WorldScribe: Towards Context-Aware Live Visual Descriptions

Ruei-Che Chang
rueiche@umich.edu
University of Michigan
Ann Arbor, MI, USA

Yuxuan Liu
liurick@umich.edu
University of Michigan
Ann Arbor, MI, USA

Anhong Guo
anhong@umich.edu
University of Michigan
Ann Arbor, MI, USA

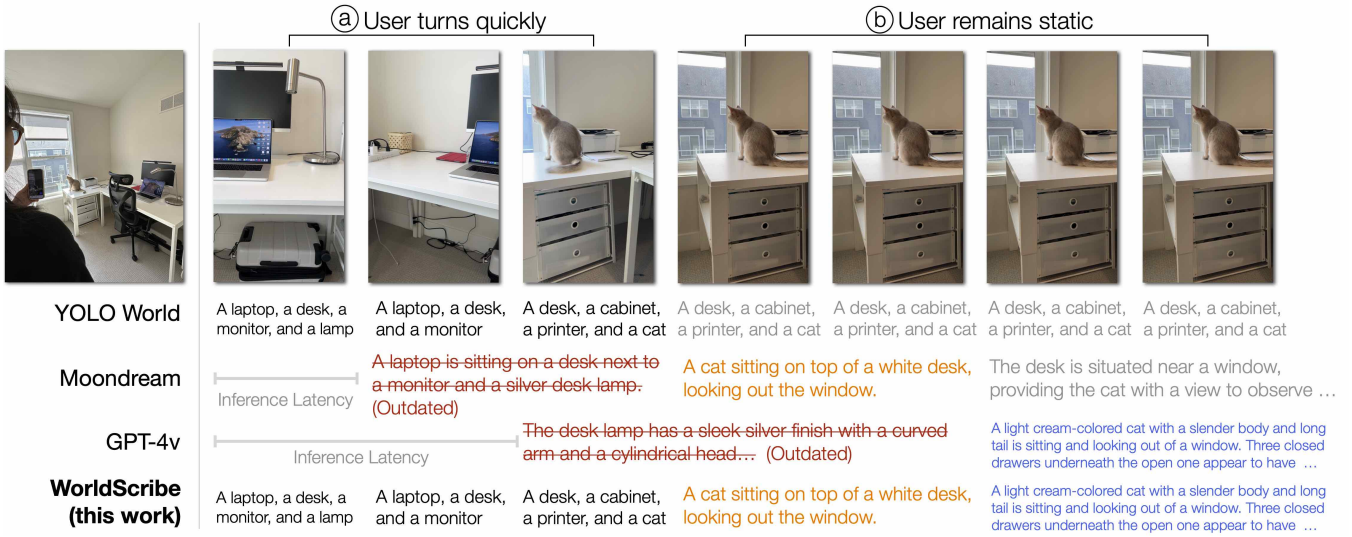


Figure 1: WorldScribe towards making the real world accessible for blind people through context-aware live visual descriptions. WorldScribe dynamically combines different vision-language models to provide live adaptive descriptions. (a) When the user turns quickly to scan the environment and yields frequent visual changes, WorldScribe generates basic descriptions with word-level labels (*e.g.,* YOLO World [27]) or general descriptions with objects and spatial relationships (*e.g.,* Moondream [10]). On the other hand, (b) when the user remains static and faces a new scene for a duration that indicates their interests, WorldScribe provides rich descriptions from an overview to details (*e.g.,* GPT-4v [7]) to facilitate their visual scene understanding.

ABSTRACT

Automated live visual descriptions can aid blind people in understanding their surroundings with autonomy and independence. However, providing descriptions that are rich, contextual, and just-in-time has been a long-standing challenge in accessibility. In this work, we develop *WorldScribe*, a system that generates automated live real-world visual descriptions that are customizable and adaptive to users' contexts: *(i)* WorldScribe's descriptions are tailored to users' intents and prioritized based on semantic relevance. *(ii)* WorldScribe is adaptive to visual contexts, *e.g.,* providing consecutively succinct descriptions for dynamic scenes, while presenting longer and detailed ones for stable settings. *(iii)* WorldScribe is adaptive to sound contexts, *e.g.,* increasing volume in noisy environments, or pausing when conversations start. Powered by a suite of vision, language, and sound recognition models, WorldScribe introduces a description generation pipeline that balances the tradeoffs between their richness and latency to support real-time use. The design of WorldScribe is informed by prior work on providing visual descriptions and a formative study with blind participants. Our user study and subsequent pipeline evaluation show that WorldScribe can provide real-time and fairly accurate visual descriptions to facilitate environment understanding that is adaptive and customized to users' contexts. Finally, we discuss the implications and further steps toward making live visual descriptions more context-aware and humanized.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Accessibility systems and tools**.

KEYWORDS

Visual descriptions, blind, visually impaired, assistive technology, accessibility, context-aware, customization, LLM, real world, sound

ACM Reference Format:
Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. 2024. WorldScribe: Towards Context-Aware Live Visual Descriptions . In *The 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24), October 13–16, 2024, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3654777.3676375

UIST '24, October 13–16, 2024, Pittsburgh, PA, USA

ACM ISBN 979-8-4007-0628-8/24/10
https://doi.org/10.1145/3654777.3676375

1 INTRODUCTION

Automated live visual descriptions can help blind or visually impaired (BVI) people understand their surroundings autonomously and independently. Imagine Sarah, who is blind, is exploring the zoo with her 2-year-old toddler. As they walk around the African grassland section, live visual descriptions provide rich information about the texture of the terrain animals are resting on and occasionally notify her about the movement of the zebras and rhinos. They join a giraffe feeding tour, and live visual descriptions narrate when a couple of giraffes reach out near her toddler. She feeds the lettuce leaves in her hand to them and snaps a nice photo. Such contextual visual descriptions can supplement their environmental understanding and support a range of ever-changing scenarios.

However, providing rich and contextual descriptions has been a long-standing challenge in accessibility. Researchers have explored ways to provide BVI individuals with visual descriptions across various visual media. For example, traditional AI captioning for digital media (*e.g.*, images, videos) offers basic information but often misses the nuanced details BVI users need in varied contexts [44, 74]. While human-powered [19, 38, 60, 61] or human-AI hybrid techniques [24, 32, 64] deliver more detailed descriptions asynchronously for digital media, they fall short in real-world scenarios that require descriptions to be timely and pertinent to the user's context. As a result, existing solutions in describing the real world have been limited to leveraging remote sighted assistance (RSA) to access BVI users' live camera streams and describe the visuals, such as Chorus:View [48] with crowd workers, BeMyEyes [2] with volunteers, and Aira [1] with trained agents. However, these human services can be extremely costly, not always available, and potentially raise privacy concerns.

The advent of vision-language models (VLMs) and large language models (LLMs) makes it possible to provide automated visual descriptions without human assistance. Off-the-shelf tools, such as SeeingAI [11], EnvisionAI [4] or ImageExplorer [49], enable BVI people to upload an image and receive detailed descriptions. However, the asynchronous and one-size-fits-all nature of the produced descriptions makes it difficult to adapt these tools to dynamic real-world scenarios. Providing seamless real-time automated descriptions is further challenging when considering user needs and contexts [44, 74]. For instance, BVI people have individual preferences [74] (*e.g.*, different visual experiences and familiarity with the environment), and the rich visual contexts may influence information priority depending on users' needs (*e.g.*, walking on the street, or visiting a museum). Furthermore, real-world sounds could hinder the perception of spoken descriptions [23]. Therefore, when providing live visual descriptions in the real world, it is crucial to collectively consider *user intent, visual and sound contexts* (which we will refer to as the user's context throughout the paper).

Informed by prior works on providing visual descriptions and a formative study with five BVI people, we identified design considerations for a system to provide live visual descriptions in the real world, such as providing descriptions with overview first then adaptive details on the fly, prioritizing descriptions based on semantic relevance, and enabling customizability based on varied user needs. We then develop *WorldScribe*, a system that generates automated visual descriptions that are adaptive to the users' contexts

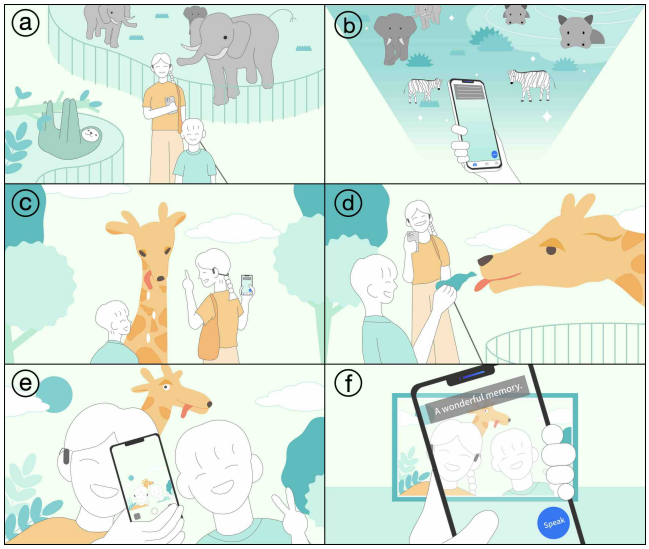


Figure 2: (a) Sarah is exploring the zoo with her toddler using WorldScribe, (b) which describes surroundings to her. (c, d) They join a giraffe feeding tour, live visual descriptions narrate when giraffes reach out near her toddler, who feeds the lettuce leaves to them and (e, f) snaps a nice photo.

in real-time, and in the real world. First, WorldScribe is adaptive to the user's intent, *e.g.*, prioritizing the most pertinent descriptions based on semantic relevance, and visual attributes based on user customizations. Second, WorldScribe is adaptive to visual contexts; for instance, it provides consecutively succinct descriptions for dynamic visual scenes while it presents longer and more detailed ones for stable settings. Third, WorldScribe is also adaptive to sound contexts, *e.g.*, increasing description volume in noisy environments, or pausing when conversations start.

Powered by a suite of vision, language, and sound recognition models, WorldScribe introduces a description generation pipeline with different VLMs that balances the tradeoffs between their richness and latency to support real-time usage (Figure 1). WorldScribe dynamically assigns prompts to VLMs encoded with user customizations on their information needs and in-the-moment visual contexts (*e.g.*, busy or static), and prioritizes descriptions based on the user intent and the proximity of the described content to the user. WorldScribe also keeps the spoken descriptions up-to-date by examining the object compositions and similarity between the VLM-referred and current camera frames and the changes in user orientations.

Our pipeline evaluation shows that WorldScribe can provide fairly accurate visual descriptions, cover important information, and prioritize descriptions based on users' intent and proximity. Furthermore, our user study with six BVI participants demonstrates that WorldScribe enables effective environment understanding that is adaptive and customizable to users' contexts. However, there is still a gap in making AI-generated descriptions humanized, user-centric, and context-aware, which we discuss and provide implications for future work. WorldScribe represents an important step towards solving this long-standing accessibility challenge, and its technical approach may find applications broadly for enhancing real-time visual assistance to promote real-world and digital media accessibility.

Table 1: Overview of research or commercial apps for describing visual media. AD denotes audio description.

App Category	Application	Description type	Enabling source	Real time	Audio presentation	Customization options
Navigation	BlindSquare [3]	Audio direction	Map	✓	Spatial audio	Landmarks
	SoundScape [9]	Audio direction	Map	✓	Spatial audio	Landmarks, route
	NavCog [16]	Audio direction	Map	✓		Landmarks
Image Understanding	Seeing AI [11]	Image description	AI			Short or detailed content
	Envision AI [4]	Image description	AI			Short or detailed content
	ImageExplorer [49]	Image description	AI			Number of info layers, accuracy, specific object info
Video Understanding	YouDescribe [38]	Audio description	Human		inline ADs	
	Rescribe [64]	Audio description	Human-AI		inline ADs	
	OmniScribe [24]	Audio description	Human-AI		Spatial Audio, inline and extended ADs	
	ShortScribe [77]	Audio description	AI			
	InfoBot [72]	Video VQA	AI			Request to AI
Remote Sighted Assistance	Aira [1]	Verbal Guidance	Human	✓		Request to sighted agents
	BeMyEyes [2]	Verbal Guidance	Human	✓		Request to sighted agents
Visual Question Answering	VizWiz [19]	Image description	Human			Request to crowd workers
	BeMyAI [8]	Image description	AI			Request to AI
Real-world Visual Understanding	**WorldScribe (this work)**	**Live visual description**	**AI**	✓	**Auto volume adjustment or pause**	**User intents, objects, visual attributes, audio presentations, verbosity**

2 RELATED WORK

Our work builds upon prior work to provide BVI people with descriptions for accessing digital media and the real world, in order to fulfill their diverse needs. We describe our motivation and insights from previous literature below.

