VIAssist: Adapting Multi-modal Large Language Models for Users with Visual Impairments

Bufang Yang[†], Lixing He[†], Kaiwei Liu[†] and Zhenyu Yan[†]
The Chinese University of Hong Kong, Hong Kong SAR, China

Abstract—**Individuals with visual impairments, encompassing both partial and total difficulties in visual perception, are referred to as visually impaired (VI) people. An estimated 2.2 billion individuals worldwide are affected by visual impairments. Recent advancements in multi-modal large language models (MLLMs) have showcased their extraordinary capabilities across various domains. It is desirable to help VI individuals with MLLMs' great capabilities of visual understanding and reasoning. However, it is challenging for VI people to use MLLMs due to the difficulties in capturing the desirable images to fulfill their daily requests. For example, the target object is not fully or partially placed in the image. This paper explores how to leverage MLLMs for VI individuals to provide visual-question answers. VIAssist can identify undesired images and provide detailed actions. Finally, VIAssist can provide reliable answers to users' queries based on the images. Our results show that VIAssist provides +0.21 and +0.31 higher BERTScore and ROUGE scores than the baseline, respectively.**

Index Terms—**Visual impairments, Foundation models, Visual question answering, Multi-modal large language models (MLLMs), Internet of Things**

I. INTRODUCTION

Visually impaired (VI) people refer to individuals who have a partial or total inability to visual perception [1]. According to the statistics from the World Health Organization (WHO), at least 2.2 billion people worldwide suffer from near or distance VI [2]. Recent years have seen growing interest in harnessing artificial intelligence (AI) technologies to support VI individuals [3], [4].

Many AI-based systems have been developed to enhance the quality of life for VI individuals [5], [6], e.g., obstacle detection and navigation. The exceptional performance and natural way of interaction exhibited by large language models (LLMs) and their multi-modal variants multi-modal large language models (MLLMs) have garnered increased attention [7]–[9]. VI individuals can query MLLMs like GPT-4V [10] to answer their daily concerns, such as "What is the expiration date written on this medicine bottle?" In the era of MLLMs, the remarkable visual understanding and reasoning capabilities of MLLMs can bring new life experiences to VI individuals.

However, it has been observed that the images captured by VI individuals are significantly lower in quality than those taken by sighted individuals [1]. This is because VI individuals, due to their limited vision, cannot visually inspect captured images on their own. Consequently, such images are highly susceptible to poor quality, often capturing only parts of the intended target. These low-quality images can lead the MLLMs to generate unreliable responses and decrease
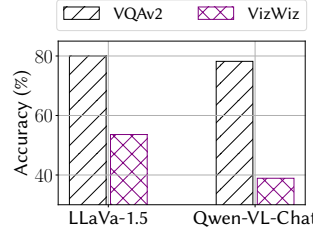

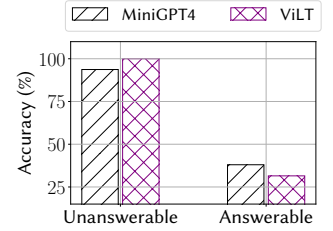Fig. 1: Performance of MLLMs on standard VQA and VI's VQA datasets.

Fig. 2: In-depth analysis of the MLLMs' performance on VI's VQA dataset.

the experience of VI users. A promising approach involves using MLLMs to guide VI individuals through a step-by-step process to capture high-quality images. However, our in-depth analysis shows that current MLLMs face challenges in providing effective and practical suggestions for VI users to re-shoot photos.

In this paper, we explore how to improve the usability of MLLMs for VI individuals. We design VIAssist, an MLLM tailored for enhanced adaptability to the unique inquiries of VI users. For low-quality images, VIAssist can provide actionable and detailed suggestions for retaking photos. Upon capturing a high-quality photo, VIAssist is capable of producing reliable answers to queries from VI users. We first collect an instruction dataset with questions and images tailored for VI individuals, alongside crafting aligned responses. Utilizing this dataset, our fine-tuned MLLM, VIAssist, showcases enhanced responsiveness to VI-specific queries. Both qualitative and quantitative results show that VIAssist can provide more reliable responses for VI users.

II. PRELIMINARY

A. Understanding VQA from VI individuals

1) Visual question answering (VQA): VQA has been a hot research topic in recent years [11]. The emergence of generative artificial intelligence, especially MLLMs [7], [9], has shifted VQA from closed-form to free-form, wherein it can generate text freely based on input images and questions.

