Evaluating Multimodal Language Models as Visual Assistants for Visually Impaired Users

Anonymous ACL submission

Abstract

This paper explores the effectiveness of Multimodal Large Language models (MLLMs) as assistive technologies for visually impaired individuals. We conduct a user survey to identify adoption patterns and key challenges users face with such technologies. Despite a high adoption rate of these models, our findings highlight concerns related to contextual understanding, cultural sensitivity, and complex scene understanding, particularly for individuals who may rely solely on them for visual interpretation. Informed by these results, we collate five user-centred tasks with image and video inputs, including a novel task on Optical Braille Recognition. Our systematic evaluation of twelve MLLMs reveals that further advancements are necessary to overcome limitations related to cultural context, multilingual support, Braille reading comprehension, assistive object recognition, and hallucinations. This work provides critical insights into the future direction of multimodal AI for accessibility, underscoring the need for more inclusive, robust, and trustworthy visual assistance technologies[1].

1 Introduction

As the capabilities of Large Language Models (LLMs) have been extended to multimodal contexts, particularly in applications that combine vision and language processing, one promising area is the use of multimodal LLMs (MLLMs) as visual assistants. MLLMs can provide valuable support, particularly for individuals with visual impairments, by accurately interpreting and describing visual content, including real-world images and videos (Karamolegkou et al., 2024).

MLLMs have already integrated into assistive technologies and services[2], such as automated captioning systems and smart devices (Yuan et al., 2025). However, these models still face limitations in acting as effective visual assistants (Tong et al., 2024). For example, a Blind or Low Vision (BLV) traveller using an MLLM-powered assistant to navigate a foreign city may receive inaccurate descriptions of street signs due to poor image quality or incomplete translations, leading to confusion or safety risks. Such scenarios raise concerns about reliability and safety in critical tasks and pose risks for users who depend on precise visual assistance. Given that modern MLLMs are hill-climbing multimodal reasoning benchmarks (Liu et al., 2024c; Li et al., 2024c; Wang et al., 2024b), a comprehensive evaluation of their effectiveness and limitations in accessibility applications is urgently needed.

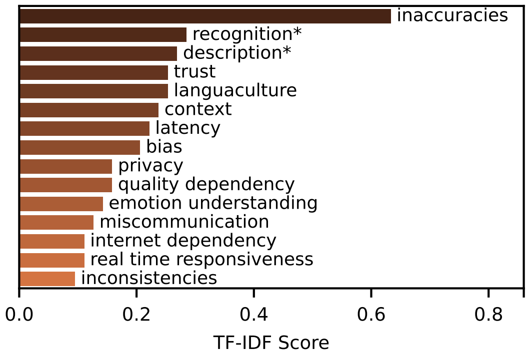


Figure 1: User survey results highlighting the 15 most important terms (measured by TF-IDF scores), representing key challenges for AI visual assistants. (*) includes tasks such as object, handwriting and face recognition; and image, scene, and video description.

To better understand these challenges, we conducted a user survey (§3) to identify the tasks and issues most relevant to individuals with visual impairments. Figure 1 summarizes the survey responses, highlighting that inaccuracies, such as hallucinations (i.e., factually incorrect or inconsistent generated content) and misleading information, are the primary concerns. Motivated by the

[1]We make our survey, evaluation data, and code publicly available at `linked removed for review`.

[2]`https://aira.io/`, `https://www.bemyeyes.com/`

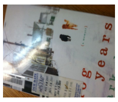
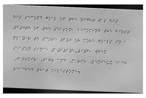
Task	Setting	Example		Model Response
Image Captioning	Original Cultural		General: A stork with Asian symbols on a silver coin. Cultural: The obverse side of a 500 won South Korean coin with a Red-crowned crane.	**Caption:** An elegant coin featuring a bird in flight, bearing inscriptions that hint at its cultural significance.
Image Question Answering	Original Multilingual		EN Q: What is the title of this book? EN A: dog years (n=9), dog years: memoir (n=1) EL Q: Ποιος είναι ο τίτλος αυτού του βιβλίου; EL A: Χρόνια σκύλου (n=9), Χρόνια σκύλου: απομνημονεύματα (n=1)	**EN Answer:** dog years **EL Answer:** Δεν μπορεί να απαντηθεί (Cannot be answered)
Optical Braille Recognition	Transcription Cross-script Question Answering		Translate the Braille to English: His second goal of the night was his 60th of the season, becoming the first [...] Q: Based on this text who had scored 60 goals in the season 1995-96? A: Jaromir Jagr and Mario Lemieux	**Transcription:** The provided content appears to already be in English. If you are asking for an interpretation or explanation of the [...] **Answer:** The provided content does not contain any information about goals scored in the 1995-96 season.
Video Object Recognition	General Objects Assistive Objects		Q: What type of assistive measuring device appears in the video? A: Liquid level indicator	**Answer:** Measuring cup
Video Question Answering	Descriptive Questions		Q: What colour are these sandals? A: Pink	**Descriptive Answer:** Blue
	Spatial Questions		Q: Where are my keys? A: in corner of doormat next to slipper	**Spatial Answer:** on mat
	Adversarial Questions		Q: What does this parcel say? A: Not enough information are depicted in video to answer this question	**Adversarial Answer:** Amazon

Figure 2: Illustration of the five key areas of our framework. We focus on tasks pertinent to BLV people covering different aspects for captioning, transcribing, and answering questions about visual content.

findings of our survey, we design an evaluation framework with tasks relevant to BLV individuals, focusing on five key areas as shown in Figure 2: 1) Image Captioning targeting cultural content, 2) multilingual Image Question Answering , 3) Optical Braille Recognition to transcribe and answer questions about Braille text rendered in images, 4) Video Object Recognition covering general usage objects as well as assistive items commonly used by BLV people, and 5) Video Question Answering covering descriptive, spatial, and adversarial questions. Importantly, we contribute datasets for multilingual and video question answering as well as Braille recognition to improve the capabilities of the next generation of MLLMs that assist BLV individuals. Our experiments emphasize the need for further advances in multimodal AI to ensure these models can reliably support individuals who rely on them for visual tasks.

2 Related Work

MLLM Evaluation Benchmarks MLLMs are mainly evaluated on general-domain benchmarks that assess visual perception, knowledge, and reasoning (Goyal et al., 2017b; Schwenk et al., 2022; Yin et al., 2023; Li et al., 2024b; Liu et al., 2024c; Lu et al., 2024). However, these benchmarks do not capture all critical dimensions of MLLM performance. One exception is the holistic evaluation by Lee et al. (2024), which examined 22 MLLMs across nine aspects, revealing that no model excels in all areas and that all lack multilingual support. Other studies highlight inconsistencies in MLLMs' responses (Chen et al., 2024) and measure performance in diverse cultural contexts (Nayak et al., 2024; Mogrovejo et al., 2024). Despite these efforts, the effectiveness of MLLMs as visual assistants in accessibility settings remains unexplored.

Multimodal Models for Assistive Applications Previous works apply *task-specific* models for assistive applications including visual question answering (Liu et al., 2024b; Huh et al., 2024; Gurari et al., 2018), image captioning (Gurari et al., 2020), object detection (Reynolds et al., 2024; Tseng et al., 2022), and private content identification (Tseng et al., 2024). Some conversational agents focus on privacy-aware assistance (Baker et al., 2021), education for BLV users (Di Nuovo et al., 2024), scenarios with low-quality images (Yang et al., 2024), while other studies integrate MLLMs into assistive devices and smartphone applications (Holiel et al., 2024; De Marsico et al., 2024). These works demonstrate the potential of MLLMs in accessibility but also highlight the need for systematic evaluation of their effectiveness and limitations.

3 User Survey

Understanding user perspectives is crucial for identifying key application areas, surfacing unmet needs, and guiding future improvements in model design and evaluation (Liao et al., 2024; Kirk et al., 2024). To gain insights into the real-world usability of MLLMs in the role of visual assistants, we conducted a survey focused on user adoption patterns and experienced challenges.

Design The survey consists of two phases: open-ended questions and Likert scale ratings. Phase one begins with a user adoption question, asking whether participants use or would consider using AI for visual assistance. The second question explores tasks where these models could be most beneficial, while the third targets challenges users have experienced in past interactions. Phase two focuses on assessing specific tasks and issues. We recruited 106 participants through Prolific[3], with varying degrees of visual impairment. By analysing responses and identifying key themes, we identified common use cases and areas where AI needs improvement. Below, we present a summary of the open-ended responses and provide further details on survey design, analysis, compensation, demographics, and Likert scale results in Appendix A.

User Adoption and Tasks The majority of respondents (87%) use or would use AI as visual assistants, while 9% declined due to concerns about accuracy, reliability, and lack of personal touch, and 4% were unsure, depending on the assistance type. Participants found AI most useful for description, transcription, translation, and recognition. Common use cases included identifying and translating products for shopping or cooking, understanding diagrams in subjects like chemistry and math, analysing food consistency, choosing outfits, and interpreting facial expressions. Some mentioned more specialized uses, such as autonomous navigation, medical imaging analysis, Braille interpretation, space planning, design assistance, artistic creation, and emotional support.

Challenges Participants were asked to list challenges they have experienced when using AI models. Responses varied in specificity, requiring grouping and qualitative analysis using an iterative open thematic approach[4] (Liao et al., 2024). The most common challenges are visualized in Figure 1. A major problem was *inaccuracies*, as it was often mentioned that users struggle to verify whether the provided information is correct. This included issues such as *incorrect directions*, *misidentification of objects*, *misinterpretation of signs and symbols*, and *misleading or incomplete responses*. Many challenges fell under *recognition* and *description* tasks, particularly difficulties with handwritten text (especially small, messy, or multilingual), Braille, currency, and signs. Participants also reported problems with object recognition in poor image conditions (low resolution, lighting issues, or background noise) and in cluttered or ambiguous settings, sometimes mistaking shadows for obstacles or misidentifying overlapping objects.

Some also mentioned that AI often fails to recognize hazards, interpret multicultural and social cues, and provide sufficiently detailed descriptions. Further challenges involved trust, language limitations, contextual understanding in scenes (e.g., understanding spatial relationships or complex environments), latency, bias, privacy concerns, dependency on high-quality data, emotional understanding, and communication barriers, such as unclear prompts or lack of adaptation to individual needs.

4 Evaluation Framework

We evaluate MLLMs across image and video understanding tasks, specifically designed to assist visually impaired individuals. Our task selection is informed by user input, reflecting use cases where users reported a high likelihood of adopting AI visual assistants (see Figure 8). We additionally emphasize high-priority needs such as cultural context awareness, multilingual support, and recognition of assistive devices and hallucinations.

Tasks The evaluation spans five tasks: `Image Captioning`, which evaluates performance in generating descriptions for images taken by visually impaired individuals (§5). `Image Question Answering` to evaluate question answering using images and questions provided by visually

[3] https://www.prolific.com/

[4] Further details are in Appendix A.

impaired individuals (§6). `Optical Braille Recognition`, where we assess the performance on transcribing and answering questions about Braille text in images (§7). Finally, using videos recorded by BLV individuals, we evaluate `Video Object Recognition` (§8), and `Video Question Answering` on descriptive, spatial, and adversarial questions (§9). In each corresponding section, we introduce the related background, specify the evaluation setup, and report our result.

Models We focus on 12 prominent models based on the following criteria: 1) strong performance in image and video understanding, 2) open access (open-source or open-weights), and 3) moderate computational overhead to balance performance and latency. Table 13 provides details about multilingual support and exposure to domain data.

5 Image Captioning

Image captioning aims at generating textual descriptions for images. Gurari et al. (2020) introduce VizWiz, the first dataset with images from visually impaired users, launching a series of multimodal challenges (Gurari et al., 2018). Since then, research has focused on improving models for assisting visually impaired users (Dognin et al., 2022; Ahsan et al., 2021; Delloul and Larabi, 2023), mostly in English settings. More recently, Karamolegkou et al. (2024) identified cultural implicatures in VizWiz that annotations and models overlook and curated a subset of 324 images and 648 captions spanning 60 cultures.

Setup We evaluate model performance on generating descriptions for images taken by visually impaired people. We use the original validation set of VizWiz-Captions (N=500) (Gurari et al., 2020) and the **multicultural** extension (Karamolegkou et al., 2024) (N=324), which filters the original set and provides re-annotations focused on culture-related content. As a metric, we use the RefCLIPScore (Hessel et al., 2021), which has shown robust alignment with human judgment.

