



ARL-TR-9906 • APR 2024



# CHRIS-Bot: A Robot for Dialogue and Scene Understanding of Anomalous Environments in Virtual Reality

by Stephanie M Lukin, Jaci South, and Shawn Bowser

DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

## **NOTICES**

### **Disclaimers**

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.



# **CHRIS-Bot: A Robot for Dialogue and Scene Understanding of Anomalous Environments in Virtual Reality**

**by Stephanie M Lukin**  
*DEVCOM Army Research Laboratory*

**Jaci South**  
*Southern Illinois University*

**Shawn Bowser**  
*Army Education Outreach Program*

# REPORT DOCUMENTATION PAGE

*Form Approved*  
**OMB No. 0704-0188**

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

**PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

<b>1. REPORT DATE (DD-MM-YYYY)</b> April 2024		<b>2. REPORT TYPE</b> Technical Report		<b>3. DATES COVERED (From - To)</b> June 14–September 5, 2023	
<b>4. TITLE AND SUBTITLE</b> CHRIS-Bot: A Robot for Dialogue and Scene Understanding of Anomalous Environments in Virtual Reality				<b>5a. CONTRACT NUMBER</b>	
				<b>5b. GRANT NUMBER</b>	
				<b>5c. PROGRAM ELEMENT NUMBER</b>	
<b>6. AUTHOR(S)</b> Stephanie M Lukin, Jaci South, and Shawn Bowser				<b>5d. PROJECT NUMBER</b>	
				<b>5e. TASK NUMBER</b>	
				<b>5f. WORK UNIT NUMBER</b>	
<b>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</b> DEVCOM Army Research Laboratory ATTN: FCDD-RLA-IC 2800 Powder Mill Rd, Adelphi, MD 20783				<b>8. PERFORMING ORGANIZATION REPORT NUMBER</b> ARL-TR-9906	
<b>9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)</b>				<b>10. SPONSOR/MONITOR'S ACRONYM(S)</b>	
				<b>11. SPONSOR/MONITOR'S REPORT NUMBER(S)</b>	
<b>12. DISTRIBUTION/AVAILABILITY STATEMENT</b> DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.					
<b>13. SUPPLEMENTARY NOTES</b> Contact author email: <stephanie.m.lukin.civ@army.mil>					
<b>14. ABSTRACT</b> Robots can play a critical role in supporting human teammates; however, there are many challenges to ensuring effective collaboration under unknown or anomalous conditions. Natural language is a useful method for allowing humans to issue instructions at a high level. However, we further enhance the human-robot dialogue paradigm by increasing the robot's ability to provide common ground for the conversation by performing scene understanding and reporting back on its findings. We offer the following contributions in this report: 1) a human-robot vignette centered around the train derailment in East Palestine, Ohio, USA, in February 2023 that we modeled in a simulated platform; 2) a robot implementation to autonomously navigate this 3-D space as dictated by natural language instructions using sentence embeddings and cosine similarity for the robot's dialogue management; and 3) scene understanding using Vision-Language Models to analyze a visual snapshot of the environment and generate a textual analysis for the human teammate. We conclude with a list of planned tasks to evaluate the models at the algorithmic level as well as for their efficacy in assisting a human in information gathering and disaster relief tasks.					
<b>15. SUBJECT TERMS</b> human-robot interaction, scene understanding, large language models, Military Information Sciences					
<b>16. SECURITY CLASSIFICATION OF:</b>			<b>17. LIMITATION OF ABSTRACT</b> UU	<b>18. NUMBER OF PAGES</b> 28	<b>19a. NAME OF RESPONSIBLE PERSON</b> Stephanie M Lukin
<b>a. REPORT</b> Unclassified	<b>b. ABSTRACT</b> Unclassified	<b>c. THIS PAGE</b> Unclassified			<b>19b. TELEPHONE NUMBER (Include area code)</b> 310-448-5396

## Contents

---

---

<b>List of Figures</b>	<b>iv</b>
<b>List of Tables</b>	<b>iv</b>
<b>1. Introduction</b>	<b>1</b>
<b>2. Background: Ohio Train Derailment</b>	<b>3</b>
<b>3. Comparison to Prior Human-Robot Dialogue Systems</b>	<b>4</b>
<b>4. Robot Design and Vignette</b>	<b>4</b>
<b>5. Implementation</b>	<b>6</b>
5.1 Simulated Robot and Environment	6
5.2 Dialogue Management	7
5.3 Navigation and Manipulation	9
5.4 Scene Understanding	9
5.5 Deployment	10
<b>6. Evaluation</b>	<b>10</b>
6.1 Accuracy of Dialogue Manager	11
6.2 Preliminary Analysis of Scene Understanding	13
<b>7. Conclusions and Future Work</b>	<b>14</b>
<b>8. References</b>	<b>16</b>
<b>List of Symbols, Abbreviations, and Acronyms</b>	<b>21</b>
<b>Distribution List</b>	<b>22</b>

## List of Figures

---

---

Fig. 1	Drone photograph of train derailment from the National Transportation Safety Board (left). Reconstructed train derailment in RIVR (right). . . . .	7
Fig. 2	Visual stimuli 1 . . . . .	10
Fig. 3	Running multiple simulations with RIVR. . . . .	11
Fig. 4	Visual stimuli 2 . . . . .	13

## List of Tables

---

---

Table 1	Subset of natural language instructions mapped to matched instruction intent with reported errors . . . . .	12
---------	---	----

