Volume 12
May 2024
ISSN:

2024

The Journal on Technology and Persons with Disabilities

Journal Track Proceedings, Online 2024

CSUN | CENTER ON DISABILITIES

2024
THE JOURNAL ON TECHNOLOGY AND PERSONS WITH DISABILITIES

Which AI Systems Create Accurate Alt Text for Picture Books?

Quan (Monica) Zhou, Nathan Samarasena, Kyle Han, Nataly López, Stacy M. Branham
University of California, Irvine
quanz13@uci.edu, nsamaras@uci.edu, kyleh1104@gmail.com,
natalyisalopez@gmail.com, sbranham@uci.edu

Abstract

In the last decade, there has been a surge in development and mainstream adoption of Artificial Intelligence (AI) systems that can generate textual image descriptions from images. However, only a few of these, such as Microsoft's SeeingAI, are specifically tailored to needs of people who are blind screen reader users, and none of these have been brought to bear on the particular challenges faced by parents who desire image descriptions of children's picture books. Such images have distinct qualities, but there exists no research to explore the current state of the art and opportunities to improve image-to-text AI systems for this problem domain. We conducted a content analysis of the image descriptions generated for a sample of 20 images selected from 17 recently published children's picture books, using five AI systems: asticaVision, BLIP, SeeingAI, TapTapSee, and VertexAI. We found that descriptions varied widely in their accuracy and completeness, with only 13% meeting both criteria. Overall, our findings suggest a need for AI image-to-text generation systems that are trained on the types, contents, styles, and layouts characteristic of children's picture book images, towards increased accessibility for blind parents.

Keywords

AI, alt text, image description, blind, screen reader user, picture book illustrations

Introduction

Advances in technology have enabled people with vision disabilities the ability to access information in digital images using AI-generated alternative texts. Alternative text, sometimes called alt text, provides a textual description of digital images that can be read aloud by a screen reader or a refreshable braille display ("Guideline 1.1 - Text Alternatives"). Yet, these AI-generated solutions are not tailored to the scenario where blind parents want to read with their sighted children. First, studies show that when blind parents read with children, they desire access to many details of the images that are typically not included in alternative text, in order to discuss the images together (Park et al.). Second, prior research shows that children's picture book images have distinct features that are not often represented in AI training sets, such that AIs may not be able to accurately detect objects in such images (Hicsonmez et al.).

Co-reading is an important scenario to design for, as it contributes to children's cognitive and literacy skills development (Mason; Fletcher and Reese). When children listen to their parents reading storybooks, they are exposed to new words in a meaningful context, which brings them "a larger, more fully featured oral vocabulary" (Mason). Images in picture books are also important; labeling objects in the images increases children's exposure to novel vocabulary and concepts, as well as helping parents guide children's attention and participation (Fletcher and Reese). Therefore, when blind parents co-read with sighted children, they desire access to not only the print text, but also the images, so they can engage their child in educational dialogues about the visuals (Park et al.).

Unfortunately, initial studies suggest that blind parents do not find images in current reading technologies very accessible (Park et al.; Storer and Branham). There have been several studies aimed at improving co-reading experiences, but none of them are designed to enable

parents with vision disabilities to read with their sighted children. Attarwala et al. designed a listening and talking e-book, which allows sighted users to audio record themselves reading a book, and people who are blind can later listen to the recording while viewing the book in large digital text. There are also several studies focusing on sighted parents reading to children with vision disabilities, utilizing tactile books to help children feel and learn about images (Kim and Yeh). Some devices that are designed for the non-disabled population show promise for interactive reading applications. For example**,** Zhang et al.'s research presents an AI-enabled system, StoryBuddy, which can automatically generate dialogic questions about the story for children without the parent being involved. However, research suggests that blind parents do not want to be cut out of the reading process (Park et al.; Storer and Branham). Methods that provide access to rich images in children's picture books—to facilitate synchronous, accessible (for blind parents), yet visually oriented (for sighted children) co-reading experiences—are still needed. Therefore, in this paper, we explore opportunities for AI to generate image descriptions as opposed to replacing the parent in co-reading interactions.

