# TableNarrator: Making Image Tables Accessible to Blind and Low Vision People

Ye Mo
Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems, College of Computer Science
Zhejiang University
Hangzhou, Zhejiang, China
moye0017@gmail.com

Gang Huang
School of Software Technology
Zhejiang University
Ningbo, Zhejiang, China
david.huang@zju.edu.cn

liangcheng li
Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems, College of Computer Science
Zhejiang University
Hangzhou, Zhejiang, China
liangcheng_li@zju.edu.cn

Dazhen Deng
School of Software Technology
Zhejiang University
Ningbo, Zhejiang, China
dengdazhen@zju.edu.cn

Zhi Yu
School of Software Technology
Zhejiang University
Ningbo, Zhejiang, China
yuzhirenzhe@zju.edu.cn

Yilun Xu
School of Software Technology
Zhejiang University
Ningbo, Zhejiang, China
yilun.xu@zju.edu.cn

Kai Ye
School of Software Technology
Zhejiang University
Ningbo, Zhejiang, China
knorrsumurskoxnm@hotmail.com

Sheng Zhou*
Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems, School of Software Technology
Zhejiang University
Ningbo, Zhejiang, China
zhousheng_zju@zju.edu.cn

Jiajun Bu
Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems, College of Computer Science
Zhejiang University
Hangzhou, Zhejiang, China
bjj@zju.edu.cn

## Abstract

The widespread use of image tables presents significant accessibility challenges for blind and low vision (BLV) people, limiting their access to critical data. Despite advancements in artificial intelligence (AI) for interpreting image tables, current solutions often fail to consider the specific needs of BLV users, leading to a poor user experience. To address these issues, we introduce TableNarrator, an innovative system designed to enhance the accessibility of image tables. Informed by accessibility standards and user feedback, TableNarrator leverages AI to generate alternative text tailored to the cognitive and reading preferences of BLV users. It streamlines access through a simple interaction mode and offers personalized options. Our evaluations, from both technical and user perspectives, demonstrate that TableNarrator not only provides accurate and comprehensive table information but also significantly enhances the user experience for BLV people.

*Corresponding author.


*CHI '25, Yokohama, Japan*

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714329

## CCS Concepts

• **Human-centered computing → Accessibility systems and tools**.

## Keywords

Accessibility, Assistive Technology, Screen Reader, Image Tables, Computer Vision, Large Language Model

**ACM Reference Format:**
Ye Mo, Gang Huang, liangcheng li, Dazhen Deng, Zhi Yu, Yilun Xu, Kai Ye, Sheng Zhou, and Jiajun Bu. 2025. TableNarrator: Making Image Tables Accessible to Blind and Low Vision People. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3706598.3714329

## 1 Introduction

Tables are essential data formats for conveying structured information. Widely used in real-world critical fields such as education, scientific research, medical diagnosis, and commercial activity, they play a pivotal role in enabling BLV users to make informed decisions and access knowledge, making their accessibility vital for promoting equitable participation [3, 23, 51, 73]. In the early stage, tables are often stored in CSV (Comma-Separated Values) format [30] or structured using HTML (HyperText Markup Language) [10]. With the increasing demand for visually appealing designs, cross-platform compatibility, *image tables* have become a widely-used tabular format [73], where the table text and other

visual information such as color, layout, and graphical elements are naturally rendered into a single image.

Although image tables enjoy advantages in information expression and propagation, they have posed significant challenges in table perception and understanding for BLV people [52]. Unlike tables stored in CSV or structured using HTML that can be directly read by screen readers [4], image tables require manually added alternative text or additional computer recognition and interpretation. Although techniques like optical character recognition (OCR) [15, 24] can capture the text information, the lack of rich hierarchical and visual information hinders users from fully accessing information in the table [40, 49], thereby leading to difficulties in understanding. Accurately identifying and understanding the rich information within image tables is important for BLV people to access information equitably [65] and integrate into the community.

Artificial intelligence technologies, such as table recognition [71, 73] and table understanding [62, 64], have opened new opportunities in this area over the past years. However, these methods still face critical challenges to meet the needs of BLV people. On one hand, most algorithms focus on specific scenarios, such as PDF documents, but struggle to accurately process diverse and complex table formats in real-world contexts like e-commerce or social media. Figure 1 shows an example of complex image table accessing. Three existing common methods are applied on it to obtain alternative text. As two results provided via recent screen reader are messy or invalid information for BLV users, even the most advanced visual understanding models (such as GPT-4V [1]) cannot correctly interpret image tables, especially lacking in the functionality of elements within the tables. On the other hand, the output of existing table understanding algorithms does not fully consider the needs of BLV people. Through our formative study with BLV users, we found that the alternative text they need are **concise and precise** descriptions. Over-processed redundant information further complicates understanding, leading to a poor user experience. To some extent, a question-answer interaction mode may provide the proper information, but it requires users to have prior knowledge of the table's content. Additionally, the information provided by visual language models often faces the risk of hallucination [25].

To tackle the above challenges, we conducted a formative study to gain a deeper understanding of the needs and preferred descriptive formats of BLV users. We recruited 8 BLV users and collected comprehensive information through semi-structured interviews. Furthermore, we designed questions to explore the accessibility issues of image tables and summarized the main barriers, encouraging BLV users to share their suggestions for functional improvements.

To enhance the accessibility of image tables, we further designed TableNarrator, a system for accessing image tables that is friendly to BLV users. Based on the conclusions of the formative study and combining information accessibility standards and guidelines, we identified the design goals of TableNarrator. These goals include **simplifying the table structure** in the alternative text while **keeping cell information distinct**, **providing necessary supplementary information**, **offering a simple and direct interaction mode**, and **providing several personalized options**. We then established an integrated architecture to enhance the comprehension of image tables. This architecture is consists of table layout analysis module and table structure recognition module to extract image tables' content, structural relationships and to classify table headers. The architecture also utilize the capabilities of the large language model to fuse the semantic information. Furthermore, according to the requirements of BLV users, we designed personalized options to accommodate users' different access purposes and incorporated a simple but effective gesture to enhance the user experience. Our system is specifically designed to convey image table content efficiently and autonomously, and make the provided information easy to understand.

We performed a technical evaluation and user study on TableNarrator with prevalent baselines, and the results confirmed that TableNarrator not only has high technical accuracy and content coverage but also received higher ratings from the 9 BLV users who participated in the evaluation in terms of its content and interaction mode. Meanwhile, we make a deep observation of TableNarrator and discuss several insights into it.

Our contributions are mainly in the following aspects:

- We design and perform an investigation on the requirements for comprehending image tables of BLV users, which can guide the design of image table accessibility technology in the future.
- We propose TableNarrator, a novel system designed to analyze structural and semantic information of image tables, with several personalized options and a simple interaction mode for helping BLV users. This provides a prototype system for the accessibility research field.
- We conduct both technical evaluation and user study on TableNarrator. The evaluation methods, datasets, and user experience results will be released publicly to support further research in this field.

## 2 Related Work

### 2.1 Image Accessibility

Images have emerged as a fundamental data type in social media. While they offer distinct advantages in expressing and disseminating information, they simultaneously present substantial accessibility challenges for BLV people [53]. More specifically, while screen readers enable BLV individuals to access textual content, these tools fall short when it comes to images lacking alternative texts. According to the Web Accessibility Initiative (WAI) by the World Wide Web Consortium (W3C), images should be accompanied by text alternatives that encapsulate the information or function they represent[1]. In response to this requirement, a range of techniques have been proposed to enhance image accessibility by providing accurate descriptions. Early solutions have utilized crowdsourcing to manually generate the alternative texts [2, 59, 75], while they suffer from high resource consumption and subjectivity.

Recent advances have utilized machine learning techniques for automatic generation, with image caption techniques [22] being a prominent example. Widely used in the literature, these techniques take images as input and automatically generate descriptive

[1] https://www.w3.org/WAI/tutorials/images/

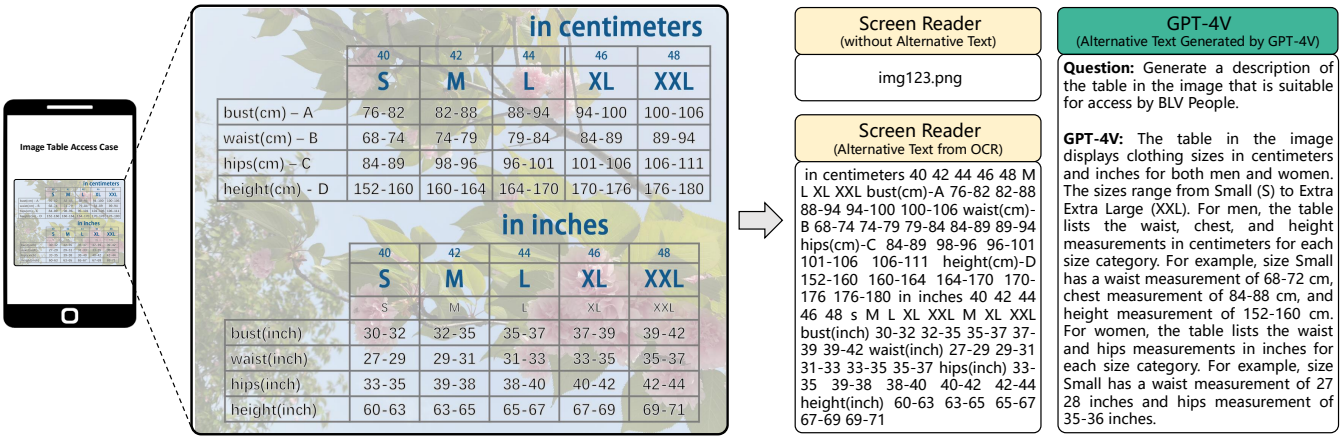


**Figure 1: Three common image table access methods. Without alternative text, the screen reader will read the filename of the image which is invalid information for the BLV users. With the help of OCR technologies, the screen reader can extract the content in all cells and read it from left to right and top to bottom sequentially. With a text prompt, the multimodal large language model such as GPT-4V can generate a paragraph of superficial descriptions according to the image table.**

summaries. Early works have utilized the image caption for alternative text generation in various scenarios, including single natural image [13, 58], meme images [46], mobile texting [14], photo sharing [72] and Regionspeak [74]. Later works have integrated the image caption techniques into practical image accessibility systems, such as Automatic Alt Text (AAT) [29], Be My Eyes, Seeing AI, Lookout and VoiceOver [63]. In recent years, the success of multi-modal large language models (MMLLM), such as Blip [32, 33], LLaVa [37] and GPT-4V [44], has provided exciting opportunities for accurate image caption, which has shown great potential in real-world scenarios. Although successful, the quality of the generated captions can vary, and they may not always capture the most important aspects of an image from a human perspective. Based on feedback from BLV users of our user study: *"Be My Eyes... even with very clear photographs... fails to recognize the text on them."* With regard to image content, existing accessibility techniques can be primarily classified into two categories: text-rich image accessibility and text-scant image accessibility. We primarily discuss the text-rich image accessibility.