## 1. Introduction

---

Robots can play a critical role in supporting human teammates by gathering information at remote locations that might be too risky, inhospitable, or inaccessible for humans due to hazardous conditions or natural disasters.<sup>1,2</sup> However, these scenarios are challenging, if not impossible, to safely gather data for and train a robot in the first place without risking system breakdown or subjecting humans in the field to danger. These robots have historically placed the burden of navigation on a human operator who must teleoperate the robot using a handheld device<sup>3-5</sup> or through a point and click interface that supports autonomous path planning,<sup>6-8</sup> which may disrupt the human operator's focus on the bigger picture. Unreliable network connectivity can further limit the human's ability to visually assess the remote location if video streaming is unavailable. At the other extreme, a human's attention may be divided if the human is using multiple robots or sifting through hours of mostly unchanging video feed, where they may be unable to isolate key events within multiple streams of incoming data.

Within this larger problem space, we identify and tackle three challenges to develop human-robot teaming systems for unreliable and dangerous scenarios: 1) the need for alternate methods of robot control to allow human operators to focus on higher-level planning and reasoning rather than lower-level robot control; 2) the need for a robot to visually understand a scene and narrate, highlight, and extrapolate events in a textual form for a human to easily digest among the myriad or lack of incoming visual data; and 3) the need for a safe environment to develop, test, and evaluate these technologies for deployment in the field.

In previous studies, researchers have explored diverse ways to control robots and found that language in particular has proven to be a natural way to communicate with a robot that is more akin to working with a teammate, rather than a tool.<sup>9-12</sup> Working within the human-robot dialogue problem space, we experiment with a new approach for computing matches for a dialogue management classifier that has the potential to significantly reduce the data collection and training efforts required in prior studies due to the use of off-the-shelf, large language models (LLMs).

When using language to communicate with a robot in the physical world, that language becomes situated in that the human or robot refers to locations, descriptions, or properties about what is seen in the environment. When the human and robot can

share visual information, whether through a live view of the robot’s point of view or by standing side-by-side, overcoming miscommunications in dialogue, such as “which door?”, can be resolved by further references to the environment, for example, “the one on the left.” Prior work in remote human-robot and human-human navigation tasks always involves a real-time sharing of visuals.<sup>13–19</sup> However, in the complete absence or oversaturation of live-feed, alternative options must be available. Some interaction paradigms allow humans the option to take photos on demand to see into the visual as a selective point of reference<sup>20</sup>; however, there may be circumstances in which sending images may be too expensive. We focus on this novel interactive mode and how a robot may support sharing visual common ground by performing scene analysis and using narrative intelligence to generate text pertaining to the robot’s environment based on a particular purpose, goal, or audience.

The capabilities we develop for the human–robot team must be repeatedly tested within ever-changing environments to allow the adaptability of the robot to be evaluated by a human operator. To this end, many roboticists use Sim2Real (i.e., simulated to real world) paradigms in which a simulated construct of a real-world environment is created and programmed to mimic real-world properties, including physics and robot capabilities. In recent work, researchers have sought to make these testbeds scalable, and we extend one such framework called RIVR (Robot Interaction in Virtual Reality) to highlight the flexibility in programming our robot features.<sup>21</sup>

A team of two interns hosted by the US Army Combat Capabilities Development Command Army Research Laboratory and recruited through the National Security Innovation Network X-Force Fellowship spent 10 weeks exploring this problem space. Over the course of the work period, we designed a robot with dialogue and visual scene understanding and explanation. We implemented these capabilities within a scalable and reusable simulation and virtual reality platform for testing robots and conducting experimentation. We focused on the train derailment disaster that occurred in East Palestine, Ohio, USA, in February 2023 as a study of anomalous and unexpected scenarios. We created CHRIS-Bot, the Chemistry Human-Robot Interactive System Robot, to navigate through the train derailment as recreated in simulation, and we designed a vignette where CHRIS-Bot can investigate the scene, explain what it sees and elements that appear anomalous or

unsafe, and navigate through it following human instruction to take air, soil, and water samples.

We offer the following contributions in this report:

1. A human-robot collaboration vignette exhibiting a train derailment modeled in Unity in the RIVR experimental platform.
2. Autonomous robot navigation of this 3-D space as dictated by human natural language instructions, using sentence embeddings and cosine similarity for dialogue management.
3. Scene understanding using Vision-Language Models (VLMs) in the context of a human-robot dialogue to analyze a visual snapshot of the environment and generate a report for the human.

We enumerate planned tasks in our ongoing and future work to evaluate the models at the algorithmic level and for their efficacy in assisting a human in particular tasks and domains. We conclude with the ways in which this framework can extend beyond the capabilities described in this report.

## **2. Background: Ohio Train Derailment**

---

---

On February 3, 2023, in East Palestine, Ohio, USA, 38 train cars derailed including 11 tank cars carrying hazardous materials of vinyl chloride, butoxyethanol, ethylhexyl acrylate, isobutylene, butyl acrylate, and benzene.<sup>22</sup> Over 100,000 gallons of hazardous materials were spilled, and several train cars burned for more than 2 days.<sup>23</sup> Clean-up efforts were immediate, and over 70 emergency agencies were mobilized. Since a number of toxic chemicals were spilled, clean-up efforts included testing for the various chemicals, including air,<sup>24</sup> soil,<sup>25</sup> and water sample tests.<sup>26</sup> A controlled burn of remaining chemicals was ordered to prevent further explosions.<sup>27</sup> We study this incident so that we may envision how robots can be a useful teammate for first responders in their relief efforts.

