International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

# AI-Powered Vision-Language Accessibility Agent for Situational Awareness and Navigation Support

**Anand Sarangam**

***Group Technical Manager, USA***

***Abstract*--Artificial Intelligence (AI) can be considered a groundbreaking technology in augmenting accessibility and navigation assistance through employing the latest computer vision and natural language processing techniques. In this paper, the readers are presented with an AI-enhanced vision-language accessibility agent on situational awareness, trying to incorporate features of detecting objects, image captioning, and environmental risk assessment on a framework. The proposed system employs the use of the COCO128 as the object recognition database, and a trained BLIP vision-language model to produce descriptions of the world around it. The research design is an experimental research design but it involves evaluating the system in different indoor and outdoor situations by preparing datasets, implementing models, image processing, classifying risks, and evaluating the system. The agent of accessibility that is generated can identify objects that had high scores on the accessibility scales, develop sensible object labels within a scene, classify the environments into low, medium and high-risk categories to present convenient information about navigation. Experimental tests have proved that the integrated framework has the capacity of supporting environmental familiarity effectively and increasing accessibility by offering real time contextual information that will help in navigating it with safety. These results suggest that vision-language models used together with computer vision can greatly improve the perception of the scene, as opposed to traditional object detection methods. Although the proposed system continues to face difficulties with dynamism in settings and real-time applications, it provides an inspiring design of the future intelligent assistive technologies. Generally, this study assists in elaborating AI-founded accessibility remedies that could motivate individuals with reduced locomotion to move freely, possessing a greater situational awareness, and living improved lives.**

***Keywords*-- Artificial Intelligence, Vision-Language Model, Computer Vision, Accessibility, Navigation Support**

## I. Introduction

The development of artificial intelligence (AI) has revolutionized computer vision, in particular, due to the ability of the intelligent systems to perceive the information from the images and help the users [1]. Vision-language models, which are one of the recent developments for applications of AI technologies, allow the understanding of images, as well as generation of natural language information and responses.

Such vision-language models are very useful for blind and physically restricted people in order to be able to orient and make decisions independently in complex environments. Recent developments in the field such as deep learning, transformers, and multimodal technologies allowed developing systems which are able to recognize objects, generate captions, detect hazards and provide the information about the scene [2].

The systems of assistive technologies, which use sensors, GPS systems, or rule-based algorithms, are sometimes able to provide only a limited amount of information about the environment. However, the vision-language agents developed using AI technology have the abilities to recognize multiple objects, understand the context of the scene, detect hazards and generate natural language information [3].

The current research project is aimed at developing an accessibility agent with AI-powered vision-language capabilities that will be useful for improving situational awareness and navigation [4]. In terms of the system design, the object detection algorithms will be combined with the pre-trained BLIP vision-language model, which is expected to be used for the analysis of the uploaded images, recognition of objects in the environment, contextual captioning, evaluation of potential risks, and provision of recommendations for improving accessibility. The COCO128 dataset will be used for the training of the object detection part of the system.

The proposed accessibility agent will be intended to facilitate the acquisition of necessary information by the users and thus help them navigate more efficiently and conveniently. Hence, the research is related to the development of assistive technologies.

## II. Literature Review

### *2.1 Artificial Intelligence for Accessibility and Assistive Technologies*

Among the latest innovations, there is an emergence of AI technology as one of the major tools of creating the technologies of accessibility and making the lives of disabled people better.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

AI technology utilizes deep learning techniques in the analysis of images, audio and environmental information in order to provide intelligent assistance in daily tasks [5]. The recent development of computer vision technologies has made a significant impact on the field of accessibility, since it helps in the automatic detection of objects, people, road signs, obstacles and surroundings. In this way, the blind people can receive the information which they cannot receive without human help.

Recently developed multimodal AI technologies allow combining computer vision and natural language processing [6]. Not only is it possible to detect the object, but also it becomes possible to create the whole image of the situation using human language.

Additionally, intelligent assistance AI applications have moved on into various areas such as medicine, education, self-driven cars, and smart homes. The intelligent assistance system comprises functions such as voice assistance, object detection, interior navigation, and contextual recommendations [7]. However, there are certain limitations in existing applications, such as insufficient contextual understanding, difficulties in detecting multiple objects simultaneously, and reduced performance in varying environmental conditions. Therefore, development of intelligent assistants capable of performing object recognition and contextual assistance remains relevant.

