Grounded Instruction Understanding with Large Language Models: Toward Trustworthy Human-Robot Interaction

Ekele Ogbadu[1], **Stephanie Lukin**[2], **Cynthia Matuszek**[1]
[1]University of Maryland, Baltimore County
1000 Hilltop Circle,
Baltimore, MD USA
[2]DEVCOM Army Research Laboratory
Playa Vista, CA USA
eogbadu1@umbc.edu, stephanie.m.lukin.civ@army.mil, cmat@umbc.edu

Abstract

Understanding natural language as a representational bridge between perception and action is critical for deploying autonomous robots in complex, high-risk environments. This work investigates how large language models (LLMs) can support this bridge by interpreting unconstrained human instructions in urban disaster response scenarios. Leveraging the SCOUT corpus, a multimodal dataset capturing human-robot dialogue through Wizard-of-Oz experiments, we construct SCOUT++, aligning over 11,000 visual frames with language commands and robot actions. We evaluate three instruction classification approaches: a neural network trained on tokenized text, GPT-4 using text alone, and GPT-4 with synchronized visual input. Results show that while GPT-4 (text-only) outperforms traditional models in accuracy, its multimodal variant exhibits degraded performance, often producing vague or hallucinated outputs. These findings expose the challenges of reliably grounding language in visual context and raise questions about the trustworthiness of foundation models in safety-critical settings. We contribute SCOUT++, a reproducible multimodal pipeline, and benchmark results that shed light on the capabilities and current limitations of vision-language models for risk-sensitive human-robot interaction.

Introduction

Natural language serves not only as a communication tool but as a powerful representational layer bridging perception, reasoning, and action in autonomous systems. For robots operating in complex, real-world environments, understanding unconstrained instructions demands interpreting linguistic semantics while integrating visual and spatial context. A seemingly simple command like "Go to the center of the red room" requires recognizing the scene visually and reasoning spatially to identify the target location. While such tasks are effortless for humans, they remain challenging for robots, particularly in high-stakes domains like urban disaster response, where situations are dynamic and consequences significant.

This work seeks to investigate if large language models (LLMs) and vision-language models (VLMs) can function as this representational bridge, enabling robots to translate natural language instructions into actionable behaviors grounded in perception. We leverage the SCOUT (Situated Corpus of Understanding Transactions) dataset (Lukin et al. 2024), a multimodal corpus collected via Wizard-of-Oz (WOZ) experiments where human "Commanders" issued spoken instructions to robots operated by two hidden "Wizards." The Dialogue Manager (DM) converted these instructions into robot commands, while the Robot Navigator (RN) executed them. The long-term goal in human-robot interaction (HRI) is to replace these human intermediaries with autonomous systems that directly interpret and act on human instructions, as we aim to do in this work.

The SCOUT dataset is particularly valuable because it captures visually grounded dialogues under constrained conditions: for example, the Commander relies on sporadic visual snapshots rather than live video. To enable systematic training and evaluation of both LLM and VLM approaches, we extend this dataset into SCOUT++ by time aligning over 11,000 visual frames with natural language instructions and robot actions, creating standardized data for model evaluation.

In this paper, we present a comparative study of three approaches for classifying human instructions into the expected robot action: a neural network trained on tokenized text, GPT-4 operating on text alone, and GPT-4 incorporating synchronized visual context. Although GPT-4 Vision underperforms in raw accuracy, it sometimes produces contextually informed responses—supported by an explicit reasoning step—that highlight both the promise and the current limitations of using language as a representational framework for grounding perception and action in robotics.

SCOUT++ serves as the foundation for our evaluations of three models: a neural network trained on word embeddings, GPT-4 with text-only input, and GPT-4 integrating both text and visual context—enabling direct comparison of symbolic, language-only, and vision-grounded approaches for human-robot interaction.

Background: The SCOUT Corpus

The SCOUT dataset (Lukin et al. 2024) is a multimodal collection of human-robot dialogue recorded during four indoor search and rescue experiments—two in virtual environments and two in real-world settings. Data was collected via

a Wizard-of-Oz (WoZ) approach, where human operators simulated robotic autonomy, allowing researchers to capture natural, task-oriented language in complex environments.

In each session, a human Commander issued spoken instructions to a robot, mediated by two hidden operators: a Dialogue Manager (DM), who interpreted and clarified commands, and a Robot Navigator (RN), who executed the robot's movements. Visual access was intentionally asymmetric: the DM and RN continuously viewed the robot's environment through a live video feed, while the Commander relied on:

- A 2D LIDAR map of obstacles and walls;
- Text responses from the DM;
- On-demand images captured by the robot, requested via commands like "Send Image" or "See view."

This limited visual access often led to ambiguity, for example the command: *"Go to the object in the middle of the room."* Without real-time visuals, the Commander could not confirm what objects were visible, requiring the DM to clarify with follow-up questions before passing instructions to the navigator. Such interactions highlight the challenges of grounding spatial and referential language when perceptual feedback is delayed.

SCOUT comprises 278 dialogues, over 89,000 utterances, 5,785 images, and 30 LIDAR-generated maps, with annotations including AMR, Dialogue-AMR, and dialogue structure labels. In Experiment 4, partial automation was introduced via a classifier trained on 183 labeled instructions using NPCEditor, achieving a 75% match rate without the use of LLMs, visual information, or dialogue history, providing a baseline target for improvement in our work.