Results Table 1 shows the image captioning evaluation results. All models achieve RefCLIPScores between 70 and 81 on the original setting indicating relatively good performance. Paligemma outperforms other MLLMs by at least 5 points, likely due to its exposure to VizWiz data during pretraining. In the cultural setting, we observe a clear performance divide. Five out of the examined MLLMs

Model	Original	Cultural
Idefics3	76.0	75.5
InternVL2.5-MPO	74.3	74.8
LLaVA-v1.6	72.3	52.2
Llama-3.2-Vision-Instruct	75.0	72.8
MiniCPM-V-2.6	78.0	74.8
Molmo	70.9	47.4
Paligemma	**81.0**	55.0
Phi-3.5-Vision-Instruct	71.9	62.6
Qwen2-VL-Instruct	75.9	**76.9**

Table 1: RefCLIPScore results on the original and cultural VizWiz image captioning validation set.

Model	Original	Multilingual
Idefics3	45.7	30.4
InternVL2.5	65.1	39.1
LLaVA-v1.6	54.8	40.8
Llama-3.2-Vision-Instruct	52.9	29.6
MiniCPM-V-2.6	72.2	30.7
Molmo	40.2	28.6
Paligemma	**75.6**	16.9
Phi-3.5-Vision-Instruct	59.0	36.5
Qwen2-VL-Instruct	61.9	**44.9**

Table 2: VQA Accuracy results on the original and multilingual VizWiz question answering validation set.

show robust performance ($\pm$4 points difference), while other models show substantial degradation (20-25 points). To assess progress in culture-aware descriptions, we inspect 100 captions from the top two models. For both Qwen2-VL-Instruct and Idefics3, approximately a third of the generated captions (31% and 33%, respectively) include accurate but generic information–while they correctly describe the scene, they miss culturally significant details such as specific names of symbols, cultural figures, or non-English language scripts. This indicates that models might still overlook cultural context, which is essential to fully describe a scene.

6 Image Question Answering

Image question answering (IQA) enables users to ask about images and receive relevant answers. As part of the VizWiz initiative, Gurari et al. (2018) created an IQA dataset capturing real-world challenges, where visually impaired users take photos that may be blurry, poorly framed, or contain unanswerable questions. Recent efforts address these issues through answer grounding (Chen et al., 2023), long-form answers (Huh et al., 2024), and models suggesting image adjustments (Liu et al., 2024b). However, no work has yet examined these challenges in a multilingual setting.

Setup We evaluate each model on visual question answering using the VizWiz validation set (Gurari et al., 2018). To assess the global accessibility of these models, we extend the evaluation to a **multilingual** setting. We use an automatic translation pipeline with human quality checks to translate 500 questions and reference answers to 34 languages. Details about the translation process are provided in Appendix B.1. The task metric is VQA accuracy (Antol et al., 2015), which takes into account multiple reference answers as the evaluation metric.

Results The results shown in Table 2 reveal large performance disparities across models and evaluation settings. In the original English setting, Paligemma and MiniCPM-V-2.6 (75.6% and 72.2% respectively), which include VizWiz VQA data in their pretraining mixture, achieve the highest performance by a large margin. However, these models also suffer the largest performance drops in the multilingual setting. We observe that they often fail to follow prompt instructions, such as answering in the language of the question, instead defaulting to English responses. The best multilingual performance is achieved by Qwen2-VL-Instruct, which shows the most consistent performance ranging between 35.4 and 49.0 for all non-English languages. Table 9 shows the VQA accuracy grouped by high-, medium-, and low-resourced languages (Joshi et al., 2020). We observe limited performance variance across the three groups, with all models performing similarly poorly regardless of language resource levels. This suggests that even high- and medium-resource languages lack reliable IQA support for blind users who do not speak English.

7 Optical Braille Recognition

Despite increasing interest in the text comprehension abilities of MLLMs (Li et al., 2024a; Liu et al., 2024d), their capacity to process Braille within images remains underexplored. Existing Braille recognition approaches focus on character-level classification where a visual component first detects the characters, followed by a character classifier (Li et al., 2020; Smelyakov et al., 2018; Gao et al., 2024). However, character-level approaches do not fully assess the reading comprehension capabilities of modern MLLMs. For this purpose, we compile two datasets focusing on sentence-level Braille-to-Text transcription and paragraph-level cross-script question-answering. Our datasets differ from prior work as they target longer context, support zero-shot and few-shot evaluation, and introduce a training split that can be incorporated in the visual instruction tuning data of an MLLM.

Model	F-200	N-128	Avg
Idefics3	1.9	2.1	2.0
InternVL2.5-MPO	8.7	8.5	8.6
LLaVA-v1.6	2.9	2.7	2.8
Llama-3.2-Vision-Instruct	8.9	8.3	8.5
MiniCPM-V-2.6	8.9	9.2	9.1
Molmo	5.3	5.44	5.4
Phi-3-vision-128k-instruct	10.2	9.5	9.9
Qwen2-VL-Instruct	**75.2**	**72.5**	**73.8**

Table 3: Zero-shot results on sentence-level Braille-to-Text transcription. Out of the eight models, only Qwen2-VL-Instruct exhibits Braille comprehension capabilities.

7.1 Dataset Creation

For sentence-level transcription, we compile a dataset using English sentences from the shared task of WMT 2024 (Haddow et al., 2024). More specifically, we use a subset of 100k sentences from the `Facebook-wikimatrix-1-deu-eng` corpus for training, as well as NTREX-128 (Federmann et al., 2022) (N=1997), and FLORES-200 (NLLB Team et al., 2024) (N=1012) for evaluation. With regards to paragraph-level question answering, we leverage SQuAD (Rajpurkar et al., 2018) (training N=130K, evaluation N=11.9K), which provides text paragraphs together with a few relevant questions. In both tasks, we render the Braille text into images (see Appendix C), and apply augmentations that correlate with quality flaws often occurring in images taken by BLV people (Yu et al., 2023). The model accepts an image containing Braille text, the input prompt including a question for SQuAD only, and needs to provide the appropriate English response, i.e. either the transcription of the rendered Braille sentence or the answer to the question.

Evaluation Metrics Since the Braille-to-Text transcription is a character-level transformation, for the sentence-level transcription, we report the chrF++ score (Popović, 2017). For SQuAD, we report the character-level F1-score based on the model's prediction and the candidate answers for each question, as well as the exact match.

7.2 Results

Can MLLMs read Braille? We prompt MLLMs to transcribe rendered Braille sentences to regular English text. Table 3 illustrates the zero-shot performance on our two English-to-Text transcrip-

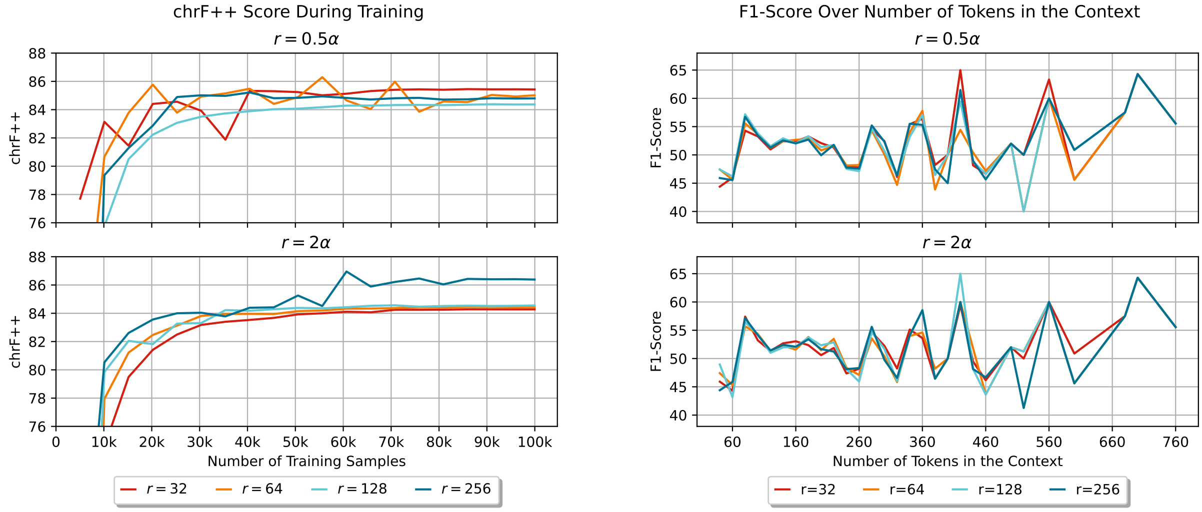


Figure 3: Left: Average chrF++ on sentence-level Braille-to-Text transcription. Right: F1-Score on cross-script question answering where results are binned based on the length of the context paragraph.

r	α	F-200	N-128	Avg	SQuAD	
			chrF++		F1	EM
32	64	88.2	82.6	85.4	51.8	49.8
64	128	88.2	81.9	85.0	51.8	49.8
128	256	87.4	81.4	84.4	52.0	50.0
256	512	87.6	81.9	84.8	51.7	49.7
32	16	87.2	81.4	84.3	51.9	49.9
64	32	87.4	81.4	84.4	51.9	**50.2**
128	64	87.5	81.7	84.6	51.9	50.1
256	128	89.2	83.5	**86.4**	**52.1**	50.1

Table 4: LoRA fine-tuning results for Llama-3.2-Vision-Instruct on sentence-level Braille-to-Text transcription, and cross-script question answering.

tion evaluation sets. Our results clearly show that most modern MLLMs are not equipped with Braille recognition capabilities. Surprisingly, out of the examined models, only Qwen2-VL-Instruct demonstrates non-trivial performance indicating its capability of reading Braille from images.

Proof of concept: Learning to Read Braille Next, we are interested in a training recipe that results in an MLLM capable of reading Braille text in images. For this purpose, we focus exclusively on Llama-3.2-Vision-Instruct, as a model with strong text comprehension capabilities but lacking the ability to read Braille text. We finetune Llama-3.2-Vision-Instruct both for sentence-level transcription, as well as paragraph-level question answering using LoRA (Hu et al., 2022) following guidelines from existing cookbook recipes[5]. For each configuration, we sweep across different hyperparameters (see Appendix C) and select the one with the best validation performance.

Table 4 illustrates the finetuning results of Llama-3.2-Vision-Instruct on both tasks. We observe that the model achieves great performance across a wide range of configurations. Additionally, Figure 3 (left) shows that model performance improves quickly–typically saturating at 30K samples. Similar results can be observed in the case of paragraph-level question answering. Finally, Figure 3 (right) shows the F1-Score of all finetuning runs according to the length of the context paragraph, i.e., the number of English tokens that have been transcribed to Braille and rendered in images. We observe that the model maintains similar performance in short as well as long paragraphs. Taken together, these results show that while most modern MLLMs are not equipped with Braille comprehension, learning to read Braille text in images is feasible with a moderate number of demonstrations. Consequently, we expect the next generation of MLLMs powering accessibility applications to take into account Braille reading comprehension capabilities as part of the visual instruction tuning stage.

8 Video Object Recognition 🎥

Video-based object recognition extends traditional image recognition (Russakovsky et al., 2015; Hu et al., 2023; Sun et al., 2021), allowing models to identify objects that appear in a video sequence. While image recognition provides a snapshot of visual content, it may miss useful contextual cues available in videos, such as gradual occlusions

[5] Practical Tips for Finetuning LLMs Using LoRA

Model	General (N=880)	Assistive (N=156)
Video LMs		
LLaVA-NeXT-Video	56.0	26.0
LLaVA-Video	65.7	41.3
VideoChat-Flash	56.0	20.8
Image + Video LMs		
InternVL2.5-MPO	59.1	36.5
MiniCPM-V-2.6	65.1	**44.2**
Phi-3.5-Vision-Instruct	52.2	25.3
Qwen2-VL-Instruct	**69.8**	39.7

Table 5: Accuracy in Video Object Recognition of general and assistive object categories.

Model	D	S	A	Avg
Video LMs				
LLaVA-NeXT-Video	56.0	49.7	13.4	39.7
LLaVA-Video	**78.2**	63.4	7.7	49.8
VideoChat-Flash	72.4	**64.1**	9.2	48.6
Image + Video LMs				
InternVL2.5-MPO	67.7	59.4	9.0	45.4
MiniCPM-V-2.6	68.7	63.3	**17.7**	**49.9**
Phi-3.5-Vision-Instruct	61.4	46.3	10.2	39.3
Qwen2-VL-Instruct	71.9	58.5	12.6	47.7

Table 6: Accuracy in Video Question Answering. D: Descriptive, S: Spatial, A: Adversarial Questions.

or varying viewpoints. Unlike video classification, which typically focuses on activity recognition (Goyal et al., 2017a; Kay et al., 2017), our task aims to identify the presence of objects in a video, making it more aligned with real-world assistive applications. Moreover, while there are a number of datasets filmed in real-world environments (Lomonaco and Maltoni, 2017; Damen et al., 2018), very few are specifically centred on visual assistance for visually impaired users (Massiceti et al., 2021; Islam et al., 2024).

