VQA-driven Event Maps for Assistive Navigation for People with Low Vision in Urban Environments

Joseph Morales[1,2] Bruk Gebregziabher[3] Alex Cabañeros[3] Jordi Sanchez-Riera[1]
[1]Institut de Robotica i Informatica Industrial, CSIC-UPC, Barcelona, Spain
[2]MIT [3]Biel Glasses

***Abstract*— We introduce a novel framework for assistive urban navigation for individuals with low vision. Utilizing a smart glasses platform developed by Biel Glasses, which provide a continuous stream of stereo images and GPS fixes, we generate an *Event Map* based on key semantic elements extracted by carefully prompted visual question-answering (VQA) models. For individuals with blurry or reduced fields of vision (low vision), traversing city streets poses a variety of challenges; they may struggle to perceive construction work, potholes, crowded sidewalks, and other ambiguous obstacles obstructing their paths. Some tasks, such as distinguishing traffic light signals, are nigh impossible without assistance from a companion or city infrastructure aimed towards accessibility. Although the majority of these problems may be solved with individually tailored traditional computer vision algorithms, developing and running a suite of these algorithms is challenging and resource demanding. Therefore, our proposed solution capitalizes on a single underlying implementation that need only be extended by adding queries. We validate our approach using a custom dataset of over 1,300 annotated images from various locations around Barcelona, reporting performance across different urban navigation tasks. We demonstrate the performance of the end to end system on a run of data collected by the Biel Glasses platform.**

I. INTRODUCTION

Wearable devices like phones and glasses are increasingly playing a key role in helping people navigate urban environments. In the case of individuals with low vision, these devices are crucial for detecting both static and dynamic elements in their surroundings and help them to move safely through city streets. Detecting stairs, crossings, other pedestrians, obstacles or any other unexpected hazard is critical for guiding them with clear directions. Moreover, these detections must be performed efficiently, using limited resources, to keep these devices lightweight and to maximize their battery life.

To date, various solutions have been attempted to address these challenges, including the integration of separate computer vision algorithms in path planning tasks [1], where these algorithms are activated or deactivated as needed to optimize computational power and battery consumption. However, this approach is still sub-optimal due to several drawbacks: 1) the need for separate training for each type of detection, 2) the complexity of tailoring algorithms to specific detections, such as identifying steps using vanishing lines [2], and 3) the inability to anticipate which objects or obstacles will need to be detected.

Fig. 1: This figure illustrates that, given images captured by the Biel Glasses platform, the proposed system answers questions about the environment and generates event labels which are then used to provide meaningful feedback to the user.

Recently introduced visual question answering (VQA) models [3], [4], [5], which take an image and a related question as inputs and output an answer with a confidence score, offer a promising alternative for object detection without the need to train a model for specific objects. These models leverage Natural Language Models (LLMs) to generalize across a wide range of questions and texts, making them well-suited for identifying objects or events on the fly. In this manner, by formulating the right set of questions about an image, these models can uncover high-level concepts [6], [7] and events, which can then be utilized for navigation purposes.

We propose to use a VQA model to create a high-level event map, a scene graph, with critical events needed to improve the navigation system of a pair of glasses developed by Biel Glasses [1], see Figure 1. Using these glasses that are equipped with a stereo-camera and a GPS unit, we are able to construct a scene graph that contains detected events needed by the path planning or safety avoidance algorithms to give directions and warnings to low vision people. These event detections are made using a VQA model [3] with pre-trained questions chosen to discover critic events relevant for the navigation. Some of these events include: identifying crosswalk lights, street crossings, construction, obstacles on the road or sidewalk, crowdedness of a sidewalk, and types of shops within view. Each event is stored with its GPS location and VQA model embedding in a frontend-backend structure to optimize storage and computational resources.

[1]https://bielglasses.com

To validate the proposed approach we collect more than 3,000 images of the city of Barcelona from distinct neighborhoods to ensure a diversity of event types and how they manifest in the city. For each detection, we perform a prompt optimization from several questions that later are tested to evaluate the best performing one and ensure a high reliable score to the event detection. The VQA model is compared with other state-of-the-art models, and a through out study of the detections are made for all dataset.

