VQAsk: a multimodal Android GPT-based application to help blind users visualize pictures

Maria De Marsico
Sapienza University of Rome
Rome, Italy
demarsico@di.uniroma1.it

Chiara Giacanelli
Sapienza University of Rome
Rome, Italy
giacanelli.1801145@studenti.uniroma1.it

Clizia Giorgia Manganaro
Sapienza University of Rome
Rome, Italy
manganaro.2017897@studenti.uniroma1.it

Alessio Palma
Sapienza University of Rome
Rome, Italy
palma.1837493@studenti.uniroma1.it

Davide Santoro
Sapienza University of Rome
Rome, Italy
santoro.1843664@studenti.uniroma1.it

ABSTRACT

VQAsk is an Android application that helps visually impaired users to get information about images framed by their smartphones. It enables to interact with one's photographs or the surrounding visual environment through a question-and-answer interface integrating three modalities: speech interaction, haptic feedback that facilitates navigation and interaction, and sight. VQAsk is primarily designed to help visually impaired users mentally visualize what they cannot see, but it can also accommodate users with varying levels of visual ability. To this aim, it embeds advanced NLP and Computer Vision techniques to answer all user questions about the image on the cell screen. Image processing is enhanced by background removal through advanced segmentation models that identify important image elements. The outcomes of a testing phase confirmed the importance of this project as a first attempt at using AI-supported multimodality to enhance visually impaired users' experience.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility systems and tools**; **Ubiquitous and mobile computing systems and tools**; **Empirical studies in interaction design**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Visual Question Answering, visually impaired users, natural language processing and computer vision for scene interpretation

ACM Reference Format:
Maria De Marsico, Chiara Giacanelli, Clizia Giorgia Manganaro, Alessio Palma, and Davide Santoro. 2024. VQAsk: a multimodal Android GPT-based application to help blind users visualize pictures. In *International Conference on Advanced Visual Interfaces 2024 (AVI 2024), June 03–07, 2024, Arenzano, Genoa, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3656650.3656677

AVI 2024, June 03–07, 2024, Arenzano, Genoa, Italy

ACM ISBN 979-8-4007-1764-2/24/06
https://doi.org/10.1145/3656650.3656677

1 INTRODUCTION

In a world where visual media are part of daily life, accessibility and inclusion become essential to assist visually impaired individuals. This is one of the goals of Visual Question Answering (VQA). VQA is a challenging research area joining advancements in Computer Vision (CV) and Natural Language Processing (NLP). VQA applications entail comparing the semantic information in a visual media against the semantic elements embedded in a question in natural language. This paper proposes a VQA application to support people with different visual abilities in exploring their surrounding environment through information about images framed by their smartphones. The aim is to help to mentally visualize or create a mental representation of what they cannot or can hardly see. This can be achieved thanks to the ability of automatic VQA to answer questions about submitted images. The goal of exploring any image represents a complex and challenging task for several reasons: 1) the questions are not predetermined; both the image and the combinations of questions change across interactions, being the users completely free to ask whatever they want; 2) visual information is usually very rich and high-dimensional; 3) VQA entails many computer vision sub-tasks, e.g., object detection, activity recognition, and scene classification) [10]. The application tackles the challenge through the powerful Generative Pre-trained Transformer (GPT) MiniGPT-4 [15], inspired by GPT-4 [11] but light enough to be embedded in a smartphone app. In addition, automatic image segmentation allows for isolating relevant scene objects for more precise questions and answers. The proposed design creates an inclusive environment through interaction via trigger words for function calls and spontaneous speech for the questions, and also through haptic feedback besides the normal touch-screen.

In summary, the contributions of this work are the following:
- a mobile app for users with different visual abilities embedding powerful AI models for VQA and automatic image segmentation;
- the design of three different interaction modes for inclusive use.

2 RELATED WORK

The reader particularly interested in VQA can refer to available surveys, e.g., [2, 12, 13] and the recent [7]. This section will only briefly summarize its use for accessibility. The increasing attention to accessibility design has stimulated the automatic analysis of

images and related questions to generate the proper answers for visually impaired people [8]. To evaluate VQA applications designed to this aim, [6] introduces the VisWiz dataset with 31,000 visual questions from blind people, who took a photo with their mobile phones and recorded a spoken question about it. Each image is labeled with 10 answers. A privacy-preserving version of the dataset, named VisWiz-priv [5], avoids image regions containing private information (e.g., credit card numbers or private addresses). A related iPhone app named VizWiz [3] allows asking a visual question and obtaining an answer in nearly real-time. The authors of [1] combine bottom-up and top-down attention mechanisms to analyze the question by a Gated Recurrent Unit (GRU), while a CNN processes the image. A further application using a reinforcement learning model to help a blind person navigate the street is described in [14].