To surpass this baseline, we reconstructed and enhanced SCOUT into **SCOUT++**, standardizing transcripts, resolving timestamp inconsistencies, aligning video frames, and normalizing labels. SCOUT++ forms the basis of our experimental evaluation, providing a standardized and richly annotated multimodal benchmark for assessing advances in instruction understanding and grounding in human-robot interaction.

Related Work

Research on natural language as a representational bridge for autonomous systems spans multiple domains, from the design of grounded instruction-following datasets to advances in language and vision-language modeling, and the integration of multimodal perception into robotic decision-making. While these areas may appear distinct, they collectively address the core challenge of enabling robots to interpret and act upon unconstrained human instructions in real-world environments. Datasets such as SCOUT, ALFRED, and R2R provide the foundational benchmarks, language and vision-language models offer the interpretive capabilities, and multimodal fusion techniques supply the perceptual grounding necessary for robust performance. Together, these strands of research form an interdependent pipeline that informs our approach in this work.

Dialogue Classification in Human-Robot Interaction. Dialogue systems for Human-Robot Interaction (HRI) aim to support natural and flexible communication between humans and robots, particularly in safety-critical settings such as search and rescue. Early approaches relied on rule-based templates or simple intent classifiers trained on small corpora (Gervits et al. 2021), which were effective for constrained commands but struggled with ambiguity and contextual understanding. Later work explored statistical and learning-based methods to achieve more robust grounding of natural language in robot actions (Thomason et al. 2020; Juluru et al. 2021). Other studies, such as Tellex et al. (Tellex et al. 2011), emphasized mapping natural language instructions to navigation tasks, demonstrating the importance of structured representations in embodied agents. Our work extends this trajectory by comparing traditional neural networks with a large generative model (GPT-4) for multimodal command classification, a capability largely unexplored in HRI datasets like SCOUT.

Word Embeddings and Large Language Models. Distributed word representations like GloVe (Pennington, Socher, and Manning 2014) and Word2Vec revolutionized NLP by capturing semantic similarity in continuous spaces, enabling improvements in dialogue modeling and intent recognition (Lebret 2016; Yu et al. 2021). Neural models powered by embeddings have been applied to SCOUT (Lukin et al. 2024) and similar datasets to classify instructions, though their reasoning capacity is limited to local patterns. In contrast, Large Language Models (LLMs) such as GPT-3 and GPT-4 (OpenAI 2023) integrate contextual reasoning over long sequences and demonstrate zero-shot or few-shot performance across diverse tasks. Recent work has explored grounding LLMs in multimodal settings, for instance by combining frozen LLM backbones with visual encoders (Tsimpoukelli et al. 2021), or leveraging models like CLIP (Radford et al. 2021) and Flamingo (Alayrac et al. 2022) for vision-language alignment. However, the application of such models to robotic instruction-following remains nascent. We benchmark GPT-4 against a GloVe-powered baseline, highlighting differences in reasoning, label consistency, and generalization.

Multimodal Understanding in Robotics. Multimodal integration (combining vision, language, and other sensory inputs) has become central to grounded instruction-following (Lucignano et al. 2013; Wiriyathammabhum et al. 2016). Research on vision-and-language navigation (VLN) has shown how models like Speaker-Follower (Fried et al. 2018) and R2R (Anderson et al. 2018) can use language as a symbolic guide to navigate unseen environments. Similarly, datasets such as ALFRED (Shridhar et al. 2020) highlight the benefits of connecting instructions with perceptual state for generalization to new tasks. In robotic dialogue, visual grounding improves disambiguation and reduces errors in dynamic settings (Anikina and Kruijff-Korbayova 2019; Sharma, Sharma, and Athaiya 2017).

Despite these advances, many models still treat perception and language as loosely coupled streams, limiting their ability to produce context-aware, actionable outputs. By

aligning 11,000 visual frames with SCOUT dialogues in SCOUT++, we do this in order to explore tighter integration using a multimodal LLM (GPT-4 Vision) for closed-label classification. This approach is intended to reveal the trade-offs between symbolic, neural, and generative approaches, as well as the challenges of using language as a unified representational layer for perception and control.

Comparison with Vision-Language Datasets. Benchmark datasets like Room-to-Room (R2R) (Anderson et al. 2018), Vision-and-Dialogue Navigation (VDN) (Thomason et al. 2019), and ALFRED (Shridhar et al. 2020) study grounded language understanding in simulated, photo-realistic environments. These datasets focus on navigation, manipulation, and visual grounding, but typically rely on scripted or templated instructions, minimal turn-taking, and lack the bidirectional dialogue and role separation found in SCOUT. Collected via interactive Wizard-of-Oz experiments in real and virtual environments, SCOUT features spontaneous, task-driven exchanges that reflect real-time decision-making under uncertainty.

SCOUT++ further aligns timestamped image frames with instruction–response pairs, creating a multimodal dataset rooted in realistic robotic deployments. Unlike simulation-based VLN and ALFRED, SCOUT and SCOUT++ operate with embodied systems, asynchronous feedback, limited visual access, and role-based communication. This makes SCOUT++ well-suited for evaluating models in real-world robotics pipelines. Our GPT-4 Vision experiments highlight both the potential and current limitations of large multimodal models for such scenarios, complementing prior work on simulated instruction following.