The main contributions of our proposed method are: 1) use a VQA model to discover relevant events for safety navigation through the city, 2) a frontend-backend framework to create a scene graph that can store and remove these events and, 3) a dataset with images and their locations within the city of Barcelona.

II. RELATED WORK

A. Pedestrian Navigation

Pedestrian navigation is a dynamic area of research, with a primary focus on path planning and safety. Whether it's a robot or a person moving through urban environments, key challenges include detecting dynamic objects and identifying correctly the sidewalks to navigate various locations effectively. Consequently, detecting other pedestrians [8], [9], recognizing sidewalk boundaries [10], and enhancing self-localization [11] are essential components for ensuring safe and efficient navigation. Additionally, navigation can be further enhanced through the use of augmented reality [12], landmarks [13], and route optimization based on user preferences [14]. In the particular case of of individuals who are blind or have low vision, navigation is even more challenging as even detecting a door shops [15] can be a problematic, and guidance during navigation must be preferably expressed in a natural language manner [16]. While developing all these solutions contribute significantly to improving navigation, the aforementioned methods still fall short in addressing the uncertainties inherent in urban environments. Streets can be obstructed by construction barriers, potholes, or other unpredictable obstacles, making it difficult to develop a system that accounts for every possible element. This complexity highlights the need for an algorithm capable of generalizing and extracting relevant information from an image of the environment. To address these challenges, we propose leveraging a Visual Question Answering (VQA) model, which can provide answers to any questions about the environment based on a visual input.

B. Visual Question Answering (VQA)

VQA models are a specialized subset of Vision-Language Models (VLMs), which aim to integrate flexible language understanding with image analysis. For example, CLIP [17] achieves this by projecting the embeddings from language and visual encoders into a shared embedding space using contrastive learning. This approach enables users to interact with both language prompts and images by comparing the similarity of the generated embeddings. VQA models, such as Vision-Language Transformers (ViLT) [3] and similar models [18], [19], [4], [5], build on this concept by generating textual responses based on an input image and a given textual prompt, offering a more context-aware analysis of visual data. This kind of models can generalize to any kind of question, eliminating the need to train individual image detectors, and are used in many kind of applications such as unknown object detection [20], [21], [22], visual grounding [23], [24] and interactive environments [25]. These capabilities have also been exploited by numerous robot applications in navigation tasks. Given natural language instructions, these models can detect relevant objects on a scene [26], extract landmarks to facilitate navigation tasks [27] and understand high level concepts to construct navigation maps [16]. While navigation maps that identify obstacles, crosswalks, traffic lights, and other critical elements are essential for safe route planning and guidance, current VQA models have yet to be applied in constructing such maps. So far, no existing work has leveraged VQA models to create these detailed, real-time navigation aids detecting these mentioned elements.

C. Scene Representation

The most common approach to representing a scene is through the use of scene graphs. Scene graphs are structured representations of visual environments, capturing the relationships between objects in a scene. In the context of navigation, scene graphs can provide a comprehensive map of an environment by identifying and categorizing objects like obstacles, pathways, and landmarks, along with their spatial relationships. This allows for more effective navigation, as systems can better understand the layout and dynamics of an environment, enabling safer and more efficient path planning. The scene graphs ability to represent semantic concepts and compress environmental information makes them ideal for use in navigation algorithms. The majority of works use scene graphs for indoor environments, providing semantic guidance [28], planning and finding objects in rooms [7] or constructing such graphs from other points of view [29] rather than the common frontal view. Scene graphs also have the potential to encode dynamic objects like elements that can be in different places (i.e. a cup of coffee) and store the information on the graph for later localization of the object [30]. These graphs can also be combined with language models like GPT-3.5 to perform queries about the elements of the graph. This allows the system to access data encoded in specific nodes that may not have been relevant initially but become important in the current context [6]. Unlike previous works that use scene graphs for querying, detecting dynamic objects, or navigating primarily in indoor environments, our approach extends these concepts to outdoor settings. Additionally, since our representations do not require the full complexity of traditional scene graphs, we employ a simplified version, namely event maps. These event maps store all detected events during navigation, and we enable querying and detection of dynamic objects by adding or removing elements from these event maps, while encoding image data to optimize spatial representation. This approach