A growing body of research has been exploring automatic alt text generation and optimization using AI image-to-text generation software (Mack et al.; Kreiss et al.; Gleason et al.). The models used by these systems are based on a wide range of image sources, including Wikipedia images (Kreiss et al.) and social media images from Facebook (Wu et al.) and Twitter (Gleason et al.). However, according to a study of AI recognition of different illustrator's styles in children's picture books, the imaginary nature of these formats may lead to "extreme characters and settings," such that illustrations are distinct from common images (Hicsonmez et al.). This begs the question: which AI systems create accurate alt text for picture books? There

has yet to be research that explores the current state of the art and opportunities to improve AI image-to-text generation systems for this problem domain.

To address this gap, we conducted a content analysis (Krippendorff) of five current AI systems: asticaVision, BLIP, SeeingAI, TapTapSee, and VertexAI. We then selected a random sample of 17 recently published children's picture books from popular recommendation lists (e.g., The New York Times Best Children's Books of 2022). We extrapolated a total of 669 digital images from the EPUBs of these books offered by Google Play Books. Then, we selected a subset of 20 images that represented a range of image *types* (i.e., decorative, informative, complex), *contents* (i.e., people/animals, abstract/patterns, background/scenery), *styles* (i.e., cartoon, realistic, stylized), and *layouts* (e.g., half page, single page, full spread, panels). We ran these images through each AI system, generating 100 descriptive texts, and then we conducted quantitative and qualitative analysis to ascertain the quality of these descriptions according to W3C's and similar image description guidelines.

Methods

Our content analysis had three phases. First, we identified a sample of digital children's book images (Phase 1). Next, we ran a search to identify viable AI tools (Phase 2). Finally, we conducted the content analysis on the generated text (Phase 3).

Phase 1: Digital Image Selection

This work borrows the digital children's book selection process used by Jinseo Kim in his 2023 master's thesis titled "Are Digital Children's Books Accessible to Blind Parents with Sighted Children?" (Kim). Because blind parents report lack of access to the wide variety of recently published books available in print, Seo identified popular books from a range of lists (e.g., New York Times from 2015 to 2022, NPR's 100 Children's Books list). Only 17 of these

120 children's books were available from a mainstream eBook vendor (i.e., Google Books). He then extracted 669 digital images from the books for accessibility research purposes. From this set of images, our research team selected a subset of 20 images that represented a range of image *types* ("Decorative Images"), *contents* (identified through inductive thematic analysis), *styles* (Guru Staff), and *layouts* (Ferreira; "Panel (Comics)"). The categories and examples of how they were applied to particular images can be found in Tables 3, 4, 5, and 6 respectively.

Phase 2: AI Tool Selection

Next, we conducted a search for off-the-shelf AI image-to-text generation software using both the Google search engine and Apple App Store search. Our search strings were: "image captions ai," "alt text ai," "alt text app," "image description app," and "computer vision." We identified 14 AI tools using this method. We then filtered our list down to five viable AI systems, including those that had distinct output and were therefore using distinct underlying AI systems, those that most reasonably adhered to W3C guidelines for alt text ("Technique G94"), and those that were specifically designed for blind people (TapTapSee and SeeingAI). Figure 1 depicts the details of our filtering process, which led us to ultimately select the following five AI systems: VertexAI by Google, BLIP by Salesforce, SeeingAI by Microsoft, TapTapSee by Cloughtsight, Inc., and asticaVision by Onomal Inc.

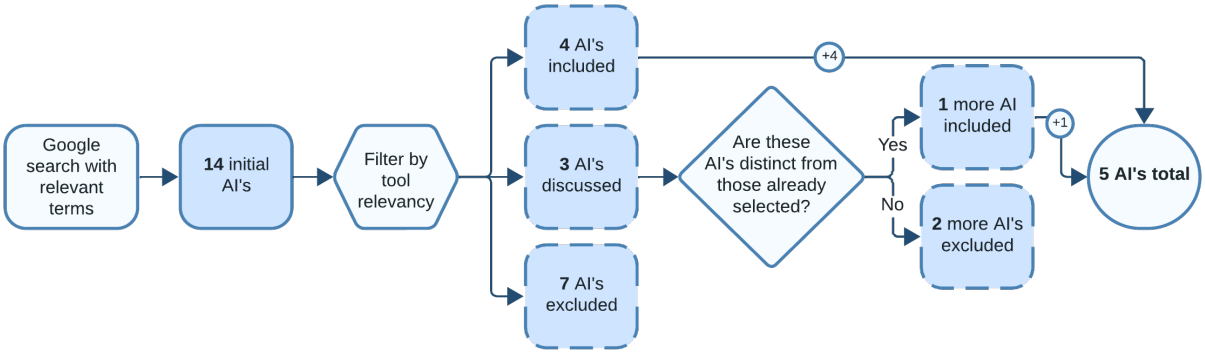
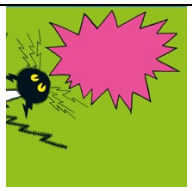
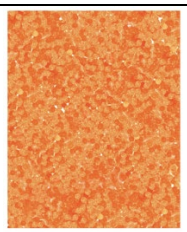