2.1 Descriptions for Digital Visual Media

To understand digital visual media, BVI people typically rely on textual descriptions. World Wide Web Consortium has established Web Content Accessibility Guidelines for creators to add proper captions to images [82] and audio descriptions to videos [62, 80, 81] for BVI people to receive equal information as sighted people. Several platforms allowed BVI people to request descriptions for images and videos that lack accessible visual descriptions from volunteer describers, such as YouDescribe [38] and VizWiz [19]. Despite the availability of these resources, learning those guidelines and providing good descriptions remain difficult [60]; the scarcity of human resources also makes it hard to address the high volume of requests from BVI people [31], who may have different information needs based on their access contexts [73, 74].

To address these challenges, semi-automatic AI systems have been developed to streamline the description authoring process, such as generating initial image captions [32] or audio descriptions [64, 72, 85]. Although these systems reduce laborious tasks, they still require human effort to make one-size-fits-all descriptions usable. Recently, VLM-powered systems can generate high-quality audio descriptions comparable to human describers [77] and allow BVI people to query visual details interactively [8, 25, 37, 72]. However, the asynchronous and one-size-fits-all nature of descriptions makes it difficult to adapt these tools to dynamic real-world scenarios. In response, this work aims to provide live contextual descriptions for BVI users by understanding their intent and visual contexts. This is achieved through a description generation pipeline with dynamic prompt assignments based on user contexts, and different VLMs that balance the tradeoffs between their richness and latency to achieve real-time purposes.

2.2 Descriptions for Real-World Accessibility

Accessing the real world through descriptions enhances BVI individuals' independence in various tasks, such as object identification [17, 36, 58], line following [46], and navigation [3, 9, 16, 30, 42, 43, 45, 47]. Navigation is especially important but challenging in unfamiliar settings, which demands extensive environmental understanding [30, 45], such as recognizing intersections [45, 47], signs [15, 45, 83], and traffic light statuses [26, 75]. While these systems offered critical task-specific guidance (*e.g.,* audio directions), they lacked visual descriptions for ever-changing surroundings. Tools, such as SeeingAI [11], ImageExplorer [49], and Envision AI [8], enabled BVI users to snap a photo and receive comprehensive visual descriptions within seconds, while BeMyAI [8] allowed BVI people to access details through turn-by-turn interactions (Table 1). Yet, their utility falls short in rapidly changing visual scenes that require live and continuous descriptions.

An alternative for understanding the dynamic real world is through human assistance, such as RSA, which connects BVI users with sighted agents via video calls to fulfill requests through verbal

guidance. However, conveying visual information in this way can be challenging and cognitively demanding [41, 51], where agents were under pressure to understand and effectively communicate key details [50–52], or they needed to tailor the level of detail to the user's needs [35, 51]. Moreover, RSA services could raise privacy concerns [20], incur high costs (*e.g.,* $65 for 20 monthly minutes with professional services, such as Aira [1]), and volunteer-based options, such as BeMyEyes [2], may not always be available. In this work, we aim to make the real world accessible to BVI people through automated live visual descriptions in order to enhance their environmental understanding beyond navigation instructions.

2.3 Fulfilling Diverse Needs of BVI People

Creating high-quality descriptions that meet BVI people's diverse information needs is challenging for different visual media. Prior research found that current one-size-fits-all approaches to image descriptions are insufficient for providing necessary details for meaningful interaction [44, 57, 67, 68, 73, 74]. Stangl et al. [73, 74] identified that the source of an image and the user's information goal impacted their information wants, and proposed universal terms (*e.g.,* having identity or names for describing people) as minimum viable content, with other terms provided on demand based on users' contexts (*e.g.,* person's height, hair color, etc.). Guidelines also indicated that the inclusion of certain visual details should be context-based, such as having general information for first access or having details (*e.g.,* color, orientations of objects) when gauging one's understanding of certain image content [13, 65].

The varied information needs of BVI people were also found when accessing different types of videos [39, 59, 72], such as different preferences on the audio description content (*e.g.,* object details, settings), and output modalities (*e.g.,* audio, tactile). For 360-degree videos, which offer richer visual information and immersion, BVI people also have varied preferences for linguistic aspects (*e.g.,* level of details, describing from first- or second-person view), or audio presentations (*e.g.,* spatial audio) [24, 40]. Thus, the way of presenting visual information and determining its richness is crucial and depends on the user's needs and context.

These findings of users' varied preferences for digital media also extended to the real world. Herskovitz et al. [33] identified the needs of BVI individuals, who often customize assistive technologies for daily activities by combining mobile apps for different visual tasks, such as obtaining clock directions from Compass or descriptions from BlindSquare, or filtering visual information (*e.g.,* text or colors). Overall, prior work in both the digital and real worlds has highlighted the importance of customization for adapting to diverse user preferences and contexts. In this work, we explore live visual descriptions that are context-aware, by enabling customization options on the description content (*e.g.,* level of details, visual attributes), and audio presentation (*e.g.,* pausing or increasing volume) to tailor to the diverse needs of BVI people.

3 FORMATIVE STUDY

We conducted a formative study to identify design requirements for a system providing live visual descriptions in the real world. We conducted semi-structured interviews with 5 BVI participants (Table 2) to gather feedback on their needs through several potential scenarios. We developed scenarios considering several aspects, including user intent, familiarity with environments, visual complexity, and sound contexts. We developed scenarios considering user intent, familiarity with environments, visual complexity, and sound contexts, as these aspects were identified in previous works as influencing information needs [16, 23, 30, 44, 45, 73, 74]. Participants were asked to imagine using a future live description system capable of capturing visuals and sounds in their surroundings and brainstorm their needs and potential solutions. From these discussions, we extracted key insights reflecting participants' needs and strategies, which we used in the design of WorldScribe.

3.1 Design Considerations

We reported design considerations derived from our participants:

D1 - Overview first, adaptive details on the fly. Participants emphasized the need for descriptions with proper levels of granularity, depending on their context. They preferred immediate and succinct information when several important events occurred simultaneously (*e.g.,* multiple barriers and directions during navigation), and longer and detailed descriptions when there was no time pressure (*e.g.,* understanding artwork in a museum). When searching for something, they wanted an overview of the space, including landmarks and spatial locations, followed by more details as they approached the target or encountered similar items requiring differentiation. This approach aligns with the *"Overview first, zoom and filter, then details-on-demand."* by Shneiderman [70]. Therefore, our solution should provide the proper level of information and delve into details when users express interest.

D2 - Prioritize descriptions based on semantic relevance. Participants mentioned strategies for filtering and prioritizing complex visual information. The most commonly noted strategy was to prioritize descriptions relevant to their goals of context, such as road signs or barriers during navigation, available stores and offerings during meal times, etc. They also emphasized that nearby objects are more important for safety and should be prioritized over distant information. Our system should present information most relevant to the user's goals and proximity to ensure timely and practical use.

D3 - Enable customizability for varied user needs. Similar to prior work in Section 2.3, we observed varied individual preferences from our participants. They expressed different information needs depending on the context. For example, descriptions should consider their mobility, such as providing information about objects out of their cane reach (*e.g.,* hanging lights, cars with high ground clearance) or focusing on dynamic obstacles but not static ones in their familiar environments. Participants also noted that sound context influences the consumption of descriptions, with some suggesting pausing or increasing the volume in noisy environments. Some preferred manual control over these options, while others pointed out that the automatic approach would benefit urgent or busy scenarios, such as navigation or if the description content is crucial. Based on these findings, WorldScribe should offer customizable options for description content and presentation to meet diverse user needs.

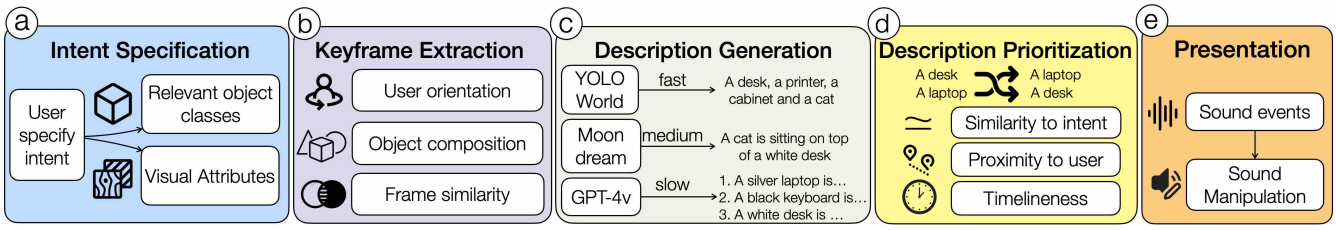


Figure 3: WorldScribe system architecture. (a) The user first specifies their intent through speech and WorldScribe decomposes it into specific visual attributes and relevant objects. (b) WorldScribe extracts keyframes based on user orientation, object compositions, and frame similarity. (c) Next, it generates candidate descriptions with a suite of visual and language models. (d) WorldScribe then prioritizes the descriptions based on the user's intent, proximity to the user, and relevance to the current visual context. (e) Finally, it detects environmental sounds and manipulates the presentation of the descriptions accordingly.

4 WORLDSCRIBE

WorldScribe is a system that provides live visual descriptions for BVI people to facilitate their environmental understanding. BVI users can specify their intent, general or specific, which will be decomposed by WorldScribe into specific visual attributes and objects of relevance (Intent Specification Layer). Then, WorldScribe extracts key frames from the camera video stream (Keyframe Extraction Layer), and employs a suite of vision and language models to generate rich descriptions (Description Generation Layer). The descriptions are prioritized based on users' intent, proximity, and timeliness (Prioritization Layer), and presented with audio manipulations based on sound context (Presentation Layer).

4.1 Scenario Walkthrough of WorldScribe

Here, we illustrate WorldScribe in an everyday scenario, taking Brook as the main character, a graduate student who is blind.

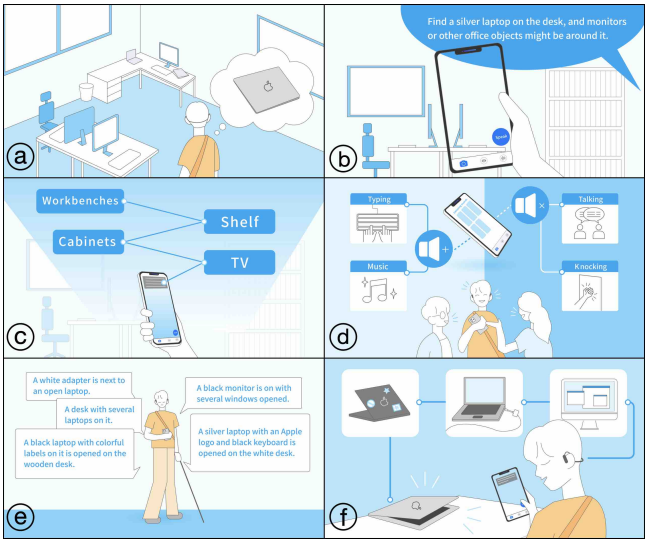


Figure 4: (a) Brook is looking for a silver laptop using WorldScribe in the lab by first (b) specifying his intent. (c) As he moves quickly, WorldScribe reads out names of fixtures, and (d) pauses or increases its volume based on environmental sounds. When approaching his seat and Brook stops to scan, (e) WorldScribe provides verbose descriptions when the visual scene is relevant to his intent, (f) allowing him to follow the cues and find the laptop.

Brook just finished his advising meeting, and he wants to find a lab laptop with powerful computational resources to proceed with his project. The lab is filled with large items like TVs, workbenches with electronics, rows of seats with monitors, personal items, cabinets, and garbage bins, with their layout changing daily based on activities. The only cues Brook has from his labmate are that the laptop is silver (Figure 4a) and located around the student seats amid office or personal objects (*e.g.,* monitors, adapters, backpacks).

Arriving at the spacious lab, Brook specifies his goal by talking to WorldScribe (Figure 4b): *"find a silver laptop on the desk, and monitors or other office objects might be around it."* He then aims his smartphone camera forward with WorldScribe on. Along the way, WorldScribe provides concise and timely descriptions of objects not directly related to his goal, where several fixtures, such as "TVs", "cabinets", "workbenchs", help him orient himself (Figure 4c).