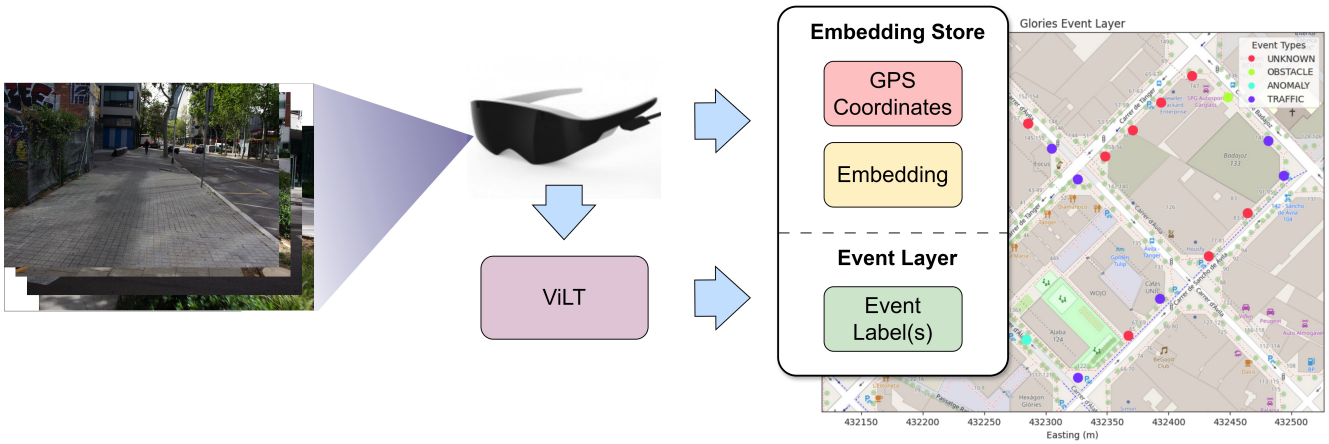

Fig. 2: This diagram illustrates the proposed end to end system that is used to produce Event Maps. The Biel Glasses platform collects GPS data and produces an image stream that is processed by ViLT (or some other VQA model) to produce embeddings and event labels. These are used to incrementally add nodes to the embedding store and the event layer, resulting in a full Event Map.

is particularly useful for navigation, as it allows us to store large-scale maps and relevant elements, which are crucial for guiding people with low vision through complex urban environments.

III. METHOD

A. Research Questions

This work aims to investigate the practical application of VQA models for assistive navigation tasks in urban environments for individuals with low vision, with a specific focus on evaluation in the city of Barcelona. The study compares the accuracy, recall, specificity, and F1 scores of VQA models on targeted binary perception tasks using a dataset of images collected and manually labeled to perform the evaluation.

The urban pedestrian navigation tasks investigated in this paper are specifically designed to assist in orienting and guiding individuals with low vision. These tasks include:

1) Distinguishing Crosswalk Signal
2) Stair Detection
3) Generic Obstacle Detection
4) Building front Identification

For each task, VQA models are queried with a set of semantically similar question prompts that would allow a user to accomplish the desired task. For example, to accomplish stair detection, the models may be queried with questions like “Are there stairs in this image?” and ”Is there a step in front of the camera?” For each of these tasks there is a ”common sense” prompt that is used to determine a baseline to compare across models. The purpose of this prompt is to emulate a person proposing a question that should allow another person or the model to accomplish the task, without attempting to prompt engineer for performance. The set of semantically similar questions are evaluated collectively to determine how sensitive each model is to prompt engineering on each task, and determine in which ways model performances can be improved.

B. Models

To compare VQA models of varying complexities and sizes, ViLT [3], BLIP [4], and BLIP-2 [5] were evaluated. Models were evaluated by providing the previous research questions as a set of semantically similar prompts and then comparing the model output against a desired answer. For each question, there is a “positive” set of images where the model is expected to provide the desired answer and a “negative” set where the model is expected to provide anything except for the desired answer. For example, in the case of “Is this crosswalk light red or green?,” the desired answer may be set as “red,” with the positive set being images of red crosswalk lights and the negative set being images of green crosswalk lights.