Language Models as Policy Generators. Recent work has begun to explore how language models can move beyond interpretation to serve as high-level policy generators for robotics. SayCan (Ahn et al. 2022) combines large language models with affordance functions to select feasible robot actions grounded in physical constraints, illustrating how language can be used to reason about what can be done in context. PaLM-E (Driess et al. 2023) extends this approach with a unified, embodied multimodal model that processes text, images, and sensor data to produce robotic behaviors. Similarly, Code-as-Policies (Bavarian et al. 2022) treats language as a programmatic interface, translating instructions into executable robot code via LLMs. These approaches position language as a rich, structured representation for both perception and control. Our proposal to explore GPT-4 Vision as an interactive decision-making agent builds on this trend by evaluating its potential to generate grounded, context-aware action sequences in real-time, using natural language as the connective tissue between visual perception and autonomous behavior.

Approach

To investigate the role of multimodal inputs in classifying human-robot dialogue within the SCOUT and SCOUT++ datasets, we implemented a structured and reproducible experimental pipeline. Our workflow spans from raw data normalization to the development of two distinct classification models: a neural network trained on tokenized text with word embeddings, and a GPT-4-based language model evaluated with both textual and visual inputs. In this section we outline the full modeling pipeline, including dataset reconstruction, visual context alignment, model architecture and training for the neural network, and prompt engineering and evaluation strategies for GPT-4.

The SCOUT++ Dataset

Although based on the original SCOUT corpus, our experiments required substantial data transformation. The raw SCOUT data existed as multiple Excel files documenting dialogues between human Commanders and a robot, with inconsistent annotations and formats, especially in earlier sessions. We developed a reconstruction pipeline to unify and clean this data for machine learning tasks.

Initially, all files were standardized into a consistent schema, consolidating key metadata fields like session date, environment type, and participant ID. Rows lacking both Commander input and DM→RN responses were removed.

To reduce label sparsity, we normalized response phrases by consolidating semantically equivalent expressions (e.g., mapping "send photo," "image," and "take picture" to "send image"), resulting in a unified set of canonical responses. After cleaning and normalization, all data was merged into a single file gpt_input.xlsx for use in model training.

For multimodal experiments, we aligned commands with visual frames extracted from session videos. This required manually calibrating timestamps for synchronization and applying fixed crop settings to isolate the robot's field of view. Frames were saved in a structured format linking image files to dialogue data. In total, we extracted 11,898 frames, creating a synchronized dataset of paired textual and visual information.

We refer to this processed dataset as SCOUT++, which underpins all experiments in this study.

Neural Network Experiment (Text-Only)

After data preprocessing, we conducted a supervised learning experiment using a neural network trained solely on textual input. Each sample comprised a Commander-issued instruction paired with a DM→RN response label. The goal was to assess whether a deep learning model could classify robot directives based on language alone, without visual or contextual cues. For example, "go to the doorway on the left" mapped to "move to door (left)," while "scan the room" corresponded to "look around," reflecting the linguistic variability and spatial references the model needed to resolve.

To ensure balanced labels, we excluded rare classes (appearing fewer than two times), removed duplicates, and discarded entries missing Commander fields. The final dataset included 2,474 unique instruction–response pairs across 55 response classes.

Text was tokenized using Keras's Tokenizer, restricting the vocabulary to the 1,000 most frequent words. Commands were converted into sequences of up to 150 tokens and embedded with 100-dimensional GloVe vectors (glove.6B.100d.txt), forming a trainable embedding matrix. The neural network architecture consisted of:

- A GloVe-initialized Embedding layer
- A bidirectional LSTM (64 units) with 0.3 dropout
- A Dense output layer with softmax activation over 55 classes

Training was carried out using the Adam optimizer and sparse categorical cross-entropy loss. To mitigate overfitting and improve generalization, we applied a 5-fold cross-validation strategy with stratified sampling. The training process incorporated early stopping (monitoring validation loss with a patience of three epochs) and adaptive learning rate reduction. Additionally, class weights were computed to balance class frequency during training.

Training configuration included:

- Batch size: 64
- Epochs: Up to 20 per fold (with early stopping)
- Evaluation metrics:
  - Accuracy
  - Macro-averaged precision
  - Macro-averaged recall
  - Macro-averaged F1 score

The best-performing model across folds, as determined by F1 score, was saved for subsequent evaluation. During each fold, we generated a confusion matrix to visualize prediction performance across labels and plotted training/validation accuracy curves.

GPT-4 Experiment (Text-Only)

To complement the neural network experiment, we evaluated OpenAI's GPT-4 large language model using the same Commander-issued instructions used for the NN training, ensuring fair comparison. And unlike the subsequent multimodal setup, this experiment focused purely on textual modality, aiming to test GPT-4's ability to classify natural language instructions into robot control responses using only the command text and a limited set of possible outputs.

