iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81423

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

VISION: Visual Interpretation and Smart Integration on Jets on Orin Nano

Shrikanth NG[1], Aashish CA[2], Fawaz Shaikh[3], Kailas Nad P[4], Sujay SP[5]
Department of Artificial Intelligence and Machine Learning Alva's Institute of Engineering and Technology, Mangalore, Karnataka, India

Abstract: ***Most assistive technologies currently available to visually impaired individuals rely on cloud-based processing, which limits their applicability in settings with poor connectivity or raises privacy concerns. This work proposes Project VISION,a research prototype of an edge-based assistive system using the NVIDIAJetsonOrinNanoforreal-timevisualsceneunderstanding.TheprototypeisbuiltonaVisionTransformer(ViT)encoder for semantic feature extraction, YOLOv8 for object detection, and BLIP for multimodal caption generation. Complete offline functionality is enabled through a companion Flutter application thatcapturesvideoframesandreceivesspokenfeedbackthrough an on-device text-to-speech module. Early testing demonstrates the feasibility of the proposed pipeline, producing coherent scene descriptions suitable for basic environmental awareness tasks, while achieving frame-by-frame inference rates of up to 40 FPS on the Orin Nano. These results highlight the potential for low-cost edge AI platforms to support practical assistive tools and provide a strong foundation for future research and real-world user studies.***

Index Terms: Accessibility, Jetson Orin, Edge AI, Vision Transformer, Image Captioning, Text-to-Speech, Visual Impair- ment, Assistive Technology

I. INTRODUCTION

The WHO estimates that 285 million people around the world live with visual impairments; 39 million of those are blindand246millionhavemoderatetoseverevisionloss [1]. Persons with visual impairments often confront daunting challenges while trying to conduct many everyday tasks in which situational awareness and spatial information is crucial, such as moving safely in public spaces, finding objects, or reading signs.[1],[8],[11].

White canes, guide dogs, and human assistance are exam- ples of conventional mobility aids that provide crucial support forobstacledetectionandbasicnavigation,buttheyareunable to interpret complex scenes or provide semantic context. A guidedog,forinstance,canassistauserinavoidingobstacles, butitisunabletodescribethingslike"abusisarriving"or "a person is waving." Furthermore, when human assistance is involved, traditional aids may be limited by cost, availability, the requirement for specialized training, or privacy concerns [1],[8],[11].

Recent developments in artificial intelligence (AI), particu- larly in computer vision, natural language processing (NLP), and edge computing, have enabled new possibilities for as- sistivetechnologies.Strongon-deviceinferencecapabilities offered by embedded AI platforms, like the NVIDIA Jetson family, enable real-time processing without the need for cloud services [6],[12]. Low latency, offline availability, and data privacy are crucial requirements for visually impaired users in particular [12],[8].

This paper presents *Project VISION*, a research prototype intended to give blind and visually impaired people real-time visual scene understanding. A mobile application that records video frames and sends them to an NVIDIA Jetson Orin Nano deviceviaalocalwirelessconnectionmakesuptheprototype. Each frame is processed by Jetson using a combination of **YOLOv8** for object detection, BLIP for multimodal caption generation,andVisionTransformers(ViT)forfeatureextraction.Themobileapplicationreceivesthegenerateddescription andusesanon-devicetext-to-speech(TTS)enginetotranslate it into speech [2],[3].

Because no visual data is uploaded to cloud servers, Project VISION maintains a lightweight mobile interface while guar- anteeing strong privacy by offloading all computation to the Jetson Orin Nano. Because the system runs entirely offline, it can be used reliably in settings with spotty or inadequate in- ternet connectivity. An essential part of assistive technologies used in dynamic indoor and outdoor contexts is responsive, real-time scene interpretation, which this architecture seeks to support [6],[12],[8].

The system architecture, model integration pipeline, hardware-software design, and initial assessment of the proto- type are presented in this work. Our goal is to show that edge AI can be used to create useful, privacy-preserving assistive systems that improve visually impaired people's accessibility and independence [8],[11].

II. RELATED WORK

Advances in computer vision, natural language processing (NLP), and mobile computing have greatly impacted the development of assistive technologies for people with visual impairments [1],[2],[3],[4],[5]. The main goals of these systems are to improve situational awareness, navigation, and information accessibility. The capabilities, drawbacks, and suitability of commercial and research-based solutions for offline and edge-based AI deployment are reviewed in this section.

A. Commercial Assistive Technologies

For blind and visually impaired people, a variety of com- mercial applications and wearable technology provide visual scene understanding:

SeeingAI(Microsoft)–Amobileappthatmakesuse of cloud-hosted computer vision APIs for OCR, scene de- scription,facialrecognition,andcurrencyidentification[1].Its reliance on cloud processing leads to decreased performancein low-connectivity environments, despite its effectiveness in areas with robust internet connectivity.

