Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/57/20241309

Action-Aware Vision Language Navigation (AAVLN): AI vision system based on cross-modal transformer for understanding and navigating dynamic environments

Jasmine Liu

Shanghai American School, Shanghai, China

jasmineliu.sh@gmail.com

Abstract. Visually impaired individuals face great challenges with independently navigating dynamic environments because of their inability to fully comprehend the environment and actions of surrounding people. Conventional navigation approaches like Simultaneous Localization And Mapping (SLAM) rely on complete scanned maps to navigate static, fixed environments. With Vision Language Navigation (VLN), agents can understand semantic information to expand navigation to similar environments. However, both cannot accurately navigate dynamic environments containing human actions. To address this challenge, we propose a novel cross-modal transformer-based Action-Aware VLN system (AAVLN). Our AAVLN Agent Algorithm is trained using Reinforcement Learning in our Environment Simulator. AAVLN's novel cross-modal transformer structure allows the Agent Algorithm to understand natural language instructions and semantic information for navigating dynamic environments and recognizing human actions. For training, we use Reinforcement Learning in our action-based environment simulator. We created it by combining an existing simulator with our novel 3D human action generator. Our experimental results demonstrate the effectiveness of our approach, outperforming current methods on various metrics across challenging benchmarks. Our ablation studies also highlight that we increase dynamic navigation accuracy with our Vision Transformer based human action recognition module and cross-modal encoding. We are currently constructing 3D models of real-world environments, including hospitals and schools, for further training AAVLN. Our project will be combined with Chat-GPT to improve natural language interactions. AAVLN will have numerous applications in robotics, AR, and other computer vision fields.

Keywords: Scene Understanding, Cross-modal Transformer, Computer Vision, Vision Language Navigation

1. Introduction

1.1. Background and Motivation

Visual impairment is a growing global concern, affecting more than half a billion individuals with varying degrees of visual acuity, from mild to severe. The World Health Organization identifies macular degeneration, uncorrected refractive error, cataract, and age-related visual impairments as some of the

DOI: 10.54254/2755-2721/57/20241309

leading causes of vision loss [1]. The challenges that visually impaired people face have led to their heavy reliance on guardians for navigating and understanding their surroundings.

Since 2020, we have been teaching visually impaired children English. Through our communication, we learned about their difficulties with navigating and understanding their environments. To further understand the difficulties and needs of visually impaired people, we designed and conducted a joint survey with the visually impaired students that we have taught and with low vision doctors from Fudan University's Affiliated Hospital. They expressed great interest in this project as they found it meaningful, especially for the visually impaired population. Through our survey, we found that over 75% of visually impaired people do not go out alone daily due to their top 3 difficulties of navigation, human action recognition, and road sign recognition. Currently over 65% of people rely on the company of parents or family members to navigate and understand their environment when navigating outdoors. This illustrates the critical challenges visually impaired people face with independent navigation. For action recognition, over 65% of visually impaired people want to recognize hand gestures, over 55% want to recognize head movements, and over 45% want to recognize upper body movements. When navigating in general, 60% are concerned about the accuracy and efficiency of their navigation route and over 70% are concerned about running into others. While we often believe visually impaired people face more difficulties with outdoor navigation, 50% of the survey respondents expressed they found indoor and outdoor navigation to be equally difficult, if not indoor navigation being even more difficult. Finally, over 50% of people expressed they often find it difficult to navigate a variety of environments, mainly hospitals, shopping malls, restaurants, and metro stations.

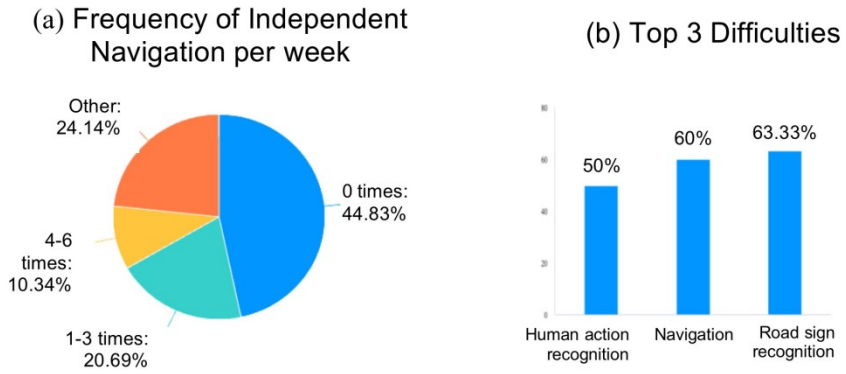


Figure 1. Survey results from our joint survey with visually impaired people and low vision doctors

Last year, we developed a scene text recognition system to assist the visually impaired in recognizing curved and distorted scene text, including road signs, restaurant signs, package labels, etc. However, accurate navigation in dynamic environments and human action recognition remain a critical issue. Aside from heavily relying on help from family members or guardians to navigate outdoors, visually impaired people only have a minimal number of potential AI systems to assist them. Unfortunately, these systems can only help with accurately navigating static environments or simply recognize text and objects. They cannot accurately navigate dynamic environments containing human actions by planning a route based on the user's destination while recognizing the actions of people along the path. This presents a crucial need to develop an AI system that can effectively navigate environments while taking into account the actions of surrounding people.

1.2. Contributions

In this paper, we aim to address these difficulties by introducing a novel navigation system for embodied AI. Our Reinforcement Learning framework for AAVLN comprises two main components: the Agent Algorithm and the Environment Simulator. The Agent Algorithm is trained in our virtual action-based simulator that contains generated 3D human actions.Our proposed system leverages a cross-modal

DOI: 10.54254/2755-2721/57/20241309

transformer structure to comprehend natural language instructions, current environmental views, historical observations, and human action sequences to understand and navigate dynamic environments.

The main contributions of our paper are threefold:

1) we develop a novel Agent Algorithm through innovatively adding an action recognition module into the Agent Algorithm's Vision Language Navigation architecture for cross-modal transformer. This novelty expands the original navigation system's ability by giving it awareness of surrounding actions. Consequently, our system can holistically consider different modalities when calculating the agent's next step to their destination. Agents can then both accurately navigate dynamic environments, and also recognize human actions.

2) we develop a novel action-based simulator through innovating a self-supervised Human Action Generator that creates a broad selection of 3D human actions. Then, we integrate these generated actions into my chosen environment simulator to provide a dynamic training environment to achieve better performance and generalization abilities.

3) we define a novel benchmark for measuring action-aware navigation performance with the Completeness and Accuracy metrics. My novel benchmark measures both the accuracy of the navigation path and also the accuracy of the action recognition results.

4) My system based on cloud-client architecture can solve the critical challenge of helping the visually impaired understand and navigate the world. The client end contains my user-friendly mobile APP connected to a pair of ordinary glasses with a small camera attached to it. The client end then calls on the cloud end, which contains our AAVLN algorithm deployed onto a cloud server.

2. Related Works

Existing papers discuss various aspects of navigation and action recognition algorithms, and environment simulators for Reinforcement training.