### *2.2 Vision-Language Models for Scene Understanding*

One of the significant progressions in the multimodal AI industry is vision-language models as they combine vision and language functionalities. Although conventional object detection models are capable of recognizing the classes of objects, vision-language models are able to generate the full description of the relations between the objects and the picture in general [8]. Models of BLIP, CLIP, Flamingo, and GPT are performing really well in image captioning, visual question answering, and scene understanding.

BLIP (Bootstrapping Language-Image Pre-training) framework has become quite popular thanks to efficient implementation of the processes of images encoding and language generation, using less computational power than some of the alternatives [9]. The model is capable of generating context-aware captions that are necessary when the task requires more information from the picture than just the list of the objects.

The vision-language model also facilitates zero-shot learning, which is a way of recognising an unseen situation through semantic reasoning. It enhances navigation assistance because it describes paths, recognizes obstacles and explains connections between objects in the environment [10].

Nevertheless, today's models experience certain difficulties in dealing with complex environments, occluded objects, changes in lighting conditions, and processing in real time. This fact encourages more research on optimizing models for accessibility.

### *2.3 Object Detection and Risk Assessment for Navigation Support*

The object detection process has served to dominate intelligent navigation as this aids in the automatic identification of objects and localization in the environment. YOLO, Faster R-CNN, SSD and DETR (some of the best object detection techniques) have been doing a great effort in ensuring object detection is efficient and effective. This process is achieved through the creation of bounding boxes along with confidence scores which enable identification and localization [11].

When it comes to navigation assistance, it is important to understand that while the process of object detection begins with object identification, it must go further to determine the level of danger in the environment. Intelligence systems rely on the results obtained from the process of object detection to evaluate the level of danger associated with the objects identified for user safety [12]. This means that moving cars, sharp objects, stairs, or crowded areas can pose greater navigation dangers than static objects and large spaces.

Current studies use object detection together with contextual analysis to enhance situational awareness. Instead of conducting object analysis separately, smart systems consider relationships between objects, their movement and characteristics of the environment before providing suggestions on accessibility [13]. Thus, the effectiveness of navigability is improved, as well as the number of false alarms. Nevertheless, the challenge of obtaining efficient detection results in various light conditions, weather and crowd situations remains unresolved.

### *2.4 Integrated Vision-Language Accessibility Systems*

Various attempts have been made in recent times to integrate vision, NLP, and decision making into one accessibility system. Such unified systems provide end-to-end support by combining the elements of object detection, image captioning, contextual analysis, and environmental hazard detection. Such systems help in situational awareness by converting visual data into descriptive instructions on how to navigate through the environment [14].

Current day accessibility systems utilize data bases such as COCO to train object detection algorithms which can detect various objects in an environment. In addition to that, vision-language models are used to analyze objects in order to describe the scene while classifying the environment hazards based on their severity.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

Additionally, technologies such as cloud computing and edge AI have brought even more advancements to the accessibility systems through their ability to perform processing with minimal delays [15]. Devices containing AI models can carry out immediate environment analyses without any additional equipment. Although there are many advantages offered by these technologies, there is still room for improvement with regard to context detection, multilingualism, customized suggestions and reliability in rapidly changing environments. Therefore, the direction of future research should be toward the development of adaptable multimodal accessibility agents.

## III. Methodology

The research is going to employ experimental research methodology in the development and evaluation of an AI-powered vision-language accessibility agent for enhancing situational awareness and navigation aids. Experimental research methodology entails the employment of computer vision, deep learning and natural language processing technologies for enabling the identification of objects in the environment, generation of captions, navigation threat assessment and accessibility aid provision. The overall research methodology involves four major steps, namely data gathering, model creation, implementation and performance evaluation of the model.

### *3.1 Research Design*

The choice of the experiment-based research design lies in the ability to conduct the experimentation process by applying the artificial intelligence models. The research will include developing the vision-language accessibility agent that will analyze the images and generate the contextual information needed for navigation. The research will include a number of consecutive steps such as data collection, setting up the model, object detection, image captioning, risks assessment, and the final recommendation on accessibility.

### *3.2 Dataset Preparation*

The COCO128 will be selected as the data set of choice due to the availability of images that are highly labeled and contain objects present in the real world. The data set is made up of different environmental conditions such as kitchen, road, building, sports ground, vehicle, animals, and public place which can be used in the application of accessibility.

Data set verification is carried out before using the data set so as to ascertain that the images are usable. The image preprocessing stage entails scaling, normalization, and validation of the images.

