Non-Event Oriented Video Assessments in Long-Form Robot Videos

Stephanie M. Lukin[1], Kimberly A. Pollard[1], Claire N. Bonial[1], Cory J. Hayes[1], Ron Artstein[2], Kallirroi Georgila[2], David Traum[2]
[1]DEVCOM Army Research Laboratory
[2]USC Institute for Creative Technologies
Correspondence: stephanie.m.lukin.civ@army.mil

Abstract

We introduce Video-SCOUT, a novel dataset of sixty 20-minute robot-recorded videos from human-robot collaborative exploration exercises, together with a new video analysis method for these types of exploration videos. Unlike video from stationary cameras where detection of motion can help identify events of interest, the camera in an exploration task is constantly in motion while the environment is stationary. Our analysis method—Non-Event Oriented Video Assessments (NOVA)—uses vision-language models to select frames relevant for supporting a particular assessment within continuous long-form videos. Results of testing with two different video-language models reveals a trade-off in precision and recall, and exhibits gains in overall recall when combined with a human's knowledge, suggesting that NOVA may improve a human analysis of robot-navigation. We outline future work to mitigate miscommunication in human-robot interaction by leveraging dialogue with NOVA in support of better collaboration.

1 Introduction

Robot camera perspectives have proven beneficial in capturing environments in-the-wild which may be unsafe for humans, e.g., in disaster response (Fernandes et al., 2019; Jayawardene et al., 2021; Chiou et al., 2022b; Chitikena et al., 2023) or search and rescue (Drew, 2021; Chiou et al., 2022a; Wang et al., 2023; Esteves Henriques et al., 2024). A human working with a robot to move through these spaces may use controls to teleoperate it or may issue verbal instructions, requiring a dialogue to ensure the human and robot establish common ground (shared beliefs and assumptions (Clark and Marshall, 1981)) throughout the journey. This communication becomes critical under bandwidth constraints where the human cannot see the robot's view in real time. Robot remote exploration videos vastly differ from those we encounter on a daily basis, such as current events in the news, movies and TV shows, and social media such as YouTube. Social Media videos are commonly clear, well-lit, shot from camera angles typical of those used in human-focused or commercial media, and are curated to depict activities or events of interest to the audience. By contrast, video recordings of robot journeys may show a non-human camera perspective, i.e., a ground robot looking up, and may be grainy and poorly-lit. The video may contain periods with no activity, or stretches where the robot moves through sparse or repetitive environments.

To develop techniques for automatic video understanding of robot-recorded videos, we must reevaluate what "understanding" means when nothing appears to be "happening." How do people talk about uncertain spaces they are trying to move around? To study these questions, we introduce a video and language resource comprised of long-form, robot-recorded videos collected through human-robot dialogue while the robot explores a remote, sparse environment without activity. **Video-SCOUT** consists of 60 videos averaging 20 minutes long with accompanying human-robot dialogue referencing the environment. Smaller video clips of the exploration are segmented by annotations of the dialogue's structure. Video-SCOUT will be made publicly available at `https://github.com/USArmyResearchLab/ARL-SCOUT` under a CC0-1.0 license. While much work has examined events and activity within videos, Video-SCOUT differs from these in video quality, perspective, length, and content, which we compare in detail.

We present a new challenge in which a robot supports a human in conducting environment-level assessments of robot-recorded videos. We call this **Non-Event Oriented Video Assessment (NOVA)**, in which videos may be utilized by human-robot exploration teams in collecting visual evidence from the environment within the video to answer specific assessment questions. We evaluate the feasibil-

Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval, pages 27–41

ity of using Vision-Language Models (VLMs) on this challenge with Video-SCOUT. Frames of high relevance are selected by VLMs from the robot-recorded video given a natural language description of the human's assessment task. Model performance is evaluated with precision and recall, and we examine to what degree these models could supplement a human's performance in making assessments. We conclude by laying the groundwork for a future framework that provides contextual, situated knowledge by leveraging the turn-by-turn dialogue structure alongside the robot recorded video and NOVA task to mitigate miscommunication for more grounded and efficient human-robot interactions.

2 Video-SCOUT

Video-SCOUT consists of robot-recorded videos from the Situated Corpus of Understanding Transactions (SCOUT) (Lukin et al., 2024), a dataset of experimental trials completed by human participants with remotely-located ground-robot partners. Participants used language to instruct their robot teammate to move through a remote environment and seek objects of interest and make assessments about the space. To model connectivity challenges, bandwidth was limited to sending linguistic messages, LiDAR (Light Detection and Ranging) information, and occasional images, rather than real-time full video or teleoperation.

2.1 SCOUT: Human-Robot Collaboration

SCOUT is comprised of four human-robot experiments with increasing automation of dialogue management and language generation while maintaining the same participant-facing affordances and tasks (Bonial et al., 2025). Participants interacted with the remotely-located robot to explore a house-like environment with unfinished walls, floors, and sparse furniture items. Participants issued unconstrained, spoken instructions to a Clearpath Jackal robot using a push-to-talk interface, and the robot responded through text messages. Participants were shown a 2D top-down LiDAR map created from the robot's Hokuyo laser scanner that updated in real time as the robot moved. While the robot had continuous access to its front-facing RGB camera (an Asus Xtion Pro Live), participants did not. Instead, participants were informed of the robot's ability to take and send still photographs. Photos of the environment could be requested at any time. Many participants used photos to see where to go and took photos to comprehensively view the environment (Lukin et al., 2023).

Participants completed two 20-minute main trials with the robot's assistance after first completing a training trial in which they could familiarize themselves with the interface, activity, and robot capabilities. A paper worksheet was provided to participants, and they were asked to count and photograph objects of interest which varied by trial (e.g., cones, shovels, and shoes). Furthermore, participants were asked to answer the following assessments: "Is there anything that indicates the environment has recently been occupied?", "Were the last occupants speakers of English or a foreign language?", and "Is there anything that you could use to coordinate operations or activities in a headquarters type environment?" Participants were not required to photograph supporting evidence encountered for these assessments, but they verbally reported their conclusion to the experimenter at the trial's end, and could cite encountered evidence.