Setup We evaluate models on their ability to identify objects from videos recorded by BLV people. Similar to the image settings, these videos pose challenges such as blurriness and non-centred objects. We use 1036 video clips from ORBIT (Massiceti et al., 2021), which show household objects from 92 categories. These objects include both general everyday objects (e.g., TV remote control) and assistive items (e.g., Braille display). Additionally, objects are recorded in *Clean* videos, which show an object in isolation, and *Clutter* videos, which show the target object in context with other items. More details about the dataset are provided in Appendix D.1. Following previous work on evaluating generated outputs with one ground truth answer, we adopt the LAVE protocol (Mañas et al., 2024), which leverages a language model to judge the generated outputs and provide a rating between 1-3. We use Llama-3.3-70B-Instruct (AI@Meta, 2024) and report average normalized ratings.

Results Table 5 reports model accuracy on recognizing general and assistive objects, revealing a clear gap: while models perform moderately well on generic object categories (52-69.8% accuracy), they struggle significantly at recognizing assistive items, achieving only 23-41% accuracy. This performance disparity might be expected from a data distribution perspective, as assistive objects are less common and current MLLMs are known to struggle with capturing long-tail knowledge (Geigle et al., 2024; Parashar et al., 2024). However, this result indicates that generalist MLLMs are still far from providing comprehensive support for visually impaired users in everyday contexts.

9 Video Question Answering 🎥

There have been a lot of works assessing descriptive and spatial understanding of models through video question-answering (Yu et al., 2019; Xiao et al., 2021; Xu et al., 2017; Li et al., 2024c), as well as more fine-grained skills like perception and reasoning (Patraucean et al., 2023), or ego-centric setups (Mangalam et al., 2023). Most datasets are compiled from existing corpora (Fabian Caba Heilbron and Niebles, 2015; Grauman et al., 2022) and crawled from open platforms (Thomee et al., 2016; Shang et al., 2019), and do not focus on videos filmed by visually impaired people. To address this gap, we curated a new video QA dataset based on videos filmed by BLV users.

Setup We evaluate models on their ability to answer questions given videos recorded by visually impaired people using the object recognition ORBIT (Massiceti et al., 2021) dataset. We annotate 98 videos and provide 882 question-answer pairs that target three types of questions: 1) descriptive questions regarding the attributes of the objects (colour, shape, number), 2) spatial Understanding about the position of items and their relation to other items, and 3) adversarial questions about items not present in the video (Li et al., 2023). Adversarial questions, which cannot be answered based on the information provided in the video, help assess whether models hallucinate responses or can reliably acknowledge uncertainty—- a critical safety feature for assistive technologies. More details about the dataset are provided in Appendix D.2. For evaluation, we follow the LAVE protocol as described above.

Results Table 6 shows the evaluation results for MLLMs that support video inputs. MiniCPM-V-2.6 and LLaVA-Video achieve the highest overall performance, although no model ranks first across all question types. Notably, we do not observe a performance advantage for models specifically fine-tuned on video data compared to models trained on both images and videos. Regarding the results per question type, we observe the following patterns. While most MLLMs show promising results on descriptive questions, spatial understanding remains challenging even for the best-performing models (VideoChat-Flash and LLaVA-Video at 63-64%). Most concerning is the behaviour on adversarial questions, where models consistently provide concrete answers rather than acknowledge uncertainty. For assistive technologies, this tendency to hallucinate responses instead of expressing an inability to answer could lead to misleading or potentially unsafe guidance (Li et al., 2023). In Appendix D.4, we show that even explicit prompting to express uncertainty as needed yields limited success: while some models improve on adversarial questions, they either achieve only modest gains or overgeneralize uncertain responses to valid questions.

10 Discussion and Conclusion

What is missing from existing evaluation frameworks? To better understand the use cases and challenges faced by visually impaired individuals, we designed a survey to collect first-hand insights. These findings provide valuable input for designing more effective, user-centred multimodal AI systems, and can add evaluation aspects to both targeted and holistic evaluation approaches (Liang et al., 2023; Lee et al., 2024). Our analysis captured a wide range of challenges that are underexplored or missing from holistic evaluation frameworks, such as 1) technical constraints (latency, real-time settings, internet dependency), 2) multilingual, cultural and contextual understanding, 3) trust and reliability issues amplified by hallucinations, misinterpretations, underspecified responses and failure in safety-critical or ambiguous scenarios.

Can existing models be used as visual assistants? We evaluated a range of multimodal models on datasets from visually impaired users, revealing notable limitations. For example, captioning becomes more challenging with culture-specific images, as models struggle to capture cultural nuances and distinctive items. Similarly, in image question answering, models show substantial performance degradation, which aligns with our survey findings. Optical Braille recognition seems to be a new challenge for almost all models, with most failing to perform the task, pointing to gaps in both training data and generalization abilities. In tasks like video object recognition, MLLMs struggle to identify assistive objects, revealing a lack of specificity in recognizing items important to BLV users. For video question answering, models have difficulty answering adversarial questions that refer to items not present in the image, which points to the models' vulnerability in real-world applications where visibility and conditions are not always ideal.

Beyond classic benchmark evaluation. Our findings suggest a pressing need for the development of datasets and models tailored to user needs and preferences. Such datasets should reflect the real-world complexities and unique challenges faced by BLV users across culturally diverse environments, multilingual settings, assistive devices, poor-quality input, and latency constraints. Furthermore, engaging with BLV users in the design and improvement of visual assistants is essential to ensure models address their needs (Caselli et al., 2021; Sloane et al., 2022). Continuously gathering feedback on usability, accuracy, trust, and preferences can help develop more accessible, contextually aware, and user-centred AI (Kirk et al., 2023).

Traditional evaluation metrics and benchmarks are insufficient in capturing the specific difficulties faced by users in practice (Liao and Xiao, 2023; Wang et al., 2024a). Existing benchmarks focus on general performance and may overlook critical aspects like real-world usability, and user satisfaction. To bridge this gap, future research should explore new, reproducible, user-centred methodologies of evaluation that go beyond conventional metrics to better assess models in everyday scenarios (Elangovan et al., 2024). By focusing on the unique challenges of visually impaired individuals and integrating their experiences into the development and evaluation of AI models, we can move towards more effective and inclusive visual assistants.

Limitations

While this work offers valuable insights into the potential of MLLMs as visual assistants for the visually impaired, several limitations should be acknowledged. First, our evaluation does not cover tasks related to navigation assistance, which is a crucial aspect of real-world applications for visually impaired individuals. Second, our experimental design focuses primarily on the performance of MLLMs in controlled environments and user-centred tasks and may not fully capture the complexities of dynamic, real-world scenarios. Lastly, our findings underscore the need for further research to address issues related to real-time responsiveness, reasoning tasks, and the inclusion of marginalized languages and cultural contexts.

Ethics Statement

This work adheres to ethical guidelines. The user survey was carried out with compensation and informed consent, ensuring that participants were fully aware of the purpose of the study and how their data would be used. We took careful measures to protect the privacy and confidentiality of all participants, with no personally identifiable information being disclosed or shared. Additionally, we acknowledge the potential bias introduced in this evaluation due to the use of datasets and models that may themselves contain inherent biases. All datasets used in this work are under CC BY 4.0 license [6]. This work emphasizes the need for AI systems that prioritize user trust and safety while acknowledging the potential limitations associated with AI deployment in sensitive contexts.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Hiba Ahsan, Daivat Bhatt, Kaivan Shah, and Nikita Bhalla. 2021. Multi-modal image captioning for the visually impaired. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 53–60, Online. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

[6]https://creativecommons.org/licenses/by/4.0/

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Katie Baker, Amit Parekh, Adrien Fabre, Angus Addlesee, Ruben Kruiper, and Oliver Lemon. 2021. The spoon is in the sink: Assisting visually impaired people in the kitchen. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 32–39, Gothenburg, Sweden. Association for Computational Linguistics.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. Association for Computational Linguistics.

Chongyan Chen, Samreen Anjum, and Danna Gurari. 2023. Vqa therapy: Exploring answer differences by visually grounding answers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15269–15279.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024. Measuring and improving chain-of-thought reasoning in vision-language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 192–210, Mexico City, Mexico. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Maria De Marsico, Chiara Giacanelli, Clizia Giorgia Manganaro, Alessio Palma, and Davide Santoro. 2024. Vqask: a multimodal android gpt-based application to help blind users visualize pictures. In *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*, pages 1–5.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Khadidja Delloul and Slimane Larabi. 2023. Towards real time egocentric segment captioning for the blind and visually impaired in rgb-d theatre images. *arXiv preprint*.

Elisa Di Nuovo, Manuela Sanguinetti, Pier Felice Balestrucci, Luca Anselma, Cristian Bernareggi, and Alessandro Mazzei. 2024. Educational dialogue systems for visually impaired students: Introducing a task-oriented user-agent corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5507–5519, Torino, Italia. ELRA and ICCL.

Pierre Dognin, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A. Young, and Brian Belgodere. 2022. Image captioning as an assistive technology: Lessons learned from vizwiz 2020 challenge. *J. Artif. Int. Res.*, 73.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Babu Bodapati, and Dan Roth. 2024. ConSiDERS-the-human evaluation framework: Rethinking human evaluation for generative large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1137–1160, Bangkok, Thailand. Association for Computational Linguistics.

Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Zhiqiang Gao, Lulu Chang, Bing Ren, Jing Han, and Jie Li. 2024. Enhanced braille recognition based on piezoresistive and piezoelectric dual-mode tactile sensors. *Sensors and Actuators A: Physical*, 366:115000.

Gregor Geigle, Radu Timofte, and Goran Glavaš. 2024. African or European swallow? benchmarking large vision-language models for fine-grained object classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2653–2669, Miami, Florida, USA. Association for Computational Linguistics.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017a. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017b. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

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

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

Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors. 2024. *Proceedings of the Ninth Conference on Machine Translation*. Association for Computational Linguistics, Miami, Florida, USA.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

Heidi Ahmed Holiel, Sahar Ali Fawzi, and Walid Al-Atabany. 2024. Assisting visually impaired subjects using large language models: A comprehensive evaluation. In *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pages 561–566. IEEE.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075.

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*.

Md Touhidul Islam, Imran Kabir, Elena Ariel Pearce, Md Alimoor Reza, and Syed Masum Billah. 2024. Identifying crucial objects in blind and low-vision individuals' navigation. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '24, New York, NY, USA. Association for Computing Machinery.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. 2020. imgaug. https://github.com/aleju/imgaug. Online; accessed 01-Feb-2020.

Antonia Karamolegkou, Phillip Rust, Ruixiang Cui, Yong Cao, Anders Søgaard, and Daniel Hershcovich. 2024. Vision-language models under cultural and inclusive considerations. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 53–66, TBD. ACL.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Hannah Rose Kirk, Andrew M. Bean, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2023. The past, present and better future of feedback learning in large language models for subjective human preferences and values. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2409–2430, Singapore. Association for Computational Linguistics.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Michael Bean, Katerina Margatina, Rafael Mosquera, Juan Manuel Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*.

Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Somerville Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, and Percy Liang. 2024. VHELM: A holistic evaluation of vision language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024a. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024b. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Renqiang Li, Hong Liu, Xiangdong Wang, Jianxing Xu, and Yueliang Qian. 2020. Optical braille recognition based on semantic segmentation network with auxiliary learning strategy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 554–555.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. 2024d. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*.

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. 2024. Llms as research tools: A large scale survey of researchers' usage and perceptions. *Preprint*, arXiv:2411.05025.

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

Li Liu, Diji Yang, Sijia Zhong, Kalyana Suma Sree Tholeti, Lei Ding, Yi Zhang, and Leilani H. Gilpin. 2024b. Right this way: Can VLMs guide us to see more to answer questions? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024d. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12).

Vincenzo Lomonaco and Davide Maltoni. 2017. Core50: a new dataset and benchmark for continuous object recognition. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 17–26. PMLR.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5, pages 4171–4179.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, volume 36, pages 46212–46244. Curran Associates, Inc.

Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. 2021. ORBIT: A Real-World Few-Shot Dataset for Teachable Object Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

David Orlando Romero Mogrovejo, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, Emilio Villa Cueva, Jinheon Baek, Soyeong Jeong, Injy Hamed, Zheng Xin Yong, Zheng Wei Lim, Paula Mónica Silva, Jocelyn Dunstan, Mélanie Jouitteau, David LE MEUR, Joan Nwatu, Ganzorig Batnasan, Munkh-Erdene Otgonbold, Munkhjargal Gochoo, Guido Ivetta, Luciana Benotti, Laura Alonso Alemany, Hernán Maina, Jiahui Geng, Tiago Timponi Torrent, Frederico Belcavello, Marcelo Viridiano, Jan Christian Blaise Cruz, Dan John Velasco, Oana Ignat, Zara Burzo, Chenxi Whitehouse, Artem Abzaliev, Teresa Clifford, Gráinne Caulfield, Teresa Lynn, Christian Salamea-Palacios, Vladimir Araujo, Yova Kementchedjhieva, Mihail Minkov Mihaylov, Israel Abebe Azime, Henok Biadglign Ademtew, Bontu Fufa Balcha, Naome A Etori, David Ifeoluwa Adelani, Rada Mihalcea, Atnafu Lambebo Tonja, Maria Camila Buitrago Cabrera, Gisela Vallejo, Holy Lovenia, Ruochen Zhang, Marcos Estecha-Garitagoitia, Mario Rodríguez-Cantelar, Toqeer Ehsan, Rendi Chevi, Muhammad Farid Adilazuarda, Ryandito Diandaru, Samuel Cahyawijaya, Fajri Koto, Tatsuki Kuribayashi, Haiyue Song, Aditya Nanda Kishore Khandavally, Thanmay Jayakumar, Raj Dabre, Mohamed Fazli Mohamed Imam, Kumaranage Ravindu Yasas Nagasinghe, Alina Dragonetti, Luis Fernando D'Haro, Olivier NIYOMUGISHA, Jay Gala, Pranjal A Chitale, Fauzan Farooqui, Thamar Solorio, and Alham Fikri Aji. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. 2024. The neglected tails in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12988–12997.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens Continente, Larisa Markeeva, Dylan Sunil Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. 2023. Perception test: A diagnostic benchmark for multimodal video models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Jarek Reynolds, Chandra Kanth Nagesh, and Danna Gurari. 2024. Salient object detection for images taken by people with vision impairments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8522–8531.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287. ACM.

Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA. Association for Computing Machinery.

Kirill Smelyakov, Anastasiya Chupryna, Dmytro Yeremenko, Anton Sakhon, and Vitalii Polezhai. 2018. Braille character recognition based on neural networks. In *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, pages 509–513. IEEE.

Zeren Sun, Yazhou Yao, Xiu-Shen Wei, Yongshun Zhang, Fumin Shen, Jianxin Wu, Jian Zhang, and Heng Tao Shen. 2021. Webly supervised fine-grained recognition: Benchmark datasets and an approach. In *IEEE International Conference on Computer Vision (ICCV)*.

Ashish Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *EMNLP*.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578.

Yu-Yun Tseng, Alexander Bell, and Danna Gurari. 2022. Vizwiz-fewshot: Locating objects in images taken by people with visual impairments. In *European Conference on Computer Vision*, pages 575–591. Springer.

Yu-Yun Tseng, Tanusree Sharma, Lotus Zhang, Abigale Stangl, Leah Findlater, Yang Wang, and Danna Gurari. 2024. Biv-priv-seg: Locating private content in images taken by people with visual impairments. *arXiv preprint arXiv:2407.18243*.

Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024a. A user-centric multi-intent benchmark for evaluating large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3612, Miami, Florida, USA. Association for Computational Linguistics.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024b. Measuring multimodal mathematical reasoning with MATH-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024c. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. 2024d. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Bufang Yang, Lixing He, Kaiwei Liu, and Zhenyu Yan. 2024. Viassist: Adapting multi-modal large language models for users with visual impairments. *arXiv preprint arXiv:2404.02508*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

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*, pages 6281–6289.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 01, pages 9127–9134.

Zhiqiang Yuan, Ting Zhang, Jiapei Zhang, Jie Zhou, and Jinchao Zhang. 2025. Walkvlm:aid visually impaired people walking by vision language model. *Preprint*, arXiv:2412.20903.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024. Pangea: A fully open multilingual multimodal llm for 39 languages. *arXiv preprint arXiv:2410.16153*.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: A strong zero-shot video understanding model.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024b. Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.

A Survey Design and Results

Our survey was designed to explore how individuals who are blind or have low vision use AI models as visual assistants. The focus was on understanding the tasks they perform and the challenges they face. The survey combined multiple-choice and open-ended questions, allowing for both quantitative and qualitative insights. The responses are useful to help identify patterns and areas for improvement in AI models to better serve individuals with vision impairments.

A.1 Survey Construction

The survey was carefully designed with input from individuals who are blind or have low vision to ensure it accurately reflected their experiences and needs. We implemented two feedback loops by engaging with blind participants during the design process, allowing us to refine questions and make sure the survey was accessible and relevant.

Demographics Participants were recruited via Prolific, and compensation was based on an average reward per hour (9 pounds) to ensure fair payment for their time. We asked for participants to be located across all countries available, and a fair distribution sample. We also added a screener that participants have no vision (found under Add Screeners<Health<No Vision). This resulted in 25,485 matching participants 'who have been active in the past 90 days'. We collected a total of 106 participants, after filtering out some participants without visual impairments. Even though our survey was completely anonymous, Prolific provides some basic demographics for participants in a .csv format. We plotted some of the participant demographics after excluding the vision "yes" and "revoked_consent" participations in Figure 4.

A.2 Survey Sections and Results

Introduction. Before beginning the survey, participants were briefed on its content and purpose: *This survey is for individuals who are blind or have low vision and use AI models like ChatGPT or Gemini as visual assistants. Our goal is to understand the tasks they perform, the challenges they encounter, and their overall experiences with AI.*

We then obtained their consent, assuring them that their responses would remain completely anonymous—no email addresses or identifying information were requested. Participants were informed that the survey would contribute to a research project leading to a scientific publication and were encouraged to answer honestly and thoughtfully. Additionally, we provided contact details for both the student and their supervisor.

At the beginning of the survey, we had some initial questions asking participants about their prolific ID, and we added an extra question to verify Prolific's screener is accurate and that we are going to get responses from visually impaired people. As shown in Figure 10, there were actually some participants who did not have a visual impairment, so we had to filter their responses.

Phase 1: Open Ended Questions. The second section, as shown in Figure 11, was about *user adoption and tasks*, asking participants whether they currently use or would consider using Artificial Intelligence models as visual assistants. After filtering the responses from individuals without impairments, we visualize the results in Figure 5. Most participants would use AI models as visual assistants, but there are some who are reluctant to use them. Participants were also asked to list situations where AI would be most helpful as a visual assistant, providing their responses as comma-separated elements. We left the question open to gather insights into the settings and tasks participants perform using AI models. The third question asked participants about problems and challenges they have experienced when using AI models as visual assistants.

We conducted an iterative thematic analysis to better understand participants' perceptions of the open-ended questions, following (Liao et al., 2024). Two authors reviewed and coded all responses into thematic categories. They then met with the research team to compare and finalize the themes. For the user tasks questions, we tried to use keywords from the responses to stay closer to the original task; for the challenges question, we tried to group the concerns under more generic themes.

In Figure 7 we present a wordcloud of term frequencies of extracted themes from the responses regarding AI problems and challenges. We also present further details of the most recurrent themes along with definitions and examples that justify their grouping in Table 7.

The last section, as shown in Figure 12, was optional and asked for any additional feedback or comments, but we only collected 40 responses, and most of them had no new insights.

Phase 2: Likert Scale Questions. For the second phase of the survey, we asked the same participants after they provided their open-ended question answers to rate specific tasks and challenges. The task was to indicate on a scale of 1 to 5 how likely they are to use AI models for any of the following tasks: Image Captioning, Image Question Answering (IQA), Braille support, Video Question Answering (VQA), and Navigation. The exact phrasing of the questions can be seen in Figure 13. We then asked them to indicate how problematic their shortcomings are related to image quality, language barriers, misinformation, latency, and bias. These categories were chosen based on the discussions we had with visually impaired users in the survey design phase. The exact phrasing of the questions can be seen in Figure 14.

The results from Phase 2, presented in Figure 8 and Figure 9, indicate a growing adoption of AI models for tasks such as image captioning, question answering, and Braille recognition. However, opinions on using AI for navigation are more varied, with responses distributed across all possible values, suggesting that participants are not certain

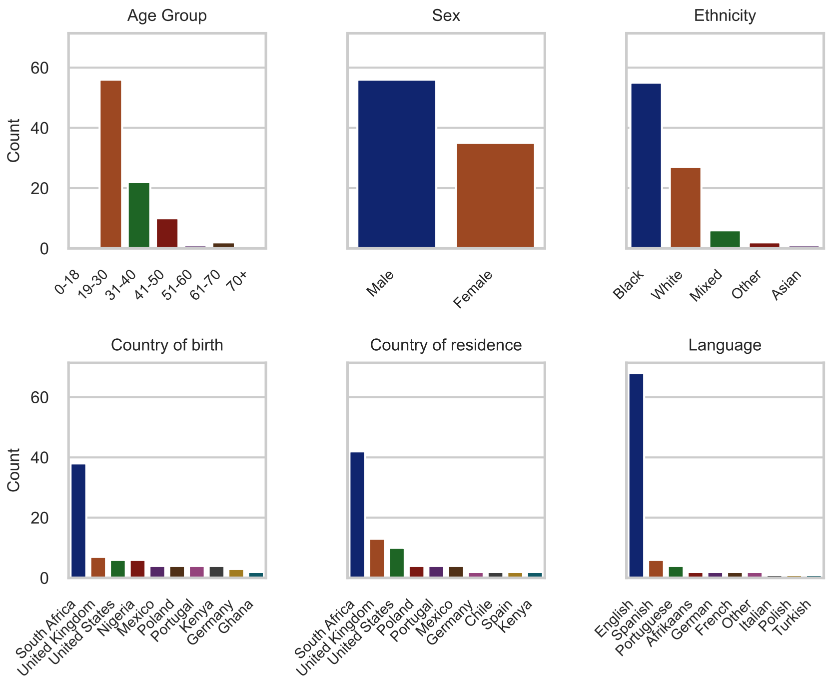


Figure 4: Age, Gender, and Ethnicity demographics extracted from Prolific after filtering the data to remove the "revoked_consent" options.

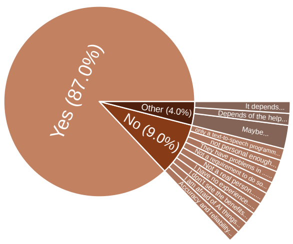


Figure 5: Responses on the potential adoption of AI models as visual assistants.

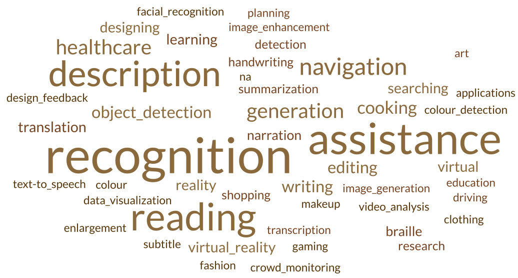


Figure 6: Visualizing all the user cases listed in our survey under the tasks open-ended question.

about using AI models as navigators.

Regarding challenges, misinformation appears to be the most common issue faced by participants, followed by language-related difficulties. Image quality also poses a problem, potentially affecting the reliability of AI-generated descriptions. While latency is a concern for many, a significant number of participants remain neutral on this issue. Notably, bias does not seem to be a major issue among respondents, indicating that, at least in their experience, AI models are perceived as relatively fair in their outputs.

We refrain from making broad generalizations and encourage readers to interpret these findings in the context of our sample size. Future research with larger and more diverse participant groups may provide further insights into these trends.

B Visual Question Answering

Given the lack of multilingual visual question-answering datasets focused on visually impaired users (Karamolegkou et al., 2024), and the challenges in evaluating multilingual models for visual question answering (Pfeiffer et al., 2022), we decided to create a multilingual version of the existing VizWiz dataset (Gurari et al., 2018). The original dataset is under CC BY 4.0 license [7].