### *3.3 Model Development*

This agent for accessibility will integrate the two tasks of object detection along with BLIP, which is a pre-trained model known as Bootstrapping Language-Image Pre-training. Object detection will be responsible for detecting the presence of different objects within an image with the coordinates for their positions, and confidence scores. BLIP vision-language model will focus on the complete image and generate description through caption generation related to the surroundings. The programming language selected for this task is Python because it has many libraries and frameworks related to artificial intelligence and computer vision.

### *3.4 Risk Assessment and Accessibility Recommendation*

Once the object is detected and the caption is produced, the environmental risk assessment phase takes place. Objects will be analyzed according to some pre-set safety standards in order to establish whether the surrounding environment is low-risk, medium-risk, or high-risk environment. For example, when vehicles, sharp objects or crowded areas are detected, this would imply a higher level of risks while normal indoor environments usually carry lower levels of risks. According to the results of the risk assessment, accessibility agents can offer appropriate recommendations.

### *3.5 System Evaluation*

The developed accessibility agent is evaluated using several images for different environmental scenarios. Object detection, confidence level, captioning abilities, environmental recognition ability, and risk assessment ability are among the areas of consideration during the evaluation. Test images considered in the evaluation include images for indoor kitchen environments, outdoor vehicles, sports stadium environments, and daily public environments images. Analysis is done on the output to determine whether the model can detect objects in the environment, caption and describe scenes accurately, and offer proper assistance. The evaluation is qualitative and is done through comparison between the captions and risk assessments and the actual scenario in the image. Confidence level is an indicator of efficient object detection.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

## IV. Result And Analysis

```
images = os.listdir(image_dir)
print(f"Total images found: {len(images)}")
print("First 5 images:")
for img in images[:5]:
    print(f"  • {img}")

Total images found: 128
First 5 images:
  • 000000000089.jpg
  • 000000000650.jpg
  • 000000000544.jpg
  • 000000000149.jpg
  • 000000000042.jpg
```

**Fig 1: COCO128 Dataset Verification Output**

The COCO128 dataset for developing the vision-language accessibility agent has been successfully verified in this figure. The system recognizes 128 images from the set and provides a sample image file name for validation. This check process helps to verify data that is available and accurate before analysis. During the study, a variety of environmental scenes are used for object detection, situational awareness and accessibility evaluation tasks, all of which are supported by the dataset.

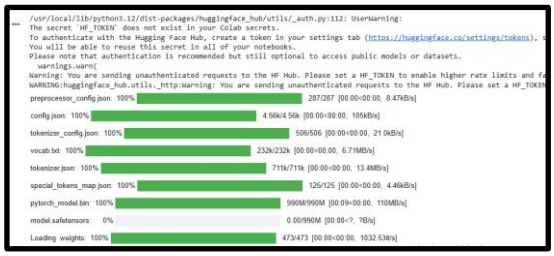


**Fig 2: Vision-Language Model Initialization Process**

This is the initialisation of the BLIP vision-language model and related configuration files. It downloads around 990 MB of pre-trained model weights and tokenizer resources needed for image captioning. This stage allows the accessibility agent to produce a description of the scene based on context. Once the model is successfully loaded, the ability to be aware of content, navigate it, and provide automatic accessibility support for the content in enterprise and assistive environments is provided.


**Fig 3: Object Detection and Risk Assessment for Kitchen Scene**

This figure illustrates an example of analysing an indoor kitchen environment with the accessibility agent. With a confidence score greater than 0.60, the system can identify objects such as an oven, microwave and knife. The created caption accurately describes the scene and has a LOW risk level because there are no immediate navigation hazards identified. This output will showcase the agent's ability to give information on environment awareness and accessibility in indoor environments.

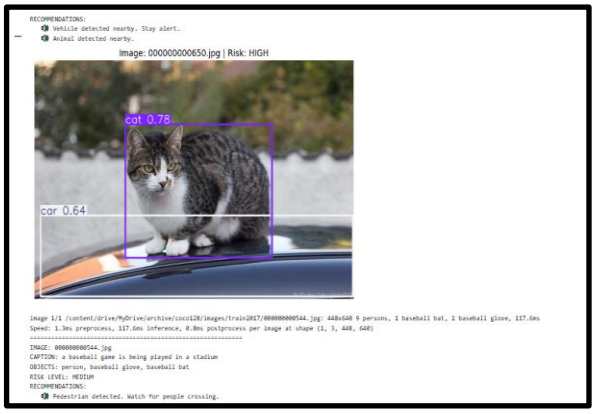


**Fig 4: Object Detection and Risk Assessment for Vehicle and Animal Scene**

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 08, August 2026)

There is a car and a cat standing on the roof. The vision-language agent not only correctly labels objects with high confidence scores (above 0.64), but it also provides a caption that describes the image. The system produces a HIGH-risk level and safety recommendations as a result of a vehicle being detected. This shows how well the agent can assist with real-time situational awareness and hazard detection.