## 2.2 Text-rich Image Accessibility

Text-rich images, which contain abundant textual content, are prevalent in various contexts, including scanned documents [17], advertisement images [54], and product manuals. Given that these images already contain readable textual information, the primary purpose of alternative text is to extract and organize this text in a comprehensible manner. To achieve this, Optical Character Recognition (OCR) [42], a technique that converts text within images into machine-readable formats, is commonly employed. With the advancement of technology, OCR has been successfully applied to various text-rich image accessibility scenarios [31, 66].

However, its success has limitations concerning the target objects. According to W3C WAI's classification of images, OCR is applicable only to common *Images of Text*, as it can recognize all text present in the image. In contrast, it struggles to perform effectively when dealing with *Complex Images*, such as tables, charts, and maps, which actually contain potential non-text information. To address this challenge, multi-modal techniques are employed to provide concise and precise text descriptions, thereby preserving the essential information within text-rich images. In this context, technologies such as document summarization [41] hold promising application potential, although they have not yet been sufficiently explored in research. Overall, despite the progress in text-rich image accessibility, existing works have primarily focused on nonstructural texts, largely overlooking the structural patterns among texts, especially in the case of image tables.

Although image accessibility has been widely studied in the literature, it still faces significant challenges in both structural text patterns and accurate description. This motivates the research of this paper and we will elaborate on the detailed image table accessibility in the following sections.

## 2.3 Image Table Accessibility

Tables serve as a fundamental tool for organizing data with logical relationships in grid structures. In their initial stages, tables were commonly presented in structured formats such as CSV or HTML. These formats allowed structural relationships to be explicitly marked and read by assistive technologies, including text-to-speech [50] and screen readers [4]. According to W3C WAI, accessible tables require HTML markup that identifies header cells and data cells and defines their relationship[2]. While this is straightforward for tables in structured formats, it poses a significant challenge for image tables. Although techniques such as OCR can readily extract text from image tables [43], the absence of critical structural information impedes comprehensive understanding [34]. For example, while the iPhone's VoiceOver has text recognition capabilities and can read out text descriptions found in images, it does so in a

[2]https://www.w3.org/WAI/tutorials/tables/

top-to-bottom, left-to-right sequence, reading text that is physically proximate to each other at a time.

In addressing this challenge, early research initiatives began with the extraction and understanding of tables in PDF files [12], a methodology that can also be applied to image tables. The subsequent launch of the ICDAR 2013 robust reading competition [28] significantly propelled advancements in this field, introducing popular tasks such as table detection [16], table structure recognition [36, 48, 67, 73], and table understanding [34]. Table detection is a process that identifies the location of tables on a page, including the coordinates and boundaries of the table within an image. Following table detection, table structure recognition aims to identify the table's structure, including its rows, columns, cells, and the relationships between cells, or reconstruct the table into semi-structured formats such as HTML or LaTeX code. Table understanding, a domain that often intersects with multimodal technologies, extends beyond table recognition. It concentrates on the extraction and analysis of table semantics to achieve a comprehensive understanding of table content. Common subtasks in the field of table understanding encompass Table-to-Text Generation [39, 45, 55], Table QA [8, 38, 70], Cell Type Classification [34], and Table-based Fact Verification [9]. Current research on table understanding primarily focuses on structured and semi-structured tables, such as database tables and spreadsheets, with less discussion centered on image tables.

Another approach to improving the accessibility of image tables is tactile embossing, where physical representations of tables are created using raised surfaces [7]. This method has been widely adopted in various fields, such as education, for presenting graphics and charts [6, 19]. Various materials and devices, such as embossed paper [11] and refreshable tactile displays (RTDs) [21], are currently available to support touch-based solutions. While tactile embossing offers BLV users a tangible means to interact with table data, it has notable limitations. It often requires specialized hardware and materials, resulting in high costs and significant time investment for creation [27]. Additionally, tactile representations are inherently static, making them unsuitable for dynamic or interactive use cases [18]. The physical size of tactile materials further restricts their usability, particularly for tables with large amounts of information [26]. These limitations underscore the need for scalable and interactive digital solutions, such as text-based approaches, to address the accessibility challenges posed by image tables.

In summary, while some existing research focuses on table accessibility, most of these efforts center on data tables (e.g., PDF tables, Web tables, spreadsheets), particularly in web environments with HTML tables. These studies offer various options for BLV users to access data tables. However, they largely overlook image tables, which are equally prevalent. Unlike data tables, image table parsing tools work with image pixels rather than semantic information and lack the ability to directly address individual cells. This limitation makes it challenging to apply traditional table accessibility methods to image tables. Thus, integrating semantic understanding and interactivity into image table accessibility tools, while aligning the capabilities of potentially applicable technologies with user needs, remains an urgent challenge. This motivates our question design in the subsequent formative study.

## 3 Formative Study

To capture the practical experiences and diverse preferences of BLV users when accessing image tables, we conduct a formative study to explore the requirements of BLV people in interacting with image tables. We recruited eight BLV users to participate in semi-structured interviews and carried out customized observation protocols that were adapted to the circumstances of each participant. Our carefully designed interview aims to answer the following questions that are pivotal to the design of image table accessibility systems:

- **Q1:** Are BLV users satisfied with the current image table accessibility?
- **Q2:** What are the most critical accessibility problems of image tables?
- **Q3:** How do BLV users prefer and expect to interact with image tables?

### 3.1 Method

**Participants.** In our study, we recruited eight BLV users to participate. To ensure accessibility, we conducted semi-structured interviews in two formats: online meetings and face-to-face interviews. Online meetings provided flexibility for participants, while face-to-face interviews facilitated communication and engagement. All participants followed the same predefined script. Each participant received a compensation of $20 for their time. We asked participants to self-assess their familiarity with tables based on their frequency of access and proficiency in use. Participants with different levels of familiarity were purposefully recruited to ensure diverse perspectives. This included seven participants (P1-P7) who were familiar with tables and one participant (P9) who was not, allowing us to consider individuals who might struggle with understanding structured data due to various factors such as educational background, age, and others. The participants' ages ranged from 27 to 50 years, representing a diverse array of professions. They have many different professions, including computer-related professionals (P2, P6), non-computer professionals (P1, P3, P5, P7, P9), a consultant well-versed in the latest advancements in accessibility technology, and the president of an association with extensive exposure to the BLV community (P4). All participants demonstrated proficiency in using the screen reader and were frequent internet users.

**Procedure.** For each participant, we initiated the process with a semi-structured interview. We introduced participants to the differences between image tables and data tables, exploring their familiarity with these structures. They were asked to describe their typical methods for accessing both tables and common images. Additionally, we inquired whether they could easily distinguish between text-containing images and image tables, and asked them to explain the reasons for their answers. Participants were also encouraged to share their experience in accessing tables and freely articulate their expectations. Subsequently, we introduced a series of pre-constructed table description demos and invited participants to suggest potential modifications. We provided the first participant with a rudimentary demo featuring only basic sequential reading of the cell text to establish a baseline to collect iterative feedback. Hierarchical reading demos were introduced in later iterations to avoid unnecessary complexity in the early stages. Subsequent

**Table 1: Participants Recruited for the Study. This table catalogs the 8 participants (P1-P7, P9) who engaged in the formative study and the 9 participants (P1, P3, P4, P5, P7-P11) who partook in the user study. Note that there was an overlap where 6 participants (P1, P3, P4, P5, P7, P9) were involved in both studies. All information was provided according to participant consent. Participants' familiarity with tables is based on their self-assessment.**

| PID | Sex | Age | Occupation | Onset | Device | Familiarity with Tables |
|---|---|---|---|---|---|---|
| P1 | Male | 27 | Music Professional | Undisclosed | PC, Mobile | High |
| P2 | Male | Undisclosed | Software Engineer | Undisclosed | PC, Mobile | High |
| P3 | Male | 37 | Self-employed | Acquired | PC, Mobile | High |
| P4 | Male | 50 | Association Chairperson, Consultant | Acquired | PC, Mobile | High |
| P5 | Female | 30 | Piano Tuner | Acquired | PC, Mobile | High |
| P6 | Male | Undisclosed | Software Engineer | Undisclosed | PC, Mobile | High |
| P7 | Female | 35 | Full-time Audio Broadcaster | Acquired | PC, Mobile | High |
| P8 | Male | 25 | Self-employed | Acquired | PC, Mobile | High |
| P9 | Male | 48 | Massage Therapist | Congenital | Mobile | Low |
| P10 | Male | 40 | Cashier | Acquired | Mobile | Low |
| P11 | Female | 46 | Cashier, Massage Therapist | Acquired | Mobile | Low |

participants received multiple demo versions that incorporated modifications based on feedback from previous participants. For each version, participants were asked to provide feedback, either approving the changes with a rationale or expressing disagreement. In cases of disagreement, we first explained the rationale behind the modifications suggested by previous participants, and then asked the current participant to clarify their objections, whether they were in complete disagreement or conditional on specific circumstances. If a participant's suggestions conflicted with previous feedback, they were classified as personalized options. Only suggestions that were widely supported by multiple participants were considered general improvements. This approach aimed to gradually refine the demo through iterative feedback, minimizing biases from individual preferences, and allowing the integration of diverse perspectives into the tool development.

**Analysis.** We first transcribed the interview content and then conducted a thematic analysis [5] using the Delve Tool [56]. Two researchers independently reviewed the transcripts to generate an initial set of codes based on recurring keywords, and participant expressions. Then, our researchers carried out an iterative coding process and grouped the codes into themes for our research questions.

## 3.2 Findings

In this subsection, we address the three questions posed prior to the commencement of the formative study:

*3.2.1* ***The accessibility of image tables is poor (Q1).*** In evaluating accessibility, we adhered to the four core principles widely recognized in the WCAG[60]: Perceivable, Operable, Understandable, and Robust. The first principle, Perceivable, mandates that information must be presented in ways that users can perceive. However, most BLV users often cannot even discern the presence of an image table. P7 stated: *"To be honest, I have not encountered image tables, or perhaps I have but simply overlooked them."* P2 remarked: *"There's certain frustration when accessing tables. There are a few reasons why I didn't realize it was a table."* The second principle, Operable, requires that users should be able to interact with interface elements. Unfortunately, this principle is often unfulfilled with image tables, as they are presented as a single image. Screen readers typically can only select this image as a whole and read out the corresponding alternative text. For instance, our participants using VoiceOver found that when the focus was set on a particular image table, the narration provided was the file name, followed by *'image, a set of black text on a white background...'* and then proceeded to vocalize all the text within the image as recognized by OCR. This approach leaves BLV users unable to manipulate or navigate individual cell data within the image table. The third principle, Understandable, emphasizes that the information and operation of the interface must be comprehensible to the user. However, some alternative texts are overly simplistic, making it difficult for BLV users to extract the information they need from the table. P2 added: *"... or even if you know it's a table, it's impossible to understand the content it displays."* The final principle, Robust, demands that content must be robust enough to be reliably interpreted by a variety of assistive technologies. Regrettably, there are currently no effective tools available to significantly improve the accessibility of image tables.

