 13 shows examples of these effects. For instance, no model correctly identified the ground beef as 90% lean, 10% fat, despite it being clearly visible

**Table 6: Examples of rotated product images where VLMs struggle to correctly identify products. Captions had to include accurate product, brand, and variety information to be coded as correct. Captions were shortened for presentation purposes only, indicated by [...].**

Image	Annotation	GPT	Gemini	Llama	Molmo
	<b>Product:</b> herb tea <b>Brand:</b> Bigelow <b>Variety:</b> I love lemon	<p><b>[Incorrect]</b> Yellow plastic bottle cap from a Bickford's <b>I Love Lemon</b> Iced Tea bottle. The text on the cap mentions the flavor (Lemon), product type (Iced Tea), and brand (Bickford's). The cap is round with textured yellow plastic.</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- <b>product: herb tea</b> (incorrectly identified as iced tea)</li> <li>- <b>brand: Bigelow</b></li> </ul>	<p><b>[Correct]</b> A yellow K-cup for <b>Bigelow I Love Lemon Herb Tea</b>. The lid is covered with a textured pattern. Purple text on the lid reads, "Bigelow I Love Lemon Herb Tea," with "Le citron... j'adore!" printed below it.</p>	<p><b>[Incorrect]</b> Yellow plastic container with circular shape and slightly raised edges. The brand name "BIGZIONY" is written in purple letters at the bottom of the container. The container has a textured surface and features purple text, including "Love, Lemon" and "High Tea", which are likely product labels or descriptions. [...]</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- <b>product: herb tea</b></li> <li>- <b>brand: Bigelow</b></li> <li>- <b>variety: I love lemon</b></li> </ul>	<p><b>[Incorrect]</b> Blue rectangular box of Land O'Lakes butter. Yellow circular lid on top with purple print. Text includes 'Land O'Lakes', 'Since 1921', 'Buttery', 'Smooth', 'Taste', 'Quality', 'From our family of farm owners to yours', '5 pounds', 'One pound rolls', '454 grams', 'Made in USA', 'Land O'Lakes, Inc.', 'Arden Hills, MN 55112'.</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- <b>product: herb tea</b></li> <li>- <b>brand: Bigelow</b></li> <li>- <b>variety: I love lemon</b></li> </ul>
	<b>Product:</b> granola bars <b>Brand:</b> Select Choice <b>Variety:</b> chewy; chocolate chip	<p><b>[Correct]</b> Box of <b>Select Choice Chewy Chocolate Chip Granola Bars</b>, 5 bars, each weighing 1.25 ounces. The box is mostly red with a picture of a granola bar containing oats and chocolate chips.</p>	<p><b>[Incorrect]</b> A rectangular red cardboard box of Quaker <b>Chewy Chocolate Chip granola bars</b>, 5 bars, 4.4 ounces (125 grams). The box is 6 inches long and 5 inches wide. The front of the box shows the "chewy" logo in white letters. Below the logo, a white box with red text reads "SELECT CHOICE." [...]</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- <b>brand: Select Choice</b></li> </ul>	<p><b>[Incorrect]</b> The product is a red and white box of <b>granola bars</b>. The box is rectangular, with a red base and a white top. The brand name "Chewy" is written in brown letters on the front of the box. The variety is "Chocolate Chip," which is written in white letters above the brand name. [...]</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- <b>brand: Select Choice</b></li> </ul>	<p><b>[Incorrect]</b> A cardboard box containing Quaker <b>Chewy Chocolate Chip granola bars</b>. It has a red-and-white design and is oriented upside down. The box contains 12 bars.</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- <b>brand: Select Choice</b></li> </ul>

in the upper left, and only Llama noticed the text. The rounded Manwich sloppy joe can partially shows the "M" from the logo and an image of prepared sloppy joe, but all models focused on the more visible tomatoes instead, inferring it was just tomato sauce.

Finally, our regression analysis shows model-wise differences in product identification performance. Compared to GPT, the best-performing model, all VLMs had significantly reduced performance (Gemini: 45.1% reduced odds; Llama: 82.9%; Molmo: 83.9%). We found a significant interaction effect for blur by Gemini ( $p < 0.05$ ). The interaction plots showed that Gemini's performance relative to GPT declines more slowly for blurred images, suggesting greater resistance to it. We also found a significant negative interaction between rotation and Molmo ( $p < 0.01$ ). The interaction plot showed that the drop in performance is steeper when images are rotated, suggesting that Molmo is worse at handling rotations than GPT is.

**4.3.3 Differences in What Each VLM Struggles With.** We now analyze each VLM separately to understand its susceptibility to image quality issues; see Table 7. Our VLM-level regression shows that GPT and Llama are less affected by rotated images than by misframed or blurred images (GPT: 86.8% versus 91.8% and 93.8% lower odds; Llama: 59.3% versus 80.3% and 84.3% lower odds). This suggests that efforts to improve GPT and Llama's performance should prioritize blurred images, which are also the most prevalent in our dataset. On the other hand, Molmo is more susceptible to rotated images (91.5% lower odds) than to blurred (82.4%) or misframed (84.3%) images, suggesting that additional training on rotated images is likely to yield the greatest benefit. Gemini was the only

model that had relatively worse performance for misframing (84.8% lower odds) than for blur (83.1%) or rotation (68.0%).

All models showed a significant interaction between blur and framing, with positive coefficients (all  $p < 0.001$ ). GPT also had a significant, positive interaction effect for framing by rotation ( $p < 0.01$ ), while Molmo had significant, positive interaction effects for blur by rotation ( $p < 0.05$ ) and framing by rotation ( $p < 0.001$ ). Inspecting the interaction plots for these revealed that when both independent variables are true (e.g., blur and misframing), the drop in performance is less steep than when only one is true, similar to the interactions between image quality issues in our prior regression.

## 5 Discussion

Despite their impressive capabilities for object recognition, our analysis reveals that VLMs struggle to provide detailed, accurate product captions that BLV people need when images have common quality issues (e.g., blur, framing, rotation). To our knowledge, this study is the first to systematically examine how image quality affects VLMs' ability to recognize products. While numerous studies have examined how VLMs can support BLV people's visual access needs, they largely sidestep image quality issues by asking for better photos (e.g., Seeing AI, Be My AI, [92]) or leaving users to triangulate facts across multiple models [27]. While such adaptive practices are creative and skillful, the normalization of errors signals a dire need to improve how VLMs (and large AI models, broadly) are adapted to applications for BLV people. Based on our findings, we first discuss how our approach moves towards disability-centered VLM evaluation and development, arguing that while VLMs are

**Table 7: Logistic regression results on a per-VLM basis that let us understand how image quality issues and product image properties affect the likelihood of correct identification. The model coefficients represent logits (i.e., log-odds). p-value significant at: \* 0.05; \*\* 0.01; \*\*\* 0.001.**

Independent Variable	GPT	Gemini	Llama	Molmo
(Intercept)	4.1068***	3.0659***	1.7902***	1.8236***
Blur = True	-2.7830***	-1.7801***	-1.8539***	-1.7344***
Framing = True	-2.4973***	-1.8848***	-1.6231***	-1.8539***
Rotation = True	-2.0067***	-1.1406**	-0.8992**	-2.4627***
Blur and Framing = True	2.0405***	1.2420***	1.0932***	1.0883***
Blur and Rotation = True	0.9562	0.4929	0.3790	0.8919*
Framing and Rotation = True	1.6500**	0.8037	0.6144	1.6175***
Blur, Framing, and Rotation = True	-0.9292	-0.1690	-0.5340	-1.0584
Null deviance (df = 1858)	1626.6	1801.4	2493.4	2550.3
Residual deviance (df = 1851)	1342.9	1581.3	2083.4	1978.0
AIC	1358.9	1597.3	2099.4	1994.0

designed for “everyone”, particular attention needs to be paid to BLV people’s specific use cases and how tools fail for them. Second, we argue that improving VLMs requires changes across the model and end-user tool development pipeline, and we propose research directions to improve VLM reliability through data curation, post-training procedures, and inference techniques to reduce errors.

## 5.1 Towards Disability-Centered Model Evaluation of AI Systems

Developing methods to evaluate model performance is an active area of research across HCI, AI, and ML communities. As such, accessibility researchers within these areas have begun to develop various approaches to disability-centered model evaluation that involve prompting [47, 103], metric assessment [67], interviews [9, 120], and more. A disability-centered approach not only depends on the creation of disability-first datasets (e.g., [111, 124]) but also on evaluation that centers on disability throughout. This includes questions of which data are focal to the study, how data are annotated to establish “ground truth”, which tasks and models are selected for evaluation, and which criteria or metrics are used to assess model performance. Below, we describe these issues and the challenges of disability-centered model evaluation.

We began by understanding the information needs of BLV people within a common yet often challenging everyday task: using VLM-based AI tools to identify household products and goods. Our approach of using a survey complemented related interview studies [3, 9, 120, 132] and allowed a relatively large sample of BLV people to share their experiences and issues with a diversity of AI tools for captioning images of products, surfacing unmet needs around details in images, and the difficulty of understanding and resolving common image quality issues. Our research team is all sighted, making it even more critical to understand and prioritize BLV people’s perspectives from the start.