The robot was controlled by Wizards-of-Oz, human confederates standing in for the robot's dialogue and navigation capabilities. The data collected from earlier SCOUT experiments with Wizards was used to incrementally automate the robot's Dialogue Management (DM) in later experiments. The DM, whether it was a DM-Wizard confederate or the automated system, listened to the participant's instructions and either responded to the participant or passed well-formed and executable instructions to a Robot Navigator (RN) Wizard who reported success or problems. The participant was only made aware they were speaking to a 'robot' and not informed of the inner workings of the robot's controls. The dialogue was annotated for Transaction Units (TUs), a dialogue structure annotation demarcating multiple dialogue turns that sequentially contributed to fulfilling the speaker's original intent across all speakers (the participant, DM, and RN) (Traum et al., 2018).

2.2 The Video-SCOUT Dataset

Video-SCOUT consists of 60 robot-centric RGB videos from the main and training trials of SCOUT Experiments 1 and 2. The total video time of the dataset is 20+ hours, averaging 20:14 minutes per video (Table 1). The videos are accompanied by a transcript file containing the dialogue between the participant and robot (DM-Wizard) which can be played as subtitles with the video. Additionally,

	# Videos	Video Length	
		Total	**Average**
Video-SCOUT	*60*	*20 hrs. 13 min.*	*20:14 min.*
Main Trials	40	14 hrs. 36 min.	21:55 min.
Train Trials	20	5 hrs. 37 min.	16:53 min.
TU Videos	1,672	15 hrs. 17 min.	32 sec.

Table 1: Video-SCOUT summary statistics

each video is split into its constituent TUs, totaling 1,672 TU video clips across main and training trials.

The *robot videos* were created by first extracting the `sensor_msgs/rgb/image_raw` topic from the SCOUT experimental bag files, then using `ffmpeg`[1] to create `.mp4` files. The average length of the main trial videos is 21:55 minutes (max: 27:28 due to network technical difficulties, min: 20:41) and the training trials average 16:53 minutes (max: 22:37, min: 9:56). All videos include an initial calibration procedure and the participant verbally reporting their counting and assessment responses to the experimenter at the conclusion of each trial. The robot had access to these videos during the experiment, while the participant could only see the robot's in-the-moment view with a photo request.

To create *subtitle files*, the time-aligned transcripts from the SCOUT dataset were converted into `.srt` text files with the participant and DM-Wizard dialogue. Each utterance was assigned a sequential ID in the `.srt` file. The participant's push-to-talk keypress timestamps determined how long the utterance subtitle remained on screen. The DM-Wizard text messages remained on screen for 5 seconds after the message was sent to the participant. Subtitle files can be added as a subtitle track to the videos using media players, revealing the robot's view of the environment in real time alongside the ongoing dialogue. An excerpt of the subtitle file from p2.02's main2 trial is below, where 'CMD' ('Commander') indicates the participant:

```
51
00:07:14.48 --> 00:07:19.48
CMD: "can you move several yards towards the
white door"
52
00:07:22.11 --> 00:07:27.11
Robot: "processing. . ."
53
00:07:53.12 --> 00:07:58.12
Robot: "I will move forward 6 feet, ok?"
54
00:07:59.63 --> 00:08:04.63
CMD: "uh i think six feet is too fff far"
55
00:08:04.46 --> 00:08:09.46
CMD: "maybe three feet"
56
00:08:08.45 --> 00:08:13.45
Robot: "ok"
57
00:08:30.54 --> 00:08:35.54
Robot: "moving. . ."
58
00:08:36.93 --> 00:08:41.93
Robot: "done"
59
00:08:38.42 --> 00:08:43.42
CMD: "can you take a photo"
60
00:08:43.33 --> 00:08:48.33
Robot: "sent"
```

Transaction Unit (TU) video clips were created by segmenting the trial video into clips containing the beginning and end of each TU. TU annotations were obtained from the SCOUT dialogue structure `.xlsx` spreadsheets.[2] The average length of the TU video clips is 32 seconds (max: 4 min. 9 sec., min: 4 sec.). Accompanying each TU video is a *TU video clip subtitle file* containing the dialogue exclusive to that TU, supplying the robot's view when participant instructions are issued, and revealing successes and discrepancies in common ground when the participant did not have full access to the video like the robot did. The timestamp for the TU's first utterance was set to 0:00:00 to play with the TU video, and the `.srt` utterance ID restarted within each TU. The excerpt above contains two TUs, therefore has two separate TU videos and transcripts (see Appendix A for the `.srt`s.)

3 Comparison with Related Work

Video understanding is typically conditioned on the categorization or detection of an 'event' visible within the video. There are differing levels of granularity in defining an event. At a high level, videos have been assigned a label depicting an overall category of the content, such as 'vehicle' or 'nature' (Thomee et al., 2016), or 'news' or 'travel' (Abu-El-Haija et al., 2016). At a more detailed level, videos have been labeled by their activity in full, such as 'changing a vehicle tire' (Smeaton et al., 2006) or 'poking a hole into a substance' (Goyal

[1] https://ffmpeg.org/

[2] TUs solely between the participant and experimenter are excluded.

et al., 2017). Others take a hierarchical approach, constructing an event template with sub-events and entity roles. A cooking video is labeled with sub-events including 'grill the tomatoes' followed by 'add oil to a pan' (Zhou et al., 2018). Disaster videos have been annotated to identify the who, what, when and where of their sub-events, such as differentiating within the video between emergency responders and people affected by a flood (Sanders et al., 2024). Many of these datasets are accompanied by natural language, including news articles written about the videos, creating rich, multimodal and multilingual datasets for event understanding, e.g., Sanders and Van Durme (2024); Kriz et al. (2025). Commonly, these event-centric tasks examine videos from YouTube, Flickr, Vimeo, movies, and TV shows. The reader can refer to Sanders and Van Durme (2024) for a comprehensive overview of recent event-centric video datasets. Kriz et al. (2025, Table 1, p2) report these videos range in length from 4 seconds to 8 minutes.

The Video-SCOUT dataset of robot-recorded videos has unique content and characteristics compared to these event-centric datasets. 'Events' at any level of granularity do not appear in Video-SCOUT, as there is no human activity depicted nor is there motion within the environment. Furthermore, the quality of Video-SCOUT videos is degraded by low-quality recording devices, and the robot's low-to-the-ground video perspective differs from human height. The average length of a Video-SCOUT video is 20 minutes, challenging how well video understanding can be conducted over a long period of time as opposed to shorter clips and in significantly different domains.

