International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

AI Assistant for Visually Impaired

Dr. Sagar BM/ Professor & HOD
Department of ISE
RV College of Engineering®
Bengaluru, India

Shashank SJ
Department of ISE
RV College of Engineering®
Bengaluru, India

Shashank V
Department of ISE
RV College of Engineering®
Bengaluru, India

Hitesh Belekeri
Department of ISE
RV College of Engineering®
Bengaluru, India

Promod J
Department of ISE
RV College of Engineering®
Bengaluru, India

***Abstract*—This paper presents an advanced assistive technology system aimed at improving accessibility and independence for visually impaired individuals. Utilizing artificial intelligence (AI) and computer vision techniques, the system provides real-time auditory feedback to users about their surroundings. The core components of the system include image captioning, face recognition, and depth estimation, all integrated to offer a comprehensive understanding of the environment. The system captures live video feed through a webcam, processes the images using pre-trained models like CLIPSeg for image segmentation and DPT for depth estimation, and generates textual descriptions of detected objects and their spatial distances. These descriptions are translated from English to Kannada using Google's translation services and converted into speech with the gTTS library, ensuring accessibility for Kannada-speaking users. Additionally, the system employs face recognition to identify known individuals in the vicinity, providing personalized auditory notifications. The combination of these technologies enables the system to offer context-aware assistance, helping visually impaired users to navigate and interact with their surroundings more effectively. Experimental results demonstrate the system's capability to deliver accurate, real-time feedback, highlighting its potential to significantly enhance the quality of life for visually impaired individuals.**

I. INTRODUCTION

In recent years, advancements in artificial intelligence (AI) have paved the way for innovative solutions aimed at improving the quality of life for individuals with disabilities. One such solution is the AI Assistant for Visually Impaired, a comprehensive tool designed to convert visual information into audio descriptions, thereby enhancing the independence and daily experiences of visually impaired users.

This AI Assistant integrates several state-of-the-art AI capabilities into a seamless web interface, providing users with a robust and user-friendly experience. It features image captioning to generate descriptive text for images, depth estimation to analyze spatial depth, and local language translation with audio feedback to ensure accessibility for non-English speakers. Additionally, it includes face recognition for identifying known faces and voice command recognition for hands-free operation. This report delves into the development, implementation, and user impact of the AI Assistant, showcasing how integrating these advanced AI functionalities can significantly empower visually impaired individuals in their everyday lives.

II. LITERATURE REVIEW

Huawei Zhang et al. [1] propose an image caption generation algorithm utilizing a Bi-directional Long Short-Term Memory (Bi-LSTM) structure. Traditional encoder-decoder models struggle with contextual information capture during word generation. The Bi-LSTM architecture addresses this limitation by integrating past and future information for enhanced prediction. It employs Forward LSTM (F-LSTM) and Backward LSTM (B-LSTM) decoders to extract semantic features. A subsidiary attention mechanism (S-Att) enhances semantic output by facilitating interaction between F-LSTM and B-LSTM. Additionally, alignment of hidden states enables semantic fusion. Experimental results on MSCOCO dataset show the effectiveness of the Bi-LSTM-s model, achieving a 9.7% improvement over LSTM-based approaches with a BLEU-4 score of 37.5.

Reshmi Sasibhooshan and et. al. [2] introduces an automatic caption generation system that employs attention mechanisms to produce descriptive captions from images with varying semantic detail. The framework utilizes a Wavelet transform-based Convolutional Neural Network (WCNN) for visual feature extraction, incorporating two-level discrete wavelet decomposition to capture spatial, spectral, and semantic details. A Visual Attention Prediction Network (VAPN) computes channel and spatial attention, enhancing feature extraction. Local features are integrated via contextual spatial relationships between objects. Coupled with a Long Short Term Memory (LSTM) decoder network, these components enable word prediction. Experimental evaluations on Flickr8K, Flickr30K, and MSCOCO datasets demonstrate the model's enhanced performance, yielding a CIDEr score of 124.2.

Simin Chen and et. al. [3] introduce a neural image caption generation (NICG) models excel in visual understanding but lack efficiency scrutiny. Real-time applications demand efficiency, yet NICG model's susceptibility to input-induced slowdowns remains overlooked.The authors propose

NICGSlowDown, an attack method, to probe NICG model's efficiency robustness. Their experiments reveal NICGSlow-Down's capability to subtly increase model latency by up to 483.86%, prompting a call for vigilance regarding NICG models' efficiency vulnerabilities in practical scenarios.