While related disability-centered approaches aim to support people with disabilities in generating “good” data for training systems [51, 60], our study examined the opposite side of this issue. We intentionally curated a disability dataset such that it targets important

but understudied cases (i.e., product images with quality issues), thus aiming to interrogate cases that are central to BLV people’s lived experiences but often set aside in research (i.e., labeled as others [19], excluded in analysis [56], or treated as a direction for future work [24]). Rather than placing the burden on BLV users to consistently capture “high-quality” photos required for successful object recognition or training, future datasets should treat image quality variability as a central design consideration, in contrast to existing datasets that overwhelmingly focus on high-quality images (e.g., ImageNet [37] and MS COCO [28, 82]) that VLMs are optimized on. Including representative quality variations that reflect the real-world conditions under which BLV people capture images can help us develop VLMs that are more resistant to such variations from the start, rather than needing to fix them in post-training.

Although academic scholars and industry corporations have emphasized the pressing need for more disability-centered datasets [20, 38, 55, 79, 97, 111, 124], annotating these datasets with meaningful “ground truth” labels so that they can be used in benchmark studies and model evaluations such as the present paper remains a challenge, particularly when the phenomena of interest are inaccessible to the people who matter most [51, 60]. Relying on crowdworkers is a common approach to annotation, but they may lack insight into disabled people’s information needs and may apply varying standards of visual interpretation in BLV-focused datasets [113]. They are also often constrained by the time allotted to each annotation and tend to move on quickly when encountering difficult cases. Using other VLMs to synthetically generate annotations is a popular approach [84, 119], but it is likely to perpetuate inaccuracies or biases that the model already has (see distribution shift [110]), rather than capturing important nuances. In other words, the most challenging use cases for machines require extensive human labor. In our case, four researchers spent more than three months reviewing, discussing, validating, and annotating low-quality images. While we developed a structured annotation framework based on BLV users’ information needs, we were still limited by the information available in images, and could not reliably code expiration dates

or product ingredient lists (other details that BLV people wanted captured and should be examined in future work).

Another challenge is selecting models to evaluate that align with disabled people's experiences and needs, and are amenable to further research. Our study selects a complementary set of VLMs: two closed-source models that power the AI image captioning tools BLV people use daily (e.g., Seeing AI, Be My AI), enabling industry relevance and application of our findings; and two open-source models because data privacy was an important concern for BLV people, and these models can be run locally, allowing greater control over privacy-sensitive data, as we discuss below. Open-source models also enable the understanding of training procedures, which can aid in interpreting evaluation results.

Finally, disability-centered approaches must contend with which measures of "success" best represent disabled people's concerns. For example, Kapur and Kreiss [67] demonstrates bias in reference-based metrics against BLV people, calling for evaluation methods based on user groups' specific needs. Towards this end, the research team manually reviewed and coded 7,436 model captions for accuracy and completeness, rather than relying on metrics that assess similarity and could lead to false positives (see Section 4.1). That is, we aimed to emphasize BLV people's information needs by requiring models to generate both necessary and accurate product details rather than settling for general category identification (e.g., "can of food") or brand recognition (e.g., "Campbell's"). Given the difficulty BLV people reported in assessing errors, let alone the risk of mis-identification, more nuanced and consistent frameworks for data annotation and error analysis are crucial for reliable VLMs, especially for high-stakes uses, such as identifying food products, medications, and household cleaners. Our annotation structure provides a pathway for annotating products, with similar structures being an important direction for future work on disability datasets.

## 5.2 Recommendations for Improving VLM Performance on Low-Quality Images

While the studied closed-source models (i.e., GPT-4.1, Gemini) perform better on low-quality images, open-source models (i.e., Llama, Molmo) are likely more fruitful for developing reliable VLMs that meet BLV people's needs. Closed-source models are limited to prompt engineering—which is insufficient for handling distorted images—and fine-tuning to improve performance. While black-box APIs for closed-source VLMs allow limited fine-tuning on provided data, they offer far less flexibility, as details about the model architecture, training data, and the tuning process (e.g., which weights are frozen and the loss function used) are not disclosed. Moreover, closed-source models may leak private data [41, 86], compromising data privacy that our survey respondents strongly desired. In contrast, open-source models make the model's architecture and training details available to researchers<sup>9</sup>, while preserving privacy when run locally. To narrow the performance gap between open- and closed-source models, we propose three areas of research across the VLM pipeline: data curation, training objectives, and inference-time techniques.

<sup>9</sup>Molmo goes further and makes training data available [35], while Llama only provides high-level descriptions of their dataset [53].

**5.2.1 Improved Post-Training of VLMs Through Data Curation.** VLM performance is heavily shaped by post-training activities, including fine-tuning on specific tasks (e.g., PixMoCap for captioning [35]) and diverse datasets [80], or training to provide answers in specific formats (e.g., instruction tuning [84]). One way to improve models at this stage is to give examples when the model lacks knowledge about a task [148]. For recognizing products and their attributes, recent research suggests that VLMs require fine-tuning for good performance [104, 125]. However, our analysis shows that off-the-shelf VLMs perform well for U.S.-based products when product images are high-quality, suggesting that the issue is not due to the model's knowledge gaps. That said, such training could help adapt models for different user populations, such as BLV people in non-English-speaking countries, which we did not study. Products in those countries are infrequently found in the U.S. or on English-written webpages, which we hypothesize are the primary sources of training data for the VLMs studied.

Better datasets could be used to train VLMs to learn more robust representations of how products look when images are degraded. While performing well on high-quality images, all models had substantially lower performance on low-quality images, suggesting they could not find enough distinguishing characteristics in those images to support successful identification (as humans could). To remedy this, future research could develop synthetic datasets in which high-quality images are systematically degraded with different image-quality issues (similar to [57]), such as a can of soda with progressively greater blur or different framing issues, and fine-tune a VLM on them. Such work can draw inspiration from research in quality-agnostic learning (e.g., [68, 142]) that has demonstrated modest improvements in handling image distortions, yet still leaves significant room for improvement in modern VLMs. For instance, Molmo already applies an overlapping cropping procedure in its training [35], which we would expect to make it more resistant to misframed images, but our findings demonstrate that further development is needed to address its sensitivity to image framing. To that end, our findings can help focus these efforts when coupled with knowledge about model training. For example, in addition to misframed images, Molmo struggled most with rotated images, suggesting that providing pairs of correctly aligned, rotated images with high-quality annotations could help the model recognize object similarities despite different orientations. Likewise, Llama struggled the most with blurred images, suggesting that providing it with pairs of blurred and non-blurred images may help. Moreover, open-source training procedures allow us to focus on fine-tuning specific parts of the model for this task, such as the vision encoder, while freezing parts that work well, like the language encoder. Synthetic datasets, however, should still be tied to and evaluated alongside user-generated datasets to help preserve the nuanced qualities of authentic data. Our existing dataset serves as a good starting point for such initiatives, as it includes high-quality images that can be altered and low-quality images for naturalistic comparison.

**5.2.2 Better Learning Objectives for Post-Training.** Alongside the data used for training, effective post-training may require reconsidering commonly used loss functions if they do not capture correctness well for the domain-specific task, such as product identification.

Our study revealed that VLMs frequently produce believable product descriptions that are subtly incorrect, affecting their meaning (e.g., “Coke Zero” versus “Diet Coke”). While VLM loss functions differ, many use cross-entropy loss between the distribution of the model’s logits and the true labels of tokens from the training data. To more directly assess whether different attributes of product annotations are preserved during fine-tuning, future work may develop evaluation metrics based on semantic relationships within the annotations. Inspiration could be taken from evaluation metrics like SPICE [11] that evaluates overlaps between scene graphs (e.g., can → on → countertop) or Cap F1 [35], which evaluates overlap between atomic concepts (e.g., “A can of soda”; “Soda is on the kitchen countertop”). Such loss functions could better steer models towards learning what constitutes good product annotations.

**5.2.3 Addressing Captioning Errors During Inference.** While improved model training can help, it is unlikely to fully resolve the issues our study reveals; instead, we hypothesize that additional inference-time techniques can enhance VLM output without burdening the BLV user to take additional photos. One way is to leverage image reconstruction techniques that repair images before captioning. For instance, with misframed images, researchers can explore inpainting techniques that produce multiple possible versions of a repaired image for captioning [4, 31], eliminating the need to take additional photos. Another is to ensure key product details are included or excluded, for which we can look to related work on reducing toxicity or enforcing lexical constraints in LLM outputs, in which constraint-based optimization can have advantages over conventional fine-tuning [87, 106]. Furthermore, these techniques can often be applied to large VLMs without costly model training, or can be combined with training smaller VLMs (which require less hardware) to improve their output beyond that of larger models.

Even after applying reconstruction techniques, a VLM may still make errors; in such cases, it should abstain from providing a caption. Simple techniques involving prompt engineering to abstain are of limited efficacy, with no guarantees that the instruction to abstain will be followed (e.g., best abstention prompting yields only 0.78 accuracy on question-answer tasks with similarly low-quality images [63]). In contrast, recent work on LLM abstention explores techniques based on self-consistency, in which the model evaluates its own outputs and level of uncertainty before returning a response, demonstrating good performance in question-answering settings [32, 70, 91, 133]. However, abstention for open-ended image captions is harder. In our study, we observed numerous cases in which image captions contained correct *parts* of our product annotations, even when the caption as a whole was incorrect. While recent work for VLMs has explored techniques to *repair* captioning errors prior to returning them during the generation process (e.g., controlling what objects are mentioned [144]; strategically adjusting model weights [78, 108, 136, 141] or fine-tuning [22, 146]; sampling multiple patches [29]; guided decoding [150]; backtracking when uncertain [40, 131]) or post-hoc verification [139, 155], these techniques can induce further errors during correction, rather than providing a higher-precision output that only includes details that are likely correct. Instead, systems for *partial abstention*, which abstain only on inconsistent caption parts, should be explored. These could help the user understand what the model knows and is confident about,

allowing them to decide whether to retake a photo to gather more information about the image or to confirm the information with someone else. Together, these techniques help make VLMs more reliable by providing high-quality responses when possible and only sharing what it is confident in when not.