As he approaches his seat, surrounded by relevant items like monitors, chargers, cables, and adapters, WorldScribe becomes more verbose (Figure 4e), providing descriptions, such as "A black monitor is on with several windows opened", "A white adapter is next to an open laptop", and "A desk with several laptops on it". These help Brook ascertain that the laptop he seeks is nearby.

Brook then scans his surroundings slowly with WorldScribe, listening to the detailed descriptions with objects' visual attributes to help distinguish the laptop he is looking for (Figure 4f), such as "A black laptop with colorful labels on it is opened on the wooden desk", and finally, "A silver laptop with an Apple logo and black keyboard is opened on the white desk".

In the lab, people talk, type on keyboards, or use power tools, generating various noises. Brook is accustomed to these sounds and has customized WorldScribe to accommodate different interference (Figure 4d). For instance, when his labmates talk, Brook wants to join the conversation, so WorldScribe immediately pauses descriptions to let him listen and resumes when they stop talking. Also, if a cellphone or clock rings suddenly, WorldScribe automatically increases the description's volume to ensure he can hear it clearly.

After working for a while, Brook takes a break on the building's balcony and uses WorldScribe to explore his surroundings (Figure 5a). The balcony has several plants, benches on the sides, and coffee tables. When Brook aims the camera at the sky (Figure 5b), WorldScribe describes "The sky is cloudless, and the sunlight is casting a warm glow over the bench". Brook then turns to the plants (Figure 5c), wondering if they have begun to germinate in this early spring period, and WorldScribe describes: "Several young plant seedlings in a stage of early growth." The beautiful view revitalizes

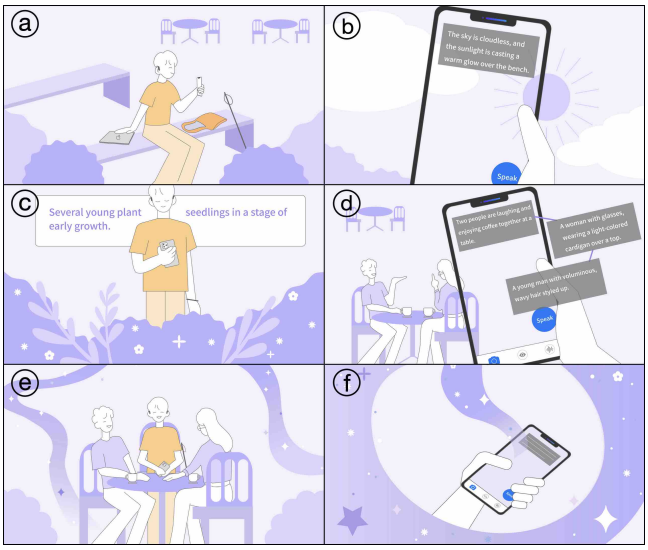


Figure 5: (a) Brook takes a break on the balcony and uses WorldScribe to explore his surroundings. (b) Through the live visual descriptions, he knows the sky is sunny, (c) plants are growing, and also notices (d) his friends are here. (e) He then joins them and has a delightful tea time. (f) WorldScribe facilitate the understanding and access of his surroundings, and make his day.

his weary mind from work. Later, Brook recognizes familiar voices coming from a coffee table. Turning towards the sound (Figure 5d), WorldScribe describes: "Two people are laughing and enjoying coffee together at a table". It then provides detailed descriptions such as "A woman with glasses, wearing a light-colored cardigan over a top" and "A young man with voluminous, wavy hair styled up". With this detailed visual information and the voices, Brook identifies them, joins their conversation, and enjoys a delightful tea time with them (Figure 5e & f).

4.2 User Interface

WorldScribe has a mobile user interface that takes the user's camera stream, environmental sounds, and user customizations as inputs for generating descriptions (Figure 6). The interface includes three pages: *(i)* a main page with a camera streaming view and speech interface, *(ii)* a customization page for visual information, and *(iii)* a page for customizing audio presentation. Users can open the camera on the main page and use speech to indicate their intent (Figure 6a), such as "find my cat, short hair and pale brown." To customize visual attributes of their interest, users can also use the speech on the main page, such as "I am interested in color" or "Tell me everything is pale brown", or manually toggle options on or off (Figure 6b). Similarly, users can verbally change the presentation of descriptions, such as "Pause when someone talks" or "Increase volume during ringtone", or manually select the options through the picker (Figure 6c).

4.3 Intent Specification Layer

In this layer, WorldScribe aims to obtain the user's intent and needs on visual information to enable customizability (**D3**). Users specify their intent on the mobile interface, and WorldScribe will classify

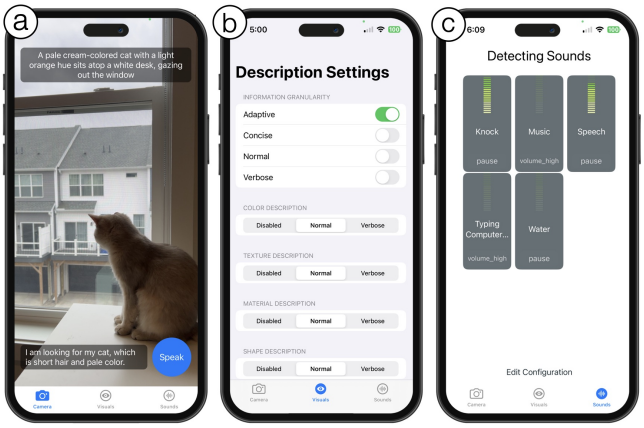


Figure 6: WorldScribe user interface. (a) The user can specify their intent and needs regarding visual attributes or audio presentation through speech input. (b) Besides speech, they can manually select options for richness and other visual attributes. (c) They can also configure pauses or increase the volume of descriptions if certain sound events are detected.

it as *general* or *specific* and generate relevant object classes and visual attributes by prompting GPT-4 [5]. If the intent is *general* or *not specified* (Figure 7a&b), WorldScribe takes object classes from existing datasets (*e.g.,* COCO [54], Object365 [69]). If the intent is *specific* (*e.g.,* "Find a silver laptop on the desk, and other office objects might be around it"), WorldScribe prompts GPT-4 [5] to generate a list of relevant objects (*e.g.,* "[laptop, desk, monitor, ...]") and adjust visual attributes of interest (*e.g., color* and *spatial information*) to *Verbose*. WorldScribe then uses YOLO World [27] and ByteTrack [86] to support open vocabulary object recognition and tracking. Users can further refine the generated object classes and other visual attributes through speech (Figure 6a) or manually on the customization page (Figure 6b).

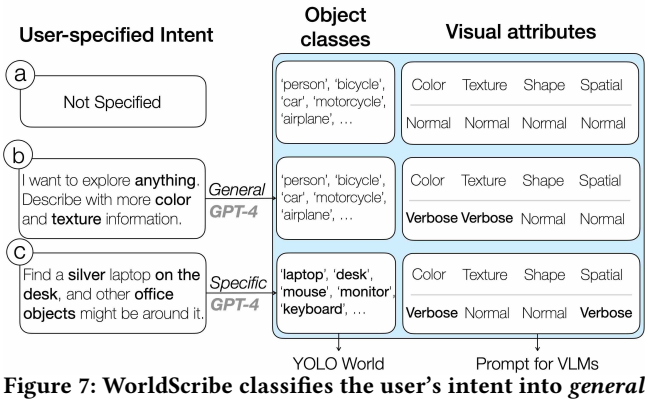


Figure 7: WorldScribe classifies the user's intent into *general* or *specific*, and generates relevant object classes and visual attributes by prompting GPT-4 [5]. (a) By default, WorldScribe uses classes from established datasets (*e.g.,* COCO [54], Object365 [69]), and sets visual attributes to *Normal*. (b) If the intent is classified as *general*, WorldScribe adjusts the visual attributes of interest to *Verbose*. (c) If the intent is classified as *specific*, WorldScribe generates relevant object classes and sets visual attributes to *Verbose*.

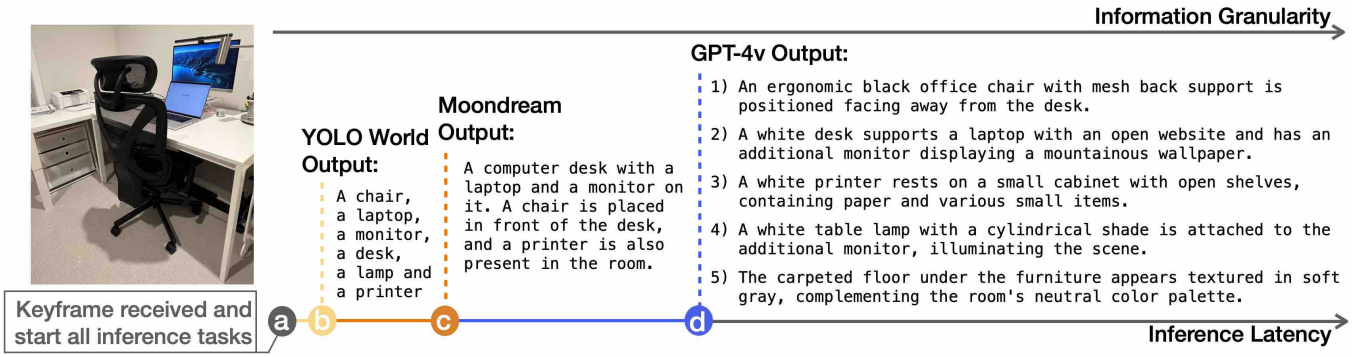


Figure 8: WorldScribe description generation pipeline with different inference latency and granularity. (a) Upon receiving a keyframe, WorldScribe starts all visual description tasks. (b) First, YOLO World [27] identifies objects as word-level labels in real-time (.1s). (c) Second, Moondream [10] generates short descriptions with objects and spatial relationships, with a small delay (~3s). (d) Finally, GPT-4v [7] provides detailed descriptions with visual attributes, with a longer delay (~9s). The estimated inference time in each model was calculated based on our computing platforms and log data in our user evaluation.

4.4 Keyframe Extraction Layer

In this layer, WorldScribe aims to identify keyframes that indicate salient visual changes or user interests in the visual scene. To achieve this, our approach uses two methods: camera orientation and visual similarity. First, WorldScribe monitors changes in the camera's orientation using the phone's inertial measurement unit (IMU). A keyframe is selected whenever the camera's orientation shifts by at least 30 degrees (one clock unit) from the previous keyframe, indicating a possible turn into a new visual scene.

Second, WorldScribe determines a keyframe by analyzing visual changes across frames. To minimize detection errors, such as misassigned object classes or IDs, we assess the consistency of object composition over n consecutive frames. In each i^{th} frame, a detected object is represented as (ID_i, C_i), where ID_i is the object's index, and C_i is the class. Thus, all objects in the i^{th} frame are represented as $O_i = (ID_i, C_i)$. A keyframe is identified if the object composition remains consistent across n frames, denoted as $O_i = O_{i+1} = ... = O_{i+n-1} \neq \emptyset$. The $(i+n-1)^{\text{th}}$ frame is then taken as the keyframe. Furthermore, to determine if the user is interested in a visual scene and requires details (conforming to **D1**), we check the m consecutive keyframes with the same composition and use the latest keyframe to prompt VLMs for detailed descriptions.

In scenarios where the object compositions across n consecutive frames are empty, denoted by $O_i = O_{i+1} = ... = O_{i+n-1} = \emptyset$, it suggests that the predefined object classes may not cover the objects in the scene, resulting in false negatives. Therefore, we still take the $(i+n-1)^{\text{th}}$ frame as the keyframe. To eliminate genuinely empty scenes (*e.g.,* aiming at a plain white wall), we measure the similarity between the candidate frame and the previous keyframe. We calculate the cosine similarity (cos_sim) between the image feature vectors of the two frames, extracted from the FC2 layer of the VGG16 [71]. If cos_sim is lower than a threshold $thres$, we count the frame as a keyframe. Furthermore, we observed situations where object compositions differ across consecutive n frames, denoted by $O_i \neq O_{i+1} \neq ... \neq O_{i+n-1} \neq \emptyset$. These changes often indicate camera drifting or objects moving in and out of view. In such cases, we check every $2n$ frame, selecting the $(i+2n-1)^{\text{th}}$ frame as a keyframe if the condition is satisfied. In our implementation, we empirically set $n = 5$, $m = 3$, and $thres = 0.6$.

