Asking for Help Using Inverse Semantics

Stefanie Tellex

Ross Knepper, Adrian Li,
Daniela Rus, Nicholas Roy

BROWN


Google
self-driving car
6TNB644


ItemFetch™
OrderFetch™
CaseFetch™


baxter
rethink robotics

How can robots collaborate with people using natural language?

- Following instructions.
  - “Put the metal crate on the truck.”

Symbol Grounding Problem

"The pallet of boxes on the left."

Move the pallet from the truck.

Remove the pallet from the back of the truck.

Offload the metal crate from the truck.

Pick up the silver container from the truck bed.

Move the pallet from the truck.

Remove the pallet from the back of the truck.

Offload the metal crate from the truck.

Pick up the silver container from the truck bed.

Move the pallet from the truck.

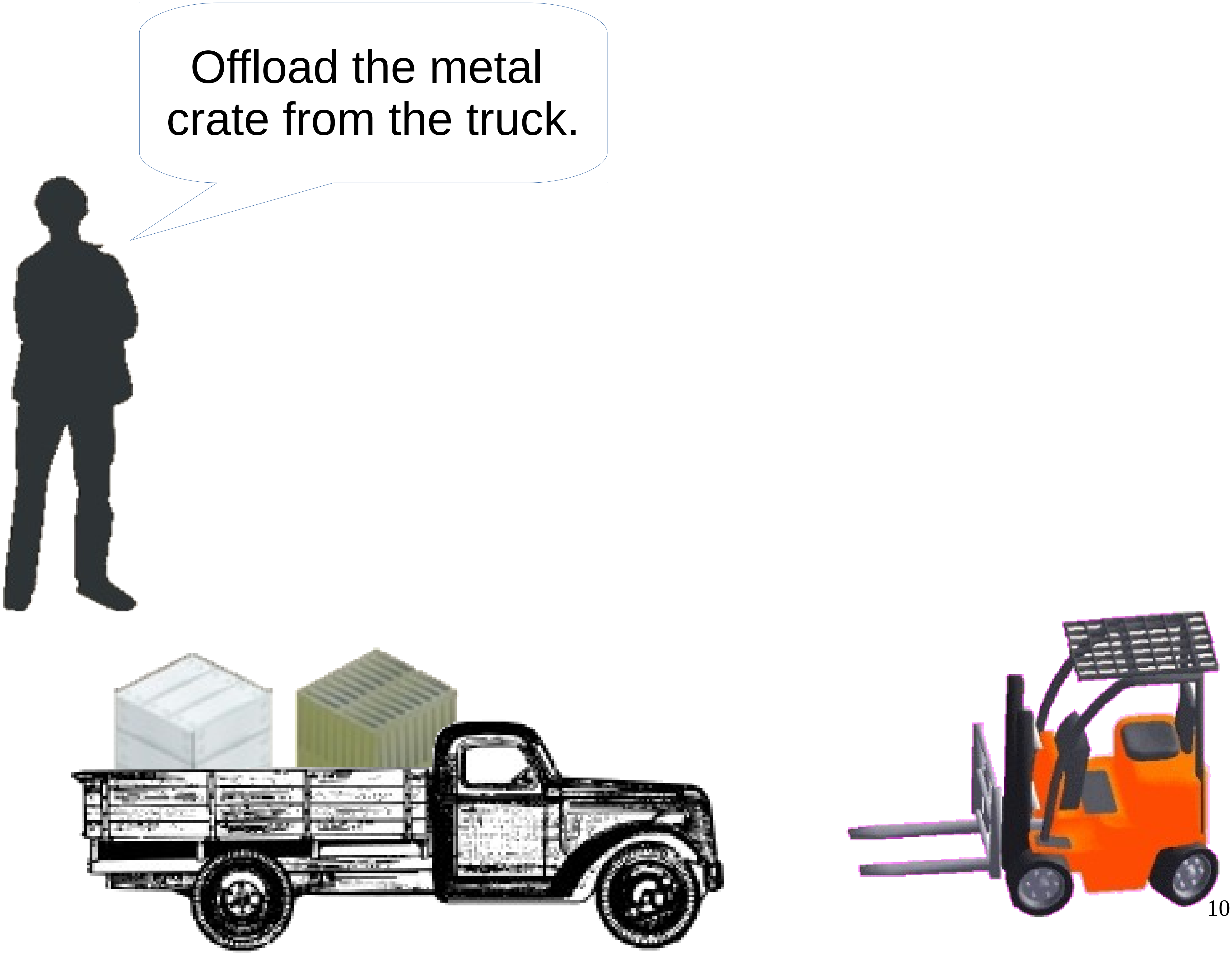
Remove the pallet from the back of the truck.

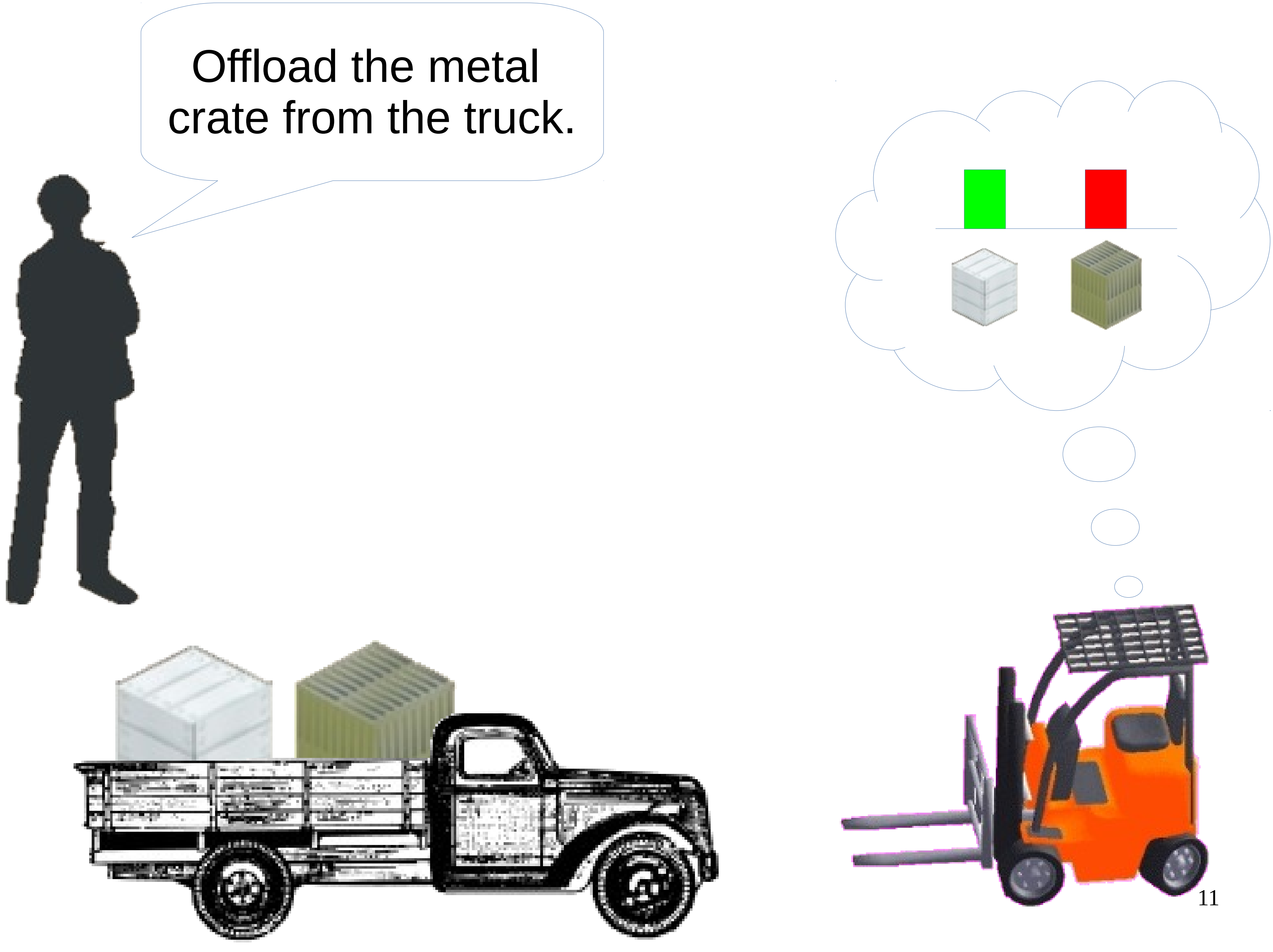
Offload the metal crate from the truck.

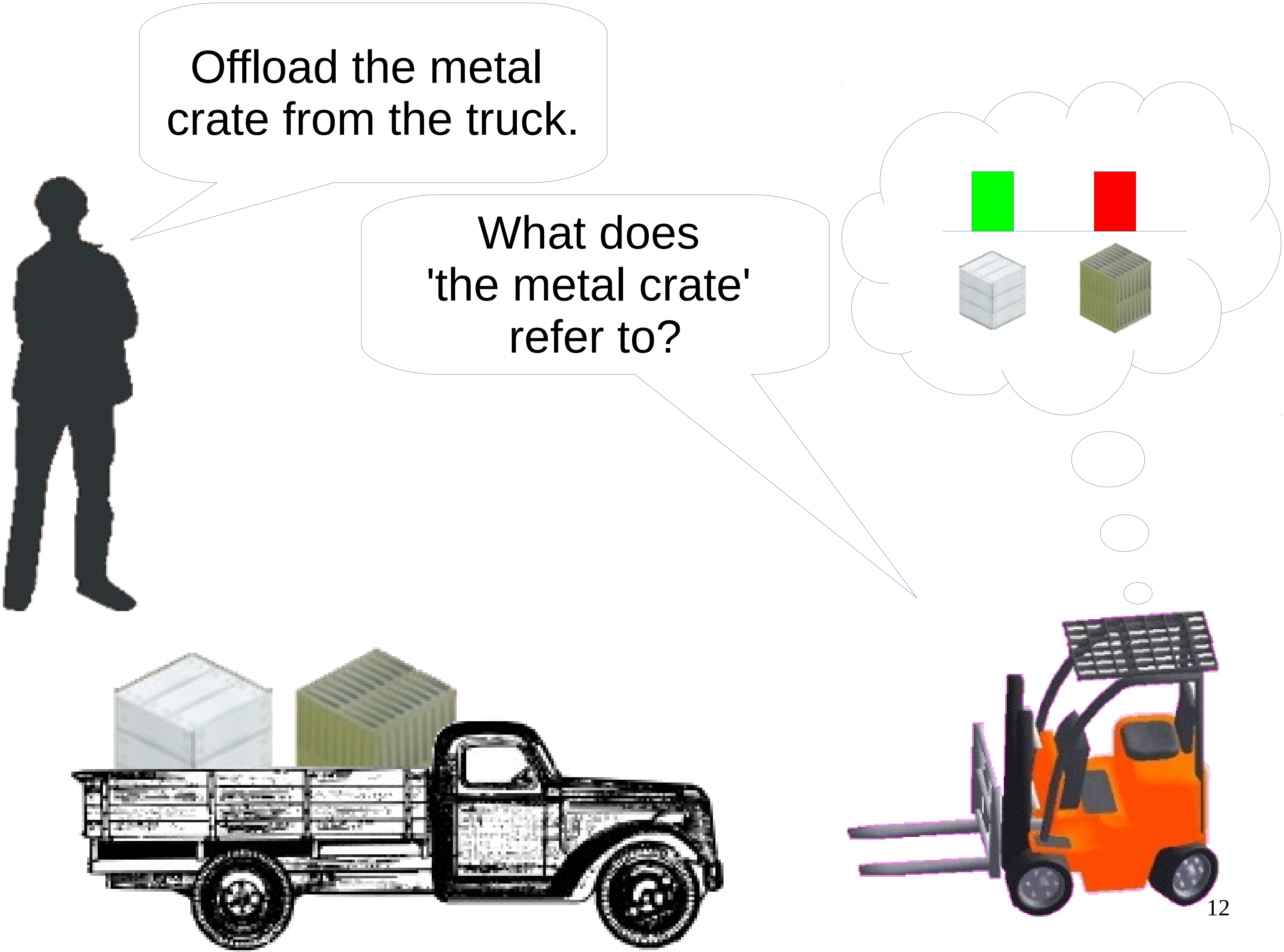
Pick up the silver container from the truck bed.

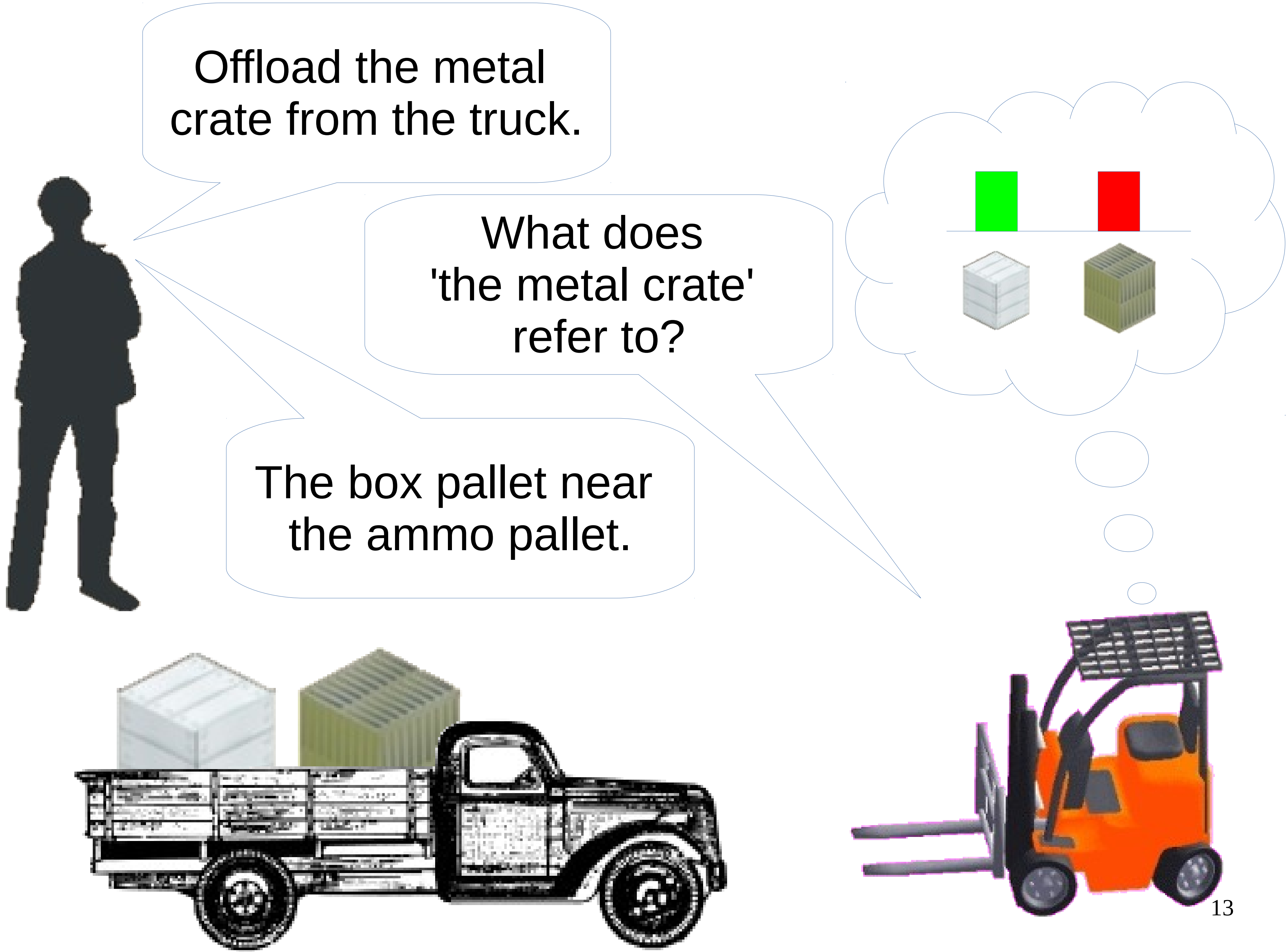
How can robots collaborate with people using natural language?