ViLT accepts an image and a textual question as inputs, and it returns k candidate responses ranked in order of likelihood, along with their logits. In this way, scores can be assigned to each response. ViLT is evaluated both by taking the top answer regardless of score, as few differences were noticed with strict confidence cutoffs. BLIP accepts an image and a textual question as inputs, and returns a generated response. There is only one candidate and no logits, so this response is directly used as the answer. BLIP-2 takes an image and a prompt in the form of “Question: Is the crosswalk light red or green? Answer:” and returns a generated sentence or phrase in response. Rather than directly comparing the response to the desired answer, the response is checked to see if it contains the desired response.

C. Event Maps

In addition to using VQA models to query images, VQA models outputs can be incorporated into persistent scene representations for the purposes of semantic mapping. In the process of generating a response to an image and a textual prompt, these VQA models first encode images into visual embeddings, which are smaller than the input images. These embeddings can be re-queried against new

textual prompts without reprocessing the original image. The embedding for a 720p image is 1/4th of the size of the image itself, motivating the storage of embeddings rather than the images themselves. To ground these embeddings and any observations derived from them in the real world, and to connect this work to previous works on scene graphs, we introduce the concept of an *Embedding Store*, an *Event Layer*, and an *Event Map*.

An embedding store is a map representation that takes the visual embedding from processing an image with a VQA model, and stores it with the associated GPS coordinates of where that image was taken. This structure allows for querying multiple embeddings at the same time, and reconstructing a map of a scene with answers evaluated across different prompts. Although images are collected and processed in series on the platform, these models allow for batch processing that makes offline inference across the whole scene much faster. Similarly, using the saved visual embeddings, each image can be queried with multiple questions concurrently offline, allowing for more semantic information to be extracted without collecting more data.

An event layer is built on top of an embedding store, and allows users to provide specific semantic labels and descriptions to embedding store nodes. Each node in the embedding store is assigned one or more "event labels," explicitly indicating useful information in the scene that has been extracted from a VQA model. For instance, if the image a node is generated from shows a car in a crosswalk, two events might be generated and associated with that node: one event indicating a crosswalk, and another event indicating an obstacle blocking the user's path. The implementer defines the priority of event types and what information is shared with the user first. An event is defined by a label or class (traffic, obstacle, etc) and an optional textual description (generated or assigned). Whereas these VQA models can provide open-set responses to input questions, events are discrete and defined by the implementer. For evaluation in this paper, we define 4 event types:

1) *Obstacle* - Something is blocking the presumed path of the user
2) *Traffic* - There is a crosswalk or road directly in front of the user
3) *Anomaly* - There is a long term change in the environment, such as construction
4) *Unknown / Generic* - None of the above, indicating a "clear" path

As will be shown in the experiments, Event Map data was collected on the Biel Glasses platform and processed offline. ViLT was chosen as the VQA model for generating these maps because it is the most lightweight model and the most feasible to implement on the platform online in the future.

IV. Experimental Results

A. Data Collection

All data for these experiments were collected manually. The images for the VQA task analysis were collected with a Canon EOS 6D at 720p, and the image stream and GPS data for the event layer analysis was collected on the Biel Glasses platform. Data collections involved continuously taking photos while on pre-planned walks of various neighborhoods in Barcelona, capturing natural urban navigation scenes.

Fig. 3: Example images from the positive and negative splits for each task. The left column represents all of the positive class images, and the right column represents all the negative. The labels on the left hand side of each row indicate which task those images were evaluated for.

Data was hand labeled by splitting images that were contextually relevant to each task into "positive" and "negative" folders. For instance, for crosswalk traversal, only images collected at crosswalks were filtered into a "red crosswalk light" folder or a "green crosswalk light" folder. Along a similar vein, for building front identification, only images of the fronts of buildings were included in either the positive or negative sets. 923 pictures were collected of sidewalk scenes, encompassing obstacle and stair related tasks, 239 photos were collected at crosswalks, and 220 photos were collected of building fronts.