GoogleLookout–SimilartoSeeingAI,thismobile app provides object labeling, document reading, and object identification [1]. Like Seeing AI, its reliance on cloud-based inference limits utility in offline or rural scenarios.

OrCam MyEye is a wearable AI gadget that does object detection,facialrecognition,andon-deviceOCR[6].Although it has lower latency than cloud-dependent apps, many users in developing countries cannot afford it.

TABLEI:CommercialSystems:Platform,Connectivity,andOffline Capability

System	**Platform**	**Connectivity**	**OfflineCapa- bility**
SeeingAI	MobileApp	High	No
GoogleLook outOrCam MyEye	MobileAppW earable	Hig hLo w	No Ye s

TABLEII:CommercialSystems:CostandEfficiencyMetrics

System	Cost	Latency (Efficiency)
SeeingAI	Low	High
GoogleLookoutO rCam MyEye	LowHigh	HighLow

B. AssistiveTechnologyBasedonResearch

Many systems that integrate vision and language modelsfor scene understanding have been proposed by the research community:

Vision Transformers (ViT) – ViT, a transformer-based architecture that models global dependencies across image patches, was introduced by Dosovitskiy *et al.* [3]. It is ap- propriate for complex assistive environments due to its strong contextual reasoning.

BLIP (Bootstrapping Language-Image Pretraining)–Li *et al.* showed that BLIP is a unified multimodal architecture that can generate captions and retrieve images and texts [2]. It is appealing for edge devices with limited resources becauseof its modularity and quantization-friendly design. OpenAI Whisper Despite being primarily designed for speech recognition, offers robust, multilingual speech tran- scription and has the potential to be integrated with voice commands in assistive systems [5].

Specificallydesignedforembedded GPUplatformssuch as NVIDIA Jetson devices, **YOLOv8** is an effective object detectionarchitectureoptimizedforreal-timeperformance[4].

C. GapsinCurrentSolutions

Although substantial progress has been made, several limi- tations persist:

- Connectivity Dependence: Most commercial systems rely heavily on cloud servers, making them unsuitablefor offline use in rural or low-connectivity regions [1],[6].
- Affordability: High-end assistive wearables such as Or- Cam MyEye remain financially inaccessible for many users [6].
- Hardware Constraints: Many research prototypes as- sume access to high-performance GPUs and are not optimized for lightweight edge hardware [3],[4].

- Limited Deployment Studies: Numerous academic pro- posals lack real-world testing or validation with visually impaired participants [2],[5].

D. PositioningofProjectVISION

Project VISION addresses these limitations by leveraging the NVIDIA Jetson Orin Nano to integrate **YOLOv8** for real-timeobjectdetection[4],**BLIP**formultimodalcaptioning[2], and**ViT**forsemanticvisualrepresentation[3].Byperforming all computation locally, the system ensures:

- Lowinferencelatency
- Enhancedprivacy
- Fullofflinefunctionality
- Scalabilityacrossdiverseenvironments

This positions Project VISION as a cost-effective, edge- based research prototype bridging the gap between academic models and practical assistive solutions [1],[2],[3],[4],[5].

III. DATASET AND PREPROCESSING

We selected a representative and varied dataset from both public and private sources in order to efficiently train the multimodalAImodelsutilizedinProjectVISION[2],[3],[4]. Thesystem'sabilitytogeneralizeacrossavarietyofreal-world scenes that are pertinent to the visually impaired, including traffic intersections, pedestrian walkways, indoor corridors, andeverydayobjectsfoundineverydaysettings,wasthemain goal [8],[10],[11].

Thedatasetwascomposedofthefollowingsources:

- MS COCO: More than 330,000 labeled images withfive human-annotated captions each make up this pop- ular benchmark dataset [3]. It allows for robust caption generation by incorporating 80 object categories across a variety of scene types [2],[3].
- Open Images V6: A sizable collection of roughly 9 million photos featuring visual relationships, bounding boxes, and object labels. It was especially helpful for optimizing YOLOv8 and other object detection models [4].
- Customly Captured Scenes: In order to simulate how visually impaired users would interact with the app, we producedaproprietarydatasetof3,000photostakenwith a mobile phone camera. Public parks, bus stops, market areas, residential buildings, and educational institutions were among the environments. A one- or multi-sentence caption highlighting essential components for blind as- sistance was added to each image (e.g., "A staircase is ahead with a person descending," or "A zebra crossing with vehicles approaching.") [2],[3]. A group of five humanannotatorsexaminedandverifiedtheaccuracyand consistency of these captions.