*3.2.2* ***The perceptibility and understanding are the most critical accessibility problems (Q2).*** In addition to the fragmented accessibility issues associated with image tables discussed above, we further investigate what are the most critical accessibility problems of image tables. Firstly, the BLV users are *not aware of the tables existing in the image*. Different from tables formatted in CSV or HTML that can be detected by existing screen readers with explicit markup or tags, the existence of image tables in an image cannot be detected. As a result, users cannot distinguish between an image table and an image with text. P5 noted: *"Essentially when I come across something that seems to be gibberish or incoherent, I consider there's a fair chance it could be a table."* Secondly, the BLV users *get confused with the alternative text of image tables* due to the lack of spatial table structure, especially in the case of multi-row/column

and multiple key-value mapping. For BLV users, particularly those with less education, they may lack the concept of structure, and even among those familiar with tables, confronting complex ones (with multiple merged cells, nested tables, etc.) or large tables can be confusing. More seriously, BLV users can only be aware of the existence of image tables when they fail to understand the text of unorganized table elements. P5 added: *"Indeed, whether it's text recognition for image tables, or data tables, both present obstacles for us.... The order is problematic."* Therefore, all our participants agreed on the idea of de-emphasizing the table structure and using semantically-enhanced text to convey the structural information.

*3.2.3* ***Preferences and expectations of image tables (Q3)***. We conducted a survey to understand participant preferences for accessing the information necessary to comprehend image tables. We classified their preferences into three primary categories: header-value relationships (P1, P2, P3, P4, P5, P6, P7, P9), table metadata (P1, P3, P4, P5, P7, P9), and personalized additional content (P1, P2, P4, P5, P6, P7). Aside from these categories, participants also voiced other expectations, with the most significant consensus centering on the reduction of information redundancy (P1, P2, P3, P4, P5, P6). Given that no existing tool supports comprehensive access to image tables, participant responses were shaped by their experiences with accessing HTML tables. Meanwhile, we encouraged them to share their opinions freely, without being constrained by the limitations of current tools.

**Header-Value Relationships.** All our participants expressed a desire for a method of accessing image tables that reduces the emphasis on the table structure. P1 stated: *'We wish for the cells to be organized... it requires the mind to build a model.'* P4 added: *'...hope not to have us think about the spatial structure, as it's somewhat energy-consuming.'* P6 advocated for minimizing the categorization of tables to avoid *'missing the most important focus.'* The relationship between headers and values fundamentally represents the table structure. When we proposed a method of combining headers and values to lessen the structural emphasis, all participants strongly endorsed it. P4 described this method *'quite good,'* P5 thought it made the reading *'clear'* and *'did not confuse the data,'* while P7 added: *'Sometimes, I indeed need to listen to the header again, as I forget what the header was.'* This method significantly reduces the pressure to understand the table structure, as users do not need to discern whether the header is a row or column, or if it serves both roles, nor do they need to remember which header corresponds to which data.

To determine whether image tables should be directly converted to HTML format, we provided participants with the audio from a traditional screen reader for reading HTML tables, as well as the audio from pre-built TableNarrator. All participants showed a preference for the latter. P9 described it as *'effortless'*. When we suggested using a complete textual summary of the table information as a second method to weaken the focus on the table structure, our participants voiced their objections, labelling it as *"overly redundant."* Additionally, this level of information processing impacted their *"Autonomy"* in accessing information.

**Image Table Metadata.** We categorized the metadata of tables into three components: the number of rows and columns (which includes indicators of the table's presence), a concise summary of the table's content, and an ordered indication of the table's headers. This categorization was guided by the W3C's tutorials on *complex images* and is classified under brief image descriptions. Our objective has been to minimize the length of these descriptions, thereby providing BLV users with a rudimentary comprehension of the scale and theme of the data they will encounter before formally accessing the information within the table. Regarding the information on the number of rows and columns, P3 stated: *"I feel this information is necessary."* P5 and P7 respectively described the concise table summary as *"essential"* with P7 adding, *"If you explain what this table is about and how many rows and columns it has, I'll have a general concept of it."* Moreover, when the number of headers is limited and users can remember them, the proposed 'value-header' reading method can be utilized, this can significantly enhance the efficiency of table access. P3 noted, *"If I hear up to this one (after the announcement of value is done), and I think that's enough for me; I don't need to listen further, and I can skip directly to the next cell."*

**Reducing Information Redundancy.** During the demonstration of the pre-built tool, the initial narration style provided was "header-value," such as Time-January, Sales-10, Time-February, Sales-12, Time-March, Sales-9. Among the eight participants, six proactively mentioned that this style of reading had too much information redundancy. P3 noted, *"My first impression was the redundancy of information."* P2 speculated about situations with many cells: *"If there are 100 cells with the same header, that would be exhausting. One would have to listen to the whole thing, when it should really be about efficiency."* Our participants called for a change in the narration style, and without prior knowledge, they offered the same suggestion. P3 suggested, *"The information should be presented in an inverted pyramid form... with the most important information at the forefront."* P2 added, *"...reversing the order... I don't deny that reading the headers before the values is the most common practice, but for the sake of efficiency, I'd prefer to hear the value first."* Six participants agreed with this suggestion. Additionally, P1 and P7 offered the same idea that the row and column numbers could be added at the end as a suffix. P1 believed, *"Those who want to listen will do so, and those who don't will ignore it,"* while P7 added, *"When we are uncertain (about the content), we can listen for a bit longer, and when it's clear, we can skip right past."*

**Personalized Content.** When we invited participants to offer additional suggestions, they unanimously emphasized the need for more personalized options. P2 said: *"Because everyone has different levels of education and personalized preferences..."* P6 also cautioned our researchers to *"be mindful of capability biases"*. Our participants proposed specific personalized content they required. These recommendations primarily involve secondary interactions, such as returning associated data after entering values in the table and filtering specific headers. We have categorized the purposes for table access proposed by users into four types: targeted search, data comparison, data analysis, and full table detail access. Targeted searches involve users focusing only on a subset of the table data, such as specific rows, columns, or cells. During data comparison tasks, users are more interested in identifying similarities and differences between data points. At this juncture, cell-granularity announcements may become less appropriate, and when cross headers are present (i.e., both rows and columns have headers), users may prefer different header comparison orientations depending

on the context. Data analysis refers to describing data trends and performing simple calculations, whereas full table detail access typically occurs in scenarios where all table content is important. In light of these classifications, our researchers propose the use of an AI voice assistant as a solution. P3 rated it as *"excellent"* while P9 believed *"this kind of thing is definitely needed."* P5 noted, *"Other forms of interaction would work as well, such as typing."*

### 3.3 System Design Goals

In our formative study, participants provided numerous constructive suggestions for designing our image table access system, based on which we designed the following four system design goals:

**G1: Retaining the structural relationships among cells.** BLV users struggle to discern table structures due to the lack of spatial information and cells relationships. As mentioned in Section 3.2.2 and the first point of Section 3.2.3, participants emphasized the difficulty of accessing table without clear structural cues. This system design goal aims to address these concerns by retaining structural relationships while simplifying their presentation.

**G2: Providing the metadata of image table.** In alignment with W3C recommendations and the requirements of our study participants (the second point of Section 3.2.3), the system is expected to offer appropriate table metadata before presenting the content, such as the number of rows and columns, concise table summary and table headers. This is critical for their general understanding of the table before further interactions, thus giving users the information of the table's scale, subject and access purpose. This system design goal aims to provide users with a clear overview before detailed interactions.

**G3: Providing a simple and direct interaction mode.** Besides the key table information above, our system will also consider the user experience by providing a simple and direct interaction mode that reduces the learning curve, allowing users to feel like interacting with a real table, rather than listening to lengthy image alternative text (the Operable principle in Section 3.2.1). Furthermore, to ensure system compatibility across APPs, our system will generate text instead of relying on other semi-structured table formats such as HTML code (the Robust principle in Section 3.2.1).

**G4: Providing personalized options.** As strongly endorsed by all users, different BLV users may have personalized preferences regarding how they access tables, such as listening to the alternative text through a screen reader, exploring the table cells interactively, and reading at different levels of granularity (the third and fourth points of Section 3.2.3). Thus, the system should provide multiple personalized options rather than a fixed one.

## 4 TableNarrator: Making Image Table Accessible

Drawing on insights from the formative study and aligning with our system design goals, we have innovatively designed TableNarrator, an intelligent system that makes image tables accessible. This system can be easily integrated into screen readers. Specifically, TableNarrator can be integrated as a modular component into existing screen readers (such as NVDA, VoiceOver) via software development APIs, enabling the real-time interpretation of image tables. Unlike traditional solutions that rely on predefined information about the development environment, TableNarrator directly utilizes state-of-the-art vision algorithms to analyze image pixels and reconstruct them into tables. This independence allows it to extract tables without requiring additional information from the development environment. Combined with its ability to integrate advanced semantic reasoning, TableNarrator can serve as an adaptive solution. To ensure compatibility, enhance interaction options, and improve system performance when handling complex tables[65], TableNarrator retains extracted information in plain text format rather than semi-structured formats like HTML.

TableNarrator comprises two information extraction modules for extracting table regions, understanding table structure, and an interaction module for conveying table information appropriately. 1) Table Layout Analysis Module: This module uses table detection to identify the table's position within the image and utilizes layout analysis to decompose the table into table title, table footer, and table body, laying the groundwork for subsequent levels of understanding. 2) Table Structure Recognition Module: This module first employs structuralization algorithms such as cell detection and RC (row-column) location to extract the table metadata, including rows and columns information, and coordinates of each cell. Based on the above information, the module further analyzes the advanced semantic contents and mines diverse relations between cells. By leveraging a Large Language Model (LLM), we also generate a concise summary with cell text in the table. 3) Personalized Options and Interaction Module: This module provides BLV users with a simple and seamless interaction mode along with personalized options. By utilizing single and double-finger gestures, it offers an experience akin to interacting with an actual table, rather than just viewing a standard image or navigating through lengthy alternative text. Additionally, it allows BLV users to customize the granularity and order of the alternative text according to their preferences.

### 4.1 Pipeline of TableNarrator

To meet the requirements, we design a table access pipeline for BLV users based on the extracted table information. The general workflow of TableNarrator is outlined below, and Figure 2 illustrates its framework.

When BLV users navigate between page elements using single-finger (or another default switching mode), if the system detects that the currently focused element is an image, the Table Layout Analysis module is invoked to determine whether the image contains a table. If a table is detected, the Table Layout Analysis module will locate and segment the table region, after which the Table Structure Recognition module will learn the structure and high-level semantic information of the table to generate alternative text to be used. First, the screen reader will read a short description [61] composed of table metadata, including a table cue, the number of rows and columns, table headers, and a concise summary. If the BLV user is not interested in the table, they can continue to swipe with single-finger, and the focus will move to the next element. If the BLV user is interested in the table, they can swipe with double-finger (or another custom switching mode), and the screen reader will read the current virtual focus's table header-table value pair and row number, column number at a cell-level granularity after each swipe. This virtual focus is implemented in the background by the screen

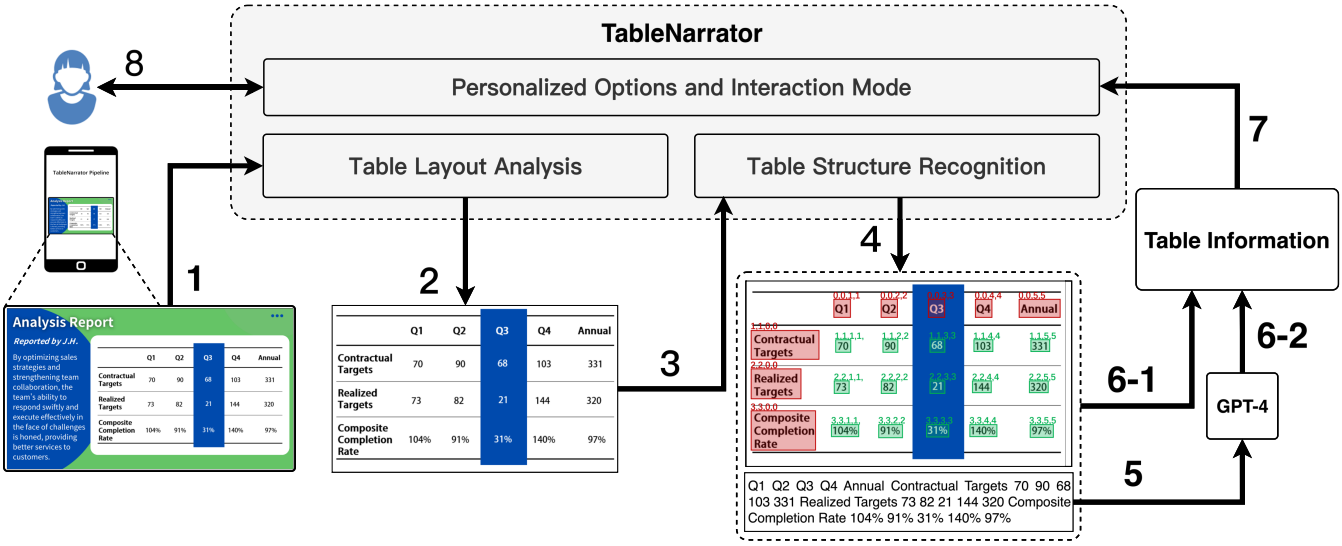


**Figure 2: The components and pipeline of TableNarrator system. It consists of the architecture of table layout analysis and table structure recognition to extract table information with steps 1, 2, 3, 4, 5, 6-1, 6-2, the personalized options and interaction mode considering user requirements can be utilized through step 7, 8. Step 1: The image is input into the Table Layout Analysis module. Step 2: The Table Layout Analysis module detects the presence of the table and locates the table region. Step 3: The table region is input into the Table Structure Recognition module for processing. Step 4: The Table Structure Recognition module extracts the table information. Step 5: The extracted text information is input into GPT-4, which generates a concise table summary. Step 6-1 and Step 6-2: The outputs from the Table Structure Recognition module and GPT-4 are combined into a readable table format. Step 7: The table information is sent to the Personalized Options and Interaction module. Step 8: Users set personalized options or directly applies the default interaction mode to access the table.**

reader and TableNarrator. BLV Users can also customize the focus granularity, such as switching the read text row-by-row with each double-finger swipe. Additionally, BLV users do not always need to browse the entire table. Once they have obtained the required information, they can switch from double-finger to single-finger to exit TableNarrator's alternative text and return to the external environment.

## 4.2 Table Layout Analysis Module

The Table Layout Analysis module is designed to detect table regions in an image and segment the table components into table title, table footer, and table body. This module plays a crucial role in automating the process of extracting structured data from images that contain tabular information. This step is accomplished based on the object detection algorithm, which outputs the coordinates and categories of objects in the image. According to our setup, if a table does not exist in the image, the algorithm will return a reminder. Specifically, we first apply CascadeTabNet (leading table detection model fine-tuned on our complex table images) [47] to detect tables in images. This approach locates the table region for further semantic and structural understanding, as opposed to plain text comprehension. However, many tables include title and footer text for complementary expressions, which are separate from the main body of the table. These elements play an indispensable role in table understanding but cannot be incorporated into table structural analysis. Therefore, we add a ResNet-based [20] module to segment the table title, table body, and table footer regions, and extract the information via an OCR module [67]. To refine the detected table regions and optimize the layout analysis results, TableNarrator also incorporated further design improvements. This may involve post-processing techniques such as boundary refinement, noise removal, and context-based analysis to enhance the accuracy of the three layout region classification results.

## 4.3 Table Structure Recognition Module

Unlike other sequential text, the text in tables has a structural organization that BLV users need to combine with the content to understand the information. However, OCR-based screen readers only read the text without providing any structural information, which leads to confusion for BLV users. For instance, in Figure 2, the table has 3 rows and 4 columns with split cells in the third row and second column. BLV users cannot obtain this specific structural information by merely reading cell content sequentially, which poses obstacles to understanding the table. To uncover their row-column relationships, we extract the position information of cells with the Table Structure Recognition module. This aims to understand the basic table structure that is used to compose the table metadata. Specifically, we first apply LORE (leading table recognition model fine-tuned on our complex table images) [67] to obtain the table structure. LORE uses a transformer-based network to recover the logistical position of each cell. Given an image region of a table, LORE will further detect the region of each cell and use

the positions of all cells to predict the local coordinates and the total number of rows and columns. Moreover, it can detect merging or splitting cells and analyze their row or column spans.

Table metadata can provide a coarse macroscopic structural understanding for BLV users. However, it is crucial for them to understand the detailed logical information among all the cells. For instance, if a BLV user wants to buy shoes and needs to browse through an image table of shoe sizes, they tend to focus on the content of a particular cell or compare it with others from different dimensions of the table. Therefore, another feature of the Table Structure Recognition module is the exploration of further semantic relation mining between different cells. It is a specialized component designed to analyze the logical relationships between cells and reconstruct the key-value relationships within the table. When considering the style of table cells, table header cells and table value cells often come in pairs, providing key-value relation information. Meanwhile, some merging or splitting cells containing table values may have other relations such as parataxis and contradiction. To mine these diverse relations, we design a transformer-based cell relation classification sub-module with multimodal inputs, including the position and content of cells. As a result, the model can identify the connections between the table headers and their associated values. This deeper understanding of the data promotes more insightful analysis and interpretation. As a complement to the table metadata, we input the combined table header-value pairs into a large language model to generate a concise table summary, which is crucial for helping users grasp the overall content of the table. Additionally, the Table Structure Recognition module also includes functionality for refining the extracted cell positions and improving the accuracy of the row-column relationships. This may involve post-processing techniques such as refinement of the cell boundary, error correction, and context-based analysis to improve the overall quality of the extracted table structure.

## 4.4 Personalized Options and Interaction Module

Figure 3, Figure 4 and Figure 5 illustrate the detailed interaction instructions for TableNarrator. The first two modules are automatically activated without any configuration, showcasing the information extraction capability of TableNarrator. The third module ensures an optimal user experience, achieved through seamless integration between the screen reader and TableNarrator. In terms of interaction, TableNarrator features simple gesture controls, allowing BLV users to enter or exit the TableNarrator environment with default single-finger or double-finger swipes, or via custom actions. From a development perspective, the screen reader interprets different gestures by leveraging gesture recognition algorithms integrated within the operating system and maps them to TableNarrator's predefined commands. Compared to traditional image description tools, TableNarrator reimagines image table accessibility through an interactive model. Traditional image description methods often leave BLV users overwhelmed with lengthy and unstructured text. TableNarrator's interaction mode not only provides BLV users with structured and hierarchical information but also offers a access experience that is immersive while preserving reading autonomy with minimal additional learning effort. Based on feedback from the formative study, we also provide personalized options. BLV users can adjust the granularity of alternative text in the settings according to their access purpose, habits, table header orientation, and table content, whether it is by cell, row, or column. The output text will be organized according to user settings and combined with TableNarrator's analysis of the table structure. For example, when a BLV user seeks a comprehensive understanding of a particular piece of information, row-level granularity may be more suitable. Additionally, BLV users can customize whether table values are read before or after table headers, effectively addressing the issue of information redundancy highlighted in the formative study.

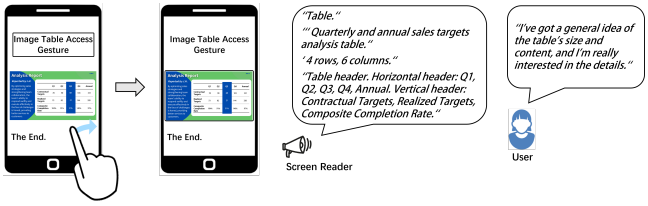


**Figure 3: Single-finger Enter. Upon detecting a table within the selected image, TableNarrator starts working. specifically, it extracts the table information and provides a brief description (table metadata).**

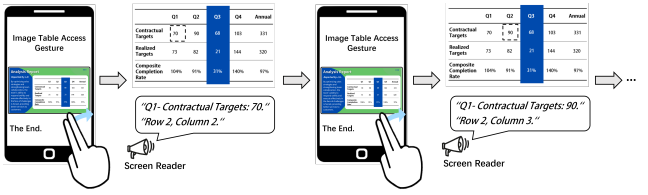


**Figure 4: Double-finger. After table metadata, the user switches using double-finger and TableNarrator accesses table content cell-by-cell. The dashed box represents virtual focus.**

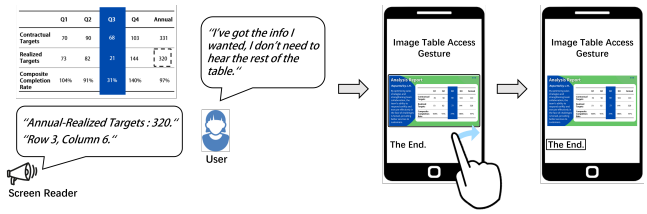


**Figure 5: Single-finger Exit. When the user intends to end the access, switching to single-finger, and the screen reader will focus on the content following the current image.**

## 4.5 Potential Limitations

There are some potential limitations among the two information extraction modules of TableNarrator.

**Table Layout Analysis Module.** One of the limitations of table layout analysis is the variability in table designs and structures. Tables can have different sizes, layouts, and formats, making it challenging to accurately detect and classify the three layouts. Designers may utilize a combination of image processing algorithms and machine learning models trained on diverse datasets to improve their accuracy and robustness in handling various table layouts.

**Table Structure Recognition Module.** One of the challenges in table structure recognition is the variability in table layouts and formats. Tables can have different numbers of rows and columns, varying cell sizes, and complex merging and splitting of cells, making it challenging to accurately extract the cell position information. Designers may use advanced algorithms for cell boundary detection, cell grouping, and row-column location inference to overcome these challenges. Another of the challenges is the variability and complexity of the table data. Tables can contain a wide range of information, including textual data, numerical data, dates, and other types of structured data, making it challenging to accurately extract and analyze key-value relationships. Designers may utilize advanced Natural Language Processing (NLP) models trained on diverse datasets to improve their accuracy and robustness in handling various types of table data.

# 5 Technical Evaluation

In this section, we conduct a technical evaluation of TableNarrator with some other table understanding systems and algorithms. Additionally, we provide error rates of TableNarrator against ground truths. The purpose of the technical evaluation is to demonstrate that the TableNarrator system can provide BLV users with comprehensive and accurate table information as we have pre-designed and intended. Specifically, we mainly evaluate the quality of two basic comprehending aspects of a given image table. 1) *Table Metadata*. Including the basic settings of a table such as whether the table exists, the number of rows and columns, the instructions of the table header, and the concise summary of the whole table. 2) *Table Content*. Including the detailed content such as the text in each cell and the relation between each cell.

## 5.1 Datasets

We collect 1000 image tables with diverse structural designs and different usages, including wired and wireless tables for e-commerce, school timetables, enterprise reports, etc. Considering the aforementioned comprehension aspects, we invite several experts and workers in these fields to label each table's information. Meanwhile, to avoid the potential comprehension problems happening to the BLV people, we then invite five BLV people and let them comprehend the labeled information of the image tables via screen reader. Therefore, we can guarantee the labels for the ground truth in technical evaluation.

## 5.2 Method

In this section, we select two powerful multimodal large language models as the main contrast methods. 1) *GPT-4V*. One of the most powerful multimodal large language models applied to the accessibility field for BLV people such as the Be My Eyes App. 2) *Claude3*. One of the most powerful large language models released by Anthropic and does well in table understanding. Since many domain-specific models can perform better than multimodal large language models in specialized fields, we also included six prevalent table recognition models for contrast. However, these models output only the most basic table information without considering readability, they are not suitable for direct use. We conducted separate evaluations of the model's performance in extracting pre-designed information.

## 5.3 Table Metadata Evaluation

We mainly evaluate the following aspects of table metadata.

1) *the Number of Rows and Columns Reminder*. The model's output of the number of rows and columns is compared with the ground truth to calculate accuracy. An image is considered correct only if both the row count and the column count are accurate. In this experiment, the prompt for two multimodal large language models is: *"Please output the number of rows and columns in the table shown in the image, in the format: number of rows-number of columns."*

2) *Table Headers Reminder*. The accuracy of table header recognition for each image is calculated based on the model's output. We calculate the proportion of correctly extracted table headers out of all table headers. In this experiment, the prompt for two multimodal large language models is: *"Please output the table headers in the table shown in the image. Output only the table headers."*

3) *Concise Table Summary*. This is evaluated by the ROUGE metric [35] on the summary results. In this experiment, the prompt for two multimodal large language models is: *"Please output a concise summary for the table shown in the image, noting that this will be provided to BLV people."* Additionally, we also conducted an anonymous manual evaluation. Specifically, we recruited 11 participants without visual impairment and randomly selected 30 tables from the dataset. Participants were asked to choose the option they believed best summarized the theme of the table and provided the best reading experience from the table summaries generated by GPT-4V, Claude3 and TableNarrator.

For GPT-4V, the percentage of both row and column counts being correct is 0.05%, while the percentage of either row or column being incorrect is 0.94%. Among these errors, row errors account for 0.60%, and col errors account for 0.87%. For Claude3, 0.19% of the outputs had both the correct number of rows and columns, while 0.82% had errors in either the row count or the column count. Among these errors, 0.58% were row wrong errors, and 0.86% were column wrong errors. For TableNarrator, 0.83% of the cases had both the correct number of rows and columns, while 0.17% had errors in either row or column count. Among these errors, 0.62% were row wrong errors and 0. 51% were col wrong errors. From the results, we find that both GPT-4V and Claude3 give wrong reminders of the number of rows and columns, which loses the coarse structural information of the table and may risk confusion in the table understanding. Meanwhile, we evaluate the accuracy of table header reminder. Figure 6 depicts the results via boxplots. TableNarrator performs better than the other two baselines.

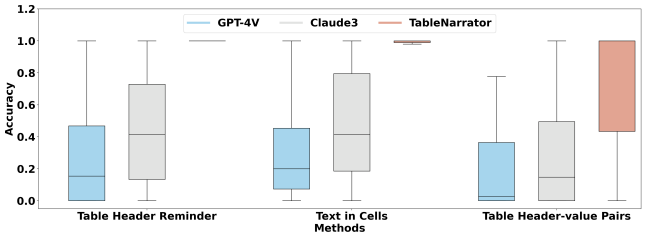


**Figure 6: The accuracy of the results from three methods on Table Header Reminder, Text in Cells and Table Header-value Pairs.**

Meanwhile, the quality of the table summary from TableNarrator is better than the other two baselines. Table 2 shows the ROUGE results. The other two baselines input image tables and apply vision-text processing algorithms to generate the summary, which focuses on the visual feature and may lose the semantics of text in cells. In contrast, TableNarrator uses the precise text in cells as model input instead of the image table, which strengthens the semantic feature and results in a more accurate summary. Furthermore, in manual evaluation, the selection rates for GPT-4V, Claude3 and TableNarrator are 12.00%, 21.14%, and 66.86%.

## 5.4 Table Content Evaluation

Furthermore, we evaluate the following aspects of the table content information. The first three metrics are calculated at the instance level with cell granularity, while the fourth metric is calculated at the single-image level with image granularity. The former is more indicative of the system's performance in recognizing table details, whereas the latter is more stringent and provides a better reflection of the overall performance of TableNarrator.

1) *Text in Cells.* This is evaluated by the 'cell accuracy' metric, which only considers a cell as correct if all the text in the cell is accurate.
2) *Table Header-Value Pairs.* This is similar to the evaluation method for text in cells and calculates the accuracy of the correct pairs.
3) *Cell Position.* This calculates the accuracy of cell positions, used to assess the module's capability in determining cell locations.
4) *Overall Evaluation.* This metric is compared to the ground truths, and a sample is considered correct only if all the output for a given image is accurate.

The results of Text in Cells and Table Header-Value Pairs are shown in Figure 6. The results of Cell Position are presented in Table 3, measured by multiple tasks of table recognition. From the results, we find that TableNarrator provides more precise content and cell relationships than the other two baselines, which is owed to the table structure recognition module. Additionally, Table 3 also compares the metrics of TableNarrator alongside domain-specific models. We selected tasks supported by all these models: table detection, cell relationship classification and cell logical location. The TableNarrator possesses enhanced capabilities due to our meticulous fine-tuning based on the characteristics of complex table data. For the Overall Evaluation, 90.16% of the images had correct Cell Logical Location for the whole image, while 91.72% of the images had correct Cell Relationship Classification for the whole image. And 85% of the images processed by TableNarrator have completely correct outputs.

# 6 User Study

## 6.1 Method

**Participants.** We recruited nine BLV participants (P1, P3, P4, P5, P7, P8, P9, P10, P11) for our user evaluation. Table 1 provides their specific information. Six of them had participated in our previous formative study. The ages of our nine participants ranged from 25 to 50 from various occupations. Six of them were familiar with the tables, while three were unfamiliar with tables. All participants were proficient in using screen readers. The evaluation was conducted both online and offline according to their preferences. Each session lasted approximately 40 minutes and we paid each participant a $20 reward for their work.

**Materials.** To cover the image tables with as many structures as possible, we collected image tables from representative styles of table structures (Figure 7): simple (T1), mered cells (T2), multi-level headers (T3), and cross headers (T4), which contained content with diverse logical semantics.

**Baseline Methods.** Each image table was evaluated using TableNarrator and two main baseline methods: GPT-4V and VoiceOver. GPT-4V represents a multimodal large language model with substantial influence, functioning as a virtual volunteer within the widely used software Be My Eyes. Our prompt was set as: *"Please generate a comprehensive and accurate description of the table in the image, noting that this will be provided to BLV people."* VoiceOver is a prevalent screen reader equipped with text recognition and image captioning capabilities, supplied by Apple for its range of devices. Furthermore, we also provided manually converted HTML code for each image table to compare user experience.

**Procedure.** We developed a prototype system to simulate the workflow of TableNarrator, which was used to conduct user evaluation with BLV people. We started our process with an overview of the background information, followed by an inquiry into user experiences when accessing tables. Subsequently, we spent 10 minutes instructing participants on how to use TableNarrator. Each participant was provided with four image tables as previously described. After each table was accessed, we conducted a semi-structured interview to gather feedback. After completing all four examples, we administered a System Usability Scale (SUS) and a Workload Assessment (WA) to evaluate the user experience. The SUS scale, reliable even with small sample sizes, consisted of 10 questions evaluated on a 5-point Likert scale. The workload assessment encompassed performance, mental demand, effort, and frustration level, rated on a 7-point Likert scale. Our methodology was specifically designed to ascertain whether TableNarrator fulfilled the four design goals established in our formative research.