Fig. 1. Flow Chart of AI tool selection process.

Phase 3: Data Collection and Analysis

We used each of the five AI systems to generate image descriptions for each of the 20 selected images, resulting in 100 image descriptions. We then qualitatively assessed the accuracy and completeness of each description. Accurate descriptions are those that are unlikely to lead to misunderstandings about the story content; complete descriptions are those that address each part of the image that contains useful information in the context of reading a children's book. The lead author applied this schema to all images, discussing and iteratively refining its application in conversation with the last author. Table 1 shows five examples of how this schema was applied. Finally, we used descriptive statistics to identify trends in images and image descriptions.

Table 1. Image classifications, AI descriptions & quality assessment data for five sample images.

Metadata	**Image #1**	**Image #7**	**Image #13**	**Image #9**	**Image #14**
Image File					
Type	Complex	Complex	Informative	Decorative	Decorative
Content	Human(s) / Animal(s)	Human(s) / Animal(s)	Human(s) / Animal(s)	Abstract / Pattern	Background / Scenery
Style	Realistic	Cartoon	Cartoon	Stylized	Stylized
Layout	Spread	Panels	Single	Single	Spread
AI Description	“A drawing of a family sitting around a table” (VertexAI)	“Mickey mouse and minnie mouse wall decor” (TapTapSee)	“A cartoon character is hitting a pink speech bubble” (BLIP)	“A close up of orange dots” (SeeingAI)	“This painting depicts a landscape with trees, animals, and a house.…” (asticaVision)
Accurate / Complete?	Accurate, Incomplete	Inaccurate, Incomplete	Inaccurate, Complete	Accurate, Complete	Accurate, Incomplete

Findings

Table 2. Accuracy and Completeness of Image Descriptions for each AI system.

AI	Accurate, Complete	Accurate, Incomplete	Inaccurate, Complete	Inaccurate, Incomplete
asticaVision	4	3	8	5
BLIP	2	4	2	12
SeeingAI	3	4	2	11
TapTapSee	0	0	0	20
VertexAI	4	8	2	6

Accuracy and Completeness Across AI systems

The text generated by VertexAI showed the best performance, with 60% accuracy across all images. SeeingAI and asticaVision tied for second, with 35%, followed by BLIP with 30%. asticaVision had a 60% completeness rate, while others were less than 30%. This is because asticaVision tends to generate a detailed yet redundant description, which raises the possibility of identifying more elements in an image, thus increasing the completeness. We also calculated the "incomplete, accurate" ratio, which is the rate of being "incomplete" for all the "accurate" text. All systems except TapTapSee had "incomplete, accurate" ratios over 40%. We speculate that AI systems tended to focus on the main object and action of an image, rather than details. Surprisingly, 100% of descriptions generated by TapTapSee—an app specifically designed for blind users—were categorized as "inaccurate, incomplete." We suspect that TapTapSee has used few or no images of cartoon and abstract art for their data training, explaining why the AI system performed poorly. Also, TapTapSee's descriptions were concise, making it less likely to contain relevant elements depicted in an image.

Image Type

Table 3. Accuracy and Completeness of Image Descriptions by Image Type.