B. VQA Model Evaluation on Common Sense Prompts

In Table I, VQA model accuracy, recall, precision, and F1 score metrics are presented on various assistance tasks using naive, common sense prompts. For the task of distinguishing crosswalk signals, images of either red or green crosswalk lights were queried for the color of the light, with pictures with red lights in the positive set and pictures with green

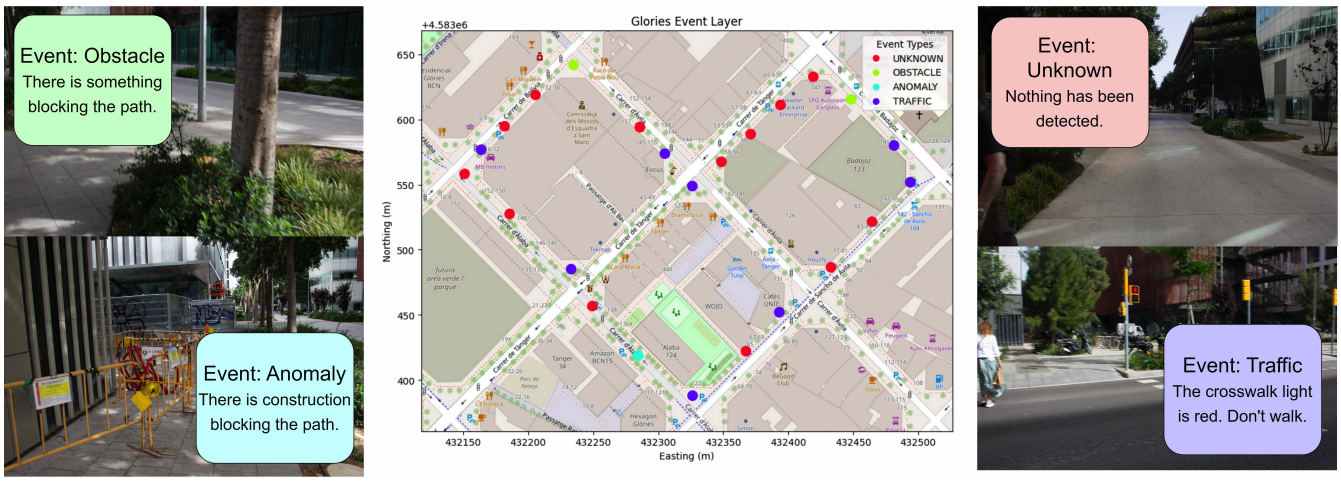

Fig. 4: Example Event Map generated with ground truth GPS fix points and labels. An example of each event label is provided to the left and right of the Event Map, demonstrating the types of images and generated descriptions that are associated with each event label type.

Common Sense Prompts					
Model	Acc	Recall	Precision	Specificity	F1
Is the crosswalk signal red or green? (red)					
ViLT	**0.80**	**0.95**	**0.78**	0.57	**0.85**
BLIP	0.77	0.94	0.75	0.51	0.83
BLIP-2	0.40	0.01	0.67	**0.99**	0.03
Are there stairs in this image?					
ViLT	0.82	0.60	0.84	**0.94**	0.70
BLIP	**0.92**	0.88	**0.88**	**0.94**	**0.88**
BLIP-2	0.83	**0.97**	0.69	0.76	0.80
Is there something blocking the sidewalk?					
ViLT	0.49	0.75	0.45	0.29	0.56
BLIP	**0.73**	**0.79**	0.65	0.68	**0.71**
BLIP-2	0.59	0.11	**0.67**	**0.96**	0.19
Is this an image of a restaurant?					
ViLT	0.64	0.94	0.45	0.51	0.60
BLIP	0.81	0.94	0.62	0.76	0.75
BLIP-2	**0.85**	**0.95**	**0.68**	**0.81**	**0.79**

TABLE I: Analysis of model performances across all tasks for each "common sense" prompt. The tasks consist of distinguishing crosswalk signal, detecting stairs, detecting generic obstacles on the sidewalk, and identifying restaurants. The accuracy, recall, precision, specificity, and F1 score are reported for each model on each prompt.

lights in the negative set. For the stair detection task, images with steps, stairs, or staircases were queried for the presence of stairs, with images of stairs in the positive set and images of sidewalks without stairs or obstacles as the negative class. For the general object detection task. the positive set consisted of images of poles, benches, and other objects that could block a person's path, and the negative set was the same as the negative set for the stairs task. For the building front identification task, images of building fronts were queried to determine if they were establishments where one could buy food, with restaurants and cafes in the positive set and apartment building fronts, garages, and non-food related shops were included in the negative set.