### **3. Comparison to Prior Human-Robot Dialogue Systems**

---

CHRIS-Bot extends prior work on human-robot interaction through dialogue for collaborative tasks; in particular, the ScoutBot system that provided a human operator the ability to speak unconstrained and natural language to a simulated robot that can execute simple navigation instructions.<sup>28</sup> In ScoutBot, the robot’s labor was divided into dialogue management and navigation components to create specialized modules for each type of processing. The dialogue management component used a retrieval-based classification algorithm that created a map of human instruction and robot-response pairs and compared new instructions to the collection of instructions. Kullback-Leibler divergence was used to calculate a similarity score between the new instruction and the instructions in the training data, and the ScoutBot system returned the highest matching robot response.<sup>29</sup> CHRIS-Bot is implemented with two separate dialogue management and navigation components, following the ScoutBot paradigm, and similarly uses a retrieval-based classification algorithm. Instead of Kullback-Leiber, which compares two raw strings, we first embed the strings using sentence embeddings and use cosine similarity to compare the vectors. Due to the recent advancements of these vector embeddings, we posit that they will make a favorable comparison.

CHRIS-Bot introduces the ability for a robot to exhibit narrative intelligence through scene understanding. To our knowledge, robot systems have yet to be integrated with grounded scene description capabilities in the interactive way we described in our design and vignette. This research direction has been proposed in prior forward-looking work on the way humans and robots interact,<sup>30</sup> and recent approaches emphasizing LLM narrative intelligence combined with multi-modal input can further refine inferences, bolstering autonomy to better align with human objectives.<sup>31–34</sup>

### **4. Robot Design and Vignette**

---

We designed CHRIS-Bot to support real-time remote collaboration in instances such as the train derailment. Under these circumstances, streaming of video or real-time teleoperation may be infeasible and thus require a dialogue between the human and robot to establish shared understanding. By enabling the robot to understand what it sees and to convey that knowledge to the human, the human may be better able to assess the situation and issue instructions to the robot. CHRIS-Bot will perform primitive maneuvers, that is, driving forward, backward, and rotating in

place. High-level capabilities can be created through sensing and avoiding obstacles and navigation to specific locations. CHRIS-Bot will take images and record and stream video from the RGB-D (red-green-blue depth) camera mounted on the chassis. From these visual media, it will analyze the visual space and generate natural language descriptions, narratives, and answer the human's questions.

CHRIS-Bot will be controlled and interacted with via natural language instructions from a human partner as opposed to other methods of robot-directed guidance, such as teleoperation via a joystick or a point-and-click computer interface. In this way, the human's hands are not encumbered by a joystick, which allows them to treat CHRIS-Bot as a teammate by issuing it instructions it can complete autonomously. The level of specificity of instructions is controlled by a dialogue manager, where, for example, prior work has required a instructions to contain end point to support network latency.<sup>29</sup> If no end point is given, the robot requests the information and will not complete the instruction until it is supplied or a new instruction is issued.

To situate CHRIS-Bot within the train derailment disaster, we plan for several capabilities that require its robotic arms to take samples in hazardous environments. These domain-specific actions include the following:

- Taking samples from soil and water using a test tube
- Holding a mobile air sample machine
- Performing field tests on samples
- Returning to command post when test tube rack is full

We outline a vignette in which we assume the human cannot see what the robot sees, yet they must assess the environment and instruct the robot. Below is an aspirational dialogue written for illustrative purposes:

**Human:** "Tell me about what you're looking at now."

**Robot:** "I'm approximately 20 feet from the majority of the overturned train cars. A liquid substance is spilling from the cargo. There is a river or stream fairly close to the cars."

**Human:** "Let's start by taking a soil sample."

**Robot:** “Where?”

**Human:** “Down by the river”

<Robot will locate the river it previously identified, drive to it, remove a test tube from the rack, scoop soil, cap the test tube, and place it back in the rack>

**Robot:** “Done. What should I do next?”

This vignette showcases the robot’s ability to understand instructions, conduct computer vision analysis, display scene understanding, generate natural language, ask for clarification of under-specified instructions, and perform high-level navigation and maneuver specific to the target domain.

We discuss our progress towards implementing a robot according to this design in Section 5. In Section 6, we evaluate two of these components.

## **5. Implementation**

---

---

Our experimental testbed requires a robot, an environment, language understanding and dialogue management, navigation and manipulation operations, and computer vision and scene understanding. To implement our planned specifications, we leverage existing infrastructure and techniques, and we enhance them with novel capabilities. These components and infrastructure are described in Sections 5.1–5.5.

### **5.1 Simulated Robot and Environment**

---

For deploying the simulated robot in a simulated environment, we use the RIVR platform.<sup>35,36</sup> RIVR models environments in Unity and supports robot implementation and integration with the Robot Operating System (ROS). Elements in the environment can be modeled to test the robot’s ability to perform in particular conditions, including fires and blockades (where areas become inaccessible or blocked to the robot), spills (where spilled chemicals need to be contained or cleaned up), smoke (where the robots sensors may be occluded or uncertain), time-sensitive elements (where a sense of urgency may be present), and unfolding information (where replanning is required as new information is reported from the field or command base stations).