B.1 Multilingual Dataset Construction

Translation and Filtering We extend the VizWiz dataset to a multilingual setting by automatically translating the original questions and answers into 35 languages from the XM3600 benchmark

[7] https://creativecommons.org/licenses/by/4.0/

Theme	Description	Example
Inaccuracies	Whether the model provides accurate predictions	*Accuracy is one issue whether this is about objects or faces, Innacurate object detection; gives wrong directions; can provide inaccurate information*
Context	Ability to interpret information based on surrounding factors, background knowledge, and situational cues	*problematic contextual understanding; inadequate localization and navigation; limited scene understanding; AI may not grasp complex spatial relationships or context; difficulty in recognizing context; limited understanding of social cues*
Recognition	Whether the model can recognize objects, faces, characters	*"Inaccurate Object Recognition; Object recognition errors"; difficulty recognizing elements in dimly lit environments; recognizing small or ambiguous objects or text; limitations in recognizing facial expressions; error in recognizing text in blurred or obstructed images*
Description	Whether the model can describe objects or scenes (from image or video)	*"Difficulty describing complex scenes with multiple objects; Bad photo descriptions; inability to describe nuanced scenes; inability to describe subjective or abstract contexts and challenges in distinguishing similar objects"*
Languaculture	Difficulties in effectively using or understanding language in multilingual or multicultural settings	*Speaking in my native language is not natural; Does not speak my language; They cannot identify culture-specific photographs*
Privacy	Whether data is kept private and stored securely	*Constant image and audio processing could raise security issues; Privacy concerns when analyzing personal images or surroundings*
Miscommunication	Communication barriers and misinterpretation of user input or intent	*They misunderstand what I mean; I dont know how to describe something to an AI model or how to get a correct response; Not effectively understanding my need or description of the question*
Quality Dependency	Reliance on high-quality training data for accurate outputs	*Dependency on high-quality data; Problems with blurry images that are too colorful or lower quality; Weather and lighting conditions and poor image quality can reduce accuracy*
Latency	The delay in processing or response time	*Speed of process is slow; Delays in processing can affect real-time assistance; Slow responses; No fast natural human-like answers*
Trust	The belief or confidence in the reliability, and truth of the model outputs	*I cannot trust it with confidence; No trust in description of images; There is no trust between users and technology; Detailed information cannot be trusted*

Table 7: Themes found in phase 1 question about problems and challenges participants have experienced when using AI models as visual assistants.

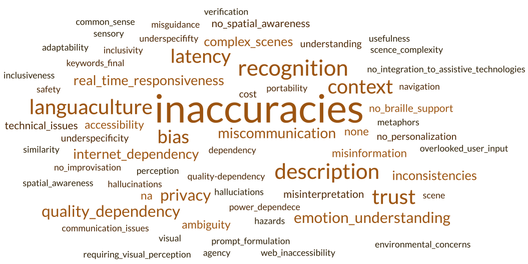


Figure 7: Visualizing all the concerns listed in our survey under the challenges open-ended question.

(Thapliyal et al., 2022), shown in Table 8. We exclude Cuzco Quechua because it is not supported by most translation models. We use the NLLB-Distilled-1.3B (Costa-jussà et al., 2022) model to translate the question-answer pairs, given its strong performance and extensive language coverage. Additionally, we sample for translation a stratified subset of 500 questions, utilizing the skill annotations to ensure representative coverage of different visual question-answering scenarios.

We follow the automatic translation process described by (Yue et al., 2024). We generate multiple translations of each question-answer pair and employ backtranslation for filtering. Specifically, we keep the translation whose backtranslation to English has the highest BLEU score (Papineni et al., 2002) with the original input as reference.

Evaluation For evaluation, we follow the VizWiz framework which relies on multiple answer references to compute the model accuracy. We extend the answer preprocessing to include non-English punctuation symbols and additionally perform unicode normalization on both predicted and ground truth answers.

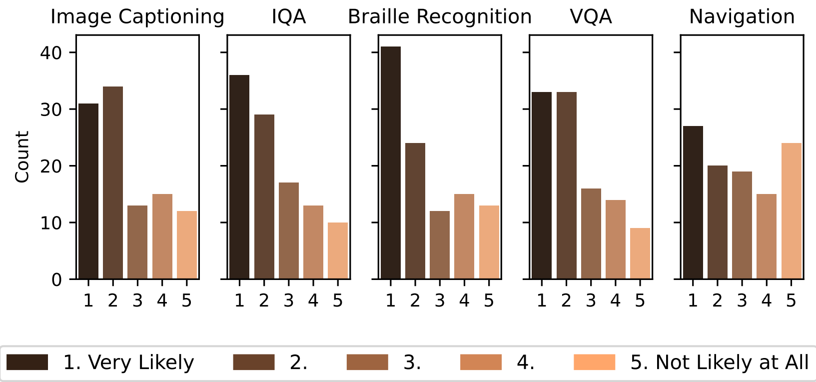


Figure 8: Likert scale responses for AI usage scenarios. The x-axis represents user responses ranging from 1 (Very Likely) to 5 (Not Likely at All), while the y-axis shows the count of responses for each question.

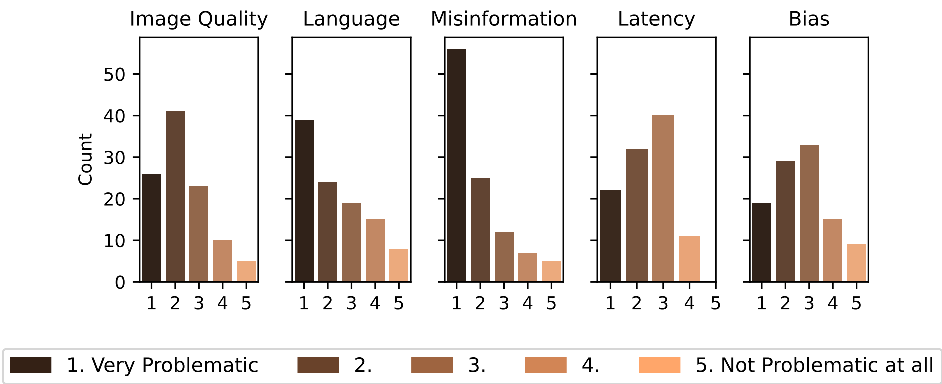


Figure 9: Likert scale responses for AI challenges. The x-axis represents user responses ranging from 1 (Very Problematic) to 5 (Not Problematic at All), while the y-axis shows the count of responses for each question.

B.2 Human Evaluation of Automatic Translation

To validate the quality of the automatically translated VizWiz QA data, we run a human evaluation. The evaluation focused on assessing the quality of machine-translated questions and answers while quantifying translation errors. At least 20 translated questions in random order were reviewed per language, each accompanied by 10 similar short answers. Evaluators assessed only the quality of the target language translation and provided relevant examples and comments in a spreadsheet tab corresponding to each language. The human evaluation guidelines are shown in Figure 15.

The study examined 7 languages selected based on the authors' fluency. Each language was assessed using 220 data points, resulting in a total of 1,540 translated questions and answers. Of these, 257 were labeled as incorrect, yielding an average translation error rate of 16.28%, representing the proportion of translations with noticeable errors.

Most errors occurred because of improper translation of English brand names, which were mistakenly translated as generic words or altered (e.g., Gevalia Coffee, Diet Coke, Dr Pepper, Windows PC, LG, Mrs. Dash, Manwich). Additionally, there were issues with yes/no questions, where some languages produced incorrect responses such as double 'yes, yes', 'I don't know', or 'I am sorry' instead of a simple yes or no.

Some errors also resulted from problematic original answers that contained typos or ungrammatical phrases, such as 'can diet' instead of 'a can of Diet Coke' or 'ginerale' instead of 'ginger ale'. A notable case involved the number 321, which was mistranslated as a random sentence rather than being retained as a numeral.

Finally, two ambiguous words in the validation set—'denomination' and 'dressing'—posed challenges. Since the responses consisted of short, context-free answers, some models translated them with one interpretation, while others chose a different meaning, leading to inconsistencies across languages and the meaning of the correct response.

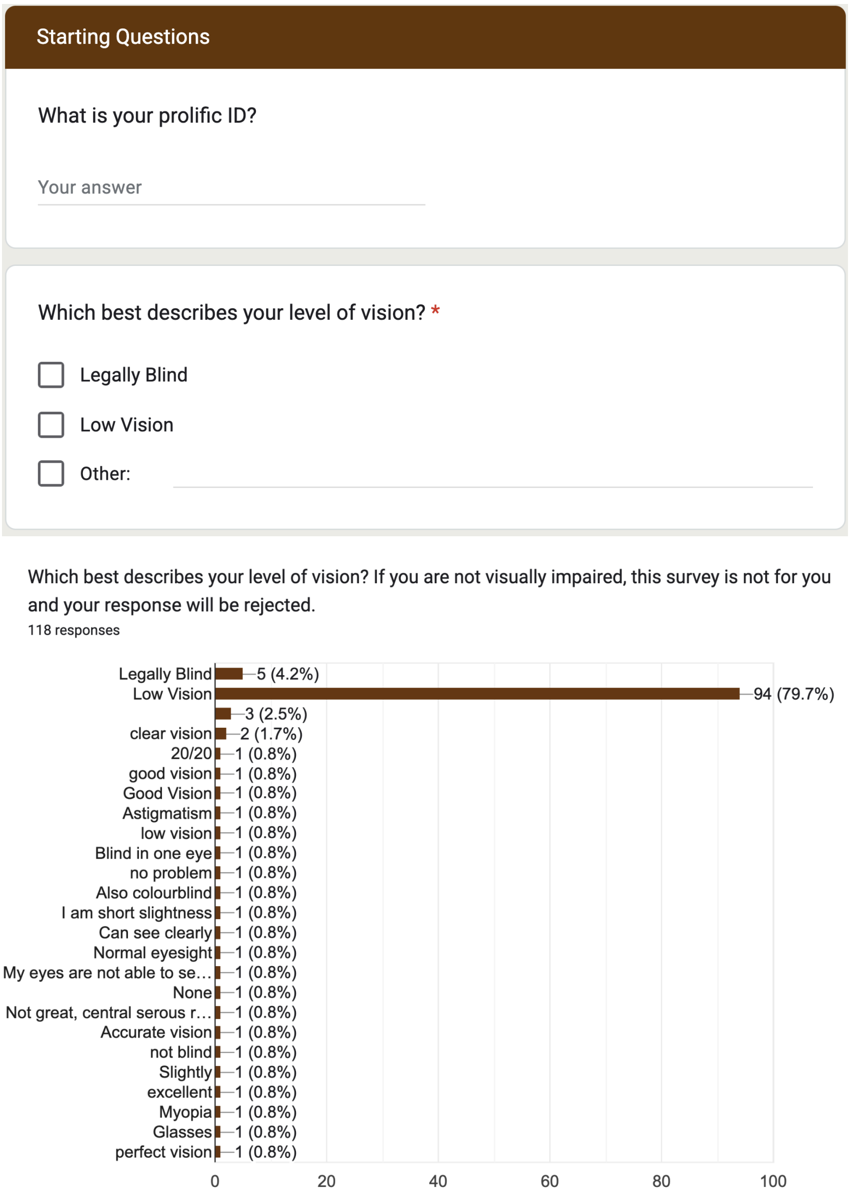
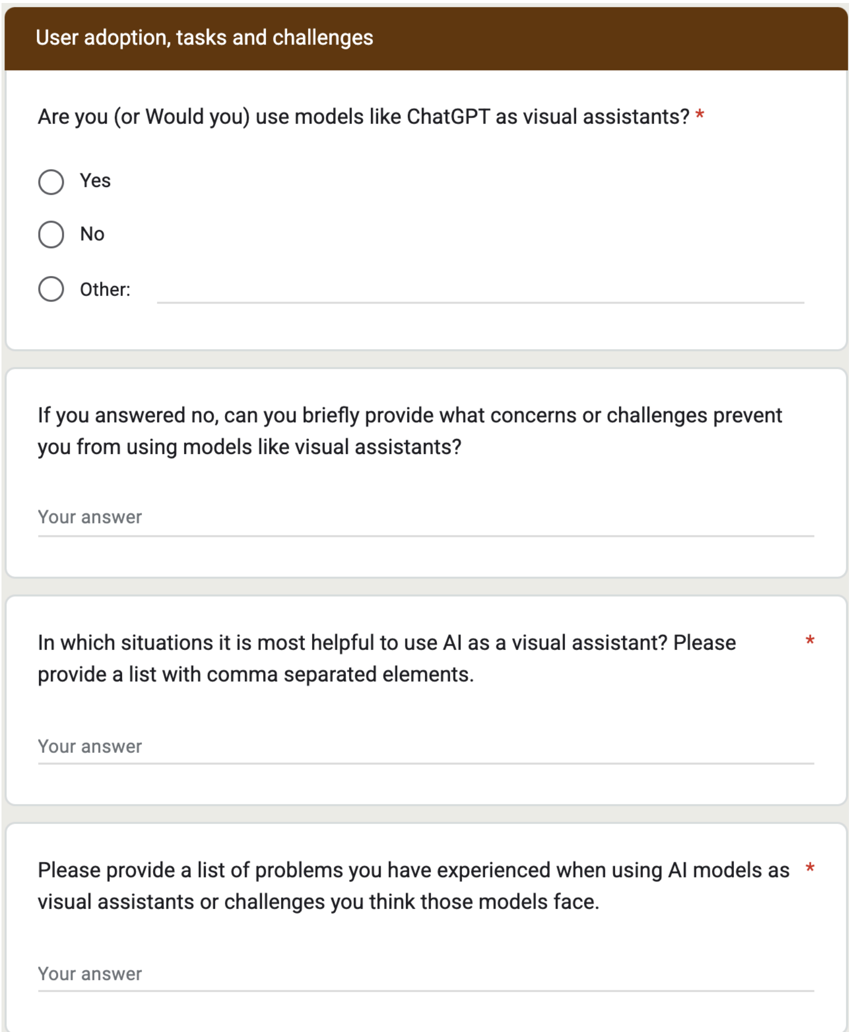


Figure 10: Phase 1: First section of the survey



User adoption, tasks and challenges

Are you (or Would you) use models like ChatGPT as visual assistants? *

- Yes
- No
- Other:

If you answered no, can you briefly provide what concerns or challenges prevent you from using models like visual assistants?