We observe that each model has varying performance across tasks, resulting in different optimal models for each task. At the same time, it is clear that the average BLIP F1 score was significantly higher than that of the other models, outperforming both ViLT and BLIP-2 (0.79 vs 0.68 and 0.45, respectively).

C. Model Prompt Sensitivity Analysis

In Table II, VQA model F1 scores are presented on the assistive tasks across a broader set of semantically similar prompts, with the goal of determining the importance of prompt engineering for each task. Although some of the new prompts may semantically indicate something subtly different than the intended task, the same labels and splits are used for evaluation to determine which prompt inputs generate the best results for the intended task.

We observe a general tendency that, other than the restaurant identification task, BLIP-2 experienced the least consistent results. At the same time, we can observe that BLIP, ignoring the one outlier in the generic obstacle detection task, performed the most consistently across prompts.

These findings in combination with the findings from Section IV-B seem to show, on a surface level, that BLIP provides the best assistive navigation performance out of the evaluated models. This suggests that for the Biel Glasses platform, it would perform well as a VQA model for general pedestrian navigation tasks, and that it would additionally perform well at live question-answering, wherein the users asks questions that are not already being processed by the system (for example: "Is there a bus in front of me at this bus stop?"). Unlike BLIP which is much more consistent, BLIP-2 performance is very dependent on prompt engineering and benefits from grounding questions in the scene, particularly for tasks that are more visual in nature than semantic (such as distinguishing the crosswalk light color).

D. Real World Mapping with Event Layer

Evaluation of real world performance can be seen in Figure 5. Data was collected using the Biel Glasses platform of a short walk around each of the sidewalks around an intersection, resulting in 15293 images, which were then down-sampled to 25 images by taking each 600th image

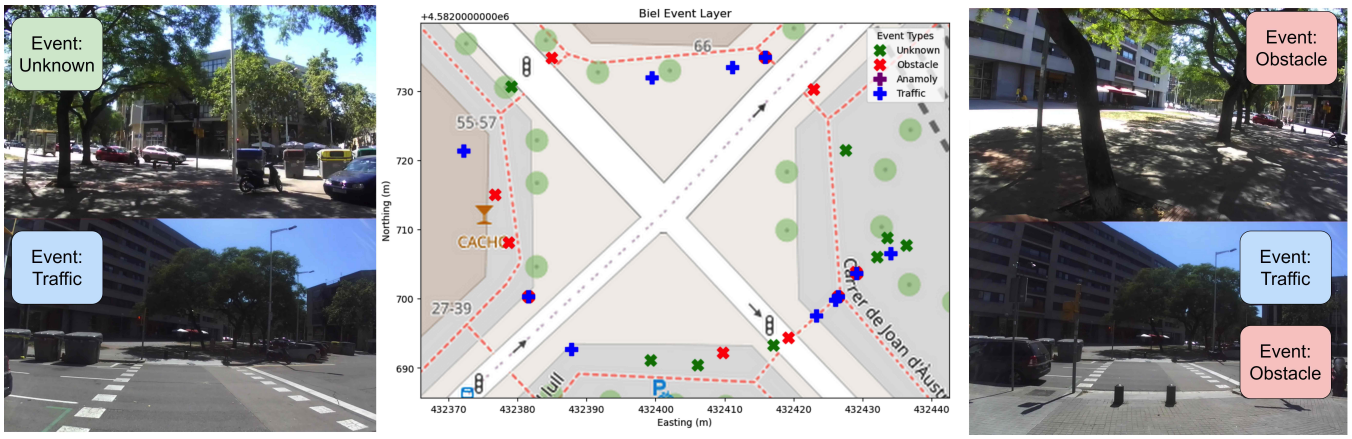

Fig. 5: Event Map generated offline using images and GPS data from the Biel Glasses platform. Examples of each event label from the captured data are shown on the sides of the map, with an example of an image with multiple event labels in the bottom right.