The paper titled Image Captioning using CNN and Transformers proposes an approach for generating descriptive captions for images by combining Convolutional Neural Networks (CNNs) and Transformers. The author Raju, K. [4] leverages the Transformer model's attention mechanisms to refine image features and generate captions effectively. They introduce adaptive attention within the Transformer-Decoder, allowing precise utilization of image information during caption generation. Through extensive training on the Flickr8K dataset, the proposed model achieves an impressive 86.21integration of CNNs and Transformers offers versatility and potential applications in various domains, promising advancements in image captioning tasks.

This paper by Indrani Vasireddy and et. al. [5] introduces an Image Caption Generator that combines Vision Transformers (ViT) and GPT-2, bridging computer vision and natural language processing (NLP). The system extracts image features using ViT and generates contextual descriptions with GPT-2, offering a user-friendly interface for receiving coherent captions. With a focus on accessibility, especially for the visually impaired, Their aim is to automate the creation of descriptive captions for images. By integrating computer vision and NLP techniques, the system analyzes image content to produce relevant descriptions, improving content accessibility and search capabilities. Additionally, it serves as assistive technology for the visually impaired, effectively interpreting and communicating image content. This paper illustrates the symbiotic relationship between computer vision and NLP, demonstrating their integration for transformative AI applications. The conference presentation will delve into technical aspects, highlighting the significance and potential impact of this integration on future AI applications.

The paper authored by Makav, Burak et al., [6] introduces image captioning approach designed to assist visually impaired individuals. It combines the VGG16 deep learning architecture for image feature extraction with the Stanford CoreNLP model for caption generation. Experimental results on the MSCOCO dataset demonstrate that the proposed approach outperforms existing methods in terms of captioning quality, as measured by metrics such as CIDEr and BLEU. The paper concludes by suggesting potential integration of the approach into smartphone or smart glasses applications for real-world use by visually impaired individuals.

Shuang Liu and et. al. [7] present a deep learning-based system for automatically generating textual descriptions of images. It outlines the challenges in accurately describing images and proposes a methodology using CNN for feature extraction and RNN for text generation. The system is evaluated using the Flickr8k dataset and achieves a satisfactory BLEU score of 0.64. The paper concludes by discussing the potential applications of the system, particularly its benefit for the visually impaired community, and suggests avenues for future research and improvement.

The paper by Manasa D and et. al. [8] addresses the gap in research concerning speech synthesis of local languages, particularly focusing on Kannada. It highlights the importance of converting linguistic scripts to speech for the benefit of individuals unable to read. While existing literature predominantly explores text-to-speech conversion in English, less attention has been paid to local languages. The authors propose an algorithm specifically designed for translating Kannada text to speech. This algorithm employs direct concatenation of speech coefficients extracted from prerecorded voice data for conversion. To assess its effectiveness, the proposed algorithm is compared with a widely used speech synthesizer, evaluating its performance against established standards.

This research by Pushpalatha K N and et. al. [9] presents a study on unidirectional translation from Kannada to English using Neural Machine Translation (NMT). RNN, particularly Long Short Term Memory (LSTM) units, are utilized within a Sequence to Sequence (Seq2Seq) framework to achieve translation. The authors compare their approach with Statistical Machine Translation (SMT) methods and demonstrate improved results, as evidenced by a higher Bi-Lingual Evaluation Study (BLEU) score of 86.32This paper by P. J. Antony and et. al.[10] explores the significant domain of language transliteration within natural language processing. Machine Transliteration involves converting characters or words from one language to another while retaining their phonological characteristics, making it crucial for orthographical and phonetic conversion. Named entity transliteration accuracy is particularly vital for machine translation and cross-language information retrieval. Hence, transliteration models must preserve the phonetic structure of words as closely as possible. This study focuses on transliterating English to Kannada using the publicly available Statistical Machine Translation (SMT) tool. The proposed technique achieves exact Kannada transliterations for 89.27% of English names, demonstrating its efficacy compared to SVM-based and Google Indic transliteration systems.