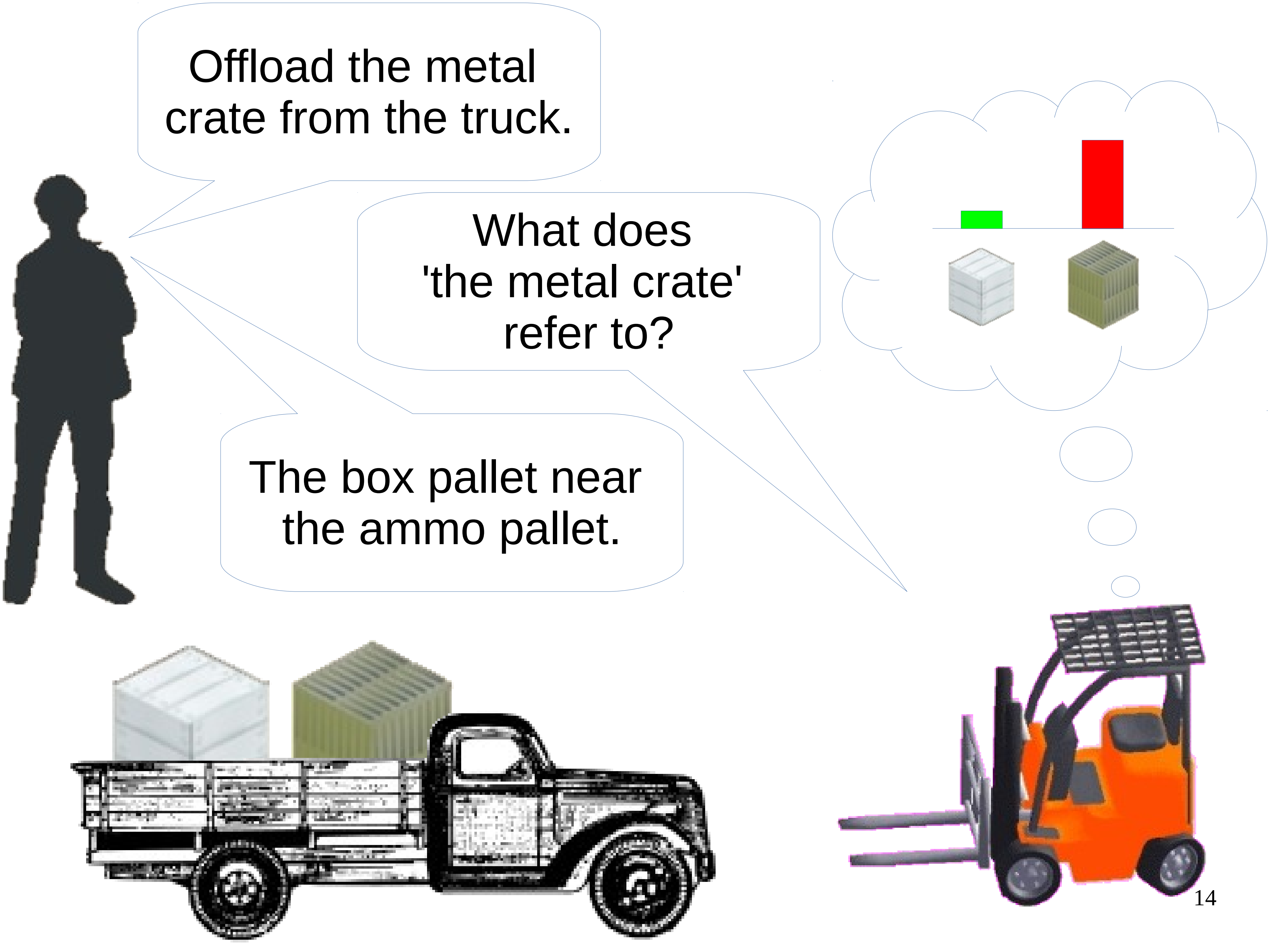
- Following instructions.
  – “Put the metal crate on the truck.”
- Asking questions.
  – “What does 'the metal crate' refer to?”
- Requesting Help.
  – “Hand me the black leg that is under the table.”


Offload the metal crate from the truck.

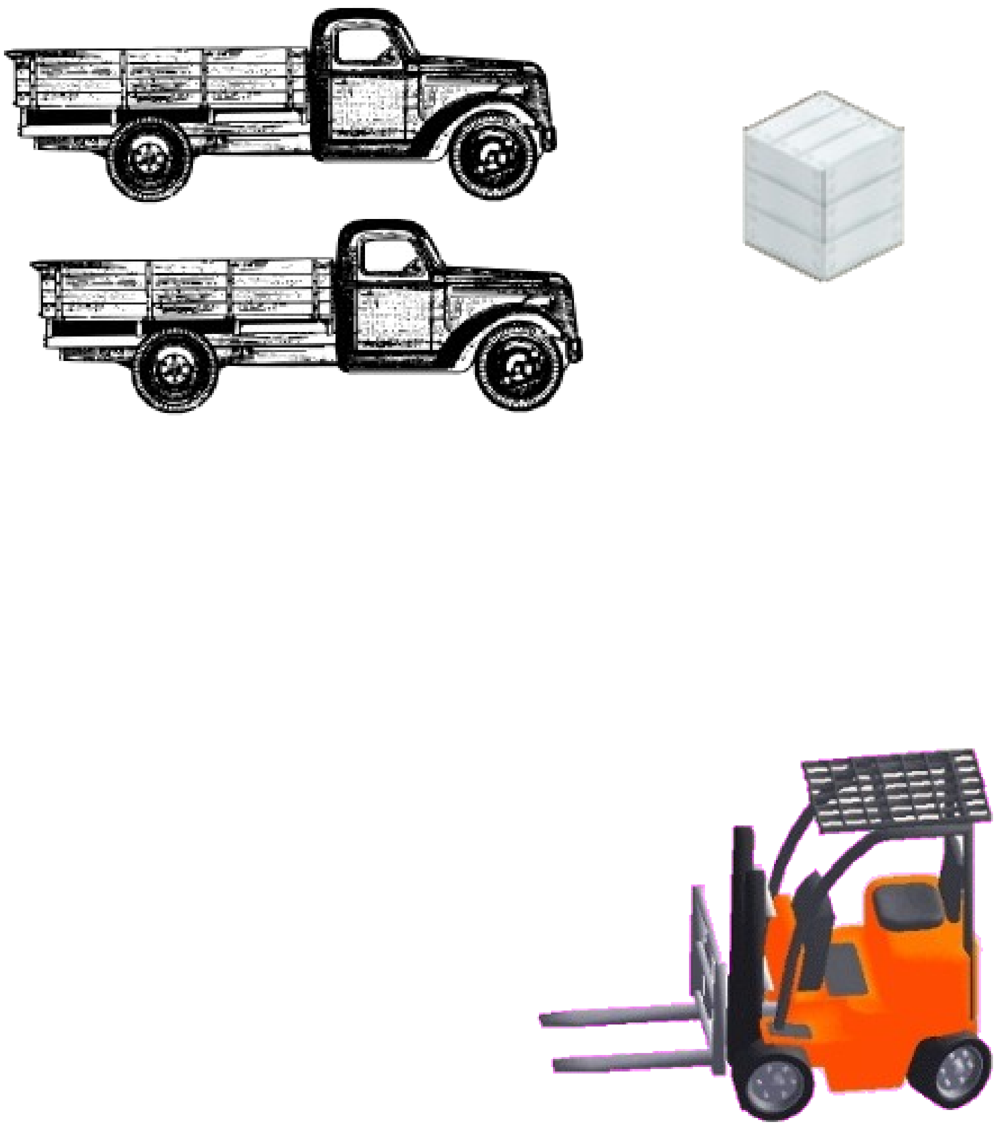

Offload the metal crate from the truck.


Offload the metal crate from the truck.
What does 'the metal crate' refer to?

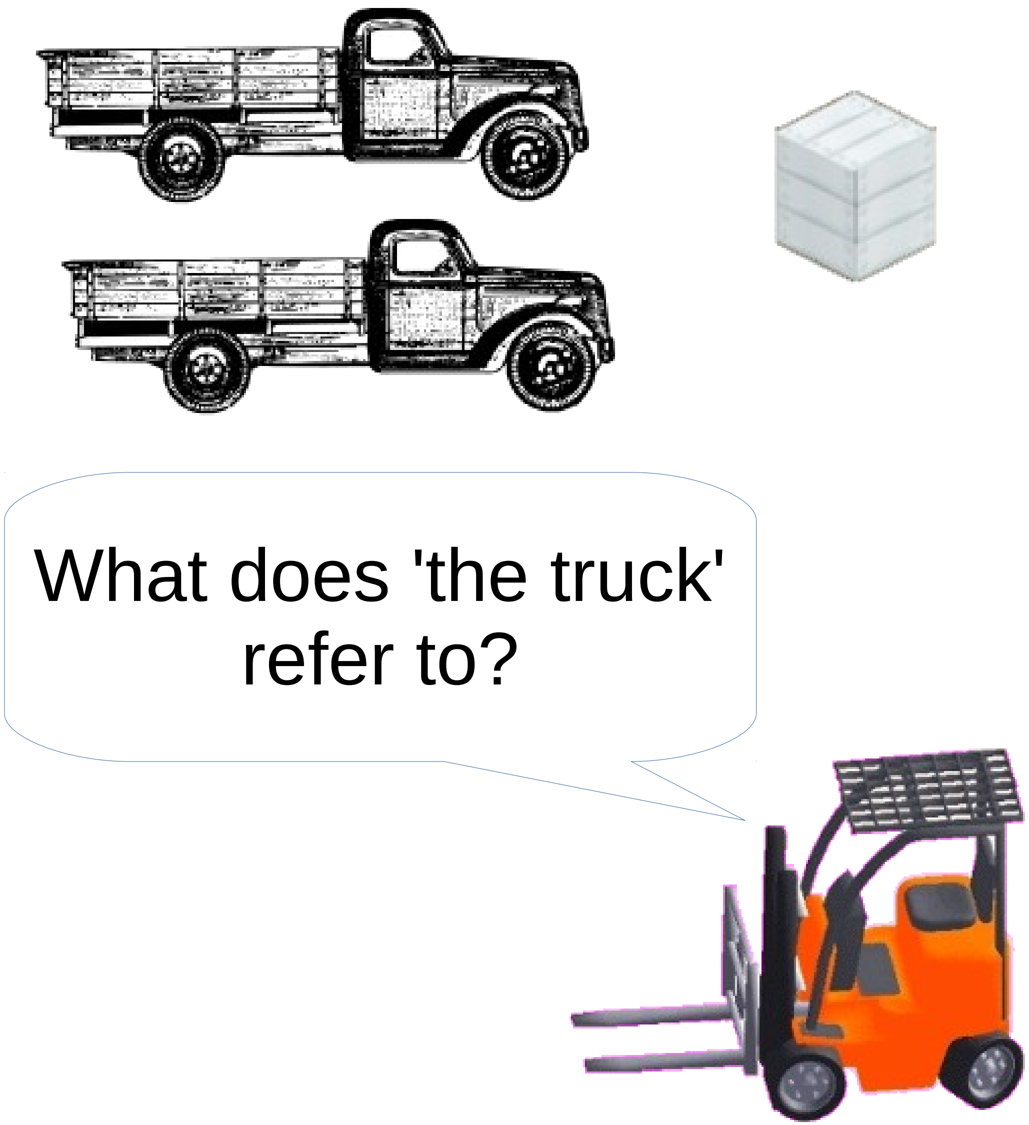

Offload the metal crate from the truck.
What does 'the metal crate' refer to?
The box pallet near the ammo pallet.


Offload the metal crate from the truck.
What does 'the metal crate' refer to?
The box pallet near the ammo pallet.

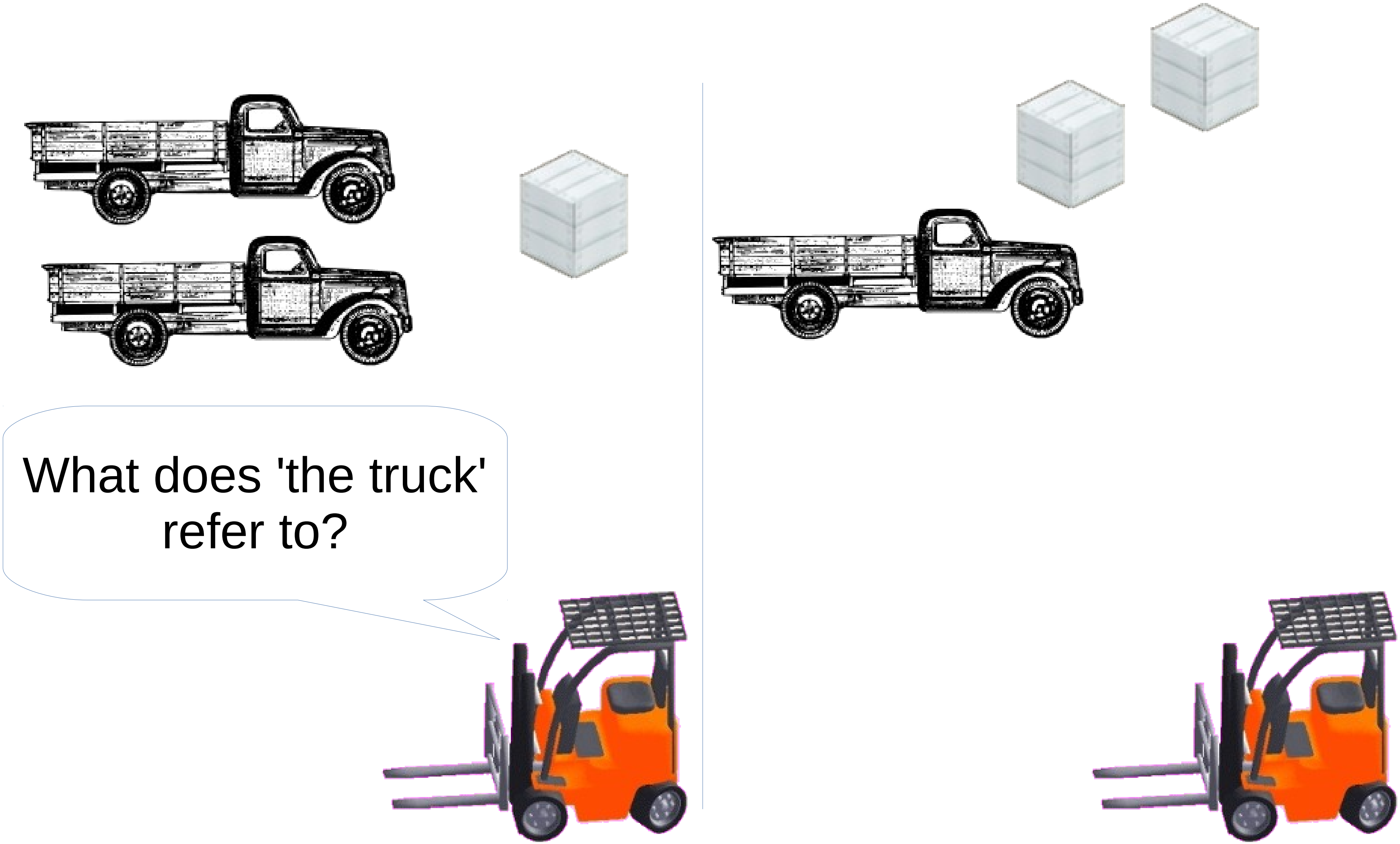
“Put the pallet on the truck.”