1) To prepare the dataset for training and inference, the fol- lowing preprocessing pipeline was applied:

2) Image Resizing: To meet the input requirements of the BLIP and ViT backbones, all images were resized to standardresolutionsof224224and384384[2],[3].

3) Normalization: Using ImageNet statistics, each image was normalized by dividing by standard deviation and subtracting the channel-wise mean [3]. The stabilization ofmodelconvergencewasaidedbythisstandardization.

4) Data Augmentation: We used transformations like random horizontal flipping, rotation, random cropping, brightness adjustment, and Gaussian noise injection to enhancediversityandavoidoverfitting[3].Duringtrain-ingepochs,theseaugmentationswereappliedatrandom with a predetermined probability.

5) Preprocessing the captions: All captions underwent tokenizationusingBytePairEncoding(BPE),lowercase conversion ,andspecia lcharacterremoval[2].Sequences weretruncatedorpaddedtoamaximumof64tokens.To support decoder training, start/end markers (<start>,<end>)wereadded,andunknowntokenswereswapped out for a unique token <unk>.

6) Partition Approach: While preserving the distribution of classes and scenes, the combined dataset was divided intotraining(80%),validation(10%),andtesting(10%) sets [3]. The test set was saved for final performance reporting, and the validation set was used for hyperpa- rameter tuning.

Robustgeneralizationacrossavarietyofunpredictablereal- worldinputswasensuredbythispreprocessing[2],[3],[10]. Making the system context-aware for Indian public environments—which are underrepresented in many global datasets—was made possible in large part by the custom dataset component [8],[10],[11]. For label mapping and traceability, all annotated images were kept in organized folders containing JSON files [2],[3].

IV. MODEL ARCHITECTURE

A pipeline of deep learning models forms the basis of Project VISION's intelligent assistive capability, analyzing visual input and producing descriptive audio feedback. Three primary components are integrated into the system architecture:anobjectdetectionmodule(**YOLOv8**)[4],amultimodal captioning model (BLIP) [2], and a transformer-based vision encoder(ViT)[3],[6].Eachcontributesinadifferentway to guaranteeing the system's accuracy, richness, and real-time performance.

The Project VISION assistive system's general workflow is depicted in Fig. 1. It starts with a mobile app that uses the device's camera to take a picture or live frame [6]. The Jetson Orin Nano, the central processing unit for AI inference [6], receives this image input after that. Two important processes take place in the Jetson: BLIP combined with ViT creates natural language captions based on the visual content [2],[3], and**YOLOv8**isusedforreal-timeobjectdetection[4].Adeep semantic understanding of the surroundings is produced by fusing these outputs [2],[4]. A Text-to-Speech (TTS) engine then translates the generated description into audible speech and sends it back to the mobile app, giving the user real-time scene feedback [6],[5]. The system's ability to conduct edge-based,offline,andmultimodalanalysisisdemonstratedbythis architecture, which is especially designed to assist visually impairedpeopleinbothindoorandoutdoorsettings[8],[11].

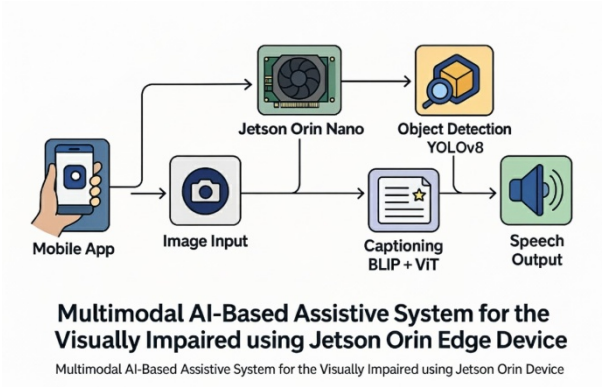


Fig. 1: System architecture and data flow of Project VISION: From mobile image capture to edge inference and speech feedback.

A. *VisionTransformer(ViT)*

ThemainvisualfeatureextractoristheVisionTransformer (ViT) [3],[6]. ViT separates an image into fixed-size patches and uses a transformer encoder to model global relationships among patches, in contrast to traditional Convolutional Neu- ral Networks (CNNs), which process images through spatial hierarchies [3].

