Leveraging Spatial Audio Feedback and Input Flexibility to Enhance BLV Users' Navigation of Virtual Environments
Final Report
By: Anna-Samsara Daefler, Demetria Mack (Group name: SpaceInt)
avd2131 , dmm2285

Synopsis

Our project aims to improve the accessibility of virtual environments for blind and low-vision (BLV) users by building and evaluating key extensions to the CEAL Lab's open-source WebNExt tool. Our primary focus was to improve both auditory feedback and input flexibility, two areas critical to a more seamless and empowering web navigation experience for BLV users. Our work can be found at https://github.com/demetriam/6156-Final-Project.

First, we investigated whether adding contextual audio cues, such as alerts when reaching visual edges or skipping over blank regions, can strengthen a user's sense of spatial orientation while browsing. Second, we explored the use of lightweight AI models, including GPT-4's vision capabilities, to generate meaningful image descriptions when alt text is missing. These additions will be evaluated for both their helpfulness and potential to distract users.

Next, we discussed how a speech-command interface can reduce the cognitive burden typically associated with navigating keyboard-driven screen readers. By enabling natural language commands like "go to the first heading" or "find the search bar," we aim to make navigation more intuitive and user-friendly.

Research Questions

RQ 1. Can additional audio feedback around visual concepts provide a more seamless experience using WebNExt?

1.1. How does building in additional spatialized audio feedback to WebNExt around visual concepts affect BLV user's sense of orientation?

Currently, WebNext renders the text of a webpage into spatial audio. However, navigating visual content comprises more than just the text on the page. Sighted users react to visual effects, animations, and the layout of the webpage itself. To this effect, we implemented spatial audio feedback when directional navigation hits the visual edge of a web page and when navigating past large blank regions. These features were implemented as an extension to the WebNExt screen reader (https://github.com/ColumbiaCEAL/spatial-interactions-extension) and can be found at https://github.com/avd2131/coms6156_final_project. These features are toggleable in the extension's settings.

A. *Hitting the visual edge of a web page*

WebNext does not provide any feedback when users try to navigate past the final element of a web page. Participants in the WebNext study expressed that audio feedback would be useful, especially in instances when the final element in a direction is not at the screen's edge. One participant expressed frustration over navigating left or right from the search bar on the Google homepage and getting no response or feedback even though the element is not at the screen boundary (see Figure 1 in the appendix). We used a beep to indicate that the user was attempting to navigate past the final element on the screen and spatialized the beep to convey the spatial positioning of the boundary that was being hit.

B. *Navigating past large blank regions*

Participants in the WebNext study reported feeling disconcerted when spatial browsing would make a large pitch or panning shift. One participant in the study reported: "Sometimes, using the arrow keys caused me to jump from the top section all the way to the bottom, which made me wonder whether I

was skipping over elements located in the middle" [2]. We used a spatialized clicking sound to indicate navigation over a large blank region. In WebNext, navigating directionally with the arrow keys triggers a loop that moves 15px in the given direction and checks for an element at that point. We added a click for every 10 steps (150px) moved in a given direction to look for an element and spatialized the sound at the current position where we are checking for an element.

C. *Evaluation*

A demo is shown in this video https://www.youtube.com/watch?v=eTPocgtyr7g. We led a structured evaluation session with ophthalmology resident Dr. Keale Cade at Howard University Hospital. The recording of this session is available in the appendix. Dr. Cade found that the modifications enhance the baseline by indicating to the user, in the first case, that they are hitting a key that will not give a response and, in the second case, providing audio-based feedback regarding the layout of the webpage.

Dr. Cade noted that modification A could alleviate the frustration that BLV users may feel when hitting a key and not getting a response, as is the case in the baseline. BLV users would feel confident that they have reached the end of the page.

Modification B, Dr. Cade noted, would be useful for people who have some vision but can not make out small details. She presented the case that when websites have different sizes of text, low-vision web-users will often report being concerned that they are missing something that is too small for them to see. This feature would let them know that there is nothing in a blank region that is too small for them to make out. Dr. Cade also observed that modification B would help BLV web users understand the structure of a site. She brought up https://www.apple.com/, ,which is formatted lengthwise with plenty of blank space and involves scrolling far down. She pointed out that this feature would let the user know that they have to keep scrolling.

Dr. Cade felt that the sounds were distinct enough from each other, such that the meaning of them is clear. She notes that these modifications are not distracting but would require users to pay close attention to sound, which increases the cognitive load of browsing a webpage due to having to interpret additional audio cues. In her experience, learning to use new technologies is not easy for certain patient populations and can lead to frustration. One suggestion that she had was that if a user repeatedly hits the end of a webpage to add verbal feedback, such as "end of page", in a pitch that would distinguish it from the rest of the webpage contents. This feature could be toggleable and would reduce the learning curve.