3 DESIGN AND IMPLEMENTATION

3.1 Requirements, Multimodality, and AI

The main goal of VQAsk is to support visually impaired users in exploring the environment and a gallery of photos, but also caters to users preferring multimodal interaction to keyboard/touchscreen-based interaction. In accessibility design, both User- and System-Functional requirements include items that are strictly related to assisting impaired people. For sight-impaired users, it is natural to ask, besides the touch-screen, for voice and haptic interaction (a Must-Have User-Functional Requirement in MoSCoW labeling), as well as for an accurate and robust speech-to-text module. The VQA-related requirement is to relate spoken/written questions with the relevant extracted image features (Must-Have System-Functional Requirements). VQask hands-free interaction addresses diverse abilities and preferences. **1) Voice Interaction**: a speaker-independent continuous speech recognizer enables users to ask questions and activate features. This offers visually impaired users an alternative to touchscreens or text-based interactions. Answers are returned by the Vocal Assistant. **2) Haptic Feedback** : tangible responses or error alerts can be received through device vibrations. **3) Visual Interaction**: a minimalist interface allows users to leverage their possible residual vision to navigate.

The VQA application core exploits AI models and tools. **MiniGPT-4** [15] is the very core module of the application. It receives both the input image and a question (text or voice). Then it combines language and vision processing by aligning a frozen visual encoder, namely the same pretrained vision components of BLIP-2 [9], with a frozen advanced large language model (LLM), namely Vicuna [4], using a linear projection layer. The architecture only requires training such a layer to align the visual features with the Vicuna language model. A proper alignment can generate GPT-4-like detailed image descriptions, which are sent back to the user. When the answer is returned, the **flutter_tts** plugin[1] is used as a wrapper around the native text-to-speech engines of the mobile Operating Systems. The user can ask more precise questions by using the object segmentation function offered by **remove.bg** APIs[2]. This web-based service employs advanced AI technology to identify foreground layers and separate them from the background.

[1]https://pub.dev/packages/flutter_tts
[2]https://www.remove.bg/it/tools-api

3.2 The interface

VQAsk is composed of two main minimalist screens, the *Homepage* and the *Editable photograph* screen (Figure 1).

In the *Homepage* the AppBar contains the name of the Application (left), the "info" icon that triggers the description of the application functions, and the "crop" that passes to the other screen. The Info Section is also synthesized to speech for visually impaired users. The app can be fully used through a few semaphoric/trigger words. The present set was meant to avoid ambiguities, due to the free-speech style of the questions, but can be easily changed. As expected, they were one of the few negative points that the users noted, as they do not evoke the corresponding functions.
PORCUPINE: the application takes a picture and loads it as input.
PICOVOICE: the following sentences will be considered as a question related to the currently loaded image; speech can be **continuous** and **spontaneous**.
JARVIS: the app will pronounce back the question asked by voice, to check the correct speech recognition.
BLUEBERRY: the question will be submitted to be answered.
GRAPEFRUIT: the Vocal Assistant will read again the Info Section.
At present, no trigger word allows also to select an image from the gallery. This is only possible by touching the button on the left.

Normal- or partially-sighted users can use icons instead of trigger words, as described below (Figure 1, left).
- The *microphone* icon, on the right of the input form, allows one to ask a question. Clicking on the *X* icon erases it.
- It is possible to listen to a question before asking it by clicking on the *volume-up* icon. The "Ask Me!" button submits the question.
- A submitted question is passed to the MiniGPT-4 module along with the image; the app passes in the *thinking stage*, a circle progress bar is shown and the Vocal Assistant tells "I'm thinking" to the user.
- The answer is shown in a green Card also including a *volume-up* button, to listen to the answer again, and a *switch button* that, when enabled, makes the app automatically read aloud a returned answer.
Two error conditions have been handled in which the app cannot answer: 1) the user clicks on the "Ask me!" button or pronounces BLUEBERRY without typing or vocally asking anything; 2) the inserted question is too short to be unambiguous. When handling the *empty question* error, **haptic feedback** is triggered along with the visual mode to enhance multimodality: when the error pops up, the phone vibrates, indicating that something went wrong.