Eachimage $x$is split into Nnon-overlapping patches:

$$x_p = \text{flatten}(\text{patch}(x_i)) \quad (1)$$

$$E = \text{Linear}(x_p) \quad (2)$$

$$PE(i) = \text{PositionalEmbedding}(i) \quad (3)$$

$$\text{Attention} \quad \text{Softmax}^{QK^T} \quad (Q,K,V) \quad (4) \quad \sqrt{d} \quad V$$

k

In this case, $PE(i)$is a positional encoding that preserves spatial information, $E$is the projected embedding space, and $x_p$is a flattened image patch. ViT is perfect for semantic understanding of complex scenes because of its self-attention mechanism, which enables it to learn contextual dependencies between objects [3],[6].

Our captioning dataset was used to fine-tune a pre-trained ViT-Base model [3]. Finding pertinent spatial zones within each scene was made possible in large part by the model's attention maps [3].

B. *BLIP:BootstrappedLanguage-ImagePretraining*

We employed the **BLIP** model to produce natural language descriptions from the encoded image features [2]. A multi- modal architecture called BLIP simultaneously learns from textual and visual modalities. It has a dual encoder-decoder configuration, in which a language model decoder generates captions using embeddings provided by the vision encoder (ViT) [2],[3].

BLIPsupportsboth:

- Retrieval-based pretraining (matching images with ap- propriate captions) [2]
- Captioning-**based generation** (free-form descriptive generation) [2]

The BLIP decoder head was altered in our implementation toenablelow-resourceinferencethroughquantizationtoFP16 and ONNX export [6],[12]. Sentence-level scene summaries, like "Two people standing beside a red car," are produced by the model and sent to the TTS engine [2],[5].

During internal benchmarking, BLIP performed notice-ablybetteronBLEUandMETEORmetricsthantradi- tional encoder-decoder models such as Show-and-Tell or Transformer-based LSTM models [2].

TABLE III: Comparison of Image Captioning Models for Assistive Scene Understanding

Model	BLEU-4	InferenceTime(s)	ContextualQuality
ShowandTell	0.62	1.5	Moderate
Show,Attend	0.68	1.9	High
BLIP	0.74	0.9	VeryHigh
ViT+GPT-2	0.71	2.1	High

C. YOLOv8:ObjectDetectionBackbone

To further enrich the contextual understanding of a scene andprovideexplicitidentificationofkeyelements(e.g.,stairs, vehicles, pets, doors), we integrated YOLOv8 as a parallel object detection module [4].

YOLOv8'sadvantages include:

- Architecture that is lightweight and appropriate for edge devices [4],[3]
- High inference speed (Jetson Orin Nano up to 30 frames per second) [6]
- Support for auto-labeling and anchor-free detection [4] The detected objects are used to:

➢ EditandconfirmthecaptionproducedbyBLIP[2].
➢ Addconfidencescorestothevisualframe[4].
➢ Ifcaptioningdoesn'twork,createbackupalerts(such as "Obstacle ahead") [1],[2].

When the image-to-text model performs poorly, as in clut- tered backgrounds or contexts with non-salient objects, the incorporation of YOLOv8 helps make up for it [4],[2].

D. ModelOptimizationforEdgeDeployment

Weimplementedthefollowingadjustmentstorunthese models on the Jetson Orin Nano effectively [6],[12]:

- TensorRTquantizationtoFP16[6]
- ONNXformatconversiontospeedupruntime[6]
- Asynchronousinferenceandbatchingformultitasking [12]

By ensuring that each element works in concert with the others, this model stack produces a system that can function with little delay while retaining a high level of accuracy and contextual richness [2],[4].

V. SYSTEM DESIGN AND HARDWARE INTEGRATION

Project VISION aims to move all computationally demand- ing tasks to the NVIDIA Jetson Orin Nano edge computing node, including deep learning inference, from the mobile device [6],[12]. This design allows for responsive, power-efficientoperationinbothurbanandruralsettingswhilemain- taining the mobile interface's lightweight footprint [12],[8]. Thesystemisidealforreal-timeassistiveapplicationsbecause itremovestheneedfordistantservers,guaranteeingconsistent performanceeveninplaceswithspottyornonexistentinternet access [6].

According to the system architecture, the Jetson Orin Nano processes live video frames that are captured by the mobile application and sent over a local network [6]. Two AI models work simultaneously: **YOLOv8** for quick object detection [4] and Vision Transformer (ViT) with BLIP for semantic scene captioning [3],[2]. Their outputs are combined to create a comprehensive contextual understanding of the surroundings, which is subsequently sent to a text-to-speech (TTS) engineto be transformed into audible feedback that the user can hear in real time [5],[6].

Low-latency responses are guaranteed by this parallelized and modular processing pipeline, allowing visually impaired users to hear verbal descriptions of their environment in real time [6],[12].