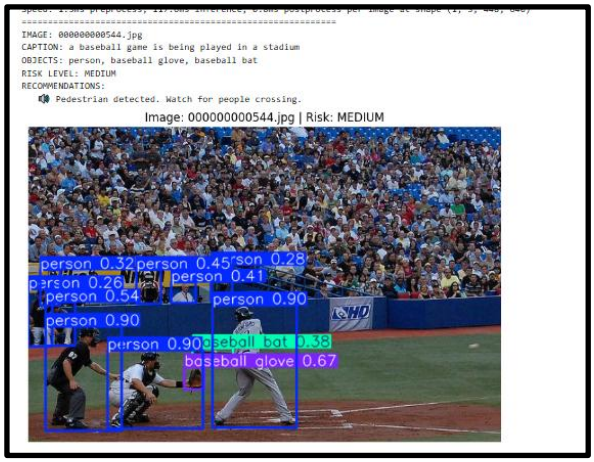


**Fig 5: Object Detection and Risk Assessment for Sports Stadium Scene**

A sports scene is shown, with multiple persons, a baseball bat, and a baseball glove detected. The generated caption is representative of the sporting event happening inside the stadium. Accessibility agent states that the environment is MEDIUM risk due to a high level of human activity and movement. The outcome shows good awareness and understanding of the environment for navigation and accessibility.

```
uploaded = files.upload()
image_path = list(uploaded.keys())[0]

# Run all modules
objects, annotated = detect_objects(image_path)
caption = generate_caption(image_path)
recommendations = get_recommendations(objects)
risk = assess_risk(objects)

# Display original image
original = plt.imread(image_path)
plt.figure(figsize=(12, 5))

plt.subplot(1, 2, 1)
plt.imshow(original)
plt.title("Original Image")
plt.axis("off")

plt.subplot(1, 2, 2)
annotated_rgb = cv2.cvtColor(annotated, cv2.COLOR_BGR2RGB)
plt.imshow(annotated_rgb)
plt.title("Detected Objects")
plt.axis("off")

plt.tight_layout()
plt.show()

# Display results
print("\n" + "=" * 60)
print("SCENE DESCRIPTION")
print("=" * 60)
print(f"  {caption}")

print("\n" + "=" * 60)
print("DETECTED OBJECTS")
print("=" * 60)
for obj in objects:
    print(f"  • {obj}")
```

**Fig 6: Image Upload and Processing Module Implementation**

This figure showed how the image upload, processing module which was developed in Python, was illustrated. The code is a combination of object detection, image captioning, accessibility recommendation generation, and risk assessment. The images that are uploaded automatically go through the module and create accessibility-related output. The architecture supports and enables autonomous agent action leveraging both computer vision and language knowledge to offer real-time accessibility support and safety advice.


**Fig 7: Final Output of the Vision-Language Accessibility Assistant**

The final results of the vision-language accessibility assistant are shown in this figure. A person and an umbrella have been detected in the image uploaded, with high confidence scores of 0.83 and 0.73, respectively. The system produces a context-specific caption, localizes recognized objects and assigns a MEDIUM risk level. This end-to-end output showcases the agent's ability to give content awareness, situational understanding, and accessibility advice to the end-user.

## V. Conclusion

Therefore, the aim of this research is the development of an AI-based vision-language accessibility agent that will increase the level of situational awareness of the user and provide him with safe navigation through the use of computer vision and NLP technologies. With the help of object detection, image captioning and risk assessment in one system, the proposed technology will provide context information for the user and improve his accessibility and safety.

Therefore, the integration of COCO128 and pre-trained BLIP vision-language models will allow identifying multiple objects and creating an appropriate environment description and improve the situational awareness of users.

As a result of this experiment, it became clear that the accessibility agent proposed is capable of analyzing different kinds of environments including kitchen, road, sports arena, and others. Besides, this agent successfully recognizes objects with high confidence scores, categorizes environmental risks into three groups (Low, Medium and High) and creates the corresponding recommendations about accessibility according to the scene type. This example proves the efficiency of using multimodal artificial intelligence solutions to interpret the environment in real-time and provide users with navigation assistance.