2) VQA for VI individuals: Building upon VQA, VQA for VI individuals (VI-VQA) is even more challenging. Since the VI individual has poor eyesight, they cannot check the content of their captured image, resulting in a lower image quality. To the best of our knowledge, VizWiz [1] is the largest VQA dataset for blind people, consisting of 31,000 photos

Fig. 3: Examples of questions and corresponding photos from VI individuals.

from blind people with ten crowd-sourced answers per visual question. It is worth noting that among the photos taken by VI individuals, 28% of them are classified as "Unanswerable," indicating the low quality of photos taken by blind people. Therefore, providing VQA for VI individuals is a challenging task.

3) Quantitative Analysis: To analyze the difference between normal people and blind people, we test the performance of several popular MLLMs on two datasets: a standard VQA dataset VQAv2 [12] and a VI-VQA dataset, VizWiz. Figure 1 shows the performance of LLaVa-1.5 [7] and Qwen-VL-Chat [9] on the two VQA datasets. We can see that there is a significant accuracy gap between the two datasets: in comparison to the VQAv2 dataset, the LLaVa-1.5 exhibits a decrease of 26.4% accuracy on the VizWiz dataset, while the Qwen-VL-Chat demonstrates a reduction of 39.3% accuracy. It indicates that the images captured by VI individuals and their corresponding questions are more challenging than the standard VQA task. Even the recently proposed MLLMs struggle to handle them effectively.

Next, we further analyze the VizWiz dataset in depth by whether the question is "Answerable" or not (2934 "Answerable" questions and 1385 "Unanswerable" questions). We test two models including one MLLM, MiniGPT4 [8], and one traditional VQA model, ViLT [11]. Figure 2 shows that both MiniGPT4 and ViLT models accurately predict "Unanswerable" questions. However, they achieve poor accuracy for "Answerable" questions, lower than 40%. Through a more detailed examination of the VizWiz dataset, we find that even for the "Answerable" questions, the corresponding images exhibit relatively lower quality, e.g., only part of the target is captured in the image.

In summary, we find that the reason for MLLMs' low accuracy lies in two aspects: 1) the poor quality of captured images by VI individuals, and 2) many photos taken by VI individuals are not "Answerable". We will investigate the features of such gaps and the reasons behind them in the next subsection.

4) Key features of VI queries: In this subsection, we outline the key characteristics of queries from VI individuals. These features are crucial for understanding the unique needs and challenges faced by VI individuals when interacting with MLLMs. We summarize the key features of VI queries as follows:

- **Incomplete target in the image.** Only a partial view of the target object mentioned in the question is captured in the image, such as the left part, right part, upper part, or bottom part.
- **Target completely out of the image.** The target object in the question is entirely absent from the image.
- **Inappropriate distance of the target.** The distance of the target object in the image is inappropriate, such as being either too far or too close.
- **Low-quality images.** The captured images suffer from blurriness, complete darkness, or low light conditions.
- **Nonsense or irrelevant questions.** The content of the pictures taken is completely unrelated to the question posed by the VI individual.
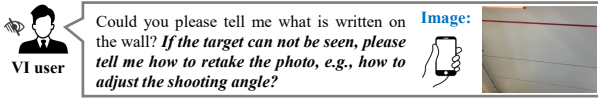
Figure 3 shows some examples of queries and captured images from VI individuals. These types of queries often arise, and they are inherently challenging or even impossible to answer, even for humans. Given that MLLMs are trained and fine-tuned based on human instruction, it is unsurprising that MLLMs may encounter difficulties in providing responses in such cases.

B. Analysis of MLLMs for VI Queries

We collect a set of low-quality images from the real world and request GPT-4V to generate responses to our inquiries. For GPT-4V to give suggestions for reshooting low-quality images, we set the prompt of GPT-4V as "<Question>. If the target can not be seen, please tell me how to retake the photo, e.g., how to adjust the shooting angle?". Figure 4 and Figure 6 show the GPT-4V performance on these low-quality images. We observe that GPT-4V's queries for VI individuals exhibit the following characteristics:

1) General and high-level guidance: Figure 4 shows the scenario where the target, such as the text on the street sign, is entirely absent in the captured image. We can see that GPT-4V has the ability to identify the targets that are not visible in the given photo. However, the generated suggestions for retaking photos remain limited to general and high-level guidance on how to capture a picture with higher quality. The response generated by GPT-4V lacks specific guidance for VI individuals on adjusting the shooting angle, including instructions on how to adjust the shooting direction and which direction to move in. This limitation restricts its practical utility in assistive systems designed for blind people.
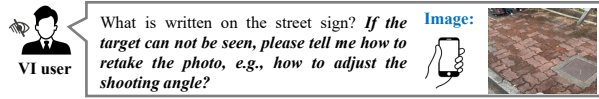
Figure 6 shows an example where the target is partially absent in the captured image. The result shows that GPT-4V can successfully detect the target's absence. However, when

Fig. 4: GPT-4V performance on VI individual queries, where the target is completely out of the image.
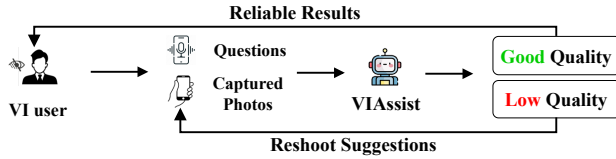

Fig. 5: System overview of VIAssist.

we add the prompt: "If the text can be partially seen, please tell me how to retake the photo, e.g., how to adjust the shooting angle?", GPT-4V only provides general and high-level guidance, failing to provide VI people helpful suggestions.

2) Impractical and ineffective suggestions for VI individuals: Figure 6 also shows that GPT-4V even suggests VI individuals look at the photo to ensure the target's visibility, which is an impractical suggestion for VI individuals.

In summary, current MLLMs like GPT-4V have the ability to detect low-quality images and identify whether the target is partially or completely absent. However, their responses are limited to high-level and general suggestions on adjusting the shooting and even impractical suggestions for VI individuals. These MLLMs can not provide a detailed or step-by-step guide for capturing high-quality images that would enable VI individuals to address their specific inquiries. The main challenge lies in enhancing MLLMs to generate detailed and practical suggestions for reshooting when the quality of the images provided by VI individuals is low.

III. VIASSIST

In this paper, we design VIAssist, which can adapt to the characteristics of VI individual queries, and generate reliable responses. Figure 5 shows the system overview of VIAssist. For low-quality images, VIAssist provides actionable and detailed advice on retaking photos. VIAssist can generate reliable answers to queries from VI users based on the given images.

A. Instruction Dataset

To enhance the MLLMs to generate reliable and practical responses for VI individuals, we need to collect an instruction dataset for fine-tuning. The format of each sample in our dataset is as follows: < Question, Image, Response >.
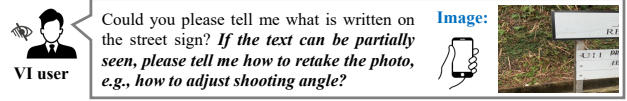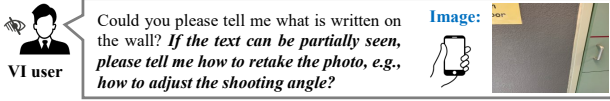
Image Collection. According to our analysis of VI individual queries in § II-A4, we collect the following types of images and questions: 1) High-quality images and questions, 2) Target completely out of the image, 3) Incomplete target in the image, 4) Inappropriate distance of the target, 5) Low-quality image, 6) Nonsense or irrelevant questions.

Since VizWiz dataset [1] lacks types 2), 3), and 4) of images and questions, we collect a real-world dataset at various locations such as campus roads, shopping malls, streets, and indoor corridors.

Question Collection. For each collected image, we first manually create questions. We also employ GPT-4V to rewrite and expand the diversity of created questions.

Response Collection. We recruited annotators to manually write responses for each image and question we collected. The annotated response contains descriptions and suggestions. The description part refers to GPT-4V's answer. In the case of images captured with good quality, the response will include the statement "the quality of this image is good." However, for images with poor shooting quality, such as only part of the target can be seen in the image, a description of the poor quality and suggestions for how to adjust the shooting will be added to the response.

Currently, we have gathered 50 images for each type of VI query. There are 300 < Question, Image, Response > pairs in our instruction dataset in total. Additionally, we have collected a separate set of 50 images for evaluation. These images are captured at various locations, distinct from the ones used in the instruction dataset.

Fig. 6: GPT-4V performance on VI individual queries, where only a partial view of the target object mentioned in the question is captured in the image.

B. Model Training

Model Architecture. The model architecture of VIAssist is based on the open-source MLLM, LLaVA [7]. It contains a pre-trained visual encoder and an LLM. The visual features are converted to the word embedding space through a projection layer. We use the frozen CLIP ViT-L/14 [13] as the visual encoder, and use Vicuna-7B [14] as the LLM.

Instruction Tuning. Since MLLM fine-tuning with full parameters requires a huge amount of data and computation resources, we use LoRA [15] technique for parameter-efficient fine-tuning. We add LoRA parameters to the projection layers within the attention layers of the LLM. During training, only the feature projectors and LoRA parameters are updated, while the remaining parameters are kept frozen.

C. VIAssist Performance

Qualitative Results. Figure 9 shows the performance of VIAssist. The results show that VIAssist's response exhibits three distinct advantages over GPT-4V. Firstly, it can assess the image quality and provide explanations for poor quality. Secondly, it offers more **detailed** and **actionable** suggestions for adjusting the shot, such as moving the camera to the left and taking another shot. In addition, it can generate fewer irrelevant or nonsensical responses. Overall, VIAssist can generate more reliable responses for VI individuals.

Quantitative Results. Following [16], we use BERTScore and ROUGE as two evaluation metrics to quantify the performance of VIAssist and existing MLLMs in terms of response quality. BERTScore [17] can be used to assess the semantic similarity between two texts. ROUGE [18] can be used to assess the degree of token overlap between two texts. For LLaVA, we use LLaVA-v1.6-34B version. For GPT-4V, we set the prompt
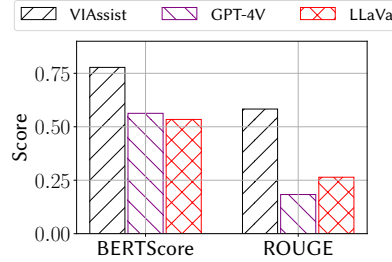
Fig. 7: Quantitative results of VIAssist and other MLLMs.
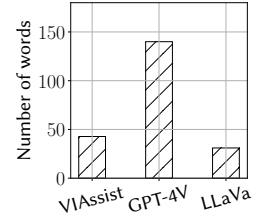
Fig. 8: The number of words in MLLMs' responses.

as "`<Question>`. If the target can not be seen, please tell me how to retake the photo, e.g., how to adjust the shooting angle?". We test the performance of LLaVA and GPT-4V under the zero-shot setting.

Figure 7 shows that VIAssist achieves +0.21 higher BERTScore and +0.31 higher ROUGE than the best baseline model. It demonstrates that VIAssist can generate more reliable responses to queries posed by VI individuals. Figure 8 shows the number of words in MLLMs' responses. We can see that GPT-4V generates the longest responses, likely due to its extensive incorporation of world knowledge, resulting in diverse content. However, as shown in Figure 4 and Figure 6, GPT-4V's responses often include a significant number of irrelevant or nonsensical suggestions. These suggestions are general and impractical for VI individuals. In contrast, VIAssist's responses exhibit higher semantic scores, moderate word count, and complexity, thereby delivering an enhanced QA experience for VI individuals.
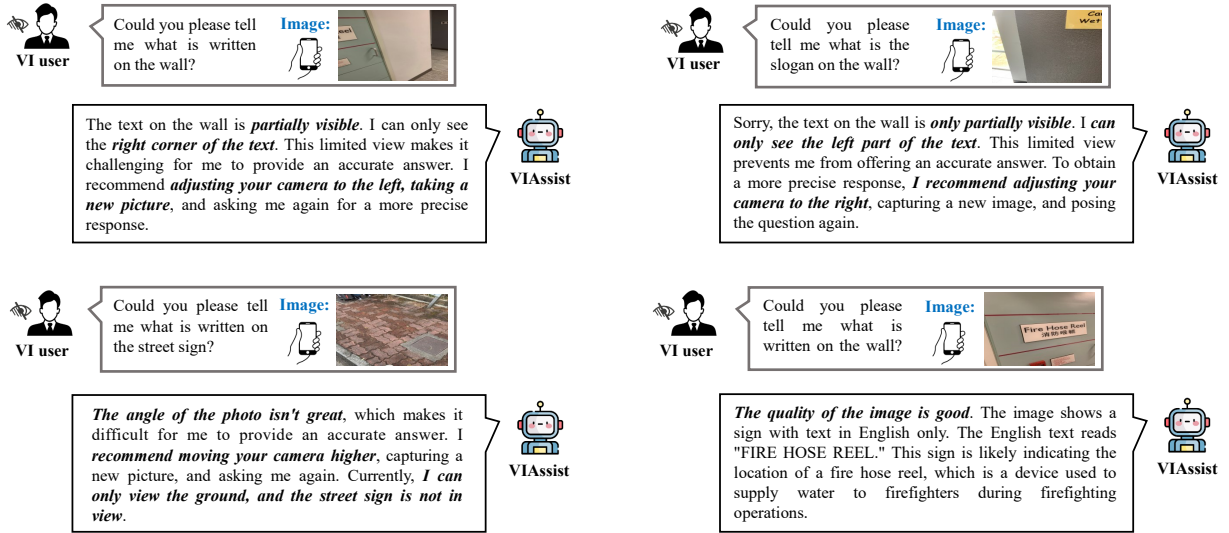
Fig. 9: VIAssist performance on VI individual queries. It can assess the image quality, generate specific suggestions for adjusting the shot, and contain fewer irrelevant responses.

IV. FUTURE DIRECTIONS

Instruction Dataset. The instruction dataset collected in this study is still limited, which occasionally leads VIAssist to provide inaccurate reshoot suggestions for new data. In the future, we will enrich our image dataset with greater diversity, including photos taken by real VI individuals, to enhance the practicality and generalization of VIAssist.

Automatic Reshooting. We also plan to enhance the usability of VIAssist. A potential approach involves refining VIAssist's output (guidance) to enable automatic camera adjustments, such as zooming in or out, eliminating the need for users to comprehend the guidance and manually retake photos

Real-time and Efficiency. According to Be My AI [3], which is the popular AI-based VI assistant app, the response time on average is still 4 minutes, which is not satisfying if the VI user wants to actively use it. Increasing computing resources is a direct approach to addressing this issue. However, this poses challenges for budget-constrained service providers like Be My Eyes [4], requiring significant effort to generate revenue. In addition, VI users can also capture videos to gather more information about their surroundings, which presents substantial challenges in terms of edge computing costs, network bandwidth, and MLLMs cloud services. Several neural network (NN) inference optimization and video streaming techniques can be considered to enhance the efficiency of the VI assist system, including edge-cloud collaboration [19], NN quantization [20], and the efficient NN architectures [21].

Dedicated Prompt Engineering. This work has not yet meticulously designed prompts for GPT-4V. Nonetheless, it is worth noting that prompt engineering [22] can significantly affect the performance of MLLMs. Future research could explore more effective prompts to enhance GPT-4V's responses to inquiries from VI users.

Other Modalities. When the target is completely absent from the photo, it poses a challenge for MLLMs to provide effective reshooting suggestions since no useful information is captured in the image. In contrast to sighted individuals, blind individuals commonly carry additional assistive devices, such as guide crutches, in addition to mobile phones. One open question is: can we leverage additional sensors and modalities, such as wireless signals, audio, and Electroencephalogram (EEG) [23], [24], to improve the performance of MLLMs?

Other Types of Impaired People. Beyond aiding VI users, LLMs can assist with other prevalent disabilities. For instance, individuals who are deaf or hard of hearing may benefit from LLMs to interpret ambient sounds. However, previous work ProtoSound [25] still relied on a relatively simple classification network, which leaves space for improvement.

V. RELATED WORK

Prior to the advent of MLLMs, researchers also explored the utilization of diverse AI models to assist VI individuals. Lin *et al.* [26] propose a multi-modal VI assist system that leverages data from RGBD cameras and earphones to aid blind individuals in navigating and understanding environments. DeepNAVI [5] runs multiple AI models simultaneously to give VI users walking instructions. As for V-eye [6], combining global localization and image segmentation, the users can be given precise location information. In summary, these works are ad-hoc designed and provide only partial information needed by VI users. Given the exceptional visual understanding and reasoning capabilities of MLLMs, exploring the application of MLLMs for VI users holds considerable promise.