The second app screen is *Editable photograph*. Actually, it is not fully usable by a visually impaired user, because it needs attention and visual awareness to select an image portion unless using automatic segmentation. Users can manually crop an image or segment it automatically to find its most important elements. This *magic cropping* allows visually impaired users to locate the salient scene or image objects and ask specific related questions. An icon activates the Information page that will be shown and read by the Vocal Assistant. Users can select photographs, and four buttons (from left to right in the center part of Figure 1) allow to: 1) *apply changes and load the image* to ask related questions; 2) *trash the image* and return to the home of *Editable Photograph*; 3) *manually crop the image*, rotate and scale it; 4) *automatically segment the image* to identify the salient elements to ask about (Figure 1, left); visually impaired users can trigger this function using the word TERMINATOR.

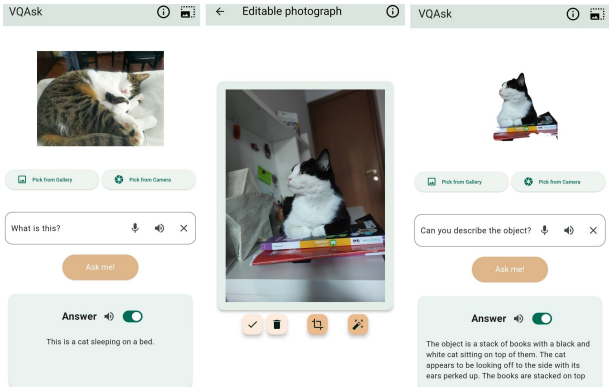

Figure 1: Main VQAsk interface screens; the last image is the follow-up of the session after cropping the image foreground.

4 TESTING AND EVALUATION

A preliminary evaluation with 10 sighted users assessed the feasibility of using the chosen AI models within a mobile application and users' preferences for the interaction modes. A simple questionnaire showed that more than half of the people preferred to interact vocally with the application. On a scale of 1 to 4, the effectiveness of voice commands got 40% of response 3 and 60% of response 4. Users appreciated the experience and the usefulness of the app.

After the preliminary test, the questionnaire was enriched to collect more detailed information, and a new group of users was enrolled. The new tests involved 12 Italian users with different visual abilities, from 20 to 35 years old. The number of users may appear quite low. However, due to the main goal of the application to assist visually impaired users, adding more normal-sighted ones would have enlarged the group of testers, yet without adding useful insight into the accessibility of the application. On the other hand, it is quite difficult to collect a group of sight-impaired users with different levels of impairment ; as a matter of fact, testing with blind people was made possible thanks to the collaboration with the *Unione Italiana dei Ciechi e degli Ipovedenti ONLUS (UICI) - Italian Union of the Blind and Visually Impaired* of Rome. Tests were authorized by the leading staff of the Union. The test session was also preceded by filling and signing forms for individual authorization to use the collected data. The users executed any task of their choice during a *Think Aloud* session. During the free task execution, it was observed from the users' reactions whether the chosen actions appeared to be intuitive and whether they encountered any bugs or difficulties in the general use of the app. The observers' notes testify that the users appeared interested, engaged, and generally amused. At the end, they filled in a short questionnaire to identify the preferred interaction modes, evaluate their experience, and suggest improvements. The questionnaire was devised to adapt to the application use cases, and to especially point out accessibility-related preferences and possible problems. It starts with a self-assessment of one's sighting ability. After this, a group of questions deals with the user's preferences when triggering each main application function, with the choice between vocal commands or icons tapping. Two following questions focus on the experience with the most critical operation for the visually impaired, i.e., cropping, and with the application in general. The next questions deal with the perceived possibility of using the application without sight, its perceived usefulness, and the user's overall satisfaction with the application answers. Precision, completeness, and accuracy of the responses refer in more detail to the users' evaluation of the way the adopted models work, and are considered in a group of specific questions. The final question is related to the anticipated future use of the application.