Previous works in the navigation field include graph-based SLAM (Simultaneous Localization and Mapping) [2], Visual SLAM (VSLAM) [3], Vision-and-Language Navigation (VLN) [4], recurrent vision-and-language BERT for navigation [5], VLN with self-supervised auxiliary reasoning tasks [6], learning to navigate unseen environments [7], and learning a generic agent for vision-and-language navigation [8].

Related action recognition algorithms [9] include video action recognition in sports [10] and representation learning for human skeleton-based action recognition [11], [12].

Environment simulators and human action generation works include SceneNN [13], Matterport3D with the Room-to-Room (R2R) navigation dataset[14], skinned multi-person linear models [15], and generating 3D people in scenes [16], [17].

2.1. Navigation Solutions

Since there are minimal developed navigation methods for visually impaired people, we first analyzed the current navigation solutions for robots. There are two main categories: Visual Simultaneous Localization and Mapping (VSLAM) and Vision Language Navigation (VLN).

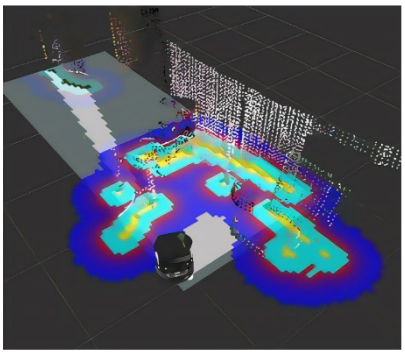
Visual Simultaneous Localization and Mapping (VSLAM), is a technique used by robots to navigate an environment using 3D vision to determine their position, orientation and create a completely scanned map of their surroundings (Figure 2a). This is achieved by tracking set points in successive camera frames. However, there are no labels for each object, so the agent cannot fully understand its surroundings. Hence, navigation with VSLAM algorithms are restricted to specific, scanned environments and does not contain scene understanding of semantic information.

Vision Language Navigation (VLN), is a technique used by robots to navigate an environment by understanding semantic information and natural language navigation instructions (Figure 2b). Natural language instructions for navigation with the destination are given to the algorithm. The algorithm comprehends the instruction then directs the agent to travel in the environment based on the algorithm's understanding of the environment's semantic information in relation to the instructions. The current challenge with VLN algorithms is that they have only been trained on simulators with static

DOI: 10.54254/2755-2721/57/20241309

environments and cannot recognize human actions. We propose an Action-Aware Vision Language Navigation system with Action Recognition to address this.

(a) Visual Simultaneous Localization and Mapping (VSLAM)



(b) Vision Language Navigation (VLN)

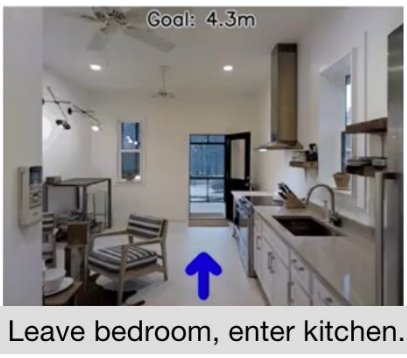


Figure 2. Existing navigation solutions VSLAM and VLN

2.2. Training Simulators

We also investigated current simulators, such as SceneNN [14] and Matterport3D [15], to use for Reinforcement training. However, these simulators do not contain human actions and are often restricted to single rooms, such as SceneNN. While Matterport3D provides floor plans and semantic annotations, it does not contain human actions. We aim to simulate human actions in the virtual environment for Reinforcement training of the Agent Algorithm.

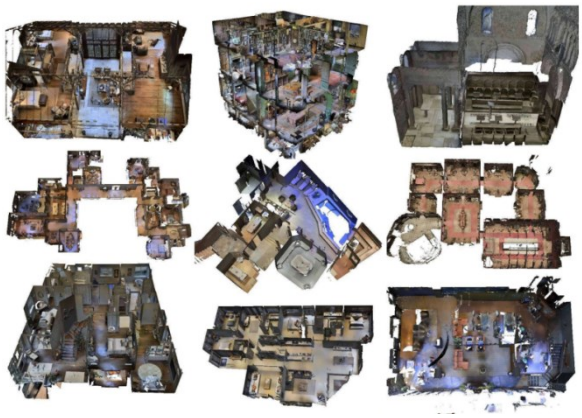


Figure 3. Matterport3D simulator environments

3. Design and Implementation of AAVLN

The Action-Aware Vision Language Navigation (AAVLN) system introduces a novel approach enabling agents to navigate and comprehend dynamic environments containing human actions.

3.1. Overview of the AAVLN System

To address visually impaired people's needs, we develop an assistance system based on cloud-client architecture. There is a user-friendly mobile APP on the client side and my AAVLN on the server side (Figure 4). AAVLN helps people understand and navigate dynamic environments containing human actions, meeting visually impaired people's needs. Other recognition functions are available as well on the server end, including text, object, currency, and face recognition.

In the client end, the mobile APP obtains the image and performs image-processing before sending the image to cloud for AI inference. The results are returned to the client end and go through json parse.

DOI: 10.54254/2755-2721/57/20241309

Finally, they are displayed on the APP screen and speech synthesis reads aloud the results.In the cloud end, the service AI consists of our main service AAVLN, which is implemented onto a Flask Web Framework. Other services call on Third Party's AI servers (Figure4).

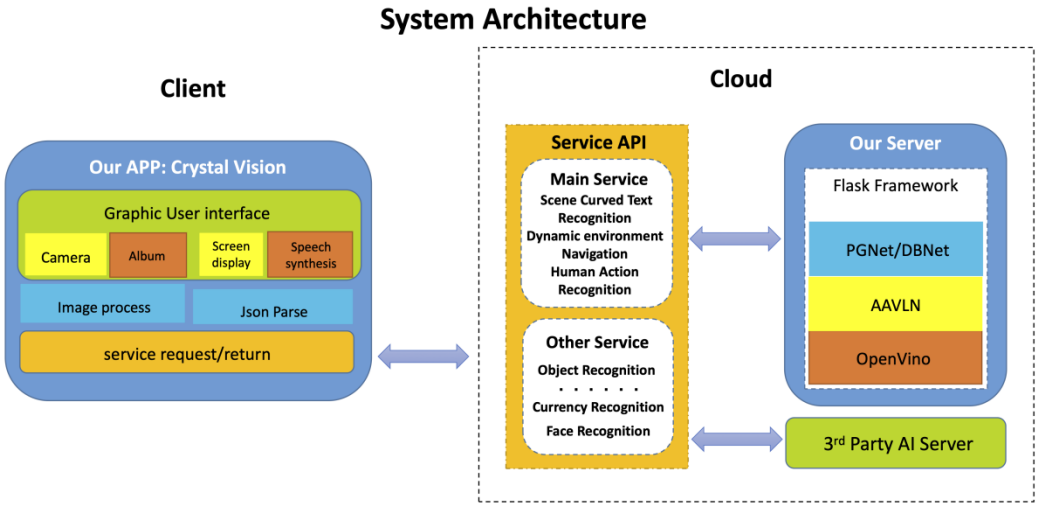


Figure 4. System Architecture of AAVLN

My AAVLN system assists visually impaired people in 5 steps as depicted in the flowchart below. First, users open the camera on their phone or attached to their glasses. Secondly, users verbally input their navigation directions through speech recognition, and our app captures it. Thirdly, our user-friendly mobile app calls on our AAVLN system that is deployed onto cloud. Fourthly, AAVLN calculates and returns the navigation and action recognition results. Finally, users hear the results from their devices to navigate dynamic environments and understand human actions.