Each prompt combined a Commander-issued instruction with a list of valid DM→RN response options (see Appendix A). Prompts contextualized the task in an urban disaster response scenario and instructed the model to select the best response and provide brief reasoning. For example:

> You are a robot operating in an urban disaster response scenario. Your task is to respond to the following command appropriately based on your training and mission goals.
>
> Commander Input: move forward and turn left
>
> Available responses:
>
> - move forward
> - turn left
> - stop
> - move forward and turn left
> - ...
>
> Select the best response from the options above and explain your choice.

Prompts were submitted to the gpt-4o model via the OpenAI Chat Completion API. The system role was defined as "You are a search and rescue robot expert," and a temperature of 0.5 balanced response diversity and determinism. GPT-4 typically returned the chosen answer, clearly marked, along with a rationale. Post-processing extracted the selected response for comparison to the ground truth label from the SCOUT dataset.

Evaluation was treated as a classification task, considering a response correct only if it exactly matched the corresponding DM→RN label. Outputs included the prompt, GPT-4's selected response, the true label, and the model's explanation. In total, GPT-4 processed 2,474 commands using, as stated earlier, the same candidate labels as in neural network training. Enabling direct comparison of classification accuracy and language understanding across both approaches.

GPT-4 Experiment (Multimodal)

Following earlier text-only experiments with a neural network and GPT-4, we extended our study to evaluate GPT-4's performance in a multimodal setting. Here, each Commander-issued instruction was paired with an image frame depicting the robot's current environment. The aim was to assess whether visual context improves GPT-4's ability to correctly interpret and classify instructions, especially those involving spatial or referential grounding.

Each prompt integrated the command text, a set of candidate DM→RN responses, and a base64-encoded image. The scenario simulated real-time decision-making for a search and rescue robot navigating urban disaster environments. An example prompt structure was:

> You are a robot operating in an urban search and rescue scenario. The following image shows your current field of view.

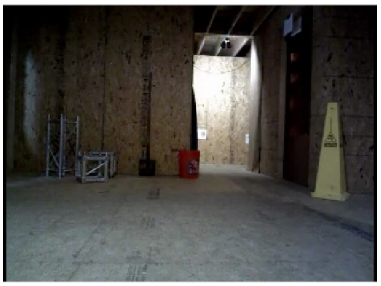

Figure 1: Example image used with the given prompt to GPT-4.

> Commander Input: move forward to the red bucket
>
> Available responses:
>
> - move forward
> - move to the red bucket
> - turn left
> - stop

Commander Instruction	GPT Output	Model Reasoning
”Go past it”	`"move forward"`	*“The instruction ’Go past it’ lacks a clear referent. Based on the visible path and obstacle layout in the image, the model assumes a safe forward motion past the nearest object.”*
”Red box maybe”	`"move to the red box"`	*“The phrase likely refers to the red box partially visible in the lower-right corner of the image. The model interprets this as a tentative command to approach that object.”*
”Turn”	`"turn right"`	*“The command lacks a direction. The model defaults to ’turn right’ based on common prior usage patterns in the training data.”*

Table 1: Examples of GPT-4 responses to ambiguous or underspecified Commander instructions. In some cases, GPT-4 Vision incorporates cues from the visual frame to ground its interpretation, illustrating the benefits of multimodal understanding.

- ...

Select the best response from the list above and explain your choice.

Prompts were processed via OpenAI’s Chat Completion API using the gpt-4o-mini model, capable of handling both text and images. The system role was set as “search and rescue robot expert,” and a temperature of 0.5 balanced determinism and variability. We processed 12,000 multimodal prompts in batches to manage dataset size.

Responses were post-processed to extract the selected action and compare it with the ground truth DM→RN label. Results, including the prompt, model response, ground truth label, correctness flag, and rationale, were recorded in a .xlsx file.

Incorporating visual data enabled us to evaluate GPT-4’s grounded language understanding and better approximate the perceptual challenges of human-robot interaction.

Results

In this section, we presented the outcomes derived from employing the three distinct natural language processing methodologies within the confines of this project. Specifically, we stated the results from the utilization of the Neural Network model trained and assessed with pre-trained GloVe word embeddings, as well as the implementation of the BERT and its distilled version, DistilBERT.

In the next section, a comprehensive analysis and interpretation of the results shown here, will be discussed at length.

Neural Network (Text-Only)

The model achieved the following average performance across five folds:

- Average Accuracy: 61.72%
- Macro-averaged Precision: 60.37%
- Macro-averaged Recall: 67.32%
- Macro-averaged F1 Score: 0.6134

The best-performing fold (Fold 3) reached the highest macro F1 score of 0.6134. Qualitative analysis of the training dynamics revealed several consistent trends. Early epochs were marked by substantial class confusion, especially between semantically similar instructions such as “move forward” and “move forward and turn left.” The model frequently struggled to resolve nuanced differences in phrasing without access to situational or visual cues. Performance improved markedly after five epochs, driven in part by the learning of domain-specific command patterns through the embedding and recurrent layers.

Despite being constrained to a single input modality (text), the neural model demonstrated consistent, interpretable performance and avoided the hallucination issues observed in some generative systems. However, its inability to incorporate contextual, multimodal or historical dialogue information limited its resolution of ambiguous or visually grounded instructions.