III. METHODOLOGY

The AI Assistant for visually impaired users integrates multiple advanced AI models to provide a comprehensive and accessible experience. The following methodology section details the functionality and rationale behind each model, including image captioning, depth estimation, local language translation with audio feedback, face recognition, and voice command recognition.

A. Image Captioning

The idea of using the image captioning model stemmed from its potential utility for the visually impaired to gain a basic understanding of their surroundings. Here is a brief working of the image captioning model used. For image captioning, the BLIP (Bootstrapping Language-Image Pre-training) model, specifically the Salesforce/blip-image-captioning-base version,

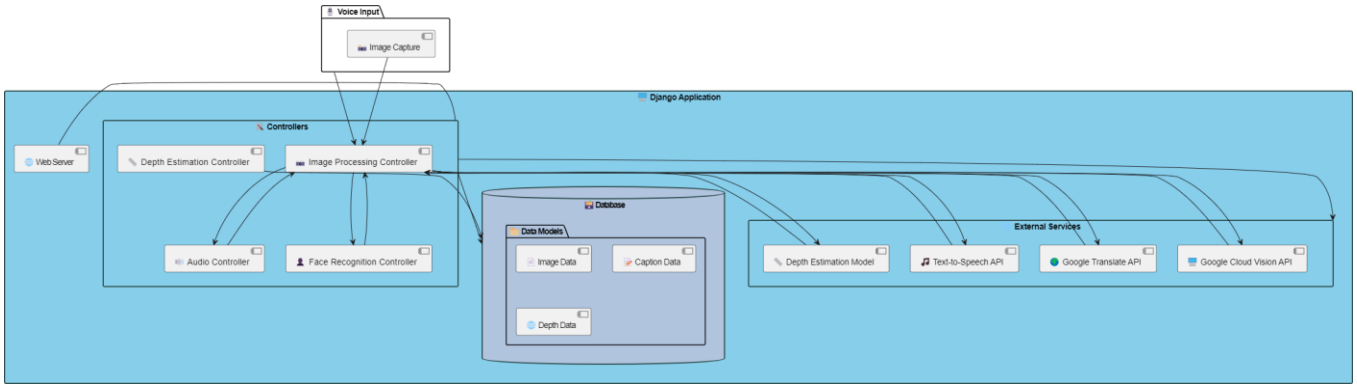

Fig. 1. Architecture Diagram of the system.

is employed. This model is chosen for its capacity to generate accurate and descriptive text for images, thereby providing visually impaired users with detailed descriptions of the image captured in their surroundings.

The image captioning model utilizes the Multimodal Mixture of Encoder-Decoder (MED) architecture within the BLIP framework. MED has multiple functions, as a unimodal encoder, image-grounded text encoder, and image-grounded text decoder. This architecture is pre-trained with three vision-language objectives—image-text contrastive learning, image-text matching, and image-conditioned language mod- eling—ensuring robust performance across various tasks.

Furthermore, BLIP incorporates CapFilt, a novel dataset bootstrapping method. This approach fine-tunes a pre-trained MED into two modules: a captioner, responsible for generating synthetic captions from web images, and a filter, tasked with removing noisy captions from original and synthetic texts. The captioner and filter synergize to enhance performance, leveraging bootstrapped captions to achieve substantial improvements across downstream tasks, particularly with a more diverse set of captions.

BLIP demonstrates state-of-the-art performance across a wide spectrum of vision-language tasks, including image-text retrieval, image captioning, visual question answering, visual reasoning, and visual dialog. It also excels in zero-shot performance when transferred to video-language tasks, specifically text-to-video retrieval and videoQA.

B. Audio Feedback in Kannada

The AI Assistant provides audio feedback in Kannada for visually impaired users. This process involves the following steps:

1) **Translation:** The output texts, generated by the models in English, are translated into Kannada using the Google Translate API.
2) **Text-to-Speech (TTS) Conversion:** The translated Kannada caption text undergoes conversion into speech using the `gTTS` (Google Text-to-Speech) library.
3) **Audio Playback:** The generated Kannada speech is played back to the user using the `pygame` library.

C. Depth Estimation

The depth estimation component plays an important role in providing spatial information to visually impaired individuals, helping them understand their surroundings and distances better. This component uses the Dense Prediction Transformer (DPT) model for depth estimation, which was trained on 1.4 million images for monocular depth estimation.