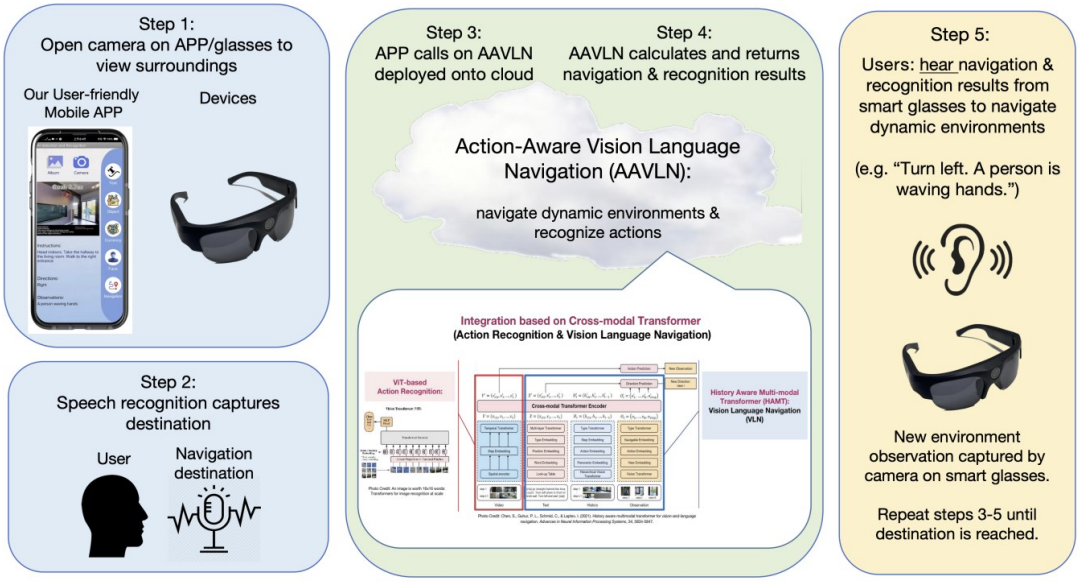


Figure 5. Action-Aware Vision Language Navigation (AAVLN) Flowchart

DOI: 10.54254/2755-2721/57/20241309

3.2. AAVLN Reinforcement Learning Framework

To develop our AAVLN agent algorithm, we innovatively combine a Vision Language Navigation algorithm and an action recognition module for cross-modal encoding. To train the algorithm, we use Reinforcement Learning with our action-based simulator, which we developed by combining our novel human action generator with the existing recognized Matterport3D simulator.

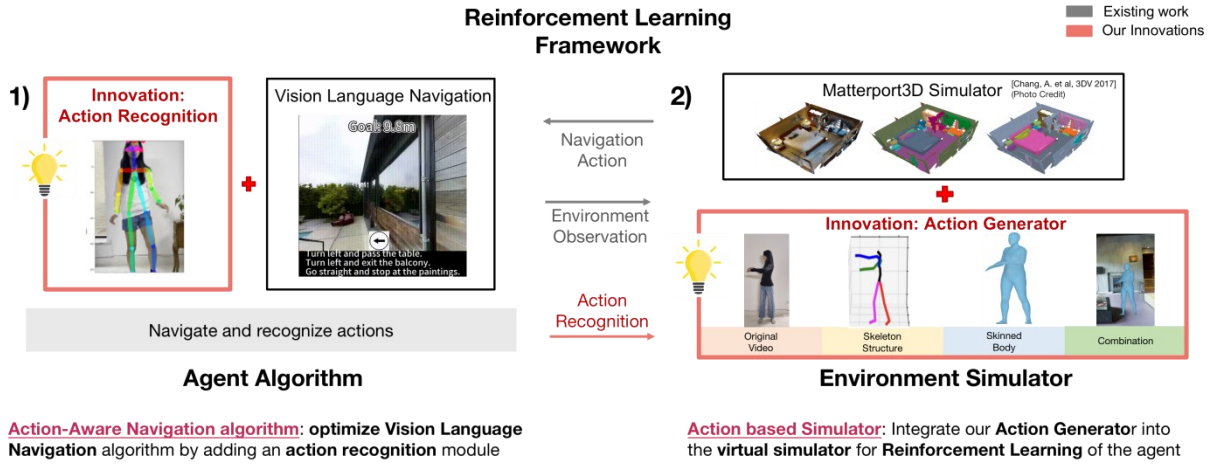


Figure 6. AAVLN Reinforcement Learning Framework

3.3. AAVLN: Agent Algorithm

AAVLN system contains an agent algorithm capable of both navigation and action recognition in environments. We added a Vision Transformer (ViT)-based action recognition branch[18] into the History Aware Multimodal Transformer (HAMT) [19], a Vision Language Navigation (VLN) algorithm. This integration allows the two algorithms to efficiently share the same backbone for cross-modal encoding. Our proposed model can jointly interpret natural language text, history observations, current view, and action videos for navigation while simultaneously achieving action recognition.

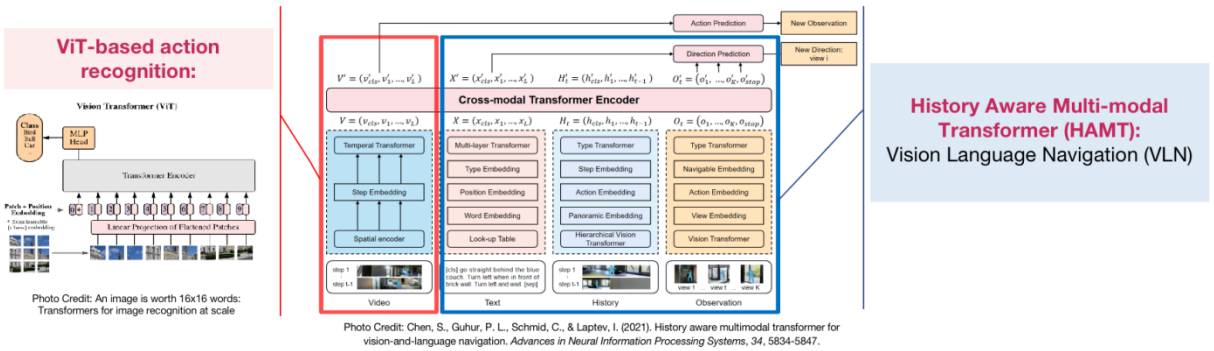


Figure 7. AAVLN Agent Algorithm Blueprint

3.3.1. Navigation: History-Aware Multimodal Transformer (HAMT)

The core of our AAVLN system, the HAMT, has a transformer-based architecture for VLN. The HAMT architecture achieves multimodal decision-making by encoding text, long-horizon history, and observation inputs together through unimodal encoders.