Prompt Sets per Task			
Question	Model F1 Scores		
	ViLT	BLIP	BLIP-2
Distinguishing Crosswalk Signal			
Is the crosswalk light red or green?	0.85	0.83	0.03
Is the walk light red or green?	0.84	0.85	0.22
Is the pedestrian light red or green?	**0.86**	0.83	0.11
Is the crosswalk light on the pole red or green?	0.83	0.85	0.55
Is the walk light on the pole red or green?	0.82	**0.86**	0.66
Detecting Stairs			
Are these stairs?	0.62	**0.89**	0.58
Are there stairs in this image?	0.70	0.88	0.80
Is this an image of stairs?	0.64	**0.89**	**0.89**
Are there steps in this image?	0.77	0.86	0.87
Are there stair steps in this image?	0.76	0.77	0.83
Detecting Generic Obstacles			
Is this an obstacle?	0.22	0.07	0.19
Is there something on the sidewalk?	0.60	0.63	0.52
Is there something blocking the sidewalk?	0.56	**0.71**	0.19
Is there something on the sidewalk blocking people from walking?	0.49	0.66	0.21
Is there something really close to the camera on the ground?	0.69	0.61	0.53
Identifying Restaurants			
Is this a restaurant?	0.60	0.75	**0.79**
Is this an image of a restaurant?	0.54	0.74	0.70
Is this a picture of a restaurant?	0.59	0.73	0.69
Is this the outside of a restaurant?	0.49	0.62	0.69
Is the outside of a restaurant in the middle of this image?	0.57	0.62	0.60

TABLE II: Comparison of F1 Scores of each model on a set of 5 prompts per task. The best model score for each task is bolded.

at regular intervals. These images were then hand labeled to each include one or more of the various event labels. The system to generate these event labels was then run on the 25 sampled and labeled images, and accuracy was calculated by comparing the number of correctly matched event labels per node to total event labels. The GPS data to ground these embeddings was directly taken without post-processing, resulting in trajectories that are a bit misaligned with the underlying OSM map.

We observe that the system achieves 71.4% accuracy (20 labels out of 28) over the Event Map. We see an accuracy of 56.3% on unknown (generic) events, 100% for traffic events, and 80% for obstacle events (no anomalies were present nor detected in this data collection). As can be gleaned from the map, the system tends to detect crosswalks much earlier than the user approaches them, and occasionally ”hallucinates” obstacles. Although early detection of crosswalks is less problematic because it can be rectified with the underlying street map, precise obstacle detection is an important target.

V. Conclusions

In the presented work, we explored how pretrained VQA models can be leveraged in assistive technology to aid urban navigation for pedestrians with low vision. We evaluated three pretrained VQA models, ViLT, BLIT and BLIP-2, on four key urban navigation tasks: recognizing crosswalk signals, detecting stairs, identifying generic obstacles, and recognizing building fronts. Our results showed that while VQA models can effectively handle these tasks, they struggle with the open-ended nature of detecting generic obstacles, and that in the majority of cases prompt engineering is beneficial to improve performance. Additionally, we introduced a Event Map framework for representing environments that users have traveled through. This framework allows users to manually define events detectable by VQA models, which can then process the resulting image embeddings either online or offline to identify and explicitly represent those events within a scene. We tested our end to end system on data collected with the Biel Glasses platform and were able to demonstrate a proof of concept. Although they require further tuning, modern VQA models provide an exciting avenue for further assistive robotics applications.

VI. Acknowledgements

This work has been partially supported by: SMARTGAZEII CPP2021-008760 funded by MCIN/ AEI /10.13039/501100011033 and by the ”European Union NextGenerationEU/PRTR”; and the industrial doctorate 2021 DI 00102 funded by the Government of Catalonia.

REFERENCES

[1] S. Teng, X. Hu, P. Deng, B. Li, Y. Li, Y. Ai, D. Yang, L. Li, Z. Xuanyuan, F. Zhu, and L. Chen, "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 6, pp. 3692–3711, 2023.

[2] C. Wang, Z. Pei, S. Qiu, and Z. Tang, "Deep leaning-based ultra-fast stair detection," *Scientific Reports*, vol. 12, p. 16124, 09 2022.

[3] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139, 18–24 Jul 2021, pp. 5583–5594.

[4] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[5] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, 23–29 Jul 2023, pp. 19 730–19 742.

[6] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard, "Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation," *Robotics: Science and Systems*, 2024.

[7] C. Agia, K. Jatavallabhula, M. Khodeir, O. Miksik, V. Vineet, M. Mukadam, L. Paull, and F. Shkurti, "Taskography: Evaluating robot task planning over large 3d scene graphs," in *Conference on Robot Learning*. PMLR, 2022, pp. 46–58.