**Analysis.** Following the evaluation, we transcribed the semi-structured interviews and conducted thematic analysis. Two researchers independently generated initial codes on the data and collaboratively reviewed and refined the themes to ensure they accurately captured the participants' feedback and were distinct from each other. Ultimately, we summarized the data collected from user evaluation and semi-structured interview into three categories:

**Table 2: The ROUGE of the content summary. We calculate the Precision(P), Recall(R) and F1-score(F) of ROUGE-1, ROUGE-2 and ROUGE-L to evaluate the quality of the table summary.**

| Method | ROUGE-1 | | | ROUGE-2 | | | ROUGE-L | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | **F** | P | R | **F** | P | R | **F** |
| GPT-4V | 0.263 | 0.267 | 0.246 | 0.068 | 0.076 | 0.070 | 0.256 | 0.261 | 0.240 |
| Claude3 | 0.244 | 0.217 | 0.209 | 0.080 | 0.094 | 0.085 | 0.240 | 0.212 | 0.205 |
| **TableNarrator** | **0.385** | **0.373** | **0.367** | **0.124** | **0.122** | **0.117** | **0.376** | **0.369** | **0.360** |

**Table 3: Comparison of TableNarrator's performance with domain-specific models in table recognition tasks, covering Table Detection, Cell Relationship Classification, and Cell Logical Location tasks, with F1 score used as the evaluation metric.**

| Method | Task | | |
|---|---|---|---|
| | Table Detection | Cell Relationship Classification | Cell Logical Location |
| Deep Split+Merge [57] | 0.6851 | 0.7921 | 0.3456 |
| LGPMA [48] | 0.7611 | 0.82 | 0.6702 |
| TGRNet [68] | 0.6953 | 0.4236 | 0.2984 |
| LORE [67] | 0.8143 | 0.8185 | 0.8904 |
| TableMaster [69] | 0.7499 | - | - |
| CTUNet [34] | 0.8904 | 0.8934 | 0.7542 |
| **TableNarrator** | **0.9851** | **0.9908** | **0.9607** |

1) The effectiveness of the processed table information; 2) The assessment of our interaction mode and personalized options; and 3) The overall system usability.

## 6.2 Results

During the user evaluation, all participants unanimously agreed that among the four tools provided (VoiceOver, GPT-4V, manual HTML format, TableNarrator), our system outperformed the rest. The performance of the manual HTML format and GPT-4V were rated second and third, and there was unanimous disapproval for the descriptions generated by the Screen Reader.

**The effectiveness of processed table information.** In terms of intelligibility of the information, all participants concurred that the alternative text provided by TableNarrator was *"No difficulties in comprehending the table content"* and expressed confidence in the comprehensiveness and accuracy of the conveyed information, trusting the descriptions provided by TableNarrator. Regarding the initial provision of the number of rows and columns in a table, P8 considered it *"indeed necessary."* P5 echoed this sentiment, stating *"It makes the scale of the table clear."* About the concise table summary, P3 pointed out that knowing this information enables one to *"decide what to do with the table,"* while P8 underscored the significance of table summary, describing it as *"necessary, definitely necessary."* The ability to quickly identify the table's structure and subject was seen as crucial by several participants, as it provides users with an overview of the table and helps them form a mental model of the table, aiding in decision-making process regarding subsequent actions.

Beyond simply providing this basic structural information, many participants found that understanding the relationship between table headers and cell values, in a format that prioritized semantic meaning over spatial layout, reduced cognitive load. As reported by many participants, comprehending tables often necessitates the mental construction of their grid structure, a process they found to be quite *"energy-consuming"*. Our solution, TableNarrator, converts spatial structural information, as visually presented, into cell-level header-value pairs. P9 described it as: *"For totally blind colleagues, ..., very friendly, ..., I'm indifferent to this or that structure, ... just listening to the text."* P7 also highly praised this type of presentation and suggested: *"I hope this feature can be extended to Excel; I need to hear the headers while filling out tables."* When we asked participants to compare TableNarrator with the manual HTML format, P11 said, *"The former (TableNarrator) is definitely better. It doesn't require memorization."* Considering that many users are accustomed to mentally arranging a grid of tables, TableNarrator added row and column numbers after each table header-value pair. P5 believed: *"Row and column numbers help establish the structure of the table, and being of the least priority, placing them last in terms of the importance of the information circumvents redundancy."* The participants broadly approved the above basic information about tables and the recitation of the table header-value correspondence, as P1 describes it as: *"just as it should be."*

As for the description texts generated by GPT-4V and Screen Readers, participants did not rate them highly. For the least appreciated screen reader, P10 stated: *"Despite the words being read out, I can only guess the table through my imagination..."* P10 commented: *"I was confused and couldn't react immediately."* In comparison, GPT-4V's descriptions were considered too general. While P3 found it helpful when only a rough idea of the table was needed, P1 found the descriptions *"rudimentary and overly general,"* making it challenging to pinpoint specific details within the table. This feedback reflected a broader concern that AI-generated information were often too vague or error-prone, as noted by P3: *"It always baffles*

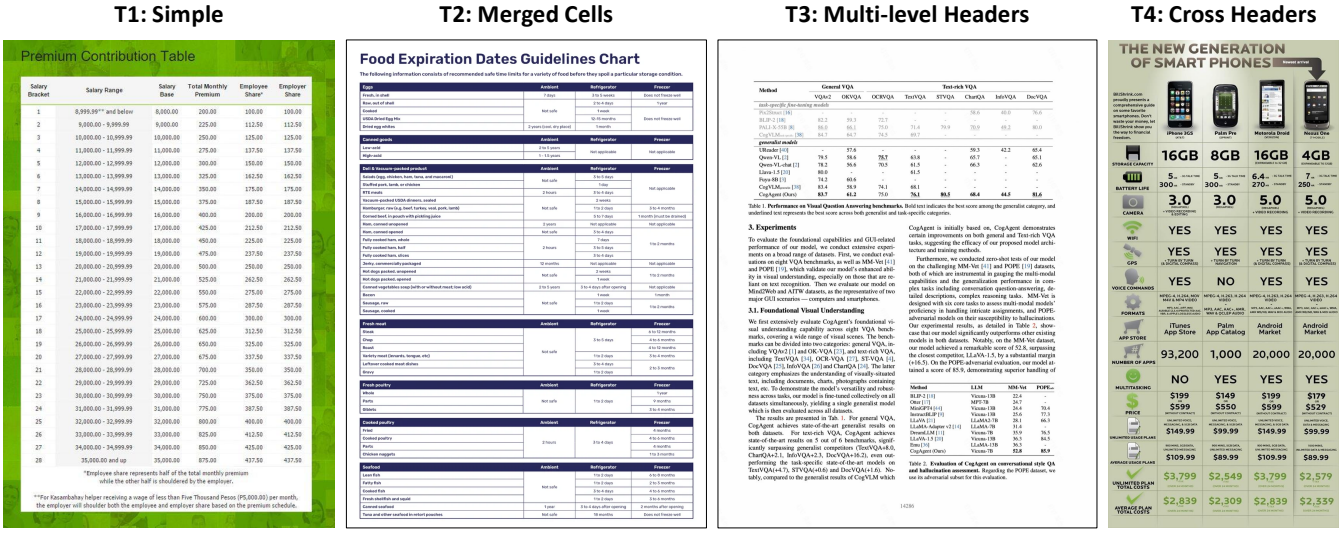


**Figure 7: The four representative real-world image tables we selected. They are categorized based on table structures into: Simple, Merged Cells, Multi-level Headers, and Cross Headers.**

*you at the most unexpected points; it keeps making mistakes where you think it shouldn't."*

**The evaluation of the interaction mode and personalized options.** Our participants gave a higher rating to the interaction mode of TableNarrator compared to the baseline methods. Concerning the convenience of information retrieval, all participants chose to read the short table metadata first and then scroll by cell. P1 reported that TableNarrator enabled them to find the specific information they needed more rapidly, stating: *"I can quickly skim through, stopping to listen only if I need to."* In contrast, GPT-4V's verbosity posed a challenge for users. P7 commented, *"You can't select a focus, and if you don't catch something, you must start over from the beginning."* P1 further critiqued GPT-4V's efficiency, stating: *"too verbose, ..., to understand this table, I have to listen to the whole thing earnestly,"* which many found *"time-consuming"*. This was an area where TableNarrator excelled, offering users a more focused, task-oriented way of interacting with table data. As for the manual HTML format, participants unfamiliar with tables also reported *"effort-related issues."* P4 emphasized again that imagining the table's layout is not easy for some individuals, especially when dealing with long tables and complex merged cells.

Participants also expressed appreciation for TableNarrator's personalized options, which allowed them to adjust the granularity of table information. P5 adjusted the granularity from cell-by-cell to group-by-group and found this method especially useful for comparing different data sets. P3, when discussing the customization of focus areas, added: *"You can use a focus box..., going through text at once..., otherwise, if the focus shifts too frequently, my hands will get tired."* This feedback illustrated the importance of allowing users to tailor the system to their own needs and preferences. When it comes to swapping table header-value pair reading order, there is no agreement on whether to read table header first or value first. P4 asserted that *"It depends on user habit."* P11 said *"If I access the scale tables for clothes or shoes, I'd better read the value first because I can quickly know the proper scales."* P7 thought *"I prefer to obtain the headers first thus I can comprehend the relation of cells clearly."* This highlighted that user preferences can vary significantly based on the task at hand and their familiarity with different table structures. Overall, most users prefer the table headers first when they want to comprehend the structure. Meanwhile, users perfer the values first when they want to search for detailed information.

**The usability of the entire system.** TableNarrator's overall usability was evaluated using the System Usability Scale (SUS), yielding an average score of 90.6, falling within the "Excellent" range (85-99). This high rating was supported by participants' qualitative feedback, which consistently praised the system's Learnability(88), Usability(89.6) and Satisfaction(94.4). P8 commented: *"very intuitive, and using this tool felt great to me."* However, the learning curve was a minor point of critique. P3 acknowledged that *"There's definitely a learning curve at the beginning,"* although the general consensus was that TableNarrator could be quickly mastered with minimal training. P2 added: *"There's no need for additional study; there's nothing much to learn, most people will pick it up quickly."*

The workload assessment also provided insights into the system's efficiency. Figure 8 displays participants' workload assessment using four tools, to compare the performance, mental demand, effort, and frustration level of TableNarrator and baseline methods. The mean scores for TableNarrator were 6, 5.78, 5.67, and 6.44 (out of a maximum of 7), reflecting the participants' recognition of its operational efficiency and the lower mental strain, effort, and frustration experienced while experiencing the system. This is particularly notable considering that the baseline methods (GPT-4V and VoiceOver) were perceived as requiring more effort and leading to higher frustration levels.