DOI: 10.54254/2755-2721/57/20241309

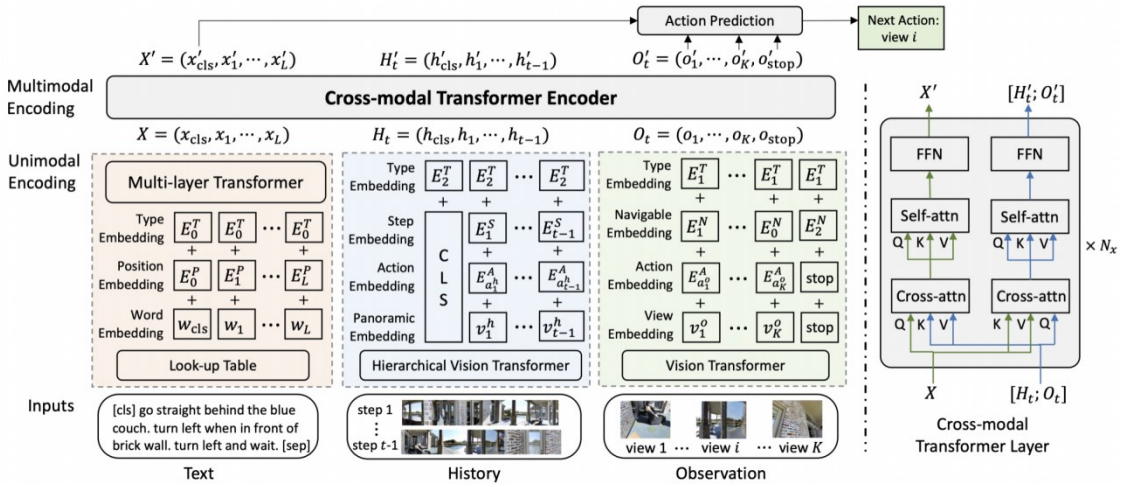


Figure 8. History-Aware Multimodal Transformer (HAMT) Structure

The textual input component embodies natural language instructions, the historical data encompasses previously navigated views, while the present observations spotlight current views. A hierarchical encoding structure within the History unimodal encoder first encodes individual images using a ViT. Following this, it models the spatial relationships between the images in each panorama, and eventually captures the temporal correlations spanning panoramas throughout history. This web of data then undergoes a cross-modal transformer encoder to capture the multimodal relationships.

Experiments conducted on various datasets with fine-grained instructions, high-level instructions, and dialogues demonstrate that HAMT outperforms other VLN algorithms and achieves state-of-the-art performance on both previously seen and unseen environments. Thus, HAMT was the selected and optimized navigation algorithm in this project.

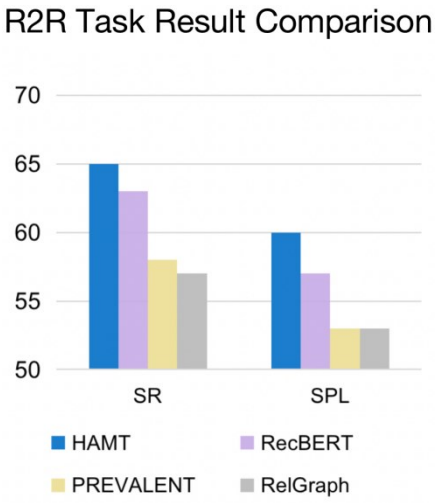


Figure 9. Comparative analysis of different Vision Language Navigation algorithms

3.3.2. Action Recognition: Novel Integration of ViT into HAMT

The proposed system innovatively incorporates an Action Recognition module to assist the agent in navigating and recognizing actions in dynamic environments. Initially, we intended to integrate the skeleton-based method Spatio-Temporal Graph Convolutional Network (STGCN) [20] into HAMT for its strong generalization capability with action recognition. It recognizes actions through extracting the skeleton structure of each person and determining the action category based on the skeleton movements.

DOI: 10.54254/2755-2721/57/20241309

However, ST-GCN did not achieve a high action recognition accuracy in our action-based simulator, reaching only 59%.

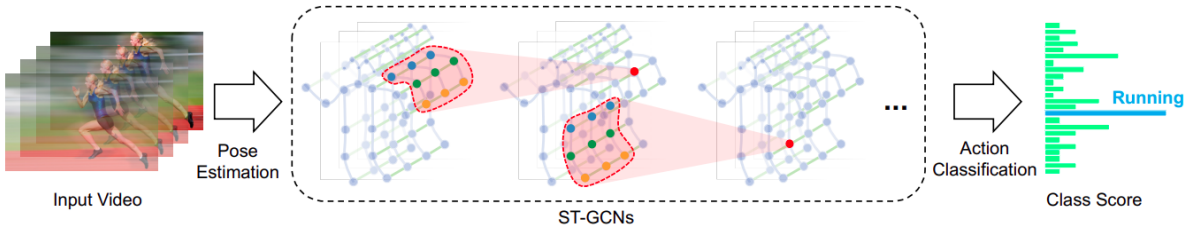


Figure 10. Spatio-Temporal Graph Convolutional Network (ST-GCN) Structure [20]

After further comparison, we selected a ViT-based Action Recognition module for its high accuracy and efficiency. The ViT-based recognition module outperforms other skeleton-based methods in our action-based simulator, achieving a 76% accuracy. Its selection was further justified by HAMT's pre-existing ViT structure. Therefore, we integrate this action recognition module into the existing HAMT algorithm so both components share a common backbone. This allows navigation and action recognition to be achieved with a greater efficiency.

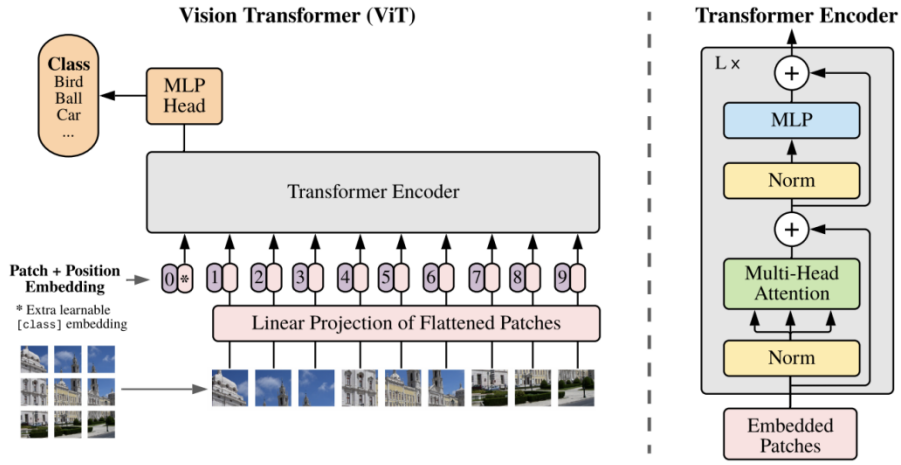


Figure 11. Vision Transformer (ViT) Structure [18]

3.4. AAVLN: Environment Simulator

To train the agent algorithm with Reinforcement Learning, we use a virtual environment simulator. For the environment simulator, we selected the recognized Matterport3D. Since existing simulators for Reinforcement training do not contain human actions, we innovate an action generator to create 3D human action videos that we integrate into our novel action-based environment simulator.