Other video understanding approaches are pattern-based, looking not to apply an 'event' annotation to a video, but rather identify where in the video a pre-defined pattern breaks. These video datasets are based around observing crowded scenes from a stationary camera over a period of time, e.g., the ShanghaiTech Campus Dataset (Luo et al., 2017), the UCSD Pedestrian Dataset (Li et al., 2013), the Subway Dataset (Adam et al., 2008), and the CUHK Avenue Dataset (Lu et al., 2013). They seek to identify anomalies, when the pattern changes, such as people fighting (Adam et al., 2008), or non-pedestrian entities entering a walkway (Pinggera et al., 2016). These videos are similar to Video-SCOUT in that our videos are recorded from a non-human perspective, yet the key difference concerns spatial movement. In Video-SCOUT, it is the robot moving, rather than entities within the environment, thus, as is the case in prior works, we must again look for how to define pattern breaking anomalies within this domain.

In an attempt to generalize an 'event,' other video understanding approaches have instead created a highlight reel (i.e., a set of individual frames extracted from a video compiled into a short video clip) or a video summary tailored for *any* video content (Song et al., 2015; Sul et al., 2023; Chang et al., 2025). Highlight detection algorithms use accompanying language from the video's metadata to extract frames relevant to the text, for example, the YouTube video's title "Killer Bees Hurt 1000-lb Hog in Bisbee AZ" (Song et al., 2015, p1) is used to condition video frame extraction relating to these keywords. To address the lack of typical 'events' in Video-SCOUT, we leverage Chang et al. (2025)'s Aha! highlight detection model and explore how other language inputs may guide video analysis of robot-recorded videos. Chang et al. (2025)'s paper analyzed eight minutes of a robot-recorded video with the input 'what objects are here?' The preliminary analysis suggested Aha! may be used on out-of-domain videos without fine-tuning, yet the paper did not conduct a thorough evaluation.

Video-SCOUT captures explorations of a space that an embodied agent, i.e., a robot, moves through. These embodied explorations are common in human-agent or human-robot exercises taking place in the physical or virtual world, where the agents are given a directive, e.g., move to a specific object or location, that the agent will complete (Das et al., 2018; Shridhar et al., 2020; Majumdar et al., 2024; Bowser et al., 2025). Dialogue within these environments allows the robot to request clarification of ambiguous instructions or referents to resolve ambiguity (Gervits et al., 2021). However, due to the fact that many real-world remote exploration contexts are bandwidth-limited, full information about the environment (including real-time streaming video) may not be available to the human issuing the robot its navigational instructions. The human must make decisions with potentially incomplete information, decisions that the robot must carry out or clarify. We formulate a new video understanding challenge around the uniqueness of our domain: in "understanding" a video without events or pattern breaking anomalies in the human-robot collaborative context.

4 Non-event Oriented Video Assessments

The Video-SCOUT dataset presents new opportunities for advancing research in human-robot collaboration by leveraging out-of-domain video datasets and task requirements. Given that the videos are of real-world quality (i.e., occasionally dark or blurry), and that they are not streamed live to the human, it is possible the human may overlook or miss something important over the course of their exploration. This is a critical opportunity for the robot teammate to provide support by analyzing its constantly running RGB camera. A robot may additionally augment a human's understanding of the long-form retrospective exploration videos in which there are periods of time where nothing is "happening." However, as previously discussed, the Video-SCOUT environment contains few action events. The robot's journey only shows still objects, and thus, new criteria must be defined.

We propose a new challenge task in video understanding on long-form videos which lack canonical 'events.' By reframing what it means to identify an 'event,' we instead analyze the environment depicted within, according to a high-level assessment question. **Non-event Oriented Video Assessments (NOVA)** is designed as a video frame retrieval experiment in which frames highly relevant to NOVA-questions are selected by an algorithm. We measure the success of this task with precision—the number of selected frames relevant to the assessment—as well as recall—the amount of relevant evidence selected from the environment of all possible relevant evidence. Video-SCOUT supplies videos for the NOVA-questions assigned to the participants: "Is there anything that indicates the environment has recently been occupied?", "Were the last occupants speakers of English or a foreign language?", and "Is there anything that you could use to coordinate operations or activities in a headquarters type environment?"

4.1 Approach

We apply two approaches for video frame retrieval tailored to NOVA-questions. The first is Aha![3], the online highlight detection algorithm from Chang et al. (2025) which achieved 91.6% in top-5 mAP (mean Average Precision) with no fine-tuning on large-scale YouTube datasets, outperforming other approaches at its time of writing. There is no formal definition of an event or activity that the algorithm looks for, providing a solid candidate for our assessment task. The algorithm assigns a relevance score to each frame with respect to the video's title, and returns a list of frames above a certain threshold. Aha! operates sequentially, not needing access to future frames to make its determination. These frames together constitute Aha!'s *Selected Frame Album* for a particular video given a NOVA-question. We ran Aha! using light modifications of the NOVA-questions as the natural language input (exact wording in Appendix B) on an NVIDIA RTX 4090.

The second approach is a general VLM, Google Gemini 2.5 Pro[4]. We give as input the full-length video, and prompt it to list timestamps of frames supporting the NOVA-question (prompt in Appendix B).[5] The authors of this paper then extracted the frames from the provided timestamps to create Gemini's *Selected Frame Albums*. Figure 1 shows a subset of selected frames, with complete albums in Appendix C.

4.2 Annotation

Annotation was conducted by the authors of this paper after familiarization with the SCOUT environment and NOVA-questions. To begin, we created an inventory of every object and area within the experiment environment that could reasonably be used to support each NOVA-question, e.g., the table with office chair and newspapers in Figure 1a might support the assessment that the space was used as a headquarters. The total inventory of these objects and areas is listed in Appendix D. As the NOVA-questions are somewhat subjective, annotation was more lenient and inclusive of different possibilities. When participants answered the questions after their trial, they were not scored on if they accounted for all possible evidence, only that they came to a conclusion. For example, when answering whether the participant believed the environment had been recently occupied, one participant answered, "Yes. There is a box of cereal in one room." Our inventory counted the cereal box and several other objects as appropriate evidence.

[3]Model accessed from publicly available repository at https://github.com/aiden200/Aha-/tree/rebuttal and system defaults accepted.

[4]Model accessed between February–March 2026 using Ask Sage and the Ask Sage Persona with temp=0.

[5]At the time of writing, VLMs are unable to process videos in an online manner, unlike Aha!. We discuss the implications of implementation further in Section 4.4, and focus here on testing the relevant frame identification capabilities.