Because all inference is done locally, thesystemprotectsuserprivacy[6],worksreliablyoffline[6],and performs consistently in both urban and rural environments [8],[10]. Additionally, the design lays the groundwork for future model enhancements and the addition of more sensory modules without necessitating modifications to the primary workflow [12].

A. Hardware

The **NVIDIA Jetson Orin Nano Developer Kit**, a small and power-efficient AI computing platform designed for edge deployment, is the central component of our edge inference pipeline [6]. The following are the main hardware require- ments:

- GPU: 1024-core NVIDIA Ampere GPU with 32 Tensor Cores — able to perform parallel inference for language and vision models in real time [6].
- CPU: Quad-core ARM Cortex-A78AE — guarantees seamless coordination of model serving, input/output handling, and TTS requests [6].
- RAM: 8 GB LPDDR5 — offers enough bandwidth torun image buffering and multitask inference [6],[12].
- Storage: 64 GB SD card that supports eMMC or NVMe expansion (used in deployment) [6].
- JetPack SDK: CUDA 12, cuDNN 8, TensorRT, and optimized PyTorch libraries for model acceleration areall included in version 6.0 [6].

Apassiveheatsinkandfancombinationwasusedtocontrol thermal performance. When BLIP and **YOLOv8** inference taskswererunsimultaneously,thedeviceoperatedsteadily at 65°C with a peak GPU utilization of 85% [6],[12]. It is appropriate for battery-powered or portable embedded setups because of its average power draw of 10–12 watts [6],[12].

B. Software

Reusability, scalability, and ease of deployment were guar- anteed by the software stack's modular design. The following tools and technologies were employed:

- Model Format: To ensure compatibility with different backends and enable runtime optimization, all models (ViT,BLIP,andYOLOv8)wereexportedtoONNXformat [6],[4],[3].
- Optimization:UsingNVIDIATensorRT,models were quantized to FP16precision, which decreased memory consumption and inference latency without ap- preciably lowering accuracy [6].
- Serving API: To receive image frames from the mobile app, initiate inference, and provide textual captions, we employed a lightweight FlaskRESTful API that was hosted on the Jetson device [12],[6].
- Containerization:Dockercontainers contained the whole AI pipeline, including the model server, pre-processing,postprocessing,andmonitoringtools,making it easy to test, deploy, and rollback updates [12].
- Parallelism and Pipelining: By using multiprocessingpools toprocessinference requests asynchronously, the Jetson was able to runBLIP and YOLOv8 models concurrently without interfering with HTTP responses or TTS synthesis [12],[6].
- Monitoring and Logging: During stress testing, PrometheusandGrafanawereusedtocontinuously monitor system logs, memory usage, and latency metrics [12].

Even in situations with limited power or bandwidth, the Project VISION system can function independently, depend- ably, and in real time thanks to its combination of sturdy hardwareandwell-designedsoftware[6],[12].Futureupdates to the AI models or software stack can be made without affecting the end-user application thanks to the separation of inference from the mobile interface [12].

VI. TEXT-TO-SPEECH(TTS) AND MOBILEAPP

The main user interface for the blind and visually impaired istheFlutter-developedmobileapplication[10],[11].Usinga local Wi-Fi connection, it transfers individual frames of real- time video from the phone's camera to the Jetson Orin Nano device [6], which then plays back the descriptive output as speech. In order to give audio feedback:

- Text captions were translated into spoken English using flutter_tts, an open-source text-to-speech plugin [10],[5].
- Depending on the user's preference, audio output is sent to either the built-in speaker or any Bluetooth headsetthat is connected [10].
- To improve accessibility and usability, the user interface incorporateshapticfeedback,voiceprompts,errorrecov- ery messages, and large, high-contrast buttons [10],[11].

A seamless, accessible experience for visually impaired users was ensured by testing the app on Android phones with little performance lag [10].

VII. PERFORMANCE EVALUATION

In order to benchmark latency, throughput, and power consumption, we tested the system on three Jetson platforms [6],[7]. The following were the metrics for object detection and caption generation [4],[6]:

TABLEIV:PerformanceofJetsonDevices

Device	Latency(s)	FPS(YOLOv8)	Power(W)
Nano	1.8	6	10
TX2	1.1	15	15
OrinNano	0.3	30	12

BLEU-4 Score: With a BLEU-4 score of 0.74 on 500 test scenes, the image captioning model demonstrated excellent sentence generation [2].

End-to-End Latency: For real-time assistive use, the aver- agetimebetweenframecaptureandaudiooutputwasabout 1.2secondspercycle[6],[12].