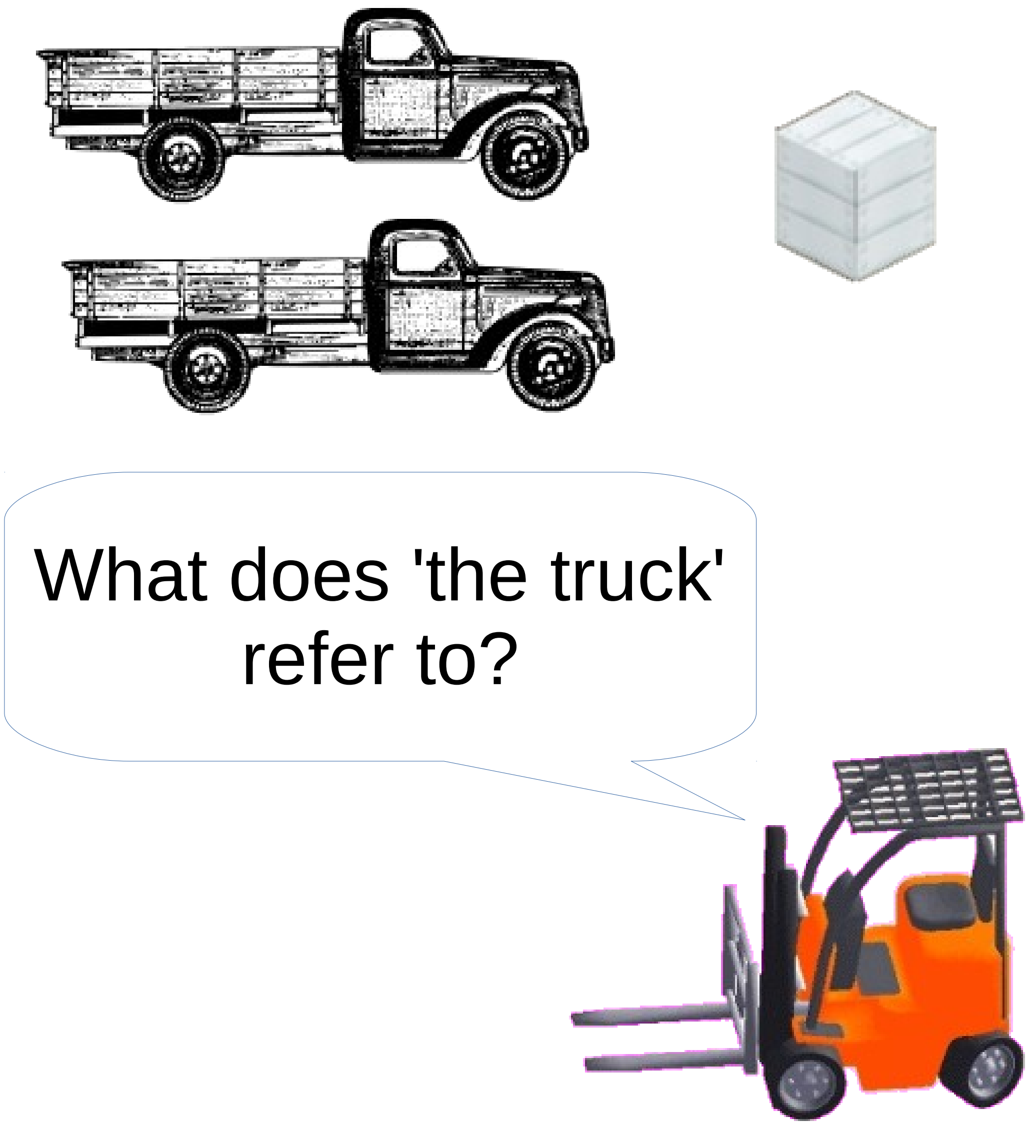
“Put the pallet on the truck.”

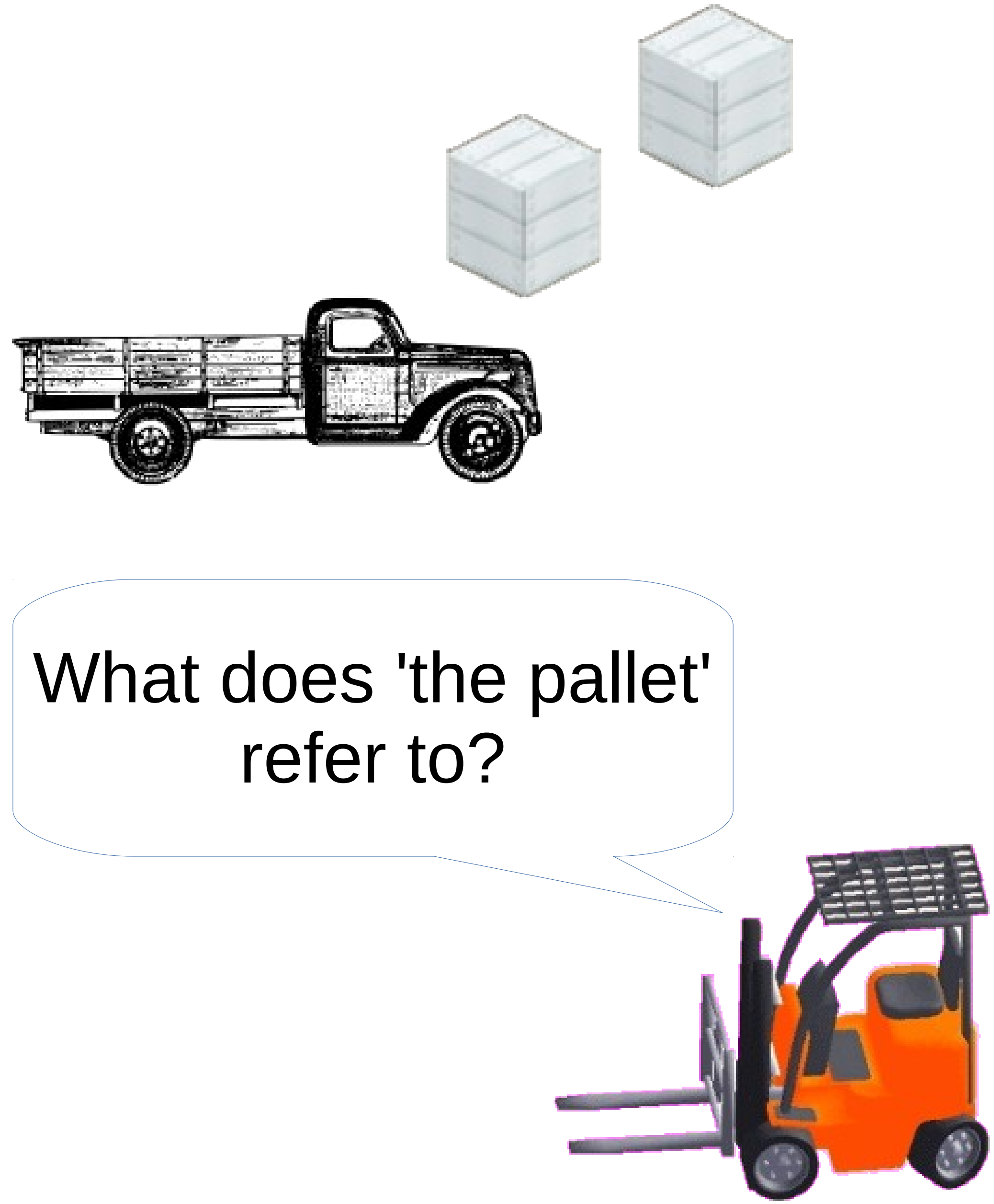


"Put the pallet on the truck."



“Put the pallet on the truck.”





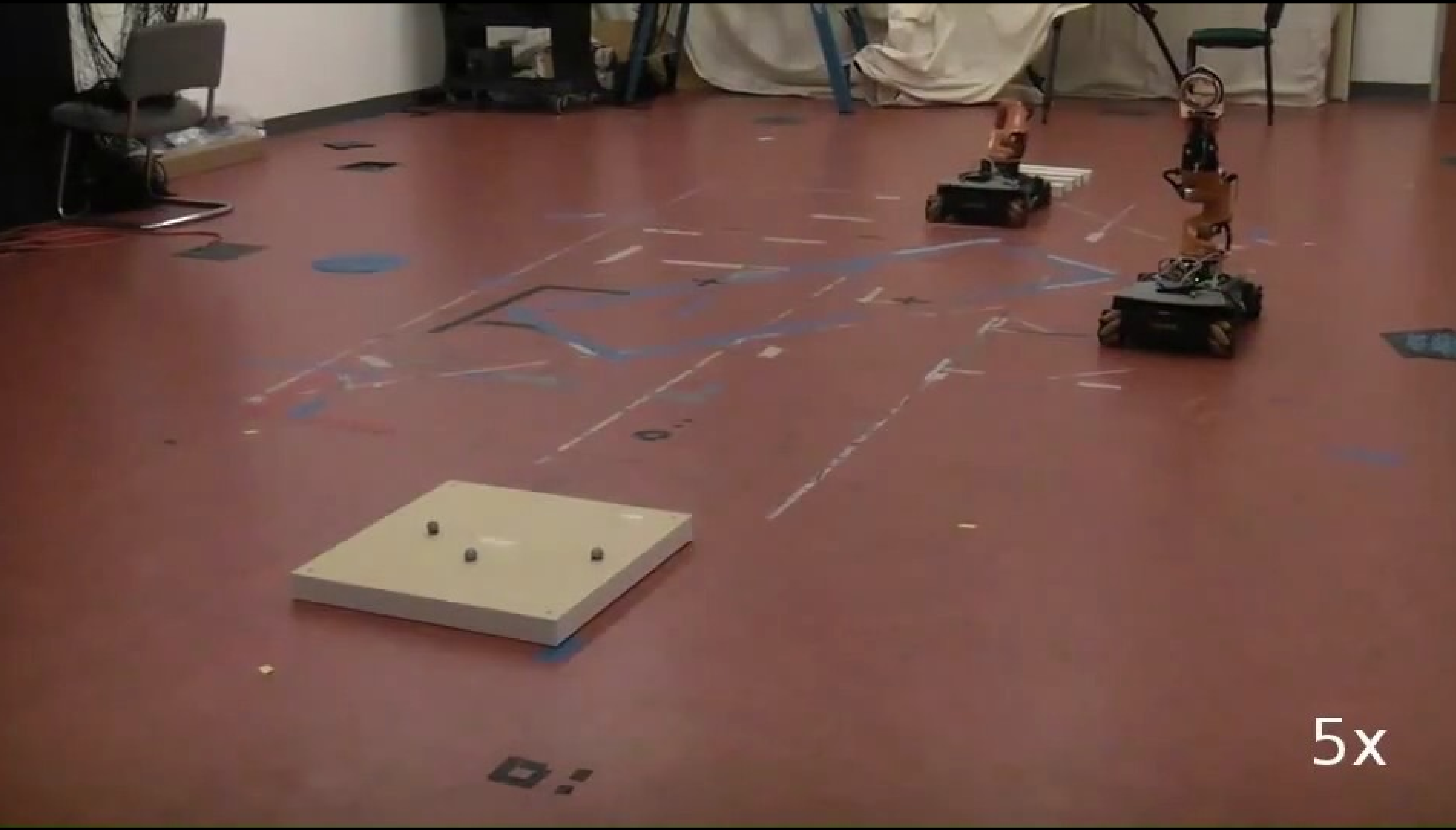
Information-theoretic Human-Robot Dialog

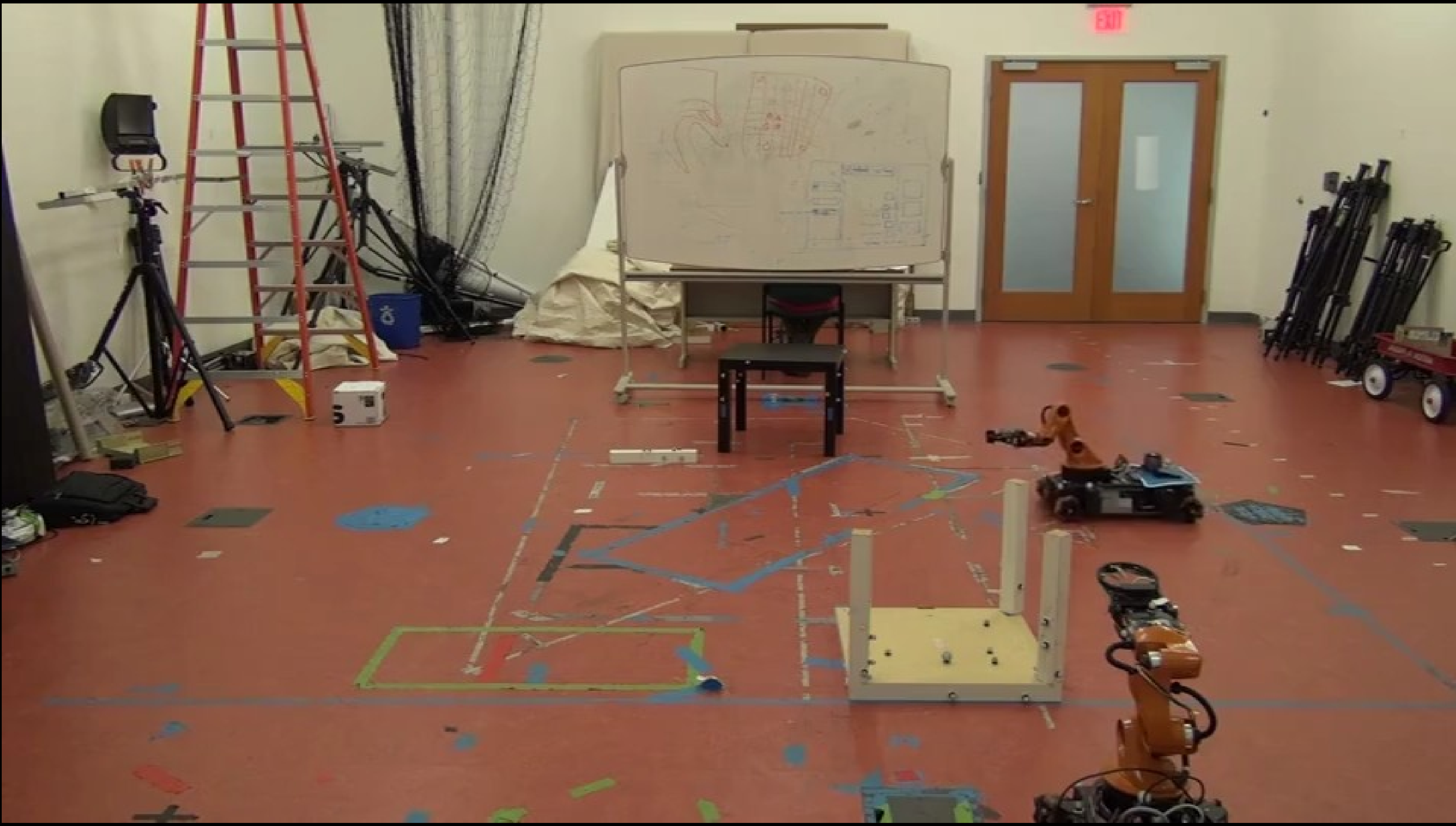
- Identify uncertain parts of the command.
- Ask a targeted question.
- Use information from the answer to infer better actions.

