Ask VR: Vision Language Model Driven Scene Descriptor for Blind and Low Vision Users in VR Environment

Jaime B. Fernandez[1,2][0000−0001−9774−3879]✉, Ali Akbar Shah Syed[1,2][0000−0002−8045−3514], and Muhammad Intizar Ali[1,2][0000−0002−0674−2131]

[1] Dublin City University, Dublin, Ireland
[2] Insight Research Ireland Centre for Data Analytics, Dublin, Ireland
jaimeboanerjes.fernandezroblero@dcu.ie, aliakbarshah.syed@dcu.ie, ali.intizar@dcu.ie

Abstract. Virtual and Mixed Reality platforms such as Meta Quest and Apple Vision Pro have accessibility challenges for Blind and Low Vision (BLV) users due to their dependence on visual cues. Existing accessibility features like color filters and text resizing have limited support which makes users with severe vision loss unable to fully engage. In this research, a novel solution has been developed that integrates 3D scenic descriptor generation within Unreal Engine User Interface using a modular client server architecture. The developed system implements a locally hosted Vision Language Model (VLM) to generate scene descriptions. During the comparative testing of VLMs, Llava 7B was identified as the most effective in balancing semantic accuracy and perceptual quality. A key innovation is a multi-prompt strategy that can guard rail the complex scenes into structured and comprehensive audio segments that cover objects, spatial layout, and mood. As the functionality of scene description is activated with a simple key press, the system provides tailored feedback that enables BLV users to integrate with VR environment independently.

Keywords: Virtual and Mixed Reality · Blind and Low Vision · Vision Language Model · Gaming Technology · Assistive Technology.

1 Introduction

Virtual and Mixed Reality environments have become an integral part of gaming, cinematic arts, design and education. These platforms offer immersive experiences by simulating the real life scenarios [1], visualising complex 3D information [2], and enabling interactions through 6 degrees of freedom (DOF) movement [3][4][5]. However, the current generation of VR and XR systems, such as Meta Quest and Apple Vision Pro, is fundamentally built upon visual interaction. Key functionalities heavily rely on on-screen text, visual gestures and colour-coded indicators to convey information and signal interactivity. This deep reliance on visual centric design creates significant accessibility barriers for Blind and Low

Vision (BLV) users [6][7][8]. Standard accessibility features are often limited, typically limited to basic functionalities like colour filters, or adjustable text sizes, which fail to serve users with conditions beyond colour vision deficiency or those requiring general magnifications [9][10]. Comprehensive tools for individuals with total blindness, central vision loss or blurred vision are scarce. Crucially, advanced tools such as 3D screen readers have yet to be introduced, and developer support for creating them remains to be complicated [11][12]. As most of the platform like VRSight uses cloud based GPT 4V to provide the real time scene description, however it requires constant internet connectivity to work and sends visual data to third party servers killough2025demonstration. However, the recent advancements in Artificial Intelligence (AI) particularly in object recognition, scene interpretation and depth sensing offer a promising path forward for enhancing the VR accessibility [13][14][15]. These technologies have already shown good results in Augmented and Mixed Reality systems, supporting features like text magnification [16] and enabling BLV users to interact with touchscreens [17]. The system developed in this study pioneers an innovative solution to address the accessibility gap by integrating Vision Language Models (VLMs) with the Unreal Engine.

2 Selection of VLM

Vision Language Models (VLMs) are the models designed to process both visual and textual data by mapping them into a shared representational space. They are available in two main deployment types which are local VLMs, that run on-premise to ensure data privacy, offline capability, and customisation [18][19], and API based VLMs, which are hosted by third parties for ease of use and scalability but raise data privacy concerns [20][21]. For this study, local VLMs were selected to meet data privacy requirements and the objective of creating a standalone plugin for Unreal Engine. Due to an 8 GB size constraint, the selection focused on compact models, resulting in the choice of LlaVA-7B, which is widely recognised for its performance [22][23][24][25], BakLLava-7B, Moondream-2B; a lightweight "Nano-LLM" with just two billion parameters [26][27].