We implement CHRIS-Bot as a Clearpath Husky, a durable ground vehicle, equipped with a Kinova Kortex arm and a built-in RGB-D camera. The train derailment environment was crafted in Unity using free assets by closely referencing real pho-

tographs of the incident that featured the train cars, rails, and smoke. The left panel in Fig. 1 shows a drone photograph taken by the National Transportation Safety Board,<sup>37</sup> and the right panel shows our recreation from a similar angle.



**Fig. 1 Drone photograph of train derailment from the National Transportation Safety Board (left). Reconstructed train derailment in RIVR (right).**

## 5.2 Dialogue Management

---

Our dialogue management policies reflect our desire for the human operator to speak unconstrained natural language to a robot that will interpret and execute navigation and exploration commands. The dialogue manager serves two functions: it ensures that the human’s instructions are well formed, and it sends a signal to the robotic processes required for completing the instructions. In this way, it is modular, rather than end-to-end, so that different dialogue management methodologies can be implemented, inspected, and enhanced independent of the target robot’s capabilities.

CHRIS-Bot’s dialogue management is implemented using retrieval-based classification. Before real-time interactions, a dataset of in-domain human instruction-robot response pairs  $(H_i, R_i)$  is curated:  $H_i$  is the  $i$ -th human instruction in the dataset, and  $R_i$  is the matching robot response to  $H_i$ . For example, the data pair (“Take a picture,” “image”) indicates that the robot’s response to the human’s utterance “Take a picture” is to use its image-taking capabilities. The robot response in the data pair (“Move forward,” “Please tell me how far to move forward”) indicates that the robot does not have enough information to execute the action from the human’s instruction “Move forward,” and it will request additional information.

Each human instruction within the dataset is embedded as a vector  $H'_i$  using Sentence-BERT. Then, when given an input sentence  $S$ , CHRIS-Bot embeds the sentence  $S'$  and performs a pairwise cosine similarity between the utterance and the set of pre-processed instructions. The cosine angle of the two vectors represents the closeness of the two embedded sentences within a multidimensional space. It is a value between 0 and 1, where 1 indicates an exact match between the sentence vectors. For the  $H'_i$  with the highest score to  $S'$ , the response  $R_i$  is selected as the most appropriate response to  $S$ . The transformations and comparisons are shown in Eq. 1.

$$\begin{aligned}
 H'_i &= SBERT(H_i) \quad \forall \quad 0 < i < |H| \\
 S' &= SBERT(S) \\
 \cos(S', H'_i) &\quad \forall \quad 0 < i < |H'|
 \end{aligned} \tag{1}$$

In future iterations, a threshold can be implemented on the cosine score to avoid returning a poorly matched response with a response “I do not understand” as well as an n-best list on which additional analysis can be performed to determine the best response. After a robot response has been matched, it is sent to the navigation and manipulation or scene understanding for execution.

Below is an example of how a user can interact with our dialogue manager in real-time. The human instructions were typed into a command line interface by the authors of this report, and the system responses are what CHRIS-Bot returned.

**Human:** rotate around  
**System:** I will turn 180  
**Human:** can you move forward  
**System:** I’m not sure where or when to stop moving forward.  
**System:** Please supply clarification.  
**Human:** one foot  
**System:** I will move forward 1 foot  
**System:** Please supply next instruction.

If a human instruction is determined to require a follow-up clarification, for example, the instruction “can you move forward” is missing an end point, the system will then notify the user of the problem, for example, “I’m not sure where or when to

stop moving forward,” and provide a prompt for the user to enter additional information. Once this information is provided, it is concatenated with the user’s original instruction and sent again to the dialogue manager for matching, and the dialogue continues.

### 5.3 Navigation and Manipulation

---

If the instruction was one of navigation or manipulation, for example, “move forward 3 feet,” the dialogue manager transformed it into a ROS command for execution. To implement movement, a continuous stream of ROS messages was published to the robot’s velocity topic to move a linear or angular vector for a specific duration that mapped approximately to feet and degrees; no direct mapping is present in ROS. For example, a linear vector of (0.3, 0, 0) for 30 iterations within the robot’s publishing loop would move the robot 3 feet forward, whereas (-0.2, 0, 0) for 20 iterations would move the robot backwards approximately 2 feet. An angular vector of (0, 0, 0.45) for 45 iterations would rotate the robot approximately 45 degrees to the left. This approach uses the same ROS messages implemented in ScoutBot.<sup>28</sup> While we had planned for CHRIS-Bot to manipulate its arm, we did not have enough time during the work period to explore this.

### 5.4 Scene Understanding

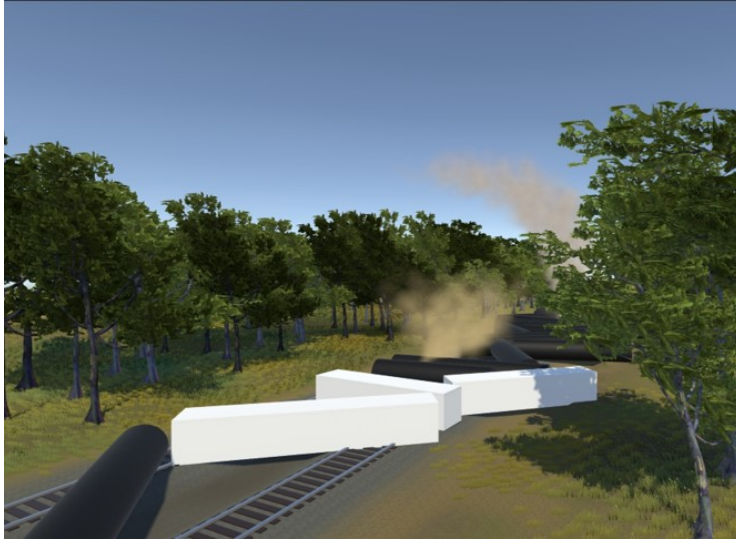
---

If the instruction was for scene understanding, for example, “tell me what you see,” the dialogue manager issued a ROS instruction to take a picture by saving the robot’s current RGB frame through the ROS image message topic. One such image is shown in Fig. 2.