GPT-4 Text-Only Classification

The following metrics summarize model performance:

- Correct classifications: 2,063 (79.9%)
- Incorrect classifications: 411
- Macro-averaged Precision: 0.35
- Macro-averaged Recall: 0.40
- Macro-averaged F1 Score: 0.35

Despite the relatively high raw accuracy, the macro-averaged metrics reflect the model’s challenges in handling a diverse label set (55 unique action classes) and the tendency to favor more frequent or generic responses.

Post-Processing and Normalization GPT-4’s outputs were not always well-aligned with the label set. Several forms of normalization were applied:

- Responses were stripped of formatting artifacts (e.g., “Selected Response:” headers, quotation marks).
- Slightly paraphrased labels were matched using fuzzy string techniques.
- Responses that failed to match any label, even approximately, were categorized under a fallback label: ”unrecognized”.

Only 13 out of 2,474 predictions fell into this fallback category. Most of these cases were due to hallucinated answers, formatting issues, or incomplete responses. For example see Table 1.

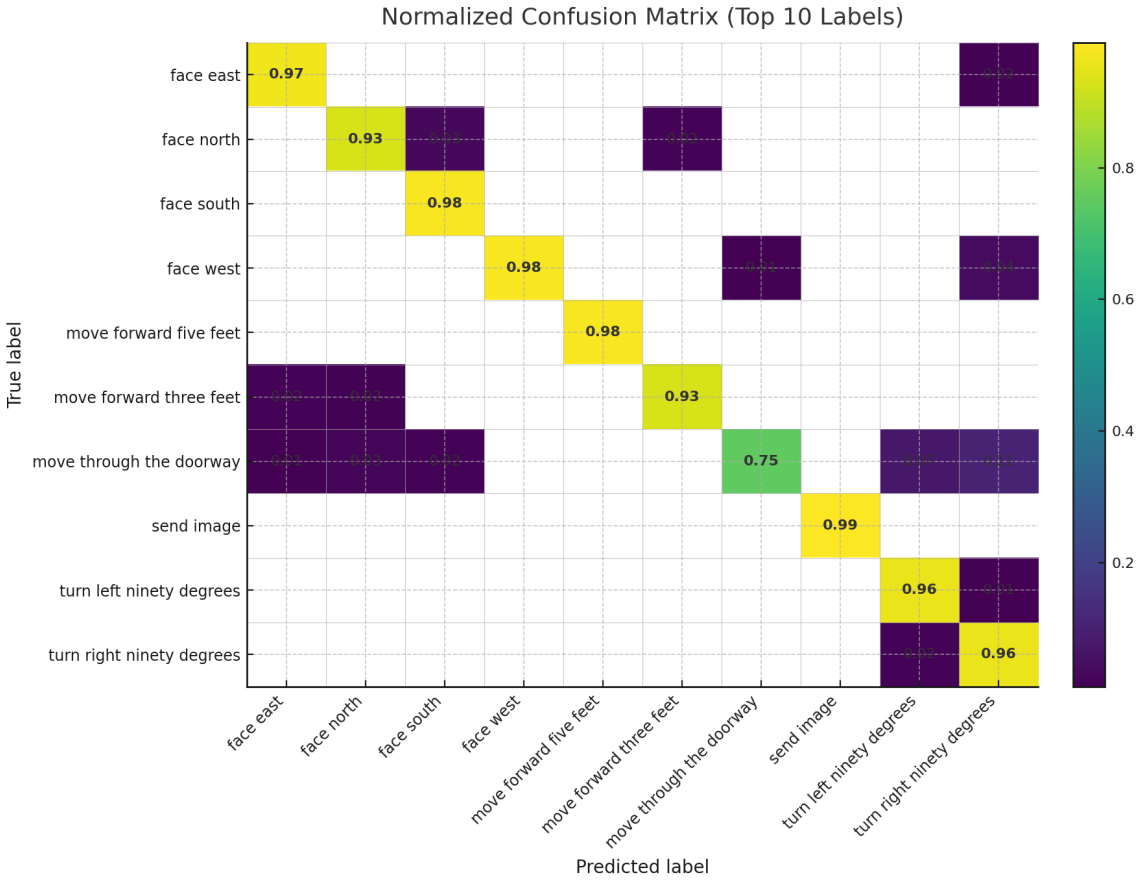

Figure 2: Normalized confusion matrix for the top 10 most frequent DM→RN instruction labels. Each row represents the true label, and each column the predicted label. Higher value diagonal cells indicate high classification accuracy for common commands such as “face south,” “send image,” and “move forward five feet.” Off-diagonal entries highlight common confusions between semantically similar actions, such as “move through the doorway” and “move forward three feet.”

In some cases, GPT-4 responded with meta-commentary indicating uncertainty or misunderstanding of the task, such:

“I’m sorry, but I can’t execute the command ‘see picture’ without additional context or a visual reference.”

These responses further highlight the limitations of text-only classification when handling commands that are inherently context-dependent or visually grounded.

To better understand model behavior, a confusion matrix was generated for the top 10 most frequent instruction labels (see Fig. 2). The visualization revealed patterns of semantic confusion, particularly among movement-based commands (e.g., misclassifying ”move forward one meter” as ”go forward” or ”proceed straight”).