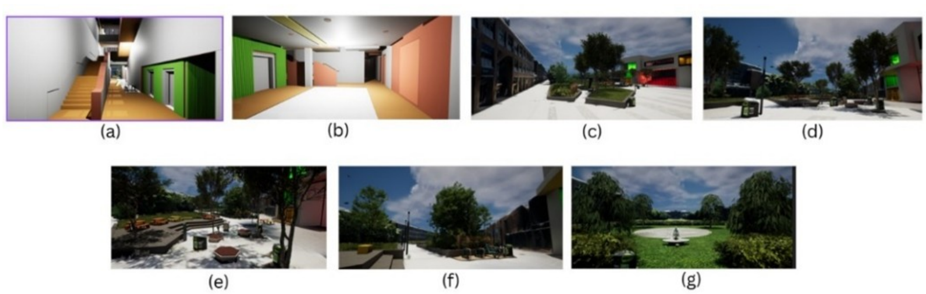

Fig. 1. Selected Images for VLM testing.

The VLM for this study was selected based on VIEscoring [28], which is a comprehensive evaluation framework for conditional image synthesis. VIEscoring combines Semantic Consistency (SC), which measures the accuracy of content understanding and alignment with image semantics and Perceptual Quality (PQ), that evaluates the realism, fluency and user comprehensibility of generated descriptions. The VLMs were tested on the image illustrated in Fig.1 and the performance scored in terms of SC, PQ abd the composite VIEscore are presented in following table.

Table 1. Model Performance Comparison

Rank	Model	VieScore	SC Score	PQ Score	Recommendation
1	Llava	0.477	0.761	0.406	Balanced Performance with highest VieScore
2	Bakllava	0.236	0.643	0.121	Good for semantic task but poor perceptual quality
3	Moondream	0.189	0.690	0.097	Inconsistent results

3 Technical Details

A modular client server architecture integrates Llava-7B into our VR environment to improve accessibility for BLV users. A custom plugin, ReadLocalTXT, imports text descriptions (.txt) into dynamic widgets, while file media source components handle VLM generated audio (.wav). This converts visual scene data into text and speech for users. The system consists of a VR client workstation and a server workstation hosting the VLM. They communicate using HuggingFace Hub, ensuring smooth VR experiences despite heavy model processing.

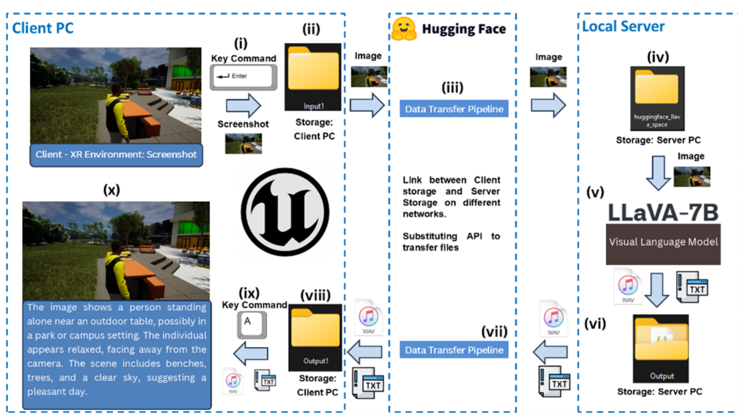

Fig. 2. General Methodology.

The methodology starts when the user presses a key to capture a screenshot of the VR screen (i, Fig.2), which is saved locally (ii, Fig.2). A Watchdog script uploads the image to the HuggingFace Hub (iii, Fig.2), where the server detects it and forwards it for VLM processing (iv, v, Fig.2). The VLM applies ten predefined prompts to generate text and audio description (vi, Fig.2). The prompts are mentioned in Table 2.