Table 1 shows both the questionnaire and the final results, that suggest an overall positive reception of the application, with users finding it intuitive and generally pleasant. In particular, users demonstrated significant appreciation for the possibility of asking questions vocally (more than 80% of them) and submitting it vocally too, without using any button and related visual awareness of the application interface. If we look at a finer partition of the answers based on visual abilities, it is interesting to notice that users with higher level of visual impairment especially prefer vocal interaction. In particular, while users with different visual abilities share this appreciation to different extents, 100% of users with serious visual impairment prefer vocal interaction. Another aspect that users enjoyed was the *magic cropping*, a novelty compared to other similar competitors, especially for the possibility to ask questions about the cropped region only. Unfortunately, this possibility cannot be exploited by the seriously visually impaired unless devising, in the future, some addition to the protocol to provide some preliminary general information able to orientate the user in the overall image content. Another interesting observation regards the overall experience with the application, where the visually impaired users provided a higher percentage of positive responses. Eventually, most users were satisfied with the quality of the model's answers to their questions. Also in this case, the visually impaired ones were the most enthusiastic. For instance, a blind user found the answer of the application really useful regarding her question about the make-up she was wearing that day. Besides questionnaire answers analysis, the observation of the users' think-aloud provides further information. As anticipated above and expected, remembering commands and pronunciation was challenging, calling for the change of the trigger words. In addition, allowing to choose the language of the commands will further make the interaction easier. Both solutions are easy to apply. Some users suggested introducing the shaking gesture to make the app repeat the question and check

Question	Sighted Users Answers	Partially sighted Users Answers	Blind Users Answers	Total Results
How do you define yourself?	**66.7%**	**16.7%**	**16.7%**	-
What did you prefer to open the Info page?	**37.5%** Using the vocal command - **63.5%** Tapping	**100%** Using the vocal command	**100%** Using the vocal command	**58.3%** Using the vocal command - **41.7%** Tapping
What did you prefer to take pictures?	**50%** Using the vocal command - **50 %** Tapping	**100%** Using the vocal command	**100%** Using the vocal command	**66.7%** Using the vocal command - **33.3%** Tapping
What did you prefer to ask the question?	**80%** Using the vocal command - **20 %** Tapping	**100%** Using the vocal command	**100%** Using the vocal command	**83.3%** Using the vocal command - **16.7%** Tapping
What did you prefer to listen to the question again?	**37.5%** Using the vocal command - **63.5%** Tapping	**100%** Using the vocal command	**100%** Using the vocal command	**66.7%** Using the vocal command - **33.3%** Tapping
What did you prefer to submit the question?	**62.5%** Using the vocal command - **37.5%** Tapping	**100%** Using the vocal command	**100%** Using the vocal command	**75%** Using the vocal command - **25%** Tapping
You prefer:	**62.5%** To listen to the answer by tapping on the icon - **37.5%** To let the answer be reproduced automatically	**100%** To let the answer be reproduced automatically	**100%** To let the answer be reproduced automatically	**75%** To listen to the answer by tapping on the icon - **25%** To let the answer be reproduced automatically
How do you rate the experience of image cropping:	**75%** Very intuitive - **25%** Quite intuitive - **0%** Not so intuitive	-	-	**75%** Very intuitive - **25%** Quite intuitive - **0%** Not so intuitive- **0%** Counter-intuitive
How do you rate the general user-experience of the app?	**50%** Very pleasant - **37.5%** Quite pleasant - **12.5%** Not so pleasant- **0%** Not pleasant	**50%** Very pleasant - **50%** Quite pleasant - **0%** Not so pleasant - **0%** Not pleasant	**50%** Very pleasant - **50%** Quite pleasant - **0%** Not so pleasant - **0%** Not pleasant	**50%** Very pleasant - **41.7%** Quite pleasant - **8.3%** Not so pleasant - **0%** Not pleasant
How much do you think you could have used the application without using your sight?	**37.5%** A lot - **62.5%** Quite - **0%** A little - **0%** Nothing	**0%** A lot - **100%** Quite - **0%** A little - **0%** Nothing	**50%** A lot - **50%** Quite - **0%** A little - **0%** Nothing	**33.3%** A lot - **66.7%** Quite - **0%** A little - **0%** Nothing
How useful do you think this application can be in everyday life?	**50%** A lot - **50%** Quite - **0%** A little - **0%** Not at all	**100%** A lot - **0%** Quite - **0%** A little - **0%** Not at all	**50%** A lot - **50%** Quite - **0%** A little - **0%** Not at all	**58%** A lot - **41.7%** Quite - **0%** A little - **0%** Not at all
How much are you satisfied with the application answers?	**63.5%** A lot - **37.5%** Quite - **0%** A little - **0%** Not at all	**100%** A lot - **0%** Quite - **0%** A little - **0%** Not at all	**100%** A lot - **0%** Quite - **0%** A little - **0%** Not at all	**75%** A lot - **25%** Quite-**0%** A little - **0%** Not at all
How precise were the application responses?	**63.5%** A lot - **25%** Quite - **12.5%** A little - **0%** Not at all	**100%** A lot - **0%** Quite - **0%** A little - **0%** Not at all	**100%** A lot - **0%** Quite - **0%** A little - **0%** Not at all	**75%** A lot - **16.7%** Quite - **8.3%** A little - **0%** I don't know
How complete were the application responses?	**63.5%** A lot - **37.5%** Quite - **0%** A little - **0%** Not at all	**100%** A lot - **0%** Quite - **0%** A little - **0%** Not at all	**100%** A lot - **0%** Quite - **0%** A little - **0%** Not at all	**75%** A lot - **25%** Quite -**0%** A little - **0%** I don't know
How accurate were the application responses?	**63.5%** A lot - **37.5%** Quite - **0%** A little - **0%** Not at all	**100%** A lot - **0%** Quite - **0%** A little - **0%** Not at all	**100%** A lot - **0%** Quite - **0%** A little - **0%** Not at all	**75%** A lot - **25%** Quite - **0%** A little - **0%** I don't know
Do you think you can use the app?	**0%** Every day - **25%** Once a week - **75%** Sometimes - **0%** I don't find it useful	**50%** Every day - **0%** Once a week - **50%** Sometimes - **0%** I don't find it useful	**50%** Every day - **50%** Once a week - **0%** Sometimes - **0%** I don't find it useful	**16.7%** Every day - **25%** Once a week - **58.3%** Sometimes - **0%** I don't find it useful