GPT-4 with Image Input

Evaluation was conducted on 12,000 instruction–response pairs, each comprising a Commander-issued natural language command and a synchronized still frame from the SCOUT video recordings. The model achieved the following performance:

- Correct classifications: 3,616 (30.13%)
- Incorrect classifications: 8,384
- Unrecognized outputs: 975 (8.13%) — often due to hallucinated or incomplete responses
- Macro-averaged Precision: 0.2296
- Macro-averaged Recall: 0.2142
- Macro-averaged F1 Score: 0.1767

During post-processing, GPT-4 outputs were normalized and compared to a set of 55 predefined DM→RN labels. Despite these efforts, 975 predictions could not be mapped to any valid label and were marked as “unrecognized.”

Confusion Matrix: The confusion matrix for the 10 most frequent labels revealed substantial overlap between unrelated categories and limited benefit from image-grounded disambiguation.

Sample Outputs: Representative cases of correct predictions, label mismatches, and unrecognized outputs are provided in the appendix for qualitative comparison.

Model	Input Type	Accuracy	Precision	Recall	F1 Score
Neural Network (GloVe + LSTM)	Text only	61.72%	0.6034	0.6650	0.6000
GPT-4 (Text Only)	Text only	79.9%	0.3526	0.3976	0.3503
GPT-4 (Text + Image)	Multimodal	30.13%	0.2296	0.2142	0.1767

Table 2: Performance Metrics Across Models. This table compares three models evaluated on SCOUT++ instruction classification: a neural network using GloVe embeddings and LSTM, GPT-4 with text-only prompts, and GPT-4 Vision using synchronized text and visual inputs. While GPT-4 (Text Only) achieved the highest accuracy, it struggled with label consistency, leading to lower precision and F1 scores. GPT-4 Vision underperformed overall, suggesting current limitations in visual grounding despite multimodal input.

Discussion

We evaluated three modeling approaches for the task of instruction-response classification in human-robot dialogue using the SCOUT++ dataset. Each model was tasked with predicting the appropriate DM→RN response given a Commander-issued instruction, with and without access to visual context. Additional evaluation details include:

- **Instruction–Response Pairs Used:** 2,474 (Neural Network and GPT-4 Text); 12,000 (GPT-4 Multimodal)
- **Unrecognized Predictions:** 0 (NN), 13 (GPT-4 Text), 975 (GPT-4 Multimodal)

While GPT-4 (text-only) achieved the highest raw accuracy, its macro-averaged metrics were significantly lower due to prediction bias toward frequent labels. The neural network model offered more balanced performance across classes, while the multimodal GPT-4 model underperformed in both accuracy and consistency.

Each approach showed distinct strengths and limitations, shaped by both model architecture and data completeness. The neural network, constrained by fixed-vocabulary embeddings and sequential modeling, performed consistently with an average macro F1 score of 0.60. Its outputs stayed within the known label space, ensuring reliability, but it struggled with semantically similar instructions where visual context could have clarified intent—for example, differentiating "turn left" from "move left and rotate."

GPT-4 in text-only mode achieved higher classification accuracy (79.9%) but lower precision and recall, due to frequent over-prediction of common labels and occasional semantic drift. Unlike the neural network, GPT-4 was not trained directly on SCOUT labels, necessitating post-processing to map responses to valid classes. Nevertheless, it demonstrated strong pattern recognition and insightful reasoning. Thirteen outputs remained unmappable despite normalization.

Unexpectedly, GPT-4's multimodal variant, which received both images and text, performed worst, with only 30.13% accuracy and an 8.1% unrecognized rate. Predictions were often vague, hallucinatory, or defaulted to disclaimers like "I cannot determine this from the image." Explanations frequently suggested confusion, hinting at underutilization of visual data or suboptimal prompt integration.

However, this apparent underperformance deserves caution. In the original SCOUT experiments, Commanders worked with limited, on-demand visual input, whereas GPT-4 Vision had continuous visual context. Thus, some "incorrect" GPT-4 predictions may actually have been more contextually appropriate or safer, reflecting fuller environmental awareness (see Fig. 3).

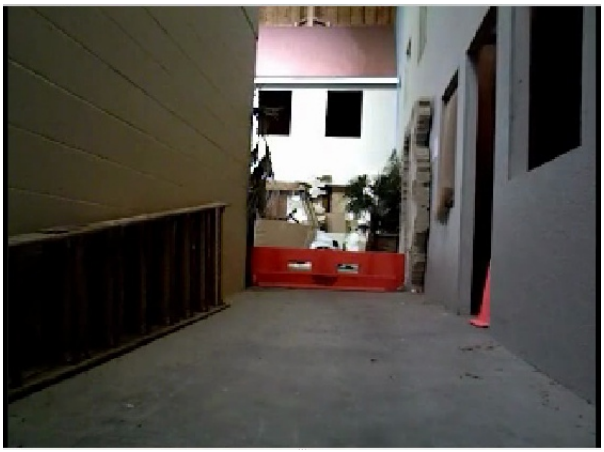

Figure 3: Visual frame from Experiment 1 aligned with the Commander-issued instruction.