Table 2. Prompt Descriptions for Image Analysis

No/Name	**Prompt**
1. Brief Description	Provide a brief, clear description of this image in exactly 30-40 words. Focus on the main subject and key elements. (Key = A)
2. Detailed Description	Provide a detailed description in exactly 60-70 words covering main subject, setting/environment, visual details, actions, text content, spatial relationships, notable features, and mood/atmosphere. (Key = B)
3. Main Subject	Identify the main subjects and objects in exactly 30-40 words. Describe their appearance and what makes them the focal point. (Key = C)
4. Setting/Environment	Describe the setting and environment in exactly 30-40 words. Include location type, indoor/outdoor, and background context. (Key = D)
5. Visual Details	Describe colours, lighting, and composition in exactly 30-40 words. Focus on visual characteristics and artistic elements. (Key = E)
6. Actions	Describe what actions or activities are happening in exactly 30-40 words. Focus on movement and behaviours. (Key = F)
7. Text Content	Read and transcribe any visible text in exactly 30-40 words. Include signs, labels, or writing. State "No text visible" if none. (Key = G)
8. Spatial Relationships	Describe object positioning and spatial arrangement in exactly 30-40 words. Focus on how elements relate to each other. (Key = H)
9. Notable Features	Point out interesting or unusual aspects in exactly 30-40 words. Highlight what stands out as unique or significant. (Key = I)
10. Mood/Atmosphere	Describe the mood and emotional feeling in exactly 30-40 words. Focus on the atmosphere and impressions the image conveys. (Key = J)

The processed .txt and .wav files are returned by the Huggingface Hub (vii, Fig.2), saved in Unreal monitored directories (viii, Fig.2), and can be accessed with designated keys (ix, Fig.2). This pipeline ensures automated scene capture, analysis and accessible output without manual intervention (x, Fig.2).

The digital 3D models used in this work were created using Bentley Systems Itwin Capture Modeler and then imported into Unreal Engine as presented in previous work [29].

4 Results

Based on the results presented in Table 1, LlaVA-7B was selected in this study for the image analysis. For testing, a sample image was taken as illustrated in

Fig. 3 and evaluated using the ten pre-defined prompts (Table 2). The results are presented in the Table 3.

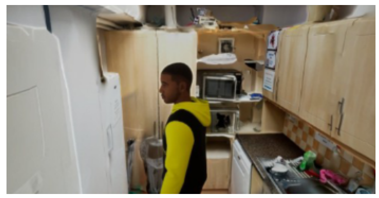

Fig. 3. Image of an indoor environment.

Table 3. Image Analysis Results by Prompt

Prompt	Results
1. Brief Description	The image shows a man standing in a kitchen, with a yellow sticker on his shirt. He appears to be looking at something out of frame. The kitchen has white cabinets and appliances, including an oven and sink.
2. Detailed Description	The results obtained from it are mentioned from prompts 3 to 10.
3. Main Subject/Objects	The primary subject in this image is a person standing in a kitchen area. The individual appears to be facing away from the camera, towards what looks like a dishwasher. The person is wearing a black shirt with a yellow stripe on the sleeve and a black cap.
4. Setting/Environment	This image seems to be taken indoors, specifically in a kitchen or a room that resembles one. There are cabinets visible, suggesting a functional space designed for food preparation and storage. The lighting is artificial, with overhead lights illuminating the scene.
5. Visual Details	The colors in the image are muted, with the person's shirt providing a contrast to the otherwise neutral tones of the kitchen environment. There is no visible text or writing in the image.
6. Actions/Activities	The individual appears to be either looking at something on the countertop or possibly interacting with the dishwasher. Their posture suggests they are standing still, not engaged in any immediate action.
7. Text Content	There is no visible text in this image.
8. Spatial Relationships	The person stands between two cabinets and to the left of what appears to be a dishwasher. The cabinets flank the individual on both sides, framing them within the kitchen space.
9. Notable Feature	One notable feature is the yellow stripe on the person's sleeve, which stands out against their black shirt. Additionally, there are various kitchen items visible in the background, such as a sink and what looks like a microwave or oven.
10. Mood/Atmosphere	The image conveys a sense of everyday life, with no strong emotions or dramatic actions taking place. It appears to be a candid moment captured within a domestic setting.

