engineering proceedings

MDPI

Proceeding Paper

Image Descriptions for Visually Impaired Individuals to Locate Restroom Facilities †

Cheng-Si He *, Nan-Kai Lo, Yu-Huan Chien and Siao-Si Lin

Department of Computer Science, National Taipei University of Education, Taipei 106, Taiwan; s110916037@stu.ntue.edu.tw (N.-K.L.); s110916039@stu.ntue.edu.tw (Y.-H.C.); s110916027@stu.ntue.edu.tw (S.-S.L.)

* Correspondence: hesisi05@gmail.com

† Presented at the 2024 IEEE 6th Eurasia Conference on IoT, Communication and Engineering, Yunlin, Taiwan, 15–17 November 2024.

Abstract: Since visually impaired individuals cannot observe their surroundings, they face challenges in accurately locating objects. Particularly in restrooms, where various facilities are spread across a limited space, the risk of tripping and being injured significantly increases. To prevent such accidents, individuals with visual impairments need help to navigate these facilities. Therefore, we designed a head-mounted device that utilized artificial intelligence (AI) to enhance its functionality. The ESP32-CAM was implemented to capture and transmit images to a computer. The images were then converted into a model-compatible format for the bootstrapping language-image pre-training (BLIP) model to process and generate English descriptions (i.e., written captions). Then, Google Text-to-Speech (gTTS) was employed to convert these descriptions into speech, which was delivered audibly through a speaker. The SacreBLEU and MOS scores indicated that the developed device produced relatively accurate, natural, and intelligible spoken directions. The device assists visually impaired individuals in navigating and locating the restroom facilities to a satisfactory level.

Keywords: ESP32-CAM; machine learning; BLIP model; artificial intelligence; text-to-speech technology; gTTS; SacreBLEU; MOS

check for updates

Academic Editors: Teen-Hang Meen, Chi-Ting Ho and Cheng-Fu Yang

Published: 25 April 2025

Citation: He, C.-S.; Lo, N.-K.; Chien, Y.-H.; Lin, S.-S. Image Descriptions for Visually Impaired Individuals to Locate Restroom Facilities. *Eng. Proc.* **2025**, *92*, 13. https://doi.org/10.3390/engproc2025092013

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Visually impaired individuals face challenges in daily life without visual aids. In restrooms, where facilities are scattered in a limited space, blind individuals are at a higher risk of being tripped or injured. When using the restrooms, they often struggle to accurately locate objects like toilets, handrails, or sinks in their surroundings. To overcome this difficulty, providing auditory assistance is a viable solution [1].

Undoubtedly, vision assumes a decisive role in the recognition of concrete objects surrounded by human beings. For individuals who lose their eyesight, developing spatial awareness can be extremely challenging. However, if these individuals rely on other senses, such as hearing, they can better perceive and understand what surrounds them for the convenience and safety of their daily lives.

We developed a device based on artificial intelligence (AI) to provide users with real-time spoken information on the locations of facilities in a restroom. The development progressed through several stages. First, we trained the bootstrapping language-image pre-training (BLIP) model [2,3] by using the common objects in context (COCO) dataset [4] developed by Microsoft. We set up a head-mounted camera to capture images, which were

then fed into the BLIP model to generate written text (i.e., captions). Finally, we adopted the Python library Google text-to-speech (gTTS) to convert the text into audio prompts delivered through a speaker [5]. The results indicated that the developed device satisfactorily produced speech to describe the locations of facilities in the scene. As suggested, an AI-powered device delivers instant audio instructions and assists the visually impaired in navigating their surroundings and accurately identifying objects.

2. Literature Review

2.1. Vision-Language Models

To effectively detect objects in images, researchers often develop models based on convolutional neural networks (CNNs) [6,7]. Typical CNN models consist of three main types of layers: convolutional layers, pooling layers, and fully connected layers. In training, they adjust the weights of each unit by using backpropagation. However, CNN-based models primarily handle visual inputs and often require integration with other architectures to process textual data. Alternatively, a computer vision model that incorporates a vision transformer (ViT) is recommended for performing tasks such as image classification or object detection, especially when trained on large datasets.