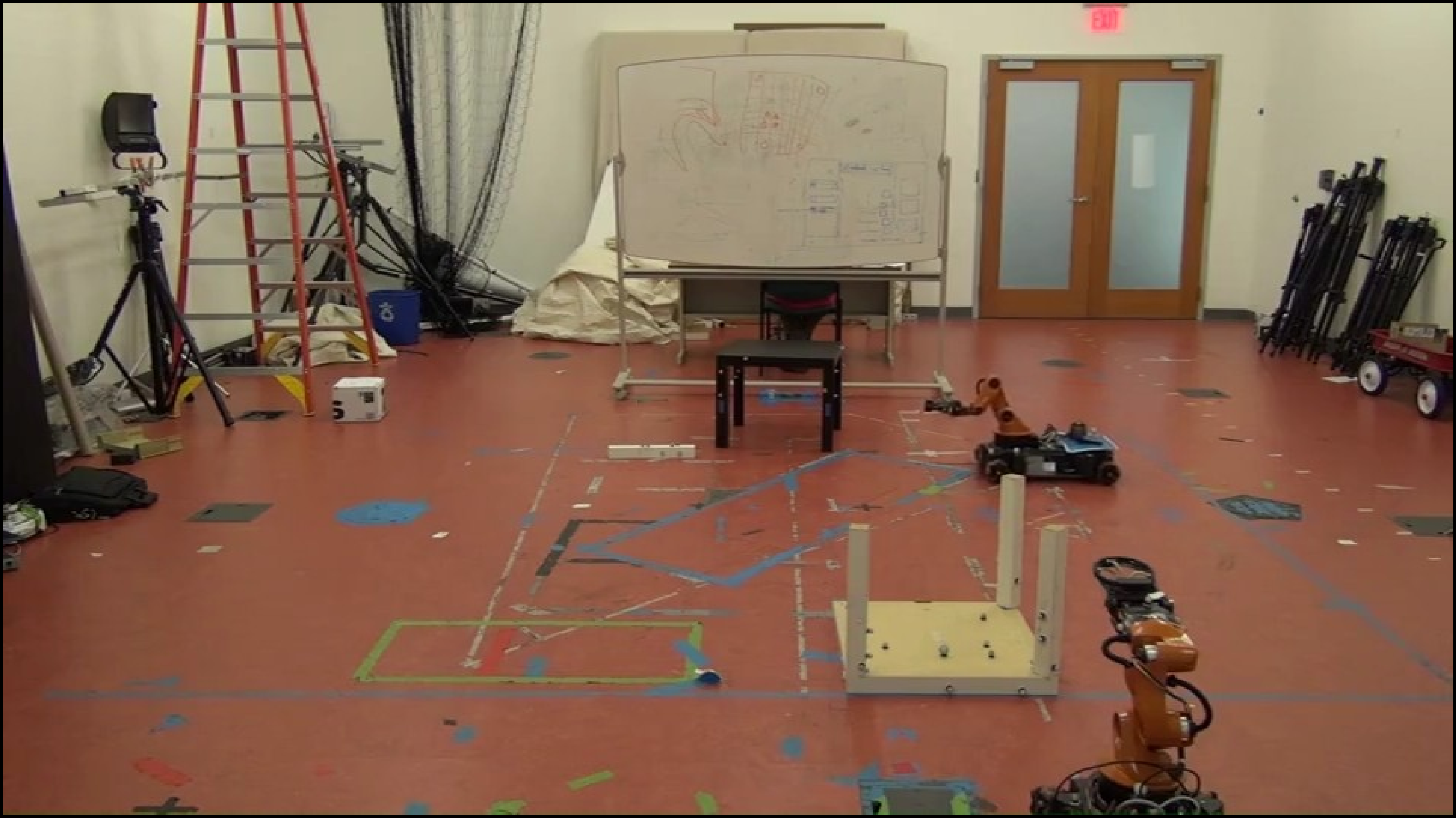
Joint work with Robin Deits, Pratiksha Thaker

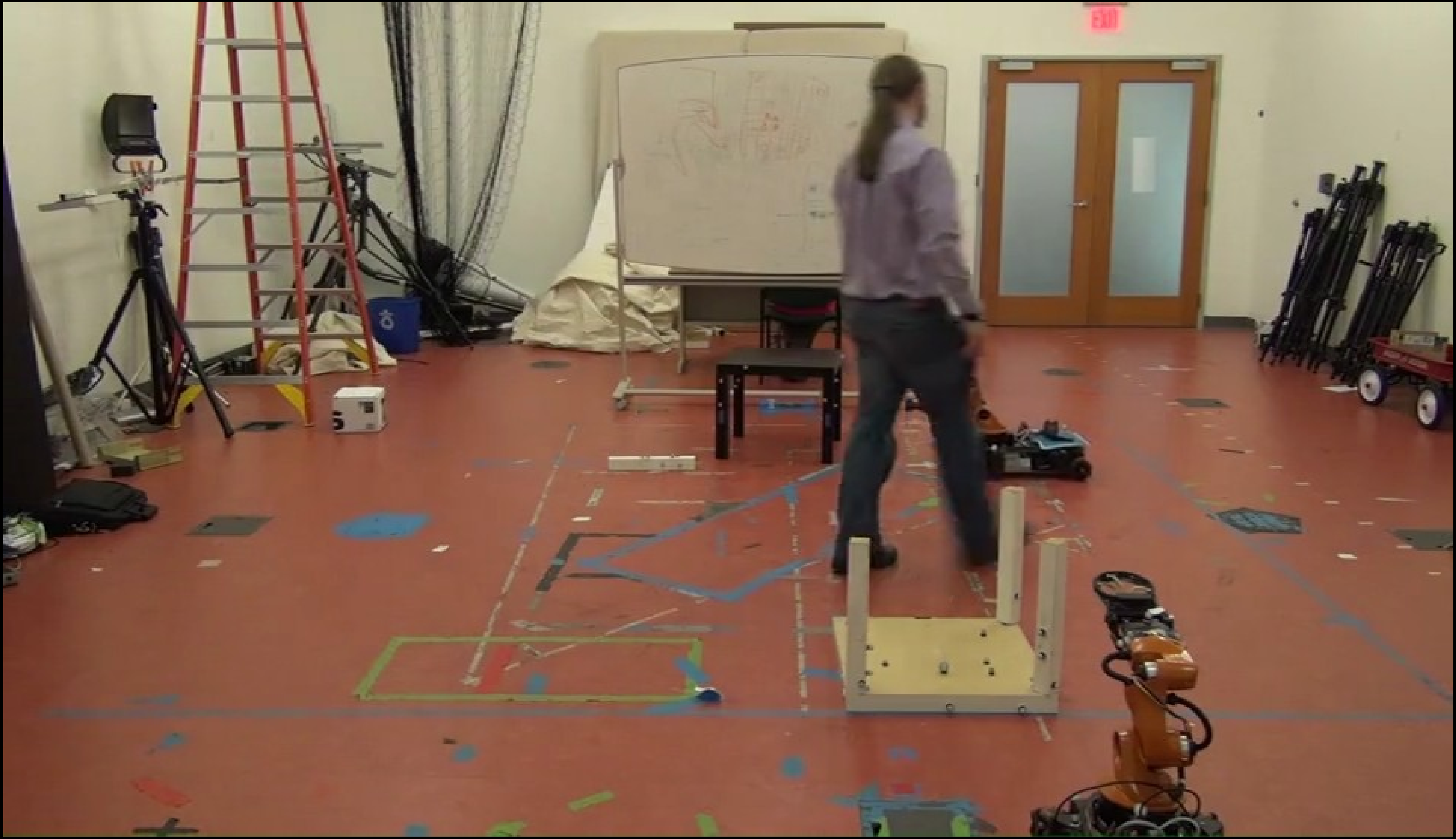
How can robots collaborate with people using natural language?

- Following instructions.
  - “Put the metal crate on the truck.”
- Asking questions.
  - “What does 'the metal crate' refer to?”
- Requesting Help.
  - “Hand me the black leg that is under the table.”


5x


EXIT


EXIT


EXIT

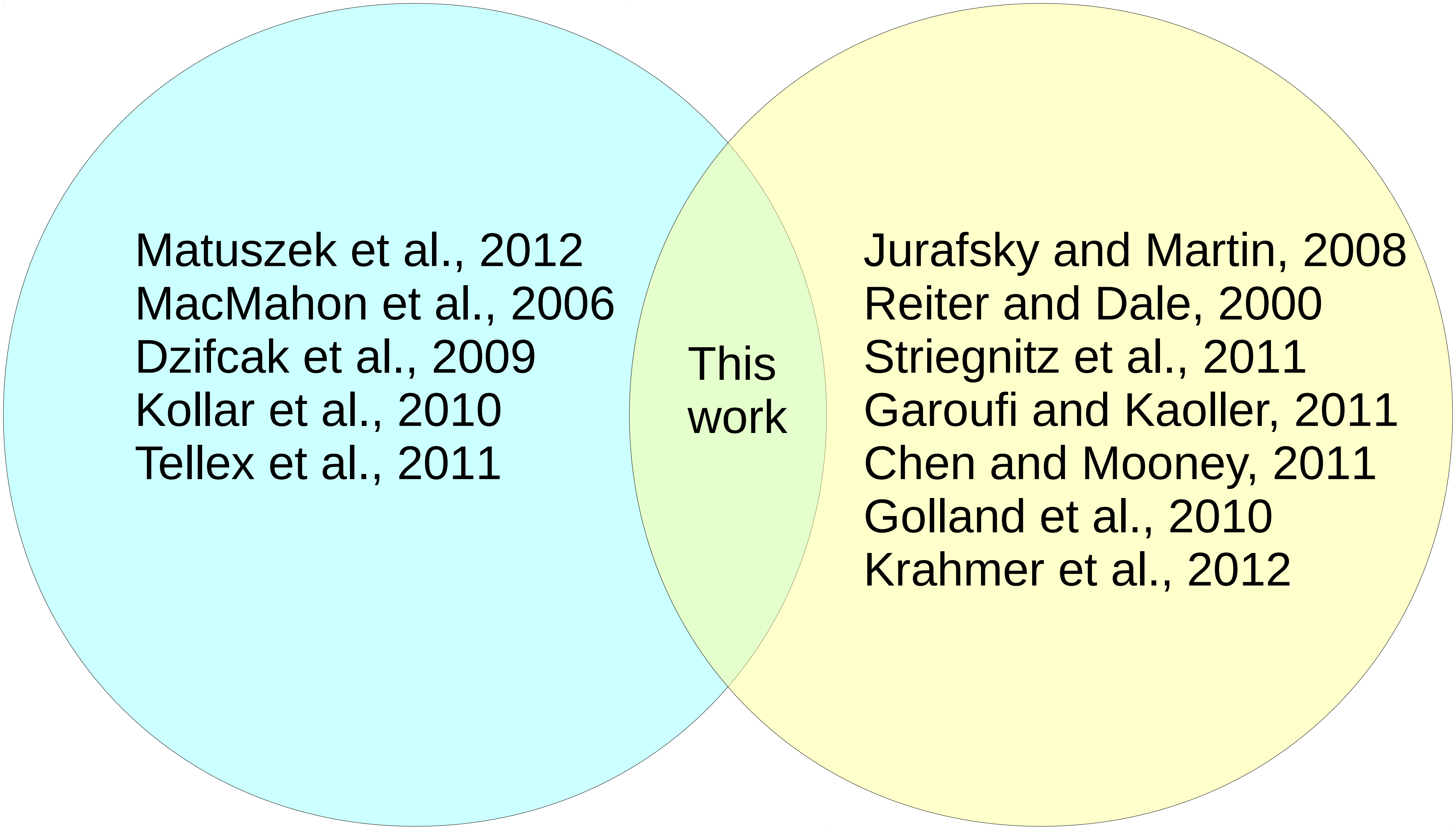

Help me!


Please hand me the white table leg.

Solution Overview

1. Detect the failure.
2. Infer an action to fix the problem.
3. Infer a natural language sentence describing the action.
4. Replan after the human has provided help based on the updated state.

Prior Work



Understanding Language Generating Language

Prior Work: Unifying Generation and Understanding

- Goodman and Stuhlmueller (2013)
  - Bayesian approach to generate and understand language.
  - Bag-of-words models of semantics demonstrated in simulation.
- Vogel et al. (2013)
  - DEC-POMDP to demonstrate Gricean maxims emerge from multiagent interaction.
  - Bag-of-words models of semantics demonstarted in simulation.
- This work
  - Bayesian approach to generate and understand grounded language for robots.
  - Compositional grounded semantics demonstrated on an end-to-end robotic system.
- Dragan and Srinivasa (2012)
  - Analogous mathematical framework for gesture interpretation and production.

Solution Overview

1. Detect the failure.
2. Infer an action to fix the problem.
3. Infer a natural language sentence describing the action.
4. Replan after the human has provided help based on the updated state.

Solution Overview

1. Detect the failure.
2. Infer an action to fix the problem.
3. Infer a natural language sentence describing the action.
4. Replan after the human has provided help based on the updated state.

Assembly System

- Strips-style symbolic planner to assemble a piece of furniture. (Knepper et al., 2013)
- Pre- and post- conditions for each action.

```
    action attach_leg_to_top(robot(Robot), leg(Leg), table_top(TableTop)) {

    pre {
        robot.arm.holding == leg;
        table_top.hole[0].attached_to == None;
    }
    post {
        robot.arm.holding = None;
        table_top.hole[0].attached_to = leg.hole;
        leg.hole.attached_to = table_top.hole[0];
    }
}
```

Solution Overview

1. Detect the failure.
2. Infer an action to fix the problem.
3. Infer a natural language sentence describing the action.
4. Replan after the human has provided help based on the updated state.

Solution Overview

1. Detect the failure.
2. Infer an action to fix the problem.
3. Infer a natural language sentence describing the action.
4. Replan after the human has provided help based on the updated state.

Infer an Action

- Rule-based heuristic to generate symbolic request.

Failed symbolic condition	Symbolic request
`part.visible == True;`	locate_part(*robot*, *part*)
`robot.arm.holding == leg;`	give_part(*robot*, *part*)
`leg.aligned == top.hole[0];`	align_with_hole(*leg*, *top*, *hole*)
`leg.hole.attached == top.hole[0];`	screw_in_leg(*leg*, *top*, *hole*)
`top.upside_down == True;`	flip(*top*)

Solution Overview

1. Detect the failure.
2. Infer an action to fix the problem.
3. Infer a natural language sentence describing the action.
4. Replan after the human has provided help based on the updated state.

Solution Overview

1. Detect the failure.
2. Infer an action to fix the problem.
3. Infer a natural language sentence describing the action.
4. Replan after the human has provided help based on the updated state.

Infer a Natural Language Sentence

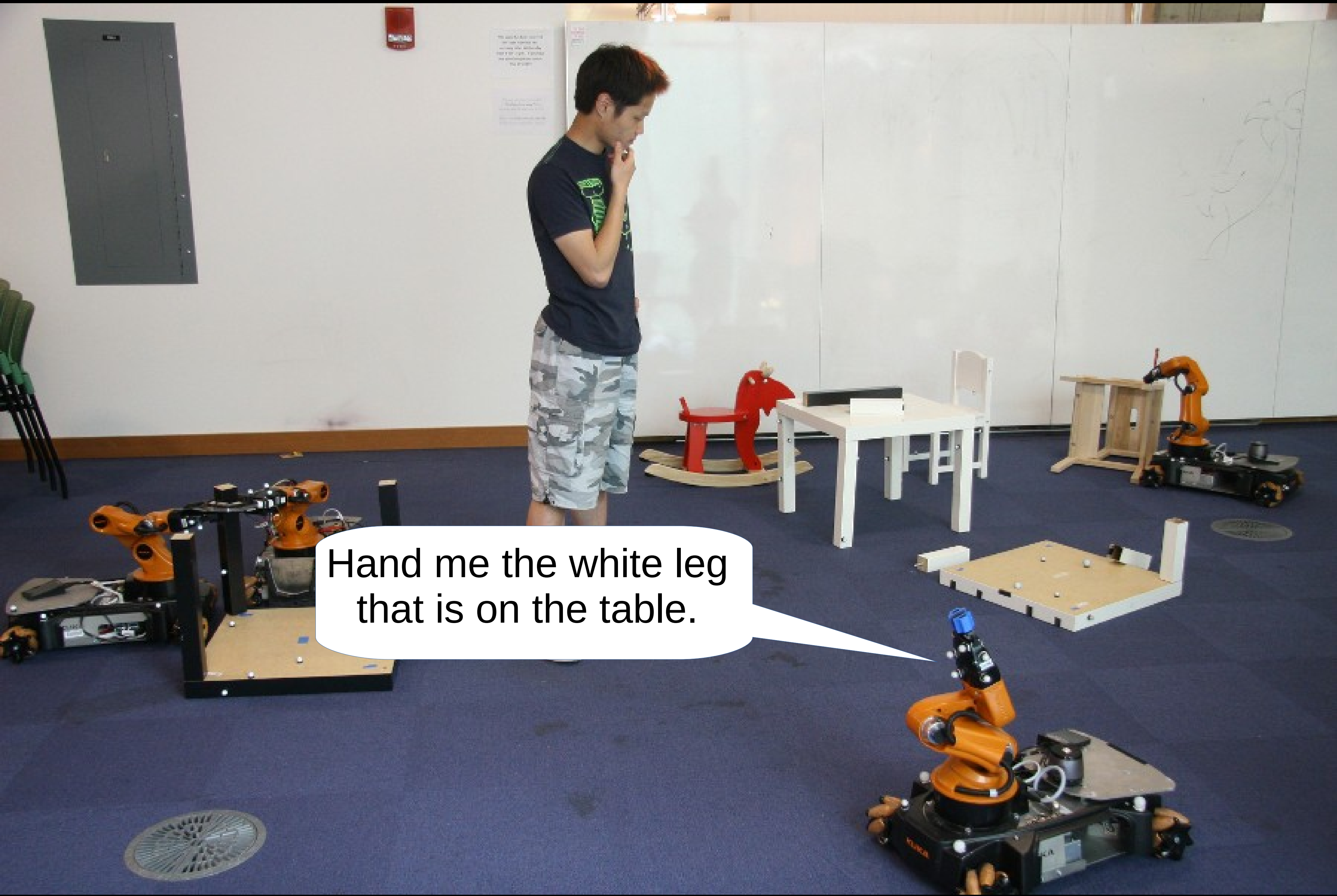
- Input: Symbolic Request
- Output: Natural language request