Commander Input	*pivot left one hundred eighty degrees*
SCOUT Label	*turn left one hundred eighty degrees*
GPT-4 Prediction	*move forward three feet*
GPT-4 Reasoning	The command to pivot left 180 degrees indicates that the robot should turn around, and moving forward three feet would help it advance into the open area visible in the image, potentially allowing for further exploration or assessment of the environment.

This example demonstrates how GPT-4 Vision inferred a more contextually informed instruction based on visual input.

This informational asymmetry complicates direct comparisons. Overall, our results highlight the trade-off between the reliability and constraint of structured models and the flexibility—but variable accuracy—of generative systems. Future work could explore hybrid architectures, domain-specific fine-tuning, and improved data alignment for vision-language models in human-robot interaction.

Future Work

This study suggests several directions for improving grounded human-robot dialogue, especially in multimodal understanding.

First, GPT-4 Vision's limited performance indicates a need for stronger visual grounding. Future work could integrate structured vision modules—like object detection or scene graphs—to preprocess images before combining them with text, improving context-aware reasoning.

Second, large language models struggle with label consistency due to generative outputs. Fine-tuning GPT-4 on constrained response formats or using prompts enforcing strict label patterns could improve accuracy. Soft-constrained decoding or post-processing may help resolve mismatches observed in this study.

Third, our findings highlight a visual asymmetry: Commanders operated with limited images, while GPT-4 Vision had full visual context. This makes direct comparison difficult, as some GPT-4 responses might reflect better-informed decisions. Future experiments could restrict GPT-4's visual input to mirror the Commander's limited perspective.

Longer-term, we plan to create a new dataset combining images with sensor data like LiDAR, depth, GPS, and IMU readings. This would help models reason not only about visual content but also spatial location and movement (see Fig. 4).

A compelling direction is exploring GPT-4 Vision as a policy generator, predicting the next action based on ongoing visual input rather than simply classifying static commands. This could enable autonomous robot control loops where GPT-4 suggests actions directly, supporting training for lighter models or reinforcement learning agents.

Such work could transform LLMs from reactive interpreters into proactive agents capable of autonomous planning and perception.

Conclusion

This research explored the challenge of interpreting unconstrained natural language instructions in human-robot dialogue, with a focus on language as a representational bridge between perception and action. Using the SCOUT dataset, an urban disaster response corpus rich in multimodal interactions, we implemented and evaluated three models: a neural network trained on GloVe-embedded text, a GPT-4 model operating on raw language input, and a GPT-4 Vision model that incorporated both text and synchronized visual frames.

While the neural model offered consistent, label-constrained outputs, it lacked contextual flexibility. GPT-4's text-only variant demonstrated stronger reasoning but sometimes deviated from the SCOUT label space. The GPT-4 Vision model, despite lower raw accuracy, produced responses that were often more contextually informed—suggesting a form of grounded reasoning that exceeds what was possible under the original Commander's limited visual access. This finding highlights not only the promise of multimodal models, but also the difficulty of defining correctness in dialogue tasks when informational asymmetries exist between human and machine agents.

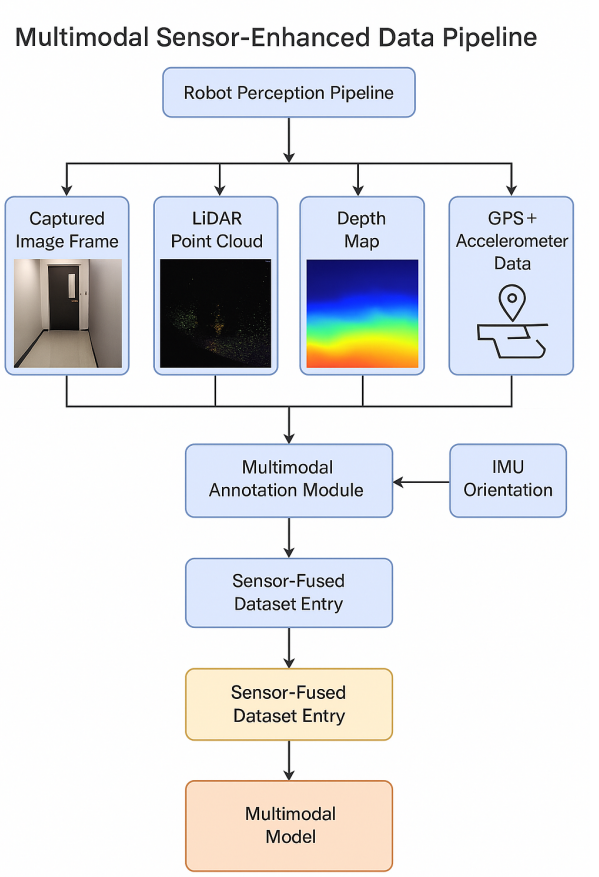

Figure 4: LiDAR Point Cloud: 3D spatial structure of environment.— Depth Map: Per-pixel depth info to understand relative distances.— GPS + Accelerometer: Absolute location + robot movement.— IMU: Orientation and heading data.— Annotation Module: Combines all sensor inputs + links to natural language instruction and action label.— Sensor-Fused Dataset Entry: A single rich training example.— Multimodal Model: The future model that processes these fused inputs.