1.2. Can meaningful image descriptions (e.g., when alt text is missing) be generated using lightweight AI models? How do vision models compare in generating alt text?

Alternative text is a valuable tool for making content accessible to users who are blind or have low vision when they cannot access images, GIFs, or other visual elements on screens. In today's world, where much of the content we engage with is presented on screens, it is essential that devices include tools capable of producing and, in some cases, even "speaking" alternative text across websites, with these features integrated into operating systems.

Although alternative text has evolved into an essential requirement for promoting web accessibility, it was not initially designed for this purpose. When first introduced in 1993, its goal was to provide text descriptions for images on websites that could not render images. In 1999, it became a requirement for the img and area tags in HTML 4.01 [11]. Over the years, this requirement has been adopted by many web pages and online applications as an essential tool, working in conjunction with screen readers to enable blind and low-vision users to access online environments regardless of their physical abilities.

Despite its success in bridging digital barriers for people with disabilities, many have noted that alternative text and algorithms used to generate and read it do not effectively access or determine the meaning of the visual components within the context of the rest of the webpage. This can cause confusion for those who rely on it. Many models are becoming increasingly better at identifying what is in an image and how it relates to the surrounding text elements. When it comes to how things are verbally described to

us as humans, context is everything. Therefore, if we want to ensure that the tools we create are more inclusive for those who rely on them, especially regarding alt text, we need to ensure that the algorithms generating this text are aware of the significance of context when describing images. In a Reddit discussion about the practicality and dependability of LLM-based alt-text generators, Reddit users note:

"LLM-based alt text generators cannot assess image nor subject context, plus are still prone to hallucinations based on the built-in biases of the AI developers. As such, while the generators can produce very detailed descriptions, you'll end up spending more time on prompt engineering for every individual image and editing the output. You will get much more accurate and specific results just by just working with a copy writer on the alt text or doing it yourself as you have full context for what the image is meant to represent in the document." [9]

"As someone who subscribes to the philosophy that the purpose of alt text is to convey the purpose or message of an image, not a literal description of it, I wouldn't use any alt text generator."[12]

"The problem is that alt text for the sake of alt text is worthless. You don't describe the picture, you describe the context in which that picture brings to the page. Simply looking at the picture in a vacuum isn't going to solve any alt text issues."[12]

In recent years, popular large language models (LLMs) on the market, such as ChatGPT and Gemini, have introduced image-processing features in their public models. This fact, coupled with the concerns expressed by Reddit users, raises an important question: have popular LLM models evolved to generate more contextual alternative text without extensive prompt engineering? In this part of our project, we will conduct a small test to evaluate how well popular LLMs consider the context when generating alt text with minimal human intervention and oversight.

A. Methodology

We will evaluate the following models: GPT4, Gemini, Qwen. We are choosing these models because they are currently the most popular and widely used options for powering chatbot interactions. We also considered using Copilot; however, the underlying model that supports that experience is GPT-4c to avoid making redundant calls, it was omitted. We only considered models that we either already had access to or could access for free.

In formulating our prompts, we followed the National Federation of the Blind's suggestions on obtaining visual information from AI [6]. They recommend stating that the description is intended for blind individuals, as this will encourage the model to provide more details about elements that a sighted person may not deem necessary to describe. We also took into account the alt text guidelines from Harvard's digital accessibility guide and encouraged adherence to these guidelines through prompting [7].

We wanted to choose websites that are widely used by the general public (including BLV users) and represent different content types (informational, commercial, ecommerce). With these guidelines in mind, we came up with a list in a ranking of hardest to easiest to generate alt text for:

- NYTimes - Images can relate to nuanced news topics, which can make their meaning not entirely obvious.
- Amazon - Products can look generic when standing alone - context like size, use or variant is usually outside the image.
- Ikea - Products here are often shown in staged rooms with products, but you won't know the product dimensions, included items, or configurations without reading the product page.
- Wikipedia - Images are usually literal enough that the LLM can often guess correctly or closely enough based on visual patterns and common sense.
- Yelp - Images are uploaded by users and therefore are of varying quality. They mostly are

described in and related to the review.

To ensure our results are well-developed, we will use two prompts to evaluate how effectively they describe images in context. All test images will be screenshots from the websites listed above, in which the image we want described is outlined with a red border. Our code and the images can be found at https://github.com/demetriam/6156-Final-Project.

Prompt 1: "*Describe the image with a red border to someone who is blind. The description will be read aloud by a screen reader, so keep the description to 1-2 sentences and follow best practices for alt text. Convey the most important visual details that are relevant in the context of the webpage this image is a part of, but don't overwhelm the user with unnecessary information. Consider why this image is included instead of describing every little detail. No need to say 'image of' or 'picture of', but do say if it's a logo, illustration, or diagram.*"

Prompt 2: "*Describe the image with a red border to someone who is blind. This screenshot includes a visual element (ex: product banner, photo, text etc.) generate appropriate alt text for that visual element only ignoring the surrounding webpage layout or article structure. Keep the description concise and relevant to the image content and purpose.*"

B. Scoring Guide

When evaluating the effectiveness of the model in generating rich contextual alt text, we will leverage best practices from notable online resources, including subreddits like /accessibility and /blind, to ensure our framework takes into account the audience that this analysis impacts to create our evaluation framework and algorithm that assigns a quality score to each response.

Our 5 factor framework for evaluating the LLMs will be called CLEAR It will be based on the

- **C**ontextual Awareness
    - Does the model leverage surrounding information like nearby text, page layout, or symbols to generate a more meaningful description?
- **L**anguage Naturalness
    - Is the alt text humanlike, using natural phrasing rather than robotic or templated language?
- **E**ssential Details
    - Does it capture the most important visual elements in the image? Focusing on what's most relevant to the user, without overwhelming them with unnecessary information.
- **A**ccuracy
    - Is the description factually and visually correct?
- **R**edundancy avoidance
    - Avoids phrases like "image of," "picture showing," or "graphic of." Starts directly with the content

[7][8][10]

Category	Score = 2 (Excellent)	Score 1 (Partial)	Score 0 (Poor)
Contextual Awareness	Effectively integrates context from surrounding text, UI elements, or inferred task to enhance the image description.	Shows some use of context but somewhat incomplete or surface-level.	No use of context; purely visual description or inaccurate assumptions.

Language Naturalness	Reads smoothly and naturally, like human speech or writing.	Somewhat natural, but slightly robotic.	Mechanical, robotic template like language.
Essential Detail Inclusion	Captures the most relevant visual content clearly and concisely. Focused on what matters.	Captures some important content, but may include some unimportant element.	Misses key elements or under describes the image.
Accuracy	Objectively correct and matches whats in the image. No hallucinations.	Minor inaccuracies, vague terms.	Incorrect or misleading - wrong objects, relationship or missing key visual elements.
Redundancy Avoidance	Avoid "image of", "graphic of". Begins with meaningful content.	Includes some redundancy.	Starts with "image of", "photo of"; verbose.

Scoring:
- **Max Score per rating:** 10 points
- **Interpretation:**
  - **9-10:** Excellent alt text that is suitable for real world deployment with minimal editing.
  - **6-8:** Usable but needs some refinement
  - **0-5:** Needs significantly improvement or rewriting.
- In the case that the description is not relevant to the image in question, a score of 0 is given.

For the responses, scores, and notes please see the tables in the appendix.

C. Discussion

Many of the descriptions contain redundant information and/or begin with phrases like "the image shows". Providing instructions in the prompt to not do this helped in many cases as demonstrated by comparing the results of prompt 1 (provides instructions) and prompt 2 (does not provide instructions). This highlights the importance of the prompt, but even with these instructions in some cases the description starts in this way (see GPT-4o with prompt 1 for Amazon for example). Ideally the model would incorporate relevant context, but not repeat information that is stated directly above or below the image. For instance, describing that an image is a still from the "Pride and Prejudice" movie (see Gemini with prompt 1 for NYTimes) is useful information, but stating that "this visual is accompanied by the headline …" is not (see Qwen with prompt 2 on NYTimes). This is a difficult task as it requires reasoning about what information is essential and how the image relates to the surrounding web page.

From the 30 descriptions, 15 were accurate, 5 had minor inaccuracies, and 6 had major inaccuracies (the other 4 failed to describe the relevant image) yielding a 50% accuracy rate. This is a concern, as BLV users will not trust AI-generated descriptions if they often prove to be inaccurate. These descriptions can be misleading for instance GPT-4o with prompt 1 for Wikipedia describes a painting as featuring a formal setting when it shows a battlefield. One approach to improve the accuracy could be to aggregate multiple descriptions from one or more LLMs to generate a consensus, but this would increase the cost and latency. Another approach to try could be to specify in the prompt to only include details that the model is sure about.

In 4 cases, the model was not able to identify the image in question and instead described another image on the webpage or described the webpage itself. This could indicate that the image was not marked distinctly enough. However, this was not universal across any image.

Gemini on average was the fastest at returning a response (1.8 seconds). On average, GPT-4o had approximately double the response time (3.6 seconds) and Qwen had over five times the response time (11.2 seconds). Practically speaking, for this functionality to be integrated into a screen reader alt text needs to be generated in real-time. It would also be possible to preload image descriptions as users navigate the rest of the web page, but this would introduce an extra cost and be unnecessary if users do not browse the images.

Overall, we found that LLMs were able to incorporate context beyond giving a literal description when shown an image in the context of the surrounding webpage and it was possible to use a generic prompt that produced reasonable results for images included in different settings. We believe there is potential to build in this kind of functionality into a screen reader with some additional modifications to improve accuracy and verbosity. While human-written alt text is still preferable, in cases where alt text is missing this could be a valuable option for BLV individuals navigating the web.

RQ 2. Can a speech-command interface that allows BLV users to navigate web pages through natural language reduce the cognitive load associated with keyboard-driven screen reader navigation? Does voice navigation feel more accessible or empowering?

The goal is to simplify navigation by enabling commands like "go to the first heading" or "find the search bar," which can be more intuitive than memorizing keyboard shortcuts. Many accessibility tools focus on auditory output, but few address input flexibility, especially via speech. We built a proof-of-concept system to evaluate the effectiveness of this approach.

A. Methodology

The code for this part can be found at https://github.com/demetriam/6156-Final-Project. For our speech-based navigation prototype, we leveraged the sounddevice Python library to capture real-time audio input from the microphone and convert it into a NumPy array. This raw audio is passed to the speech_reconginiton toolkit and then processed using the Google Web Speech API to produce text. Once the text is transcribed, it is passed it a Seleium-based handler, which is a toolkit that simulates user interactions on a web page. We implemented common webpage-tasks: scrolling, navigating to an element on a page (e.g. the first header), and finding and selecting the search bar on a page. The result was a working pipeline where spoken input is captured, transcribed, interpreted, and executed as a real-time web navigation action.

B. Evaluation

A demo is shown in this video https://www.youtube.com/watch?v=bc73DenpiL4. We led a structured evaluation session with ophthalmology resident Dr. Keale Cade at Howard University Hospital. The recording of this session is available in the appendix.

Dr. Cade found that the speech-command interface takes out a lot of the learning curve, frustration, and cognitive load associated with using a traditional screen reader since it uses natural language instead of keyboard shortcuts. She noted that this interface would be suitable for those with motor impairments as well as those with vision impairment since it allows web-users to navigate a page without using a mouse or typing on a keyboard. In her experience, patients with a recent disability are more prone to get frustrated when learning to use assistive technologies so she valued that the technology is intuitive. Dr. Cade pointed out that it is easier with this interface to go back to something you saw previously without having to remember how far you have to scroll and where exactly the element is located on the page. She points out that people with vision impairment may have trouble with spatial recognition so they could have a harder time knowing and remembering where something is on a webpage than a sighted user would. Even with spatial audio, she notes that it is challenging to process the

text of the webpage and also remember where an element is. The speech-command interface takes off this pressure. An additional benefit is that verbalizing a command saves time compared with the spatial navigation implemented in WebNExt or a traditional screen reader where a user would have to navigate through all the elements on the page until they reach the one that they are looking for.

Dr. Cade believes that this interface could be integrated with spatial audio features like those of WebNExt and those that we implemented in RQ 1, but one challenge would be to sort out what sounds are playing at what time to not confuse users and to be able to process speech commands.

Self-Evaluation + Contributions
Anna-Samsara Daefler

I implemented the additional audio feedback discussed in RQ 1.1, recorded a demo of the new features, and came up with questions to guide our evaluation session with Dr. Cade. I evaluated RQ 1.2 with Demetria. I also added additional functionality to the speech-command interface of RQ2 and recorded a demo for this interface.

Through this project, I learned about technologies that make the internet more accessible. The field of human-computer interactions was new to me, so I found it really interesting to learn and think about how people interact with computers and how these interactions can be made more seamless and accessible. It was challenging to find a satisfying way to evaluate the viability of our project, but the feedback that we got from ophthalmologist Dr. Keale Cade was very helpful in guiding this work.

Demetria Mack
Gen AI alt text work:

For this project, I was responsible for building the code for the AI-powered alt-text generation system. To do this, I wrote a script with a multi-model pipeline to generate alt text for the images that we wanted the LLMs (Gemini, GPT-4, and Qwen) to analyze. I began by initializing each model API client, which was relatively straightforward since the API information for each app was available online through documentation. Both Anna and I collaborated on designing prompts that would guide the LLMS to generate context-aware image descriptions. Once the prompts were created, I wrote the core functionality to process batches of images and automatically generate ALT text output. This included writing logic to iterate through the folder with the screenshots of websites we were evaluating, applying prompts, and storing each model's response. The biggest challenge I encountered while working on this part of the project was dealing with API rate limits and token quotas, especially since we were using free developer tiers.

To evaluate the quality of the ALT text, I created the CLEAR evaluation framework for assessing response quality. For this work, I leveraged existing best practices when writing alt text. I think the most difficult part of this work was creating a cool acronym name for the framework.

Speech Processing work:

I also worked to build the foundation for our speech-based navigation prototype. This part of the project drew heavily from what I've been learning in my speech recognition course (COMS 6998). I leveraged the sounddevice library to capture real-time audio input from my microphone and convert it into a NumPy array. This raw audio was passed to the speech_reconginiton toolkit and then processed using the Google Web Speech API to produce text. Once the text was transcribed, I passed it a Seleium-based handler, which is a toolkit that simulates user interactions on a web page. Here, I was able to implement some initial behavior like scrolling Anna later extended this functionality with additional interaction options. The result was a working pipeline where spoken input is captured, transcribed, interpreted, and executed as a real-time web navigation action.

This part of the project was really fun for me because I was able to connect the concepts I'm actively learning in other classes to a real-world accessibility use case. Although Anna-Samsara and I struggled to pinpoint the exact direction we wanted to take this project, I am proud of the progress we

made. We remained committed to focusing on HCI and accessibility, drawing inspiration from the ongoing work in the CEAL lab.

Smaller General Tasks
- Contributing to evaluating the responses from LLMs in the Gen AI work
- Setting up group calls
  - Helping to set project milestones
- Setting up the github repo
- Contributing to the content written in the project proposal, progress report, and final report
- Facilitating communication with the ophthalmologist who evaluated our speech processing prototype - Dr.Cade of Howard University.
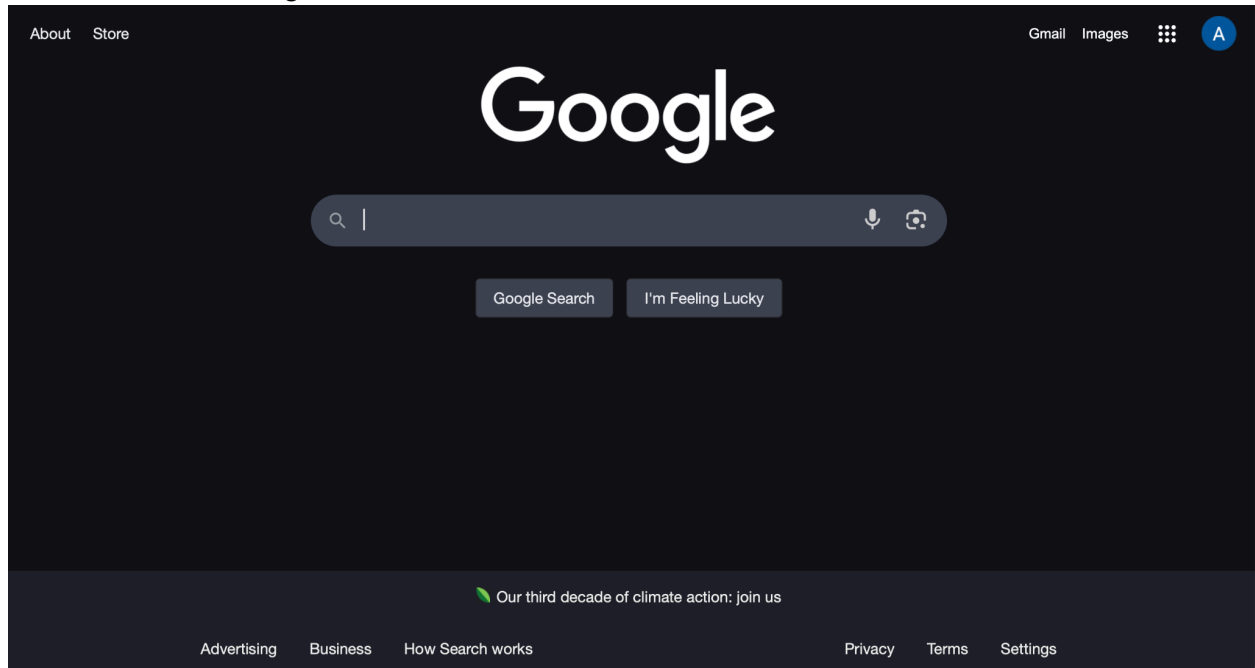
References