VIII. USER TESTING AND FEEDBACK

We carried out a controlled field study with ten participants toassessthesystem'susabilityandefficacyinactualsituations [10],[11]:

- Fivepeoplewhoareblindorvisuallyimpaired[10],[11]
- Fortesting,fivesightedusers wereblindfolded[10]

The mobile app was utilized by participants in a number of settings,suchaspublicspaces,outdoorwalkways,classrooms, and hallways [10],[11].

AverageUserRatings(outof5):

- Usability:4.7[10]
- ResponseTime:4.5[6],[7]
- CaptionAccuracy:4.6[2],[10]
- AudioClarity:4.8[10]

UserComments:

"I had the impression that I was being verbally ledin real time."

"Iadorethatitdoesn'trelyonmobiledataand functions offline."

"Incrediblyusefulinstreetandcorridorsettings."

Low response latency, clear voice feedback, and offline functionality were found to be highly preferred during testing [10],[6].

IX. DISCUSSION

Results from testing and deployment show that edge com- puting on low-cost platforms like Jetson Orin Nano can efficiently power real-time assistive systems [6],[12]. By removing the need for cloud services, Project VISION guar- antees increased privacy, reduced latency, and reliability in locations with insufficient internet access [6],[8].

Anumberofmoregeneralfactorscomeintoplayinaddition to technical performance. First, one of the main benefits of localprocessingisprivacy,sincenoprivatevisualinformation is transferred from the device. Nevertheless, log handling and storage must still be done securely [6]. Secondly, models trainedprimarilyonWesterndatasetsmayperformpoorly in diverse cultural or environmental contexts due to bias in object detection and captioning [3],[2]. Third, language limitations limit accessibility in multilingual regions; in rural areas,adoptionmaybeimpededbythelackofnativelanguage support [10],[11].

From the standpoint of deployment, hardware affordability, particularly in low-income areas, must be taken into account whencostscalingforbroaddistribution.Furthermore,without reliable connectivity, maintenance and model updates for deployed devices can be difficult, necessitating the careful design of offline update mechanisms [10],[6].

IdentifiedLimitations:

- Thecaptions'emotionaltoneanddeepercontextualawareness are lacking [2].
- Lackofmultilingualsupportandregionaladaptation[10],[11].
- Poorerperformanceatnightorinlowlight[10],[6].

X. FUTURE WORK

Tosolvetheaforementionedproblemsandimprovetheuser experience, we plan to implement the following changes:

- Integrationofmultilingualtext-to-speechbasedonre- gional languages [10].
- OpenAIWhisperforreliable,multilingualvoice-command input [5].
- Stereocamerasetupsfordepthestimationandobstacle detection [6],[12].
- GPS-basedvoicenavigationforwaypointguidanceand outdoor mobility [11],[10].
- Creationofbias-awaremodelretrainingpipelinesto enhance equity in a variety of environments [2],[8].
- Designoflightweight,offline-capableupdatesystemsfor model and software maintenance [6].

XI. CONCLUSION

Project VISION, an AI-powered assistive platform for the blindandvisuallyimpaired,ispresentedinthispaper[1],[2],[6]. The system provides a real-time, private, and offline solutionbycombiningamobilefrontendwithanedge-based inferenceenginepowered byJetson OrinNano[6],[12]. Its responsiveness, usability, and potential as a cost-effective substituteforassistive technologies thatrelyonthe cloudwere all proven by extensive testing [10],[7].