Table 1: Answers of the questionnaire

whether it was correctly understood. Other users suggested introducing an offline mode enabling the use of the app without an internet connection. However, this would require storing too much resources locally and is still not feasible.

5 CONCLUSIONS

The VQAsk application aims at allowing people with different visual abilities to effectively and mostly independently explore visual data, either stored of captured in real-time. One of VQAsk's standout features is its voice interaction, which allows users to ask questions about visual data through voice commands, and to receive spoken responses. The application relies on machine learning models, enabling the app to effectively respond to users' questions about images or relevant parts of them. The evaluation with 12 users with different visual abilities demonstrated that voice interaction was the preferred method for inputting questions. Additionally, the majority of users preferred automatic playback of responses over manual tapping, highlighting the importance of multimodal accessibility. Possible future improvements surely include: 1) allowing personalized semaphoric/trigger words to increase the quality of the user experience of the application; 2) incorporating acoustic interactions, such as beeps or other auditory cues, as well as user gestures, which can provide additional feedback and guidance for users; translating the application in other languages, such as Italian.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[2] Silvio Barra, Carmen Bisogni, Maria De Marsico, and Stefano Ricciardi. 2021. Visual question answering: Which investigated applications? *Pattern Recognition Letters* 151 (2021), 325–331.

[3] 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.

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 09 January 2024)* (2023).

[5] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. 2019. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 939–948.

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

[7] Md Farhan Ishmam, Md Sakib Hossain Shovon, MF Mridha, and Nilanjan Dey. 2024. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Information Fusion* (2024), 102270.

[8] Walter S Lasecki, Phyo Thiha, Yu Zhong, Erin Brady, and Jeffrey P Bigham. 2013. Answering visual questions with conversational crowd assistants. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–8.

[9] 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. *arXiv preprint arXiv:2301.12597* (2023).

[10] S. Manmadhan and B. C. Kovoor. 2020. Visual question answering: a state-of-the-art review. In *Artifcial Intelligence Review (2020)*.

[11] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[12] Himanshu Sharma and Anand Singh Jalal. 2021. A survey of methods, datasets and evaluation metrics for visual question answering. *Image and Vision Computing* 116 (2021), 104327.

[13] Yash Srivastava, Vaishnav Murali, Shiv Ram Dubey, and Snehasis Mukherjee. 2021. Visual question answering using deep learning: A survey and performance analysis. In *Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5*. Springer, 75–86.

[14] Martin Weiss, Simon Chamorro, Roger Girgis, Margaux Luck, Samira E Kahou, Joseph P Cohen, Derek Nowrouzezahrai, Doina Precup, Florian Golemo, and Chris Pal. 2020. Navigation agents for the visually impaired: A sidewalk simulator and experiments. In *Conference on Robot Learning*. PMLR, 1314–1327.

[15] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592* (2023).