The results have both brief and detailed analyses. The detailed analysis includes the following prompts: main subjects/objects, setting/environment, visual details, actions/activities, text content, spatial relationships, notable features, and mood/atmosphere. Accordingly, selecting prompt 2 yields results that integrate the output obtained from all eight prompts (from 3 to 10).

5 Relevance and Impact

The developed system addresses a significant gap in the accessibility of the VR by providing a platform to the BLV users to experience an interactive environments through VLM driven scenic descriptions. With the integration of a locally hosted VLM with Unreal Engine, the system demonstrates how multimodal AI can transform visual content into meaningful auditory feedback. The system's approach scopes participation in VR and XR platforms by offering inclusive design possibilities that could be reached to underrepresented communities. In addition to that, the developed system highlights a novel intersection of accessibility, AI and multimedia systems for the multimedia modeling community by providing a strong foundation for future research on inclusive content generation in a VR environment.

6 Conclusions and Future Works

This study addresses the accessibility gaps in VR for BLV users by integrating a VLM driven 3D screen reader into Unreal Engine. The system is built on a modular client server architecture which automates scene capture, analyses it and processes the audio and text descriptions. LlaVA-7B was selected over Moondream and BakLLaVa for its balanced accuracy and quality which enables scenic descriptions without hindering VR performance. A multi-prompt strategy deconstructs complex visuals into clear insights, allowing users to explore objects, layouts, and environmental context through auditory feedback. However, there is a significant latency due to separate client server deployment with an average response time of 36 seconds per command. This latency could be reduced by deploying the complete system in a single workstation. Future works includes developing an Unreal Engine plugin, BLV user testing, user feedback system and expanding functionality to support tasks like multimodal integration and real time navigation system.

Acknowledgments. This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland [12/RC/2289_P2] at Insight Research Ireland Centre for Data Analytics, Dublin City University; The authors also thank Bentley Systems for their support and funding.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Lim, C. P., Nonis, D., Hedberg, J.: Gaming in a 3D multiuser virtual environment: Engaging students in science lessons. British Journal of Educational Technology, vol. 37, no. 2, pp. 211–231 (2006)
2. Bowman, D. A., North, C., Chen, J., Polys, N. F., Pyla, P. S., Yilmaz, U.: Information-rich virtual environments: theory, tools, and research agenda. In: Proceedings of the ACM symposium on Virtual reality software and technology, pp. 81–90 (2003)
3. Jones, P. E.: Three-dimensional input device with six degrees of freedom. Mechatronics, vol. 9, no. 7, pp. 717–729 (1999)
4. Argelaguet, F., Andujar, C.: A survey of 3D object selection techniques for virtual environments. Computers and Graphics, vol. 37, no. 3, pp. 121–136 (2013)
5. Cao, X., Ju, K. P., Li, C., Jain, D.: SceneGenA11y: How can Runtime Generative tools improve the Accessibility of a Virtual 3D Scene?. In: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, pp. 1–10 (2025)
6. Wong, A., Gillis, H., Peck, B.: VR Accessibility Survey: Survey for People with Disabilities. https://drive.google.com/file/d/0B0VwTVwReMqLMFIzdzVVaVdaTFk/view (2018)
7. Teófilo, M., Lucena, V. F., Nascimento, J., Miyagawa, T., Maciel, F.: Evaluating accessibility features designed for virtual reality context. In: 2018 IEEE international conference on consumer electronics (ICCE), pp. 1–6. IEEE (2018)
8. Trewin, S., Laff, M., Hanson, V., Cavender, A.: Exploring visual and motor accessibility in navigating a virtual world. ACM Transactions on Accessible Computing (TACCESS), vol. 2, no. 2, pp. 1–35 (2009)
9. Hollister, J. M.: Virtual libraries in the new metaverse: An exploratory study on community libraries in Meta Horizon Worlds. International Journal of Knowledge Content Development and Technology, vol. 14, no. 4, pp. 113–133 (2024)
10. Naikar, V. H., Subramanian, S., Tigwell, G. W.: Accessibility feature implementation within free vr experiences. In: Extended abstracts of the CHI conference on human factors in computing systems, pp. 1–9 (2024)
11. Killough, D. et al.: Demonstration of VRSight: AI-Driven Real-Time Descriptions to Enhance VR Accessibility for Blind People. In: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, pp. 1–5 (2025)
12. Killough, D. et al.: XR for All: Understanding Developer Perspectives on Accessibility Integration in Extended Reality. arXiv preprint arXiv:2412.16321 (2024)
13. Desai, K., Johnson, J.: Virtex: Learning visual representations from textual annotations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11162–11173 (2021)
14. Zhang, Y. et al.: Recognize anything: A strong image tagging model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1724–1732 (2024)
15. Yoon, H., Kim, B.-H., Mukhriddin, M., Cho, J.: Salient region extraction based on global contrast enhancement and saliency cut for image information recognition of the visually impaired. KSII Transactions on Internet and Information Systems (TIIS), vol. 12, no. 5, pp. 2287–2312 (2018)