When a picture is captured , it goes through a pre-processing stage to ensure compatibility and optimal performance with the depth estimation model. The DPT model and its processor are then loaded from the transformers library. The DPTImageProcessor helps prepare the input data, and the DPTForDepthEstimation model generates depth predictions. The processor transforms the image into a tensor and normalizes it before passing it through the DPT model. The model then produces a depth map, which is a 2D array showing the estimated distance of each pixel from the camera.

The depth map is combined with the output of the image segmentation model, CLIPSeg. This integration helps estimate the distance of specific objects detected in the image. CLIPSeg generates binary masks for these objects, indicating their presence in corresponding pixel locations. These masks are applied to the depth map to isolate depth values for the detected objects, and the mean depth value of these areas is calculated to estimate their distance from the camera.

Using the calculated depth values, the system generates auditory feedback to alert the user about nearby objects. If an object is detected to be close, a warning message is translated into Kannada and spoken out loud using text-to-speech (TTS) technology. This feedback mechanism helps visually impaired individuals navigate their environment more effectively.

D. Face Recognition

Face detection technology was integrated to enhance situational awareness for visually impaired users. Utilizing the Face Recognition library, built with advanced face recognition capabilities powered by deep learning algorithms from dlib, which boasts an impressive accuracy of 99.38% on the Labeled Faces in the Wild benchmark. The system employs face detection algorithms to identify faces within the image. Upon

successful recognition, the system promptly provides auditory feedback to the user, announcing the name of the recognized individual.

E. Voice Commands and Controls

The voice command recognition system, built using the 'speech_recognition' library, captures and transcribes spoken instructions from users, facilitating intuitive interaction with the AI Assistant through hands-free commands.

F. Workflow of the finally integrated Model

This workflow operates within a Computer environment, and differs when operated with a Phone. Below is the description of the workflow of the finally integrated Model.

1) **Initialization:** Upon execution, the Python application initializes, importing necessary libraries and setting up configurations after which the camera module is activated, enabling real-time image capture from the device's camera.
2) **Image Captioning and Depth Estimation:** Upon pressing the space button, the image captioning and depth estimation processes are triggered. The captured image undergoes image captioning using the Salesforce/blip-image-captioning-base model, followed by depth estimation using the Intel/dpt-large model. Descriptive captions and depth information are generated for the captured scene.
3) **Face Recognition:** Upon pressing the 'f' button, the face recognition module is activated. Users initialize the system with images of known faces, associating them with corresponding identities. Real-time face detection and recognition are performed using the face_recognition library, identifying known individuals within the captured frames.
4) **Voice Command Recognition:** Upon pressing the 'j' button, the voice command recognition module is activated. Users issue commands verbally, which are transcribed and processed using the speech_recognition library. Regular expressions (regex) ensure precise recognition and execution of commands, preventing repetitions or unintended activations.

Throughout the workflow, the assistant provides feedback and responses to user inputs and system outputs. Auditory feedback in Kannada is synthesized and played back using the gTTS and pygame libraries, enhancing user accessibility and comprehension.

IV. RESULTS

The initial approach involved utilizing the VGG16 model for extracting image features and employing an RNN to generate captions. While this method performed adequately with images from the Flickr8k dataset on Kaggle, it failed to provide accurate descriptions for new images. The generated captions often lacked finer details and nuances, resulting in generic or inaccurate descriptions. Given the importance of precise image descriptions for visually impaired individuals, a shift was made to pretrained models. Figure 3 shows an example image taken by the system, and Figure 4 presents the corresponding caption generated for that image. The pretrained model provided a more accurate and detailed caption, demonstrating its effectiveness in capturing the essential elements and context of new images. For face images, the model identifies faces based on known face encodings. When a new face image is provided, the system matches it with the stored encodings to recognize the individual. Figure 5 shows the output when a picture of Ronaldo is shown to the system. The model accurately identifies Ronaldo, demonstrating its effectiveness in face recognition. For all recognized faces, the system provides auditory feedback in Kannada, ensuring accessibility for visually impaired users. This feedback includes the name of the identified individual, enhancing situational awareness and aiding in personal interactions. In Figure 6, the binary mask for object detection is used to estimate depth. Integrating depth estimation and face recognition enhances situational awareness for visually impaired users, providing accurate spatial information and identifying known individuals through auditory feedback, thus improving safety and interaction. This approach demonstrates the potential of advanced AI models for creating accessible solutions for those with visual impairments.