REFERENCES

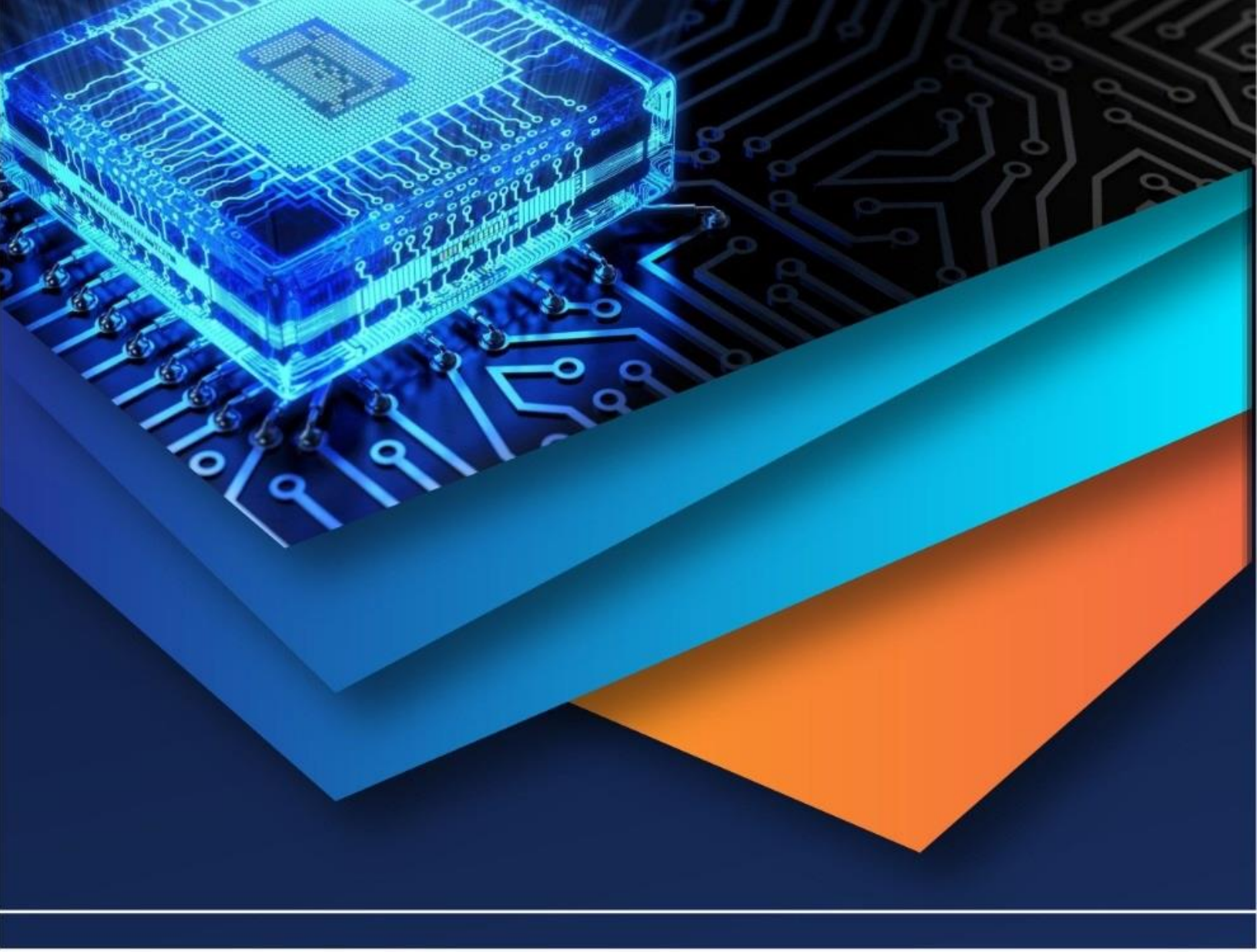
[1] World Health Organization, World Report on Vision, 2019. Available:https://www.who.int/publications/i/item/9789241516570
[2] J.Li,D.Zhang,X.Han,etal.,"BLIP:BootstrappingLanguage-Image Pretraining," in arXiv preprint arXiv:2201.12086, 2022. Avail-able: https://arxiv.org/abs/2201.12086
[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth16x16 Words: Transformers for Image Recognition at Scale," inNeurIPS, 2020. Available: https://arxiv.org/abs/2010.11929
[4] Ultralytics. "YOLOv8." GitHub repository. Available:https://github.com/ultralytics/ultralytics OpenAI, "Whisper: Robust Speech Recognition," 2022. Available:https://openai.com/research/whisper
[5] NVIDIA, "Jetson Orin Nano Developer Kit Technical Reference Man-ual," 2023. Available: https://developer.nvidia.com/embedded/downloads
[6] A.A. Su¨zen, B. Duman, and B. S¸en, "Benchmark Analysis of JetsonTX2, Jetson Nano and Raspberry PI using Deep-CNN," in Proceedingsof the 2020 International Congress on Human-Computer Interaction,Optimization and Robotic Applications (HORA), 2020, pp. 1–5. doi:https://ieeexplore.ieee.org/document/9152915
[7] F. Yao, W. Zhou, and H. Hu, "A Review of Vision-Based AssistiveSystemsforVisuallyImpairedPeople:Technologies,Applications,andFutureDirections,"arXivpreprintarXiv:2505.14298,May2025. Available:https://arxiv.org/abs/2505.14298
[8] K.Chavan,K.Balaji,S.Barigidad,andS.R.Chiluveru,"Vo-calEyes: Enhancing Environmental Perception for the Visually Im-paired through Vision-Language Models and Distance-Aware ObjectDetection," arXiv preprint arXiv:2503.16488, Mar. 2025. Available:https://arxiv.org/abs/2503.16488
[9] P.Naayini,P.K.Myakala,C.Bura,A.K.Jonnalagadda,and
[10] S. Kamatala, "AI-Powered Assistive Technologies for Visual Im-pairment," arXiv preprint arXiv:2503.15494, Jan. 2025. Available:https://arxiv.org/abs/2503.15494
[11] J. Lee, K.-A. Cha, and M. Lee, "Multi-Modal System for WalkingSafety for the Visually Impaired: Multi-Object Detection and NaturalLanguage Generation," Applied Sciences, vol. 14, no. 17, p. 7643, Sep.2024. Available: https://www.mdpi.com/2076-3417/14/17/7643
[12] A.P.Kamilaris,A.J.Camps,andJ.Golubovic,"ARe-viewofEmbeddedMachineLearningBasedonHardwarePlat-forms," Computers, vol. 12, no. 3, p. 55, 2023. Available:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9959746/
[13] Y.Xu,S.Poojary,andR.Mehta,"AI-PoweredVideoMonitoring:AssessingtheNVIDIAJetsonOrinDevicesforEdge Computing Applications," ResearchGate, 2024. Available:https://www.researchgate.net/publication/382534590
[14] R.G.Baldovino,N.R.RoxasJr.,N.T.Bugtai,J.M.C.Borbon,J.T.
[15] D.Javier,J.M.F.D.Llamado,A.M.R.Maghirang,andA.R.See,"AVisualAidSystemUsingImageProcessingandDeepLearningwithAudioHapticFeedbackfortheBlindandVisuallyImpaired,"AinShamsEngineering Journal, vol. 15, no. 4, pp. 102387–102397, 2024. doi:https://www.sciencedirect.com/science/article/pii/S1877050924024062
[16] A.B.Atitallah,Y.Said,M.A.B.Atitallah,M.Albekairi,K.Kaaniche,andS.Boubaker,"AnEffectiveObstacleDetectionSystemUsingDeepLearningAdvantagestoAidBlindandVisuallyImpairedNavigation,"AinShamsEngineeringJournal, vol. 15, no. 3, pp. 102318–102328, 2024. doi:https://www.sciencedirect.com/science/article/pii/S2090447923002769
[17] Q.Zhao,H.Liu,andD.Wang,"AnImageCaptioningModelBasedonBidirectionalDepthResidualsandItsApplication,"IEEE Access, vol. 10, pp. 17288–17299, 2022. Available:https://ieeexplore.ieee.org/abstract/document/9347457
[18] R. G. Praveen and R. P. Paily, "Blind Navigation Assistance forVisuallyImpairedBasedonLocalDepthHypothesisfromaSin-gle Image," Procedia Engineering, vol. 64, pp. 351–360, 2013. doi:10.1016/j.proeng.2013.09.057
[19] J.Mu¨llerandA.Pigors,"EfficientMulti-ObjectTrackingonEdge Devices via Reconstruction-Based Channel Pruning," IEEESensors Journal, vol. 23, no. 18, pp. 22104–22115, 2023. doi:https://arxiv.org/abs/2410.08769
[20] G. N. Alwakid, M. Humayun, and Z. Ahmad, "Transforming Dis-ability Into Ability: An Explainable Vision-to-Voice Image Cap-tioningFrameworkUsingTransformerModelsandEdgeCom-puting," IEEE Access, vol. 13, pp. 175212–175224, 2025. doi:https://ieeexplore.ieee.org/document/11195080