1. Arnavi Chheda-Kothary, David A Rios, Kynnedy Simone Smith, Avery Reyna, Cecilia Zhang, and Brian A. Smith. 2023. Understanding Blind and Low Vision Users' Attitudes Towards Spatial Interactions in Desktop Screen Readers. In Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23). Association for Computing Machinery, New York, NY, USA, Article 83, 1–5. https://doi.org/10.1145/3597638.3614490

2. Arnavi Chheda-Kothary, Ather Sharif, David A. Rios, and Brian A. Smith. 2025. "It Brought Me Joy": Opportunities for Spatial Browsing in Desktop Screen Readers. In CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3714125

3. Juan Nino, Sherezada Ochoa, Jocelyne Kiss, Geoffreyjen Edwards, Ernesto Morales, James Hutson, Frederique Poncet, and Walter Wittich. 2024. Assistive Technologies for Internet Navigation: A Review of Screen Reader Solutions for the Blind and Visually Impaired. International Journal of Recent Engineering Science 11, (December 2024), 260–274. https://doi.org/10.14445/23497157/IJRES-V11I6P122

4. Venkatesh Potluri, Tadashi E Grindeland, Jon E. Froehlich, and Jennifer Mankoff. 2021. Examining Visual Semantic Understanding in Blind and Low-Vision Technology Users. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 35, 1–14. https://doi.org/10.1145/3411764.3445040

5. Michael F. Chiang, Roy G. Cole, Suhit Gupta, Gail E. Kaiser, and Justin B. Starren. 2005. Computer and World Wide Web Accessibility by Visually Disabled Patients: Problems and Solutions. Survey of Ophthalmology 50, 4 (2005), 394–405. https://doi.org/https://doi.org/10.1016/j.survophthal.2005.04.004

6. National Federation of the Blind. AI prompt engineering. Retrieved April 19, 2025 from https://nfb.org/resources/publications-and-media/access-on-podcast/ai-prompt-engineering-how-communicate-ai-get

7. Harvard University. Write helpful Alt Text to describe images. Retrieved April 19, 2025 from https://accessibility.huit.harvard.edu/describe-content-images

8. WebAIM. (n.d.). *Alternative Text*. WebAIM. Retrieved April 24, 2025, from https://webaim.org/techniques/alttext/

9. u/AnonymousUser. (2013, March 18). *Alt text generation for PDFs*. Reddit. Retrieved April 24, 2025, from https://www.reddit.com/r/accessibility/comments/1bomlb3/alt_text_generation_for_pdfs/

10. American Library Association. (n.d.). *Context is Important to Writing ALT Text*. ALA Support. Retrieved April 24, 2025, from https://www.ala.org/support/context-important-alt-text

11. University of Washington. (n.d.). *A brief history of HTML*. DO-IT, AccessComputing. Retrieved April 24, 2025, from https://www.washington.edu/accesscomputing/webd2/student/unit1/module3/html_history.html

12. u/AnotherUser. (2013, March 10). *Stumbled upon a neat alt text generator*. Reddit. Retrieved April 24, 2025, from

Appendix

Figure 1. Shows the Google homepage. The search bar is in the center of the webpage and there are no elements to the left or right of it.

RQ 1.2 Evaluation.
Prompt 1
GPT-4o results:

	NYTimes	Amazon	Ikea	Wikipedia	Yelp	Average
Response	Two hands clasping in a famous gesture from ""Pride & Prejudice,"" signifying an emotionally significant moment.	The image with a red border shows a package of cat food featuring a white cat on the front, surrounded by various household items in the ""Buy again"" section on a shopping webpage."	A technical diagram of a bed frame shows dimensions and measurements for assembly, with numbers indicating lengths and widths in various parts of the frame.	The image with a red border is a historical painting depicting a significant scene from the American Revolutionary War, featuring military figures in a formal setting. It's part of a gallery highlighting key moments of	A brightly colored drink topped with a lime and orange slice, placed on a wooden table, evoking a festive atmosphere.	

10

	NYTimes	Amazon	Ikea	Wikipedia	Yelp	Average
				the war on a Wikipedia page about the conflict.		
Time Elapsed (seconds)	3.527193784713745	3.527193784713745	2.865338087081909	4.796469211578369	3.9247028827667236	3.72817955
CLEAR Score	C=2; L=1; E=2; A=2; R=1 **Overall: 8**	C=2; L=1; E=1; A=2; R=0 **Overall: 6**	C=1; L=2; E=2; A=1; R=2 **Overall: 8**	C=1; L=1; E=1; A=0; R=0 **Overall: 3**	C=0; L=2; E=2; A=0; R=2 **Overall: 5**	6
Notes	Stating that this signifies an "emotionally significant moment" is somewhat redundant.	Considers context but includes redundant information.	The purpose of the measurements is not really for assembly.	The painting shows a battlefield not a formal setting.	There is no lime slice.	

Gemini results:

	NYTimes	Amazon	Ikea	Wikipedia	Yelp	Average
Response	A still image from the 2005 film adaptation of *Pride and Prejudice*, showing two hands clasped together.	A diagram showing a sale on Ring security products, as part of Amazon's spring home improvement sale.	Diagram showing the dimensions of a queen-size bed frame with storage drawers.	An illustration depicting a pivotal moment in the American Revolutionary War, showing soldiers and key figures.	Four photos of food and drinks served at the restaurant, Dleña	
Time Elapsed (seconds)	1.7915589809417725	2.1098921298980713	1.6065800189971924	1.5894896984100342	1.6895420551300049	1.757412577
CLEAR Score	C=2; L=2; E=2; A=2; R=2 **Overall: 10**	 **Overall: 0**	C=2; L=2; E=2; A=2; R=2 **Overall: 10**	C=1; L=2; E=2: A=2; R=2 **Overall: 9**	 **Overall: 0**	5.8
		Does not describe the relevant image.		Concise, but could use context clues to be more specific.	Does not describe the relevant image.	

Qwen results:

	NYTimes	Amazon	Ikea	Wikipedia	Yelp	Average
Response	The red-bordered illustration shows two hands gently holding each other, symbolizing connection and support. This image accompanies an article about the enduring popularity of a specific hand gesture from the 2005 adaptation of ""Pride & Prejudice.""	The illustration shows a red border around a product image of Fancy Feast cat food pouches, featuring a white cat. This section is part of the ""Buy again"" recommendations on the Amazon homepage.	The diagram shows the dimensions of the TONSTAD bed frame with storage, including length, width, and drawer measurements. This visual helps you understand the bed's size and storage capacity.	The illustration depicts a pivotal moment in the American Revolutionary War, showing the surrender of Lord Cornwallis after the siege of Yorktown. The scene includes soldiers and leaders on horseback, symbolizing the American victory.	A vibrant tamarindo margarita is served in a glass with a slice of orange on the rim, placed on a wooden table. The drink's bright orange and pink hues suggest its fresh and refreshing nature, complementing the review's mention of its delightful flavor.	
Time Elapsed (seconds)	7.400306940078735	7.175092935562134	27.75198268890381	8.906224966049194	19.751281261444092	14.19697776
CLEAR Score	C=2; L=1; E=1; A=0; R=0 **Overall: 4**	C=2; L=1; E=1; A=0; R=0 **Overall: 4**	C=2; L=1; E=2; A=2; R=1 **Overall: 8**	C=2; L=2; E=2; A=1; R=2 **Overall: 9**	C=2; L=2; E=2; A=1; R=2 **Overall: 7**	**6.4**
Notes	Repeats information presented in the text below. The image is not an illustration.	Not an illustration.		More accurate to describe this as a painting than an illustration.	The orange is not on the rim, but demonstrates good use of contextual information such as naming the drink mentioned in the review that includes this image.	

Prompt 2
GPT-4o results:

	NYTimes	Amazon	Ikea	Wikipedia	Yelp	Average
Response	A close-up image showing one hand holding another hand gently, expressing a subtle gesture of connection.	The image in the red border shows a pack of Fancy Feast cat food featuring a white cat on the box. The pack includes 30 cans of wet cat food.	A technical drawing of a bed frame with measurements labeled for various dimensions, such as length, width, and height. The drawing shows a three-dimensional view of the bed, focusing on structural aspects and design features.	The image with a red border depicts a historical scene, likely an 18th-century painting, showing a significant moment during the American Revolutionary War. It features several figures, including soldiers, gathered around what seems to be a surrender or negotiation, capturing the essence of a military event.	Four photos of food and drinks served at a restaurant. The images show a cocktail, an empanada, enchiladas, and a dessert.	
Time elapsed	3.5453929901123047	4.2026321887976997	3.7764649391174316	4.337281942367554	1.543571949005127	3.481068802
CLEAR Score	C=0; L=2; E=2; A=2; R=1 **Overall: 7**	C=0; L=2; E=2; A=2; R=0 **Overall: 6**	C=2; L=2; E=2; A=2; R=2 **Overall: 10**	C=1; L=2; E=1; A=1; R=0 **Overall: 5**	**Overall: 0**	**5.6**
Notes	Does not mention the image being from the 'Pride and Prejudice' movie.			Unnecessarily surmises that the painting is from the 18th century (it is from the 19th century).	Does not describe the relevant image.	

Gemini results:

	NYTimes	Amazon	Ikea	Wikipedia	Yelp	Average
Response	Close-up of two hands clasped together, suggesting	A box of Purina Fancy Feast Poultry & Beef Grilled cat	Diagram showing the dimensions of a queen-size bed	A painting depicting a pivotal battle scene of the	Four photos of food and drinks at a restaurant: a cocktail, an	

	NYTimes	Amazon	Ikea	Wikipedia	Yelp	Average
	intimacy and longing.	food, containing 30 cans.	frame with storage drawers.	American Revolutionary War, showing soldiers on horseback and foot, with a prominent flag.	empanada, enchiladas, and crème brûlée.	
Time Elapsed (seconds)	1.8492140769958496	1.7412250041 96167	1.63822078704 83398	1.89050412178 03955	1.73751187324 52393	1.771335173
CLEAR Score	C=0; L=1; E=2; A=2; R=2 **Overall: 7**	C=0; L=2; E=2; A=2; R=2 **Overall: 8**	C=2; L=2; E=2; C=2; A=2; R=2 **Overall: 10**	C=1; L=2; E=2; A=0; R=2 **Overall: 7**	**Overall: 0**	**6.4**
	Does not mention the image being from the 'Pride and Prejudice' movie.			I am not sure which flag the description is referring to as prominent as there are two in the painting.	Does not describe the relevant image.	

Qwen results:

	NYTimes	Amazon	Ikea	Wikipedia	Yelp	Average
Response	The image with the red border shows two hands clasped together in a gesture of connection or support. The hands appear to belong to two different individuals, with one hand gently holding the other. This visual is accompanied by the headline: ""The 'Pride & Prejudice' Hand Flex: One Gesture and the	The image with the red border shows a product listing for ""Fancy Feast Poultry & Grill Variety Pack"" cat food. The packaging features a white cat and indicates that the pack contains 30 pouches. The product is listed under the ""Buy again"" section, suggesting it is a previously	The image with the red border is a detailed diagram of a bed frame with storage, specifically the TONSTAD model. It shows the bed frame from a side view, highlighting the dimensions of various parts. The headboard height is 45 5/8 inches, the footboard height is 18 1/8 inches, and the mattress length	The image with the red border is a painting titled ""Surrender of Lord Cornwallis."" It depicts a significant moment from the American Revolutionary War, showing the British surrender after the siege of Yorktown. The scene includes soldiers and military figures, with a central figure on	The image with the red border shows a vibrant, layered drink served in a tall glass. The drink has a gradient of colors, transitioning from a deep orange at the bottom to a lighter yellow at the top. It is garnished with a slice of orange on the rim of the glass. The drink is placed on a wooden	

	Web Is Still Swooning,"" suggesting that this hand gesture has gained significant attention and popularity online.	purchased item.	is 79 1/2 inches. The mattress width is 59 7/8 inches. The diagram also illustrates the dimensions of the storage drawers, with the drawer depth inside being 22 7/8 inches and the drawer width inside being 28 3/8 inches. The overall length of the bed frame is 83 1/8 inches, and the width is 63 3/8 inches. The diagram includes measurements in both inches and centimeters for clarity.	horseback, likely representing a key military leader. The painting captures the formal and solemn atmosphere of the surrender ceremony.	surface, and there are some decorative elements in the background, including what appears to be a rolled-up menu or napkin holder.	
Time Elapsed (seconds)	6.914019823074341	7.299041032791138	11.459800958633423	8.187477111816406	7.561936855316162	8.284455156
CLEAR Score	C=2; L=0; E=1; A=2; R=0 **Overall: 5**	C=2; L=0; E=1; A=1; R=0 **Overall: 4**	C=2; L=0; E=1; A=2; R=0 **Overall: 5**	C=2; L=2; E=1; A=2; R=0 **Overall: 7**	C=0; L=1; E=1; A=0; R=0 **Overall: 2**	**4.6**
Notes	Includes redundant information like stating that "the image with the red border shows…" and restating the headline directly below the image.	Product name has minor inaccuracy: it is not labelled as a variety pack anywhere.	The description is much too long and includes redundant information that is printed above the photo in a section labeled 'Measurements'.	Picks up on the name of the painting from context.	Incorrectly describes the color gradient of the drink. The drink transitions from orange at the bottom to red at the top. The orange slice is not on the rim of the glass but	Descriptions are too long.

					on top of the drink. There is also no rolled-up menu or napkin holder in the background.	

Evaluation Session with Dr. Keale Cade.

🎬 *Research Review Sync - 2025/04/24 17:32 EDT - Recording*