The saved image was then passed to a VLM with a prompt requesting the system to explain the scene. We piloted several systems and prompts, which we discuss more in Section 6. Here, we show the input prompt given to LLaVA (Large Language and Vision Assistant),<sup>38</sup> a VLM capable of multi-modal chat capabilities, and its output:

**Prompt:** What do you see in this image?

**LLaVA:** *“The image features a train traveling down the tracks, surrounded by trees and a forest. The train is located near the center of the scene, with its engine and several cars visible. The tracks appear to be on a slight curve, adding a sense of depth to the scene. In addition to the train, there are several people dispersed throughout the image, likely observing the train or enjoying the natural surround-*



**Fig. 2 Visual stimuli 1**

*ings. Some of them are closer to the train, while others are further away, near the edge of the forest.”*

Although most of the description is accurate—there are a train, tracks, and a forest—other elements of the description are not. For instance, and most prominently, the train is no longer traveling down the tracks. In Section 6, we describe our pilot testing and preliminary analysis.

## **5.5 Deployment**

---

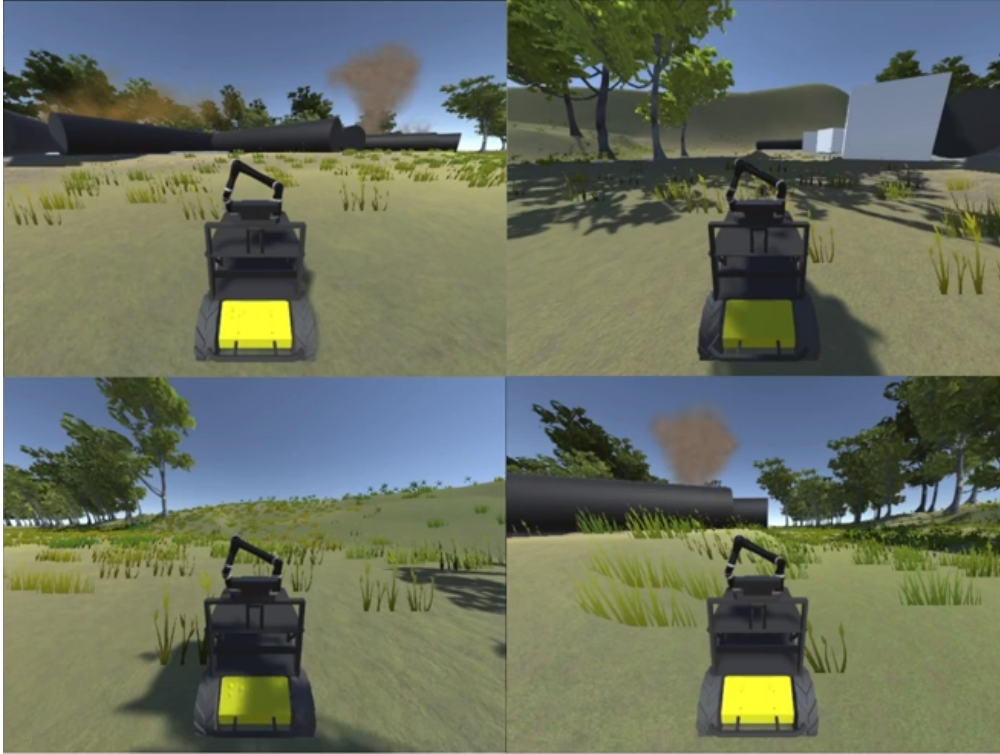
We developed a docker workflow to significantly reduce the time to setup and run the simulation by easily launching multiple simulations to run in parallel without the need for human intervention. Figure 3 shows four simulations running in parallel. The viewpoint is a fixed camera following the robot. This may be the view humans see if they are controlling many robots at once, or this view may represent a behind-the-scenes snapshot that compiles visual information from human experiments.

## **6. Evaluation**

---

---

We conducted a thorough evaluation of the dialogue management component of CHRIS-Bot and a preliminary examination into the scene understanding component. The evaluations are described in Sections 6.1 and 6.2.



**Fig. 3 Running multiple simulations with RIVR**

## **6.1 Accuracy of Dialogue Manager**

---

Our dialogue management classifier was trained and tested on the same dataset of human-robot dialogue utterances<sup>39</sup> used by Gervits et al.<sup>29</sup> in their evaluation of ScoutBot, which used Kullback Leibler Divergence. The training set consisted of over 900 pairs of human utterances and robot responses, and the test set consisted of 45 unseen pairs.

We also adapt Gervits et al.’s<sup>29</sup> leniency evaluation by annotating the selected responses that are not a perfect match to the ground truth robot response. We annotate for two categories: *Felicitous Responses*—in which the classifier’s selected response is appropriate in the larger context of the dialogue and would have the same effect as the correct response—and *Approximate Responses*—in which the classifier’s selected response differed only slightly from the correct one (e.g., variation in turn radius or movement distance).