Recent studies on MLLMs [7], [9] mainly cater to users with unimpaired vision, offering scant consideration for individuals with visual impairments. Liu *et al.* [27] incorporates SAM [28] and image captioning to build a system to help blind people. Be My Eyes [4] is a free mobile application designed to enhance accessibility for individuals who are blind or have low vision. The application has released its new function: Be

My AI [3], which utilizes AI models to guide users instead of volunteers. Compared to the original Be My Eyes, it provides a 90%+ successful rate with a one-third response time (4 minutes on average). In other words, only 10% of calls will be handled by volunteers. However, the response time is still too long to be applied in the real world. VIALM [16] is the first benchmark to test the performance of MLLMs for queries from VI individuals. However, it primarily focuses on evaluating the success rate of fulfilling requests from VI individuals, overlooking the issues of low-image quality conditions.

VI. Conclusion

This paper aims to enhance the usability of MLLMs for VI individuals. Our preliminary experiments show the limitations of current MLLMs when generating responses for inquiries from VI individuals. In this paper, we design VIAssist, a MLLM that exhibits enhanced adaptability to the unique characteristics of VI queries. For low-quality images, VIAssist provides actionable and detailed advice on retaking photos. Once a high-quality photo is captured, VIAssist can deliver reliable answers to queries from VI users. Both qualitative and quantitative results show that VIAssist can generate more reliable and practical responses for VI users.

References

[1] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

[2] WHO, "Blindness and vision impairment," Aug 2023.

[3] Bemyeyes, "Be my eyes integrates be my aitm into its first contact center with stunning results," Mar 2024.

[4] B. M. Eyes, "The story about be my eyes," Mar 2024.

[5] B. Kuriakose, R. Shrestha, and F. E. Sandnes, "Deepnavi: A deep learning based smartphone navigation assistant for people with visual impairments," *Expert Systems with Applications*, vol. 212, p. 118720, 2023.

[6] P.-J. Duh, Y.-C. Sung, L.-Y. F. Chiang, Y.-J. Chang, and K.-W. Chen, "V-eye: A vision-based navigation system for the visually impaired," *IEEE Transactions on Multimedia*, vol. 23, pp. 1567–1580, 2020.

[7] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[8] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[9] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.

[10] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[11] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*, pp. 5583–5594, PMLR, 2021.

[12] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[14] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[15] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2021.

[16] Y. Zhao, Y. Zhang, R. Xiang, J. Li, and H. Li, "Vialm: A survey and benchmark of visually impaired assistance with large models," *arXiv preprint arXiv:2402.01735*, 2024.

[17] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[18] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74–81, 2004.

[19] B. Yang, L. He, N. Ling, Z. Yan, G. Xing, X. Shuai, X. Ren, and X. Jiang, "Edgefm: Leveraging foundation model for open-set learning on the edge," *arXiv preprint arXiv:2311.10986*, 2023.

[20] S. Ma, H. Wang, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, R. Wang, J. Xue, and F. Wei, "The era of 1-bit llms: All large language models are in 1.58 bits," *arXiv preprint arXiv:2402.17764*, 2024.

[21] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, "Flashattention: Fast and memory-efficient exact attention with io-awareness," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16344–16359, 2022.

[22] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[23] B. Yang, W. Wu, Y. Liu, and H. Liu, "A novel sleep stage contextual refinement algorithm leveraging conditional random fields," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022.

[24] B. Yang, X. Zhu, Y. Liu, and H. Liu, "A single-channel eeg based automatic sleep stage classification method leveraging deep one-dimensional convolutional neural network and hidden markov model," *Biomedical Signal Processing and Control*, vol. 68, p. 102581, 2021.

[25] D. Jain, K. Huynh Anh Nguyen, S. M. Goodman, R. Grossman-Kahn, H. Ngo, A. Kusupati, R. Du, A. Olwal, L. Findlater, and J. E. Froehlich, "Protosound: A personalized and scalable sound recognition system for deaf and hard-of-hearing users," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2022.

[26] Y. Lin, K. Wang, K. Yi, and S. Lian, "Deep learning based wearable assistive system for visually impaired people," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pp. 0–0, 2019.

[27] R. Liu, J. Zhang, K. Peng, J. Zheng, K. Cao, Y. Chen, K. Yang, and R. Stiefelhagen, "Open scene understanding: Grounded situation recognition meets segment anything for helping people with visual impairments," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1857–1867, 2023.

[28] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.