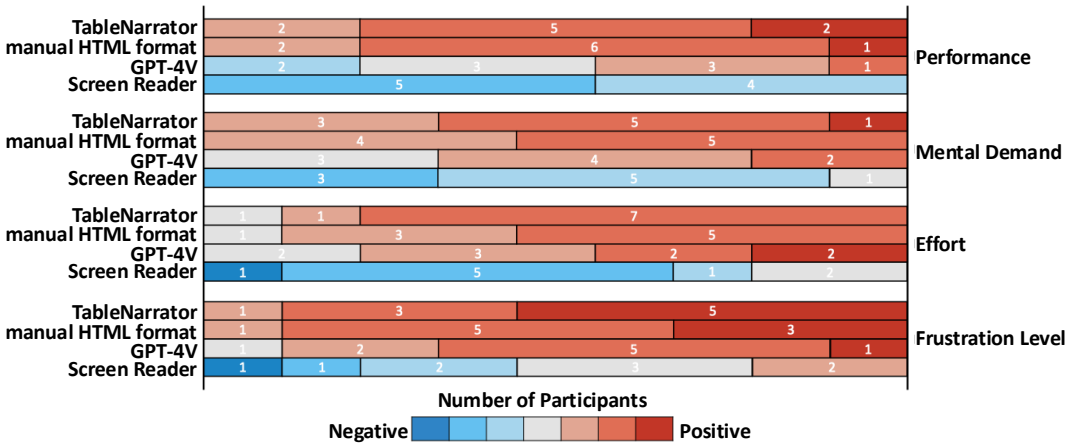


**Figure 8: Overall workload assessment of TableNarrator, manual HTML format, GPT-4V and Screen Reader by participants. Best viewed on screen.**

## 7 Discussion

In this section, we revisit the design goals proposed in our formative study and discuss the performance and limitations of TableNarrator based on the evaluation results. We also discussed a series of valuable insights gained during the design and development of TableNarrator, and analyzed the potential and limitations of multimodal large language models in the accessibility of image tables.

### 7.1 Limitations of TableNarrator When Encountering Various Types of Tables

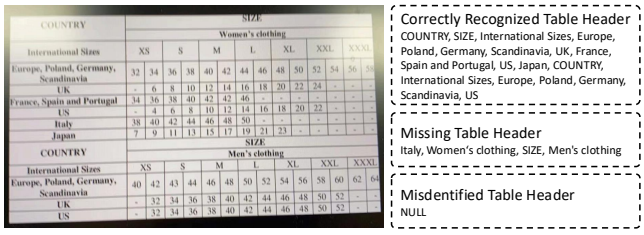


**Figure 9: Results of a photographic table image recognized by TableNarrator. Errors are concentrated in the classification of the table headers.**

*7.1.1* ***Tables in photographic images.*** According to both technical and user evaluations, while TableNarrator delivers excellent results on Internet image tables, they are more prone to errors when dealing with tables in photographic images. In particular, inherent aberrations such as camera shake, distortion, and skew introduce additional analytical complexities. Preliminary evaluations with photographic table images indicate that TableNarrator lacks the robustness necessary to provide precise referential support for BLV users. Figure 9 illustrates an example of this. This is a critical issue for table accessibility in real-world scenarios, such as the real shot of product manuals or instructions for public facilities. However, current research usually focuses on manually designed image tables, such as advertisements on social media. In the future, we aim to evolve TableNarrator into a universally adaptable table narration tool. This will be achieved by embracing a wide spectrum of image tables, augmenting our training datasets with photographic samples, and refining the underlying visual perception model.

*7.1.2* ***Tables with complex layouts.*** Although TableNarrator can effectively recognize and understand table structure information in a majority of cases, the complexity of table layouts in real-world scenarios can easily result in misclassification and misrecognition of cells, leading to an incorrect semantic understanding of the table as a whole. Figure 10 illustrates how such errors can affect the overall usability of a table representation. Typical complexities in layout include the absence of clear boundaries between cells (unlike the comma-separated format of CSV), and merged cells where one key corresponds to multiple values or vice versa. Moreover, TableNarrator is subject to the risk of error accumulation over multiple stages. The entire process involves several steps from recognizing table structures to classifying cells and their content. Errors at any stage can be compounded as the process progresses. While erroneous recognition of individual cells may not hinder the overall understanding of the table for sighted users, it becomes exceedingly challenging for BLV users to identify potential errors via one-dimensional alternative text. Therefore, to prevent significant misunderstandings due to incorrect recognition, systems designed for image table access can not only improve the accuracy and robustness of table recognition and understanding, but also address issues promptly by increasing multi-round interactions with users, such as through a QA module based on large language models.

### 7.2 Deeper Understanding of User Needs

*7.2.1* ***Elderly users needs.*** While TableNarrator is primarily designed to provide alternative text for BLV users, it's worth noting that certain demographics, such as the elderly, may primarily struggle with reading and identifying text in smaller font sizes, even though they can recognize and understand the overall structure of tables. Therefore, they may not need to use alternative text to interact with tables. However, image tables often remain at a fixed size

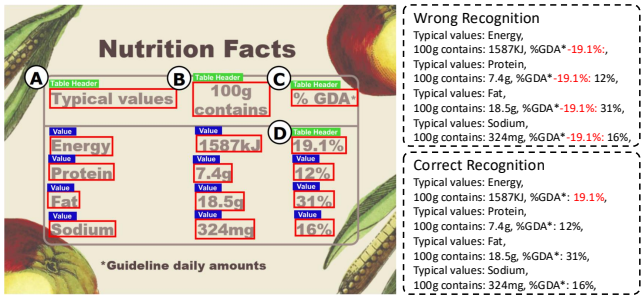


**Figure 10: Impact of detail errors on TableNarrator's output. The visual model misclassifies D as a header alongside correct headers A, B, and C. Textual narration creates confusion by sequentially vocalizing cells wrongly identified as headers.**

and lack the capability to adjust font size to accommodate the visual requirements of the elderly. Recognizing this, the application scope of TableNarrator can be expanded to serve this user demographic by providing them with the ability to understand and access the textual content in images of tables. This represents a significant expansion from its original focus, addressing the needs of a broader audience.

*7.2.2* ***Different information processing depth needs.*** During the process of accessing table data, TableNarrator allows users to select varying degrees of alternative text granularity according to their needs. Thus, TableNarrator can provide an appropriate information-conveying strategy whether the user is quickly skimming through data or engaging in an in-depth study. However, TableNarrator does not encompass all user needs, as many of their requirements fall outside the operational scope of TableNarrator, such as the sorting function and the QA conversation feature, necessitating collaboration with other works. Such requirements imply a deeper level of information processing. Although the current design of TableNarrator ensures the autonomy of BLV users in accessing information, there are times when more deeply processed information is also needed. Therefore, a better option is to provide information at varying levels of processing for users to choose from. In the future, we will endeavour to further refine table-type classification based on multidimensional metrics including data characteristics, table structure, and user interaction patterns. This will enable TableNarrator to adapt to a broader range of usage scenarios and provide more precise and personalized data processing services.

*7.2.3* ***User-friendly interaction.*** In our formative study, the frequently highlighted issues are related to user-friendly interaction. Firstly, the redundant alternative text has brought a negative user experience, where the user-friendly formats are usually concise and precise. Our system addresses these concerns by offering several personalization options, allowing users to alter the presentation of information according to their specific needs. Users can customize the system to first announce the name followed by headings, enabling them to skip to the next piece of data after hearing the name. This design philosophy reduces the broadcast of redundant information. Additionally, for tables with straightforward and comprehensible headings (e.g., name, sex), users may opt for this approach as well. Secondly, feedback from two BLV users suggested a desire for a more natural and emotive voice interaction, rather than mechanical. This has prompted us to consider how to foster an emotional connection between technology and users, promoting Human-centered computing. In the future, we aim to achieve this by adopting more advanced voice technologies. This shift from merely fulfilling functional needs to enhancing experiences will propel assistive technologies towards more lively and personalized directions, ultimately fostering a deeper connection with users.

## 7.3 Image Table Accessibility Meets Large Language Models

During our research, our BLV participants repeatedly highlighted the impact of AI technologies on their lives, particularly in areas such as intent understanding and user-friendly interaction simulations. With the rapid development of large language models, we are confident that these technologies hold great promise for improving the accessibility of image tables.

However, as previously discussed, present large language models still face significant challenges, including understanding tables in various languages and reasoning about intricate table relationships. While existing multimodal large language models have demonstrated potential for enhancing image table accessibility, we were surprised to find that performance can vary dramatically when processing tables in various languages. This discrepancy stems from the inherent limitations of language biases in the training data of large language models and highlights a critical inequity for BLV users from different linguistic backgrounds, potentially limiting the global adoption of assistive tools. Another critical problem that deserves attention is the "hallucination" issue of large models, where they sometimes generate information that is unrelated or even incorrect in relation to the actual data. Such instances could seriously impact BLV users' understanding of information and their trust in AI tools. Incorporating appropriate supervision and verification mechanisms into the system can help mitigate this issue.

As such, TableNarrator ensures the efficacy of each step to the greatest extent possible by amalgamating multiple steps, rather than relying on the multimodal large language model to generate results. Our technical and user evaluations have further substantiated the effectiveness of this approach. Despite these limitations, we maintain that with the rapid evolution of large language models, more steps could be supplanted in the future. In summary, while large-scale models have shown clear advantages in enhancing the accessibility of image tables and images, it is still necessary to carefully evaluate and improve upon their limitations, including deep consideration for fairness and reliability, to ensure that the BLV community from varied linguistic and cultural backgrounds can equitably benefit from the progress in accessibility technologies.

## 8 Conclusion

In this paper, we introduce TableNarrator, an innovative system designed to enhance the accessibility of image tables. This system has found extensive application in social media and online education contexts. Guided by our formative study, we gathered

invaluable suggestions regarding the accessibility of the image tables from BLV users. In compliance with the design objectives, we present TableNarrator, a system composed of several meticulously designed modules to extract tables regions, comprehend the table structure, and convey table information appropriately. We also offer personalized options and a simple and direct interaction mode that enables users to access image tables. Both technical and user evaluations demonstrate the effectiveness of TableNarrator. In the future, our aim is to further explore tables with multilingual content and complex layouts to enhance the experience for diverse user groups.

## Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Grant No.62372408).

## References

[1] 2023. GPT-4V(ision) System Card. https://api.semanticscholar.org/CorpusID:263218031

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

[3] Angelique Boekelder and Michael Steehouder. 2015. Selecting and switching: some advantages of diagrams over tables and lists for presenting instructions. *Writing and Speaking in the Technology Professions: A Practical Guide* (2015), 161–173.

[4] Yevgen Borodin, Jeffrey P Bigham, Glenn Dausch, and IV Ramakrishnan. 2010. More than meets the eye: a survey of screen-reader browsing strategies. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*. 1–10.

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[6] Justin R Brown, Stacy A Doore, Justin K Dimmel, Norbert Giudice, and Nicholas A Giudice. 2023. Comparing Natural Language and Vibro-Audio Modalities for Inclusive STEM learning with blind and low vision users. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–17.

[7] Matthew Butler, Leona M Holloway, Samuel Reinders, Cagatay Goncu, and Kim Marriott. 2021. Technology developments in touch-based accessible graphics: A systematic review of research 2010-2020. In *Proceedings of the 2021 Chi conference on human factors in computing systems*. 1–15.