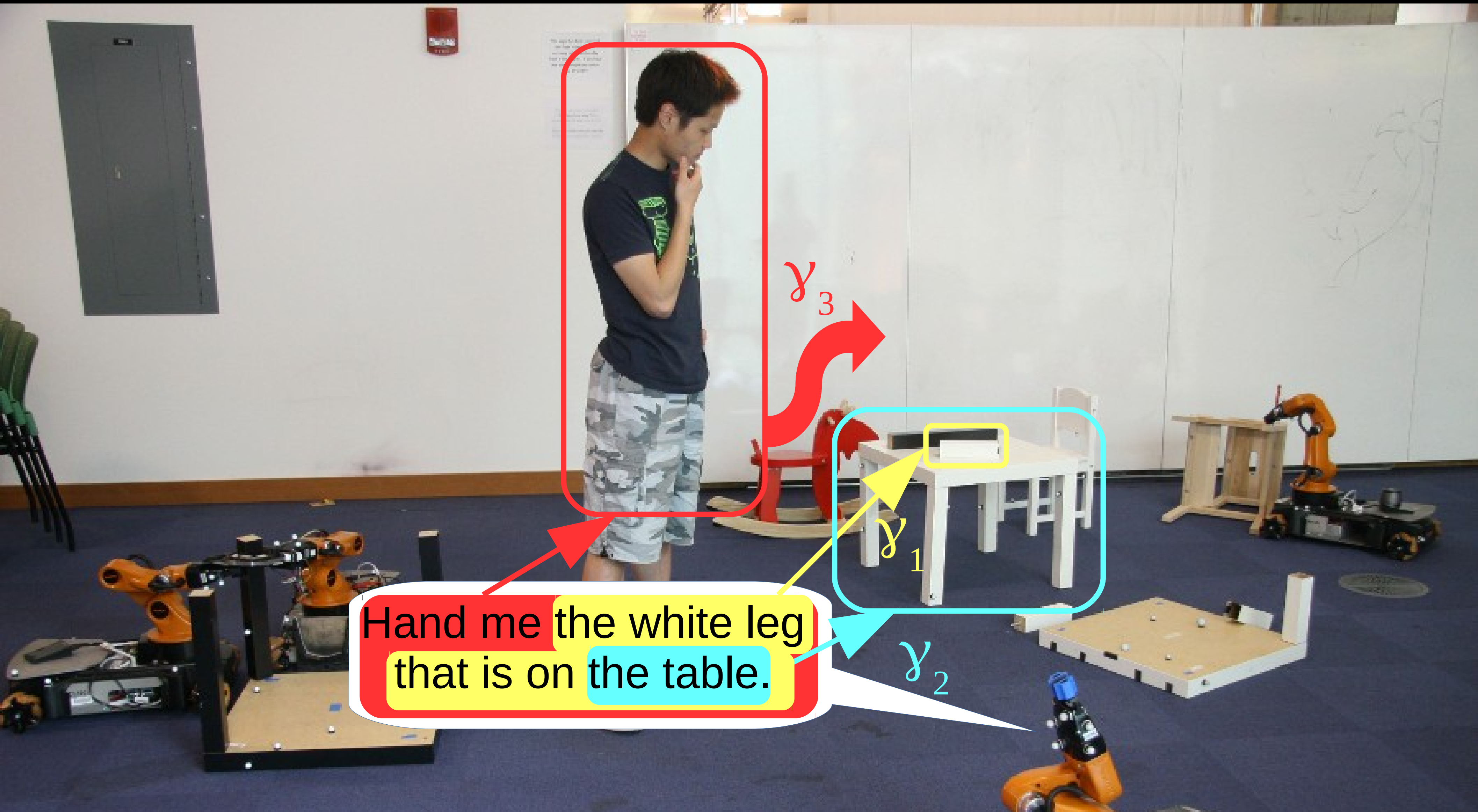
Symbolic request	Natural Language Request
`locate_part(`*`robot`*`, `*`part`*`)`	`Find the part.`
`give_part(`*`robot`*`, `*`part`*`)`	`Give me the part.`
`align_with_hole(`*`leg`*`, `*`top`*`, `*`hole`*`)`	`Align the part with the hole.`
`screw_in_leg(`*`leg`*`, `*`top`*`, `*`hole`*`)`	`Screw in the leg.`
`flip(`*`top`*`)`	`Flip the table.`

Template baseline


Hand me the part.


Hand me the white leg.


Hand me the white leg that is on the table.



γ_k are *groundings*, or objects, places, paths, and events in the external world. Each γ_k corresponds to a constituent phrase in the language input.

Semantics

Understanding (Forward Semantics):

What groundings match the language?

$$\underset{\gamma_1 \ldots \gamma_N}{argmax} \; f(\gamma_1 \ldots \gamma_N, language)$$

γ_k are *groundings*, or objects, places, paths, and events in the external world. Each γ_k corresponds to a constituent phrase in the language input.

Semantics

Understanding (Forward Semantics):

What groundings match the language?

$$\underset{\gamma_1 \ldots \gamma_N}{argmax} \; p(\gamma_1 \ldots \gamma_N | language)$$

$$\underset{\gamma_1 \ldots \gamma_N}{argmax} \; \frac{1}{Z} \prod_i g_i(\gamma_1 \ldots \gamma_N, language)$$

Tellex et al. 2011

γ_k are *groundings*, or objects, places, paths, and events in the external world. Each γ_k corresponds to a constituent phrase in the language input.

Semantics

Understanding (Forward Semantics):

What groundings match the language?

$$\underset{\gamma_1 \ldots \gamma_N}{argmax} \; p(\gamma_1 \ldots \gamma_N | language)$$

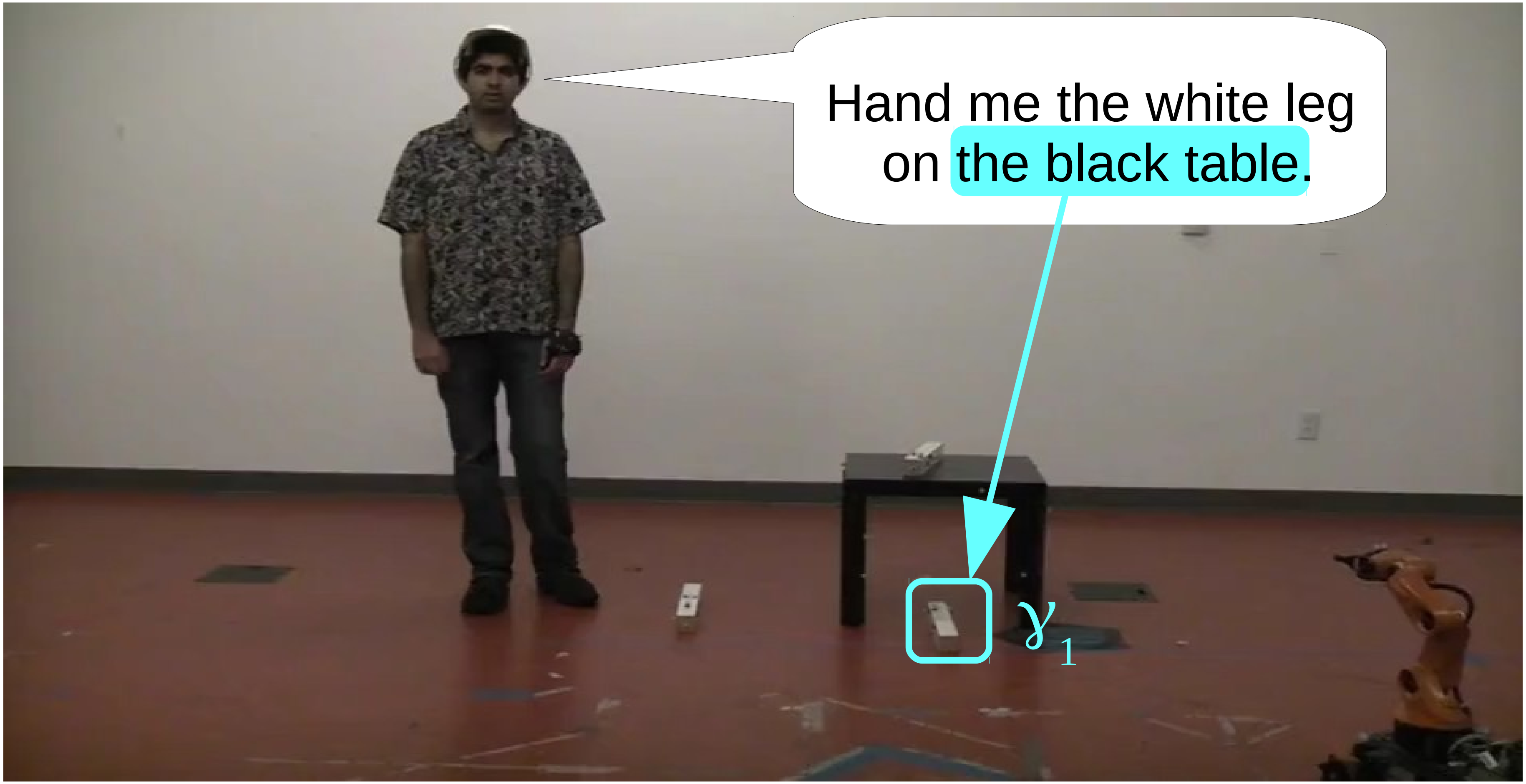
$$\underset{\gamma_1 \ldots \gamma_N}{argmax} \; \frac{1}{Z} \prod_i g_i(\gamma_1 \ldots \gamma_N, language)$$

Tellex et al. 2011

γ_k are *groundings*, or objects, places, paths, and events in the external world. Each γ_k corresponds to a constituent phrase in the language input.

Searching for Groundings

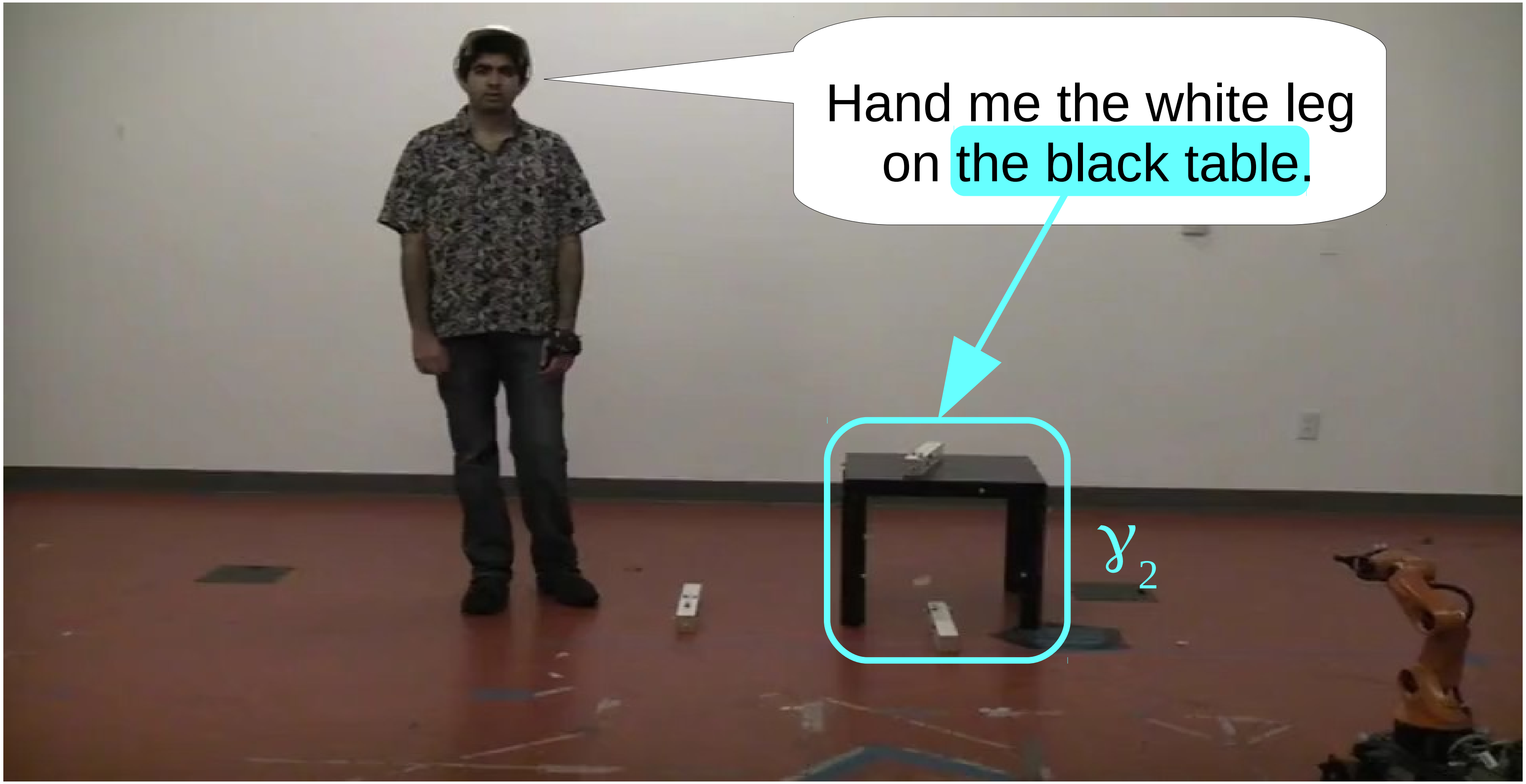
$$\underset{\gamma_1 \ldots \gamma_N}{argmax} \; \frac{1}{Z} \prod_i g_i(\gamma_1 \ldots \gamma_N, language)$$



$$g_1(\gamma_1, \text{the black table}) = 0.1$$

Searching for Groundings

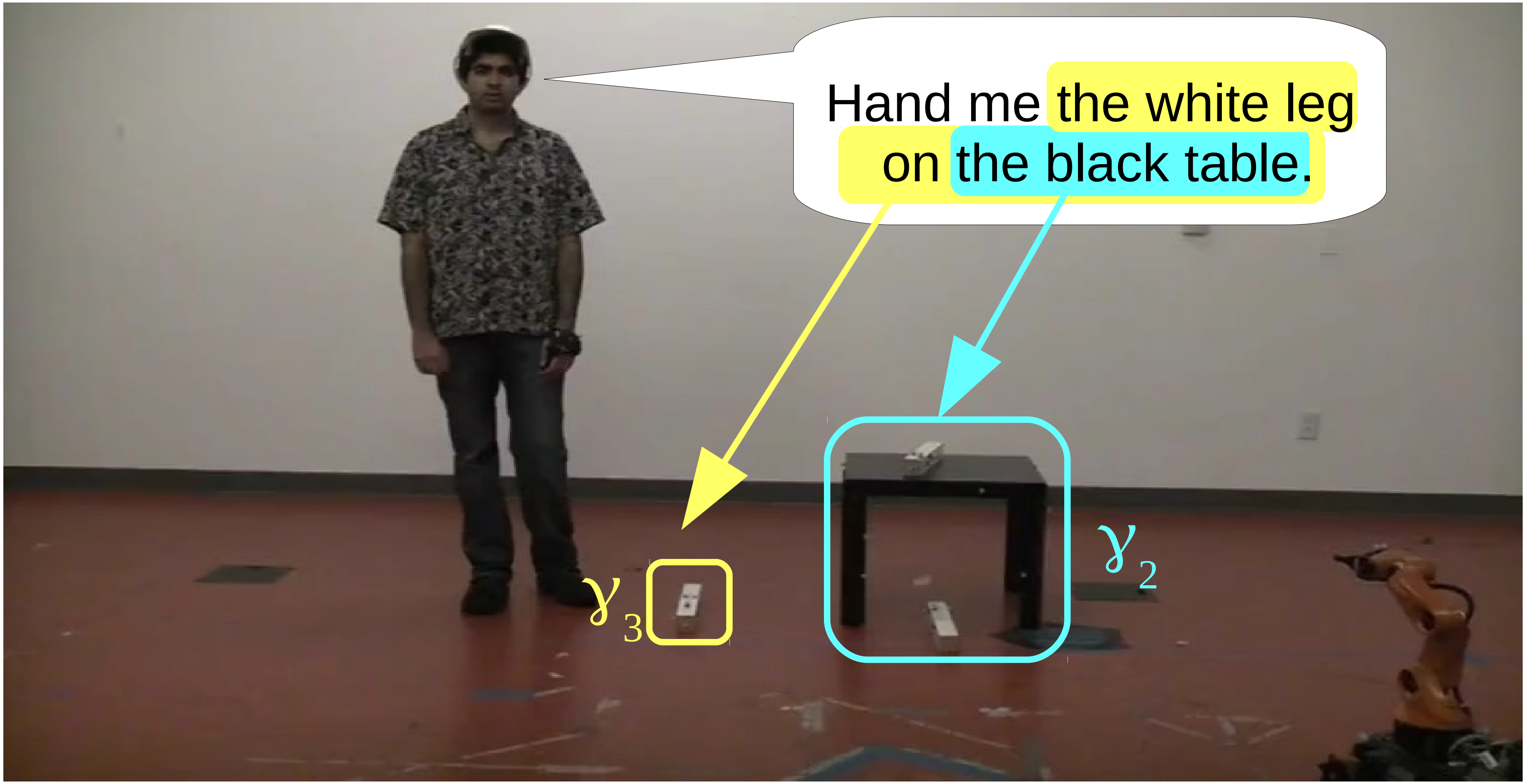
$$\underset{\gamma_1 \dots \gamma_N}{argmax} \; \frac{1}{Z} \prod_i g_i(\gamma_1 \dots \gamma_N, language)$$