### 5.3 Recommendations for Supporting Better User Understanding of Image Quality Issues

While we emphasize multiple ways to improve VLM performance on low-quality images, BLV people may still need to re-take photos, which participants in our study wanted better guidance on. Thus, we must continue to design applications that provide richer feedback on the photo-taking experience, helping users understand their environment and potential image quality issues, and guiding them in resolving them. For example, as our participants suggested, a multi-faceted approach could provide feedback *before* taking the photo, pointing out lighting conditions and environmental details that may affect the process; *during* photo taking, offering continuous feedback to the user about the camera angle and object positioning to capture relevant parts of products (e.g., product logo, back of the box, nutrition label) [6, 64, 74, 127]; and *after* taking the photo, informing users about image quality issues to help them learn what might affect captioning and how to make adjustments. However, survey participants also raised concerns that people with multiple disabilities may find such interventions more difficult. For example, participants mentioned difficulty holding the camera steady enough and carefully controlling their breathing to prevent blur. Others mentioned their dexterity makes it difficult to orient the camera in particular ways. While improving the photo-taking experience is important, the complexities of photo-taking for disabled users underscore the need for technical improvements first and foremost, rather than placing the labor of taking good photos on the users.

### 5.4 Limitations and Future Work

Our study has a few limitations that future work should address. First, we focus on evaluating product identification *accuracy* rather than the *caption quality* of VLMs generally. We focus on products because BLV respondents in our survey strongly wanted to know which products they had photographed. However, VLMs provide numerous details in image captions, including key product information (e.g., a can of Coca-Cola), plus visual details of the product and nearby objects (e.g., the can is red; the can is on the counter), which BLV people want in captions (shown by our survey and prior work [67, 97]). Moreover, how information is presented can change its interpretation. For instance, humans often use *hedging language* to indicate uncertainty about information (e.g., “likely is” Diet Coca-Cola); as VLMs can also use such language, understanding how it affects BLV people’s interpretation of uncertain information with respect to helpfulness and safety—such as if key dietary information is missing, leading to less trust in the output—may inform how a VLM should present captions. Existing work shows that expressions of uncertainty can meaningfully influence users’ reliance on model outputs [140]. However, current VLMs struggle to communicate their internal uncertainty through natural language [69, 117]. This misalignment becomes particularly problematic for BLV people when models use overly confident language

despite uncertainty, or, conversely, when they hedge even when the information is accurate. Future studies should examine caption quality in this more holistic manner.

A second limitation is reducing image quality issues to a binary variable. Our dataset included a count of crowdworkers who identified an image quality issue, but treating the count as continuous or ordinal over-interprets it (i.e., 5 is not necessarily more blurry than 2), which is why we converted it to a binary variable. In reality, image degradation occurs on a spectrum, likely affecting VLMs differently as it worsens. For instance, low blur may cause no issues with captioning, while higher blur is problematic. Future work can draw from computer vision research to quantify image degradations (e.g., blur kernel estimation [45, 118, 149]; occlusion-robust object detection and segmentation [105, 145]; rotation-robust object and text detection [88, 109, 138]) and, for instance, use these values in regression analysis similar to ours.

Finally, our experiment focused on VLMs and data with a U.S. and English-speaking bias. These VLMs would likely perform worse on product photos from a non-English-speaking country. Previous research has identified cross-cultural bias as a significant limitation of VLMs perceived by BLV users [9]. Future work should consider how well the VLMs we studied perform in cross-cultural contexts and may also explore other open-source models that explicitly train on other languages (e.g., Qwen [134] or Deepseek [34] for Chinese).

## 6 Conclusion

As blind and low-vision (BLV) people increasingly rely on Vision-Language Model (VLM)-based tools to generate image captions for product identification, we need a more nuanced understanding of how these systems handle the image-quality issues common in BLV people's photographs. Our survey of 86 BLV people reveals their perspectives on understanding image-quality issues and errors when using VLM-based tools for product captioning, and the difficulties BLV people face in recovering from those errors. We then constructed an annotated dataset of 1,859 images taken by BLV people (729 high-quality, 1,130 low-quality images that are blurred, misframed, or rotated) with detailed product annotations—including product type (e.g., soup), brand (e.g., Campbell's), and variety (e.g., tomato, low-sodium)—and evaluated four different VLMs on it. We found that all VLMs experience a decline in product identification accuracy when image quality issues are present, with performance worsening when multiple issues are present. Moreover, we showed that each VLM is more or less susceptible to the studied image quality issues, suggesting ways to prioritize improving its performance. Making VLM-based captioning tools reliable will require collaboration among HCI and ML researchers and tool designers. Together, we will need to revisit the datasets used to evaluate these models; improve model performance through fine-tuning or inference-time techniques, especially for privacy-preserving open-source models; and design systems to provide richer feedback on VLM errors.

## Acknowledgments

We thank the Accessibility Research Collective at the University of California, Irvine, and the CollabLab at Northwestern University for helpful discussions. Research funding was provided by the National Science Foundation through awards SES-2326023 and SES-2326024.

## References

- [1] 2025. *The American Foundation for the Blind*. <https://www.afb.org/home>
- [2] 2025. *National Federation of the Blind*. <https://nfb.org/>
- [3] Rudaiba Adnai and Maitraye Das. 2024. “I Look at It as the King of Knowledge”: How Blind People Use and Understand Generative AI Tools. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's NL Canada, 2024-10-27). ACM, 1–14. doi:10.1145/3663548.3675631
- [4] Sakshi Agarwal, Gabe Hope, and Erik B. Sudderth. 2024. *VIPaint: Image Inpainting with Pre-Trained Diffusion Models via Variational Inference*. doi:10.48550/arXiv.2411.18929 arXiv:2411.18929 [cs]
- [5] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. 2024. ARNIQA: Learning Distortion Manifold for Image Quality Assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 189–198.
- [6] Dragan Ahmetovic, Daisuke Sato, Uran Oh, Tatsuya Ishihara, Kris Kitani, and Chiaki Asakawa. 2020. ReCog: Supporting Blind People in Recognizing Personal Objects. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376143
- [7] aira. Retrieved March 2025. Aira. <https://aira.io/>
- [8] Rahaf Alharbi, Robin N. Brewer, and Sarita Schoenebeck. 2022. Understanding Emerging Obfuscation Technologies in Visual Description Services for Blind and Low Vision People. 6 (2022), 1–33. Issue CSCW2. doi:10.1145/3555570
- [9] 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* (New York, NY, USA, 2024-10-27) (ASSETS '24). Association for Computing Machinery, 1–17. doi:10.1145/3663548.3675659
- [10] AllenAI. 2024. AllenAI/Molmo-72B-0924, Hugging Face. <https://huggingface.co/allenai/Molmo-72B-0924>.
- [11] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *Computer Vision – ECCV 2016* (Cham, 2016), Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, 382–398. doi:10.1007/978-3-319-46454-1\_24
- [12] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (Ann Arbor, Michigan, 2005-06), Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (Eds.). Association for Computational Linguistics, 65–72. <https://aclanthology.org/W05-0909/>
- [13] Be My Eyes. 2023. Introducing: Be My AI.
- [14] Be My Eyes. 2025. Be My Eyes. <https://www.bemyeyes.com/>.
- [15] Be My Eyes. Retrieved April, 2025. How Do I Use Be My AI? <https://support.bemyeyes.com/hc/en-us/articles/18133134809105-How-do-I-use-Be-My-AI>
- [16] Andy Beatman and Ailsa Leen. 2024. 6 Ways Generative AI Helps Improve Accessibility for All with Azure.
- [17] Nilavra Bhattacharya, Qing Li, and Danna Gurari. 2019. Why Does a Visual Question Have Different Answers?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019). 4271–4280. [http://openaccess.thecvf.com/content\\_ICCV\\_2019/html/Bhattacharya\\_Why\\_Does\\_a\\_Visual\\_Question\\_Have\\_Different\\_Answers\\_ICCV\\_2019\\_paper.html](http://openaccess.thecvf.com/content_ICCV_2019/html/Bhattacharya_Why_Does_a_Visual_Question_Have_Different_Answers_ICCV_2019_paper.html)
- [18] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandy White, Samuel White, and Tom Yeh. 2010. VizWiz: Nearly Real-Time Answers to Visual Questions. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology* (New York New York USA, 2010-10-03). ACM, 333–342. doi:10.1145/1866029.18666080
- [19] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. 2013. Visual Challenges in the Everyday Lives of Blind People. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris France, 2013-04-27). ACM, 2117–2126. doi:10.1145/2470654.2481291
- [20] Danielle Bragg, Naomi Caselli, Julie A Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E Ladner. 2021. The fate landscape of sign language ai datasets: An interdisciplinary perspective. *ACM Transactions on Accessible Computing (TACCESS)* 14, 2 (2021), 1–45.
- [21] Yang Trista Cao, Kyle Seelman, Kyungjun Lee, and Hal Daumé III. 2022. What's Different between Visual Question Answering for Machine "Understanding" Versus for Accessibility?. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 1025–1034. doi:10.18653/v1/2022.acl-main.75
- [22] Peter Carragher, Nikitha Rao, Abhinand Jha, R. Raghav, and Kathleen M. Carley. 2025. SegSub: Evaluating Robustness to Knowledge Conflicts and Hallucinations

in Vision-Language Models. *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media* 2025 (June 2025), 27. doi:10.36190/2025.27

[23] David M. Chan, Austin Myers, Sudheendra Vijayanarasimhan, David A. Ross, and John Canny. 2023. IC3: Image Captioning by Committee Consensus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8975–9003. doi:10.18653/v1/2023.emnlp-main.556

[24] 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 (UIST '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3654777.3673675

[25] Ruei-Che Chang, Yuxuan Liu, Lotus Zhang, and Anhong Guo. 2024. EditScribe: Non-Visual Image Editing with Natural Language Verification Loops. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3663548.3675599

[26] Ruei-Che Chang, Rosiana Natalie, Wenzian Xu, Jovan Zheng Feng Yap, and Anhong Guo. 2025. Probing the Gaps in ChatGPT's Live Video Chat for Real-World Assistance for People Who Are Blind or Visually Impaired. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3663547.3746319

[27] Meng Chen, Akhil Iyer, and Amy Pavel. 2025. Surfacing Variations to Calibrate Perceived Reliability of MLLM-generated Image Descriptions. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3663547.3746393

[28] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. *Microsoft COCO Captions: Data Collection and Evaluation Server*. doi:10.48550/arXiv.1504.00325 arXiv:1504.00325 [cs]

[29] 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 (ICML '24, Vol. 235)*. JMLR.org, Vienna, Austria, 7824–7846.

[30] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. 2020. Assessing Image Quality Issues for Real-World Problems. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3643–3653. doi:10.1109/CVPR42600.2020.00370

[31] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. 2023. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*.

[32] Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively Answering Ambiguous Questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 530–543. doi:10.18653/v1/2023.emnlp-main.35

[33] Nathan Davis, Bo Xie, and Danna Gurari. 2020. Quality of Images Showing Medication Packaging from Individuals with Vision Impairments: Implications for the Design of Visual Question Answering Applications. 57, 1 (2020), e251. doi:10.1002/pra2.251

[34] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damao Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fulu Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghai Lu, Shangyan Zhou, Shanhuan Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuteng Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjiu Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyu Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun

[35] Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Chen Huang, Chen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zhihu Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. DeepSeek-V3 Technical Report. doi:10.48550/arXiv.2412.19437 arXiv:2412.19437 [cs]

[36] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiazen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Witlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihis, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2025. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 91–104.

[37] Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or Vision: Do Vision-Language Models Have Blind Faith in Text?. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3867–3876. doi:10.1109/CVPR52734.2025.00366

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. doi:10.1109/CVPR.2009.5206848

[39] Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2023. ASL citizen: a community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems* 36 (2023), 76893–76907.

[40] Tim Dettmers and Luke Zettlemoyer. 2023. The Case for 4-Bit Precision: K-Bit Inference Scaling Laws. In *Proceedings of the 40th International Conference on Machine Learning (ICML '23, Vol. 202)*. JMLR.org, Honolulu, Hawaii, USA, 7750–7774.

[41] Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025. TruthPrInt: Mitigating Large Vision-Language Models Object Hallucination Via Latent Truthful-Guided Pre-Intervention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7372–7382.

[42] Emily Forlini. 2025. After Backlash, ChatGPT Removes Option to Have Private Chats Indexed by Google. *PCMAG* (August 2025).

[43] En-Vision America, Inc. 2025. Talking Prescription Labels | ScripTalk. <https://www.scriptability.com/scriptalk-talking-labels>.

[44] Zhiyuan Fan, Yumeng Wang, Sandeep Polisetty, and Yi R. Fung. 2025. Unveiling the Lack of LVLML Robustness to Fundamental Visual Variations: Why and Path Forward. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 20222–20242. doi:10.18653/v1/2025.findings-acl.1037

[45] Zilin Fang, Andrei Ignatov, Eduard Zamfir, and Radu Timofte. 2023. SQAD: Automatic Smartphone Camera Quality Assessment and Benchmarking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20532–20542.

[46] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T. Roweis, and William T. Freeman. 2006. Removing Camera Shake from a Single Photograph. 25, 3 (2006), 787–794. doi:10.1145/1141911.1141956

[47] Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. 2023. OPTQ: Accurate Quantization for Generative Pre-trained Transformers. In *The Eleventh International Conference on Learning Representations*. <https://iclr.cc/virtual/2023/poster/10855>

[48] Vinitha Gadireaju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Remi Denton, and Robin Brewer. 2023. “I wouldn’t say offensive but...”: Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) (FAccT ’23). Association for Computing Machinery, New York, NY, USA, 205–216. doi:10.1145/3593013.3593989

[49] 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* (New York, NY, USA, 2023-10-22) (ASSETS ’23). Association for Computing Machinery, 1–21. doi:10.1145/3597638.3608955

[50] S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. 2022. No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications*

of Computer Vision. 1220–1230.

[50] 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 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3613904.3642211

[51] Steven M. Goodman, Ping Liu, Dhruv Jain, Emma J. McDonnell, Jon E. Froehlich, and Leah Findlater. 2021. Toward User-Driven Sound Recognizer Personalization with People Who Are d/Deaf or Hard of Hearing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 63 (June 2021), 23 pages. doi:10.1145/3463501

[52] Google. 2025. Gemini 2.5 Flash. <https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash>.

[53] Aaron Graftafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferre, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esioubi, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillen Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Ilyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, prefix=van der useprefix=false family=Linde, given=Jelmer, Jennifer Billock, Jenny Hong, Jena Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasudevan Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Kirthika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, prefix=van der useprefix=false family=Maaten, given=Laurens, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, prefix=de useprefix=false family=Oliveira, given=Luke, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Celebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Saha Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rarapthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sooto, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gouquet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhao Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Couderet, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anani Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arka Bandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardt, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beatty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testugine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcuate, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Gunu Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Harouf Habeeb, Harrison Rudolph, Helen Suk, Henry Asperegn, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelema, Kegian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wherstedt, Madiam Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mihay Vyatatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghatham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadhi Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjin Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wencheng Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. *The Llama 3 Herd of Models*. doi:10.48550/arXiv.2407.21783 arXiv:2407.21783 [cs]

[54] Danna Gurari and Kristen Grauman. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017-05-02) (CHI '17). Association for Computing Machinery, 3511–3522. doi:10.1145/3025453.3025781

[55] 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3608–3617. doi:10.1109/CVPR.2018.00380

[56] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning Images Taken by People Who Are Blind. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Vol. 12362. Springer International Publishing, 417–434. doi:10.1007/978-3-030-58520-4\_25

[57] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*.

[58] Jaylin Herskovitz, Andi Xu, Rahaf Alharbi, and Anhong Guo. 2024. ProgramAlly: Creating Custom Visual Access Programs via Multi-Modal End-User Programming. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3654777.3676391

[59] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Performance-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7514–7528. doi:10.18653/v1/2021.emnlp-main.595

[60] Jonggi Hong, Jaina Gandhi, Ernest Essuah Mensah, Farnaz Zamiri Zeraati, Ebrima Jarjue, Kyungjun Lee, and Hernisa Kacorri. 2022. Blind Users Accessing Their Training Images in Teachable Object Recognizers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens Greece, 2022-10-23). ACM, 1–18. doi:10.1145/3517428.3544824

[61] Jonggi Hong and Hernisa Kacorri. 2024. Understanding How Blind Users Handle Object Recognition Errors: Strategies and Challenges. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA, 2024-10-27) (ASSETS '24). Association for Computing Machinery, 1–15. doi:10.1145/3663548.3675635

[62] 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*. 1–17.

[63] Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Hansika Murugu, Danna Gurari, Eunsol Choi, and Amy Pavel. 2024. Long-Form Answers to Visual Questions from Blind and Low Vision People. In *First Conference on Language Modeling*.

[64] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P. Bigham. 2011. Supporting Blind Photography. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '11). Association for Computing Machinery, New York, NY, USA, 203–210. doi:10.1145/2049536.2049573

[65] Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. TIGER: Text-to-Image Grounding for Image Caption Evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 2141–2152. doi:10.18653/v1/D19-1220

[66] Hernisa Kacorri, Kris M. Kitani, Jeffrey P. Bigham, and Chieko Asakawa. 2017. People with Visual Impairment Training Personal Object Recognizers: Feasibility and Challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2017-05-02) (CHI '17). Association for Computing Machinery, 5839–5849. doi:10.1145/3025453.3025899

[67] Rhea Kapur and Elisa Kreiss. 2024. Reference-Based Metrics Are Biased Against Blind and Low-Vision Users’ Image Description Preferences. In *Proceedings of the Third Workshop on NLP for Positive Impact*. 308–314.

[68] Insoo Kim, Seungju Han, Ji-won Baek, Seong-Jin Park, Jae-Joon Han, and Jinwoo Shin. 2021. Quality-Agnostic Image Recognition via Invertible Decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12257–12266.

[69] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. “I’m Not Sure, But...”: Examining the Impact of Large Language Models’ Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24)*. Association for Computing Machinery, New York, NY, USA, 822–835. doi:10.1145/3630106.3658941

[70] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*.

[71] Patrick E Lanigan, Aaron M Paulos, Andrew W Williams, Dan Rossi, and Priya Narasimhan. 2006. Trinetra: Assistive Technologies for Grocery Shopping for the Blind. In *ISWC*. 147–148.

[72] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. Vilbertscore: Evaluating image caption using vision-and-language bert. In *Proceedings of the first workshop on evaluation and comparison of NLP systems*. 34–39.

[73] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 2022. Image-Explorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions. In *CHI Conference on Human Factors in Computing Systems* (New Orleans LA USA, 2022-04-29). ACM, 1–15. doi:10.1145/3491102.3501966

[74] Kyungjun Lee, Jonggi Hong, Simone Pimento, Ebrima Jarjue, and Hernisa Kacorri. 2019. Revisiting Blind Photography in the Context of Teachable Object Recognizers. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA, 2019-10-24) (ASSETS '19). Association for Computing Machinery, 83–95. doi:10.1145/3308561.3353799

[75] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*. 201–216.

[76] Sooyeon Lee, Madison Reddie, and John M. Carroll. 2021. Designing for Independence for People with Visual Impairments. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 149 (April 2021), 19 pages. doi:10.1145/3449223

[77] Sooyeon Lee, Madison Reddie, Chun-Hui Tsai, Jordan Beck, Mary Beth Rosson, and John M Carroll. 2020. The emerging professional practice of remote sighted assistance for people with visual impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[78] 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. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2024), 13872–13882. doi:10.1109/CVPR52733.2024.01316

[79] Qisheng Li and Shaomei Wu. 2024. “I Want to Publicize My Stutter”: Community-led Collection and Curation of Chinese Stuttered Speech Data. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–27.

[80] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, Nadine Chang, Karan Sapra, Amala Sanjay Deshmukh, Tuomas Rintamaki, Matthieu Le, Ilia Karmanov, Lukas Voegtle, Philipp Fischer, De-An Huang, Timo Roman, Tong Lu, Jose M. Alvarez, Bryan Catanzaro, Jan Kautz, Andrew Tao, Guilin Liu, and Zhiding Yu. 2025. Eagle 2: Building Post-Training Data Strategies from Scratch for Frontier Vision-Language Models. doi:10.48550/arXiv.2501.14818 arXiv:2501.14818 [cs]

[81] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out* (Barcelona, Spain, 2004-07). Association for Computational Linguistics, 74–81. https://aclanthology.org/W04-1013/

[82] 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*. David Fleet, Tomas Pajdla, Bernt Schiele, and Timme Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755. doi:10.1007/978-3-319-10602-1\_48

[83] Allison Linn. 2016. Decades of Computer Vision Research, One ‘Swiss Army Knife’.

[84] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *Advances in Neural Information Processing Systems* 36 (December 2023), 34892–34916.

[85] Huan Liu, Lingyu Xiao, Jiangjiang Liu, Xiaofan Li, Ze Feng, Sen Yang, and Jingdong Wang. 2024. Revisiting MLLMs: An In-Depth Analysis of Image Classification Abilities. *CoRR* abs/2412.16418 (December 2024).

[86] Connie Loizos. 08-28-25. Anthropic Users Face a New Choice – Opt out or Share Your Chats for AI Training. *TechCrunch* (08-28-25).

[87] Ximing Lu et al. 2023. Inference-Time Policy Adapters (IPA): Tailoring Extreme-Scale LMs without Fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 6863–6883.

[88] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. 2018. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia* 20, 11 (2018), 3111–3122.

[89] Xiaoyu Ma, Chenxi Feng, Jiaoqiao Wang, Qiang Lin, Suiyu Zhang, Jinchi Zhu, Xiaodiao Chen, Chang Liu, and Dingguo Yu. 2023. A Model-Agnostic Semantic-Quality Compatible Framework Based on Self-Supervised Semantic Decoupling. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. Association for Computing Machinery, New York, NY, USA, 6774–6784. doi:10.1145/3581783.3613775

[90] 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* (New York, NY, USA, 2017-05-02) (CHI '17). Association for Computing Machinery, 5988–5999. doi:10.1145/3025453.3025814

[91] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9004–9017. doi:10.18653/v1/2023.emnlp-main.557

[92] Maniratnam Mandal, Deepthi Ghadiyaram, Danna Gurari, and Alan C. Bovik. 2023. Helping Visually Impaired People Take Better Quality Pictures. 32 (2023), 3873–3884. doi:10.1109/TIP.2023.3282067

[93] H. B. Mann and D. R. Whitney. 1947. On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. 18, 1 (1947), 50–60. jstor:2236101 https://www.jstor.org/stable/2236101

[94] Meta. 2024. Meta-Llama/Llama-3.2-90B-Vision-Instruct · Hugging Face. https://huggingface.co/meta-llama/Llama-3.2-90B-Vision-Instruct.

[95] Meta. 2025. Introducing the Meta AI App: A New Way to Access Your AI Assistant.

[96] Ananya Gubbi Mohanbabu and Amy Pavel. 2024. Context-Aware Image Descriptions for Web Accessibility. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA, 2024-10-27) (ASSETS '24). Association for Computing Machinery, 1–17. doi:10.1145/3663548.3675658

[97] Cecily Morrison, Martin Grayson, Rita Faia Marques, Daniela Massiceti, Camilla Longden, Linda Wen, and Edward Cutrell. 2023. Understanding Personalized Accessibility through Teachable AI: Designing and Evaluating Find My Things for People Who Are Blind or Low Vision. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA, 2023-10-22) (ASSETS '23). Association for Computing Machinery, 1–12.

doi:10.1145/3597638.3608395

[98] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2023. Improving Multimodal Datasets with Image Captioning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, 22047–22069.

[99] OpenAI. 2025. *Prompt Engineering - OpenAI API*. <https://platform.openai.com/docs/guides/prompt-engineering>

[100] OpenAI. 2025. ChatGPT. <https://chatgpt.com/>.

[101] OpenAI. 2025. GPT-4.1. <https://platform.openai.com/docs/models/gpt-4.1>.

[102] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02* (Philadelphia, Pennsylvania, 2001). Association for Computational Linguistics, 311. doi:10.3115/1073083.1073135

[103] Sohyeon Park, Aehong Min, Jesus Armando Beltran, and Gillian R Hayes. 2025. “As an Autistic Person Myself.” The Bias Paradox Around Autism in LLMs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 774, 17 pages. doi:10.1145/3706598.3713420

[104] Vishnu Prabhakaran, Purav Aggarwal, Vishruth Kulshreshtha, Arunita Das, Sahini Venkata Sitaram Sruti, and Anoop Saladi. 2025. VIT-Pro: Visual Instruction Tuning for Product Images. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, Weizhu Chen, Yi Yang, Mohammad Kachuee, and Xue-Yong Fu (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 695–707.

[105] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoya Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. 2022. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision* 130, 8 (2022), 2022–2039.

[106] Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. *Advances in Neural Information Processing Systems* 35 (2022), 9538–9551.

[107] Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. 2024. Benchmarking Robustness of Multimodal Image-Text Models under Distribution Shift. *Journal of Data-centric Machine Learning Research* [Jan. (2024)]. doi:10.48550/arXiv.2212.08044 arXiv:2212.08044 [cs]

[108] Sreetama Sarkar, Yue Che, Alex Gavin, Peter Anthony Beeler, and Souvik Kundu. 2025. Mitigating Hallucinations in Vision-Language Models through Image-Guided Head Suppression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 12492–12511. doi:10.18653/v1/2025.emnlp-main.631

[109] Ashutosh Saxena, Justin Diemeyer, and Andrew Y Ng. 2009. Learning 3-d object orientation from images. In *2009 IEEE International conference on robotics and automation*. IEEE, 794–800.

[110] Hope Schroeder, Deb Roy, and Jad Kabbara. 2025. Just Put a Human in the Loop? Investigating LLM-Assisted Annotation for Subjective Tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 25771–25795.

[111] Tanusree Sharma, Abigale Stangl, Lotus Zhang, Yu-Yun Tseng, Inan Xu, Leah Findlater, Danna Gurari, and Yang Wang. 2023. Disability-First Design and Creation of A Dataset Showing Private Visual Information Collected With People Who Are Blind. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3544548.3580922

[112] Arielle M Silverman, Sarah Elizabeth J. Baguhn, Mei-Lian Vader, Emily M. Romero, and Chung Ho Philip So. 2025. *Empowering or Excluding: Expert Insights on Inclusive Artificial Intelligence for People With Disabilities*. Technical Report. American Foundation for the Blind.

[113] Rachel N. Simons, Danna Gurari, and Kenneth R. Fleischmann. 2020. “I Hope This Is Helpful”: Understanding Crowdworkers’ Challenges and Motivations for an Image Description Task. 4 (2020), 1–26. Issue CSCW2. doi:10.1145/3415176

[114] Abigale Stangl, Emma Sadjo, Pardis Emami-Naeini, Yang Wang, Danna Gurari, and Leah Findlater. 2023. “Dump It, Destroy It, Send It to Data Heaven”: Blind People’s Expectations for Visual Privacy in Visual Assistance Technologies. In *Proceedings of the 20th International Web for All Conference* (New York, NY, USA, 2023-04-30) (W4A '23). Association for Computing Machinery, 134–147. doi:10.1145/3587281.3587296

[115] Abigale Stangl, Kristina Shiroma, Nathan Davis, Bo Xie, Kenneth R. Fleischmann, Leah Findlater, and Danna Gurari. 2022. Privacy Concerns for Visual Assistance Technologies. *ACM Trans. Access. Comput.* 15, 2, Article 15 (May 2022), 43 pages. doi:10.1145/3517384

[116] Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who Are Blind or Have Low Vision. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA, 2021-10-17) (ASSETS '21). Association for Computing Machinery, 1–15. doi:10.1145/3441852.3471233

[117] Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyu Hu, Lukas W. Mayer, and Padhraic Smyth. 2025. What Large Language Models Know and What People Think They Know. 7, 2 (2025), 221–231. doi:10.1088/s42256-024-00976-7

[118] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. 2015. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 769–777.

[119] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large Language Models for Data Annotation and Synthesis: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 930–957. doi:10.18653/v1/2024.emnlp-main.54

[120] 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*, 1–17.