2.2. The BLIP Model

BLIP is a large multimodal model (LMM), featuring a hybrid encoder-decoder architecture to process and generate both visual and textual data effectively. Its architecture comprises four components: an image encoder, a text encoder, an image-grounded text encoder (i.e., a cross-modal encoder), and an image-grounded text decoder (i.e., a cross-modal decoder) [2].

BLIP utilizes ViT as the image encoder. ViT is a deep learning model that uses global self-attention and patch-based processing to carry out tasks such as image classification [8]. The transformer divides the image into fixed-size patches and flattens them into a sequence of one-dimensional vectors (tokens). It then adds positional embeddings and processes these tokens through a self-attention mechanism, followed by feedforward layers. The self-attention layer dynamically reweights the importance of specific features by employing a query-key-value mechanism. When applied for input tokens, it enables the model to capture relationships between spatially distant patches, which helps to understand the image's overall structure. BLIP detects objects within a larger context, enables interactions, and distinguishes between similar objects based on their location. The self-attention mechanism allows the model to recognize patterns across the entire image with an improved ability to handle complex or cluttered scenes by leveraging long-range dependencies and contextual information.

Second, bidirectional encoder representations from transformers (BERT) can be integrated into BLIP's architecture as a text encoder. BERT is a natural language processing (NLP) model built on the transformer architecture [9]. It adopts bidirectional pre-training by considering the context of words both preceding and following the target word. The BERT model consists of a transformer encoder, segment embeddings, and positional embeddings that establish relationships between words within the broader context. Moreover, BERT is pre-trained on a large corpus and later fine-tuned for specific tasks on smaller datasets. In pre-training, BERT learns a wide range of linguistic patterns and structures through tasks such as masked language modeling (MLM) and next-sentence prediction (NSP). After pre-training, BERT is fine-tuned by employing task-specific datasets with a low learning rate. This helps the model adapt to new tasks while much of its pre-trained linguistic knowledge is still retained. The ability to be fine-tuned for distinct tasks considerably improves BERT's performance in various NLP applications. Therefore, multimodal trans-

former models incorporate BERT or its derivatives for text processing as BERT combines visual and textual information to generate descriptive audio outputs to assist visually impaired individuals [10].

To integrate visual features into textual representations, BLIP uses a cross-attention mechanism that strengthens interactions between image and text encoders. This addition is crucial as it links textual information with the content and structure of visual data [11]. The cross-attention mechanism allows textual features (e.g., word embeddings from BERT) to attend to visual features (e.g., image embeddings from ViT), which facilitates interactions between the two modalities. Accordingly, the model aligns the outputs of the visual encoder with the text encoder by focusing on visual features relevant to the words or phrases in the text. In other words, the model utilizes image embeddings to influence the processing of text embeddings to integrate visual and textual data. As such, through training on a benchmark dataset, a model with a cross-attention layer becomes capable of (1) aligning images and text embeddings, (2) learning richer representations of text grounded in image data; and (3) better understanding the relationship between visual elements and the words that describe them [12].

Fourth, a causal self-attention mechanism is integrated into the model as a part of the text decoder [13] to increase its capabilities in image captioning and visual question answering (VQA). The inclusion of the causal self-attention layer plays a critical role in ensuring logical sequentiality in generated outputs. While the cross-attention mechanism manages the alignment of words with visual features to enhance their contextual relevance, the causal self-attention ensures that each word in the description is based on the preceding ones. By effectively combining these mechanisms, the model maintains a logically sequential language flow while grounding its outputs in relevant visual information.

2.3. Joint Pre-Training of BLIP Model

The BLIP model is jointly trained with image-text contrastive learning (ITC), image-text matching (ITM), and image-conditioned language modeling (LM). To optimize its performance in the pre-training phase, ITC is applied between the image encoder and text encoder to align feature spaces. ITM is employed between the image encoder and image-grounded text encoder to learn joint image-text features by capturing fine-grained alignment between vision and language. Finally, LM is integrated into the image encoder and image-grounded text decoder to generate text descriptions based on the given image.

In particular, BLIP treats VQA as an answer-generation task [14]. The VQA model uses the correct answer as the target during training, and it is fine-tuned by minimizing a loss function associated with the language model.