Our system achieved a 95% accuracy when accounting for these leniency categories. Gervits et al.<sup>29</sup> (p. 10) do not report a leniency score, but they note that

“over half of the incorrect responses in the test set were either Felicitous or Approximate to the correct response.” Using a strict accuracy, we achieve 62% compared to Gervits et al.’s<sup>29</sup> 75%. Table 1 shows a subset of instructions from the test set in which the selected response is an exact match to the ground truth (“Correct”), the selected response is considered correct if using the lenient methods (“Felicitous” or “Approximate”), or the selected response is wrong (“Incorrect”).

**Table 1 Subset of natural language instructions mapped to matched instruction intent with reported errors**

Test Set Input	Selected Response	Ground Truth	Accuracy
robot move three feet past the doorway	move forward 3 feet	move forward 3 feet	Correct
robot turn a hundred eighty degrees	turn 180	turn 180	Correct
turn north	turn to face North	turn to face North	Correct
mkay turn left forty five degrees	turn left 45 degrees	turn left 45	Felicitous
send picture	send image	image	Felicitous
turn left	You can tell me to turn a number of degrees or to face something.	How far left should I turn?	Felicitous
turn left one hundred degrees	turn left 90 degrees	turn left 100	Approximate
robot face southwest	turn to face West	turn to face South-west	Incorrect

Despite the lower strict accuracy of CHRIS-Bot compared to ScoutBot, when looking at the individual examples, it is clear that responses such as “send image” and “image” achieve the same intent within the particular domain. Thereby, we are encouraged by the 95% leniency accuracy.

We plan to test the performance by running the classifier on other human-robot and human-human navigation datasets, including CReST,<sup>15</sup> ArtWalk,<sup>13</sup> SCARE,<sup>17</sup> HuRDL,<sup>19</sup> and others, and comparing the scores against existing benchmarks.

## 6.2 Preliminary Analysis of Scene Understanding

---

In Section 5, we noted that although the LLaVA model accurately identified a train, tracks, and a forest, other elements of the description were incorrect—the most obvious being that the train was not on the tracks. LLaVA hallucinates several elements in the scene that are verifiable by conferring with the source image. First, it is unclear whether the tracks are “on a slight curve,” and the fact that it would “add a sense of depth to the scene” is irrelevant. Next, there are no people visible in the scene whatsoever, despite its claims that there are “several people dispersed throughout the image.” Finally, the model attributes false or unknowable motivations and emotions; if people really were present, it is reasonable that they would be “observing the train.” However, given the nature of the event, it seems unreasonable that they would be “enjoying the natural surroundings” any more given the incident, which could reasonably be inferred to have recently occurred given the dark smoke actively rising from the cars.

Next, we tested a scene understanding prompt with instructions meant to set the stage for the system in generating its responses. Here, we add to the prompt an indicator of “who” the system should be responding as, that is, an autonomous robot, and the task to provide context for the generation, that is, investigation of a train derailment. We test Google’s Gemini VLM\* with the following prompt for Fig. 4:



**Fig. 4 Visual stimuli 2**

---

\*<https://gemini.google.com/app>

**Prompt:** You are an autonomous robot that has been tasked with investigating the scene of a train derailment. Describe what you see and make a decision about where you will go next.

**Gemini:** *“I approach the scene of the derailment and see that several train cars have been overturned. There is a large amount of smoke and debris in the area. I can also smell a strong chemical odor. I use my sensors to scan the area and identify several hazardous materials.”*

Gemini hallucinates with this prompt as well; in this case, it creates sensors for the robot. CHRIS-Bot does not have a sense of smell, so the predicted action is not viable. In future work, we will continue to test different VLM’s and undergo extensive prompt design and engineering to establish our own benchmarks and measures of success with respect to the scene understanding and hallucinations to train these powerful generative models to describe and discuss the scenes we are providing.

## **7. Conclusions and Future Work**

---

---

In this report, we present a design and implementation for a robotic agent, CHRIS-Bot, that is capable of 1) movement through a simulated environment following natural language instructions from human input and 2) scene analysis of the visuals. These capabilities brought together by the RIVR testbed enable us to pilot the technologies and establish a future testing framework.

To better align with prior human-human and human-robot collaboration paradigms, we further plan to allow for the user of spoken language to emulate more natural communication; time did not allow for this during the work period. Speech-to-text performance will further be incorporated into the evaluation scope. Additionally, we plan to measure human task performance with different parameters of the system in complex environments and conditions. To facilitate the creation of these environments, we will explore the use of synthetic scene generation, rather than hand-crafting scenes with Unity. One such potential technology is neural radiance fields (NeRFs), which have produced photorealistic renderings of complex scenes, capturing reflections, refractions, and other view-dependent effects not possible with photogrammetry.<sup>40,41</sup> NeRFs use a fully connected deep neural network to represent static scenes, typically using 50–150 input images of the scene taken from different viewpoints. Although training a single NeRF is computationally expensive, methods using image-conditioned diffusion models can be used to edit NeRF scenes us-

ing text-instructions.<sup>42</sup> This presents opportunities for quickly generating variations of a single 3-D scene to introduce different combinations of adverse conditions and test how the human-robot team can best address the situation, for example, smoke and fire, smoke no fire, and no smoke and fire. Finally, we plan to conduct human subjects evaluation with the robot teammate to measure how the performance of search and navigation tasks is affected by the automated capabilities.