(a) *Relevant* frame for Occupied, Language, and Headquarters (table, chair, cup, newspapers, bottle).

(b) *Relevant* frame for Occupied (plant). *Distractor* frame for Language and Headquarters.

(c) *Distractor* frame for Occupied, Headquarters, and Language.

Figure 1: Selected frame examples and annotations for NOVA-Questions

Instead of our NOVA algorithms providing the *conclusion* in a natural language statement as the participants did, we designed our experiment to exhaustively analyze the environment for *any* possible supporting evidence, taking the form of a set of selected frames. Annotation of the *Selected Frame Albums* was conducted by one annotator (an author of the paper) after the inventory was finalized through group discussion, and was verified by another author. Each frame in Aha! and Gemini's *Selected Frame Albums* were assigned a label based on the NOVA-question: *Relevant* if at least one piece of evidence in that frame was present (e.g., Figure 1a is *Relevant* for all NOVA-questions), or *Distractor* if no evidence was present in that frame (e.g., Figure 1b is a *Distractor* for the Language and Headquarters NOVA-questions, and Figure 1c is a *Distractor* for all NOVA-questions).

To determine the added value each algorithm provides to a human's analysis, it is necessary to measure what the human reasonably could have deduced on their own. To calculate this, the visual content of the participant's photo requests was examined. This represented the total possible knowledge of the environment the participant could have obtained by the end of their trial, because they did not have access to the live video stream. A *Relevant* or *Distractor* label was assigned to the photos regarding each NOVA-question.

Within each *Relevant* frame in the *Selected Frame Albums* and in the participant's photo requests, the annotator indicated which items or areas provided evidence for the NOVA-question from the assessment inventory. Finally, the annotator reviewed each full-length robot video to count the maximum number of observed evidence within that trial. This was conducted to avoid penalizing recall if the robot never passed by particular evidence cues. In this way, if a participant never instructed the robot to explore the room with the table with office chairs, the *Selected Frame Albums* and photo request annotations would not be penalized for not having seen this space. The maximum evidence score was used to compute a customized recall of the *Selected Frame Albums* and the participant's requested photos.

Cohen's Kappa was computed for the *Relevant* and *Distractor* frame annotation within the Participant Photo Requests. Across all NOVA-questions, $\kappa = 0.74$. Agreement varied by NOVA-question (Occupied $\kappa = 0.97$; Language $\kappa = 0.62$; Headquarters $\kappa = 0.49$). Adjudication revealed discrepancies in overlooking items that were mutually agreed to be relevant, e.g., wall signs appearing in multiple frames. Additionally, Cohen's Kappa was computed for the object inventory within each *Relevant* frame from the participant's photo requests. Across all NOVA-questions, $\kappa = 0.85$. Again, agreement varied by NOVA-question (Occupied $\kappa = 0.91$; Language $\kappa = 0.61$; Headquarters $\kappa = 0.92$). Adjudication revealed challenges pertaining to the Language task in accounting for illegibility due to photo noise, distance from the camera, dark photos, and bad angles, i.e., edge-on shots. A consensus was reached regarding edge-on shots: regardless of how powerful a computer vision algorithm or human eyesight, the text cannot be read if it is not properly in the frame, therefore frames with, for example, the calendar shown from the side perspective, were not counted. Regarding the required quality of the text for legibility, annotation was based on human-determination rather than how good the algorithms may be at OCR (Optical Character Recognition). Because the models were only returning the frame, it is up to the human (the annotators or a future user) to determine its useful-

ness. Furthermore, the Language task only asked to determine the language, not fully read the text; therefore, if an annotator determined enough letters or characters were legible, the text was counted. We discuss this more in the Limitations section, and frame annotations with visuals are given in Appendix C. Other discrepancies in annotation were agreed to be oversights, and annotation was revised on both *Relevant* frames and object inventory until agreement was reached. Subsequently, with the new adjudicated guidelines, annotation was verified and adjusted on Aha! and Gemini's *Selected Frame Albums*.

4.3 Results

We examined the 20 main trials from SCOUT's Experiment 1 (ten participants completed two main trials each). A *Selected Frame Album* was created for each NOVA-question (Occupied, Language, Headquarters) for each approach (Aha!, Gemini). This resulted in a total of 60 *Selected Frame Albums*.

	Part. Photos	**Aha!**	**Gemini**
% Relevant Frames	57.23	63.43	90.21
Occupied	79.27	85.53	100
Language	39.70	40.87	89.86
Headquarters	52.73	63.89	80.77
% Distractor Frames	42.77	36.57	9.79
# Photos Requested or Album Length	33.9	12.73	6.03

Table 2: Average precision of participant photos and *Selected Frame Albums*

Table 2 shows the average percent of *Relevant* and *Distractor* annotations. Of participant photos (first column), 57.23% were relevant, with scores varying based on the NOVA-question. These scores match with observations from prior work in which participants used photo requests to answer NOVA-questions and for navigation (refer to Section 2.1 and Lukin et al. (2023) for photo strategies). The other columns show Aha! and Gemini's *Selected Frame Albums* scores. Across all NOVA-questions, Gemini showed high precision: 90.21% of the frames selected by Gemini were annotated as containing relevant evidence, whereas Aha! averaged 63.43% precision. Within NOVA-questions, both Aha! and Gemini saw drastic differences. For Aha!, the Language NOVA-question had the lowest precision score (40.87%), compared to the Occupied and Headquarters NOVA-questions (85.53% and 63.89% respectively). Meanwhile, Gemini achieved 100% precision on the Occupied NOVA-question, compared to 80.77% on the Headquarters NOVA-question and 89.86% on the Language NOVA-question. The two approaches differed further in the length of the *Selected Frame Albums*, with Aha! selecting an average of 12.73 frames from the full-length video, and Gemini an average of 6.03 frames. On average, 33.9 photos were taken by participants.

Table 3 is organized around the individual markers of evidence, measuring recall conditioned on the total number of evidence items that could possibly be found within the full-length video. The first column in Table 3 shows the percentage of evidence found in the participant photo requests. Across all NOVA-questions, the average recall is high at 87.31%, representing the highest possible score the participant could have achieved without having access to the full robot video. Columns *Aha!* and *Gemini* report their respective *Selected Frame Albums* relevance recall. On their own, recall is drastically lower than the participant photos, with Aha! achieving an average recall of 59.50%, and Gemini 54.42%. We interpret the participant's high recall as a result of taking photos to aid navigation, whereas Aha! and Gemini were tasked only to complete the assessments. The participant photos exhibit a trade-off in high recall at the cost of lower precision, whereas Aha! and Gemini are more balanced.