3. Methodology

3.1. Algorithm Architecture

We developed an AI-powered device to assist visually impaired individuals in navigating restroom facilities by providing real-time spoken messages. To establish its architecture, we followed the recommendations proposed by Li et al. [2] and used the BLIP model. The BLIP model consisted of an image encoder, an image-grounded text encoder (i.e., a cross-modal encoder), and an answer decoder.

3.2. Speech Synthesis Program

In this study, the captions (i.e., descriptions) produced by the BLIP model were converted into speech using speech synthesis technology. We selected gTTS to request Google to convert the target text into speech and return the audio output. Given the easy

access to an internet connection in the restroom where the study was conducted, the TTS library was used.

3.3. Training BLIP Model

We applied the COCO dataset to train the BLIP model for image captioning. COCO is a benchmark dataset with a vast collection of images that feature complex content and diverse scenes. The effectiveness of TTS models heavily depends on the datasets used for training [15]. To perform the VQA task, a large benchmark dataset such as COCO was essential for training the model to understand and answer questions about visual content. This study uses images from the COCO dataset, which is released under a Creative Commons Attribution 4.0 License (CC BY 4.0). Figure 1 shows the sample images.



Figure 1. Samples of COCO dataset.

To train the BLIP model, we set the activation function to rectified linear unit (ReLU) and adjusted the batch size and number of epochs to accelerate the experimental process. Figure 2 illustrates the training process.

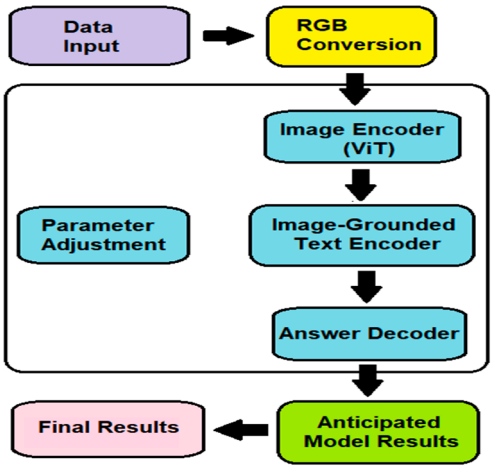


Figure 2. Training process of BLIP model.

3.4. Execution Environment

The developed model was trained and executed in a specific environment. Table 1 summarizes the execution environment.

Table 1. Execution environment.

Operational System	Environment	Programming Language	Framework	Training Color Space	Image Resolution	CPU	Router
Windows	Anaconda	Python	Torch	RGB	QVGA (320 × 240)	Intel Core i7 2.60 GHz	TP-Link (LAN)

3.5. Device

We implemented the ESP32-CAM in the head-mounted device. The in-built camera captured and transmitted images to the server over a Wi-Fi connection. These images were used as inputs and processed by the BLIP model to generate descriptive captions. The captions were then converted into speech by gTTS, and finally, the speech was played aloud through the speaker.

3.6. Metrics for Evaluation of Model Performance

To evaluate the performance of the developed, two metrics were adopted: SacreBLEU scores and mean opinion score (MOS). SacreBLEU is an improved version of the bilingual evaluation understudy (BLEU). It increases the consistency of the old version by standardizing calculations and employing a fixed tokenization scheme. The higher the SacreBLEU scores are, the better model performances. SacreBLEU scores serve as a quantitative metric for vision-language models such as BLIP [16]. To obtain SacreBLEU scores, we selected a Python library SacreBLEU. The scores were calculated by comparing the reference (i.e., a list of one or more descriptions of an image) with the hypothesis (i.e., the caption generated by the BLIP model). An English professor was invited to write the references for all the images. MOS is a qualitative standard to evaluate the quality of spoken messages generated by TTS programs. To obtain MOSs, human raters were invited to assess speech based on criteria such as intelligibility, accuracy, fluency, and naturalness [17]. The arithmetic average of all raters' responses on the Likert scale was computed to represent MOS. To follow these procedures, we invited five human raters to answer a five-point scale that assessed the accuracy and intelligibility of the speech ("1" indicating "Bad" and "5" indicating "Excellent"). The mean score was computed and used as MOS.