$$g_1(\gamma_{1,} \text{the black table}) = 0.9$$

Searching for Groundings

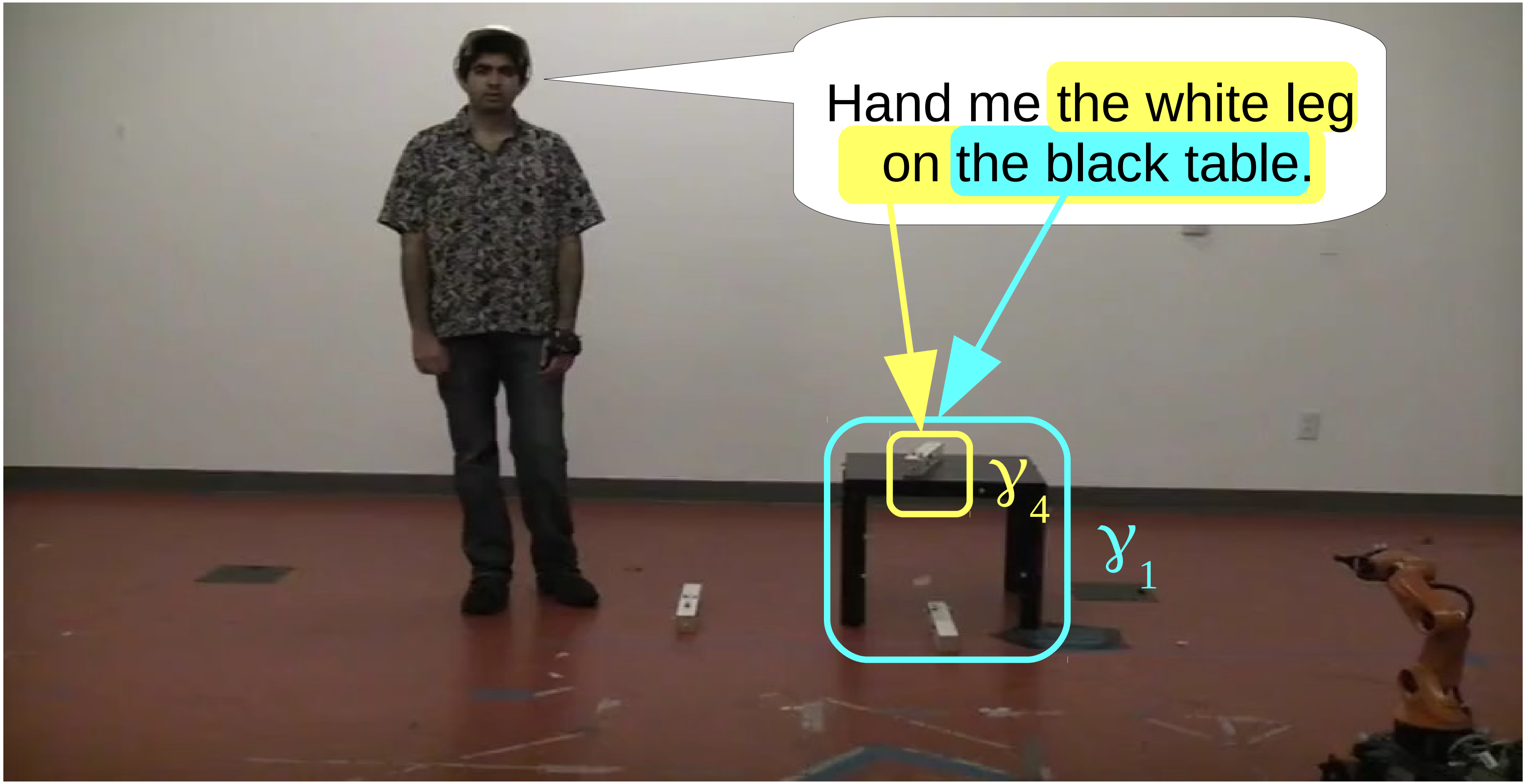
$$\underset{\gamma_1 \dots \gamma_N}{\operatorname{argmax}} \; \frac{1}{Z} \prod_i g_i(\gamma_1 \dots \gamma_N, language)$$



$$g_1(\gamma_1, \text{the white leg on the black table}) = 0.1$$

Searching for Groundings

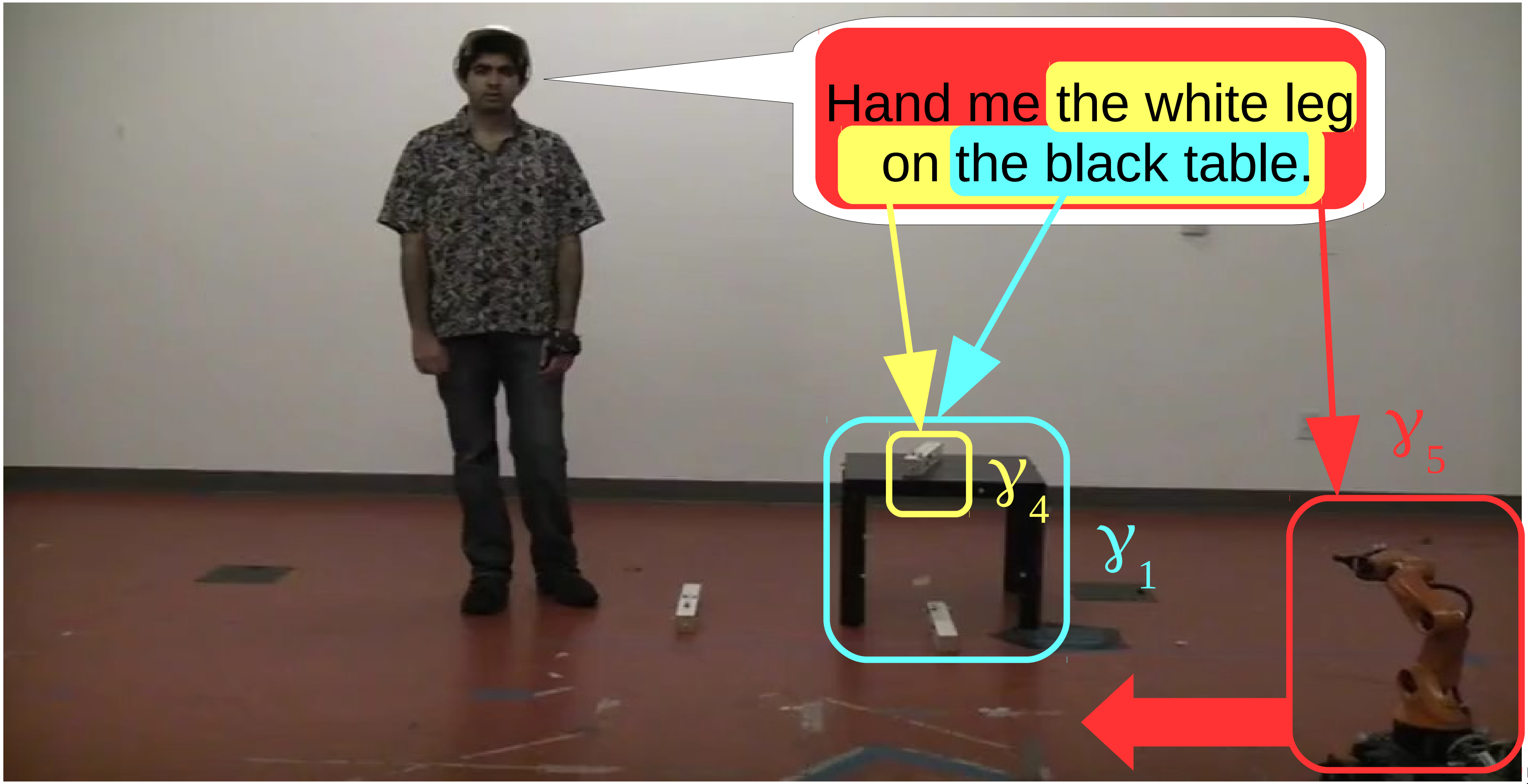
$$\underset{\gamma_1 \dots \gamma_N}{argmax} \; \frac{1}{Z} \prod_i g_i(\gamma_1 \dots \gamma_N, language)$$



$$g_1(\gamma_{1,} \text{the white leg on the black table}) = 0.9$$

Searching for Groundings

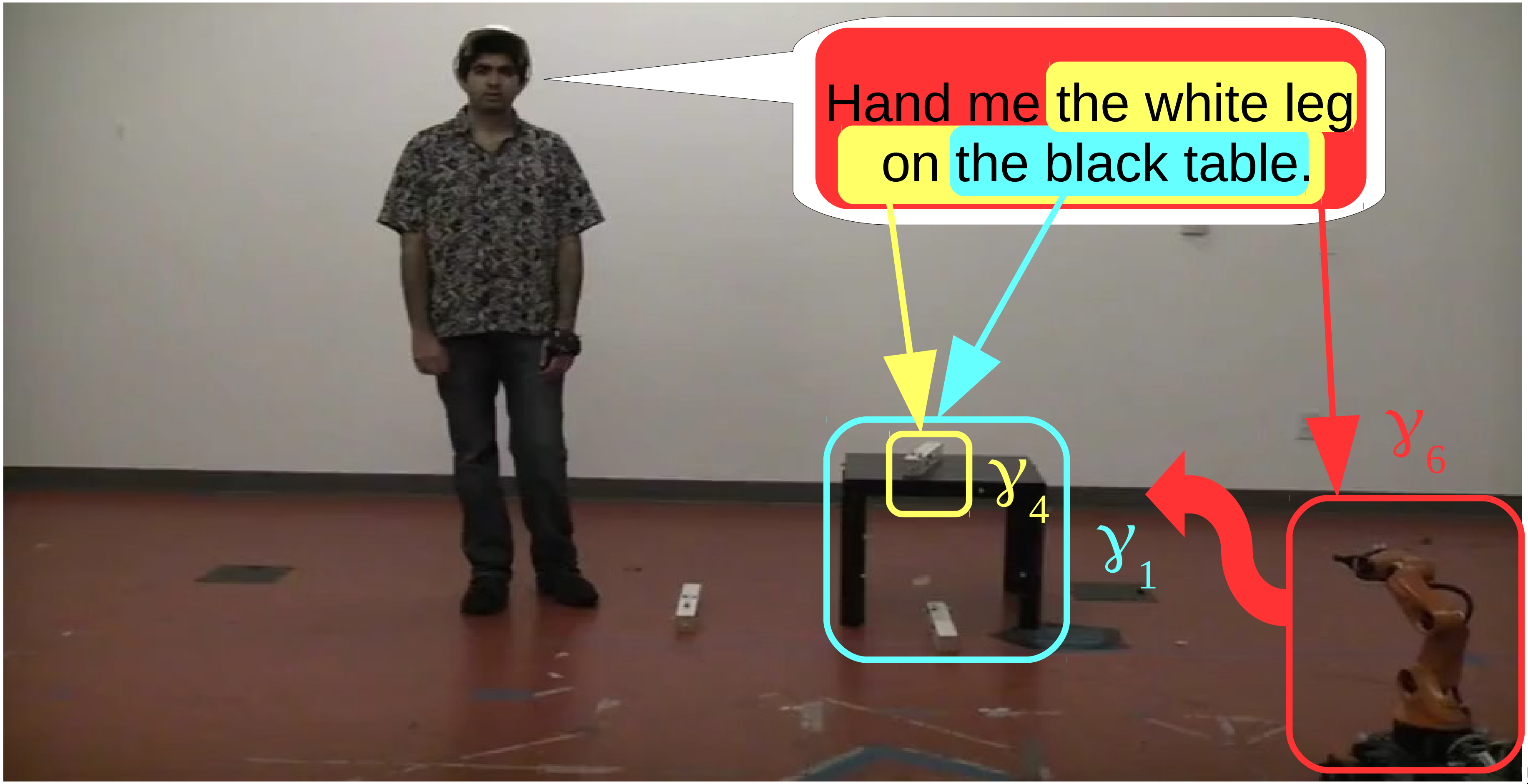
$$\underset{\gamma_1 \dots \gamma_N}{argmax} \; \frac{1}{Z} \prod_i g_i(\gamma_1 \dots \gamma_N, language)$$



$$g_1(\gamma_1, \text{Hand me the white leg on the black table}) = 0.1$$

Searching for Groundings

$$\underset{\gamma_1 \ldots \gamma_N}{\mathrm{argmax}} \; \frac{1}{Z} \prod_i g_i(\gamma_1 \ldots \gamma_N, \mathit{language})$$



$$g_1(\gamma_{1,} \text{Hand me the white leg on the black table}) = 0.9$$

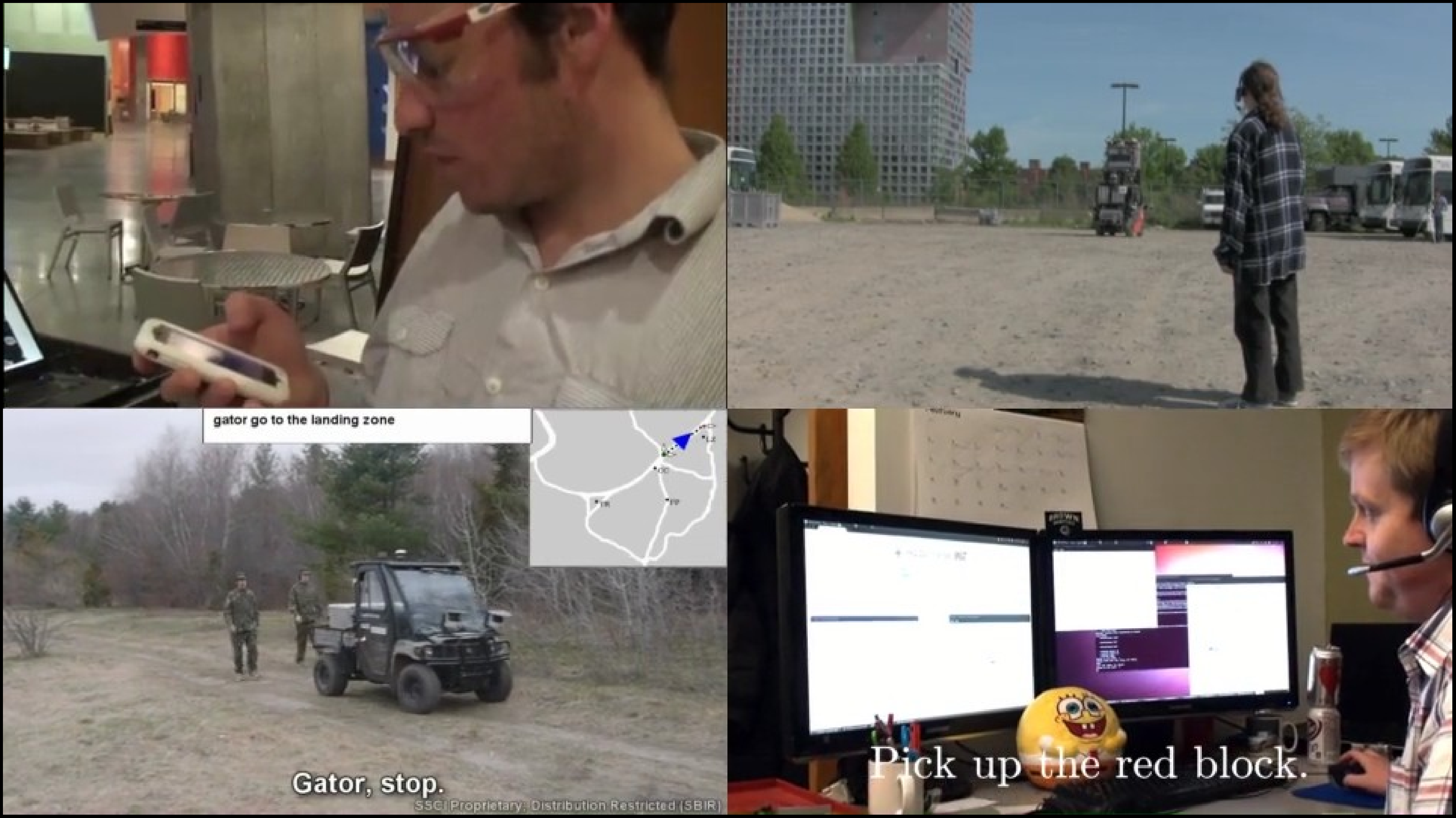
Training the Semantics Model

Type the words you would use to ask a person
to carry out the action you see in this video.