Your answer

In which situations it is most helpful to use AI as a visual assistant? Please provide a list with comma separated elements. *

Your answer

Please provide a list of problems you have experienced when using AI models as visual assistants or challenges you think those models face. *

Your answer

Figure 11: Phase 1: Second section of the survey

Mission accomplished! Thank you!

Feel free to provide any additional feedback/problems/comments one should take into account regarding using AI models as visual assistants.

40 responses

none

N/A

None

Make longer surveys so that one can learn more.

nothing at the moment

None.

It was great to be participating

I'm glad AI's are being made to assist people with low vison

n/a

should be clear

No feedbacks, thank you.

Okay

NA

aI might overlook crucial but subtle elements in an image

No comment everything was clear

No, thank you

no comment

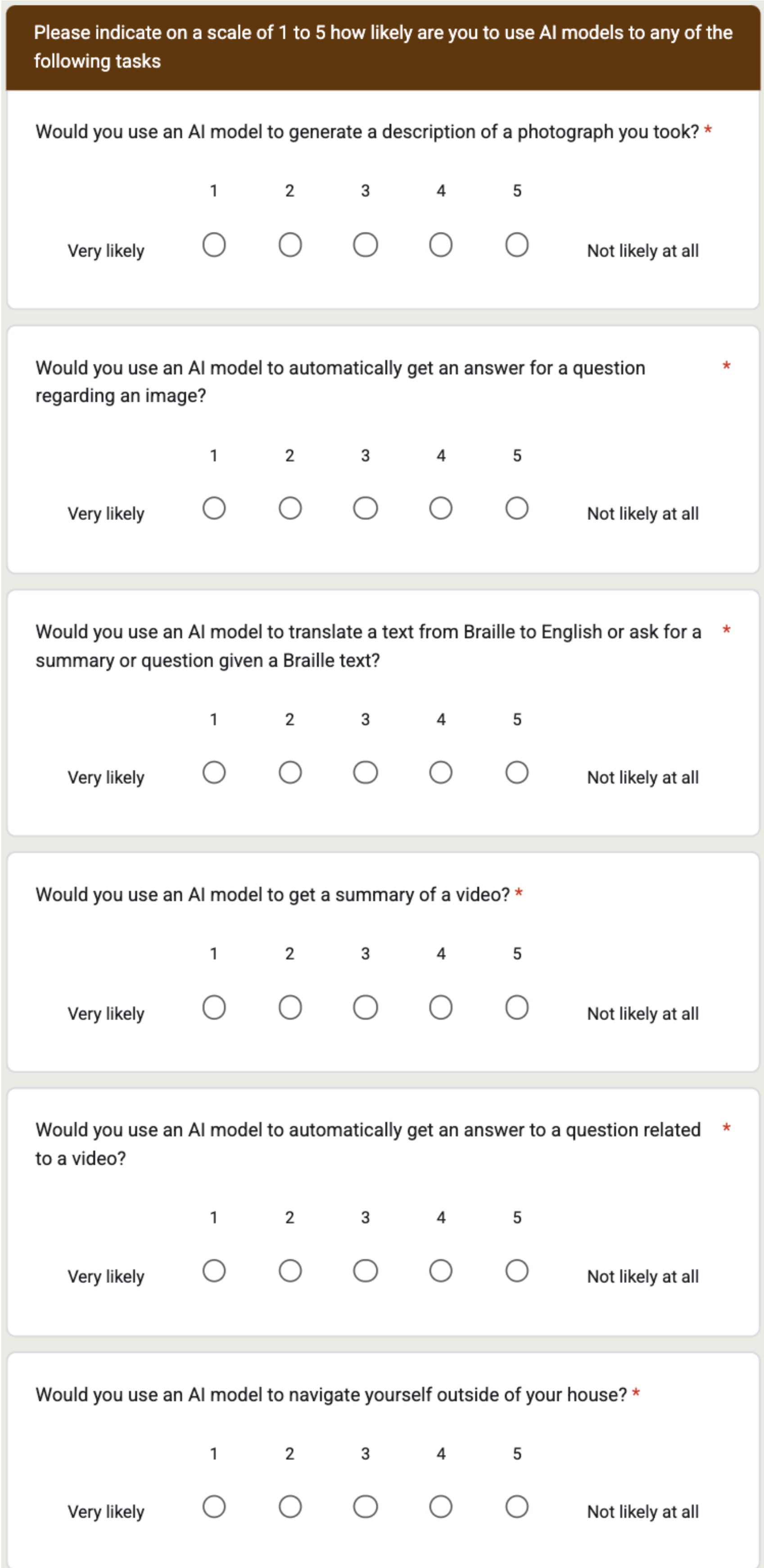
good

one should avoid taking an imagine in direct sun light

-

it was a good survey,i enjoyed it

Figure 12: Phase 1: Last question before we direct the participants to the second phase.



Please indicate on a scale of 1 to 5 how likely are you to use AI models to any of the following tasks

Would you use an AI model to generate a description of a photograph you took? *

	1	2	3	4	5	
Very likely	○	○	○	○	○	Not likely at all

Would you use an AI model to automatically get an answer for a question regarding an image? *

	1	2	3	4	5	
Very likely	○	○	○	○	○	Not likely at all

Would you use an AI model to translate a text from Braille to English or ask for a summary or question given a Braille text? *

	1	2	3	4	5	
Very likely	○	○	○	○	○	Not likely at all

Would you use an AI model to get a summary of a video? *

	1	2	3	4	5	
Very likely	○	○	○	○	○	Not likely at all

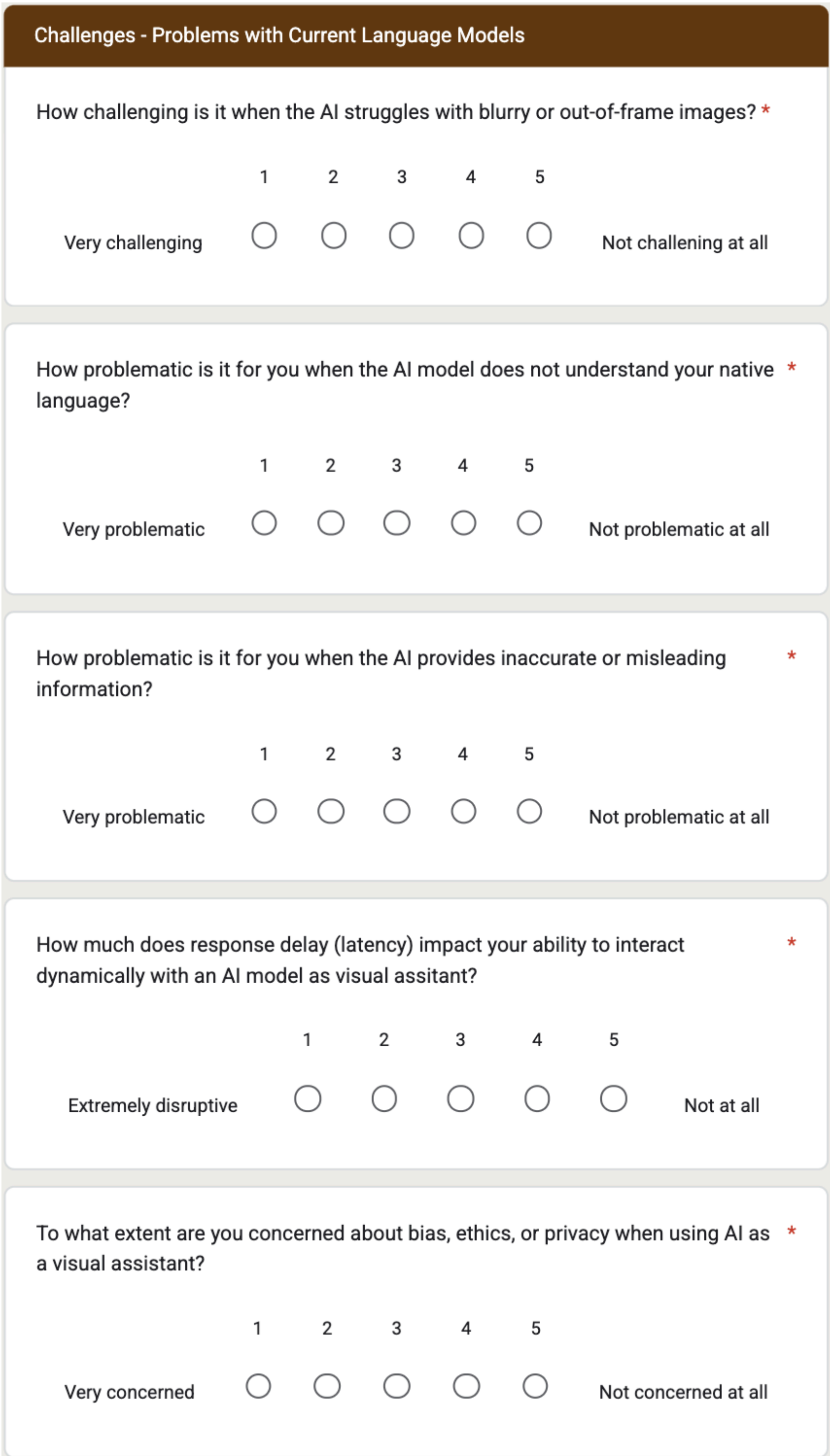
Would you use an AI model to automatically get an answer to a question related to a video? *

	1	2	3	4	5	
Very likely	○	○	○	○	○	Not likely at all

Would you use an AI model to navigate yourself outside of your house? *

	1	2	3	4	5	
Very likely	○	○	○	○	○	Not likely at all

Figure 13: Phase 2: Asking participants to rate tasks that AI can be used for.



Challenges - Problems with Current Language Models

How challenging is it when the AI struggles with blurry or out-of-frame images? *

	1	2	3	4	5	
Very challenging	○	○	○	○	○	Not challening at all

How problematic is it for you when the AI model does not understand your native language? *

	1	2	3	4	5	
Very problematic	○	○	○	○	○	Not problematic at all

How problematic is it for you when the AI provides inaccurate or misleading information? *

	1	2	3	4	5	
Very problematic	○	○	○	○	○	Not problematic at all

How much does response delay (latency) impact your ability to interact dynamically with an AI model as visual assitant? *

	1	2	3	4	5	
Extremely disruptive	○	○	○	○	○	Not at all

To what extent are you concerned about bias, ethics, or privacy when using AI as a visual assistant? *

	1	2	3	4	5	
Very concerned	○	○	○	○	○	Not concerned at all

Figure 14: Phase 2: Asking participants to rate challenges they have encountered when using AI models.

Language	ISO Code	Script	Resource
Arabic	ar	Arabic	High
Bengali	bn	Bengali	Mid
Czech	cs	Latin	High
Danish	da	Latin	Mid
German	de	Latin	High
Greek	el	Greek	Mid
English	en	Latin	High
Spanish	es	Latin	High
Persian	fa	Arabic	High
Finnish	fi	Latin	High
Filipino	fil	Latin	Mid
French	fr	Latin	High
Hebrew	he	Hebrew	Mid
Hindi	hi	Devanagari	High
Croatian	hr	Latin	Mid
Hungarian	hu	Latin	High
Indonesian	id	Latin	Mid
Italian	it	Latin	High
Japanese	ja	Japanese	High
Korean	ko	Hangul	High
Māori	mi	Latin	Low
Dutch	nl	Latin	High
Norwegian	no	Latin	Low
Polish	pl	Latin	High
Portuguese	pt	Latin	High
Romanian	ro	Latin	Mid
Russian	ru	Cyrillic	High
Swedish	sv	Latin	High
Swahili	sw	Latin	Low
Telugu	te	Telugu	Low
Thai	th	Thai	Mid
Turkish	tr	Latin	High
Ukrainian	uk	Cyrillic	Mid
Vietnamese	vi	Latin	High
Chinese	zh	Han	High

Table 8: Language, ISO-Codes, script, and resource levels.

We are going to release the dataset validation spreadsheet with the translation errors and the languages after the anonymity period.

B.3 Further Results

We report performance per language script in Table 10, and per language in Table 11.