Looking ahead, we advocate for the creation of richer multimodal datasets that fuse visual input with spatial and dynamic sensor data, such as LiDAR, depth, GPS, and IMUs (see Fig. 4). Such data will allow models to reason not only about what is seen, but also about where the robot is and how it is moving, enabling more robust situational grounding. Beyond classification, we propose reframing large language models as proactive policy generators: agents that can interpret context and suggest next actions, transforming static instruction-following into dynamic task planning.

Together, these contributions lay the foundation for a new generation of dialogue systems that treat language as a multimodal, perceptually grounded representation, one that supports more autonomous, capable, and context-aware human-robot collaboration in mission-critical environments.

References

Ahn, M.; Brohan, A.; Chebotar, Y.; Hou, K.; Hsu, J.; Lee, B.; et al. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *Robotics: Science and Systems (RSS)*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Cabi, S.; Driess, D.; et al. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*.

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Gould, S.; and van den Hengel, A. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 3674–3683.

Anikina, T.; and Kruijff-Korbayova, I. 2019. Dialogue Act Classification in Team Communication for Robot Assisted Disaster Response. In Nakamura, S.; Gasic, M.; Zuckerman, I.; Skantze, G.; Nakano, M.; Papangelis, A.; Ultes, S.; and Yoshino, K., eds., *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 399–410. Stockholm, Sweden: Association for Computational Linguistics.

Bavarian, M.; Garg, A.; Lee, B.; Chebotar, Y.; Hsu, J.; Brohan, A.; et al. 2022. Code as Policies: Language Model Programs for Embodied Control. In *Conference on Robot Learning (CoRL)*.

Driess, D.; Srinivas, A.; Chen, M.; Colburn, J.; Ichter, B.; Lynch, C.; et al. 2023. PaLM-E: An Embodied Multimodal Language Model. *arXiv preprint arXiv:2303.03378*.

Fried, D.; Hu, R.; Cirik, V.; Rohrbach, A.; Andreas, J.; Morency, L.-P.; Berg-Kirkpatrick, T.; Saenko, K.; Klein, D.; and Darrell, T. 2018. Speaker-follower models for vision-and-language navigation. In *NeurIPS*, volume 31, 3318–3329.

Gervits, F.; Leuski, A.; Bonial, C.; Gordon, C.; and Traum, D. 2021. A classification-based approach to automating human-robot dialogue. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, 115–127. Singapore: Springer Singapore.

Juluru, K.; Shih, H.-H.; Keshava Murthy, K. N.; and Elnajjar, P. 2021. Bag-of-Words Technique in Natural Language Processing: A Primer for Radiologists. *RadioGraphics*, 41(5): 1420–1426. PMID: 34388050.

Lebret, R. P. 2016. Word embeddings for natural language processing. Technical report, EPFL.

Lucignano, L.; Cutugno, F.; Rossi, S.; and Finzi, A. 2013. A Dialogue System for Multimodal Human-Robot Interaction. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, 197–204. New York, NY, USA: Association for Computing Machinery. ISBN 9781450321297.

Lukin, S. M.; Bonial, C. N.; Marge, M.; Hudson, T.; Hayes, C. J.; Pollard, K. A.; Baker, A.; Foots, A.; Artstein, R.; Gervits, F.; Abrams, M.; Henry, C.; Donatelli, L.; Leuski, A.; Hill, S. G.; Traum, D.; and Voss, C. R. 2024. SCOUT: A Situated and Multi-Modal Human-Robot Dialogue Corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

OpenAI. 2023. GPT-4 Technical Report. urlhttps://openai.com/research/gpt-4. Accessed: 2025-08-02.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Sharma, S.; Sharma, S.; and Athaiya, A. 2017. Activation functions in neural networks. *Towards Data Sci*, 6(12): 310–316.

Shridhar, M.; Thomason, J.; Mooney, R.; Mottaghi, R.; Zettlemoyer, L.; and Fox, D. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *CVPR*, 10740–10749.

Tellex, S.; Knepper, R. A.; Li, A.; Rus, D.; and Roy, N. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, volume 25, 1507–1514.

Thomason, J.; Gordon, D.; Bisk, Y.; Han, X.; Zettlemoyer, L.; and Mottaghi, R. 2019. Vision-and-dialog navigation. In *Conference on Robot Learning (CoRL)*.

Thomason, J.; Padmakumar, A.; Sinapov, J.; Walker, N.; Jiang, Y.; Yedidsion, H.; Hart, J.; Stone, P.; and Mooney, R. 2020. Jointly Improving Parsing and Perception for Natural Language Commands through Human-Robot Dialog. *Journal of Artificial Intelligence Research*, 67: 327–374.

Tsimpoukelli, M.; Menick, J.; Cabi, S.; Eslami, S. A.; Vinyals, O.; and Hill, F. 2021. Multimodal few-shot learning with frozen language models. In *NeurIPS*, volume 34, 200–212.

Wiriyathammabhum, P.; Summers-Stay, D.; Fermüller, C.; and Aloimonos, Y. 2016. Computer Vision and Natural Language Processing: Recent Approaches in Multimedia and Robotics. *ACM Comput. Surv.*, 49(4).

Yu, B.; Deshpande, A.; McLeod, A. M.; Patha, N. S. L.; and Dreyer, M. 2021. Word embeddings for natural language processing. US Patent 11,030,999.