Training the Semantics Model

Pick up a black table leg off of the floor.
Pick up the black table leg.
Pick up the black table leg.
Walk over to the white table.
Place black leg on white table bottom.
Locate the black table leg on the floor by the white table.
Find the black table leg and attach it to the white table.
Hand me the black table leg

Robotic Demonstrations of G^3



Tellex et al. AAAI 2011, Kollar, Tellex et al. HRI 2010, Huang, Tellex et al., IROS 2010, Tellex et al., JHRI 2013, Tellex et al., MLJ 2013

Semantics

Understanding (Forward Semantics):

What groundings match the language?

$$\underset{\gamma_1 \ldots \gamma_N}{argmax} \; f\left(\gamma_1 \ldots \gamma_N, language\right)$$

Generation (Inverse Semantics):

What language specifies the groundings?

$$\underset{language}{argmax} \; f\left(\gamma_1 \ldots \gamma_N, language\right)$$

γ_k are *groundings*, or objects, places, paths, and events in the external world. Each γ_k corresponds to a constituent phrase in the language input.

Searching for Sentences

Context Free Grammar

$$S \rightarrow VP \quad NP$$

$$S \rightarrow VP \quad NP \quad PP$$

$$NP \rightarrow NP \quad PP$$

$$PP \rightarrow TO \; NP$$

$$VP \rightarrow \text{flip}|\text{give}|\text{pickup}|\text{place}$$

$$NP \rightarrow \frac{\text{the white leg}|\text{the black leg}|\text{me}}{\text{the white table}|\text{the black table}}$$

$$TO \rightarrow \text{under}|\text{on}|\text{near}$$

Searching for Sentences

- *S*
  - *VP*
    - Pick up
  - *NP*
    - NP
      - the white leg
    - *PP*
      - *TO*
        - near
      - NP
        - the black table