Model	High	Mid	Low
Idefics3	24.8	20.8	21.7
InternVL2.5-MPO	40.3	36.4	39.6
Llava-v1.6	41.9	37.7	43.3
Llama-3.2-Vision-Instruct	29.8	27.6	33.4
Molmo	28.2	28.5	32.6
MiniCPM-2.6	32.1	31.0	22.7
Paligemma	19.5	13.7	11.2
Phi-3-Vision-Instruct	36.9	36.3	34.6
Qwen2-VL-Instruct	44.8	44.9	42.9

Table 9: Accuracy on multilingual VizWiz grouped based on the language characterization as High, Mid, and Low resource.

C Optical Braille Recognition

Dataset Creation We generate rendered images of Braille text as summarized in Section 7.1. We apply augmentations to the images from both transcription and cross-script QA tasks using the imgaug library (Jung et al., 2020). More specifically, we use color, edge, geometric, contrast, and blur transformations families, where an image can be transformed with multiple of these augmentations at the same time. For color, we select one of posterize, color quantization, and color temperature. With regards to edge transformation, we either sharpen, emboss the image, or convert edges into black or white and overlay the resulted transformation with the original image. For geometric transformations, we shear the image over the width or height, or rotate the image. Additionally, we scale pixel values by a fixed gamma constant. Finally, we apply either gaussian, bilataral, motion or mean shift blur. All augmentations are applied in random order. The values for all of the parameters along with scripts to reproduce the augmentations is available `linked removed for review`.

Table 14 illustrates examples of inputs-outputs for both tasks where the Braille text is rendered in images that have been augmented and the model needs to output plain English text. Note that in all cases, the correct output cannot be inferred unless the model is able to read the Braille content from the image.

Training Logs & Hyperparameters Table 12 illustrates the hyperapameters used to finetune Llama-3.2-Vision-Instruct on both tasks for Optical Braille Recognition. Note that the LoRA adapters are applied to the key and value weight matrices in each transformer layer following the default implementation (Hu et al., 2022). We expect that applying the adapters to other linears can further improve performance. All experiments were conducted using 1xH100 GPU. Training logs for all runs are available `linked removed for review`.

D Video Object Recognition and Question Answering

ORBIT dataset ORBIT (Massiceti et al., 2021) is a dataset of videos collected by people who are blind/low-vision, originally collected for few-shot object recognition. The dataset includes "clean" videos, which show an object in isolation, and "clutter" videos, which show the target object in the con-

Human Evaluation Instructions

MULTILINGUAL TRANSLATION EVALUATION

Objective:
Your task is to evaluate the translation quality of at least 20 machine-translated questions and their corresponding answers. (1 question has 10 similar short answers.) Focus only on the quality of the target language translation, not the accuracy of the question-answer content. Translations are provided in a JSON file, and results should be recorded in the spreadsheet tab labeled with your language name.

Evaluation Process:

1. **Review Translations:**
   - Read the translated answers for each question in the JSON file.
   - If unsure about a translation, retrieve the original question using the image ID on this platform: `https://vizwiz.cs.colorado.edu/VizWiz_visualization/view_dataset.php`.
2. **Identify Errors:** For each translated question-answer pair, check for errors. For example you can identify:
   - **Grammatical Errors:** Incorrect grammar or sentence structure.
   - **Lexical Errors:** Incorrect word choices or omissions.
   - **Formatting Errors:** Issues with punctuation or capitalization.
3. **Assign Error Severity:**
   - **Minor:** Small errors that do not impact the meaning.
   - **Moderate:** Errors that partially affect clarity or meaning.
   - **Severe:** Errors that significantly alter or obscure the meaning.
4. **Count Errors:**
   - Track the total number of errors for each translated QA pair.
   - Provide a severity score for each error identified.

Report Results:
Record your results in the spreadsheet using the provided structure:

`ImageID, Error Count, Error Type, Comments, Examples`

If a QA pair has multiple error types, separate them with commas under *Error Type*.

Figure 15: Guidelines for Multilingual Translation Evaluation.

Model	Latin	Han	Japanese	Hangul	Cyrillic	Arabic	Devanagari	Hebrew	Thai	Telugu	Greek	Bengali
Idefics3	26.1	11.2	8.9	19.2	17.5	23.7	27.0	25.9	11.5	13.4	20.8	20.3
InternVL2.5-MPO	40.5	44.1	39.7	25.9	27.2	41.3	35.6	43.6	42.0	39.6	39.3	29.3
Llava-v1.6	42.9	43.2	34.3	42.0	24.9	42.5	41.2	44.0	42.6	42.9	41.7	19.0
Llama-3.2-Vision-Instruct	31.1	35.9	23.5	20.7	23.2	30.7	16.9	34.3	37.9	30.9	26.8	18.0
Molmo	31.8	23.6	6.8	32.1	25.5	22.0	12.8	42.3	19.4	30.1	33.2	12.5
MiniCPM-2.6	34.0	43.4	32.2	21.4	24.4	22.1	11.0	31.8	33.5	14.4	34.9	10.3
Paligemma	18.4	15.5	26.9	22.9	13.3	8.9	9.9	10.8	13.1	20.5	9.9	12.5
Phi-3-Vision-Instruct	38.7	33.7	31.5	34.3	26.0	33.0	30.4	37.7	37.7	30.1	35.5	35.5
Qwen2-VL-Instruct	45.5	42.7	36.9	44.6	43.0	43.9	44.2	42.2	46.3	42.9	46.4	42.1

Table 10: Accuracy on multilingual VizWiz per script.

Model	ar	bn	cs	da	de	el	en	es	fa	fi	fil	fr	he	hi	hr	hu	id	it
Idefics3	25.7	20.3	22.4	27.8	37.4	20.8	50.4	26.8	21.7	25.6	14.1	22.4	25.9	27.0	21.8	19.7	30.9	31.9
InternVL2.5-MPO	42.9	29.3	38.9	36.3	43.3	39.3	60.2	37.8	39.8	41.0	43.3	40.7	43.6	35.6	41.8	32.4	35.8	41.6
Llava-v1.6	43.6	19.0	40.8	43.3	46.5	41.7	60.1	42.0	41.4	40.7	43.7	44.0	44.0	41.2	39.4	41.8	45.1	44.7
Llama-3.2-Vision-Instruct	41.3	18.0	32.0	20.8	33.4	26.8	45.1	35.8	20.1	33.3	15.7	32.2	34.3	16.9	28.3	23.7	37.2	36.9
Molmo	38.9	12.5	24.6	27.2	23.6	33.2	43.6	31.7	5.2	25.4	38.4	32.6	42.3	12.8	29.7	32.4	40.8	31.6
MiniCPM-2.6	39.8	37.6	16.9	3.6	8.3	36.8	30.8	44.0	32.4	34.1	30.3	39.5	55.3	70.2	34.3	40.8	5.0	6.0
Paligemma	5.6	12.5	12.3	13.5	26.2	9.9	78.3	10.6	12.2	26.0	16.7	4.7	10.8	9.9	9.9	16.6	19.4	12.0
Phi-3-Vision-Instruct	33.6	35.5	36.6	38.3	37.5	35.5	51.0	37.4	32.5	36.9	41.0	36.3	37.7	30.4	38.2	39.6	41.1	39.4
Qwen2-VL-Instruct	45.3	42.1	41.6	47.2	48.2	46.4	68.6	43.7	42.5	45.6	47.8	44.8	42.2	44.2	44.7	35.4	48.9	44.7

Model	ja	ko	mi	nl	no	pl	pt	ro	ru	sv	sw	te	th	tr	uk	vi	zh
Idefics3	8.9	19.2	28.9	27.8	25.8	25.3	26.1	22.3	22.7	21.8	18.8	13.4	11.5	17.2	12.4	28.8	11.2
InternVL2.5-MPO	39.7	25.9	43.3	44.4	40.5	40.5	40.2	42.6	44.8	30.3	34.8	39.6	42.0	40.0	9.6	41.8	44.1
Llava-v1.6	34.3	42.0	43.0	45.3	44.0	41.4	41.6	40.8	32.3	36.5	43.2	42.9	42.6	35.0	17.4	40.6	43.2
Llama-3.2-Vision-Instruct	23.5	20.7	32.2	30.8	39.9	26.2	39.3	32.4	21.3	21.9	30.6	30.9	37.9	25.6	25.0	30.7	35.9
Molmo	6.8	32.1	39.2	28.1	33.0	28.5	38.8	30.8	40.2	37.7	28.0	30.1	19.4	17.4	10.8	36.0	23.6
MiniCPM-2.6	36.3	20.4	38.3	27.5	36.5	37.0	35.8	27.7	16.7	5.3	34.7	38.3	21.5	33.7	41.0	42.0	33.8
Paligemma	26.9	22.9	1.2	41.2	12.7	12.1	18.2	21.2	16.9	19.2	10.5	20.5	13.1	13.8	9.7	9.3	15.5
Phi-3-Vision-Instruct	31.5	34.3	39.2	36.5	38.5	38.3	34.5	43.5	37.5	43.8	30.5	30.1	37.7	42.0	14.5	31.7	33.7
Qwen2-VL-Instruct	36.9	44.6	41.9	46.4	43.7	45.3	37.9	46.8	49.0	44.9	43.1	42.9	46.3	42.3	37.1	47.1	42.7

Table 11: Accuracy on multilingual VizWiz per language.

Hyperparameter	**Values**
global batch size	64
LR	{1e-5, 1e-4, 5e-4}
lr schedule	cosine decay
LR warmup	0.03
number of epochs	1
optimizer	AdamW
LoRA rank	{32, 64, 128, 256}
LoRA alpha	{16, 32, 64, 128, 256, 512}

Table 12: Hyperparameters during both finetuning on both sentence-level and paragraph-level tasks.

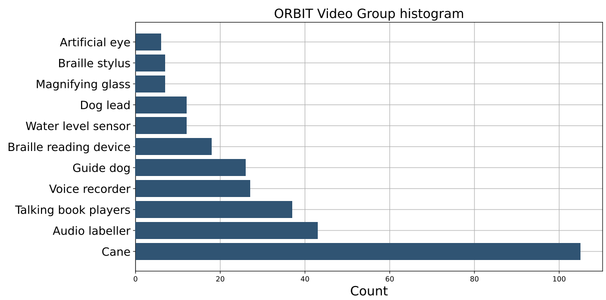


Figure 17: Number of videos per assistive category.

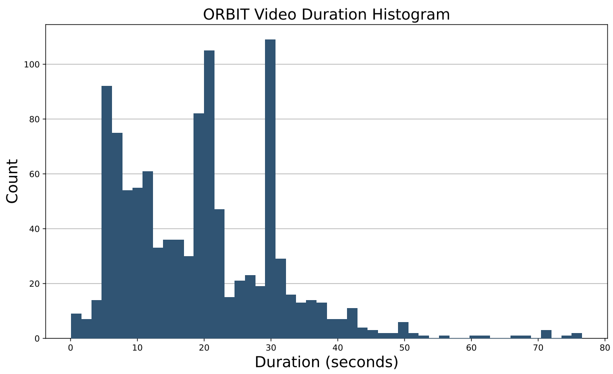


Figure 16: Video duration histogram.

text of other items. The target objects are labelled by the participants and grouped into clusters by the dataset authors. Videos are provided at 1080x1080 frame resolution and 30 frames per second. We utilize 1069 video clips from 51 participants and 92 object clusters, with a median duration of 19.7 seconds (see Figure 16 for the video duration distribution). The videos include household objects, which are general everyday objects (e.g., TV remote, house keys, wallet) and assistive items (e.g., Braille display, white cane, liquid level indicator), as illustrated in Figure 17.

D.1 Video Object Recognition Dataset Construction

For video object recognition, we use the dataset provided from (Masssiceti et al., 2021). The dataset is under CC BY 4.0 license [8]. We select 512 "clean" and 514 "clutter" videos through stratified sampling across object categories. We convert the dataset into a question-answering format using a two-step semi-automatic process. First, we prompt a language model to extract a representative keyword for each object cluster. Second, based on these keywords and object labels, we generate an object recognition question for each group. The prompts used for the dataset creation are shown in Figure 20. Finally, the generated questions are reviewed manually and adjusted if needed.