The columns *Part. Photos + Aha!* and *Part. Photos + Gemini* in Table 3 report how much of a recall boost could have been achieved if the participant's photo requests were combined with the automatically created *Selected Frame Albums*. These were computed by taking the unique complement of the object and area inventory between the participant's photo requests and the approaches' *Selected Frame Albums*. This evidence complement yields gains of 2.39% for Aha! and 2.35% for Gemini averaged across NOVA-questions, showing that these *Selected Frame Albums* contribute a small set of unique evidence not observed by the participant in their exploration. As an example of this, in one trial, a participant instructed the robot to move down a hallway, and a sleeping bag and a movie poster were passed in the process. The participant did not photograph these mid-movement views, but these frames were selected from the full video by Aha! and by Gemini respectively.

% Relevance Recall	Part. Photos	Aha!	Gemini	Part. Photos + Aha!	Part. Photos + Gemini
All Assessments	**87.31**	**59.50**	**54.42**	**89.70 (+2.39)**	**89.66 (+2.35)**
Occupied Assessment	89.37	67.62	57.39	92.71 (+3.34)	89.93 (+0.56)
Language Assessment	79.17	52.71	56.16	82.24 (+3.07)	84.89 (+5.72)
Headquarters Assessment	93.39	58.18	49.71	94.16 (+0.77)	94.16 (+0.77)

Table 3: Evidence recall from the participant's photos and Aha! and Gemini's *Selected Frame Albums* (first three columns). The new recall percentage achieved by combining the unique instances of evidence found in participant's photos and the *Selected Frame Albums*, with gains in parentheses (fourth and fifth columns.)

4.4 Discussion

Model Performance. A precision-recall trade-off manifested in our evaluation. Of Gemini's *Selected Frame Albums*, 90.21% of frames were relevant to the NOVA-question. However, and possibly because it only extracted an average of 6.03 frames, its ability to capture all possible evidence in the video was lower, only 54.42%. The metrics leaned differently for Aha! It achieved a lower average precision of 63.43% and recall of 59.50%. The Language NOVA-question was particularly challenging for Aha! (40.87% precision), whereas Gemini achieved 89.86% precision, showing significant ability to identify texts in the videos. On the other hand, Aha! achieved almost 5% higher recall than Gemini on average, and in particular, about 10% recall increase on the Occupied and Headquarters NOVA-questions. Despite these trade-offs in recall and precision, the models' overall gains in supplemental evidence of participant photos is comparable: 2.39% for Aha! and 2.35% for Gemini. This suggests there may be different opportunities for employing Aha! or Gemini for the NOVA task, given that they can perform well under different conditions for different NOVA-questions.

We outline several strategies to increase and balance precision and recall. Because Gemini's *Selected Frame Album* length was short, future iterations could consider chunking the 20-minute video into smaller segments so that its high precision may still be achieved while increasing recall by combining the results of the chunked videos. Some assessments may be harder for the model to reason about than others; for example, it scored lower in deducing what constituted as a headquarters compared to identifying visible written language.

A key strength of Aha! is its adaptability without fine-tuning. While it achieved a remarkable 91.6% on the event-centric video datasets using the accompanying video titles reported in its paper (Chang et al., 2025), it scored high in *Distractors* within the SCOUT domain (36% of Aha!'s frames were annotated as *Distractor*). To maintain its moderate recall on the Occupied and Headquarters NOVA-questions, we posit that *Distractors* may be minimized with an additional filtering process tailored or fine-tuned to the SCOUT domain. Observationally, most of Aha!'s *Distractors* focused on empty spaces, such as Figure 1c. In both models, the Language NOVA-question scored lower than the others, which, as we discussed in Section 4.2 and more in the Limitations section, comes with legibility challenges.

New HRI Challenges. The identification of frames in non-event videos introduces new opportunities for human-robot collaboration. The high recall scores of both Aha! and Gemini is encouraging and may enable both real-time assisted exploration and retrospective exploration analysis. In the former, it will be critical to minimize disruption and maintain or increase the human's trust in the robot if it is providing real-time alerts on locations possibly relevant to the assessment. A robot with high *Distractors* could quickly erode trust, and after a few incorrectly flagged frames, a human may ignore the robot's suggestions as it interrupts their attempt to complete the assessment. However, with adequate precision and recall checks, a future experiment could allow the robot to take the initiative and offer to stop and look.

There remain gaps in running both Aha! and Gemini in real time. Aha! requires considerable processing power that may be unavailable onboard a robot, and Gemini cannot process videos fully in real time. By contrast, both models would serve well in retrospective exploration analysis. A future experiment could provide a human the *Selected Frame Albums* in addition to their photo requests for greater post-exploration recall. Additionally, it remains to be seen whether a high number of *Distractors* may be less critical in a post-experiment review where real-time interruption is not an issue.

5 Future of Human-Robot Collaboration

NOVA represents a foundational step in supporting human-robot collaboration through video understanding. By enabling a robot to analyze its surroundings with respect to a human's goals, the robot is poised to understand other elements of the collaboration at a more granular level. While the NOVA-question guided the entirety of the *Selected Frame Album* creation, the human is involved in the robot's journey every step of the way by speaking to the robot about what they want to do. There are often cases of a mismatch between the human's language about the environment and what the robot is actually seeing.

We envision future human-robot paradigms which leverage the robot's understanding of a human's instructions in close combination with its visuals to identify when common ground may be lost and to seek strategies to prevent or mitigate it. We thus begin to develop a new challenge leveraging the NOVA framework and the Video-SCOUT dataset called the **Common Ground Alignment Problem (Common-GAP)**. Common-GAP is a decision problem for proactive multimodal repair in bandwidth-limited human-robot exploration. Given the robot's video, the dialogue history, and current dialogue structure annotation, we propose three different challenges for the robot: i) predict misalignment, ii) resolve reference ambiguity, and iii) take initiative. In i), the robot should preemptively detect misalignment in what the human and the robot believe about the world, such as the presence or absence of an object mentioned. In ii) the robot should instigate effective disambiguation and repair strategies to restore common ground. Finally, in iii) the robot should learn when to take initiative and proactively report an observation based on the dialogue history and its continuous video feed. We plan to leverage Video-SCOUT and in particular, the TU Video Clips. See Appendix E for full examples of each decision state.

