

7 Appendix

7.1 Waypoint Selection Experiments

We perform preliminary experiments across various VLMs which explores providing scene understanding as context for the task of waypoint selection. We construct this VQA task by using future goal points as ground-truth and generating various annotated waypoints [18, 45] where the task is to select the annotated waypoint that makes progress towards the goal while being considerate of the humans in the scene. We condition the prompt on social reasoning context in the form of how the agents in the scene are interacting, which comes directly from answers on SOCIALNAV-SUB. An example of this VQA task is shown in Figure 4. The results for the experiments are shown in Table 3. Overall, when provided with scene understanding context extracted from the human oracle’s answers, the models have better performance than when no scene context or random scene context is provided. Note that the advantage is insignificant for Gemini 2.5, as its performance is relatively poor with large variance. While preliminary, this experiment shows that more accurate social scene context can provide VLMs with information that is helpful for the models to infer the ground-truth waypoints. It further implies that improving a VLM’s scene understanding capabilities could potentially help it understand the scene context more accurately and then select reasonable navigation actions conditioning on the inferred scene context.



Figure 4: **An example of the waypoint selection VQA task.** This particular example highlights using scene context from the human oracle. Having no context removes the middle portion of the text prompt that includes the context, and having random context randomizes each relational action for the context (such as “avoiding”).

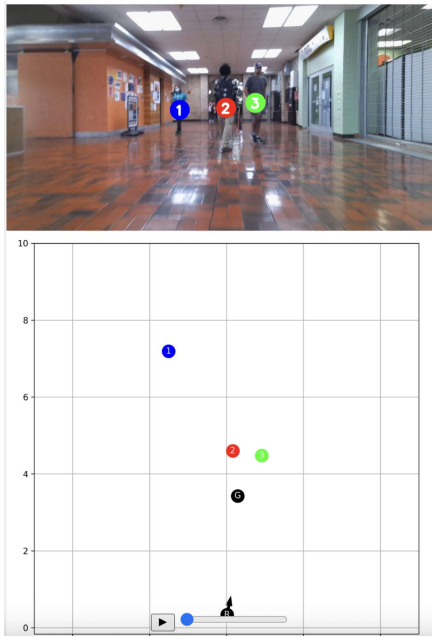
7.2 Human-Subject Study Details

As mentioned in Section 3.4, we conducted a human-subject study under an IRB-approved protocol to collect human data to establish an evaluation method for SOCIALNAV-SUB. We conducted our human-subject study using Prolific [50] with 75 humans participants who were shown 4 scenarios each and provided labels for all questions. The survey contained attention-check questions and were manually inspected to validate data quality. Figure 5 shows an example of the interface the humans

Table 3: Comparison of model performance with and without scene context. Each result is averaged across 3 separate runs. The metric is accuracy of selecting the same waypoint as the human operator. We average accuracy across 3 separate runs and include standard error.

Model	No Scene Context	Random Context	Human Context
Gemini 2.0	58.76% \pm 0.56%	58.76% \pm 0.56%	61.02% \pm 1.69%
Gemini 2.5	37.85% \pm 4.62%	38.42% \pm 3.70%	39.55% \pm 4.62%
o4-mini	35.59% \pm 0.98%	44.07% \pm 1.69%	47.46% \pm 3.53%

were provided for the human-subject study. Humans sequentially answered questions for each scenario in the following order: spatial reasoning questions, spatiotemporal reasoning questions, and social reasoning questions.



Movements

Pages left: 16

Instructions: This series of questions will ask you about the locations and movements of the robot and people shown in the video.

If you can't see the person at the start of the video, please estimate their starting location based on the first seen location. For the end location, please do the same based on the last seen location.

The robot is _____. (Select all that apply)

☒ moving ahead
☐ turning left
☐ turning right

In the beginning, person X is _____ the robot.

	ahead of	to the left of	to the right of	behind
person 1	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
person 2	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
person 3	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

At the end, person X is _____ the robot.

	ahead of	to the left of	to the right of	behind
person 1	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
person 2	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
person 3	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>

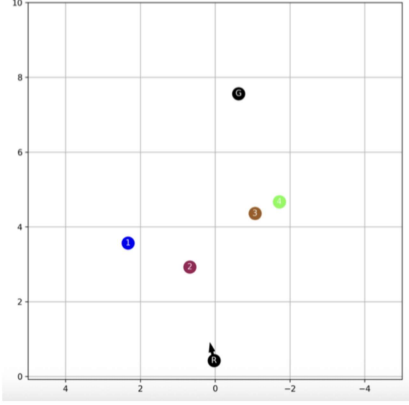
Figure 5: An example of a survey page shown to human participants. Before this, humans were given human-subject study participation instructions, requirements, and instructions about the survey content.

7.3 VQA Prompt Details

To provide fair comparison between humans and VLMs, we provided VLMs with highly similar input. In Figure 6, we provide a full VQA example of what the VLM receives as input. Chain-of-thought reasoning was used in the main experiments outlined in Section 4.3 and this particular usage consisted of sequentially asking the VLM questions, where later questions require higher-level reasoning, and providing the VLM its answers to the relevant questions within the prompt.

7.4 Main Experiment Qualitative Results

As mentioned in Section 4.2, we found cases of all VLMs in the experiment failed on questions with high human consensus in all reasoning categories, especially in cases of high crowd densities; we show these failure cases in Figure 7. We also highlight cases where VLMs can provide success, shown in Figure 8.



Title: Navigating Through Crowds Survey

This is a research survey about walking through crowds. You will be shown sequences of images of a robot navigating around people. The survey consists of multiple choice questions about the behavior of people and robots in a shared space.

Survey Description & Instructions:

The sequences of images are from the robot's perspective. Some visible people in the sequences of images have a unique circled number. The robot is generally moving forward but may turn or change its speed. We will ask you questions about these different sequences of images representing different scenarios.

Instructions: This series of questions will ask you about the locations and movements of the robot and people shown in the sequence of images.

If you can't see the person at the start of the video, please estimate their starting location based on the first seen location. For the end location, please do the same based on the last seen location.

In the beginning, Person 2 is _____ the robot.

Possible answers: "ahead of", "to the left of", "to the right of", "behind"

Please provide the answer to the single question in JSON format (NOT a list) as follows:

`{"answer": "<one of the possible answers>"}`

Ensure the response is in JSON format and includes only one key, "answer"

Figure 6: **An example of a full VQA prompt shown to VLMs.** This context closely resembles the instructions that were provided to human participants for the human-subject study. In addition to the image shown on the left, the VLM also receives the next 9 images in the sequence.



Figure 7: **Examples of failure cases for all VLMs.** *Top-left:* Failing to recognize that person 5 is on the left. *Top-right:* Failing to recognize that person 4 ends up further away. *Bottom-left:* Answering that the distant person 3 should be avoided. *Bottom-right:* Incorrectly answering that an action should be taken with respect to person 7, although all humans did not think they were relevant.



Figure 8: **Examples of success cases for VLMs.** *Top-left:* All VLMs correctly infer that person 1 is not obstructing the path to the goal. *Top-right:* Gemini correctly predicts that person 1 should be avoided (the other VLMs incorrectly predict this). *Bottom-left:* GPT-4o correctly answers that person 5 is on the left, whereas both Gemini and LLaVa-Next-Video answer that person 5 is behind the robot. *Bottom-right:* Most VLMs (but not all) predict that person 6 is being considered as the robot is moving towards the goal, similar to the distribution among human answers.

509 7.5 Survey Question Descriptions

510 Here we show the qualitative descriptions of questions used throughout the benchmark by providing
 511 a question for each VQA prompt, shown in Table 4. We categorize these questions according to
 512 their reasoning capability.

Table 4: **Qualitative descriptions of the text components for questions used in SOCIALNAV-SUB**, their pertaining primary reasoning capability, and the number of unique questions through SOCIALNAV-SUB. All questions are multiple-choice questions, with each VQA prompt providing the possible answers. An example of a VQA prompt can be found in Figure 2 and a full example can be found in Appendix 7.3.

VLM Reasoning Capability	Qualitative Description of Question	# of Questions
Spatial	Person Initial Position: The position of the person at the beginning of the video.	399
	Person Ending Position: The position of the person at the end of the video.	399
	Goal Initial Position: The initial position of the goal with respect to the robot’s view.	60
	Goal End Position: The end position of the goal with respect to the robot’s view.	60
	Person End Goal Obstruction: Whether the person is obstructing the robot’s path towards the goal at the end of the video.	399
Spatiotemporal	Robot Moving Direction: The direction the robot is moving in the video.	60
	Person Distance Change: The relative distance change of the person to the robot from the beginning of the video to the end.	399
	Person Goal Obstruction: Whether the person is obstructing the robot’s path towards the goal during the video.	399
Social	Robot Affected by Person: Whether the robot’s (human operator’s) actions are affected by the person.	399
	Robot Action to Person: The high-level relational action of the robot with respect to the person (e.g., the robot avoided person 2).	399
	Person Affected by Robot: Whether the robot’s (human operator’s) actions are affected by the person.	399
	Person Action to Robot: The high-level relational action of the person with respect to the robot (e.g., person 2 avoided the robot).	399
	Robot Affected by Person at End: Whether the robot’s (human operator’s) actions are affected by the person at the end of the video.	399
	Robot Action to Person at End: The high-level relational action of the robot with respect to the person at the end of the video.	399
	Person Action to Robot at End: The high-level relational action of the person with respect to the robot at the end of the video.	399

7.6 Main Experiment Question Results

We provide the question-level performance for the main experiment results from Section 4.2 for the VLMs shown in Table 5, reasoning-based VLMs shown in 6, and the baselines shown in Table 7.

Table 5: Performance Across Individual Questions for non-reasoning VLMs. These results highlight the deficiencies of non-reasoning VLMs: 1) Gemini [7] has stronger social reasoning than other non-reasoning VLMs for most questions but has worse spatial reasoning performance across most tasks compared to GPT-4o; 2) LLaVa-Next-Video [17] has poor spatial reasoning performance for most questions, determining the moving direction of the robot, and poor ability to infer the future action of the robot, but performs well for certain questions such as determining whether somebody is obstructing the goal and some social reasoning questions; 3) GPT-4o [6] has moderate performance across tasks but lacks strong social reasoning.

Category	Question Name	Gemini 2.0		GPT-4o		LLaVa-Next-Video	
		PA	CW PA	PA	CW PA	PA	CW PA
Spatial	Person Initial Position	0.52 ± 0.01	0.81 ± 0.01	0.54 ± 0.01	0.84 ± 0.01	0.05 ± 0.00	0.10 ± 0.01
	Person Ending Position	0.38 ± 0.01	0.64 ± 0.02	0.43 ± 0.01	0.71 ± 0.02	0.24 ± 0.01	0.44 ± 0.02
	Goal Initial Position	0.69 ± 0.04	0.85 ± 0.04	0.74 ± 0.03	0.92 ± 0.03	0.14 ± 0.02	0.20 ± 0.04
	Goal End Position	0.56 ± 0.04	0.73 ± 0.05	0.65 ± 0.04	0.83 ± 0.04	0.15 ± 0.02	0.22 ± 0.04
	Person End Goal Obstruction	0.74 ± 0.01	0.86 ± 0.02	0.68 ± 0.02	0.78 ± 0.02	0.80 ± 0.01	0.93 ± 0.01
Spatiotemporal	Robot Moving Direction	0.46 ± 0.05	0.64 ± 0.05	0.57 ± 0.04	0.81 ± 0.04	0.24 ± 0.04	0.38 ± 0.05
	Person Distance Change	0.31 ± 0.01	0.53 ± 0.02	0.46 ± 0.01	0.74 ± 0.02	0.47 ± 0.01	0.75 ± 0.02
	Person Goal Obstruction	0.62 ± 0.02	0.76 ± 0.02	0.54 ± 0.02	0.67 ± 0.02	0.74 ± 0.01	0.89 ± 0.01
Social	Robot Affected by Person	0.64 ± 0.02	0.78 ± 0.02	0.50 ± 0.02	0.63 ± 0.02	0.75 ± 0.01	0.91 ± 0.01
	Robot Action to Person	0.51 ± 0.01	0.75 ± 0.02	0.37 ± 0.01	0.57 ± 0.02	0.25 ± 0.01	0.42 ± 0.02
	Person Affected by Robot	0.74 ± 0.01	0.88 ± 0.01	0.58 ± 0.02	0.71 ± 0.02	0.79 ± 0.01	0.94 ± 0.01
	Person Action to Robot	0.62 ± 0.01	0.86 ± 0.02	0.45 ± 0.02	0.65 ± 0.02	0.67 ± 0.01	0.92 ± 0.01
	Robot Affected by Person at end	0.72 ± 0.01	0.87 ± 0.01	0.55 ± 0.02	0.68 ± 0.02	0.79 ± 0.01	0.94 ± 0.01
	Robot Action to Person at end	0.60 ± 0.01	0.85 ± 0.02	0.41 ± 0.02	0.59 ± 0.02	0.08 ± 0.01	0.14 ± 0.01
	Person Action to Robot at end	0.62 ± 0.01	0.87 ± 0.01	0.40 ± 0.02	0.59 ± 0.02	0.03 ± 0.00	0.05 ± 0.01

Table 6: Performance Across Individual Questions for Large Reasoning Models. These results indicate that o4-mini displays worse performance across most spatial reasoning question but has strong performance on determining if a person is obstructing the path to the goal. We hypothesize, with evidence in Appendix 7.1, that better performance in these questions can result in better social reasoning performance and may be a limiting factor for o4-mini. Gemini 2.5 shows worse performance across spatiotemporal reasoning and social reasoning compared to o4-mini but comparable performance in spatial reasoning. Gemini 2.5 has a particularly difficult time in determining the moving direction of the robot compared to other models. Although we evaluated using o4-mini and Gemini 2.5 flash, we expect that these may be lower bounds on the performance for their higher-end model variations.

Category	Question Name	Gemini 2.5		o4-mini	
		PA	CW PA	PA	CW PA
Spatial	Person Initial Position	0.49 ± 0.01	0.78 ± 0.02	0.49 ± 0.01	0.76 ± 0.02
	Person Ending Position	0.36 ± 0.01	0.59 ± 0.02	0.36 ± 0.01	0.58 ± 0.02
	Goal Initial Position	0.70 ± 0.04	0.86 ± 0.04	0.48 ± 0.05	0.58 ± 0.06
	Goal End Position	0.52 ± 0.04	0.67 ± 0.05	0.48 ± 0.05	0.60 ± 0.06
	Person End Goal Obstruction	0.66 ± 0.02	0.77 ± 0.02	0.81 ± 0.01	0.93 ± 0.01
Spatiotemporal	Robot Moving Direction	0.34 ± 0.04	0.50 ± 0.06	0.56 ± 0.04	0.80 ± 0.04
	Person Distance Change	0.47 ± 0.01	0.75 ± 0.02	0.45 ± 0.01	0.71 ± 0.02
	Person Goal Obstruction	0.60 ± 0.02	0.73 ± 0.02	0.73 ± 0.01	0.87 ± 0.01
Social	Robot Affected by Person	0.57 ± 0.02	0.71 ± 0.02	0.73 ± 0.01	0.89 ± 0.01
	Robot Action to Person	0.44 ± 0.02	0.67 ± 0.02	0.58 ± 0.01	0.84 ± 0.02
	Person Affected by Robot	0.70 ± 0.02	0.83 ± 0.02	0.77 ± 0.01	0.91 ± 0.01
	Person Action to Robot	0.58 ± 0.02	0.80 ± 0.02	0.60 ± 0.02	0.83 ± 0.02
	Robot Affected by Person at end	0.56 ± 0.02	0.68 ± 0.02	0.77 ± 0.01	0.91 ± 0.01
	Robot Action to Person at end	0.44 ± 0.02	0.62 ± 0.02	0.62 ± 0.01	0.87 ± 0.01
	Person Action to Robot at end	0.58 ± 0.01	0.81 ± 0.02	0.58 ± 0.02	0.82 ± 0.02

Table 7: **Performance Across Individual Questions for Baselines.** For the Human Oracle and Average Human baselines, these results highlight questions that humans disagreed on more often, showing that determining spatial labels for humans was more disagreeable than social reasoning questions. The rule-based baseline performance indicates that it struggles with determining what the initial and ending position of humans are as well as determining if a person gets closer to further away, showing that it is not as trivial as determining a cutoff value for this from rules described in 7.8.

Category	Question Name	Human Oracle		Average Human		Rule-Based	
		PA	CW PA	PA	CW PA	PA	CW PA
Spatial	Person Initial Position	0.64 ± 0.01	1.00 ± 0.00	0.46 ± 0.01	0.73 ± 0.00	0.49 ± 0.01	0.78 ± 0.01
	Person Ending Position	0.61 ± 0.01	1.00 ± 0.00	0.43 ± 0.01	0.71 ± 0.01	0.41 ± 0.01	0.67 ± 0.02
	Goal Initial Position	0.80 ± 0.02	1.00 ± 0.00	0.68 ± 0.03	0.85 ± 0.01	0.68 ± 0.04	0.83 ± 0.05
	Goal End Position	0.77 ± 0.02	1.00 ± 0.00	0.62 ± 0.02	0.82 ± 0.01	0.56 ± 0.04	0.72 ± 0.05
	Person End Goal Obstruction	0.86 ± 0.01	1.00 ± 0.00	0.77 ± 0.01	0.89 ± 0.00	0.80 ± 0.01	0.93 ± 0.01
Spatiotemporal	Robot Moving Direction	0.69 ± 0.03	1.00 ± 0.00	0.52 ± 0.03	0.74 ± 0.02	0.62 ± 0.04	0.87 ± 0.03
	Person Distance Change	0.63 ± 0.01	1.00 ± 0.00	0.46 ± 0.01	0.74 ± 0.00	0.48 ± 0.01	0.76 ± 0.02
	Person Goal Obstruction	0.83 ± 0.01	1.00 ± 0.00	0.73 ± 0.01	0.88 ± 0.00	0.78 ± 0.01	0.94 ± 0.01
Social	Robot Affected by Person	0.82 ± 0.01	1.00 ± 0.00	0.72 ± 0.01	0.87 ± 0.00	0.76 ± 0.01	0.91 ± 0.01
	Robot Action to Person	0.67 ± 0.01	1.00 ± 0.00	0.50 ± 0.01	0.74 ± 0.01	0.57 ± 0.01	0.82 ± 0.02
	Person Affected by Robot	0.84 ± 0.01	1.00 ± 0.00	0.74 ± 0.01	0.88 ± 0.00	0.79 ± 0.01	0.94 ± 0.01
	Person Action to Robot	0.72 ± 0.01	1.00 ± 0.00	0.56 ± 0.01	0.77 ± 0.01	0.67 ± 0.01	0.93 ± 0.01
	Robot Affected by Person at end	0.84 ± 0.01	1.00 ± 0.00	0.73 ± 0.01	0.88 ± 0.00	0.79 ± 0.01	0.94 ± 0.01
	Robot Action to Person at end	0.70 ± 0.01	1.00 ± 0.00	0.53 ± 0.01	0.75 ± 0.01	0.67 ± 0.01	0.94 ± 0.01
	Person Action to Robot at end	0.71 ± 0.01	1.00 ± 0.00	0.54 ± 0.01	0.76 ± 0.01	0.70 ± 0.01	0.98 ± 0.01

516 7.7 Ablation Experiments

Table 8: **Ablation experiment of querying strategies.** The metric used is Probability of Agreement (PA). The baseline row BEV+CoT represents the VLM’s performance with both CoT and BEV prompts enabled, while the subsequent rows show the effects of removing either CoT or BEV components.

Model	Ablation	Spatial Reasoning	Spatiotemporal Reasoning	Social Reasoning
GPT-4o	CoT+BEV	0.56 ± 0.01	0.51 ± 0.01	0.47 ± 0.01
	No CoT	0.58 ± 0.01	0.53 ± 0.01	0.35 ± 0.01
	No BEV	0.51 ± 0.01	0.44 ± 0.01	0.42 ± 0.01
LLaVa-Next-Video	CoT+BEV	0.35 ± 0.01	0.58 ± 0.01	0.48 ± 0.01
	No CoT	0.35 ± 0.01	0.58 ± 0.01	0.38 ± 0.01
	No BEV	0.35 ± 0.01	0.61 ± 0.01	0.46 ± 0.01
Gemini 2.0	CoT+BEV	0.55 ± 0.01	0.46 ± 0.01	0.63 ± 0.01
	No CoT	0.56 ± 0.01	0.48 ± 0.01	0.58 ± 0.01
	No BEV	0.56 ± 0.01	0.46 ± 0.01	0.64 ± 0.01

517 To understand the impact of specific querying strategies on model performance, we conducted ab-
518 lation experiments, systematically removing components such as chain-of-thought (CoT) reasoning
519 and BEV prompts. Table 8 summarizes how these ablations affect PA in spatial, spatio-temporal,
520 and social reasoning tasks.

521 **CoT reasoning.** The results indicate that removing the CoT component does not significantly affect
522 spatial and spatiotemporal reasoning performance. However, the removal of CoT leads to a notable
523 decrease in social reasoning performance across all models. We hypothesize that social reasoning
524 tasks more often require multi-step reasoning to which CoT can help structure complex chains of
525 inference.

526 **BEV visual prompts.** The results from removing BEV prompts indicate that there is not a signif-
527 icant effect across the capabilities for LLaVa-Next-Video and Gemini 2.0, but provides a notable
528 decrease in performance for GPT-4o across all capabilities. While these results may not indicate a
529 clear winner for all models, it suggests that prompt design remains an open question which needs to
530 be further studied, an endeavor that can be pursued using our benchmark.

Table 9: **Gemini ablation experiments when using ground truth spatial and spatiotemporal answers for CoT reasoning.** Our results indicate that better spatial reasoning and spatiotemporal reasoning leads to better performance on social reasoning questions.

Question Name	CoT		CoT with Ground-Truth Spatial(Temporal) Reasoning	
	PA	CW PA	PA	CW PA
Robot Affected by Person	0.64 \pm 0.02	0.78 \pm 0.02	0.78 \pm 0.01	0.94 \pm 0.01
Robot Action to Person	0.51 \pm 0.01	0.75 \pm 0.02	0.60 \pm 0.01	0.88 \pm 0.01
Person Affected by Robot	0.74 \pm 0.01	0.88 \pm 0.01	0.78 \pm 0.01	0.94 \pm 0.01
Person Action to Robot	0.62 \pm 0.01	0.86 \pm 0.02	0.65 \pm 0.01	0.90 \pm 0.01
Robot Affected by Person at end	0.72 \pm 0.01	0.87 \pm 0.01	0.78 \pm 0.01	0.93 \pm 0.01
Robot Action to Person at end	0.60 \pm 0.01	0.85 \pm 0.02	0.65 \pm 0.01	0.91 \pm 0.01
Person Action to Robot at end	0.62 \pm 0.01	0.87 \pm 0.01	0.65 \pm 0.01	0.91 \pm 0.01

Spatial Reasoning’s Affect on Performance. We ran an additional experiment to see if a lack of strong performance for spatial and spatiotemporal reasoning was affecting performance on social reasoning questions. Table 9 shows the results of running this experiment where we used the human consensus answer’s for the answers for spatial and spatiotemporal reasoning questions for the VLM, which was also provided as chain-of-thought reasoning to the VLM in the form of context; the VLM was then evaluated on social reasoning questions. These results indicate that a strong spatial and spatiotemporal reasoning capabilities can lead to significantly better performance on social reasoning questions. The “Person Goal Obstruction” question may provide sufficient information for the VLM to easily answer the “Robot Affected By Person” question, to which we run an additional experiment and empirically found that, although it was not as drastic, there were performance gains across all questions. These results indicate that hybrid VLM systems that help VLM’s with their weaknesses (such as dedicated perception modules) may be more effective rather than entirely relying on the VLM for all questions.

7.8 Rule-Based Baseline Details

As mentioned in Section 4.1, we developed a rule-based baseline which uses a set of hand-crafted rules to determine answers for VQA questions. Although our simple approach demonstrates better performance than VLMs, it is by no means comprehensive and more complex rules can be devised to further push performance. We briefly summarize the simple rules to determine answers for our Rule-Based baseline:

- **Spatial Reasoning Position Questions:** Determine deviation in the horizontal direction and use it along with cutoff values to determine whether to answer they are to the left, ahead, or behind.
- **Goal Obstruction Questions:** Draw a line from the robot to the goal and a line from the person’s trajectory, if the lines intersect, consider the person obstructing the goal.
- **Person Distance Change:** Look at the initial relative position and end relative positions for the person, determine the appropriate answer based on the distance between the two points.
- **Robot Moving Direction:** Use the horizontal deviation between the the initial relative position of the robot and the end relative position to determine if the robot is turning.
- **Social Reasoning “Affected” Questions:** If the person is obstructing the goal, then answer that the robot will be affected by them.
- **Social Reasoning “Action” Questions:** If the person is obstructing the goal, then avoid the person. For person action questions, use the same answer as the robot action questions.