[121] 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 *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3746059.3747597

[122] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM<sup>2</sup>)*, Ofir Ariv, Miruna Clinciu, Kaustubh Dhole, Rotem Dror, Sebastian Gehrmann, Eliya Habba, Itay Itzhak, Simon Mille, Yotam Perlitz, Enrico Santus, João Sedoc, Michal Shmueli Scheuer, Gabriel Stanovsky, and Oyvind Tafjord (Eds.). Association for Computational Linguistics, Vienna, Austria and virtual meeting, 404–430.

[123] The R Foundation. 2025. R: The R Project for Statistical Computing. <https://www.r-project.org/>.

[124] Lida Theodorou, Daniela Massiceti, Luisa Zintgraf, Simone Stumpf, Cecily Morrison, Edward Cutrell, Matthew Tobias Harris, and Katja Hofmann. 2021. Disability-first dataset creation: Lessons from constructing a dataset for teachable object recognition with blind and low vision data collectors. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, 1–12.

[125] Ameni Trabelsi, Maria Zontak, Yiming Qian, Brian Jackson, Suleiman Khan, and Umit Batur. 2025. What Matters When Building Vision Language Models for Product Image Analysis? (2025).

[126] Tess Van Daele, Akhil Iyer, Yuning Zhang, Jalyn C Derry, Mina Huh, and Amy Pavel. 2024. Making short-form videos accessible with hierarchical video summaries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–17.

[127] Marynel Vázquez and Aaron Steinfeld. 2014. An Assisted Photography Framework to Help Visually Impaired Users Properly Aim a Camera. *ACM Trans. Comput.-Hum. Interact.* 21, 5 (Nov. 2014), 25:1–25:29. doi:10.1145/2651380

[128] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566–4575. doi:10.1109/CVPR.2015.7299087

[129] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA, USA, 2015-06). IEEE, 3156–3164. doi:10.1109/CVPR.2015.7298935

[130] Tess Winlock, Eric Christiansen, and Serge Belongie. 2010. Toward Real-Time Grocery Detection for the Visually Impaired. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 49–56. doi:10.1109/CVPRW.2010.5543576

[131] Tsung-Han Wu, Heekyung Lee, Jiaxin Ge, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. 2025. Generate, but Verify: Reducing Hallucination in Vision-Language Models with Retrospective Resampling. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*.

[132] Jingyi Xie, Rui Yu, He Zhang, Syed Masum Billah, Sooyeon Lee, and John M. Carroll. 2025. Beyond Visual Perception: Insights from Smartphone Interaction of Visually Impaired Users with Large Multimodal Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3706598.3714210

[133] Yasin Abbasi Yakhori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba

Szepesvári, et al. 2024. Mitigating LLM Hallucinations via Conformal Abstention. *arXiv preprint arXiv:2405.01563* (2024).

[134] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Kegin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yingfei Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. doi:10.48550/arXiv.2505.09388 arXiv:2505.09388 [cs]

[135] Chun-Ju Yang, Kristen Grauman, and Danna Gurari. 2018. Visual Question Answer Diversity. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (2018), Vol. 6. 184–192. <https://ojs.aaai.org/index.php/HCOMP/article/view/13341>

[136] Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. 2025. Nullu: Mitigating Object Hallucinations in Large Vision-Language Models via HalluSpace Projection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14635–14645. doi:10.1109/CVPR52734.2025.01364

[137] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. 2022. MANIQA: Multi-Dimension Attention Network for No-Reference Image Quality Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1191–1200.

[138] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. 2012. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE conference on computer vision and pattern recognition*. 1083–1090.

[139] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination Correction for Multimodal Large Language Models. *Science China Information Sciences* 67, 12 (December 2024), 220105. doi:10.1007/s11432-024-4251-x

[140] Gal Yona, Roei Aharoni, and Mor Geva. 2024. Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words?. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 7752–7764. doi:10.18653/v1/2024.emnlp-main.443

[141] Dokyoon Yoon, Youngsook Song, and Wooyong Park. 2025. Stop Learning It All to Mitigate Visual Hallucination, Focus on the Hallucination Target. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2025), 4200–4208. doi:10.1109/CVPR52734.2025.00397

[142] Lu Yu, Malvina Nikandrou, Jiali Jin, and Verena Rieser. 2023. Quality-Agnostic Image Captioning to Safely Assist People with Vision Impairment. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI '23)*. Macao, P.R.China, 6281–6289. doi:10.24963/ijcai.2023/697

[143] Xiaoyu Zeng, Yanan Wang, Tai-Yin Chiu, Nilavra Bhattacharya, and Danna Gurari. 2020. Vision Skills Needed to Answer Visual Questions. 4 (2020), 1–31. Issue CSCW2. doi:10.1145/3415220

[144] Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, Chunyuan Li, and Manling Li. 2024. HallE-Control: Controlling Object Hallucination in Large Multimodal Models. doi:10.48550/arXiv.2310.01779 arXiv:2310.01779 [cs]

[145] Guanqi Zhan, Weidi Xie, and Andrew Zisserman. 2022. A Tri-Layer Plugin to Improve Occluded Detection. In *BMVC*.

[146] Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. 2024. Reflective Instruction Tuning: Mitigating Hallucinations in Large Vision-Language Models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXVIII*. Springer-Verlag, Berlin, Heidelberg, 196–213. doi:10.1007/978-3-031-73113-6\_12

[147] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*. <https://openreview.net/forum?id=SkeHuCVFDr>

[148] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. 2024. Why Are Visually-Grounded Language Models Bad at Image Classification? *Advances in Neural Information Processing Systems* 37 (December 2024), 51727–51753.

[149] Youjian Zhang, Chaoyue Wang, Stephen J Maybank, and Dacheng Tao. 2021. Exposure trajectory recovery from motion blur. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 7490–7504.

[150] Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2025. Mitigating Object Hallucination in Large Vision-Language Models via Image-Grounded Guidance. In *Forty-Second International Conference on Machine Learning*.

[151] Yuhang Zhao, Sarit Szpiro, Jonathan Knighten, and Shiri Azenkot. 2016. CueSee: Exploring Visual Cues for People with Low Vision to Facilitate a Visual Search Task. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg Germany, 2016-09-12). ACM, 73–84. doi:10.1145/2971648.2971730

[152] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2018. A Face Recognition Application for People with Visual Impairments: Understanding Use Beyond the Lab. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC Canada, 2018-04-21). ACM, 1–14. doi:10.1145/3173574.3173789

[153] Yi Zhao, Yilin Zhang, Rong Xiang, Jing Li, and Hillming Li. 2024. VIALM: A Survey and Benchmark of Visually Impaired Assistance with Large Models. doi:10.48550/arXiv.2402.01735 arXiv:2402.01735 [cs]

[154] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, 46595–46623.

[155] 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. In *The Twelfth International Conference on Learning Representations*.

## A Crowdworker Ratings for Captionability of Images and Image Quality Issues in Dataset

Table 8 details how many crowdworkers found images captionable (from Gurari et al. [55]) and the presence of image quality issues (from Chiu et al. [30]) for our subset of 1,859 images.

## B Additional Examples of Product Captioning Performance on Low-Quality Images

Tables 9, 10, 11, 12, and 13 provide additional examples of how the studied VLMs fail for various image quality issues.

## C Image Captioning Prompt for All VLMs

You are a helpful assistant who identifies products in images for blind and low-vision individuals. Identify the product in the image while following these guidelines:

- (1) Identify crucial features about the product, including:
  - (a) Object type (can, bag, plastic container, etc.)
  - (b) Product type (prepared or frozen meal, seasoning mix, soda, coffee)
  - (c) Brand (Heinz, Coca-Cola, Starbucks, etc.)
  - (d) Variety (specific flavors, sizes, count of items, etc.)
  - (e) Visual features (color, shape, size, etc.)
- (2) Use clear, direct, and objective language. Do not use vague adjectives like ‘large’ or ‘small’, or vague adverbs like ‘prominently’ or ‘clearly’.
- (3) DO NOT mention camera artifacts (e.g., blur) or if an object is partially visible.
- (4) DO NOT use introductory phrases (e.g., ‘The image shows’, ‘The object is’, ‘The primary object is’).

**Table 8:** The final dataset for Study 2 included 1,859 images taken by BLV people, with 729 images being high-quality images and 1,130 being low-quality images. Each image has at least three captions from crowdworkers (i.e., no more than 3 people said the image was Unrecognizable). High-quality images have no image quality issues > 1; low-quality images have at least one issue for which  $\geq 4$  crowdworkers reported it. The 17 high-quality images with rotation  $\geq 4$  were images that only had a rotation issue (noted by the crowdworkers) but were actually not rotated, as checked by two researchers (see Section 4.2.2); since they had no other issues, we moved these into the high-quality subset. Each row indicates the number of crowdworkers who reported that an image was unrecognizable or had the specified image-quality issue. Percentages are column-wise.

Image Type	Num. Crowdworkers	Unrecog.	Blur	Framing	Rotation	Obstruction	Too Dark	Too Bright	Other	No Issue
High-Quality (729 Images)	0	706 (96.8%)	620 (85.1%)	529 (72.6%)	687 (94.2%)	718 (98.5%)	729 (100.0%)	726 (99.6%)	721 (98.9%)	17 (2.3%)
	1	18 (2.5%)	109 (15.0%)	200 (27.4%)	25 (3.4%)	11 (1.5%)	0	3 (0.4%)	8 (1.1%)	0
	2	5 (0.7%)	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0
	4	0	0	0	14 (1.9%)	0	0	0	0	346 (47.5%)
	5	0	0	0	3 (0.4%)	0	0	0	0	366 (50.2%)
Low-Quality (1,130 Images)	0	936 (82.8%)	311 (27.5%)	168 (14.9%)	585 (51.8%)	1018 (90.1%)	973 (86.1%)	960 (85.0%)	1081 (95.7%)	1130 (100.0%)
	1	194 (17.2%)	173 (15.3%)	137 (12.2%)	112 (9.9%)	88 (7.8%)	133 (11.8%)	131 (11.6%)	48 (4.3%)	0
	2	0	106 (9.4%)	132 (11.7%)	90 (8.0%)	6 (0.5%)	18 (1.6%)	23 (2.0%)	1 (0.1%)	0
	3	0	76 (6.7%)	136 (12.0%)	75 (6.6%)	4 (0.4%)	2 (0.2%)	8 (0.7%)	0	0
	4	0	278 (24.6%)	395 (35.0%)	209 (18.5%)	10 (0.9%)	3 (0.3%)	5 (0.4%)	0	0
	5	0	186 (16.5%)	162 (14.3%)	59 (5.2%)	4 (0.4%)	1 (0.1%)	3 (0.3%)	0	0

**Table 9:** Examples of blurred product images where VLMs may only provide high-level information or incorrectly infer what the product is. Captions were shortened for presentation purposes only, indicated by [...].

Image	Annotation	GPT	Gemini	Llama	Molmo
	<b>Product:</b> [Correct] Box of Quaker brand instant oatmeal in the Raisin, Date & Walnut variety, with a yellow and blue color scheme and the Quaker logo featuring a man in a hat. <b>Brand:</b> Quaker <b>Variety:</b> raisin; date; walnut	[Correct] Box of Quaker brand instant oatmeal in the Raisin, Date & Walnut variety, with a yellow and blue color scheme and the Quaker logo featuring a man in a hat.	[Correct] A yellow box of Quaker Instant Oatmeal in the Raisin, Date, and Walnut flavor. The Quaker logo, a man in a hat, is shown above the word "Quaker" on a blue stripe. Below this, a red rectangle states "Raisin, Date & Walnut" in white text.	[Incorrect] The product is a box of Quaker Old Fashioned Oats, featuring a blue and yellow color scheme with a red rectangle on the front. The box has a rectangular shape and a medium size.	[Incorrect] A rectangular box of Quaker oats, specifically oatmeal packets. The brand name "QUAKER" is visible on the packaging, with additional text indicating the flavor "raisin, date, and walnut".
	<b>Product:</b> [Incorrect] Aluminum can with a pink top, featuring a white awareness ribbon design, commonly associated with breast cancer awareness campaigns. Specific brand and variety are not identifiable from visible features. <b>Brand:</b> Yoplait <b>Missing:</b> - product: yogurt - brand: Yoplait	[Incorrect] Aluminum can with a pink top, featuring a white awareness ribbon design, commonly associated with breast cancer awareness campaigns. Specific brand and variety are not identifiable from visible features.	[Incorrect] A pink, spherical chocolate candy in a shiny metallic wrapper. A white ribbon design is on the front, along with the word "GODIVA" in white lettering.	[Incorrect] A cylindrical, red can with a white logo and text. The logo is a stylized letter "D" with the brand name "Dr Pepper" written in white text next to it. [...]	[Incorrect] Coffee bag from Eight O'Clock Coffee brand. Red packaging with white branding visible.  Missing: - product: yogurt - brand: Yoplait

**Table 10: Examples of images illustrating how framing affects product identification and resulting captions. In the Corn Pops and McCormick Great Guacamole examples, all VLMs fail to fill in the missing information needed for correct identification. The Honey Nut Cheerios example provides two alternate framings, with varying amounts of the text visible. Despite the cereal's mascot being visible on both, Llama and Molmo fail to correctly identify the product when more of the product text is hidden. Captions were shortened for presentation purposes only, indicated by [...].**

Image	Annotation	GPT	Gemini	Llama	Molmo
	<b>Product:</b> cereal <b>Brand:</b> Kellogg's <b>Variety:</b> Corn Pops	<b>[Incorrect]</b> Cardboard box of breakfast cereal, likely corn flakes, with yellow flakes pictured in a bowl of milk and bold red and blue outlined letters on a yellow background. <b>Missing:</b> - brand: Kellogg's - variety: Corn Pops	<b>[Incorrect]</b> A rectangular box of Cap'n Crunch's Crunch Berries cereal. The box is yellow and red with an image of crunchy cereal pieces. <b>Missing:</b> - brand: Kellogg's - variety: Corn Pops	<b>[Incorrect]</b> The product is a box of Kraft macaroni and cheese. The box is yellow and features the brand name in red letters, with a black and white logo in the top-left corner. The background of the box is yellow, and it has a white border around the edges. <b>Missing:</b> - product: cereal - brand: Kellogg's - variety: Corn Pops	<b>[Incorrect]</b> Coffee-Mate French Vanilla coffee creamer bottle, white with yellow label, on yellow background. 32 fl oz (946 ml) size. Red lettering on left side. Partial yellow and black text on right, "CASH" visible. <b>Missing:</b> - product: cereal - brand: Kellogg's - variety: Corn Pops
	<b>Product:</b> seasoning mix <b>Brand:</b> McCormick Produce Partners <b>Variety:</b> Great Guacamole	<b>[Incorrect]</b> Packet of McCormick Produce Partners Great Sample seasoning mix, featuring a yellow label with black and red text, outlined in green and black. <b>Missing:</b> - variety: Great Guacamole	<b>[Incorrect]</b> A rectangular pouch of McCormick Produce Partners Great American Potato seasoning mix. The top is black, and a red banner with the white McCormick logo is below it. The main part of the pouch is yellow with green borders. <b>Missing:</b> - variety: Great Guacamole	<b>[Incorrect]</b> The product is a bag of McCormick's "Great Sample" Produce Partners seasoning mix, featuring the brand name "McCormick" and the product name "Produce Partners" in black text on a yellow background. [...] <b>Missing:</b> - variety: Great Guacamole	<b>[Incorrect]</b> McCormick Produce Partners Great (partially visible) package. Black top, yellow background with green outline. Red and black text, McCormick logo with blue M and red C. Red banner with white text "Produce Partners". <b>Missing:</b> - product: seasoning mix - variety: Great Guacamole
	<b>Product:</b> cereal <b>Brand:</b> Cheerios <b>Variety:</b> Honey Nut	<b>[Correct]</b> Cereal box, Honey Nut Cheerios, General Mills brand, honey nut flavor, features a cartoon bee mascot and a bowl of cereal with honey being drizzled, predominantly orange and yellow color scheme.	<b>[Correct]</b> A box of General Mills Honey Nut Cheerios cereal. The box is orange, with a cartoon bee character holding a honey dipper over a red bowl of cereal with milk.	<b>[Correct]</b> The product is a red and orange cereal box featuring the Honey Nut Cheerios brand and variety. The box is adorned with a yellow bee character and contains a bowl of cereal with a honey dipper on top.	<b>[Correct]</b> Cereal box: Type: Box Product: Cereal Brand: Kellogg's Variety: Honey Nut Cheerios Shape: Rectangular Color scheme: Orange and yellow Prominent graphic: Honeybee character Main graphic: Bowl of Cheerios cereal with milk, honey drizzled on top
	<b>Product:</b> cereal <b>Brand:</b> Cheerios <b>Variety:</b> Honey Nut	<b>[Correct]</b> Cardboard box of Honey Nut Cheerios cereal by General Mills, featuring the honey bee mascot holding a honey dipper, with round oat cereal pieces and text highlighting it as "America's Favorite Cereal." The box has an orange and blue color scheme. Net weight is 1 lb 5 oz (595 g).	<b>[Correct]</b> Cereal box of Honey Nut Cheerios. The box is red with blue and yellow accents. It features a cartoon bee holding a honey dipper, and a bowl of cereal. The text "America's Favorite Cereal" is written on the bottom. The net weight is 1 pound, 1 ounce (481 grams).	<b>[Incorrect]</b> Product Description - Object Type: Plastic container - Product Type: Prepared meal - Brand: Heinz - Variety: Baked beans[...] <b>Missing:</b> - product: cereal - brand: Cheerios - variety: Honey Nut	<b>[Incorrect]</b> Cereal box: Kellogg's Froot Loops. Red background. Yellow text on blue banner: "America's Favorite Cereal". Animated bee character holding a Froot Loop. Visible net weight: 11oz. Standing on wooden floor. <b>Missing:</b> - brand: Cheerios - variety: Honey Nut

**Table 11: Examples of images with rotation issues where different VLMs may only provide high-level information or incorrectly infer what the product is. Captions were shortened for presentation purposes only, indicated by [...].**

Image	Annotation	GPT	Gemini	Llama	Molmo
	<p><b>Product:</b> Deli Napoli Meatballs</p> <p><b>Brand:</b> Lite n' Easy</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- product: Deli Napoli Meatballs</li> <li>- brand: Lite n' Easy</li> </ul>	<p><b>[Incorrect]</b> Round plastic container of DEKA brand Angel Cookies. The lid is white with a red and white label that features the brand logo and product name.</p>	<p><b>[Incorrect]</b> A round, plastic, white container of <b>Deli Napoli Meatballs</b>, 100g. The lid is white with a red stripe across the top and the words "Deli Easy" in white font. Below that, the words "Napoli Meatballs" are written in black cursive font. The bottom of the lid has a red recycling symbol and the words "Store below 4°C."</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- brand: Lite n' Easy</li> </ul>	<p><b>[Incorrect]</b> This is a plastic container of meatballs from the brand IDEAL, with a white label and red accents. The label features black text, including the product name "Meatballs" and a handwritten note. The container has a round shape and is placed on a beige surface, accompanied by a black object in the bottom right corner.</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- product: Deli Napoli Meatballs</li> <li>- brand: Lite n' Easy</li> </ul>	<p><b>[Incorrect]</b> Kraft Philadelphia cream cheese container, 8 oz.</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- product: Deli Napoli Meatballs</li> <li>- brand: Lite n' Easy</li> </ul>
	<p><b>Product:</b> juice</p> <p><b>Brand:</b> Dawn</p> <p><b>Variety:</b> Smooth 'n Juicy</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- brand: Dawn</li> </ul>	<p><b>[Incorrect]</b> Carton of Dalin <b>Smooth n' Juicy</b> orange juice, featuring an image of orange slices and juice splash on a blue and orange background.</p>	<p><b>[Incorrect]</b> A blue and orange juice box is oriented vertically. [...] The juice box is Dairymen brand orange juice. The top of the juice box is blue and has a logo in white that says Dairymen. [...] Beneath the logo, the words "Smooth 'n Juicy" are written in yellow and orange. [...]</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- brand: Dawn</li> </ul>	<p><b>[Incorrect]</b> The product is a rectangular-shaped carton of orange juice with the top-left corner of the carton facing the viewer. [...] The brand name "Dair" is written in white text on the top-right corner of the carton, while the product name "Smooth 'n Juicy" is displayed in yellow and white 3D-style text on the center-left side. [...]</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- brand: Dawn</li> </ul>	<p><b>[Incorrect]</b> A blue rectangular container of orange juice smoothie. The container displays an image of oranges and has "Smooth" in white text and "Juice" in yellow text with a drop shadow. The brand "Dewon" is written in cursive script, with the "W" featuring a leaf design. The container is placed at an angle on a light brown-orange countertop.</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- brand: Dawn</li> <li>- variety: Smooth 'n Juicy</li> </ul>

**Table 12: Examples of images with multiple issues (blur, framing, and rotation) with variations in product captions. Two slightly different frames of the same product, a package of Mucinex Expectorant Maximum Strength medication, are shown along with generated captions. Captions were shortened for presentation purposes only, indicated by [...].**

Image	Annotation	GPT	Gemini	Llama	Molmo
	<p><b>Product:</b> expectorant</p> <p><b>Brand:</b> Mucinex</p> <p><b>Variety:</b> maximum strength</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- variety: maximum strength</li> </ul>	<p><b>[Incorrect]</b> Box of <b>Mucinex expectorant</b>, blue and white packaging, designed to relieve chest congestion, thin and loosen mucus, and intended for adults and children over 12 years.</p>	<p><b>[Incorrect]</b> A rectangular box of <b>Mucinex expectorant</b>, extended-release bi-layer tablets. The box is white with the "Mucinex" brand name printed in blue letters. It contains 40 tablets and lists "Relieves chest congestion" and "Thins and loosens mucus" in blue text on a blue background on the side of the box.</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- variety: maximum strength</li> </ul>	<p><b>[Incorrect]</b> The product is a white box with a red stripe and blue accents, featuring the brand name "Aucinex" in dark blue text. [...] along with additional text in smaller white letters on a blue background. The background of the image is a light-colored wood surface, likely a table or countertop.</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- product: expectorant</li> <li>- brand: mucinex</li> <li>- variety: maximum strength</li> </ul>	<p><b>[Incorrect]</b> Aucinex expectorant box. It's a medication product, likely used for respiratory conditions. The box has a white background with a blue strip and red curvy line. The brand name "Aucinex" is visible in blue text.</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- brand: mucinex</li> <li>- variety: maximum strength</li> </ul>
	<p><b>Product:</b> expectorant</p> <p><b>Brand:</b> Mucinex</p> <p><b>Variety:</b> maximum strength</p>	<p><b>[Correct]</b> Box of <b>Mucinex Maximum Strength Expectorant</b>, 1200 mg guaifenesin extended-release bi-layer tablets, labeled for 12-hour relief, with 14 tablets indicated on the blue section, primarily white packaging with blue and red accents.</p>	<p><b>[Correct]</b> A white and blue box of <b>Maximum Strength Mucinex Expectorant</b> contains 1200 mg guaifenesin extended-release bi-layer tablets. [...] A white-outlined circle with a blue arrow pointing rightward and the words "12 HOUR" [...] It reads: "Relieves Chest Congestion" and "Thins And Loosens Mucus." Below this in white, it states: "10 extended-release bi-layer tablets."</p>	<p><b>[Incorrect]</b> The product is a box of <b>Mucinex Expectorant</b> tablets. The box is white with blue and red accents, featuring the brand name "Mucinex" in large blue letters on the front. [...]</p> <p><b>Missing:</b></p> <ul style="list-style-type: none"> <li>- variety: maximum strength</li> </ul>	<p><b>[Correct]</b> Boxed <b>expectorant</b> medication. Brand: <b>Mucinex</b>. Variety: <b>Maximum Strength</b> 12-Hour Extended Release Chest Congestion, packaging features red and blue color scheme on a white background.</p>

**Table 13: Examples of images with text panels (rows 1–2) and rounded labels (rows 3–4). VLMs can read text panels, but often incorrectly or miss key information. Rounded objects often obscure the label, requiring more inference about the product, which humans do well, but VLMs still struggle with. Captions were shortened for presentation purposes only, indicated by [...].**

Image	Annotation	GPT	Gemini	Llama	Molmo
	<b>Product:</b> almond milk <b>Brand:</b> O Organics <b>Variety:</b> vanilla	<b>[Correct]</b> Carton of <b>O Organics Almond Milk, Vanilla variety</b> . The package includes the <b>O Organics</b> brand logo and text indicating it is an organic <b>almond milk</b> product with <b>vanilla flavor</b> . The carton features a combination of white and other colors, with descriptive product information on the side.	<b>[Incorrect]</b> A rectangular white carton contains organic <b>vanilla almond milk</b> . The carton has a light brown banner with the words <b>“Almond Milk”</b> in white sans-serif font. Below the banner, the word <b>“Vanilla”</b> is printed in white text. At the top of the carton, an orange and blue logo is visible, with the word <b>“ORGANICS”</b> printed underneath.	<b>[Incorrect]</b> The image is out of focus, but it appears to depict a product label for <b>vanilla almond milk</b> . [...] It features a rectangular shape with a prominent orange banner at the top, displaying the brand name <b>“Almond Milk”</b> in large black letters. Below the banner, the flavor <b>“Vanilla”</b> is written in smaller black text. [...]	<b>[Incorrect]</b> <b>Almond milk vanilla</b> carton, possibly by Horizon Organic. White carton with black text and color accents. <b>Missing:</b> - <b>brand: O Organics</b>
	<b>Product:</b> ground beef <b>Variety:</b> 90% lean; 10% fat <b>Missing:</b> - <b>variety: 90% lean; 10% fat</b>	<b>[Incorrect]</b> Plastic-wrapped foam tray of <b>ground beef</b> , labeled as “Ground Beef 93% Lean 7% Fat” and “Beef Ground Round USA.” [...]	<b>[Incorrect]</b> <b>Ground beef</b> , 96% lean/4% fat, in a plastic and Styrofoam tray with a Nutrition Facts label and safe handling instructions.	<b>[Incorrect]</b> The product is a package of <b>ground beef</b> . The package is blue and rectangular with a clear plastic window showing the red ground beef inside. [...]	<b>[Incorrect]</b> <b>Ground beef</b> in blue plastic container with clear plastic top. Nutrition facts label on left side shows 93% lean/7% fat. [...]
	<b>Product:</b> sloppy joe sauce <b>Brand:</b> manwich <b>Missing:</b> - <b>product: sloppy joe sauce</b> - <b>brand: Manwich</b>	<b>[Incorrect]</b> Metal can of what appears to be a tomato-based product, featuring an image of a tomato and other vegetables on the label. [...] The specific brand and variety are not fully visible.	<b>[Incorrect]</b> A red metal can of Campbell's brand condensed tomato soup. The label features a large, realistic image of a red tomato.	<b>[Incorrect]</b> The product is a can of tomato sauce. The can is red with a blue and white label. The label has a picture of a tomato on it. The brand name is not visible. [...]	<b>[Incorrect]</b> Small metal can of Hunt's tomato sauce. White label with red color block. Recipes on label for dishes using product. Image of red tomatoes on label. <b>Missing:</b> - <b>product: sloppy joe sauce</b> - <b>brand: Manwich</b>
	<b>Product:</b> peach; slices <b>Brand:</b> Sweet Harvest <b>Variety:</b> heavy syrup	<b>[Correct]</b> Metal can of <b>Sweet Harvest</b> brand yellow cling <b>peach slices in heavy syrup</b> . The can has an orange and yellow label with images of peach slices and green accents.	<b>[Correct]</b> A can of <b>Sweet Harvest</b> Yellow Cling <b>Peach Slices in heavy syrup</b> . The can is primarily yellow and orange, with a picture of peach slices at the top. The product name is displayed in white text against an orange background.	<b>[Incorrect]</b> Product Description - Object Type: Can - Product Type: Prepared meal - Brand: Great Value - Variety: <b>Peach Slices</b> in Syrup - Visual Features: Yellow and orange can with a green leaf logo on the front. <b>Missing:</b> - <b>brand: Sweet Harvest</b> - <b>variety: heavy syrup</b>	<b>[Incorrect]</b> A yellow can of Del Monte Sweet Peas. <b>Missing:</b> - <b>product: peach; slices</b> - <b>brand: Sweet Harvest</b> - <b>variety: heavy syrup</b>