[8] https://creativecommons.org/licenses/by/4.0/

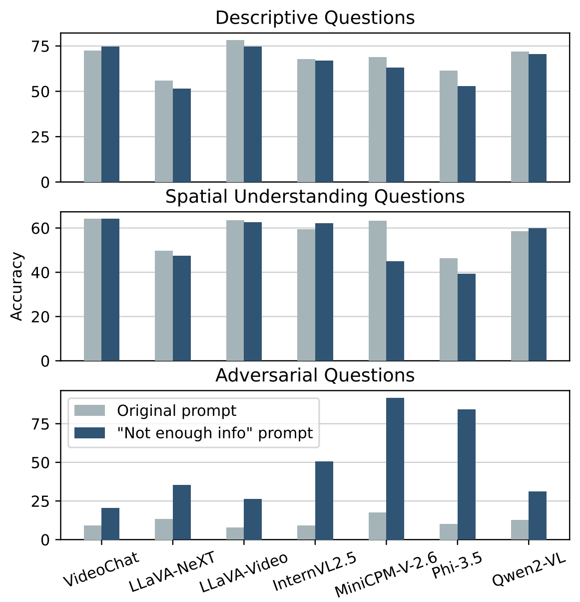


Figure 18: Accuracy on VideoQA when prompting the model to output “Not enough information" as needed.

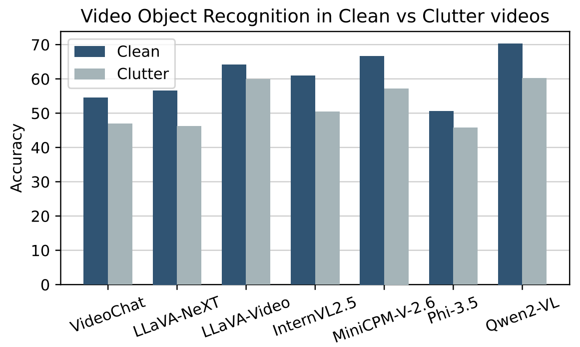


Figure 19: Accuracy in Video Object Recognition in Clean vs Clutter videos.

D.2 Video Question Answering Dataset Construction

Since there is currently no dataset with question answering for videos filmed by visually impaired users, we decided to curate such a dataset using videos from ORBIT (Massiceti et al., 2021). We use only “clutter" videos that provide a more naturalistic setting.

We generate three types of questions: 1) Descriptive Questions, such as questions about color and number of objects, 2) Spatial Understanding, such as questions about the location or spatial relationship between objects and 3) Adversarial Questions which cannot be answered based on the information provided in the video. To generate the questions, we used a manual approach where three of the authors of the paper followed the guidelines provided in Figure 21. We create a total of 882 question-answer pairs (294 per question type).

D.3 Evaluation Metric

Given that only one label is available for each question, we adopt the LAVE metric (Mañas et al., 2024) for evaluation. LAVE uses a language model judge to provide a rationale and a rating between 1-3. Ratings are then normalized in the range [0, 1]. We use Llama-3.3-70B-Instruct (AI@Meta, 2024) as the language model.

D.4 Further Results

Given the low performance on adversarial questions, we explore whether explicit prompting can mitigate this shortcoming. We modify the prompt to instruct models to respond with “Not enough information" when the video content is insufficient to answer the question. As shown in Figure 18, performance in Adversarial Questions consistently improves with the “Not enough information" prompt. For most models, however, performance remains poor (at most 50% accuracy) with minimal effect on other question types. This suggests that models continue to hallucinate answers frequently. While performance increases drastically for MiniCPM-V-2.6 and Phi-3.5-Vision-Instruct, this comes at the cost of performance in other categories, as models tend to unnecessarily over-generate the “Not enough information" response. These results suggest that current prompting strategies alone cannot reliably prevent hallucination in video question answering—a critical safety concern for assistive applications.

E Models

Table 13 reports the details for selected models, including the Huggingface tag used when accessing the model, the total number of model parameters, and whether models support image, video, and multilingual inputs. Additionally, we report whether VizWiz is included in the model’s training data. We find no evidence that any models are exposed to the ORBIT dataset. Note that Paligemma is the only model that is not instruction fine-tuned, which is why we exclude it from zero-shot results in optical Braille recognition (Table 3).

F Examples

Table 14 shows example input-output pairs for Optical Braille Recognition.

Model	Huggingface Tag	Param	Image	Video	Multilingual (# Langs)	Trained on VizWiz
Idefics3 (2024)	HuggingFaceM4/Idefics3-8B-Llama3	8B	✓	✗	✗	✗
InternVL2.5-MPO (2024d)	OpenGVLab/InternVL2_5-8B-MPO	8B	✓	✓	✓(11)	✗
LLaVA-NeXT-Video (2024a)	llava-hf/LLaVA-NeXT-Video-7B-hf	7B	✗	✓	✗	✗
LLaVA-Video (2024b)	lmms-lab/LLaVA-Video-7B-Qwen2	7B	✗	✓	✗	✓*
LlaVA-v1.6 (2024a)	llava-hf/llava-v1.6-mistral-7b-hf	8B	✓	✗	✗	✗
Llama-3.2-Vision-Instruct (2024)	meta-llama/Llama-3.2-11B-Vision-Instruct	11B	✓	✗	✗	–
MiniCPM-V-2.6 (2024)	openbmb/MiniCPM-V-2_6	8B	✓	✗	✓(36)	✓
Molmo (2024)	allenai/Molmo-7B-D-0924	7B	✓	✗	✗	✗
Paligemma (2024)	google/paligemma-3b-mix-448	3B	✓	✗	✓(35)	✓
Phi-3.5-Vision-Instruct (2024)	microsoft/Phi-3.5-vision-instruct	4B	✓	✓	✓(–)	–
Qwen2-VL-Instruct (2024c)	Qwen/Qwen2-VL-7B-Instruct	8B	✓	✓	✓(–)	–
VideoChat-Flash (2024d)	OpenGVLab/VideoChat-Flash-Qwen2-7B_res448	8B	✓	✓	✗	✗

Table 13: Model Details. The model pool is limited to 1) open-source/weights models with 2) strong image or video understanding capabilities and 3) medium computational overhead. '–' is used when there is insufficient public information to determine the value. * VizWiz included in the image training phase before video instruction tuning.

G Acknowledgments

We acknowledge the use of GitHub Copilot[9] in the implementation of our research. All final code is verified by the authors.

[9]https://github.com/features/copilot

Prompts for ORBIT Video Object Recognition Data Generation

KEYWORD EXTRACTION

You will be given a list of objects, and you have to answer with one short word or phrase that can be used to describe the group.

Examples
Objects: [watch, wrist watch, apple watch, apple wath, risk watch, my apple watch]
Answer: watch

Objects: [black small wallet, my purse, my wallet, ladies purse, money pouch, coin purse, wallet for bus pass cards and money, i d wallet, ipod in wallet, walletv, wallet, purse]
Answer: wallet

Objects: [orbit Braille reader and notetaker, orbit reader 20 Braille display, Braillepen slim Braille keyboard, Braille orbit reader, Braille note, my Braille displat]
Answer: Braille reading device

Generate the answer for the following objects:

QA DATA GENERATION

You will be given a list of objects and a common label that describes the group. Your task is to generate a question that can be asked to identify an instance of this group in a video.

Examples:
Objects: [slippers, nike trainers, my shoes, boot, trainers, trainer shoe, slipper, my trainers, shoes, running shoes]
Group: shoes
Question: What type of clothing do you see in the video?

Objects: [orbit Braille reader and notetaker, orbit reader 20 Braille display, Braillepen slim Braille keyboard, Braille orbit reader, Braille note, my Braille displat] Group: Braille reading device
Question: What kind of assistive device was there?

Objects: [black small wallet, my purse, my wallet, ladies purse, money pouch, coin purse, wallet for bus pass cards and money, i d wallet, ipod in wallet, walletv, wallet, purse]
Group: wallet
Question: What type of accessory appears in the video?

Generate the question for the following:

Figure 20: Prompts for the Video Object Recognition dataset.

Instructions for ORBIT Video QA Data Generation

ANNOTATION GUIDELINES FOR VIDEO-BASED QUESTION GENERATION

Step1: Video Access Open this link containing short videos: `url` and the json file attached. Watch 50 short video clips and generate 3 short questions + short answers about the clip.
Step2: Question Generation Questions should be designed to help a Visually Impaired Person (VIP) understand and interact with their environment. They should be based on information that can be visibly inferred from the video. The focus should be on:

- Descriptive Questions (D) These questions ask about the appearance, quantity, or basic attributes of objects. Examples: "What is the colour of this item?", "How many X items do you see?", "What shape is this object?"
- Spatial Understanding Questions (S) These questions focus on the location and relationships between objects and people. Examples: "What is next to this item?", "Where is item X?", "Is there an item Y next to item X?"
- Adversarial Questions (A) These questions ask for items or cues not present in the video. Examples: "Is there an X in the image?", "Is there an X item next to the Y item?", "Is the colour of X item green?"

  The answer to this question is always: "Not enough information are depicted in the video to answer this question."

Step3: Answer Generation Answers should be grounded in the information provided in the video. They should be short, clear, concise, and based on the video footage. For example Q: "How many X items are there", A: "four", Q: "Where is X item placed?" A: "inside a kitchen cabinet", Q: "Are there any mangoes next to the toy?" A: "No".
Step4: Write the video name id (eg. "P100–exercise bench–clutter-pan–P100–exercise-bench–clutter"), question, question_type, and answer, in the json file. We provide some examples there for your guidance.

Figure 21: Guidelines for Video Question Answering Data Generation

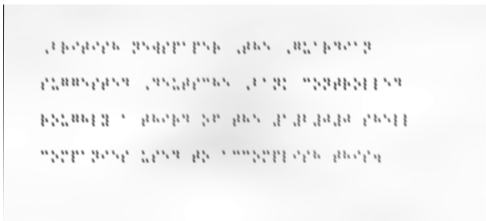
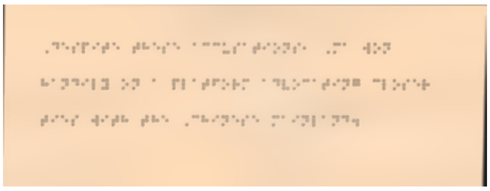
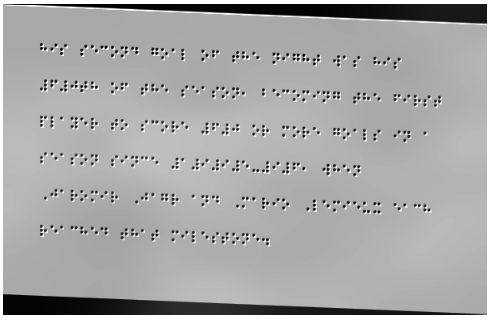
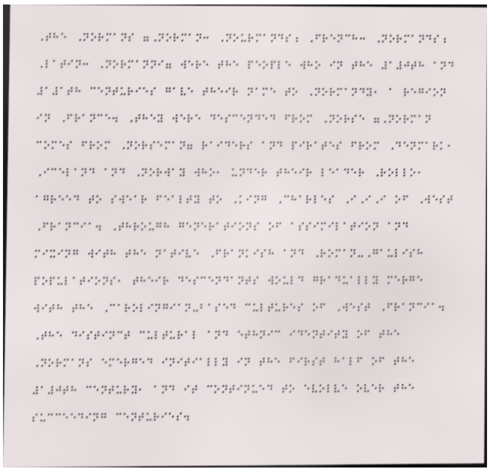
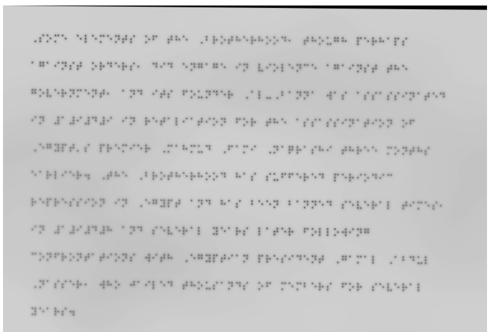
	**Source: WMT2024** **Input:** Transcribe the Braille to English. **Output:** British newspaper The Guardian suggested Deutsche Bank controlled roughly a third of the 1200 shell companies used to accomplish this.
	**Source: WMT2024** **Input:** Transcribe the Braille to English. **Output:** Despite these accusations, Ma won handily on a platform advocating closer ties with the Chinese mainland.
	**Source: WMT2024** **Input:** Transcribe the Braille to English. **Output:** his second goal of the night was his 60th of the season, becoming the first player to score 60 or more goals in a season since 1995-96, when Jaromir Jagr and Mario Lemieux each reached that milestone.
	**Source: SquAD** **Input:** Answer the following question based on the image. If the question is not answerable, output 'unanswerable'. When were the Normans in Normandy? **Output:** 10th and 11th centuries
	**Source: SquAD** **Input:** Answer the following question based on the image. If the question is not answerable, output 'unanswerable'. What Egyptian president jailed hundreds of members of the Brotherhood? **Output:** unanswerable

Table 14: Illustration of inputs-outputs for the two Optical Braille Recognition tasks: Braille-to-Text Transcription and Cross-Script Question Answering. Highlighted text indicates the target output. Note that the model cannot provide the correct response unless it can map Braille symbols to English.