[8] C. Cao, P. Trautman, and S. Iba, "Dynamic channel: A planning framework for crowd navigation," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019.

[9] S. Buckeridge, P. Carreno-Medrano, A. Cosgun, E. Croft, and W. Chan, "Mapless urban robot navigation by following pedestrians," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, 2023.

[10] Y. Du, N. Hetherington, C. Oon, W. Chan, C. Quintero, E. Croft, and H. MacHiel Van Der Loos, "Group surfing: a pedestrian-based approach to sidewalk robot navigation," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019.

[11] H. Fu, V. Renaudin, Y. Kone, and N. Zhu, "Analysis of the recent ai for pedestrian navigation with wearable inertial sensors," *IEEE Journal of Indoor and Seamless Positioning and Navigation*, vol. PP, pp. 1–13, 01 2023.

[12] D. Kumar, S. Iyer, E. Raja, R. Kumar, and V. P. Kafle, "Improving pedestrian navigation in urban environment using augmented reality and landmark recognition," *IEEE Communications Standards Magazine*, vol. 8, no. 1, pp. 20–26, 2024.

[13] L. Zhu, J. Shen, J. Zhou, Z. Stachoň, S. Hong, and X. Wang, "Personalized landmark adaptive visualization method for pedestrian navigation maps: Considering user familiarity," *Transactions in GIS*, vol. 26, no. 2, pp. 669–690, 2022.

[14] Y. Akasaka and T. Onisawa, "Personalized pedestrian navigation system with subjective preference based route selection," *Intelligent Decision and Policy Making Support Systems*, pp. 73–91, 2008.

[15] M. Weiss, S. Chamorro, R. Girgis, M. Luck, S. E. Kahou, J. P. Cohen, D. Nowrouzezahrai, D. Precup, F. Golemo, and C. Pal, "Navigation agents for the visually impaired: A sidewalk simulator and experiments," in *3rd Annual Conference on Robot Learning*, 2019.

[16] Z. Huang, Z. Shangguan, J. Zhang, G. Bar, M. Boyd, and E. Ohn-Bar, "Assister: Assistive navigation via conditional instruction generation," in *17th European Conference in Computer Vision (ECCV)*, 2022.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763.

[18] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[19] H. Ben-Younes, R. Cadène, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering." in *ICCV*, 2017, pp. 2631–2639.

[20] W. Kuo, Y. Cui, X. Gu, A. Piergiovanni, and A. Angelova, "Open-vocabulary object detection upon frozen vision and language models," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=MIMwy4kh9lf

[21] P. Du, Y. Wang, Y. Sun, L. Wang, Y. Liao, G. Zhang, E. Ding, Y. Wang, J. Wang, and S. Liu, "Lami-detr: Open-vocabulary detection with language model instruction," in *Proceedings of the European conference on computer vision (ECCV)*, 2024.

[22] C. Zhu and L. Chen, "A survey on open-vocabulary detection and segmentation: Past, present, and future," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[23] L. Xiao, X. Yang, F. Peng, M. Yan, Y. Wang, and C. Xu, "Clip-vg: Self-paced curriculum adapting of clip for visual grounding," *Trans. Multi.*, vol. 26, p. 4334–4347, oct 2023.

[24] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[25] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "Iqa: Visual question answering in interactive environments," in *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018.

[26] F. Kenghagho Kenfack, F. Ahmed Siddiky, F. Balint-Benczedi, and M. Beetz, "Robotvqa — a scene-graph- and deep-learning-based visual question answering system for robot manipulation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[27] D. Shah, B. Osiński, b. ichter, and S. Levine, "Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action," in *Proceedings of The 6th Conference on Robot Learning*, 2023.

[28] Z. Dai, A. Asgharivaskasi, T. Duong, S. Lin, M.-E. Tzes, G. J. Pappas, and N. Atanasov, "Optimal scene graph planning with large language model guidance," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[29] R. Liu, X. Wang, W. Wang, and Y. Yang, "Bird's-eye-view scene graph for vision-language navigation," in *ICCV*, 2023.

[30] F. Zhou, H. Liu, H. Zhao, and L. Liang, "Long-term object search using incremental scene graph updating," *Robotica*, vol. 41, no. 3, p. 962–975, 2023.