Image Type	# of Image	Accurate, Complete	Accurate, Incomplete	Inaccurate, Complete	Inaccurate, Incomplete
Decorative	6	9	6	2	13
Informative	5	4	4	4	13
Complex	9	0	9	8	28

Decorative images' generated text shows the highest accuracy, while that of informative and complex images are both below 40%. Decorative images usually contain symbolic objects and tend to precede or succeed the story itself, so they will not substantially affect readers' understanding of the story. Therefore, we did not require "accurate" image descriptions for decorative images to include the symbolic meaning, leading to high rates of accuracy among this image type. For example, Image #9 (Table 1) is composed of countless orange dots. A close reading of the story reveals that the orange dots symbolize connection and hope, as the color orange appears in scenes where the boy spends time with the dog; further, as the image appears on the last page of the book, it suggests hope that the relationship will continue. We labeled descriptions "accurate" if they mention the "large number of orange dots." In contrast, informative and complex images contain more information that are critical to the main story (e.g., Table 1., Images #1 and #7). In order to be considered "accurate," descriptions for these image types needed to be both correct and complete, leading to lower levels of accuracy among this image type.

Image Content

Table 4. Accuracy and Completeness of Image Descriptions by Image Content.

Image Content	# of Image	Accurate, Complete	Accurate, Incomplete	Inaccurate, Complete	Inaccurate, Incomplete	Accuracy
human(s) / animal(s)	13	4	8	13	40	18%
abstract / pattern	3	9	1	1	4	67%
background / scenery	4	0	10	0	10	50%

Our data suggests that image content is correlated with accuracy and completeness of the generated text. As shown in Table 4, descriptions for images that were coded as “abstract / pattern” or “background / scenery” were accurate at rates of over 50%. However, the descriptions for images coded as “human / animal” were only accurate at a rate of 18%. This may be a result of these images portraying more complex stories, including character identities and actions (e.g., Table 1, Images #1 and #7), leading to higher standards of accuracy and completeness. In contrast, images coded as “abstract / pattern” tended to have simple stories (e.g., Table 1, Image #14), which led to a lower standard of accuracy. Examining completeness for images coded as “background / scenery,” the descriptions were all determined to be “incomplete.” These visuals contained many elements and thus had an elevated standard for complete descriptions. As an example, Image #14 (Table 1) depicts a landscape with rich details, including palm trees, houses, a river, and a boat. None of the AI systems generated descriptions with all of the elements above. The description closest to “complete,” generated by asticaVision, still misses the river and boat.

Image Style

Table 5. Accuracy and Completeness of Image Descriptions by Image Style.

Image Style	# of Image	Accurate, Complete	Accurate, Incomplete	Inaccurate, Complete	Inaccurate, Incomplete	Accuracy
Cartoon	10	3	6	9	32	18%
Realistic	4	4	6	1	9	50%
Stylized	6	6	7	4	13	43%

Our data also suggest that differences in image style are correlated with differences in accuracy and completeness of generated descriptions. The text generated for images classified as "realistic" had an accuracy of 50%, the highest figure. For "stylized," accuracy fell to 43%, while that of "cartoon" - the most plentiful in our sample - was only 18%. We consider several reasons for such a result. First, "stylized" images in our sample tended to represent simple aspects of the narrative (e.g., Table 1, Image #9 and #14), making it easier for AI systems to provide accurate descriptions. "Cartoon" and "realistic" images, conversely, tended to depict complex stories that included human and animal characters, which raised the standard for accuracy. Second, some "cartoon" images tended to be abstract, leading to a wide variety of "inaccurate" descriptions generated by AI systems. For example, Image #13 (Table 1) depicts a black cat with a pink speech bubble. Yet, this was interpreted as a "green and purple owl illustration" (TapTapSee), a monster (asticaVision), and a spider (asticaVision, SeeingAI). Only one AI recognized the cat (VertexAI). Third, "cartoon" images in our sample tended to include "panel" layouts; all AI systems performed poorly on such images.

Image Layout

Table 6. Accuracy and Completeness of Image Descriptions by Image Layout.

Image Layout	# of Image	Accurate, Complete	Accurate, Incomplete	Inaccurate, Complete	Inaccurate, Incomplete	Accuracy
Spread	4	0	8	2	10	40%
Single	12	13	10	8	29	38%
Panels	4	0	1	4	15	5%

While none of the image layouts attained more than 50% accuracy on their descriptions, it is worth noting that "panel" images were by far the lowest, with a mere 5% accuracy. Further, though content in these images was accurately detected, none of the AI systems could distinguish between scenes of a paneled image. Consider Image #7 (Table 1), where both asticaVision and SeeingAI recognized "skeletons," "dog," and the corresponding action, "running." Yet, VertexAI did not detect the "running" action. Instead, it generated the *only* accurate translation of the words inside the speech bubbles. None of the AI systems could successfully identify the full story, composed of the dog chasing two skeletons and scaring them by barking.