## 8. References

---

1. Nagatani K, Kiribayashi S, Okada Y, Otake K, Yoshida K, Tadokoro S, Nishimura T, Yoshida T, Koyanagi E, Fukushima M, et al. Emergency response to the nuclear accident at the Fukushima Daiichi nuclear power plants using mobile rescue robots. *Journal of Field Robotics*. 2013;30(1):44–63.
2. Kanazawa K, Sato N, Morita Y. Considerations on interaction with manipulator in virtual reality teleoperation interface for rescue robots. *International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2023;32:386–391.
3. Kang S, Cho C, Lee J, Ryu D, Park C, Shin KC, Kim M. ROBHAZ-DT2: design and integration of passive double tracked mobile manipulator system for explosive ordnance disposal. *International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2003;3:2624–2629.
4. Ryu D, Kang S, Kim M, Song JB. Multi-modal user interface for teleoperation of ROBHAZ-DT2 field robot system. *International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2004;1:168–173.
5. Yamauchi BM. Packbot: a versatile platform for military robotics. *Unmanned Ground Vehicle Technology VI*. SPIE. 2004;5422:228–237.
6. Williams SB, Pizarro OR, Jakuba MV, Johnson CR, Barrett NS, Babcock RC, Kendrick GA, Steinberg PD, Heyward AJ, Doherty PJ. Monitoring of benthic reference sites: using an autonomous underwater vehicle. *IEEE Robotics & Automation Magazine*. 2012;19(1):73–84.
7. Arvidson RE, Bell JF III, Bellutta P, Cabrol NA, Catalano J, Cohen J, Crumpler LS, Des Marais D, Estlin T, Farrand W. Spirit Mars rover mission: overview and selected results from the northern Home Plate winter haven to the side of Scamander crater. *Journal of Geophysical Research: Planets*. 2010;115(E7).
8. Camilli R, Reddy CM, Yoerger DR, Van Mooy BA, Jakuba MV, Kinsey JC, McIntyre CP, Sylva SP, Maloney JV. Tracking hydrocarbon plume transport and biodegradation at deepwater horizon. *Science*. 2010;330(6001):201–204.
9. Mavridis N. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*. 2015;63:22–35.
10. Marge M, Gervits F, Briggs G, Scheutz M, Roque A. Let’s do that first! A comparative analysis of instruction-giving in human-human and human-robot

- situated dialogue. Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue (SemDial); 2020 July 18–19; Virtual.
11. Marge M, Espy-Wilson C, Ward NG, Alwan A, Artzi Y, Bansal M, Blankenship G, Chai J, Daumé H III, Dey D, et al. Spoken language interaction with robots: recommendations for future research. *Computer Speech & Language*; 2022;71(101255):1–13.
  12. Ahn M, Brohan A, Brown N, Chebotar Y, Cortes O, David B, Finn C, Fu C, Gopalakrishnan K, Hausman, et al. Do as I can, not as I say: grounding language in robotic affordances. *arXiv:2204.01691*; 2022.
  13. Liu K, Tree JEF, Walker M. Coordinating communication in the wild: the artwalk dialogue corpus of pedestrian navigation and mobile referential communication. *Language Resources and Evaluation Conference (LREC)*; 2016 May 23–28; Portorož, Slovenia.
  14. De Vries H, Shuster K, Batra D, Parikh D, Weston J, Kiela D. Talk the walk: navigating New York City through grounded dialogue. *arXiv:1807.03367*; 2018.
  15. Eberhard KM, Nicholson H, Kübler S, Gundersen S, Scheutz M. The Indiana “Cooperative Remote Search Task” (CReST) corpus. *Language Resources and Evaluation Conference (LREC)*; 2010 May 17–23; Valletta, Malta.
  16. Padmakumar A, Thomason J, Shrivastava A, Lange P, Narayan-Chen A, Gella S, Piramuthu R, Tur G, Hakkani-Tur D. TEACH: task-driven embodied agents that chat. *AAAI Conference on Artificial Intelligence*. 2022;36(2):2017–2025.
  17. Stoia L, Shockley DM, Byron DK, Fosler-Lussier E. SCARE: a situated corpus with annotated referring expressions. *Language Resources and Evaluation Conference (LREC)*; 2008 May 26–June 1; Marrakech, Morocco.
  18. Suhr A, Yan C, Schluger J, Yu S, Khader H, Mouallem M, Zhang I, Artzi Y. Executing instructions in situated collaborative interactions. *Conference of Empirical Methods for Natural Language Processing (EMNLP) and the International Joint Conference on Natural Language Processing (IJCNLP)*; 2019 3–7 Nov; Hong Kong, China.
  19. Gervits F, Roque A, Briggs G, Scheutz M, Marge M. How should agents ask questions for situated learning? An annotated dialogue corpus. *Special Interest Group on Discourse and Dialogue (SIGDIAL)*; 2021:353–359.
  20. Lukin SM, Pollard KA, Bonial C, Hudson T, Artstein R, Voss C, Traum D. Navigating to success in multi-modal human-robot collaboration: corpus and

- analysis. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN); 2023 Aug 28–31; Busan, Korea.
21. Higgins P, Barron R, Engel D, Lukin S, Matuszek C. A collaborative building task in VR vs. reality. International Symposium on Experimental Robotics (ISER); 2023 26–30 Nov; Chiang Mai, Thailand.
  22. Yan H, Maxouris C, Brown N. The Ohio toxic train wreck was ‘100% preventable’ – but there’s no evidence the crew did anything wrong, investigators say. [accessed 2023 Oct 13]. <https://www.cnn.com/2023/02/23/us/ohio-train-derailment-east-palestine-thursday/index.html>.
  23. Elamroussi A. A first report on the Ohio toxic train wreck was released. Here’s what it found — and what investigators are still looking into. [accessed 2023 Oct 13]. <https://www.cnn.com/2023/02/24/us/ohio-train-derailment-east-palestine-friday/index.html>.
  24. Environmental Protection Agency. East Palestine, Ohio train derailment, air sampling data. [accessed 2023 Oct 13]. <https://www.epa.gov/east-palestine-oh-train-derailment/air-sampling-data>.
  25. Environmental Protection Agency. East Palestine, Ohio train derailment, soil and sediment sampling data. [accessed 2023 Oct 13]. <https://www.epa.gov/east-palestine-oh-train-derailment/soil-and-sediment-sampling-data>.
  26. Ohio Environmental Protection Agency. East Palestine public drinking water testing results. [accessed 2023 Oct 13]. <https://epa.ohio.gov/monitor-pollution/pollution-issues/east-palestine>.
  27. Singh K, Gorman S. Derailed train cars in Ohio drained of toxic chemical amid mass evacuation. [accessed 2023 Oct 13]. <https://www.reuters.com/world/us/ohio-carry-out-controlled-release-chemicals-train-derailment-site-2023-02-06/>.
  28. Lukin SM, Gervits F, Hayes CJ, Leuski A, Moolchandani P, Rogers JG III, Amaro CS, Marge M, Voss CR, Traum D. ScoutBot: a dialogue system for collaborative navigation. Association for Computational Linguistics (ACL). 2018; 93–98.
  29. Gervits F, Leuski A, Bonial C, Gordon C, Traum D. A classification-based approach to automating human-robot dialogue. Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems; 2019 Apr 24–26; Sicily, Italy.

30. Lukin SM, Bonial C, Voss CR. Visual understanding and narration: a deeper understanding and explanation of visual scenes. arXiv:1906.00038; 2019.
31. Halperin BA, Lukin SM. Envisioning narrative intelligence: a creative visual storytelling anthology. CHI Conference on Human Factors in Computing Systems; 2023 Apr 23–28; Hamburg, Germany.
32. Malakan ZM, Anwar S, Hassan GM, Mian A. Sequential vision to language as story: a storytelling dataset and benchmarking. IEEE Access. 2023;11;70805–70818.
33. Yang D, Jin Q. Attractive storyteller: stylized visual storytelling with unpaired text. Association for Computational Linguistics; 2023 July 9–14; Toronto, Canada. p. 11053–11066.
34. Lukin SM, Eum S. SEE&TELL: controllable narrative generation from images. Proceedings of the AAAI Creative AI Across Modalities Workshop; 2023 Feb 13; Washington DC.
35. Higgins P, Kebe GY, Berlier A, Darvish K, Engel D, Ferraro F, Matuszek C. Towards making virtual human-robot interaction a reality. 3rd International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions; 2021 Mar 8; Boulder, CO.
36. Higgins P, Barron R, Matuszek C. Head pose for object deixis in VR-based human-robot interaction. IEEE International Conference on Robot and Human Interactive Communication; 2022 29 Aug–2 Sep; Napoli, Italy.
37. National Transportation Safety Board. (2023, Oct 13) NTSB B-Roll: train derailment in East Palestine, OH. [Video]. Youtube. <https://www.youtube.com/watch?v=7AyXTVkvBT4>.
38. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. arXiv:2304.08485; 2023.
39. Lukin SM, Bonial CN, Marge M, Hudson T, Hayes CJ, Pollard, KA, Bakerm A, Fouts A, Arstein R, Gervits F, et al. SCOUT: a situated and multi-modal human-robot dialogue corpus. The Joint International Conference on Computational Linguistics and Language Resources and Evaluation (COLING-LREC); Forthcoming 2024.
40. Cohen-Bar D, Richardson E, Metzger G, Giryes R, Cohen-Or D. Set-the-scene: global-local training for generating controllable NeRF scenes. arXiv:2303.13450; 2023.

41. Metzger G, Richardson E, Patashnik O, Giryes R, Cohen-Or D. Latent-NeRF for shape-guided generation of 3D shapes and textures. Conference on Computer Vision and Pattern Recognition (CVPR); 2023:12663–12673.
42. Haque A, Tancik M, Efros AA, Holynski A, Kanazawa A. Instruct-NeRF2NeRF: editing 3D scenes with instructions. arXiv:2303.12789; 2023.

## List of Symbols, Abbreviations, and Acronyms

---

3-D	three-dimensional
BERT	Bidirectional Encoder Representations from Transformers
CHRIS-Bot	Chemistry Human–Robot Interaction System Robot
LLaVa	Large Language and Vision Assistant
LLM	large language model
NeRF	neural radiance field
RGB-D	red-green-blue depth
RIVR	Robot Interaction in Virtual Reality
ROS	Robot Operating System
SBERT	Sentence BERT
Sim2Real	simulated to real world
VLM	Vision-Language Model

1 DEFENSE TECHNICAL  
(PDF) INFORMATION CTR  
DTIC OCA

1 DEVCOM ARL  
(PDF) FCDD-RLB-CI  
TECH LIB

1 DEVCOM ARL  
(PDF) FCDD-RLA-IC  
SM LUKIN

1 SIU  
(PDF) J SOUTH

1 AEOP  
(PDF) S BOWSER