4. Results and Discussion

The SacreBLEU scores ranged between 35.08 and 91.22, with an average of 59.43. Although SacreBLEU is a more effective metric for image captioning models such as BLIP than other models [18]. There is no consensus among researchers regarding its threshold values. Nevertheless, most scholars agree that a SacreBLEU score above 30 indicates, at the very least, an acceptable level of model performance [19]. According to this criterion, the developed BLIP model generated captions to satisfactorily describe the images. The BLIP model described an image in which a toilet and a pack of toilet paper were present. The generated caption "there is a toilet in a small bathroom with a toilet paper dispenser" appeared (Figure 3).



Figure 3. Model-generated description of the toilet and pack of toilet paper.

The caption indicated that the BLIP model successfully detected the toilet and toilet paper. However, it mistakenly identified a pack of toilet paper as a toilet paper dispenser. As the shape of the pack was similar to that of a dispenser, and the model was trained to

recognize the dispenser rather than the toilet paper pack, it did not accurately describe the latter. Another image was described with the caption "there is a white shower and a door with a knob and a sink". The BLIP model adequately described the objects (Figure 4). The BLIP model correctly described the door with a knob. Moreover, visual features from the reflection of the sink allowed the model to detect its presence accurately. The model mistook a towel ring for a white shower. "A white shower" in the description must be "a white showerhead". As a showerhead and a towel ring look similar, the BLIP model could not differentiate between them. Nevertheless, most objects were successfully recognized and described, implying that the BLIP model accurately output descriptions of the majority of individual items in the image.

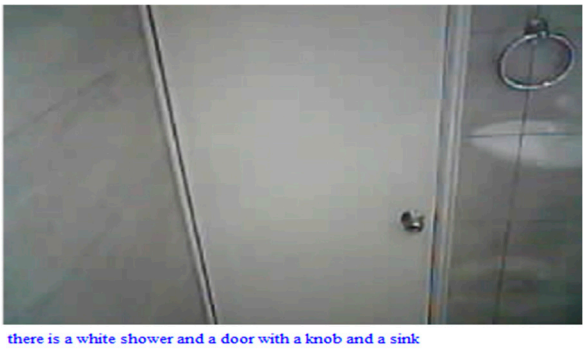


Figure 4. Model-generated description of the door, doorknob, and shower ring.

As Figure 5 illustrates, the BLIP model accurately detected the white door with a knob but failed to identify the remote control and switches on the wall. The image lacked prominent visual features for these objects and the model was not trained on enough instances of remote controls and switches. The small size or subtle features of these objects might be overlooked. However, the BLIP model successfully recognized and described the door and the knob.

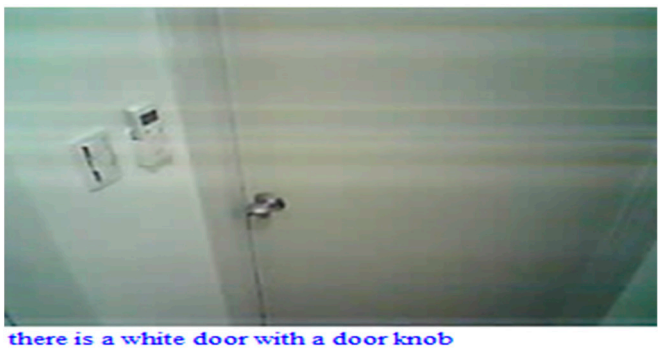


Figure 5. Model-generated description of the remote control, switches, door, and doorknob.

Collectively, the SacreBLEU scores suggested that the BLIP model precisely described most objects in the images taken and transmitted by the camera. Although several items were mistaken or undetected, the model performed well. We utilized MOS to estimate the quality of spoken language produced by gTTS. The MOS was 3.8, indicating that the speech quality was at an acceptable level [20]. The synthetic speech was natural, clear, and understandable. Using the device, users can obtain clear and instant spoken directions while navigating their surroundings.

The developed AI-powered device provides reliable, accurate, and comprehensible speech by integrating the BLIP model with gTTS into the head-mounted camera. It provides real-time audio descriptions to help the visually impaired navigate and locate restroom

facilities, thereby preventing accidents or injuries. The SacreBLEU and MOS scores strongly attest to the achievement of this goal.