6 Conclusion

We propose a new challenge in video understanding specific to human-robot collaboration: Non-event Oriented Video Assessment (NOVA). This challenge utilizes our novel Video-SCOUT dataset of 60 robot-recorded, long-form videos with accompanying dialogue transcripts and dialogue structure video clips. Our experiments show promise in using VLMs on out-of-domain videos without fine-tuning, yet reveal gaps in implementation within practical applications. We begin to develop a future task leveraging NOVA and the dialogue structure and video affordances of Video-SCOUT to address the Common-GAP in continuous, human-robot exercises. We invite the community to use this data and contribute algorithms to these challenges.

Limitations

While we designed the Aha! and Gemini prompts to be as similar to the NOVA-Questions as possible, dissimilarities arose in the required input format for each model. Additionally, variants in prompts (e.g., instructions or rephrases of the NOVA-questions) were not exhaustively tested for improved performance. We did not provide examples to the models, operating instead in a zero-shot setting, and thus performance may be improved through future iterations. The full set of inputs and prompts are listed in Appendix B. Furthermore, we did not specify or fine-tune the number of frames for Aha! and Gemini to select. This made the comparison challenging, as Aha! selected approximately twice as many frames as Gemini. We consider what new evaluation metrics are appropriate for this evaluation that are sensitive to the number of frames selected as well as the fact that they are unordered and require different ways to reward prioritization.

Adjudication revealed challenges in assigning annotation in the Language NOVA-question. Aha! and Gemini were only instructed to retrieve the frame or timestamp of a frame to answer the question. We cannot infer understanding of the text visible in the frame on behalf of the model since that was not the question asked of it. In future work, these models may be prompted differently to transcribe the text in images they retrieve to assess model legibility vs. human annotator legibility.

Acknowledgments

This work was supported in part by the U.S. Army Research Laboratory under the Advanced Research Technology, Inc. contract.

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. 2008. Robust real-time unusual event detection using multiple fixed-location monitors. *Transactions on Pattern Analysis and Machine Intelligence*, 30(3):555–560.

Claire Bonial, Stephanie M Lukin, Mitchell Abrams, Anthony Baker, Lucia Donatelli, Ashley Foots, Cory J Hayes, Cassidy Henry, Taylor Hudson, Matthew Marge, Kimberly A. Pollard, Ron Artstein, David Traum, and Clare Voss. 2025. Human–robot dialogue annotation for multi-modal common ground. *Language Resources and Evaluation*, 59(2):1525–1575.

Shawn Bowser, Cynthia Matuszek, and Stephanie Lukin. 2025. Towards integrated multimodal interaction: Merging immersive 3D worlds with language based retrieval for 3D scene understanding. In *Proceedings of the 6th Workshop on Intelligent Cross-Data Analysis and Retrieval*, pages 32–37.

Aiden Chang, Celso de Melo, and Stephanie M. Lukin. 2025. Aha—predicting what matters next: Online highlight detection without looking ahead. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Erin K Chiou, Mustafa Demir, Verica Buchanan, Christopher C Corral, Mica R Endsley, Glenn J Lematta, Nancy J Cooke, and Nathan J McNeese. 2022a. Towards human–robot teaming: Tradeoffs of explanation-based communication strategies in a virtual search and rescue task. *International Journal of Social Robotics*, 14(5):1117–1136.

Manolis Chiou, Georgios-Theofanis Epsimos, Grigoris Nikolaou, Pantelis Pappas, Giannis Petousakis, Stefan Mühl, and Rustam Stolkin. 2022b. Robot-assisted nuclear disaster response: Report and insights from a field exercise. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4545–4552. IEEE.

Hareesh Chitikena, Filippo Sanfilippo, and Shugen Ma. 2023. Robotics in search and rescue (SAR) operations: An ethical and design perspective framework for response phase. *Applied Sciences*, 13(3):1800.

Herbert H. Clark and Catherine R. Marshall. 1981. Definite reference and mutual knowledge. In *Elements of Discourse Understanding*. Cambridge University Press.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10.

Daniel S Drew. 2021. Multi-agent systems for search and rescue applications. *Current Robotics Reports*, 2(2):189–200.

Bernardo Esteves Henriques, Mirko Baglioni, and Anahita Jamshidnejad. 2024. Camera-based mapping in search-and-rescue via flying and ground robot teams. *Machine Vision and Applications*, 35(5):117.

Odair Fernandes, Robin Murphy, David Merrick, Justin Adams, Laura Hart, and Jarrett Broder. 2019. Quantitative data analysis: Small unmanned aerial systems at Hurricane Michael. In *2019 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 116–117. IEEE.

Felix Gervits, Gordon Briggs, Antonio Roque, Genki A Kadomatsu, Dean Thurston, Matthias Scheutz, and Matthew Marge. 2021. Decision-theoretic question generation for situated reference resolution: An empirical study and computational model. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 150–158.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. 2017. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5842–5850.

Vimukthi Jayawardene, Thomas J Huggins, Raj Prasanna, and Bapon Fakhruddin. 2021. The role of data and information quality during disaster response decision-making. *Progress in Disaster Science*, 12:100202.

Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Hannah Recknor, Jimena Guallar-Blasco, Alexander Martin, Eugene Yang, and Benjamin Van Durme. 2025. MultiVENT 2.0: A massive multilingual benchmark for event-centric video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24149–24158.

Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. 2013. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32.

Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal event detection at 150 FPS in Matlab. In *2013 International Conference on Computer Vision (ICCV)*.

Stephanie Lukin, Claire Bonial, Matthew Marge, Taylor A Hudson, Cory Hayes, Kimberly Pollard, Anthony Baker, Ashley N Foots, Ron Artstein, Felix Gervits, Mitchell Abrams, Cassidy Henry, Lucia Donatelli, Anton Leuski, Susan G. Hill, David Traum, and Clare Voss. 2024. SCOUT: A situated and multimodal human-robot dialogue corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14445–14458.

Stephanie M Lukin, Kimberly A Pollard, Claire Bonial, Taylor Hudson, Ron Artstein, Clare Voss, and David Traum. 2023. Navigating to success in multi-modal human-robot collaboration: Analysis and corpus release. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1859–1865. IEEE.