3.4.1. Virtual Simulator: Matterport 3D

The Matterport3D simulator consists of virtual environments for the Reinforcement Training of Vision Language Navigation tasks. The simulator dataset contains 90 buildings and 43,200 environment images. It features annotator-specified floor plans and instance-level semantic annotations, which offer a significant advantage for training agents. However, one of the major limitations of the Matterport3D environment simulator is it lacks human actions, a common limitation among existing simulators. This means they can not effectively train Action-Aware agent algorithms with Reinforcement Learning.

DOI: 10.54254/2755-2721/57/20241309

3.4.2. Action Generator

To train the agent's action recognition module using Reinforcement Learning, we introduce an innovative Action Generator that generates 3D human actions for the environment simulator. We proposed a self-supervised Action Generator based on Variational Autoencoder (VAE) [21] and Skinned Multi-Person Linear Model (SMPL).

The VAE, with its self-supervised nature, offers the advantage of generating a more diverse set of action skeletons within a single category. SMPL then generates 3D human actions with variations in size and poses for each action category based on the VAE's output. The Action Generator successfully generates 3D human skinned bodies performing 52 distinct types of actions with over 20,000 poses.

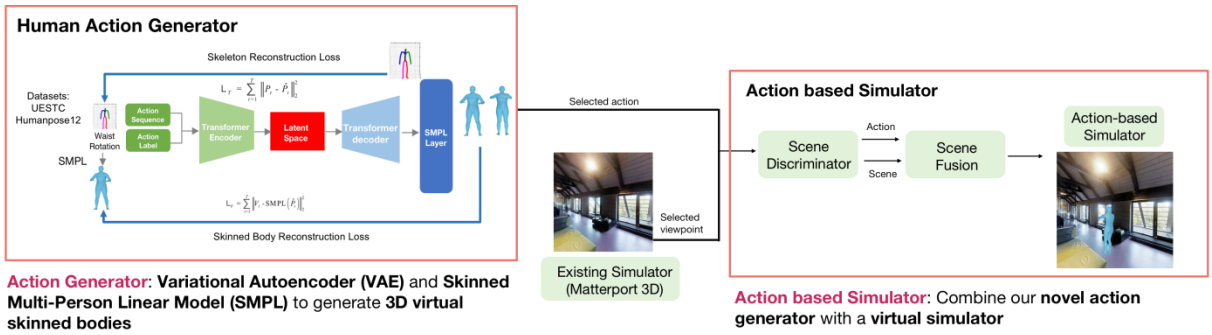


Figure 12. Integration of Novel 3D Human Action Generator with Environment Simulator

We successfully included 3D human figures doing different poses from different perspectives into our environment simulator. This allows our action-based simulators to simulate the variations that can occur in real-world dynamic environments to train our agent's navigation and action recognition abilities accordingly. The synthesis of our human action generator and the acclaimed Matterport3D simulator creates an action-centric simulator tailored for the agent's Reinforcement training. Key components in this integration include the Scene Discriminator, which determines the optimal trajectories for the incorporation of generated actions; the Scene Fusion then seamlessly weaves these actions into the simulator. This process successful integrates 3D human actions across 1,050 viewpoints, which is over 10,000 images. The Scene Discriminator and Scene Fusion ensure that the action sequences are harmoniously combined into the environment.

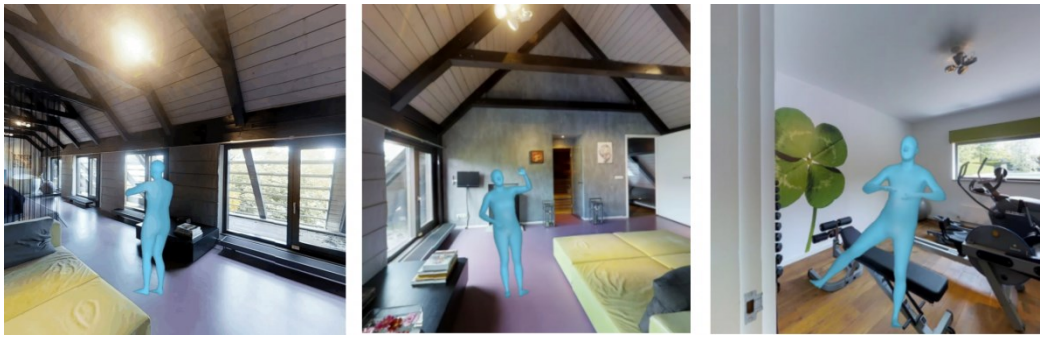


Figure 13. Action-based Simulator: Integration of Generated 3D Human Actions into the Environment Simulator

4. Evaluation

4.1. Panoramic Demonstrations

After training our AAVLN in our action-based simulator for over 200,000 iterations, we obtained panoramic demonstrations of our agent traversing through a given environment. These panoramic images are from the agent's perspective. The box within the two red lines indicate the agent's current

DOI: 10.54254/2755-2721/57/20241309

view of the environment when looking directly forward. The top left "Instruction" is the natural language navigation instructions given by the user to AAVLN, informing the AI system of their destination. Each arrow indicates the direction that AAVLN computes and returns to the user. The top right "Observation" is AAVLN's action recognition results after observing the surrounding people (simulated by the generated 3D blue figures).

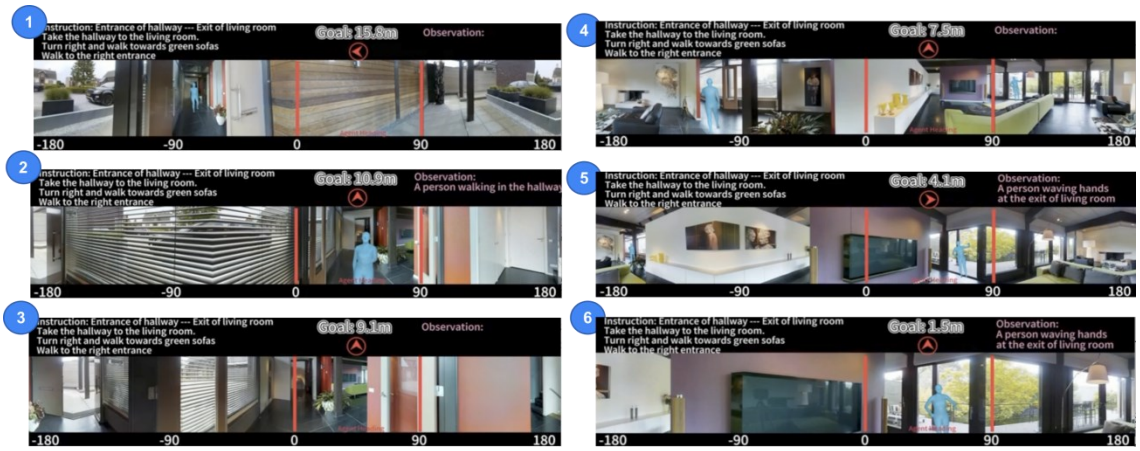


Figure 14. Panoramic demonstrations of our AAVLN in our action-based environment simulator

4.2. Evaluation Metrics

The main evaluation metrics adopted include Success Rate (SR) and Success weighted by Path Length (SPL). SR computes the proportion of trajectories that successfully reach their destination within a 3-meter margin of error from the target. On the other hand, SPL standardizes the success rate by considering the ratio of the optimal path length to the predicted path's length, considering the efficiency of the algorithm. The Action-Aware Vision Language Navigation (AAVLN) system underwent training on our novel 3D human action-based simulator for over 200,000 iterations. We compared its performance with prevailing VLN algorithms using the Room-to-Room navigation dataset (R2R), which was co-developed with the Matterport3D simulator. This dataset features 7189 paths (average length: 10 meters), containing 21,567 navigation instructions. For training and validation (seen), we used 61 scenes, with 14,025 and 1020 instructions, respectively. For validation in unseen environments, we used 11 scenes (2349 instructions). Finally for testing, we used 18 scenes (4173 instructions).