[8] Xiang Deng, Ahmed Hassan Awadallah, Christopher Meek, Oleksandr Polozov, Huan Sun, and Matthew Richardson. 2020. Structure-grounded pretraining for text-to-sql. *arXiv preprint arXiv:2010.12773* (2020).

[9] Julian Martin Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. *arXiv preprint arXiv:2010.00571* (2020).

[10] David W. Embley, Cui Tao, and Stephen W. Liddle. 2005. Automating the extraction of data from HTML tables with unknown structure. *Data Knowl. Eng.* 54 (2005), 3–28. https://api.semanticscholar.org/CorpusID:5406402

[11] Christin Engel and Gerhard Weber. 2017. Analysis of tactile chart design. In *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments*. 197–200.

[12] Nosheen Fayyaz, Shah Khusro, and Shakir Ullah. 2021. Accessibility of Tables in PDF Documents: Issues, Challenges and Future Directions. *Information Technology and Libraries* 40, 3 (2021).

[13] Yansong Feng and Mirella Lapata. 2012. Automatic caption generation for news images. *IEEE transactions on pattern analysis and machine intelligence* 35, 4 (2012), 797–812.

[14] Brian Frey, Caleb Southern, and Mario Romero. 2011. Brailletouch: mobile texting for the visually impaired. In *Universal Access in Human-Computer Interaction. Context Diversity: 6th International Conference, UAHCI 2011, Held as Part of HCI International 2011, Orlando, FL, USA, July 9-14, 2011, Proceedings, Part III 6*. Springer, 19–25.

[15] Basilios Gatos, Dimitrios Danatsas, Ioannis Pratikakis, and Stavros J. Perantonis. 2005. Automatic Table Detection in Document Images. In *International Conference on Advances in Pattern Recognition*. https://api.semanticscholar.org/CorpusID:11485240

[16] Azka Gilani, Shah Rukh Qasim, Imran Malik, and Faisal Shafait. 2017. Table detection using deep learning. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, Vol. 1. IEEE, 771–776.

[17] Cagatay Goncu and Kim Marriott. 2015. Creating eBooks with Accessible Graphics Content. 89–92.

[18] Stanislav Gyoshev, Dimitar Karastoyanov, Nikolay Stoimenov, Virginio Cantoni, Luca Lombardi, and Alessandra Setti. 2018. Exploiting a graphical Braille display for art masterpieces. In *Computers Helping People with Special Needs: 16th International Conference, ICCHP 2018, Linz, Austria, July 11-13, 2018, Proceedings, Part II 16*. Springer, 237–245.

[19] Michael E Hahn, Corrine M Mueller, and Jenna L Gorlewicz. 2019. The comprehension of stem graphics via a multisensory tablet electronic device by students with visual impairments. *Journal of Visual Impairment & Blindness* 113, 5 (2019), 404–418.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[21] Leona Holloway, Swamy Ananthanarayan, Matthew Butler, Madhuka Thisuri De Silva, Kirsten Ellis, Cagatay Goncu, Kate Stephens, and Kim Marriott. 2022. Animations at your fingertips: using a refreshable tactile display to convey motion graphics for people who are blind or have low vision. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–16.

[22] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)* 51, 6 (2019), 1–36.

[23] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems* 51, 1 (2011), 141–154.

[24] Noman Islam, Zeeshan Islam, and Nazia Noor. 2017. A survey on optical character recognition system. *arXiv preprint arXiv:1710.05703* (2017).

[25] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 1827–1843.

[26] Nicole E Johnson. 2023. *Making Tactile Pictures More Available: Techniques & Communities*. Ph. D. Dissertation. University of Colorado at Boulder.

[27] Shakila Cherise S Joyner, Amalia Riegelhuth, Kathleen Garrity, Yea-Seul Kim, and Nam Wook Kim. 2022. Visualization accessibility in the wild: Challenges faced by visualization designers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19.

[28] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. 2013. ICDAR 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*. IEEE, 1484–1493.

[29] Natalie Kerby. [n. d.]. Infrastructuring Vision: An Examination of Automatic Alt Text on Facebook. ([n. d.]).

[30] Elvis Koci, Dana Kuban, Nico Luettig, Dominik Olwig, Maik Thiele, Julius Gonsior, Wolfgang Lehner, and Oscar Romero. 2019. XLIndy: Interactive Recognition and Information Extraction in Spreadsheets. *Proceedings of the ACM Symposium on Document Engineering 2019* (2019). https://api.semanticscholar.org/CorpusID:202728714

[31] Jonathan Lazar, Suranjan Chakraborty, Dustin Carroll, Robert Weir, Bryan Sizemore, and Haley Henderson. 2013. Development and Evaluation of Two Prototypes for Providing Weather Map Data to Blind Users Through Sonification. *Journal of Usability Studies* 8, 4 (2013).

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[34] Zaisheng Li, Yi Li, Qiao Liang, Pengfei Li, Zhanzhan Cheng, Yi Niu, Shiliang Pu, and Xi Li. 2022. End-to-End Compound Table Understanding with Multi-Modal Modeling. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4112–4121.

[35] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:964287

[36] Weihong Lin, Zheng Sun, Chixiang Ma, Mingze Li, Jiawei Wang, Lei Sun, and Qiang Huo. 2022. Tsrformer: Table structure recognition with transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*. 6473–6482.

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[38] Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2021. Tapex: Table pre-training via learning a neural sql executor. *arXiv preprint arXiv:2107.07653* (2021).

[39] Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[40] Daniel Lopresti and George Nagy. 1999. A tabular survey of automated table processing. In *International Workshop on Graphics Recognition*. Springer, 93–120.

[41] Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *Comput. Surveys* 55, 5 (2022), 1–37.

[42] Shunji Mori, Hirobumi Nishida, and Hiromitsu Yamada. 1999. *Optical character recognition*. John Wiley & Sons, Inc.

[43] Shunji Mori, Ching Y Suen, and Kazuhiko Yamamoto. 1992. Historical review of OCR research and development. *Proc. IEEE* 80, 7 (1992), 1029–1058.

[44] OpenAI and Josh Achiam et. al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[45] Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373* (2020).

[46] KR Prajwal, CV Jawahar, and Ponnurangam Kumaraguru. 2019. Towards increased accessibility of meme images with the help of rich face emotion captions. In *Proceedings of the 27th ACM International Conference on Multimedia*. 202–210.

[47] Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. 2020. CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 572–573.

[48] Liang Qiao, Zaisheng Li, Zhanzhan Cheng, Peng Zhang, Shiliang Pu, Yi Niu, Wenqi Ren, Wenming Tan, and Fei Wu. 2021. Lgpma: Complicated table structure recognition with local and global pyramid mask alignment. In *International conference on document analysis and recognition*. Springer, 99–114.

[49] Ramana Rao and Stuart K Card. 1994. The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 318–322.

[50] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems* 32 (2019).

[51] Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. PubTables-1M: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4634–4642.

[52] 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*. 1–13.

[53] 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* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376404

[54] Abigale J Stangl, Esha Kothari, Suyog D Jain, Tom Yeh, Kristen Grauman, and Danna Gurari. 2018. Browsewithme: An online clothes shopping assistant for people with visual impairments. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 107–118.

[55] Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards table-to-text generation with numerical reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1451–1465.

[56] Delve Team. 2022. *Delve Tool*. https://delvetool.com/

[57] Chris Tensmeyer, Vlad I Morariu, Brian Price, Scott Cohen, and Tony Martinez. 2019. Deep splitting and merging for table structure decomposition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 114–121.

[58] Tejal Tiwary and Rajendra Prasad Mahapatra. 2023. An accurate generation of image captions for blind people using extended convolutional atom neural network. *Multimedia Tools and Applications* 82, 3 (2023), 3801–3830.

[59] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 319–326.

[60] W3C. 2008. *WCAG 2.0 Layers of Guidance*. https://www.w3.org/TR/WCAG20/#intro-layers-guidance

[61] W3C Web Accessibility Initiative (WAI). 2023. *Complex Images in Images Tutorial*. https://www.w3.org/WAI/tutorials/images/complex/

[62] Howard Wainer. 1992. Understanding graphs and tables. *ETS Research Report Series* 1992, 1 (1992), 4–20.

[63] Hélène Walle, Cyril De Runz, Barthélemy Serres, and Gilles Venturini. 2022. A survey on recent advances in AI and vision-based methods for helping and guiding visually impaired people. *Applied Sciences* 12, 5 (2022), 2308.

[64] Jingjing Wang, Haixun Wang, Zhongyuan Wang, and Kenny Q Zhu. 2012. Understanding tables on the web. In *Conceptual Modeling: 31st International Conference ER 2012, Florence, Italy, October 15-18, 2012. Proceedings 31*. Springer, 141–155.

[65] Yanan Wang, Ruobin Wang, Crescentia Jung, and Yea-Seul Kim. 2022. What makes web data tables accessible? Insights and a tool for rendering accessible tables for people with visual impairments. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–20.

[66] R Michael Winters, Neel Joshi, Edward Cutrell, and Meredith Ringel Morris. 2019. Strategies for auditory display of social media. *ergonomics in design* 27, 1 (2019), 11–15.

[67] Hangdi Xing, Feiyu Gao, Rujiao Long, Jiajun Bu, Qi Zheng, Liangcheng Li, Cong Yao, and Zhi Yu. 2023. LORE: logical location regression network for table structure recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 2992–3000.

[68] Wenyuan Xue, Baosheng Yu, Wen Wang, Dacheng Tao, and Qingyong Li. 2021. Tgrnet: A table graph reconstruction network for table structure recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1295–1304.

[69] Jiaquan Ye, Xianbiao Qi, Yelin He, Yihao Chen, Dengyi Gu, Peng Gao, and Rong Xiao. 2021. PingAn-VCGroup's solution for ICDAR 2021 competition on scientific literature parsing task B: table recognition to HTML. *arXiv preprint arXiv:2105.01848* (2021).

[70] Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. Grappa: Grammar-augmented pre-training for table semantic parsing. *arXiv preprint arXiv:2009.13845* (2020).

[71] Richard Zanibbi, Dorothea Blostein, and James R Cordy. 2004. A survey of table recognition: Models, observations, transformations, and inferences. *Document Analysis and Recognition* 7 (2004), 1–16.

[72] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2017. The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.

[73] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*. Springer, 564–580.

[74] Yu Zhong, Walter S Lasecki, Erin Brady, and Jeffrey P Bigham. 2015. Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2353–2362.

[75] Ying Zhong, Masaki Matsubara, and Atsuyuki Morishima. 2018. Identification of important images for understanding web pages. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 3568–3574.