Discussion

In our data analysis, we found several patterns which led us to believe that AI models are poor at describing children's book images. TapTapSee, an AI system designed for blind users, had 100% of its descriptions categorized as "inaccurate, incomplete." "Panel," "cartoon," and "complex" images most commonly led to poor descriptions. Prior work has shown that AI models are largely trained on utilitarian image types: Wikipedia images (Kreiss et al.) and social media images from Facebook (Wu et al.) and Twitter (Gleason et al.). Our work indicates a gap in the capabilities of AI models to generate appropriate children's book image descriptions. The

gap is particularly pronounced for images with properties that are distinct or common among children's picture books (e.g., "panel" and "cartoon" images). We propose that AI models have substantial room for improvement within this domain.

For particular images, AI systems performed relatively well in accurately detecting visual content. VertexAI was the highest performer in terms of accuracy, with 60% of descriptions coded as "accurate." AsticaVision was the highest performer in terms of completeness, with 60% of descriptions coded as "complete." Unfortunately, when considering both measures, VertexAI and AsticaVision tied for top performance, with a mere 20% of descriptions coded as "accurate, complete." Research shows that when parents read with children, they do not simply want to know the presence of a character or object; they additionally want to access details like the emotions on a character's face, the actions taking place, as well as colors and shapes (Storer and Branham; Park et al.). Therefore, we consider that descriptions which are accurate but lack completeness are not viable for co-reading scenarios.

We found images in our sample that are best described as "decorative" which contain details that parents might enjoy talking about with their children. As described in Sections 2 and 3 of the Findings, Images #7 and #14 precede or follow the actual story itself. Thus, the image content is not necessary for readers to access the main story. However, as co-reading is an opportunity for children to learn, parents may want to explore this type of image with their child. The detailed landscape in Image #14, for instance, could allow parents to guide children to label the objects, practice and expand their vocabulary, and have educational dialogues around each element. With Image #7, parents may want to discuss the abstract and symbolic relationship of the orange dots to the main story. However, W3C states that decorative images "don't add information to the content of a page" ("Decorative Images"), and thus, they "require empty

alternative text that convey their ornamental purpose" ("Module 4: Images"). Therefore, we argue that the decorative images in children's storybooks should be considered important in the co-reading process, though it may not add information to the story itself. In other words, decorative images *should* have alt text that is both accurate and complete.

Conclusion

While AI systems have seen tremendous growth over past years, no research studies have looked into the prospect of utilizing AI systems to generate image descriptions for children's picture book illustrations. Our study analyzed the descriptions of 20 images from 17 digital children's picture books. Five AI systems (asticaVision, BLIP, SeeingAI, TapTapSee, and VertexAI) each produced 20 descriptions. The accuracy and completeness of descriptions varied greatly by AI system, with variations apparent across image type, content, style, and layout. Our research highlights the necessity for AI systems—especially those like TapTapSee, which are widely adopted within the blind community—to be trained on children's picture books and address the needs of the blind community. We foresee a need to not only improve our AI systems, but to also revise alt text guidelines in ways that are sensitive to co-reading use cases. Progress on these fronts is the first step toward full inclusion of blind parents in the important educational and familial bonding experience of reading print books with children at home.

Work Cited

"Alternative Text." Digital Accessibility at Princeton, The Trustees of Princeton University, accessibility.princeton.edu/how/content/alternative-text. Accessed 12 Oct. 2023.

Attarwala, A., Munteanu, C., and Baecker, R. 2013. An accessible, large-print, listening and talking e-book to support families reading together. In Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services. 440–443.

"Decorative Images." W3C, 27 July 2019, www.w3.org/WAI/tutorials/images/decorative/.

Ferreira, Karen. "Types of Illustrations for Children's Books." Get Your Book Illustrations, 7 Feb. 2020, getyourbookillustrations.com/types-of-illustrations-for-childrens-books/.

Fletcher, K. L. and Reese, E. 2005. Picture book reading with young children: A conceptual framework. Developmental Review, 25(1), 64–103. https://doi.org/10.1016/j.dr.2004.08.009

Gleason, Cole, et al. "Twitter A11y: A browser extension to make Twitter images accessible." Proceedings of the 2020 chi conference on human factors in computing systems. 2020.

"Books on Google Play." Google, Google, play.google.com/store/books?hl=en_US&gl=US.

"Guideline 1.1 – Text Alternatives." W3C. 2019. https://www.w3.org/WAI/WCAG21/Understanding/text-alternatives

Guru Staff. "Types of Illustration Styles for Children's Books." Guru Insights, 4 Nov. 2021, www.guru.com/blog/types-of-illustration-styles-for-childrens-books/.

Hicsonmez, S., Samet, N., Sener, F., and Duygulu, P. 2017. Draw. Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. https://doi.org/10.1145/3078971.3078982

Onamal Inc. "Cognitive Intelligence API: See, Speak, and Hear with Astica." Astica AI, https://astica.ai/.

Kim, J. (2023). "Are Digital Children's Books Accessible to Blind Parents with Sighted Children?" Master's thesis, University of California, Irvine.

Kim, Jeeeun and Yeh, Tom. 2015. Toward 3d-printed movable tactile pictures for children with visual impairments. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pages 2815–2824, 2015.

Krauss, Jennifer. "The Best Children's Books of 2022." The New York Times, The New York Times, 8 Dec. 2022, www.nytimes.com/2022/12/08/books/review/the-best-childrens-books-of-2022.html.

Kreiss, Elisa, Goodman, N. D., and Potts, C. 2021. "Concadia: Tackling image accessibility with context." https://arxiv.org/pdf/2104.08376v1)%3C%22

Krippendorff, Klaus. (2018). Content Analysis: An Introduction to Its Methodology, 4th ed., SAGE Publications, pages 51–52.

Li, Junnan, et al. "Blip: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation." Salesforce AI, 5 Nov 2022, https://blog.salesforceairesearch.com/blip-bootstrapping-language-image-pretraining/.

Mack, K., Cutrell, E., Lee, B., and Morris, M. R. 2021. Designing tools for high-quality alt text authoring. The 23rd International ACM SIGACCESS Conference on Computers and Accessibility. https://doi.org/10.1145/3441852.3471207

Mason, Jana M. 1990. Reading Stories to Preliterate Children: A Proposed Connection to Reading. Technical Report No. 510.

"Module 4: Images." W3C, 29 Sept. 2022, www.w3.org/WAI/curricula/content-author-modules/images/.

"Panel (Comics)." Wikipedia, Wikimedia Foundation, 5 Aug. 2023, https://en.wikipedia.org/wiki/Panel_(comics)#:~:text=A%20panel%20is%20an%20individual,drawing%20depicting%20a%20frozen%20moment.

Park, S., Cassidy, C., and Branham, S. M. 2023. "It's All About the Pictures: Understanding How Parents/Guardians with Visual Impairments Co-Read with Their Child(ren)." In The 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23), October 22–25, 2023, New York, NY, USA. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3597638.3614488

"Seeing AI App from Microsoft." Microsoft, www.microsoft.com/en-us/ai/seeing-ai

Storer, Kevin and Branham, Stacy M. 2019. "That's the Way Sighted People Do It" What Blind Parents Can Teach Technology Designers About Co-Reading with Children. In Proceedings of the 2019 on Designing Interactive Systems Conference. 385–398.

"TapTapSee - Blind and Visually Impaired Assistive Technology - powered by CloudSight.ai Image Recognition API." TapTapSee, https://taptapseeapp.com/.

"Technique G94: Providing Short Text Alternative for Non-Text Content That Serves the Same Purpose and Presents the Same Information as the Non-Text Content." W3C, 20 June 2023, www.w3.org/WAI/WCAG21/Techniques/general/G94.

"Vertex AI." Google, https://cloud.google.com/vertex-ai?hl=en.

Vezzoli, Y., Kalantari, S., Kucirkova, N., and Vasalou, A. 2020. Exploring the design space for parent-child reading. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3313831.3376696

Wu, S., Wieland, J., Farivar, O., and Schiller, J. 2017. Automatic alt-text. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. https://doi.org/10.1145/2998181.2998364

Zhang, Z., Xu, Y., Wang, Y., Yao, B., Ritchie, D., Wu, T., Yu, M., Wang, D., and Li, T. J.-J. 2022. StoryBuddy: A human-ai collaborative chatbot for parent-child interactive storytelling with flexible parental involvement. CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3491102.3517479