4.3. Experiment I: Evaluation of AAVLN on R2R Task

In the pursuit of evaluating navigation algorithms' proficiency in dynamic environments, we present Experiment we to benchmark the performance of our novel AAVLN system against traditional VLN algorithms. This experiment was conducted using our novel action-based simulator with dynamic 3D human actions.

The table demonstrates how our AAVLN outperforms the baseline HAMT on both the SR and SPL metrics in seen and unseen environments (Table 1). Compared to existing prominent VLN algorithms, our AAVLN achieves an average of 11.9% to 22.9% improvement in SR for seen environments, 21.7% to 14.4% improvement in SPL for seen environments, 12.8% to 22.8% improvement in SR for unseen environments, and 15% to 24.3% improvement in SPL for unseen environments.

DOI: 10.54254/2755-2721/57/20241309

Table 1. A Comparative Analysis of R2R Task Performance in our Action-Based Simulator (Bracketed Values Indicate Original R2R Task Results in Static Environments)

Method	SR Seen(↑)	SPL Seen(↑)	SR Unseen(↑)	SPL Unseen(↑)
PREVALENT [22]	$50.3_{\pm 2.1}$ (69)	$47.8_{\pm 1.9}$ (65)	$39.2_{\pm 2.5}$ (58)	$34.5_{\pm 0.8}$ (53)
RelGraph [23]	$49.1_{\pm 2.9}$ (67)	$46.4_{\pm 1.6}$ (65)	$37.2_{\pm 1.7}$ (57)	$34.0_{\pm 2.0}$ (53)
RecBERT [5]	$52.5_{\pm 2.7}$ (72)	$48.2_{\pm 1.8}$ (68)	$41.6_{\pm 1.3}$ (63)	$37.2_{\pm 0.9}$ (57)
HAMT (without mask) [6]	$54.7_{\pm 1.5}$ (76)	$50.9_{\pm 2.1}$ (72)	$45.5_{\pm 1.8}$ (66)	$41.2_{\pm 1.1}$ (61)
HAMT (with mask) [6]	$61.3_{\pm 2.5}$ (76)	$55.1_{\pm 1.7}$ (72)	$49.2_{\pm 1.5}$ (66)	$43.8_{\pm 1.3}$ (61)
Ours	$73.2_{\pm 1.9}$	$69.5_{\pm 1.3}$	$62.0_{\pm 1.1}$	$58.8_{\pm 1.6}$

It is worth highlighting the divergence in performance between traditional VLN algorithms and AAVLN. Notably, when exposed to the dynamic action-based environment of our simulator, conventional VLN algorithms demonstrate a marked decrease in efficacy. This is depicted by their decrease in accuracy from the bracketed values to those above it. In contrast, the AAVLN system exhibits exemplary navigation capabilities. This superior performance is due to the cross-modal transformer that gives AAVLN the ability to understand dynamic human actions while navigating.

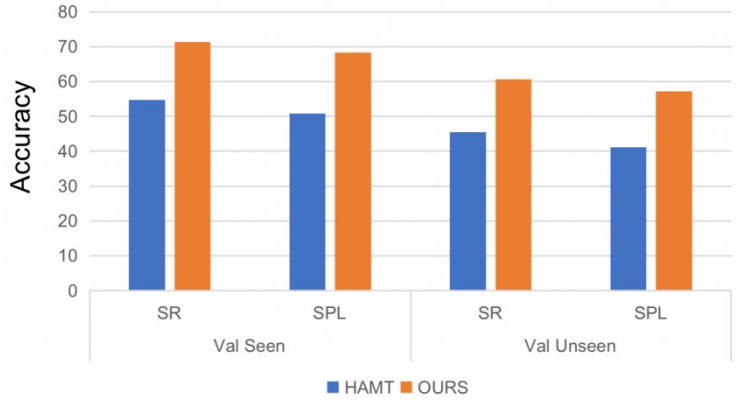


Figure 15. Comparison of baseline (HAMT) with our AAVLN

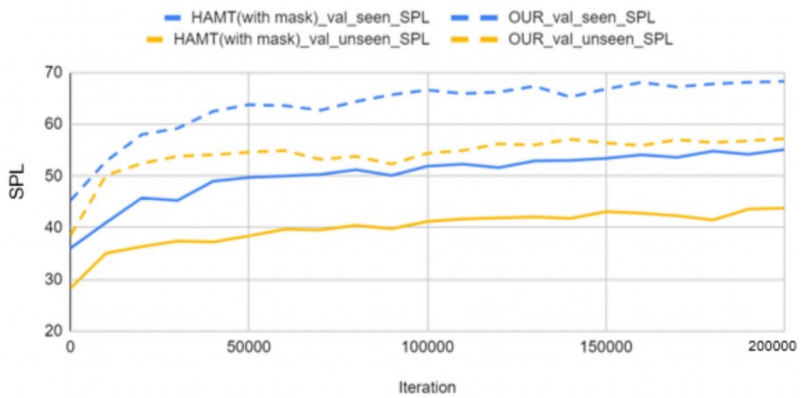


Figure 16. Comparison of HAMT with our AAVLN for over 200,000 iterations

This graph further demonstrates that throughout the 200,000 iterations, our AAVLN constantly maintains a higher navigation accuracy in dynamic environments than the baseline HAMT. Even as the graphs begin to plateau, our AAVLN remains more accurate. This is due to HAMT's lack of ability to

DOI: 10.54254/2755-2721/57/20241309

understand dynamic environments when navigating. Even though HAMT achieves exceptional performance when navigating static environments, its performance will severely decrease when unfamiliar obstacles are presented along navigation paths, such as dynamic actions. There is a significant difference between actions and static objects, but traditional VLN algorithms such as baseline HAMT have not been previously trained to understand dynamic actions

4.4. Experiment II: Evaluation of Action Recognition and Cross-modal Encoding on Navigation

In Experiment II, we conducted ablation studies to investigate the effect of various action recognition methods and the usage of cross-modal encoding on navigation performance in our action-based simulator

Our first ablation study's findings indicate that the integration of ViT-based action recognition as a proxy task significantly improved both navigation and action recognition accuracy. Compared to the combination of HAMT with other action recognition modules, our AAVLN combines HAMT with a ViT-based recognition module, achieving an average improvement of 1.8% to 7.8% in SR for seen environments, 1.3% to 11.2% improvement in SPL for seen environments, 0.3% to 9.7% improvement in SPL for unseen environments, and 0.7% to 12.3% in SPL for unseen environments. This highlights the important role of considering action recognition in navigation, and the potential for ViT-based action recognition methods to enhance overall performance. Without the ability to understand and recognize human actions, it is difficult for an algorithm to navigate dynamic environments because the difference between moving actions and static objects is drastic.