Figure 7 showcases a snapshot of the web application, integrating image captioning, depth estimation, face recognition, and voice commands. Tailored for visually impaired users, it features intuitive buttons for function initiation. Users can activate voice commands with a single click, enabling hands-free interaction with the assistant. Real-time camera input enhances situational awareness, while a double-click triggers image captioning and depth estimation processes.

Fig. 2. BLEU score of VGG16 model

Fig. 3. Test Image

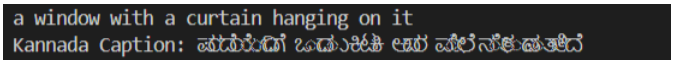

Fig. 4. Caption for the test Image

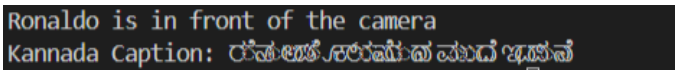

Fig. 5. Face Recognition result for Ronaldo's picture

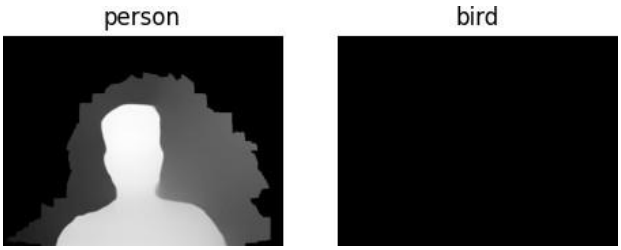

Fig. 6. Detected Objects

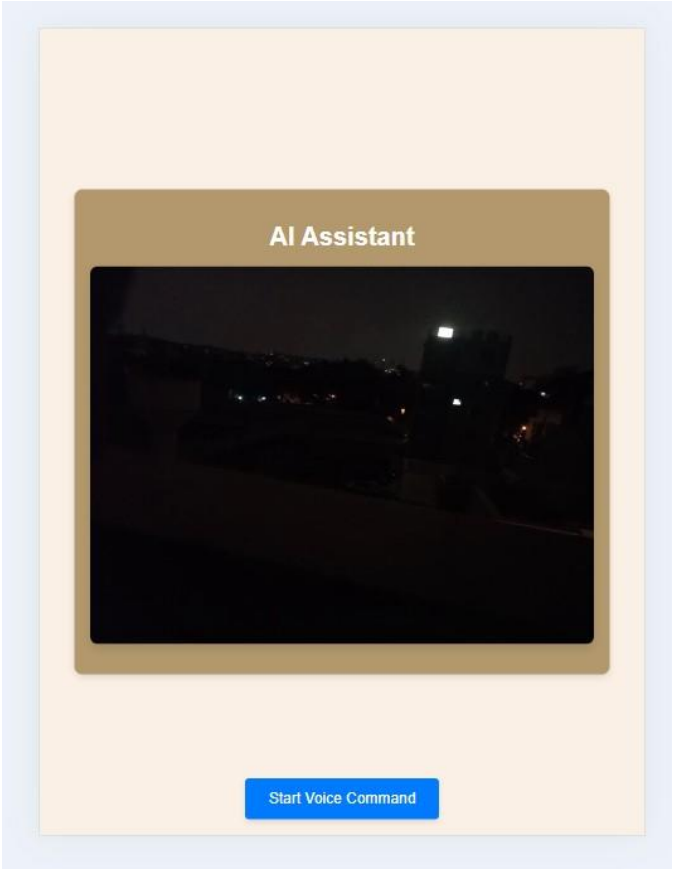

Fig. 7. Snapshot of web application

V. Conclusion

This project successfully uses advanced AI models to create a helpful tool for visually impaired people. The system combines image captioning, depth estimation, face recognition, and voice command recognition to provide a complete and easy-to-use experience. Using pretrained models improved the accuracy and detail of image captions and face recognition, making the feedback reliable and useful. The system gives spoken feedback in Kannada, making it practical for everyday use. The web application has an easy-to-use interface with real-time camera input and simple control buttons, allowing hands-free operation with voice commands. Combining depth estimation and face recognition improves situational awareness by giving accurate spatial information and identifying known faces, enhancing safety and interaction. This project shows how AI can create tools that improve the lives of visually impaired people.