Yet, even though the proposed system holds promising performances, there are certain limitations that can be identified in it. The accessibility agent relies on pre-trained models and static images in terms of processing, and therefore it may have problems operating in dynamic environments or environments that lack proper lighting. Limitations of the proposed system can be overcome through future research which would entail implementation of real-time video processing, depth estimation, multilingual voice assistance, personalization of the navigation experience and edge computing for fast inference on the mobile devices. Other advancements may also include application of GPS, wearables and reinforcement learning. Overall, the proposed vision-language agent is a remarkable innovation in the area of intelligent assistive technologies.

## REFERENCES

[1] Bastola, A., Wang, H., Boroujeni, S.P.H., Brinkley, J., Moshayedi, A.J. and Razi, A., 2025. Driving towards inclusion: A systematic review of ai-powered accessibility enhancements for people with disability in autonomous vehicles. IEEE Access.

[2] Malatji, M., 2025. Augmented Intelligence Framework for Human–Artificial Intelligence Teaming in Cybersecurity. Human-Centric Intelligent Systems, 5(2), pp.151-180.

[3] Lisi, M.P., Fusaro, M. and Aglioti, S.M., 2024. Artificial agents delivering pain and touch on the embodied avatar of human participants are socially evaluated depending on the stimulus valence. A Metaverse for the Good, 51.

[4] Parrales-Bravo, F., Arias-Flores, H., Caguana-Alvarez, L., Dávila-Medina, M., Parrales-Bravo, C. and Vasquez-Cevallos, L., 2026. GUM: Gum Understanding Mission—A Serious Game to Improve Periodontitis Literacy Among University Students. Dentistry Journal, 14(4), p.242.

[5] Wu, M., 2024. Performing Objects: Ephemerality and Liveness in Contemporary Dance and Nonhuman Performance, 1975–2022 (Doctoral dissertation, The Ohio State University).

[6] Krupáš, M., Urblík, Ľ. and Zolotová, I., 2025. Multimodal ai for uav: Vision–language models in human–machine collaboration. *Electronics*, *14*(17), p.3548.

[7] Jain, G., Findlater, L. and Gleason, C., 2025. Scenescout: Towards ai agent-driven access to street view imagery for blind users. *arXiv preprint arXiv:2504.09227*.

[8] Yang, S., Huang, Y., Cai, W., Sun, S., Fang, F., He, Y., Xie, Y., Deng, J., Zhang, H., Song, J. and Zhang, Z., 2026, April. Egocentric Co-Pilot: Web-Native Smart-Glasses Agents for Assistive Egocentric AI. In *Proceedings of the ACM Web Conference 2026* (pp. 8862-8873).

[9] Yu, R., Lee, S., Xie, J., Billah, S.M. and Carroll, J.M., 2024. Human–AI collaboration for remote sighted assistance: perspectives from the LLM era. *Future internet*, *16*(7), p.254.

[10] Hu, J., Yan, C., Zheng, Y., Wang, Z. and Zhang, J., 2026. Position: Assistive Agents Need Accessibility Alignment. *arXiv preprint arXiv:2605.13579*.

[11] Emami, Y., Zhou, H., Reddy, R., Arani, A.H., Wang, B., Li, K., Almeida, L. and Han, Z., 2026. Large Language Model-Assisted UAV Operations and Communications: A Multifaceted Survey and Tutorial. *arXiv preprint arXiv:2602.19534*.

[12] Javaid, S., Khan, M.A., Fahim, H., He, B. and Saeed, N., 2025. Explainable AI and monocular vision for enhanced UAV navigation in smart cities: prospects and challenges. *Frontiers in Sustainable Cities*, *7*, p.1561404.

[13] Taloba, A.I., Alanazi, I.A., Shahin, O.R., Abass, I.A.M., Hussein, L.F., Abdelfattah, W.M., Bizon, N., Sallah, M. and Bedair, K., 2026. AI-Powered Robotics Framework for Assistive Navigation Industrial Automation and Smart Service Applications. *Appl. Math*, *20*(3), pp.697-708.

[14] Tami, M.A., Elhenawy, M. and Ashqar, H.I., 2025, May. Multimodal large language models for enhanced traffic safety: a comprehensive review and future trends. In *International Conference on Intelligent Systems, Blockchain, and Communication Technologies* (pp. 132-147). Cham: Springer Nature Switzerland.

[15] Rangarajan, V.T., Implementing Machine Learning for AI-Powered Solutions in Robotics, Computer Vision, and Natural Language Processing.