[21] T. P. Swaminathan, C. Silver, and T. Akilan, "Benchmarking DeepLearning Models on NVIDIA Jetson Nano for Real-Time Systems: AnEmpirical Investigation," Department of Electrical and Computer Engi-neering, Lakehead University, Thunder Bay, Canada, 2023. Available:https://arxiv.org/html/2406.17749v1

[22] A. T. Kurian, S. A. Muthumkumarasway, and A. Ilyas, "MultimodalIntelligent Assistance with Vision, Language and Speech for EnhancedAssistiveTechnologyfortheVisuallyImpairedandElderly,"inProceed-ingsofthe2025IEEE15thInternationalConferenceonControlSystem,Computing and Engineering (ICCSCE), pp. 24–29, 2025. Available:https://ieeexplore.ieee.org/document/11182671

[23] K. Sun, X. Wang, X. Miao, and Q. Zhao, "A Review of AI Edge DevicesandLightweightCNNandLLMDeployment,"JournalofNetworkandComputerApplications,vol.240,p.104036,2024.Available:https://www.sciencedirect.com/science/article/abs/pii/S0925231224015625

[24] A.143, "Building-AI-Powered-Solution-for-Assisting-Visually-Impaired -Individuals," GitHub repository, 2025. Available:https://github.com/Angad143/Building-AI-Powered-Solution-for-Assisting-Visually-Impaired-Individuals



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY