Action-Aware Vision Language Navigation (AAVLN): AI Vision System based on Cross-Modal Transformer for Understanding and Navigating Dynamic Environments

Jasmine Liu, Sophia Liu

Shanghai American School, Shanghai, China

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

Contents

1 Introduction **3**
- 1.1 Background and Motivation . . . 3
- 1.2 Contributions . . . 4

2 Related Works **5**
- 2.1 Navigation Solutions . . . 5
- 2.2 Training Simulators . . . 6

3 Design and Implementation of AAVLN **6**
- 3.1 Overview of the AAVLN System . . . 7
- 3.2 AAVLN Reinforcement Learning Framework . . . 8
- 3.3 AAVLN: Agent Algorithm . . . 9
  - 3.3.1 Navigation: History-Aware Multimodal Transformer (HAMT) . . . 9
  - 3.3.2 Action Recognition: Novel Integration of ViT into HAMT . . . 10
- 3.4 AAVLN: Environment Simulator . . . 11
  - 3.4.1 Virtual Simulator: Matterport 3D . . . 11
  - 3.4.2 Action Generator . . . 11

4 Evaluation **13**
- 4.1 Panoramic Demonstrations . . . 13
- 4.2 Evaluation Metrics . . . 13
- 4.3 Experiment I: Evaluation of AAVLN on R2R Task . . . 14
- 4.4 Experiment II: Evaluation of Action Recognition and Cross-modal Encoding on Navigation . . . 15
- 4.5 Experiment III: Evaluation of Action-Aware Navigation on novel benchmark . 17

5 Discussion **17**

6 Conclusion **19**

7 Acknowledgements **20**

References **20**

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 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.

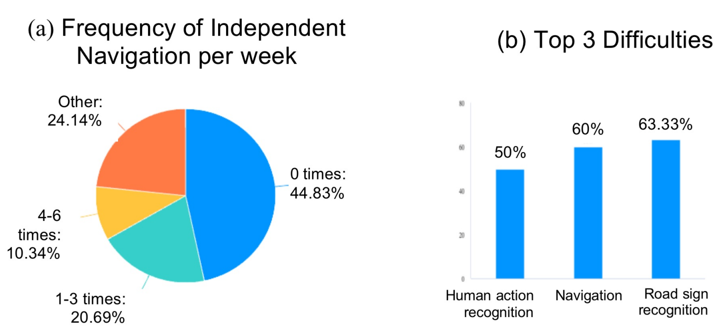

Figure 1: Survey results from our joint survey with visually impaired people and low vision doctors. (a) Over 75% of visually impaired people do not go out alone daily, and 44.83% never independently navigate outdoors. (b) The top 3 difficulties visually impaired people face are road sign recognition, navigation, and human action recognition. This causes them to encounter challenges and safety risks when independently navigating and understanding dynamic environments.

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. Through this project, we gained access to various opportunities where we received feedback and suggestions for upcoming steps. After winning Global Grand Award in Intel's AI Competition, our scene text recognition system was demonstrated at the China International Import Expo (CIIE). There, we encountered multiple AI experts in the field of Computer Vision who offered positive comments for our project, along with suggestions to leverage more recent technological advancements, including Transformers, in upcoming projects. We also received the opportunity to communicate with more visually impaired users and low vision doctors. For the people with visual impairments, they believe this project could bring great value for them, and wished there could be a system that could further help them navigate complex environments more independently. For the expert doctors in the field of low vision, they expressed their interest in providing support for my upcoming developments with this project. Similarly, during the 2022 Yau Shing Tung High School Science Competition, I received positive feedback from the professors and AI experts about my project, winning Bronze Award in Computer Science. They also suggested I can expand the system to assist visually impaired people in more ways. This inspired me to develop a new project that could help them navigate and understand dynamic environments.

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

There are 4 main contributions of our paper:

1) We develop a novel Agent Algorithm through innovatively adding an action recognition module 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 our 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. Our novel benchmark measures both the accuracy of the navigation path and also the accuracy of the action recognition results.

4) Our 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 our 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.

Our system can be applied to smart glasses to solve the critical challenge of helping the visually impaired understand and navigate the world.

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 [3], Vision-and-Language Navigation (VLN) [4], recurrent vision-and-language BERT for navigation [5], VLN with self-supervised auxiliary reasoning tasks [7], learning to navigate unseen environments [8], and learning a generic agent for vision-and-language navigation [9].

Related action recognition algorithms [10] include video action recognition in sports [11] and representation learning for human skeleton-based action recognition [12], [13].

Environment simulators and human action generation works include SceneNN [14], Matterport3D [15], skinned multi-person linear models [16], and generating 3D people in scenes [17], [18].

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) [3] and Vision Language Navigation (VLN) [4].

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, VSLAM is 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 given by an oracle (Figure 2b). The current challenge with VLN algorithms is that they have only been trained on simulators with static 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)

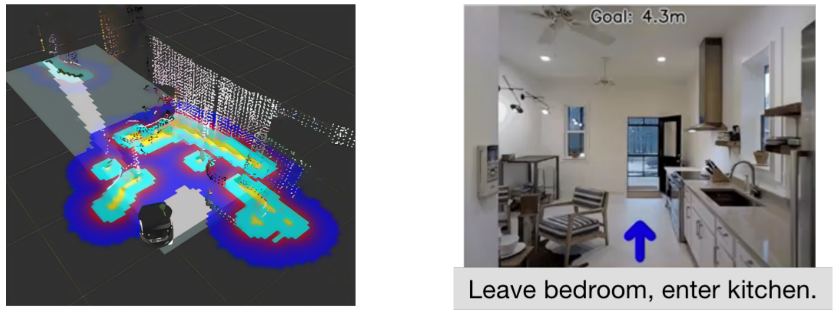

Figure 2: (a)VSLAM visualization. A map of the surroundings is created with obstacles localized in the environment. However, there are no labels for each object, so the agent cannot fully understand its surroundings.[3]. (b) Vision Language Navigation. Natural language instructions for navigation with 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.[4].

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.

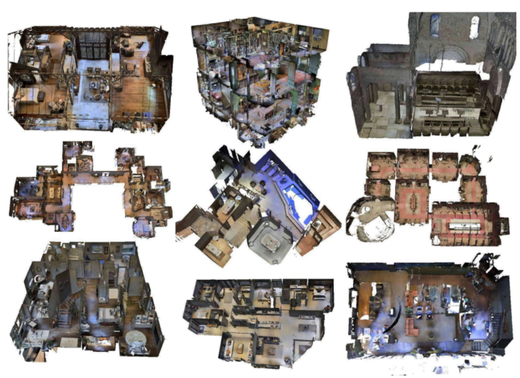

Figure 3: Matterport3D simulator environments are all static and do not contain 3D human actions.[15].

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 our 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. Finally, they are displayed on the APP screen and speech synthesis reads aloud the results.

In the cloud end, the service API consists of our main service AAVLN, which is implemented onto a Flask Web Framework. Other services call on Third Party's AI servers (Figure 4).

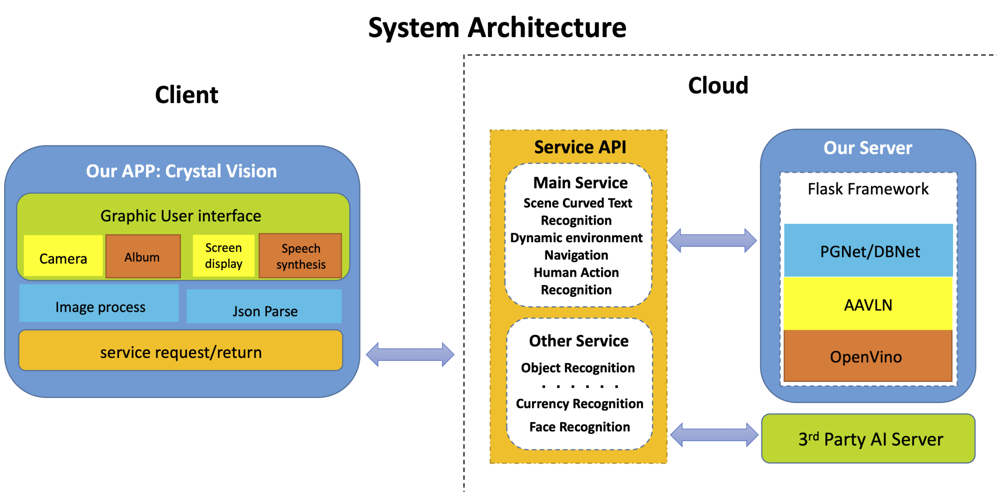

Figure 4: System Architecture of AAVLN

Our 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).

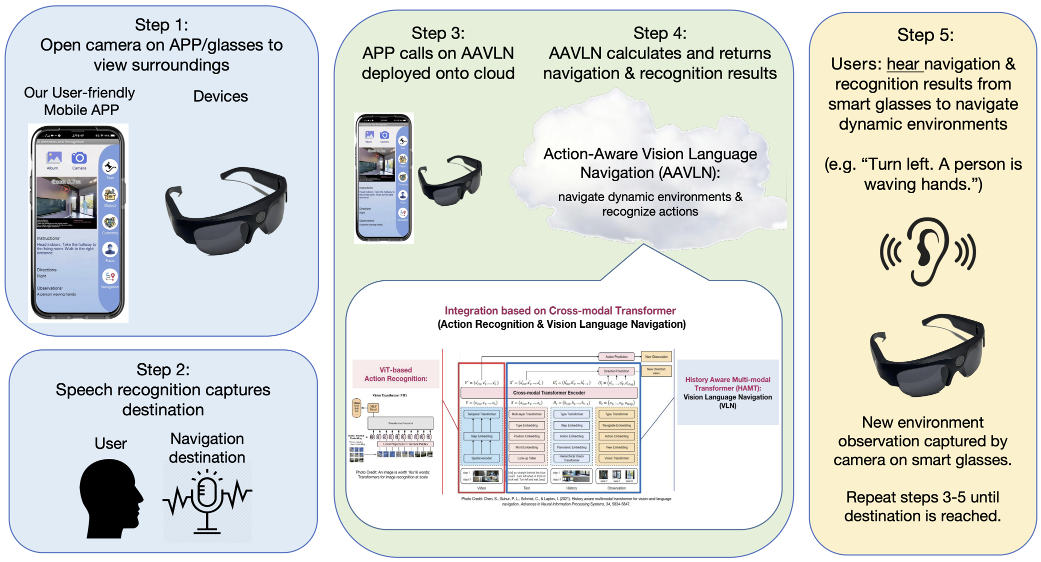

Figure 5: Action-Aware Vision Language Navigation (AAVLN) Flowchart. This flowchart illustrates how our AAVLN can be applied to help the visually impaired with navigation and recognition.

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.

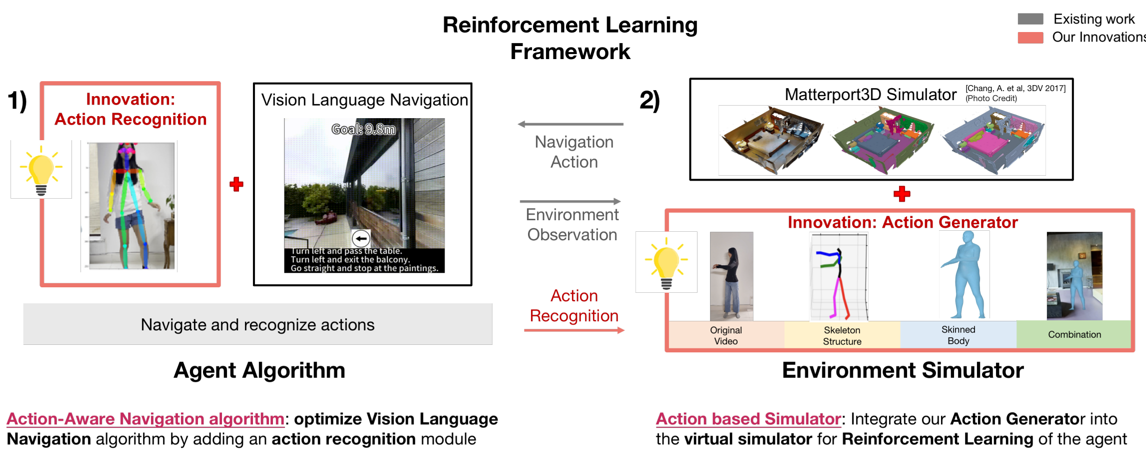

Figure 6: AAVLN Reinforcement Learning Framework. AAVLN: 1) amplifies the Vision Language Navigation (VLN) algorithm by converging it with an Action Recognition module through cross-modal encoding, leading to a comprehensive Action-Aware Navigation system, and 2) develops an action-based virtual simulator supplemented by a 3D Human Action Generator for Reinforcement training in dynamic environments.

3.3 AAVLN: Agent Algorithm

AAVLN system contains an agent algorithm capable of both navigation and action recognition in environments. We integrate a Vision Transformer (ViT)-based action recognition branch [20] into History Aware Multimodal Transformer (HAMT), a Vision Language Navigation (VLN) algorithm [6]. 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.

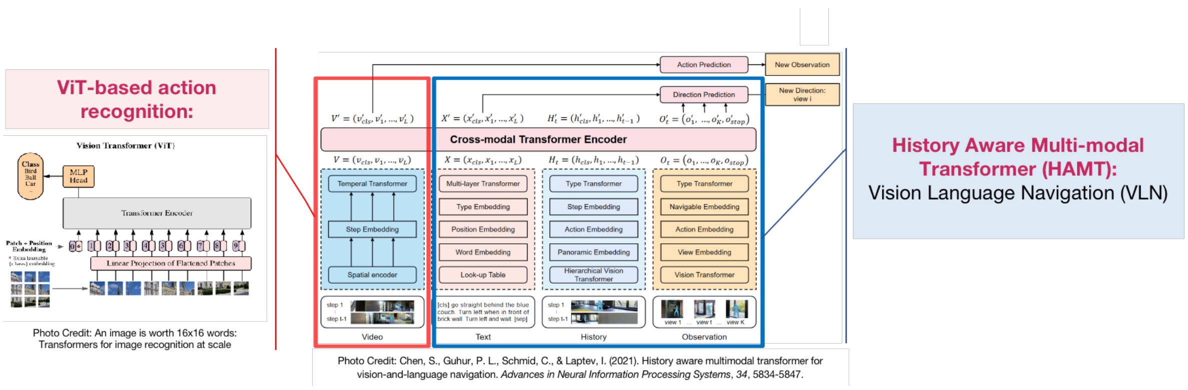

Figure 7: Agent Algorithm Blueprint: Seamless integration of the Vision Transformer-based action recognition with the History Aware Multi-modal Transformer [6] Vision Language Navigation approach for optimized cross-modal encoding.

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.

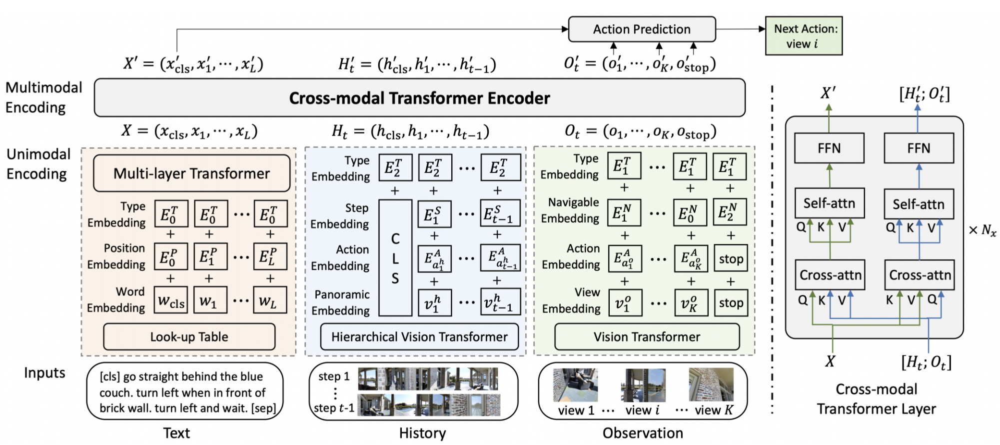

Figure 8: Depiction of the History Aware Multi-modal Transformer: Integrated unimodal encoders for text, historical data, and current observations, followed by a cross-modal transformer encoder [6].

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.

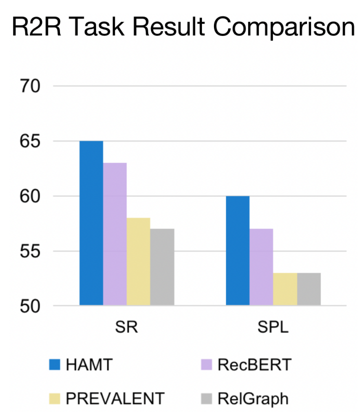

Figure 9: Comparative analysis illustrating the proficiency of different Vision Language Navigation models on the Room-to-Room dataset. Metrics employed include Success Rate (SR) and Success rate normalized by Path Length (SPL). Higher percentage indicates better performance, so HAMT was selected as our Vision Language Navigation algorithm.

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 ST-GCN [19] 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. However, ST-GCN did not achieve a high action recognition accuracy in our action-based simulator, reaching only 59%.

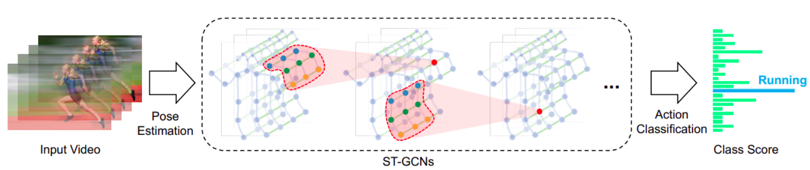

Figure 10: Preliminary design leveraging the skeleton-driven ST-GCN methodology for human action recognition. ST-GCN encompasses pose estimation followed by action categorization [19].

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.

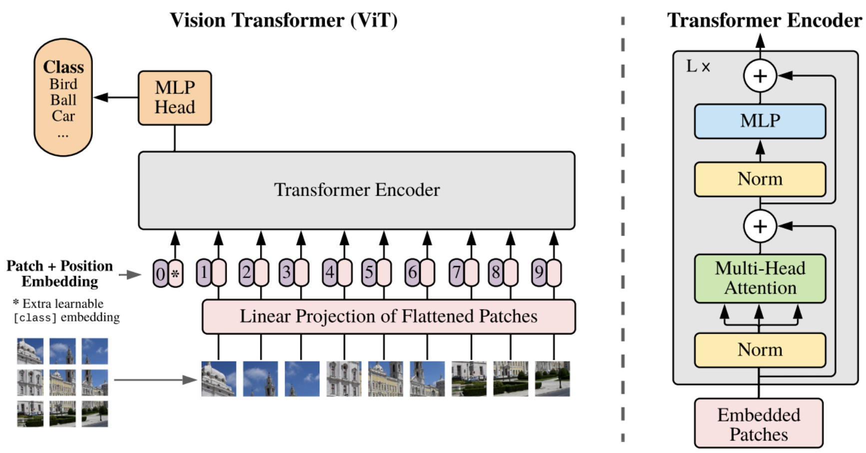

Figure 11: Design schematics of the Vision-Transformer (ViT) Action Recognition module, now seamlessly fused with the HAMT navigational algorithm [20].

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.

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) [16].

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.

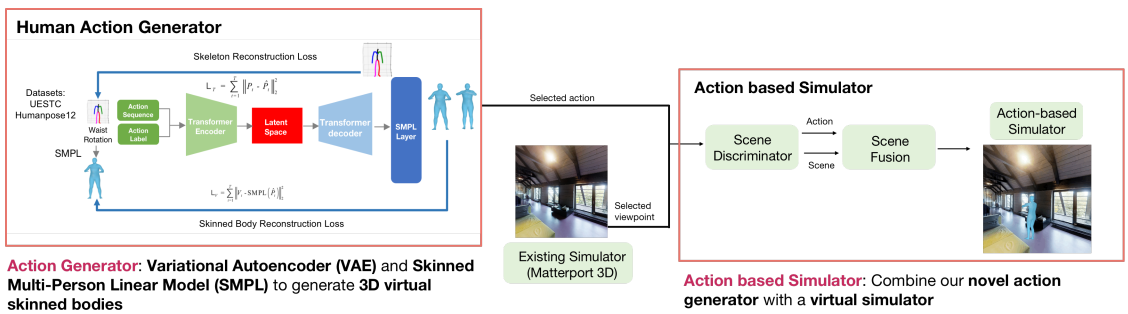

Figure 12: The self-supervised Action Generator based on a Variational Autoencoder (VAE) and the Skinned Multi-Person Linear Model (SMPL) has the capability to produce virtual 3D human models. Its integration into our environment simulator paves the way for a dynamic environment for Reinforcement training.

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.

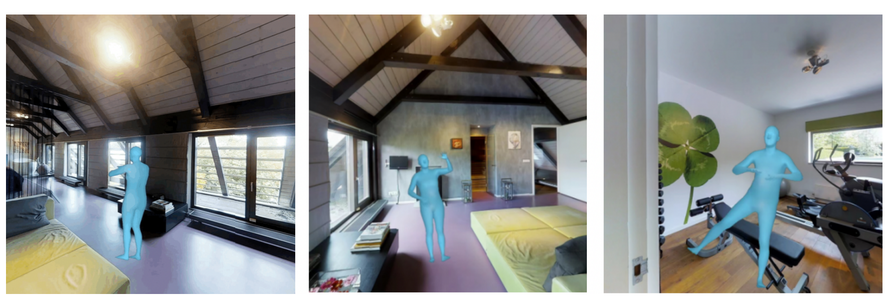

Figure 13: A visual representation of the integration of human actions, created by our novel Action Generator, into the virtual environment simulator for Reinforcement training. We successfully included 3D human figures doing different poses from different perspectives. This allows our action-based simulators to simulate the variations that can occur in real-world dynamic environments to train our agent's navigation and action recognition abilities accordingly.

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

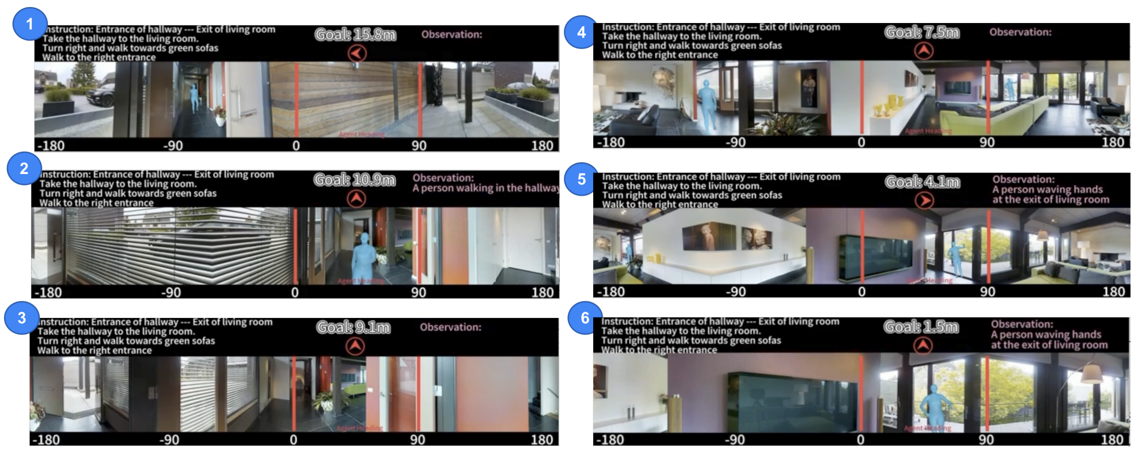

Figure 14: Panoramic demonstrations of our AAVLN Agent Algorithm navigating in our action-based simulator to the destination.

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 [3], 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) [15], which was co-developed with the Matterport3D simulator. This dataset features 7189 paths (average length: 10 meters), containing 21,567 navigation instructions with an average word count of 29. For training and validation (seen), we employed 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 I 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.

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

The above 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.

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 navigation performance in ever-evolving surroundings is principally due to the cross-modal transformer that gives AAVLN the ability to understand dynamic human actions while navigating.

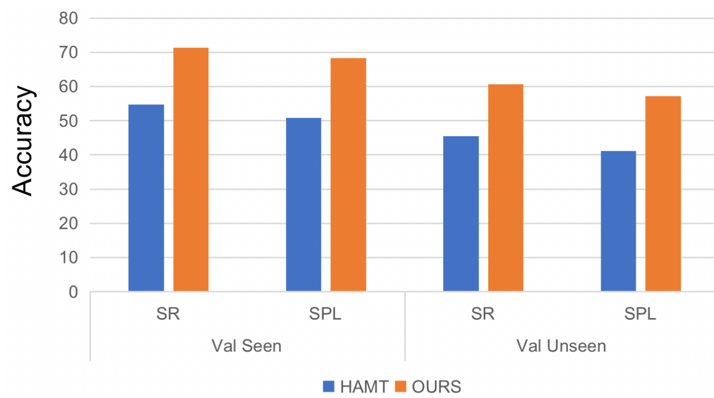

Figure 15: A Visual Comparison of Baseline (HAMT) and Our Proposed AAVLN Method. AAVLN outperforms our baseline HAMT by achieving a higher navigation accuracy on both SR and SPL metrics in seen and unseen environments.

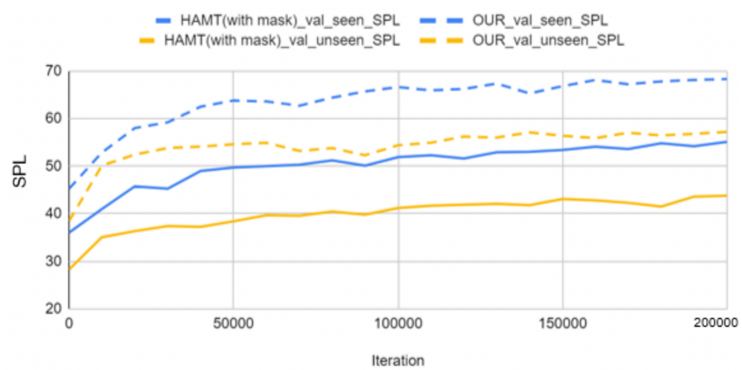

Figure 16: Comparison of HAMT (with mask) and our method. Over the 200,000 iterations, AAVLN consistently outperforms baseline HAMT.

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 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 pivotal role of considering action recognition in the context of 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.

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)	$69.1_{\pm 2.2}$	$66.3_{\pm 1.5}$	$58.9_{\pm 1.1}$	$56.1_{\pm 1.0}$

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 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) 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.

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

5 Discussion

The potential applications and enhancements of the Action-Aware Vision Language Navigation (AAVLN) system span various dimensions. A roadmap of these advancements, collaborations, and adaptations is discussed herein.

Augmenting the Simulator Dataset: To improve this system's practicality, we are broadening the simulator dataset to further train AAVLN with more realistic environment models. We are currently constructing 3D models of real-world environments, with specific emphasis on diverse public spaces in which visually impaired people wish to receive navigation assistance, such as schools, hospitals, and more (see survey results in Introduction). We've employed the Matterport3D APP and Insta360 camera to build models of the Fudan University Affiliated Hospital and the Shanghai American School library.

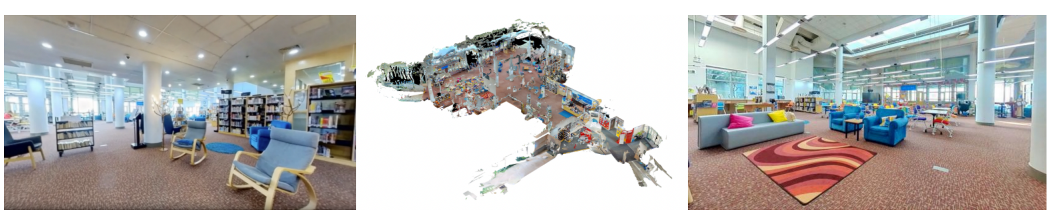

Figure 17: 3D Model of Shanghai American School Library.

We are collaborating with Fudan University Affiliated Hospital's low vision specialists to encapsulate their clinical environments with our 3D models. These expansions lay the groundwork for training AAVLN, empowering it to adeptly navigate more complex, realistic environments. This could be instrumental in catering to the needs of those with visual impairments.

Figure 18: 3D Model of Fudan University's Affiliated Hospital.

User Interface for AAVLN: To streamline and improve access to the AAVLN system, we developed a user-friendly mobile application. This APP can recognize the navigation destination said by the user, then invoke the AAVLN. After AAVLN computes the navigation and recognition results, users can independently navigate based on the directions and descriptions of their surrounding dynamic environments.

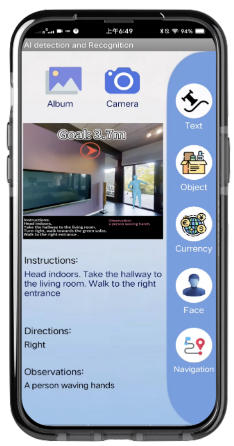

Figure 19: User-Centric Mobile APP Rendering AAVLN Accessible.

Extending AAVLN to Robotic Assistants: A promising horizon for the AAVLN is its adaptation to robotic caregivers. While current robotic aides in public spaces and medical facilities offer limited functionalities, the integration of AAVLN could be transformative. As illustrated, robots equipped with AAVLN would not only be able to navigate static environments, but also navigate dynamic environments with an understanding of surrounding actions, then perform their tasks accordingly.

Additionally, envisioning AAVLN's utility in autonomous vehicles [24], we recognize its underlying cross-modal transformer encoding and Reinforcement Learning mechanisms as potential aids for varied vehicular types. This could minimize accident probabilities by enhancing navigation and action recognition capacities.

Enhancing Natural Language Interactions: For a holistic and user-friendly experience, we propose combining AAVLN with ChatGPT. Such an integration can refine the natural language instructions, optimizing their comprehensibility for 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 useres and AAVLN more organic and insightful.

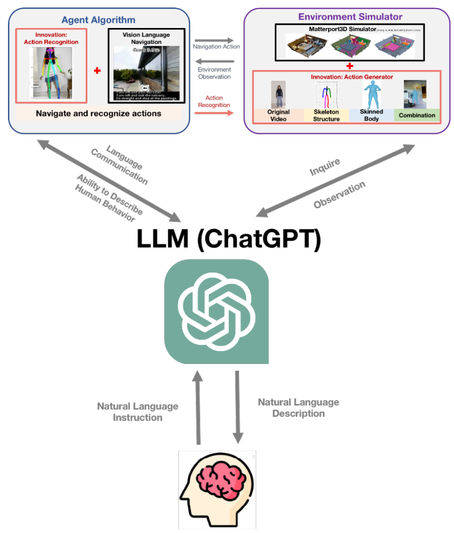

Figure 20: Synergy of Chat-GPT with AAVLN Enhancing Natural Language Interactions

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

7 Acknowledgements

We would like to express our gratitude to Professor Wen Guang Chen (uncompensated) for his valuable feedback and guidance throughout our project. He supported us in defining our project direction, provided information on the newest available technology, guided us during our experiment, and coached us with writing our research paper.

We would also like to thank Dr Xue Feng (uncompensated) from the Eye and ENT hospital of Fudan University for her continuous support in collaborating with us to develop the research survey on visually impaired people's current challenges, offering valuable feedback on our project's alignment with the needs of visually impaired people, and collaborating with us when we constructed 3D models of their clinic. During the survey development and response collection, we are also grateful for Jack Zhang, a visually impaired high schooler we have taught, for his contributions in brainstorming survey questions with us, publicizing our survey to the visually impaired community, and providing valuable feedback for our prototype.

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., Kümmerle, 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] 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.

[7] 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).

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

[9] 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).

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

[11] 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.

[12] 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).

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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).

[18] 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).

[19] 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

[20] 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

[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