Weixin Luo, Wen Liu, and Shenghua Gao. 2017. A revisit of sparse coding based anomaly detection in stacked RNN framework. In *2017 International Conference on Computer Vision (ICCV)*.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, and 5 others. 2024. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498.

Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. 2016. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 International Conference on Intelligent Robots and Systems (IROS)*.

Kate Sanders, Reno Kriz, David Etter, Hannah Recknor, Alexander Martin, Cameron Carpenter, Jingyang Lin, and Benjamin Van Durme. 2024. Grounding partially-defined events in multimodal data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15905–15927.

Kate Sanders and Benjamin Van Durme. 2024. A survey of video datasets for grounded event understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7314–7327.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10740–10749.

Alan F Smeaton, Paul Over, and Wessel Kraaij. 2006. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. TVSum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5179–5187.

Jinhwan Sul, Jihoon Han, and Joonseok Lee. 2023. Mr. HiSum: A large-scale dataset for video highlight detection and summarization. In *Advances in Neural Information Processing Systems*, volume 36, pages 40542–40555.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

David Traum, Cassidy Henry, Stephanie Lukin, Ron Artstein, Felix Gervits, Kimberly Pollard, Claire Bonial, Su Lei, Clare Voss, Matthew Marge, Cory J. Hayes, and Susan G. Hill. 2018. Dialogue structure annotation for multi-floor interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Gongcheng Wang, Weidong Wang, Pengchao Ding, Yueming Liu, Han Wang, Zhenquan Fan, Hua Bai, Zhu Hongbiao, and Zhijiang Du. 2023. Development of a search and rescue robot system for the underground building environment. *Journal of Field Robotics*, 40(3):655–683.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

A Transcript files

Below are the TU video clip subtitle files for the TUs in the subtitle excerpt presented in Section 2.2.

TU14:

```
0
00:00:00.00 --> 00:00:05.00
CMD: "can you move several yards towards the
white door"
1
00:00:07.63 --> 00:00:12.63
Robot: "processing. . ."
2
00:00:38.64 --> 00:00:43.64
Robot: "I will move forward 6 feet, ok?"
3
00:00:45.15 --> 00:00:50.15
CMD: "uh i think six feet is too fff far"
4
00:00:49.98 --> 00:00:54.98
CMD: " maybe three feet"
5
00:00:53.97 --> 00:00:58.97
Robot: "ok"
6
00:01:16.06 --> 00:01:21.06
Robot: "moving. . ."
7
00:01:22.45 --> 00:01:27.45
Robot: "done"
```

TU15:

```
0
00:00:00.00 --> 00:00:05.00
CMD: "can you take a photo"
1
00:00:04.91 --> 00:00:09.91
Robot: "sent"
```

B Natural Language Inputs and Prompts

We crafted natural language inputs and prompts to closely preserve the wording given to participants in the SCOUT experiments, which are reprinted below:

- "Is there anything that indicates the environment has recently been occupied?"
- "Were the last occupants speakers of English or a foreign language?"
- "Is there anything that you could use to coordinate operations or activities in a headquarters type environment?"

The Aha! natural language input utilized a randomly selected template from the following list:

- "'[NOVA-Question-Aha]."'
- "What segment of the video addresses the topic '[NOVA-Question-Aha]?"'
- "At what timestamp can I find information about '[NOVA-Question-Aha]' in the video?"
- "Can you highlight the section of the video that pertains to '[NOVA-Question-Aha]?"'
- "Which moments in the video discuss [NOVA-Question-Aha] in detail?"
- "Identify the parts that mention '[NOVA-Question-Aha]."'
- "Where in the video is [NOVA-Question-Aha] demonstrated or explained?"
- "What parts are relevant to the concept of '[NOVA-Question-Aha]?"'
- "Which clips in the video relate to the query '[NOVA-Question-Aha]?"'
- "Can you point out the video segments that cover '[NOVA-Question-Aha]?"'
- "What are the key timestamps in the video for the topic '[NOVA-Question-Aha]?"'

The variable [NOVA-Question-Aha] was abbreviated and reworded from the SCOUT assessments to fit these predetermined templates, and was selected from the following for the appropriate video assessment:

- "the environment has been recently occupied"
- "the written language"
- "evidence of a meeting headquarters"

The Gemini prompt was as follows: "You are an expert video analyst. Review the video and create a list of frames that support the provided hypothesis. Output your results as a list of timestamps. Hypothesis: [NOVA-Question-Gemini]."

The variable [NOVA-Question-Gemini] was selected from the following for the appropriate video assessment:

- "The environment has been recently occupied"
- "The last occupants were speakers of English or a foreign language"
- "There is something which could be used to coordinate operations or activities in a headquarters type environment"

C Selected Frame Albums

Figures 2 and 3 show Gemini and Aha!'s *Selected Frame Albums* for p1.08's main1 trial on the Language NOVA-question.

D Relevant Inventory Lists

Occupied relevant inventory list (27 observations): shoes, cooking items, shopping bag, plants, newspaper, solo cup on chair, conference table, chairs around table, clock, bottle water, calendar, map, monitor, desk, desk chair, sleeping bag, luggage, clothes, cleaning supplies, posters, TV, books by TV, wall signs, construction items, fire extinguisher, water cooler jug, stop sign.

Language relevant inventory list (16 observations; must be legible): cereal box, newspaper, calendar, map, posters, books by TV, room numbers, no smoking sign, plywood writing, cleaning supplies, yellow caution cone, stop sign, blue wall sign, luggage logo, broom logo, orange bucket.

Headquarters relevant inventory list (16 observations): cooking items, newspaper, solo cup on chair, conference table, chairs around table, clock, bottle water, calendar, map, monitor, desk, desk chair, cleaning supplies, wall signs, fire extinguisher, water cooler jug.

E Common-GAP Examples

In Section 5, we proposed a challenge task in which we seek to leverage the robot's understanding of a

(a) Frame at timestamp 00:00. *Distractor* frame (orange bucket not legible).

(b) Frame at timestamp 04:22. *Relevant* frame (yellow caution cone; cleaning supplies not legible).

(c) Frame at timestamp 10:39. *Relevant* frame (cleaning supplies).

(d) Frame at timestamp 13:25. *Relevant* frame (wall sign).

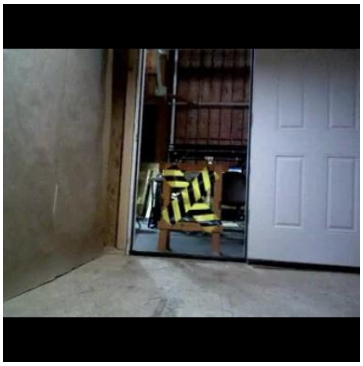
(e) Frame at timestamp 18:21. *Relevant* frame (wall sign).

(f) Frame at timestamp 20:40. *Relevant* frame (blue wall sign).

Figure 2: *Selected Frame Album* assembled by Gemini for p1.08's main1 trial on the Language NOVA-question

(a) Frame at timestamp 02:32. *Distractor* frame (yellow caution cone and orange bucket not legible).

(b) Frame at timestamp 04:18. *Distractor* frame (cleaning supplies not legible).



(c) Frame at timestamp 05:01. *Distractor* frame.

(d) Frame at timestamp 19:51. *Distractor* frame.

(e) Frame at timestamp 20:40. *Relevant* frame (blue wall sign).

Figure 3: *Selected Frame Album* assembled by Aha! for p1.08's main1 trial on the Language NOVA-question

human's instructions in close combination with its visuals to identify when common ground may be lost. We hypothesize that we can use the live robot visual along with the human instruction to automatically detect issues with the instruction. Specifically, we aim to enable i) prediction of misalignment in what the user and the robot believe about the world, ii) resolving reference ambiguity, and iii) taking initiative and proactively reporting an observation. This analysis extends beyond referring expression resolution—the robot must recognize whether or not there is a clear, ambiguous, or even incorrect reference to the environment, and then work within that space or attempt to re-establish common ground. We refer to this challenge to detect and repair as the **Common-GAP: Common Ground Alignment Problem**. Common-GAP is a decision problem for proactive multimodal repair in bandwidth-limited human-robot exploration. Given the current robot video, the dialogue history, and dialogue-structure annotations, decide whether to:

a) execute the command directly,

b) ask a clarification question,

c) offer a photo,

d) send a photo without waiting, or

e) explain why the command cannot yet be grounded.

In an initial exploration of Common-GAP, we manually reviewed the 60 transcripts from SCOUT's Experiments 1 and 2, and selected three reference episodes that parallel the system capabilities described above that we hope to enable. We then viewed the TU video clips alongside the dialogue. We observe cases where the collaborative exploration flows fairly smoothly despite a mismatch in common ground in the form of fully shared visual information. For example, in the following case, the participant would like to enter a room ahead, but there is a road cone blocking the entrance. The offer and subsequently provided visual information make this clear to the participant, who then quickly moves on to explore the other doorways in the environment.

p1.01 training trial, TU 4:

```
CMD: "go in the room"
Robot: "There's an obstruction at the entrance.
  Would you like me to send a picture?"
CMD: "send a picture"
Robot: "sent"
```

p1.01 training trial, TU 5:

```
CMD: "go to the other door"
Robot: "executing..."
Robot: "done"
```

This example showcases how the data might support enabling i): predicting misalignment in what the user and the robot believe about the world, e.g., presence or absence of an object.

We also observe cases where there are longer-horizon miscommunications that persist over several TUs due to the complexity of the physical environment. This is especially prevalent as participants explore areas with multiple adjacent doorways, as is the case in the following example. Again, the supplied visual information enables, over the course of five TUs, the participant to successfully move through the appropriate, disambiguated doorway.

p2.06 main1 trial, TU 4:

```
CMD: "turn around one hundred and eighty degrees
CMD: "and travel through the door"
Robot: "processing..."
Robot: "I will turn around 180 degrees"
Robot: "but..."
Robot: "I'm not sure which doorway you are
  referring to."
```

p2.06 main1 trial, TU 5:

```
Robot: "Should I send a picture?"
CMD: "yes"
Robot: "done, sent"
```

p2.06 main1 trial, TU 6:

```
CMD: "travel straight down the hallway"
Robot: "Which doorway?"
CMD: "four feet"
Robot: "ok. moving..."
Robot: "done"
```

p2.06 main1 trial, TU 7:

```
CMD: "take a picture"
Robot: "sent"
```

p2.06 main1 trial, TU 8:

```
CMD: "travel to the end of the hallway"
CMD: "and enter the doorway on the right"
Robot: "processing..."
Robot: "moving..."
Robot: "done"
```

Thus, this example demonstrates how the data might be used to learn ii): resolving reference ambiguity by instigating effective disambiguation and repair strategies. The above situation and others like it suggest that one strategy may be to automatically provide visual information when there are multiple of the same type of referent in the visual field, such as clustered doorways.

Furthermore, there are cases where automatically sending visual information would be helpful as some operators do not accept offers in dialogue, even when it would be helpful. This is the case in the next interaction, where the participant declines the offer for a picture:

p2.08 training trial, TU 3:

```
CMD: "go ahead.
CMD: "we're looking for doorways"
Robot: "Hmm. . ."
CMD: "go ahead. move for"
```

p2.08 training trial, TU 4:

```
Robot: "Would you like me to send a picture?"
CMD: "no thank you not right now"
```

This user then goes on to struggle with the interaction as exhibited by the fact that the operator issues two more ambiguous, unactionable commands (not shown in the exchange above). This operator only successfully moves to the desired location later, notably after requesting a picture of the environment. Subsequently, the operator leverages a clear pattern of move instructions followed immediately by requests for images, demonstrating the efficacy of this strategy for this particular operator. Thus, this example showcases both the importance and complexity of capability iii): learning when to take initiative and proactively report an observation. From the above example, we see that different kinds of users will react distinctly to different levels of proactive behaviors from the robot—different types of strategies must be leveraged with different operators to supply critical information in a way that maintains conversational norms and expectations.

In developing the Common-GAP and identifying episodic examples from Video-SCOUT, we plan to outline metrics for success, including the time it takes to overcome misunderstanding using the different strategies, to avoid it in the first place, and establishing the threshold for when the robot should employ a strategy. We plan to test different models and ablations (with LLMs, VLMS; with and without domain knowledge, dialogue history, etc.) implementing the Common-GAP on the Video-SCOUT TU clips to detect miscommunication at the earliest point. We will design a human-robot experiment in which such strategies are employed in real time, and human performance and perceptions of the communication are collected.