Our second ablation study demonstrates that the utilization of cross-modal information between navigation and action recognition significantly increases navigation accuracy. Compared to without using cross-modal encoding, our AAVLN with cross-modal encoding achieves a 4.1% improvement for SR in seen environments, 3.2% higher for SPL in seen environments, 3.1% improvement for SR in unseen environments, and 2.7% higher accuracy for SPL in unseen environments. This cross-modal information allows agents' navigation decisions to be influenced by their consideration of surrounding actions, thus increasing navigation accuracy in dynamic environments.

Table 2. Ablation Studies of Navigation Performance Without Cross-Modal Encoding. Our AAVLN based on HAMT and ViT outperforms other combinations of HAMT and other action recognition modules in navigation

Action Encoding	SR Seen(↑)	SPL Seen(↑)	SR Unseen(↑)	SPL Unseen(↑)
HAMT(Baseline)	$61.3_{\pm 2.5}$	$55.1_{\pm 1.7}$	$49.2_{\pm 1.5}$	$43.8_{\pm 1.3}$
HAMT + STRNN	$66.2_{\pm 2.3}$	$63.9_{\pm 1.2}$	$56.8_{\pm 1.5}$	$53.4_{\pm 1.6}$
HAMT + InfoGCN	$66.8_{\pm 1.7}$	$64.1_{\pm 1.5}$	$57.3_{\pm 0.8}$	$54.0_{\pm 1.3}$
HAMT + STGCN	$67.3_{\pm 2.1}$	$65.0_{\pm 2.3}$	$58.6_{\pm 1.9}$	$55.4_{\pm 0.7}$
HAMT + ViT (OURS)	$69.1_{\pm 2.2}$	$66.3_{\pm 1.5}$	$58.9_{\pm 1.1}$	$56.1_{\pm 1.0}$

Table 3. Comparison of AAVLN performance with and without cross-modal encoding

Cross-modal: HAMT & ViT	SR Seen(↑)	SPL Seen(↑)	SR Unseen(↑)	SPL Unseen(↑)
Without	$69.1_{\pm 2.2}$	$66.3_{\pm 1.5}$	$58.9_{\pm 1.1}$	$56.1_{\pm 1.0}$
With	$73.2_{\pm 1.9}$	$69.5_{\pm 1.3}$	$62.0_{\pm 1.1}$	$58.8_{\pm 1.6}$

4.5. Experiment III: Evaluation of Action-Aware Navigation on novel benchmark

In Experiment III, we create a new benchmark for Action-Aware Navigation by adding 3D human actions as ground truth. Our novel task not only tests navigation performance, but also the accuracy of action observations along the path, making it a more challenging and comprehensive evaluation. Completeness (Com) measures the percentage of complete observations along the path. Accuracy (Acc)

DOI: 10.54254/2755-2721/57/20241309

measures the correct action observations along the path. The results of the experiment show that AAVLN exhibits high accuracy in both seen and unseen environments, as measured by our Completeness (Com) and Accuracy (Acc) evaluation metrics. This indicates that the system is capable of effectively navigating environments while accurately recognizing human actions, making it a promising solution for robots and AR systems.

5. Discussion and Future Work

Augmenting the Simulator Dataset: To improve this system's practicality, we are broadening the simulator dataset to further train AAVLN. Currently, I'm constructing 3D models of real-world environments, especially of diverse public spaces in which visually impaired people wish to receive assistance, by leveraging the Matterport3D APP and Insta360 camera.

Table 4. AAVLN performance on our novel evaluation metrics. AAVLN demonstrates high accuracy based on our Completeness and Accuracy metrics

Method	Com Seen(↑)	Acc Seen(↑)	Com Unseen(↑)	Acc Unseen(↑)
AAVLN	91	79	78	73

Through communicating with low vision expert doctors from the Fudan University Affiliated Hospital, they expressed great interest in this project and willingly offered support. We are collaborating with them to construct 3D models of their clinic. These expansions lay the groundwork for empowering AAVLN with the ability to navigate more complex, realistic environments.

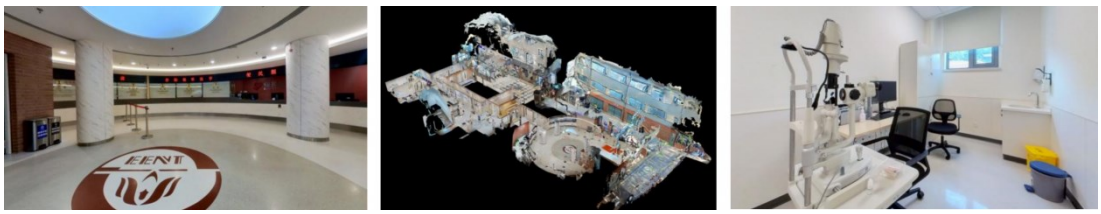


Figure 17. 3D Model of Fudan University's Affiliated Hospital.

Extending AAVLN to Robotic Assistants: Current service robots commonly rely on SLAM-based algorithms, which view all surrounding objects and humans as obstacles that should be avoided. This inability to recognize and interact with surrounding human actions limits their functions. This currently causes them to focus on simple package delivery tasks in spaces like compounds, restaurants, or hospitals. The integration of AAVLN into service robots could bring significant benefits. Robots equipped with AAVLN would not only be able to navigate static environments, but also accurately navigate dynamic environments while understanding surrounding actions, and then perform their tasks accordingly. For example, the robot could interact with a person's gestures in a shopping mall, travel over to ask for their need, then guide them to the destination.

Additionally, envisioning AAVLN's utility in autonomous vehicles [24], my underlying cross-modal transformer structure and Reinforcement Learning mechanisms could potentially aid varied vehicular types. This could minimize accidents by enhancing navigation and action recognition capacities.

Enhancing Natural Language Interactions: During clinical trials with visually impaired users and low vision expert doctors, they suggested that the heard results could be more descriptive of their environment as opposed to just the navigation direction and action recognition. Based on this user feedback, an improvement would be to include more information about the environment expressed in natural language, including further information about surrounding objects and each object or person's position in relation to the visually impaired user.

For this more holistic and user-friendly experience, we propose combining AAVLN with ChatGPT. Such an integration can refine the natural language instructions, optimizing their comprehensibility for

DOI: 10.54254/2755-2721/57/20241309

the underlying HAMT Vision Language Navigation model. Moreover, it can provide users with a more descriptive, natural narration of their surroundings, making interactions between visually impaired users and AAVLN more organic and insightful.