4.5 Description Generation Layer

In this layer, WorldScribe aims to generate descriptions with adaptive details to the user's intent and visual contexts. To achieve this, WorldScribe leverages a suite of VLMs that balances the trade-offs between their richness and latency to support real-time usage (Figure 8).

To provide overview first and details on the fly (**D1**), WorldScribe recognizes objects by YOLO World [27] and structures its results into short phrases, *e.g.,* "A chair, a laptop, a monitor, ...", allowing it to provide an overview of objects in the visual scene in real time (Figure 8b). Then, WorldScribe describes objects and their spatial relationship by prompting Moondream [10] (Figure 8c), a compact vision language model, that can achieve a decent performance in terms of latency and accuracy on this information based on our observation. Finally, WorldScribe prompts GPT-4v [7] to generate descriptions of different levels of details based on user contexts (**D1**). It offers three levels of detail: *Verbose*, *Normal*, and *Concise*, associated with prompts specifying *e.g.,* "over 15 words", "at least 10 words", and "less than 5 words", respectively. WorldScribe dynamically adjusts these length constraints based on visual complexity. For instance, it becomes *Verbose* if the visual scene is focused on for a long period, indicating user interest, with consecutive keyframes identified (See Section 4.4). It becomes *Concise* when multiple objects of interest are detected in consecutive keyframes to ensure timely coverage; otherwise, it remains *Normal* by default. Along with the recognized visual attributes of interest, WorldScribe dynamically creates prompts to suit users' intent (**D3**):

> "You are a helpful visual describer, who can see and describe for BVI people. You will not mention this is an image; just describe it, and also don't mention camera blur or motion. Please ensure you provide these adjectives to enrich the descriptions *[desired visual attributes (e.g., color, texture, shape, spatial), with examples adjectives]*, you should describe each object with ONLY ONE sentence at maximum. Don't use 'it' to refer to an object. Most importantly, each sentence should be *[sentence length constraint e.g., verbose, normal or concise]*."

Each short phrase from YOLO World [27], general description from Moondream [10], and detailed description from GPT-4v [7] are stored as a packet in the description buffer, and WorldScribe selects the most relevant and up-to-date one based on the user's contexts, which we describe next.

4.6 Description Prioritization Layer

In this layer, WorldScribe aims to select a description based on the match to the user's intent, proximity to the user, and relevance to the current visual context (**D2**).

4.6.1 Sorting GPT-4v detailed descriptions by semantic relevance. Descriptions from YOLO World [27] and Moondream [10] provide an overview and general information, which should always be prioritized to help the user construct an initial understanding of the new visual scene (**D1**). In contrast, detailed descriptions from GPT-4v [7] may contain information irrelevant to the user's intent. Therefore, WorldScribe ranks the descriptions generated by GPT-4v [7] based on their relevance to the user's intent and the proximity of the described content to the user; the nearer, the earlier to describe.

To achieve this, we first get a set of descriptions from GPT-4v [7] $S = \{s_i\}$. For a description s_i, we compute the sentence similarity score SIM(s_i) to the user's intent, and a depth score $Depth(s_i)$. To calculate the depth score, we extract the subject and its descriptors from a description (*e.g.,* "Two cylindrical trash receptacles" in Figure 9), locate each in the frame using CLIPSeg [56] (Figure 9b), and crop the salient area on the depth map generated by Depth Anything [84]. We then compute the average depth $Depth(s_i)$ of the cropped area for each description s_i.

To sort the descriptions in S, We divide them into two sets $S_a = s_i | SIM(s_i) >=$ Threshold, and $S_b = s_j | SIM(s_j) <$ Threshold.

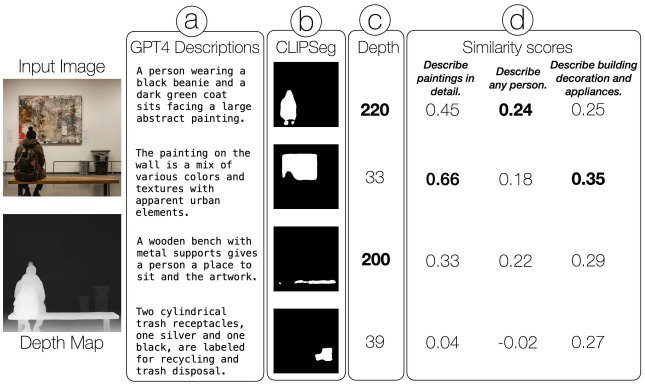


Figure 9: WorldScribe pipeline to prioritize descriptions based on semantic information. (a) Given an input keyframe image, GPT-4v [7] generates descriptions on individual objects. (b) CLIPSeg [56] generates the cropped image content based on each description. (c) The average depth value of the CLIPSeg-cropped area was computed for each described content with the depth map generated by Depth Anything [84]; the higher the depth map value, the nearer to the user. (d) The similarity score was also computed between each description and the user's specified intent. Finally, WorldScribe prioritize GPT-4v [7] generated descriptions based on semantic relevance to the user's intent, then proximity to the user (~1s).

We then sort the descriptions in S_a as a sorted sequence $A = (s_0, s_1, ..., s_{n-1})$ based on the similarity score such that $SIM(s_i) >= SIM(s_{i+1})$. We also sort sentence in S_b as $B = (s_n, s_{n+1}, ...s_{n+m-1})$ based on the depth score such that $Depth(s_j) >= Depth(s_{j+1})$. Finally, we concatenate the two sequences into final one, $A^\frown B = (s_0, ..., s_{n-1}, s_n, ..., s_{n+m-1})$. This approach ensures the initial descriptions are highly relevant to the user's intent, regardless of proximity. The remaining descriptions are sequenced by their proximity, with nearer elements described sooner.

4.6.2 Selecting up-to-date description based on current context. In reality, users may move or turn frequently, and the environment can change significantly, leading to frequent visual changes and generating multiple keyframes and VLM requests in a short time. As VLM inference times vary, some descriptions may become outdated if they take too long to process and will not be useful to the user's current visual context. Thus, we need to select descriptions that best represent the current visual scene. We consider four criteria: camera orientation, object compositions, frame similarity, and similarity to previous descriptions. A description is selected if it satisfies any of these criteria. First, we check if the candidate description's referenced object composition matches the current scene. Second, we compare the user's orientation from the description's referenced frame to the current orientation. Third, we compare the frame similarity between the description's referenced frame and the current frame using feature vectors extracted through VGG16 [71]. Descriptions similar to preceding spoken descriptions are skipped, and the description buffer is renewed when a description is omitted. Descriptions generated by GPT-4v [7] are prioritized, followed by those from Moondream [10] and YOLO World [27].

4.7 Presentation Layer

In this layer, WorldScribe aims to make descriptions audibly perceivable to the user by considering users' sound context. To achieve this, WorldScribe runs a sound detection module in the background and automatically manipulates the presentation of the descriptions accordingly. Based on our formative study, WorldScribe enables two audio manipulations on descriptions for the user to better perceive the description content in the noisy environment: *(i) Pausing* and *(ii) Increasing volume.* Users can find sound events that interest them and customize the corresponding manipulations on descriptions in WorldScribe app (Figure 6c).

4.8 Implementation Details

WorldScribe servers included a local server running on a Macbook M1 Max and another remote server with two embedded Nvidia GeForce RTX 4090. WorldScribe mobile app was built on an iPhone 12 Pro and streamed the camera frames to the local server through a Socket connection. YOLO World [27] and ByteTrack [86] were run on the local server with 5 frames per second (FPS) along with other algorithms, while other models in description generation and prioritization pipeline, including Moondream [10], Depth Anything [84] and CLIPSeg [56] were run on the remote server for each keyframe, as well as the pre-trained model "all-MiniLM-L6-v2" for sentence similarity from an open-sourced implementation on Huggingface. Overall, based on the data collected in our user study (Section 5), WorldScribe achieved an overall latency of 1.44s, and

Table 2: Participant demographics information. Participants in our formative study were marked as F1-F5. Participants in our user evaluation were marked as P1-P6.

ID	Age	Gender	Self-Reported Visual Ability	Assistive App Use	Self-Defined Goal in User Evaluation
F1	34	Male	Blind, since birth. Light perception.	BeMyEyes	N/A
F2	25	Male	Blind, later in life. Light perception.	BeMyEyes, ENVision, TaptapSee, VoiceVISTA	N/A
F3	23	Female	Blind, later in life. Light perception.	None	N/A
F4	35	Male	Blind, later in life. Light perception.	BeMyEyes and Aira	N/A
F5	24	Male	Blind, since birth. Light perception.	BeMyEyes	N/A
P1	62	Male	Low Vision, can't pick up details, using magnifiers.	None	Describe things on the wall.
P2	53	Female	Blind, since birth.	SeeingAI, BeMyEyes	Describe posters and people in detail.
P3	60	Female	Low Vision, can't pick up details.	None	Describe paintings or pictures in detail.
P4	40	Male	Blind, since birth. Light perception.	SeeingAI, BeMyEyes and SoundScape	Describe any person.
P5	87	Female	Low Vision, can't pick up details.	None	Describe artworks or paintings.
P6	72	Female	Blind, since birth. Light perception.	SeeingAI, BeMyEyes, BeMyAI, Aira	Describe things in general with more color and texture information.

each component took an average: YOLO World 0.1s, Moondream 2.87s, GPT-4v 8.78s, and prioritization pipeline 0.83s. For sound recognition, we used Apple's Sound Analysis example repository [12], which provides a visualization interface (Figure 6c) and can identify over 300 sounds.

5 USER EVALUATION

We conducted a user evaluation with six BVI participants, where they used WorldScribe in three different contexts. This study aimed to explore **RQ1:** How do users perceive WorldScribe descriptions in various contexts? and **RQ2:** What are the gaps between WorldScribe descriptions and users' expectations? We detail our study method and results below.

5.1 Participants

We recruited six BVI participants (2 Male and 4 Female) through public recruitment posts on local associations of the blind. Participants aged from 40 to 87 (Avg. 62.3) and described their visual impairment as blind (N=3) or having residual vision (N=3). Some participants had prior experiences using RSA services and used AI-enabled services, such as BeMyEyes [2] or SeeingAI [11] in their daily lives (Table 2).

5.2 Study Sessions

We enacted three different scenarios: *(i)* specific intent, *(ii)* general intent, and *(iii)* user-defined intent. In each session, the descriptions were automatically paused if the speech was detected, including the conversation between participants and the experimenter, and the volume was automatically increased if a ringtone occurred.

Scenario with specific intent. The first scenario, similar to the walkthrough scenario (Section 4.1), happened in our lab space, which is furnished with glass walls, wall-mounted TVs, several work benches with electronics and equipment, several rows of seats with monitors and scattered personal items, and a small kitchen area with microwave, fridge, sink and a lot of cabinets and garbage bins at the corners. The user's intent is "find a silver laptop on the desk, and monitors or other office objects might be around it." This scenario was designed to encourage them to think about the descriptions they need for specific purposes and whether WorldScribe supplements or obscures their intent.

Scenario with general intent. This scenario happened on one of our building's floors, which has many common objects on the intricate hallways, such as poster stands, carts for construction, trash cans, desks, and sofas. On the wall or doors, there were several artworks, paintings, posters or emergency plans, and TVs. Random people were also walking in the hallway or meeting at public tables during the study. The intent of the scenario is "I am exploring a school building. Describe general information on the appliances and the building decorations." This scenario was designed to prompt users to think if WorldScribe descriptions support their understanding of the environment.

User-defined scenario. After experiencing the previous scenarios, participants were asked to develop their own defined scenarios. They can also customize their desired visual attributes in WorldScribe mobile app based on their needs. We then took participants to the place they wanted to explore near our experiment sites.

Limitations. Though we tried to create different real-world scenarios, our study was conducted within our local environment and buildings. This setting may not fully capture the diversity and complexity of real-world environments, potentially limiting the generalizability of our findings to other contexts.

5.3 Procedure

After providing informed consent, participants were introduced to WorldScribe and the functionalities they could customize, and experienced through each session. Participants opted to either hold the camera on their front or wear the lanyard smartphone mount we prepared. To facilitate the study progress and avoid fatigue, we kept each exploration for around ten minutes or until participants paused spontaneously. At the end of each session, we interviewed our participants about their experiences with WorldScribe. The study took about two hours, and participants were compensated $50 for their participation. This study was approved by IRB in our institution.