We trained the BLIP model by using the COCO dataset. However, the dataset contains a limited number of images that feature restroom facilities, amenities, or objects. To improve the model's performance, additional images of restroom-related items must be included in the training data. Second, the use of gTTS requires a stable Wi-Fi connection. In an enclosed space such as a restroom, this connection is unreliable. Therefore, pyttsx3, a TTS library that works offline, needs to be adopted. As the developed device outputs spoken directions in English, descriptions in other languages are essential.

Author Contributions: Conceptualization, C.-S.H., N.-K.L., Y.-H.C. and S.-S.L.; methodology, C.-S.H., N.-K.L., Y.-H.C. and S.-S.L.; software, C.-S.H.; formal analysis, C.-S.H. and N.-K.L.; investigation, Y.-H.C. and S.-S.L.; data curation, C.-S.H. and N.-K.L.; writing—original draft preparation, C.-S.H.; writing—review and editing, C.-S.H.; visualization, C.-S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data may be available upon request.

Acknowledgments: The authors express their gratitude to Yuan-Chen Liu for his insightful suggestions for both the design of the device and earlier version of this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Abhishek, S.; Sathish, H.; K, A.K.; T, A. Aiding the visually impaired using artificial intelligence and speech recognition technology. In Proceedings of the 4th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 21–23 September 2022; pp. 1356–1362.
2. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for unified vision-language understanding and generation. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
3. Song, H.; Song, Y. Target research based on BLIP model. *Acad. J. Sci. Technol.* **2024**, *9*, 80–86. [CrossRef]
4. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft Coco: Common objects in context. In Proceedings of the Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
5. Patil, K.; Kharat, A.; Chaudhary, P.; Bidgar, S.; Gavhane, R. Guidance system for visually impaired people. In Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 25–27 March 2021; pp. 988–993.
6. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
7. He, C.S.; Wang, C.J.; Wang, J.W.; Liu, Y.C. UY-NET: A two-stage network to improve the result of detection in colonoscopy images. *Appl. Sci.* **2023**, *13*, 10800. [CrossRef]
8. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the Ninth International Conference on Learning Representations (ICLR 2021), Virtual Event, Austria, 3–7 May 2021.
9. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
10. Thomas, A.; U, S.; Barman, S. Third eye: AI based vision system for visually impaired using deep learning. In *Futuristic Trends in Artificial Intelligence*; Interactive International Publishers: Bangalore, India, 2023; Volume 2, pp. 101–112.
11. Tang, X.; Wang, Y.; Ma, J.; Zhang, X.; Liu, F.; Jiao, L. Interacting-enhancing feature transformer for crossmodal remote-sensing image and text retrieval. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5611715. [CrossRef]
12. Zeng, R.; Ma, W.; Wu, X.; Liu, W.; Liu, J. Image-text cross-modal retrieval with instance contrastive embedding. *Electronics* **2024**, *13*, 300. [CrossRef]

13. Yang, X.; Zhang, H.; Qi, G.; Cai, J. Causal attention for vision-language tasks. In Proceedings of the IEEE/CVF on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9842–9852.
14. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vis.* **2016**, *127*, 398–414. [CrossRef]
15. Orynbay, L.; Razakhova, B.; Peer, P.; Meden, B.; Emeršič, Ž. Recent advances in synthesis and interaction of speech, text, and vision. *Electronics* **2024**, *13*, 1726. [CrossRef]
16. Post, M. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation, Brussels, Belgium, 31 October–1 November 2018; pp. 186–191.
17. Streijl, R.C.; Winkler, S.; Hands, D.S. Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives. *Multimed. Syst.* **2016**, *22*, 213–227. [CrossRef]
18. González-Chávez, O.; Ruiz, G.; Moctezuma, D.; Ramirez-delReal, T. Are metrics measuring what they should? An evaluation of image captioning task metrics. *Signal Process. Image Commun.* **2024**, *120*, 117071. [CrossRef]
19. Computing and Reporting BLEU Scores. Available online: https://bricksdont.github.io/posts/2020/12/computing-and-reporting-bleu-scores/ (accessed on 22 August 2024).
20. Ping, W.; Peng, K.; Gibiansky, A.; Arik, S.; Kannan, A.; Narang, S.; Raiman, J.; Miller, J. Deep Voice 3: Scaling text-to-speech with convolutional sequence learning. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018; pp. 1–16.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.