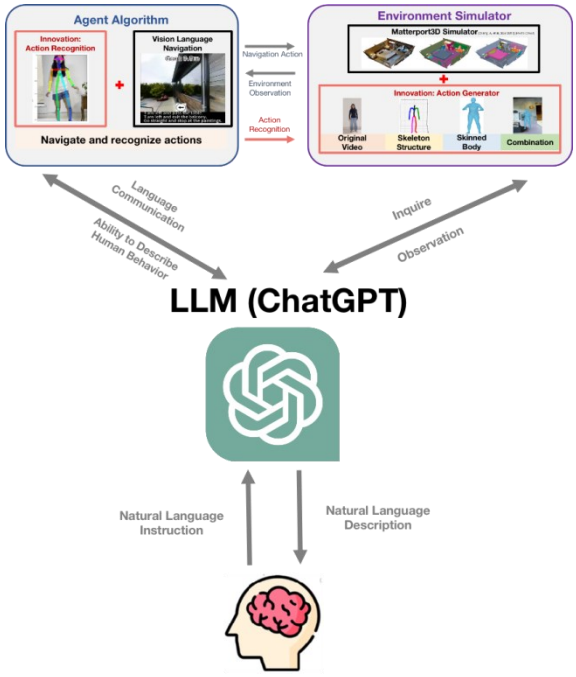


Figure 18. Combination of Chat-GPT with AAVLN to enhance natural language interactions

Augmented Reality (AR): Since AR glasses integrate virtual elements into the real-world, it can help visually present the navigation results of my AAVLN, and my AAVLN can improve AR's abilities. When navigating, the AR glasses could call on my AAVLN deployed onto cloud servers to retrieve navigation directions. Then, virtual arrows could be displayed to let the user know how they should take the next step. Since my system is action-aware, the arrows should also tell users about surrounding people and avoid bumping into them. This could be helpful for all users. For those with visual impairments, the arrows could be magnified to allow them to navigate with the arrows while hearing information about their environments. Furthermore, AAVLN can help AR glasses understand of dynamic real-world environments more comprehensively, including the semantic information and 3D human actions, to then accurately integrate virtual elements in specific locations.

In summary, the trajectory of AAVLN, as mapped in this discussion, reflects a blend of technological advancements, human-centric designs, and collaborations. As we forge ahead, our emphasis remains on creating a system that is not only sophisticated in its functionalities but also inclusive in its applications.

6. Conclusion

In this research, we developed the Action-Aware Vision Language Navigation (AAVLN) system, a cross-modal transformer-based system for navigating and understanding dynamic environments accurately.

Our novel AAVLN agent algorithm optimizes existing VLN algorithms by integrating a distinct action recognition branch, enhancing the algorithm's performance so it can navigate dynamic environments with an understanding of its surroundings. For its Reinforcement Training, we enhanced conventional simulators with 3D human actions generated by our novel action generator for a dynamic environment simulator. The results of our evaluations have shown that our AAVLN system outperforms state-of-the-art methods in both navigation and action recognition tasks, making it a major advancement in the field. Through trials with visually impaired users and doctors in the field of low vision, we

DOI: 10.54254/2755-2721/57/20241309

received positive feedback as they expressed this system could greatly improve the convenience for visually impaired people's daily independent navigation.

References

[1] World Health Organization. (2023, August 10). Vision Impairment and blindness. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

[2] Grisetti, G., K¨ummerle, R., Stachniss, C., Burgard, W. (2010). A tutorial on graph-based SLAM.IEEE Intelligent Transportation Systems Magazine,2(4), 31-43.

[3] Taketomi, T., Uchiyama, H., Ikeda, S. (2017). Visual SLAM algorithms: A survey from 2010 to 2016.IPSJ Transactions on Computer Vision and Applications,9(1), 1-11.

[4] Gu, J., Stefani, E., Wu, Q., Thomason, J., Wang, X. E. (2022). Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions.arXiv preprint arXiv:2203.12667.

[5] Hong, Y., Wu, Q., Qi, Y., Rodriguez-Opazo, C., Gould, S. (2020). A recurrent vision-and-language bert for navigation. arXiv preprint arXiv:2011.13922.

[6] Zhu, F., Zhu, Y., Chang, X., Liang, X. (2020). Vision-language navigation with self-supervised auxiliary reasoning tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10012-10022).

[7] Tan, H., Yu, L., Bansal, M. (2019). Learning to navigate unseen environments: Back translation with environmental dropout. arXiv preprint arXiv:1904.04195.

[8] Hao, W., Li, C., Li, X., Carin, L., Gao, J. (2020). Towards learning a generic agent for vision-and-language navigation via pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13137-13146).

[9] Ulhaq, A., Akhtar, N., Pogrebna, G., & Mian, A. (2022). Vision Transformers for Action Recognition: A Survey. arXiv preprint arXiv:2209.05700.

[10] Wu, F., Wang, Q., Bian, J., Ding, N., Lu, F., Cheng, J., ... Xiong, H. (2022). A Survey on Video Action Recognition in Sports: Datasets, Methods and Applications. IEEE Transactions on Multimedia.

[11] Chi, H. G., Ha, M. H., Chi, S., Lee, S. W., Huang, Q., Ramani, K. (2022). Infogcn: Representation learning for human skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20186-20196).

[12] Xing, Z., Dai, Q., Hu, H., Chen, J., Wu, Z., & Jiang, Y. G. (2022). SVFormer: Semi-supervised Video Transformer for Action Recognition. arXiv preprint arXiv:2211.13222.

[13] Hua, B. S., Pham, Q. H., Nguyen, D. T., Tran, M. K., Yu, L. F., Yeung, S. K. (2016, October). Scenenn: A scene meshes dataset with annotations. In2016 fourth international conference on 3D vision (3DV)(pp. 92-101). Ieee.

[14] Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., ... Zhang, Y. (2017). Matterport3d: Learning from rgb-d data in indoor environments.arXiv preprint arXiv:1709.06158.

[15] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M. J. (2015). SMPL: A skinned multi-person linear model.ACM transactions on graphics (TOG),34(6), 1-16.

[16] Zhang, Y., Hassan, M., Neumann, H., Black, M. J., Tang, S. (2020). Generating 3d people in scenes without people. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6194-6204).

[17] Hassan, M., Choutas, V., Tzionas, D., Black, M. J. (2019). Resolving 3D human pose ambiguities with 3D scene constraints. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 2282-2292).

[18] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. CoRR, abs/2010.11929. Retrieved from https://arxiv.org/abs/2010.11929

DOI: 10.54254/2755-2721/57/20241309

[19] Chen, S., Guhur, P. L., Schmid, C., Laptev, I. (2021). History aware multimodal transformer for vision-and-language navigation.Advances in Neural Information Processing Systems,34, 5834-5847.

[20] Yan, S., Xiong, Y., Lin, D. (2018). Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. CoRR, abs/1801.07455. Retrieved from http://arxiv.org/abs/1801.07455

[21] Petrovich, M., Black, M. J., Varol, G. (2021). Action-conditioned 3D human motion synthesis with transformer VAE. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10985-10995).

[22] W. Hao, C. Li, X. Li, L. Carin, J. Gao. (2020). Towards learning a generic agent for vision-and-language navigation via pre-training.IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. IEEE, 2020, pp. 13 134– 13 143.

[23] Y. Hong, C. R. Opazo, Y. Qi, Q. Wu, S. Gould. (2020). Language and visual entity relationship graph for agent navigation.ACM Digital Library.

[24] Team, T. A. (2022, May 27). Tesla's self driving algorithm explained. Towards AI. https://towardsai.net/p/l/teslas-self-driving-algorithm-explained