16. Stearns, L., Findlater, L., Froehlich, J. E.: Design of an augmented reality magnification aid for low vision users. In: Proceedings of the 20th international ACM SIGACCESS conference on computers and accessibility, pp. 28–39 (2018)
17. Lang, F., Machulla, T.: Pressing a button you cannot see: evaluating visual designs to assist persons with low vision through augmented reality. In: Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology, pp. 1–10 (2021)
18. Deng, W., Qi, M., Ma, H.: Global-local tree search in vlms for 3d indoor scene generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 8975–8984 (2025)
19. Doveh, S. et al.: Teaching VLMs to Localize Specific Objects from In-context Examples. arXiv preprint arXiv:2411.13317 (2024)
20. Zha, J., Fan, Y., Yang, X., Gao, C., Chen, X.: How to enable llm with 3d capacity? a survey of spatial reasoning in llm. arXiv preprint arXiv:2504.05786 (2025)
21. Wang, X. et al.: Spatial 3D-LLM: Exploring Spatial Awareness in 3D Vision-Language Models. arXiv preprint arXiv:2507.16524 (2025)
22. Cocchi, F. et al.: LLaVA-MORE: A Comparative Study of LLMs and Visual Backbones for Enhanced Visual Instruction Tuning. arXiv preprint arXiv:2503.15621 (2025)
23. Hoque, M., Hasan, M. R., Emon, M. I. S., Khalifa, F., Rahman, M. M.: Medical image interpretation with large multimodal models. In: CEUR workshop proceedings, CEUR Workshop Proceedings 3740, CEUR-WS. org (2024)
24. Zhang, S., Fang, Q., Yang, Z., Feng, Y.: Llava-mini: Efficient image and video large multimodal models with one vision token. arXiv preprint arXiv:2501.03895 (2025)
25. Cai, M. et al.: Vip-llava: Making large multimodal models understand arbitrary visual prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12914–12923 (2024)
26. Howard, P., Bhiwandiwalla, A., Fraser, K. C., Kiritchenko, S.: Uncovering bias in large vision-language models with counterfactuals. arXiv preprint arXiv:2404.00166 (2024)
27. Selvam, S., Rajendran, R. K., Sankaradas, M., Raghunathan, A., Chakradhar, S. T.: SimCache: Similarity Caching for Efficient VLM-based Scene Understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 3327–3336 (2025)
28. TIGER AI Lab: VIEScore: Towards Explainable Metrics for Conditional Image Synthesis Evaluation. https://tiger-ai-lab.github.io/VIEScore/ (2024)
29. Fernandez, Jaime B.; ALI, Muhammad Intizar. System Demo of Modeling Smart University Campus Virtual Environments. En International Conference on Multimedia Modeling. Singapore: Springer Nature Singapore, p. 218-224 (2025)