5.4 Measures and Analysis

We asked our participants to comment on their perceived accuracy and quality of descriptions, their confidence in WorldScribe descriptions, and several other open-ended questions. We recorded and transcribed the interviews and recorded all interactions with WorldScribe, which was also used for our pipeline evaluation (Section 6). Two researchers coded all qualitative interview feedback received in all sessions for further analysis via affinity diagramming.

5.5 Results

5.5.1 Perceived accuracy and skepticism towards the descriptions. **Participants perceived WorldScribe descriptions as accurate based on the contextual clues they ascertained but remained skeptical due to a few observed erroneous instances.** Participants generally commended WorldScribe for providing information otherwise unavailable in their everyday lives. They appreciated its constant descriptive capabilities, finding them useful for daily tasks such as grocery shopping, locating dropped items, and exploring the outdoors. Participants valued the real-time feedback and considered the descriptions accurate and responsive. For instance, some (P2, P4) tested the system by placing their hands in front of the camera and received immediate descriptions of their hands and accessories:

> *"I just wanted to test if it can describe the rings on my hand, it's like wow it did describe, and did a pretty good job and so responsive, so I think it's accurate for what it sees."* - P2

Despite acknowledging the accuracy and timeliness of WorldScribe's descriptions, participants expressed tentative skepticism about its practical use due to several factors. For example, occasional hallucinations, such as detecting motorcycles in the building lobby or bikes in the office space, impacted their confidence in WorldScribe's descriptions. Other instances where WorldScribe failed to mention essential information also led to doubts:

> *"I am not confident because I put my eyedrop in front of me to see if it would pick it up, but it did not, which is fine as I guess it is not programmed for that. But it will be very useful in this case."* - P1

However, some pointed out the walk-up-and-use study design made them unable to fully explore and get used to WorldScribe:

> *"Honestly, I remain conservative using it off the street tomorrow. But being used to the systems, I think if I had some time to get used to it, I could work with it."* - P4

But in general, they foresaw the promise and benefits WorldScribe can bring otherwise unavailable from existing apps, such as the real-time experiences and the adaptive level of details:

> *"It's closer to SeeingAI and BeMyEyes descriptive mode. Initially, it's like desk, chair, ... and become descriptive like a human, more color and context, if you are looking at things longer"* - P2

5.5.2 Perceived quality and customized visual information on the AI-generated descriptions. **Participants found WorldScribe descriptions useful with adaptive and customized visual information, but felt overwhelmed in some situations.** Participants noted several useful aspects of WorldScribe's descriptions. For instance, WorldScribe starts with an overview and provides details on the fly for each new visual scene. Hence, if a participant's quick movements or turns lead to a succession of new scenes, they receive an overview for each. In contrast, if they focus on the same scene for a while, they receive detailed descriptions. One participant noted:

> *"It's interesting when it just provided only a few words when I moved or turned, like a desk, a chair, a person, it's nice to know what is included in this space. And I got details if I faced that for a little longer. I like the switch between these low-level and high-level descriptors. If I'm in the moment that I should picture things myself; it'll just give me low-level descriptors. I appreciate that ... But if I'm looking for something and I'm trying to figure out where I'm near, or get some landmarks and stuff. Then I appreciate the higher-level stuff."* - P4

Aside from the level of granularity, participants also perceived the increased descriptions in their customized visual attributes. For instance, P2 made our system verbose on color and spatial relationships and remarked:

> *"This session did a better job at giving color descriptions. Also, it described more things like I said, location of things like in front of you, next to you, behind, you know, to your right or things of left."* - P2

Moreover, participants found WorldScribe offered unique and enriching experiences, *"[WorldScribe] used strong words, so beautiful."* (P3). They (P2, P4, P6) also pondered the balance between WorldScribe's detailed visual descriptions and their practical use, suggesting that the descriptions should be more colloquial to mimic a human describer who provides the minimum viable information:

> *I've never thought of a building being lit by tubes like a pattern or a line. It's all interesting for a blind person to have their eyes open to this stuff because I've never seen it before. It's all interesting information for me, but as far as practical use, I could get overwhelmed with it. Part of my brain loves it. Part of my brain is, Oh, I don't need it. So it's really interesting to be in this position. It really depends on the environment or your goal.* - P4

5.5.3 Alignment between users' mental model on the real world and what WorldScribe sees. **Participants desired the descriptions responsive to their physical reach, and the spatial information should center on them but not the image.** During study sessions, we found that camera aiming issues caused misalignment between what users thought and what WorldScribe described. For instance, some participants (P2, P3) held the smartphone in their hands to explore their surroundings. They were confused if WorldScribe missed describing the objects they touched, perceiving them to be something in their front captured by the camera. P2 frequently questioned during the study *"It did not describe when I touched it [laptop]. I wasn't sure if I was getting it within the camera."* P4, who thought the camera did not capture what he needed, wanted to change the camera mount *"Maybe next time I can use another camera headset that is over my eyes"*. Our subsequent video analysis found that their hands and the touched objects were beneath the smartphone and not captured within the frame.

Some participants also mentioned that WorldScribe could supplement their mobility with a white cane. For instance, they could confirm they had arrived at the exit upon hitting a chair, as WorldScribe had previously described the "exit" and "chair" together.:

> *I have my cane and am able to follow the directions to the exit. When I go over that area, you see where the stack of chairs is, as [WorldScribe] mentioned that before. So like it would say chairs when I hit it. Okay I'm going the right way.* - P1

However, what WorldScribe saw was based on a single image, limiting its understanding of the user's surroundings and creating erroneous spatial information. For example, the 'left' or 'right' was determined by the spatial relationships within an image rather than the user's point of view. Some participants observed this discrepancy but understood that WorldScribe focused on describing their surroundings instead of providing directions. Consequently, they proposed integrating WorldScribe with other apps to provide more comprehensive experiences, which we described in the next section.

5.5.4 Desires on more concrete information for practical use. **Participants desired concrete information for practical use in their daily lives, such as distance, directions, or pre-loaded map information.** Participants found WorldScribe useful for general environmental understanding. However, for high-stakes scenarios such as navigation, participants believed WorldScribe could supplement their experiences with existing navigation apps. For instance, navigation apps often provide general directions such as *"turn right at the next intersection"*, but it can be hard for BVI individuals to determine if they have reached the intersection or are at the crossroads. WorldScribe can assist this by describing their visual surroundings. Additionally, some asked for more concrete information along with visual descriptions, such as spatial relationships from their perspective, exact distances to objects, and their continuous updates:

> *" It would be like pre-loading the space. when I was looking for a classroom for approximately 3 doors or how many feet, and then you're gonna turn right. And then 4 doorways down on the right, that gives me a directional type of thing. When you're blind you have to do it by calling and checking the space, having that description and that context can help."* - P2

Additionally, they wanted to have more control over the descriptions and integrate with spatial audio, as one participant, who used SoundScape [9], mentioned:

> *"I could have [WorldScribe] running in the background. It'd be almost like a lucid dream if you had it on the spatial audio. Okay, that's over there. I want to know more about that. So I turn toward it. Then, it changes the environment to show me that I'm facing that exact thing. That'd be really beautiful."* - P4

Although WorldScribe was not designed for real-world navigation, these insights from users are invaluable for guiding our next steps by designing and making visual descriptions integrated into more practical and high-stake scenarios. We discuss the lessons learned from the study and potential improvements of live visual descriptions in Section 7.

6 PIPELINE EVALUATION

In this evaluation, we measured the accuracy, coverage of user-desired content, and description priority based on users' intent and proximity of described content. We collected data from our user evaluation, such as descriptions and their generative models, timestamps, referenced frames and prompts, customization settings, screen recordings, etc. These videos and frames were naturally captured by BVI participants in our user evaluation, resulting in camera motions or slants that impacted image quality, but it preserved the authenticity and relevance of our findings to real-world experiences. Each study session's video recording lasted around ten minutes, as described in Section 5.

6.1 Accuracy

We measured the accuracy of WorldScribe descriptions by inspecting the description content and the referenced frame.

Dataset & Analysis. In total, we collected 2,350 descriptions from our user study. The description sources included YOLO World [27], Moondream [10], and GPT-4v [7]. We inspected each description and considered descriptions incorrect if they could not be justified through their referenced camera frame.

Results. Overall, we found that 370 of 2,350 instances (15.74%) were incorrect in relation to the referenced camera frame. Specifically, 122 of the 638 descriptions (19.12%) from YOLO World were incorrect, 72 of the 549 (13.11%) descriptions from Moondream were incorrect, and 176 of the 1,157 (15.21%) descriptions from GPT-4v [7] were incorrect.

We observed several reasons for incorrect descriptions. For instance, when prompted to generate object classes based on users' intent for YOLO World [27], GPT-4 [5] sometimes generated classes that could not be identified in our study environment, such as *museum*, *exhibition*, and *classroom*. Additionally, the low image quality significantly impacted the accuracy of descriptions. For example, YOLO World [27] often mistook whiteboards, paper, walls, or illuminated monitor screens for other objects. Moondream [10], prompted to provide general descriptions, performed well in covering common objects and their spatial relationships but sometimes included hallucinated content. For GPT-4v [7], lighting conditions or capture angles affected the results. For example, a cabinet was mistaken for a washer and dryer when shot from the side, cluttered folding chairs were mistaken for motorcycles or bicycles, and a male with long hair was identified as female.

6.2 Coverage of User Intent

We measured if WorldScribe descriptions covered important content of the users' intent in a timely fashion.

Dataset & Analysis. We used the smartphone video recordings of the P1-P5 self-defined scenario due to their concrete intent. To determine whether the descriptions covered the essential content related to users' specified intent, we developed video codes to annotate the objects that should be described in the footage. One of the authors carefully reviewed each recording and annotated the objects relevant to users' intent. For example, we labeled items on the wall for P1's intent "Describe things on the wall" and labeled people for P4's intent "Describe any person." In each video, we observed that participants sometimes turned around or moved frequently, or

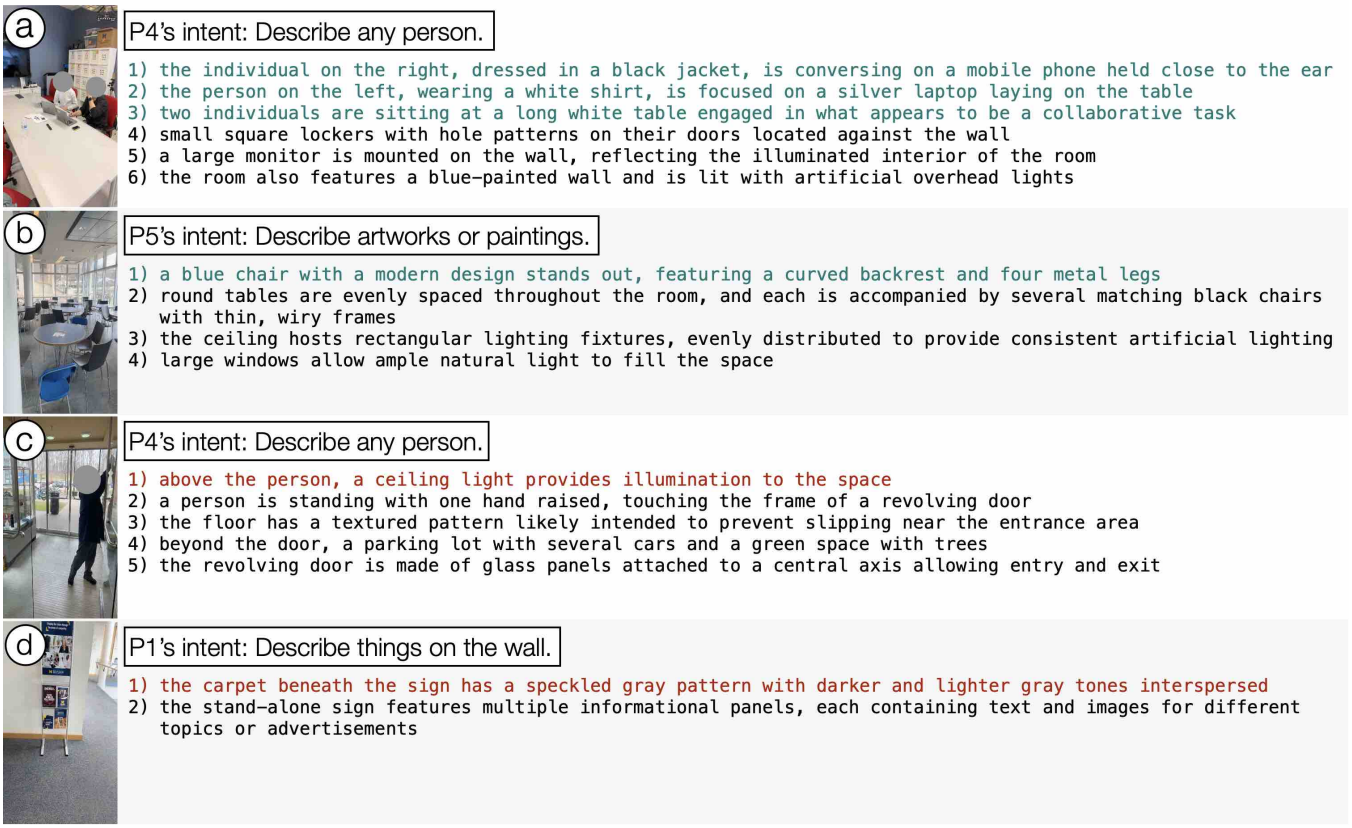


Figure 10: Example results of description priority. (a) The descriptions relevant to the user's intent were successfully prioritized. (b) If no subjects meet the user's intent, descriptions would be ordered based on the distance to the user. (c) The priority of similarity to the user's intent failed if a description involves the user's intent as supplement information such as spatial relationship. (d) The distance priority to the user failed if the camera was angled down and WorldScribe took the floor as nearest to the user.

people in the video also moved dynamically, leaving some objects of interest to appear only briefly in the video. Thus, we annotated an object only if it lasted long enough and was clear enough to identify without pausing the video. Each label covered a time range from when the object appeared to when it disappeared from view. Another author then examined each label and marked it as *covered* if the descriptions within the time range included the annotated objects. We had 64 labels in total from the five video recordings.

Limitations. Due to the small pool of participants and the limited study time, we did not have a comparable number of ground truth labels to those standard video datasets. We will discuss more about this in Section 7.3.

Results. We found that 75% of annotated objects (48 out of 64) were covered by WorldScribe descriptions. Based on our post examination, WorldScribe successfully described and covered objects of users' intent in most cases if users faced for a long period, but failed in a few cases that an object was partially occluded. We also observed that WorldScribe failed to describe an object of users' intent in time if WorldScribe was still describing the previous visual scene while users had already moved to a new one.

6.3 Description Priority

We measured if WorldScribe prioritized descriptions based on users' intent or the proximity of described content to users.

Dataset & Analysis. In total, we collected 120 descriptions by randomly selecting 20 descriptions generated by GPT-4v [7] from each participant's self-defined scenario. We marked an instance as correct if the presented description was relevant to the user's intent, or if the described content was nearest to the user.

Results. Our analysis revealed that 97 out of 120 descriptions (80.83%) aligned with user intent or were prioritized based on the proximity of the described content to the user. We found that errors commonly occurred when the relevant information was present but not the focus, such as considering the user's intended object as spatial reference (Figure 10c). Additionally, camera angles often varied, sometimes tilting towards the floor or ceiling, which was recognized as the nearest content to the user (Figure 10d).

7 DISCUSSION AND FUTURE WORK

In this section, we discuss our lessons learned and design implications for context-aware and customizable live visual descriptions.

7.1 Challenges in Describing the Real World

Describing the real world is more challenging than digital visual media due to the need for timely descriptions aligned with users' intent and the higher standards and expectations in high-stakes situations. While WorldScribe made an important step toward providing context-aware live descriptions, participants brought up several aspects and challenges for future research to address.

First, while WorldScribe simulated short-term memory by avoiding repeated descriptions within preceding sentences (Section 4.6.2), participants expressed a need for more sophisticated long-term memory for visual descriptions. They suggested that previously navigated spaces or paths should not be reintroduced upon revisiting; instead, only visual changes or new elements since the last visit should be described. Spatial information should reference a series of camera frames or more complete data source to construct and represent users' environment (*e.g.,* real-time NeRF [29, 53]), rather than relying on a single video frame.

Second, it is hard for users to express their intent in a few sentences, which may implicitly change over time when exploring an area. This need to update intents was highlighted by our participants, who hoped to converse with the system to update their intent or clarify the confusing descriptions, similar to how they interact with human describers. Ideally, aside from such turn-by-turn interactions, a context-aware live visual description system should implicitly learn and adapt to the user's intent and environment through long-term interactions with users to reduce friction and increase usability. Future works could incorporate other data sources and modalities, such as GPS data, maps, visual details in videos or images, and description history to enable long-term memory.

7.2 Towards More Humanized Descriptions

Besides the challenges in crafting useful descriptions for the real world, the way to present descriptions could also influence users' understanding or engagement with visual media or scenes. For instance, describing from a first-person or third-person perspective could affect immersion in the environment [24]. Tone, voice, and syntax [14, 18, 21, 22, 28, 66] could also significantly impact experiences and comprehension. During our study, we received varied comments and preferences on these presentation aspects. For example, participant P4 described WorldScribe's voice as "hoarse," making the content unclear and uncomfortable. While some appreciated WorldScribe's current tone, others found the descriptions "artistic" or "poetic" and too wordy for practical use. Also, while participants appreciated WorldScribe's pauses or increased volume for presentation clarity, they hoped WorldScribe could provide transitioning descriptions or earcons when shifting to a new visual scene to ensure it was describing the current but not the previous scene. They further noted that human describers use more colloquial language than WorldScribe when conveying useful information, and prioritize their clarity over grammatical nuances. Future works should consider and enable more customizations of presentation.

7.3 Benchmarking Dataset for Live Descriptions

Our pipeline evaluation was limited to data from six BVI people to reflect real-world experiences with WorldScribe, which are different from the current video captioning dataset in several aspects.

First, the quality of video is quite different from that of standard datasets. For instance, frames occasionally appear at unusual angles or become blurred due to several factors such as users' movement, whether the camera is handheld, attached to a swinging lanyard, or tucked inside a pocket. Second, objects of interest may not appear consistently across consecutive frames and could be partially obscured or located to the periphery, as camera aiming is particularly challenging for BVI people [19, 20, 34, 78, 79]. Third, to provide useful live visual descriptions, it is important to provide concrete details beyond describing visual events, such as providing distances or sizes in concrete units (*e.g.,* feet, meters). Fourth, spatial relationships of objects should pivot to users' perspective (*e.g.,* using clock directions), rather than the image itself (*e.g.,* something on your left but not something on the left of the image).

To enable appropriate evaluation of live descriptions, additional datasets and metrics are needed. First, a potential metric is *contextual responsiveness*, which evaluates if the utility of descriptions aligns with the current context, such as having directions during navigation, having rich adjectives when viewing artworks, or describing objects reached by the user physically. The second potential metric is *contextual timeliness*. For instance, high-stakes scenarios may require a higher timeliness to signal potential danger before it happens (*e.g.,* the status of traffic light, whizzing car), while low-stakes scenarios could have much room for latency. Third, a potential metric is *contexual detailedness*, which evaluates whether a description provides only the necessary information without excess visual detail (*e.g.,* using multiple adjectives when the user is only interested in color, or describing the status of all three traffic lights instead of just the lit one). Overall, evaluating the context-awareness of live descriptions involves multiple factors. To fulfill such high demand for live visual descriptions in the real world, future works should develop ways to collect and annotate video datasets shot by BVI people. It is also notable for including such datasets in existing ones to carefully build a universal and unbiased dataset that is not skewed toward any particular group.

7.4 Generalizing WorldScribe

We envision expanding WorldScribe to other media formats and integrating the rapidly evolving AI capabilities in the future. First, WorldScribe could be tailored to visual media that require immediate descriptions. For instance, when describing 360-degree videos [24], it was hard for describers to pre-populate audio descriptions for the different fields of view with rich information due to the user's unpredictable viewing trajectory. It would be beneficial for WorldScribe to generate live descriptions responsive to the user's current view, while automatically pausing descriptions when important sounds happen (*e.g.,* narration), or increasing volume when unimportant sounds occur (*e.g.,* background music) in the video.

Second, WorldScribe could also be extended to support low-vision users who may use wearables to receive visual aids in the real world [63, 88, 89]. For instance, the type of visual enhancement could be determined based on the user's mobility states and visual scenes, similar to WorldScribe detecting the user's orientation and frame similarity to provide corresponding descriptions. Also, live visual descriptions could confirm the visual scene for low-vision users, and WorldScribe could use the visually enhanced image

frame from the wearables to increase description accuracy. Future works could explore integrating different assistive technologies (*e.g.,* wearables, navigation systems) as an ecosystem to provide corresponding contextual support.

7.5 Directions with Rapid Evolution of Future Large Models

WorldScribe provided live visual descriptions by leveraging LLMs to understand users' intent and an architecture that balances the tradeoffs between latency and richness of different VLMs. Given the rapid evolution of VLMs and LLMs and computing power in recent years, it is foreseeable that accuracy and latency will significantly improve. This progress may possibly lead to the reliance on fewer or even a single large multimodal model (*e.g.,* GPT-4o [6]) to generate descriptions of varying granularity, but raising further questions about which inputs beyond images should be included when prompting. Additional contextual factors, such as environmental sounds, GPS data, and users' state and activities, could be considered, and the prompt structure could be dynamically changed based on these inputs and user needs. Future work could also explore incorporating more advanced AI models into the description generation process. For instance, we could improve descriptions with additional verification [37], deblur or increase resolution of the image for clarity, construct a 3D scene on the fly to provide accurate spatial information [29, 53], and explain sound causality by cross-grounding visual and audio data [55, 76, 87]. Future works should explore these possibilities in such a rapidly evolving landscape of computational platforms and AI model capabilities.

8 CONCLUSION

We have presented WorldScribe, a system towards providing context-aware live visual descriptions to facilitate the environmental understanding for BVI people. Through a formative study with five BVI people, we identified several design goals of providing context-aware live visual descriptions. We implemented several components to tailor user's contexts, such as enabling users to specify their intent and generate descriptions tailored to their needs, providing consecutive short or long detailed descriptions based on visual context, and presenting descriptions with pausing or volume increased based on the sound context. Through an evaluation with six BVI people, we demonstrated how they perceived the WorldScribe descriptions and identified gaps in fulfilling their expectations for using WorldScribe descriptions in practice. Through a pipeline evaluation, we showed WorldScribe can provide fairly accurate visual descriptions, cover information about the user's intent, and prioritize descriptions based on the user's intent. Finally, we discussed more challenges in describing the real world, how to make descriptions more humanized and usable, and potential benchmark datasets. Through this work, we also recognized promoting real-world accessibility through live descriptions will be a long-term overarching problem, considering the diversity of people's needs and the complexity of real-world environments.

ACKNOWLEDGMENTS

We thank our anonymous reviewers and all the participants in our study for their suggestions, as well as Andi Xu for helping facilitate our user studies.