VI. Future Enhancements

Provided below are few of the Future enahancements that could be done to the application.

- **Build an Android Application:** Develop an Android application to make the AI assistant accessible on mobile devices.
- **Integrate Real-time GPS and Navigation System:** Incorporate real-time GPS and navigation functionality into the application. The camera can be utilized to detect if the user is following the correct route.
- **Implement Panic Button:** Introduce a panic or danger button feature that enables users to send an SMS containing their last seen location to predefined emergency contacts, enhancing personal safety and security.
- **Integration with Google Search:** Enable users to perform Google searches directly within the application.
- **Enhanced Face Recognition:** Improve face recognition capabilities by leveraging large datasets from social media platforms.
- **Multilingual Support:** Expand language support to include additional languages, ensuring the AI assistant can cater to users from diverse linguistic backgrounds.
- **Advanced Natural Language Understanding:** Enhance the AI assistant's natural language processing capabilities to better understand and respond to complex user queries and commands.

References

[1] H. Zhang, C. Ma, Z. Jiang, and J. Lian, "*Image Caption Generation Using Contextual Information Fusion With Bi- LSTM-s*," IEEE Access, vol. 11, pp. 134–143, 2023, doi: https://doi.org/10.1109/access.2022.3232508.

[2] R. Sasibhooshan, S. Kumaraswamy, and S. Sasidharan, "*Image caption generation using Visual Attention Prediction and Contextual Spatial Relation Extraction*," Journal of Big Data, vol. 10, no. 1, Feb. 2023, doi: https://doi.org/10.1186/s40537-023-00693-9.

[3] S. Chen, Z. Song, M. Haque, C. Liu, and W. Yang, "*NICGSlow-Down: Evaluating the Efficiency Robustness of Neural Image Caption Generation Models*," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2022, doi: https://doi.org/10.1109/cvpr52688.2022.01493.

[4] K Lakshmipathi Raju, "*Image Captioning using CNN and Transformers*," International journal of advanced research in computer and communication engineering, vol. 13, no. 4, Mar. 2024, doi: https://doi.org/10.17148/ijarcce.2024.13469.

[5] Indrani Vasireddy, G.Hima Bindu, and Ratnamala. B, "*Transformative Fusion: Vision Transformers and GPT-2 Unleashing New Frontiers in Image Captioning within Image Processing*," International journal of innovative research in engineering and management, vol. 10, no. 6, pp. 55–59, Dec. 2023, doi: https://doi.org/10.55524/ijirem.2023.10.6.8.

[6] B. Makav and V. Kilic, "*A New Image Captioning Approach for Visually Impaired People*," 2019 11th International Conference on Electrical and Electronics Engineering (ELECO), Nov. 2019, doi: https://doi.org/10.23919/eleco47770.2019.8990630.

[7] S. Liu, L. Bai, Y. Hu, and H. Wang, "*Image Captioning Based on Deep Neural Networks*," MATEC Web of Conferences, vol. 232, p. 01052, 2018, doi: https://doi.org/10.1051/matecconf/201823201052.

[8] Manasa Dhananjaya, B. Krupa, and RK Sushma, "*Kannada text to speech conversion: A novel approach*," Dec. 2016, doi: https://doi.org/10.1109/iceeccot.2016.7955208.

[9] Pushpalatha Kadavigere Nagaraj, Kshamitha Shobha Ravikumar, Mydugolam Sreenivas Kasyap, H. Srinivas, and J. Paul, "*Kannada to English Machine Translation Using Deep Neural Network*," Ingénierie des systèmes d'information/Ingénierie des systèmes d'Information, vol. 26, no. 1, pp. 123–127, Feb. 2021, doi: https://doi.org/10.18280/isi.260113.

[10] P. J. Antony, V. P. Ajith, and K. P. Soman, "*Statistical Method for English to Kannada Transliteration*," Communications in computer and information science, pp. 356–362, Jan. 2010, doi: https://doi.org/10.1007/978-3-642-12214-9_57.

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "*BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*", arXiv preprint arXiv:2201.12086, 2022.