REFERENCES

[1] 2024. Aira. https://aira.io/
[2] 2024. BeMyEyes. https://www.bemyeyes.com/
[3] 2024. BlindSquare. https://www.blindsquare.com/
[4] 2024. Envision AI. https://www.letsenvision.com/
[5] 2024. GPT-4. https://openai.com/index/gpt-4/
[6] 2024. GPT-4 Omni. https://openai.com/index/hello-gpt-4o/
[7] 2024. GPT-4 Vision. https://platform.openai.com/docs/guides/vision
[8] 2024. Introducing Be My AI (formerly Virtual Volunteer) for People who are Blind or Have Low Vision, Powered by OpenAI's GPT-4. https://www.bemyeyes.com/blog/introducing-be-my-eyes-virtual-volunteer
[9] 2024. Microsoft Soundscape. https://www.microsoft.com/en-us/research/product/soundscape/
[10] 2024. Moondream. https://moondream.ai/
[11] 2024. SeeingAI. https://www.seeingai.com/
[12] 2024. Sound Analysis: Classify various sounds by analyzing audio files or streams. https://developer.apple.com/documentation/soundanalysis/
[13] 2024. Specific Guidelines: Art, Photos and Cartoons. http://diagramcenter.org/specific-guidelines-final-draft.html
[14] 3PlayMedia. 2020. Beginner's Guide to Audio Description. https://go.3playmedia.com/hubfs/WP%20PDFs/Beginners-Guide-to-Audio-Description.pdf. Accessed: 2021-01-13.
[15] Mouna Afif, Yahia Said, Edwige Pissaloux, Mohamed Atri, et al. 2020. Recognizing signs and doors for indoor wayfinding for blind and visually impaired persons. In *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. IEEE, 1–4.
[16] Dragan Ahmetovic, Cole Gleason, Chengxiong Ruan, Kris Kitani, Hironobu Takagi, and Chieko Asakawa. 2016. NavCog: a navigational cognitive assistant for the blind. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Florence, Italy) *(MobileHCI '16)*. Association for Computing Machinery, New York, NY, USA, 90–99. https://doi.org/10.1145/2935334.2935361
[17] Dragan Ahmetovic, Daisuke Sato, Uran Oh, Tatsuya Ishihara, Kris Kitani, and Chieko Asakawa. 2020. ReCog: Supporting Blind People in Recognizing Personal Objects. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376143
[18] Audio Description Project American Council of the Blind. 2017. Guideline for Audio Describers. https://www.acb.org/adp/guidelines.html. Accessed: 2020-11-6.
[19] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology* (New York, New York, USA) *(UIST '10)*. Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/1866029.1866080
[20] Erin L. Brady, Yu Zhong, Meredith Ringel Morris, and Jeffrey P. Bigham. 2013. Investigating the appropriateness of social network question asking as a resource for blind users. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 1225–1236. https://doi.org/10.1145/2441776.2441915
[21] Northern German Broadcasting. 2023. Audio description guidelines. https://www.ndr.de/fernsehen/barrierefreie_angebote/audiodeskription/Audio-description-guidelines,audiodeskription142.html. Accessed: 2023-04-09.
[22] Media Access Canada. 2023. DESCRIPTIVE VIDEO PRODUCTION AND PRESENTATION BEST PRACTICES GUIDE FOR DIGITAL ENVIRONMENTS. http://www.mediac.ca/DVBPGDE_V2_28Feb2012.asp. Accessed: 2023-04-09.
[23] Ruei-Che Chang, Chia-Sheng Hung, Bing-Yu Chen, Dhruv Jain, and Anhong Guo. 2024. SoundShift: Exploring Sound Manipulations for Accessible Mixed-Reality Awareness. In *Designing Interactive Systems Conference* (IT University of Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 116–132. https://doi.org/10.1145/3643834.3661556
[24] Ruei-Che Chang, Chao-Hsien Ting, Chia-Sheng Hung, Wan-Chen Lee, Liang-Jin Chen, Yu-Tzu Chao, Bing-Yu Chen, and Anhong Guo. 2022. OmniScribe: Authoring Immersive Audio Descriptions for 360° Videos. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 15, 14 pages. https://doi.org/10.1145/3526113.3545613

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

[26] Qiang Chen, Yinong Chen, Jinhui Zhu, Gennaro De Luca, Mei Zhang, and Ying Guo. 2020. Traffic light and moving object detection for a guide-dog robot. *The Journal of Engineering* 2020, 13 (2020), 675–678.

[27] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. 2024. YOLO-World: Real-Time Open-Vocabulary Object Detection. *arXiv preprint arXiv:2401.17270* (2024).

[28] Described and Captioned Media Program. 2020. Described and Captioned Media Program (DCMP). http://www.descriptionkey.org/quality_description.html. Accessed: 2019-03-19.

[29] Daniel Duckworth, Peter Hedman, Christian Reiser, Peter Zhizhin, Jean-François Thibert, Mario Lučić, Richard Szeliski, and Jonathan T Barron. 2023. SMERF: Streamable Memory Efficient Radiance Fields for Real-Time Large-Scene Exploration. *arXiv preprint arXiv:2312.07541* (2023).

[30] Christin Engel, Karin Müller, Angela Constantinescu, Claudia Loitsch, Vanessa Petrausch, Gerhard Weber, and Rainer Stiefelhagen. 2020. Travelling more independently: A Requirements Analysis for Accessible Journeys to Unknown Buildings for People with Visual Impairments. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 27, 11 pages. https://doi.org/10.1145/3373625.3417022

[31] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. "It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 549–559. https://doi.org/10.1145/3308558.3313605

[32] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376728

[33] Jaylin Herskovitz, Andi Xu, Rahaf Alharbi, and Anhong Guo. 2023. Hacking, Switching, Combining: Understanding and Supporting DIY Assistive Technology Design by Blind People. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 57, 17 pages. https://doi.org/10.1145/3544548.3581249

[34] Naoki Hirabayashi, Masakazu Iwamura, Zheng Cheng, Kazunori Minatani, and Koichi Kise. 2023. VisPhoto: Photography for People with Visual Impairments via Post-Production of Omnidirectional Camera Imaging. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 6, 17 pages. https://doi.org/10.1145/3597638.3608422

[35] Nicole Holmes and Kelly Prentice. 2015. iPhone video link as an orientation tool: Remote O&M for people with vision impairment. *Vision Rehabilitation International* 7, 1 (2015), 60–67.

[36] Jonggi Hong, Jaina Gandhi, Ernest Essuah Mensah, Farnaz Zamiri Zeraati, Ebrima Jarjue, Kyungjun Lee, and Hernisa Kacorri. 2022. Blind Users Accessing Their Training Images in Teachable Object Recognizers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 14, 18 pages. https://doi.org/10.1145/3517428.3544824

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

[38] The Smith-Kettlewell Eye Research Institute. 2022. YouDescribe. https://youdescribe.org/

[39] Lucy Jiang, Crescentia Jung, Mahika Phutane, Abigale Stangl, and Shiri Azenkot. 2024. " It's Kind of Context Dependent": Understanding Blind and Low Vision People's Video Accessibility Preferences Across Viewing Scenarios. *arXiv preprint arXiv:2403.10792* (2024).

[40] Lucy Jiang, Mahika Phutane, and Shiri Azenkot. 2023. Beyond Audio Description: Exploring 360° Video Accessibility with Blind and Low Vision Users Through Collaborative Creation. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 50, 17 pages. https://doi.org/10.1145/3597638.3608381

[41] Rie Kamikubo, Naoya Kato, Keita Higuchi, Ryo Yonetani, and Yoichi Sato. 2020. Support Strategies for Remote Guides in Assisting People with Visual Impairments for Effective Indoor Navigation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376823

[42] Seita Kayukawa, Keita Higuchi, João Guerreiro, Shigeo Morishima, Yoichi Sato, Kris Kitani, and Chieko Asakawa. 2019. BBeep: A Sonic Collision Avoidance System for Blind Travellers and Nearby Pedestrians. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300282

[43] Seita Kayukawa, Tatsuya Ishihara, Hironobu Takagi, Shigeo Morishima, and Chieko Asakawa. 2020. BlindPilot: A Robotic Local Navigation System That Leads Blind People to a Landmark Object. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3382925

[44] Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context Matters for Image Descriptions for Accessibility: Challenges for Referenceless Evaluation Metrics. arXiv:2205.10646 [cs.CL]

[45] Masaki Kuribayashi, Tatsuya Ishihara, Daisuke Sato, Jayakorn Vongkulbhisal, Karnik Ram, Seita Kayukawa, Hironobu Takagi, Shigeo Morishima, and Chieko Asakawa. 2023. PathFinder: Designing a Map-Less Navigation System for Blind People in Unfamiliar Buildings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 41, 16 pages. https://doi.org/10.1145/3544548.3580687

[46] Masaki Kuribayashi, Seita Kayukawa, Hironobu Takagi, Chieko Asakawa, and Shigeo Morishima. 2021. LineChaser: A Smartphone-Based Navigation System for Blind People to Stand in Lines. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 33, 13 pages. https://doi.org/10.1145/3411764.3445451

[47] Masaki Kuribayashi, Seita Kayukawa, Jayakorn Vongkulbhisal, Chieko Asakawa, Daisuke Sato, Hironobu Takagi, and Shigeo Morishima. 2022. Corridor-Walker: Mobile Indoor Walking Assistance for Blind People to Avoid Obstacles and Recognize Intersections. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 179 (sep 2022), 22 pages. https://doi.org/10.1145/3546714

[48] Walter S. Lasecki, Phyo Thiha, Yu Zhong, Erin Brady, and Jeffrey P. Bigham. 2013. Answering visual questions with conversational crowd assistants. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (Bellevue, Washington) *(ASSETS '13)*. Association for Computing Machinery, New York, NY, USA, Article 18, 8 pages. https://doi.org/10.1145/2513383.2517033

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

[50] Sooyeon Lee, Madison Reddie, Krish Gurdasani, Xiying Wang, Jordan Beck, Mary Beth Rosson, and John M Carroll. 2018. Conversations for Vision: Remote Sighted Assistants Helping People with Visual Impairments. *arXiv preprint arXiv:1812.00148* (2018).

[51] Sooyeon Lee, Madison Reddie, Chun-Hua Tsai, Jordan Beck, Mary Beth Rosson, and John M. Carroll. 2020. The Emerging Professional Practice of Remote Sighted Assistance for People with Visual Impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376591

[52] Sooyeon Lee, Madison Reddie, Chun-Hua Tsai, Jordan Beck, Mary Beth Rosson, and John M. Carroll. 2020. The Emerging Professional Practice of Remote Sighted Assistance for People with Visual Impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376591

[53] Chaojian Li, Sixu Li, Yang Zhao, Wenbo Zhu, and Yingyan Lin. 2022. RT-NeRF: Real-Time On-Device Neural Radiance Fields Towards Immersive AR/VR Rendering. arXiv:2212.01120 [cs.AR]

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

[55] Xingyu "Bruce" Liu, Ruolin Wang, Dingzeyu Li, Xiang Anthony Chen, and Amy Pavel. 2022. CrossA11y: Identifying Video Accessibility Issues via Cross-modal Grounding. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 43, 14 pages. https://doi.org/10.1145/3526113.3545703

[56] Timo Lüddecke and Alexander Ecker. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7086–7096.

[57] Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. 2018. Rich Representations of Visual Content for Screen Reader Users. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3173633

[58] Cecily Morrison, Martin Grayson, Rita Faia Marques, Daniela Massiceti, Camilla Longden, Linda Wen, and Edward Cutrell. 2023. Understanding Personalized Accessibility through Teachable AI: Designing and Evaluating Find My Things for People who are Blind or Low Vision. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 31, 12 pages. https://doi.org/10.1145/3597638.3608395

[59] Rosiana Natalie, Ruei-Che Chang, Smitha Sheshadri, Anhong Guo, and Kotaro Hara. 2024. Audio Description Customization. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility* (St. John's, Newfoundland and Labrador, Canada) *(ASSETS '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3663548.3675617

[60] Rosiana Natalie, Jolene Loh, Huei Suen Tan, Joshua Tseng, Ian Luke Yi-Ren Chan, Ebrima H Jarjue, Hernisa Kacorri, and Kotaro Hara. 2021. The Efficacy of Collaborative Authoring of Video Scene Descriptions. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 17, 15 pages. https://doi.org/10.1145/3441852.3471201

[61] Rosiana Natalie, Joshua Tseng, Hernisa Kacorri, and Kotaro Hara. 2023. Supporting Novices Author Audio Descriptions via Automatic Feedback. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 77, 18 pages. https://doi.org/10.1145/3544548.3581023

[62] American Council of the Blind. 2022. The Audio Description Project. https://adp.acb.org/guidelines.html

[63] Cristian Pamparău and Radu-Daniel Vatavu. 2021. FlexiSee: flexible configuration, customization, and control of mediated and augmented vision for users of smart eyewear devices. *Multimedia Tools and Applications* 80, 20 (2021), 30943–30968.

[64] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 747–759. https://doi.org/10.1145/3379337.3415864

[65] Helen Petrie, Chandra Harrison, and Sundeep Dev. 2005. Describing images on the web: a survey of current practice and prospects for the future. *Proceedings of Human Computer Interaction International (HCII)* 71, 2 (2005).

[66] Audio Description Project. 2023. Recommendation of the Federal Communications Commission disability ... https://adp.acb.org/docs/DAC%20Recommendation%20on%20Audo%20Description%20Quality%20Adopted%20October%2014%202020.pdf

[67] Elliot Salisbury, Ece Kamar, and Meredith Morris. 2017. Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5. 147–156.

[68] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. 2018. Evaluating and Complementing Vision-to-Language Technology for People who are Blind with Conversational Crowdsourcing.. In *IJCAI*. 5349–5353.

[69] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8430–8439.

[70] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*. Elsevier, 364–371.

[71] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[72] Abigale Stangl, Shasta Ihorn, Yue-Ting Siu, Aditya Bodi, Mar Castanon, Lothar D Narins, and Ilmi Yoon. 2023. The Potential of a Visual Dialogue Agent In a Tandem Automated Audio Description System for Videos. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 32, 17 pages. https://doi.org/10.1145/3597638.3608402

[73] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376404

[74] Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who Are Blind or Have Low Vision. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 16, 15 pages. https://doi.org/10.1145/3441852.3471233

[75] Haobin Tan, Chang Chen, Xinyu Luo, Jiaming Zhang, Constantin Seibold, Kailun Yang, and Rainer Stiefelhagen. 2021. Flying guide dog: Walkable path discovery for the visually impaired utilizing drones and transformer-based semantic segmentation. In *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 1123–1128.

[76] Yapeng Tian, Di Hu, and Chenliang Xu. 2021. Cyclic Co-Learning of Sounding Object Visual Grounding and Sound Separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2745–2754.

[77] Tess Van Daele, Akhil Iyer, Yuning Zhang, Jalyn C Derry, Mina Huh, and Amy Pavel. 2024. Making Short-Form Videos Accessible with Hierarchical Video Summaries. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 895, 17 pages. https://doi.org/10.1145/3613904.3642839

[78] Marynel Vázquez and Aaron Steinfeld. 2012. Helping visually impaired users properly aim a camera. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility* (Boulder, Colorado, USA) *(ASSETS '12)*. Association for Computing Machinery, New York, NY, USA, 95–102. https://doi.org/10.1145/2384916.2384934

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

[80] World Wide Web Consortium (W3C). 2022. Audio Description or Media Alternative. https://www.w3.org/TR/2008/REC-WCAG20-20081211/#media-equiv-audio-desc

[81] World Wide Web Consortium (W3C). 2022. Providing a movie with extended audio descriptions. https://www.w3.org/TR/WCAG20-TECHS/G8.html

[82] World Wide Web Consortium (W3C). 2022. W3C Image Concepts. https://www.w3.org/WAI/tutorials/images/

[83] Yutaro Yamanaka, Seita Kayukawa, Hironobu Takagi, Yuichi Nagaoka, Yoshimune Hiratsuka, and Satoshi Kurihara. 2021. One-Shot Wayfinding Method for Blind People via OCR and Arrow Analysis with a 360-degree Smartphone Camera. In *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*. Springer, 150–168.

[84] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891* (2024).

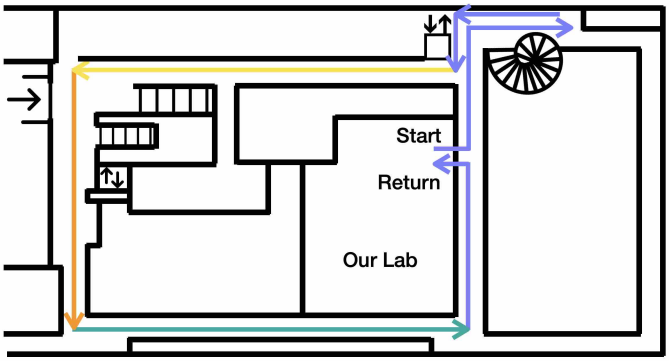
[85] Beste F. Yuksel, Soo Jung Kim, Seung Jung Jin, Joshua Junhee Lee, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Ilmi Yoon, Yue-Ting Siu, and Joshua A. Miele. 2020. Increasing Video Accessibility for Visually Impaired Users with Human-in-the-Loop Machine Learning. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3382821

[86] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*. Springer, 1–21.

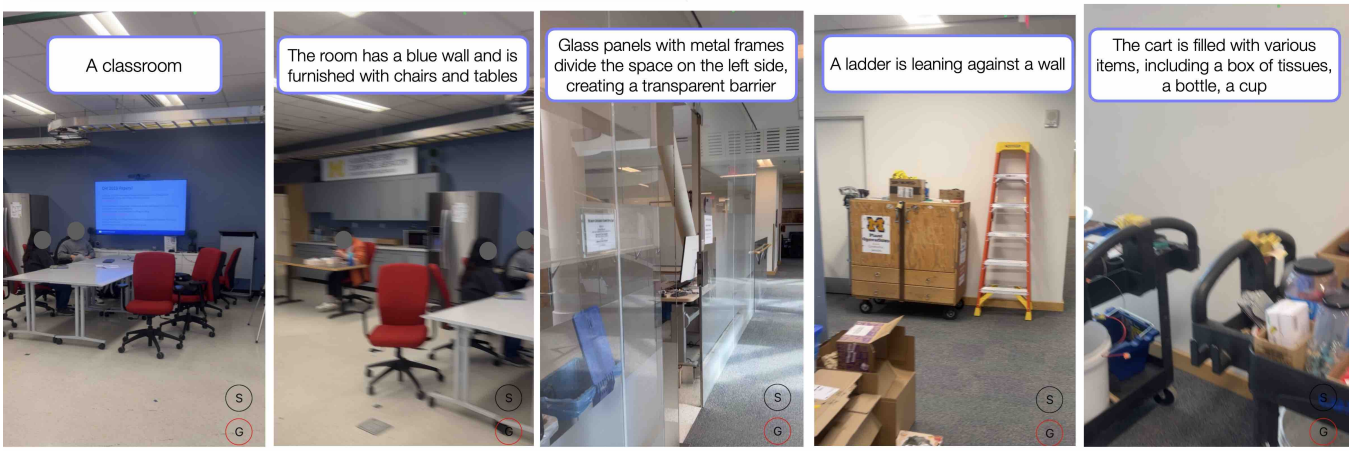
[87] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023. BuboGPT: Enabling Visual Grounding in Multi-Modal LLMs. arXiv:2307.08581 [cs.CV]

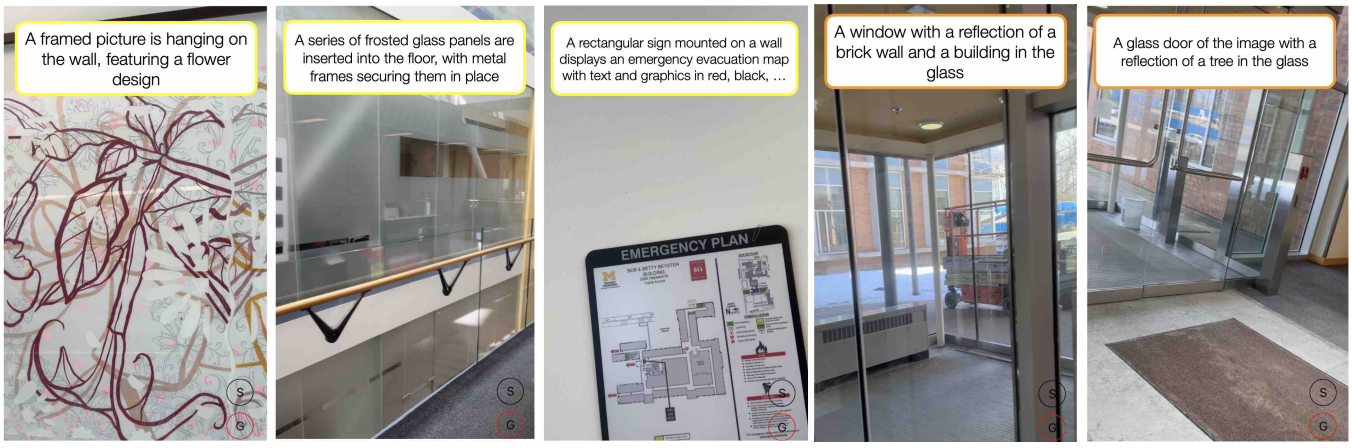
[88] Yuhang Zhao, Sarit Szpiro, and Shiri Azenkot. 2015. ForeSee: A Customizable Head-Mounted Vision Enhancement System for People with Low Vision. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility* (Lisbon, Portugal) *(ASSETS '15)*. Association for Computing Machinery, New York, NY, USA, 239–249. https://doi.org/10.1145/2700648.2809865

[89] Yuhang Zhao, Sarit Szpiro, Jonathan Knighten, and Shiri Azenkot. 2016. CueSee: exploring visual cues for people with low vision to facilitate a visual search task. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 73–84. https://doi.org/10.1145/2971648.2971730



The floor plan of 3rd floor of our building. And the route we went with participants in the scenario with general intent.





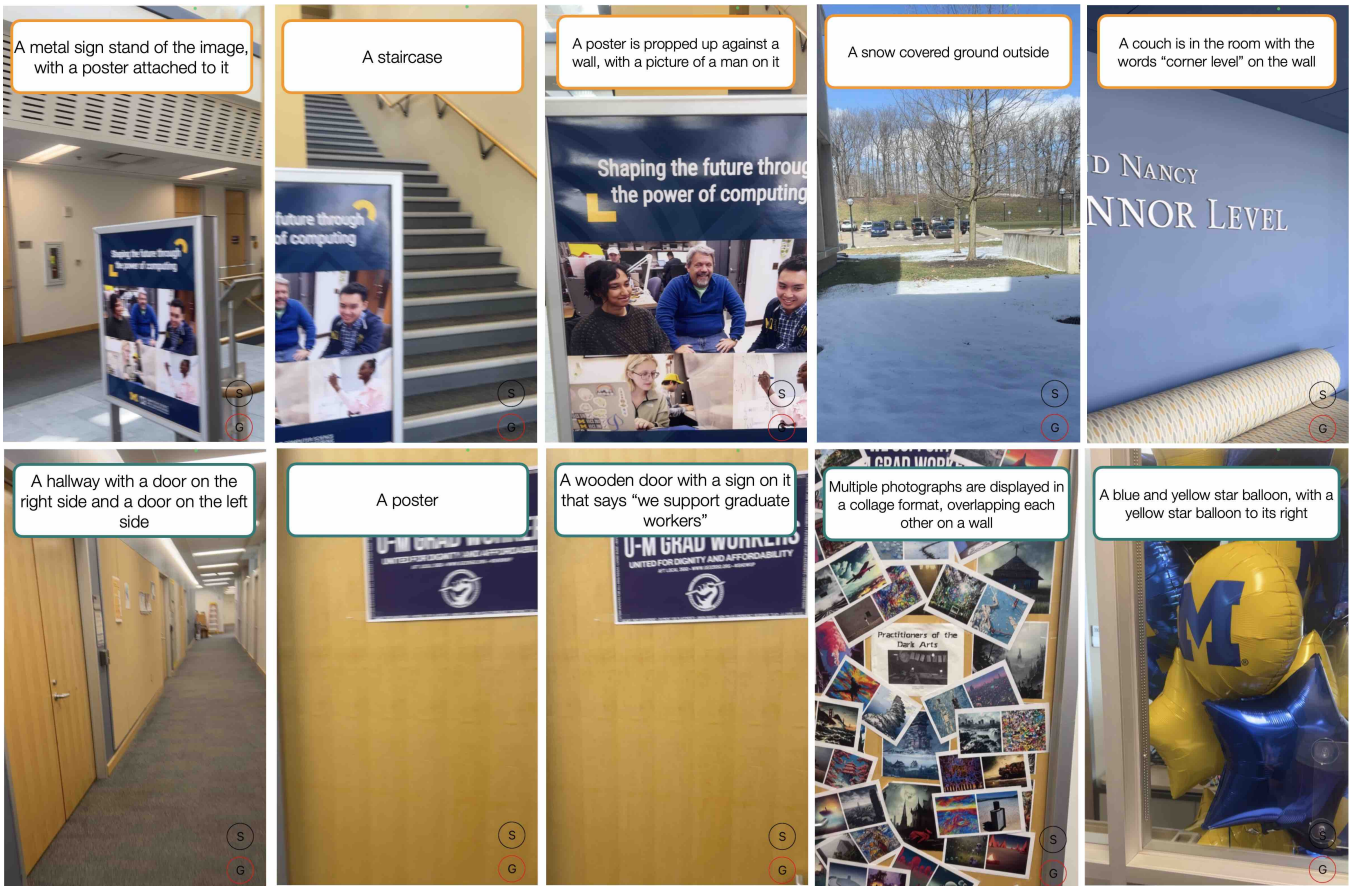


Figure 11: Example study route, keyframes, and WorldScribe-generated descriptions in our user study with P6.