Searching for Sentences

S

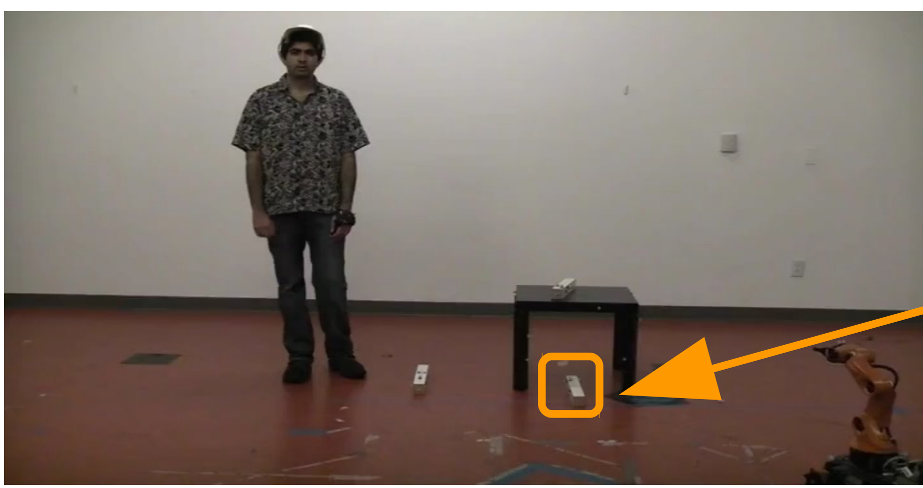
VP *NP*

Pick up NP *PP*

the white leg *TO* NP

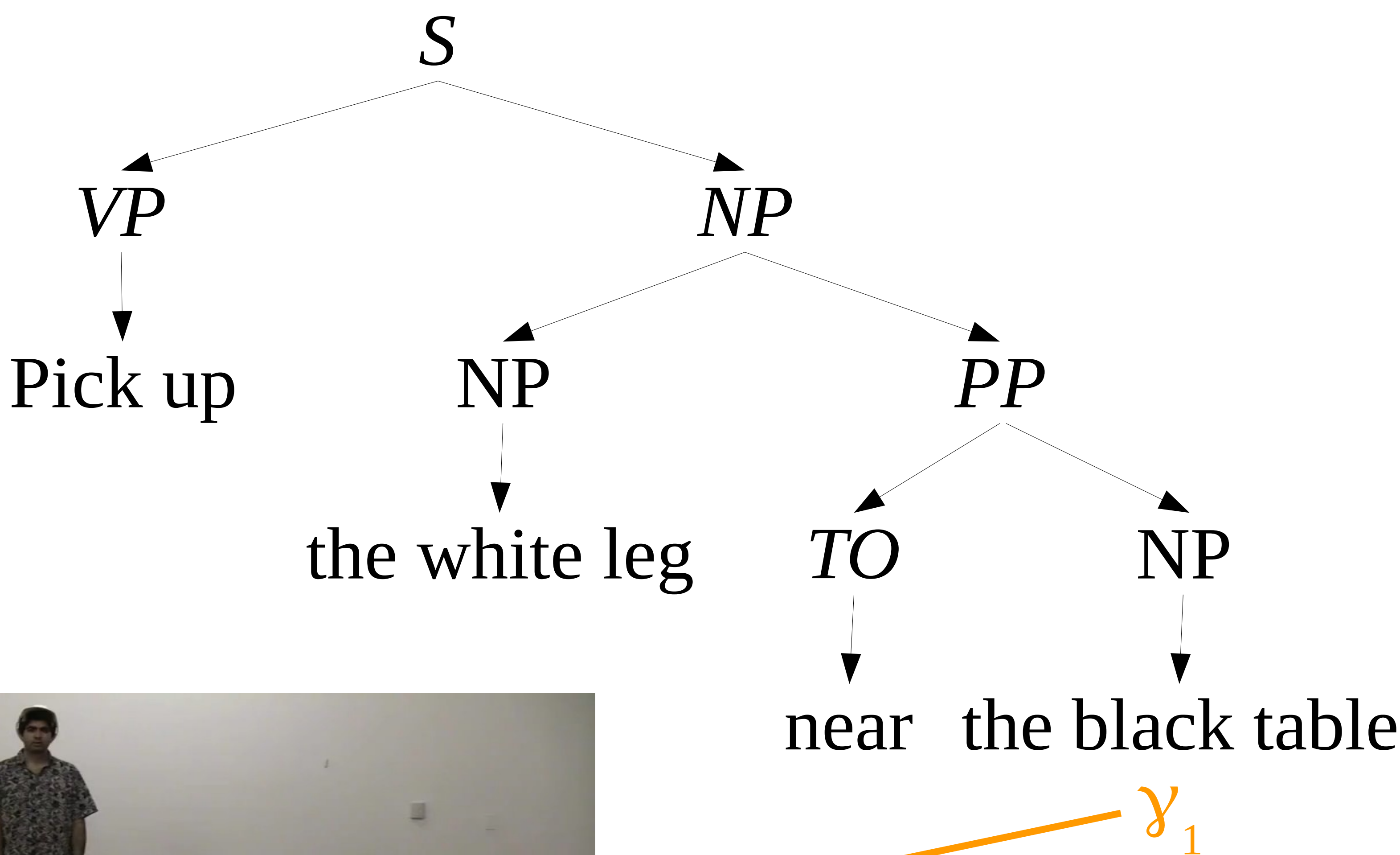
near the black table

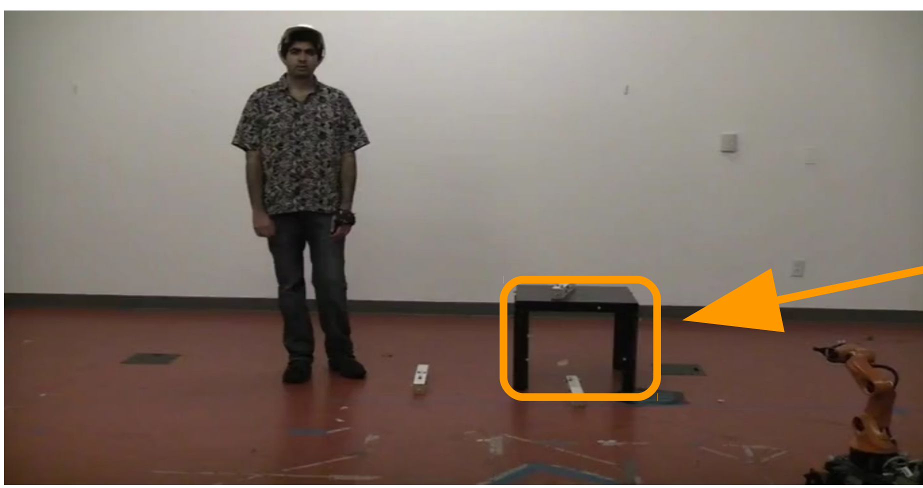
γ_1



$$g_1(\gamma_{1,} \text{the black table}) = 0.1$$

Searching for Sentences

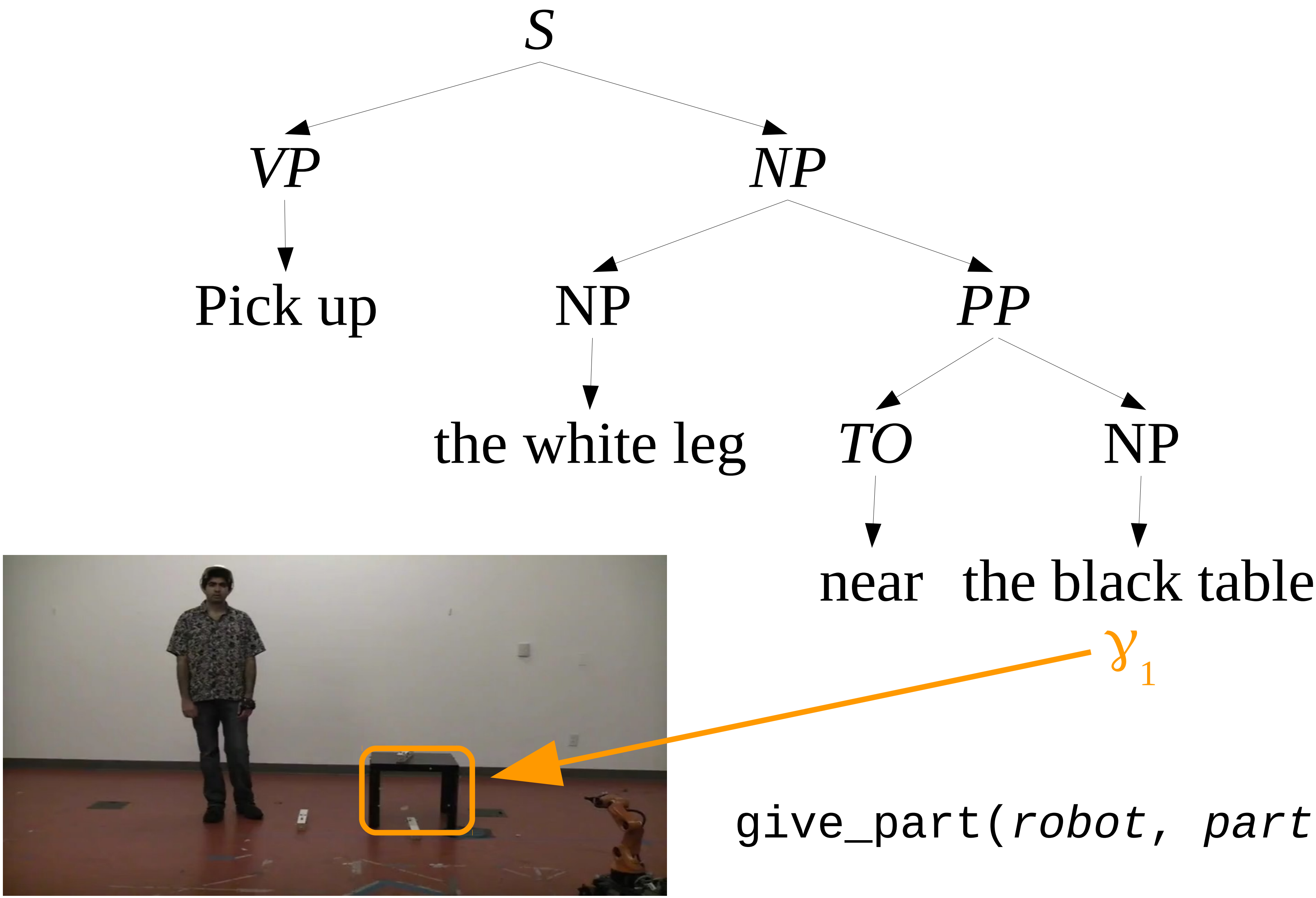




$$g_1(\gamma_1, \text{the black table}) = 0.9$$

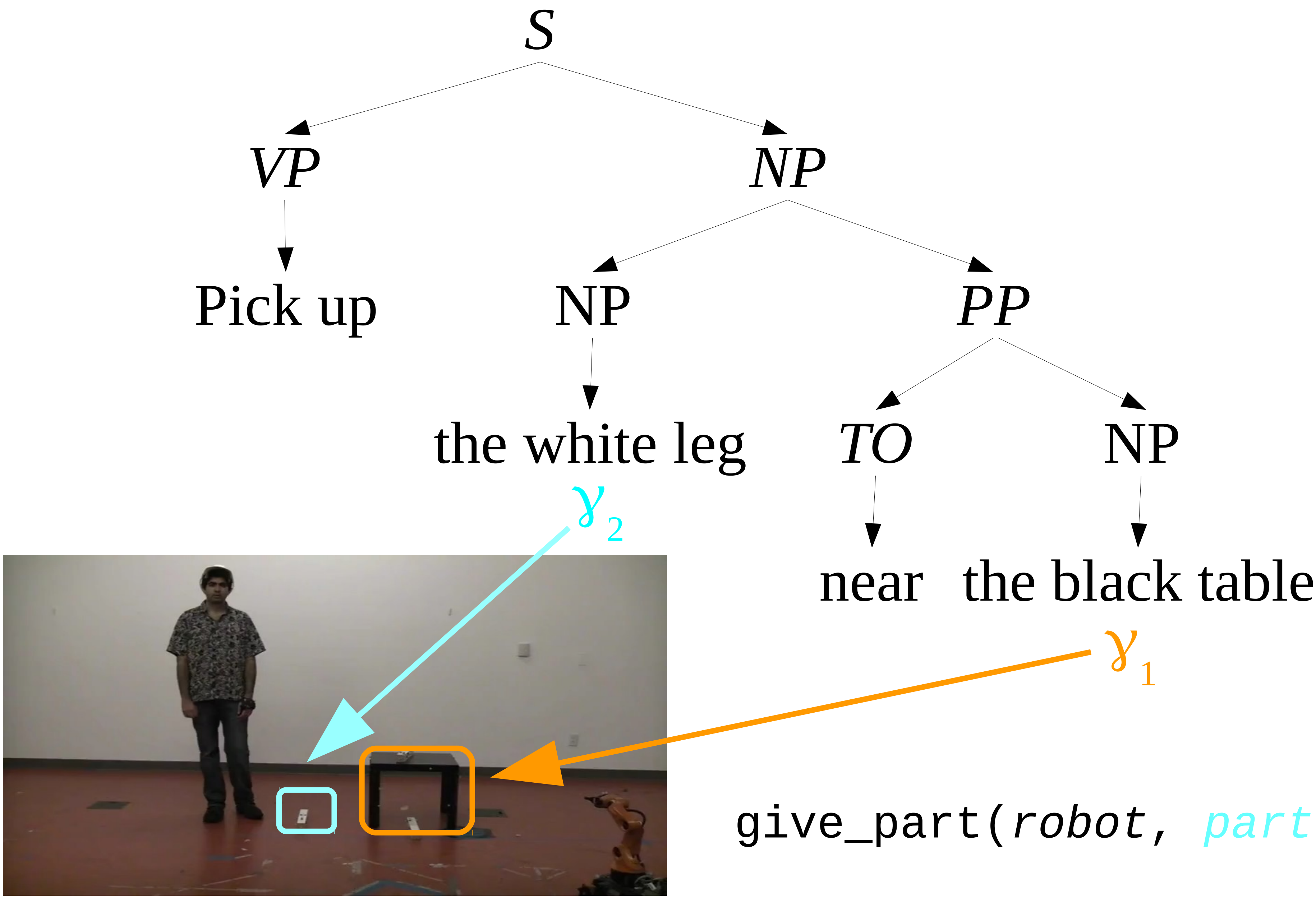
Beam search

Searching for Sentences



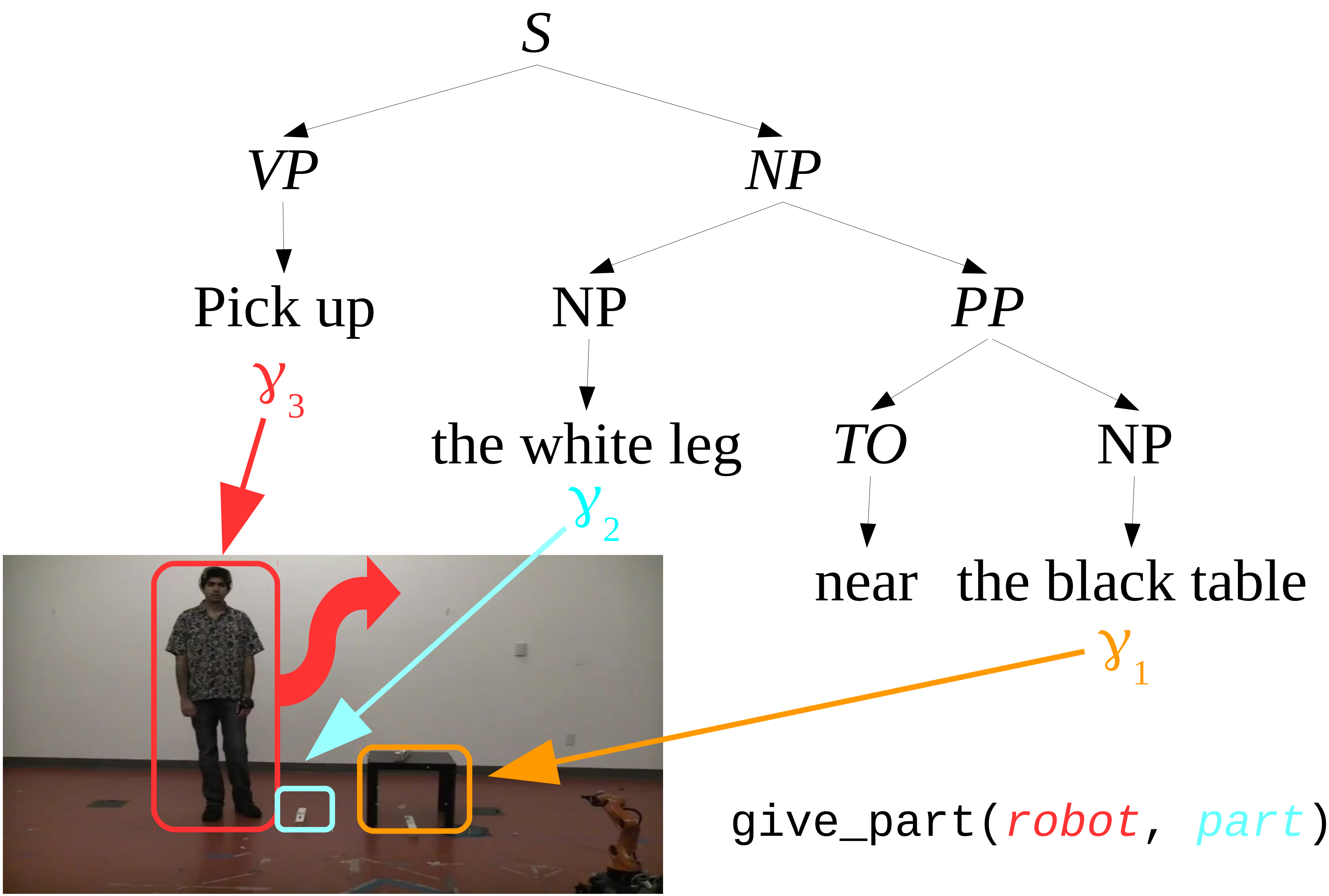
`give_part(`*`robot`*`, `*`part`*`)`

Searching for Sentences




```
give_part(robot, part)
```

Searching for Sentences



Forward Semantics

Understanding: What groundings match the language?

$$\underset{\gamma_1 \ldots \gamma_N}{argmax} \; \frac{1}{Z} \prod_i g_i(\gamma_1 \ldots \gamma_N, language)$$

Generation: What language specifies the groundings?

$$\underset{language}{argmax} \; f(\gamma_1 \ldots \gamma_N, language)$$

γ_k are *groundings*, or objects, places, paths, and events in the external world. Each γ_k corresponds to a constituent phrase in the language input.

Forward Semantics

Understanding: What groundings match the language?

$$\underset{\gamma_1 \ldots \gamma_N}{argmax} \; \frac{1}{Z} \prod_i g_i(\gamma_1 \ldots \gamma_N, language)$$

Generation: What language specifies the groundings?

$$\underset{language, \gamma_1 \ldots \gamma_k}{argmax} \; f(\gamma_1 \ldots \gamma_N, language)$$

γ_k are *groundings*, or objects, places, paths, and events in the external world. Each γ_k corresponds to a constituent phrase in the language input.

Inverse Semantics

$$\underset{language,\gamma_1\ldots\gamma_k}{\operatorname{argmax}} \; p(\gamma_1 \ldots \gamma_N | language)$$

$$\underset{language,\gamma_1\ldots\gamma_k}{\operatorname{argmax}} \; \frac{\prod_i g_i(\gamma_1 \ldots \gamma_N, language)}{\sum_{\Gamma'} \prod_i g_i(\gamma_1' \ldots \gamma_N', language)}$$

Equivalent to the problem of language understanding!

Inverse Semantics

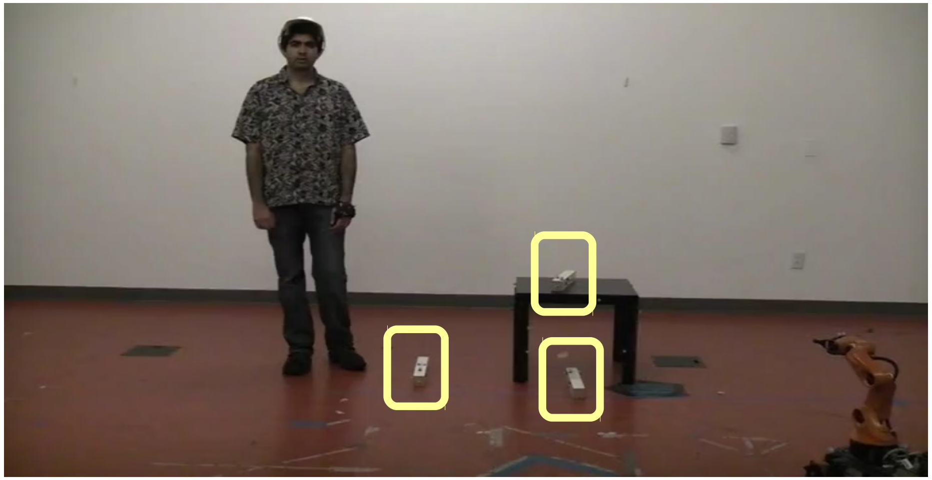
$$\underset{language,\gamma_1\ldots\gamma_k}{argmax} \frac{\prod_i g_i(\gamma_1\ldots\gamma_N, language)}{\sum_{\Gamma'}\prod_i g_i(\gamma_1'\ldots\gamma_N', language)}$$

Inverse Semantics

$$\underset{language,\gamma_1 \ldots \gamma_k}{\operatorname{argmax}} \frac{\prod_i g_i(\gamma_1 \ldots \gamma_N, language)}{\sum_{\Gamma'} \prod_i g_i(\gamma_1' \ldots \gamma_N', language)}$$

$$\underset{language,\gamma_1 \ldots \gamma_k}{\operatorname{argmax}} \frac{0.9}{0.9+0.9+0.9+K}$$

$$\underset{language,\gamma_1 \ldots \gamma_k}{\operatorname{argmax}} 0.33$$



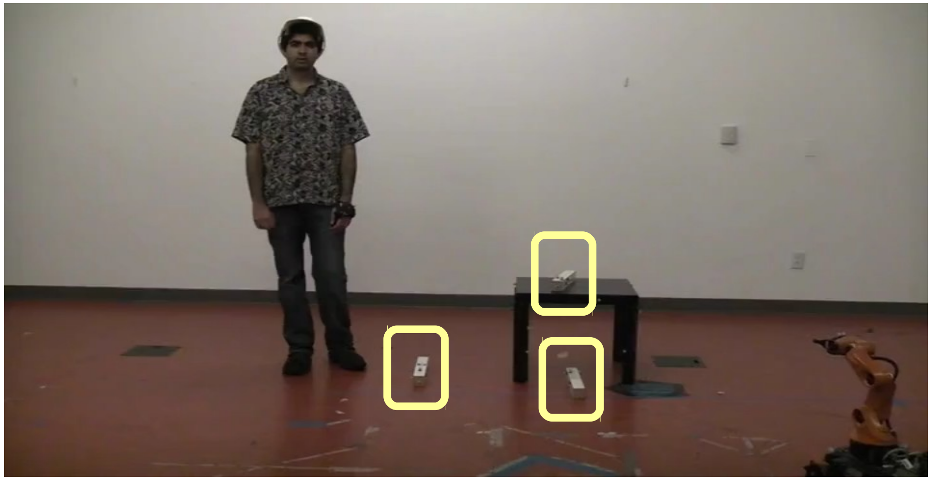
give the robot the white leg.

Inverse Semantics

$$\underset{language,\gamma_1\ldots\gamma_k}{\operatorname{argmax}} \frac{\prod_i g_i(\gamma_1 \ldots \gamma_N, language)}{\sum_{\Gamma'} \prod_i g_i(\gamma_1' \ldots \gamma_N', language)}$$

$$\underset{language,\gamma_1\ldots\gamma_k}{\operatorname{argmax}} \frac{0.7}{0.7+0.1+0.1+K}$$

$$\underset{language,\gamma_1\ldots\gamma_k}{\operatorname{argmax}} 0.78$$



give the robot the white leg that is on the black table.

Solution Overview

1. Detect the failure.
2. Infer an action to fix the problem.
3. Infer a natural language sentence describing the action.
4. Replan after the human has provided help based on the updated state.

Solution Overview

1. Detect the failure.
2. Infer an action to fix the problem.
3. Infer a natural language sentence describing the action.
4. Replan after the human has provided help based on the updated state.

Replan From Current State

- Human may have
  - helped differently than expected.
  - failed to help in time.
  - caused side-effects.

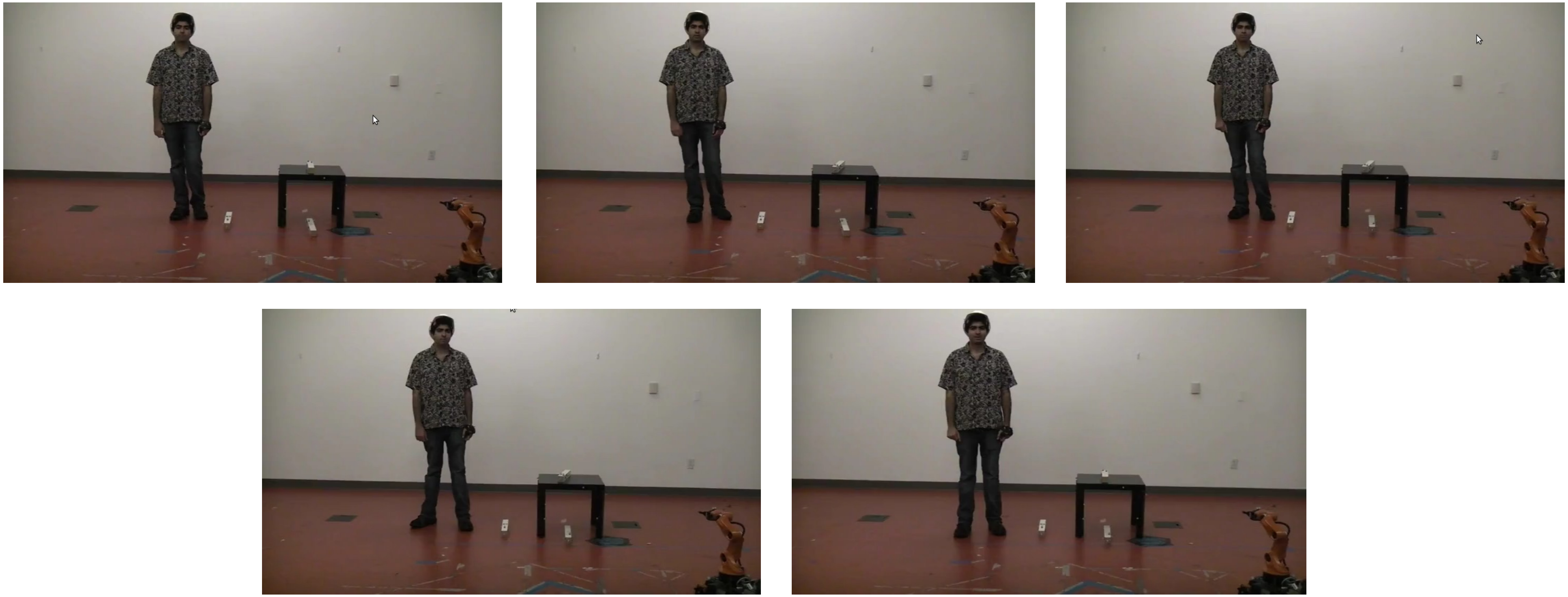
Evaluation Overview

- Does the inverse-semantics method generate requests that are easier to understand than other methods?
  - Online corpus-based evaluation.
- Does our approach work in an end-to-end-system?
  - User study in a real-world furniture assembly system.

Evaluation Overview

- Does the inverse-semantics method generate requests that are easier to understand than other methods?
  - Online corpus-based evaluation.
- Does our approach work in an end-to-end-system?
  - User study in a real-world furniture assembly system.

Corpus-Based Evaluation



“Give me the white leg that is on the black table.”

Which video is the best response to the natural language request?

Corpus-Based Evaluation

"Give me the white leg that is on the black table."

Which video is the best response to the natural language request?

Corpus-Based Evaluation: Results

Generation Algorithm	Example	Success Rate(%)	
Chance		20	
“Help me”	*“Help me.”*	21	±8.0

Corpus-Based Evaluation: Results

Generation Algorithm	Example	Success Rate(%)	
Chance		20	
“Help me”	*“Help me.”*	21	±8.0
Hand-written Request	*“Take the table leg that is on the table and place it in the robot’s hand.”*	94	± 4.7

Corpus-Based Evaluation: Results

Generation Algorithm	Example	Success Rate(%)	
Chance		20	
“Help me”	*“Help me.”*	21	±8.0
Templates	*“Hand me part 2.”*	47	±5.7
Hand-written Request	*“Take the table leg that is on the table and place it in the robot’s hand.”*	94	± 4.7

Evaluation Overview

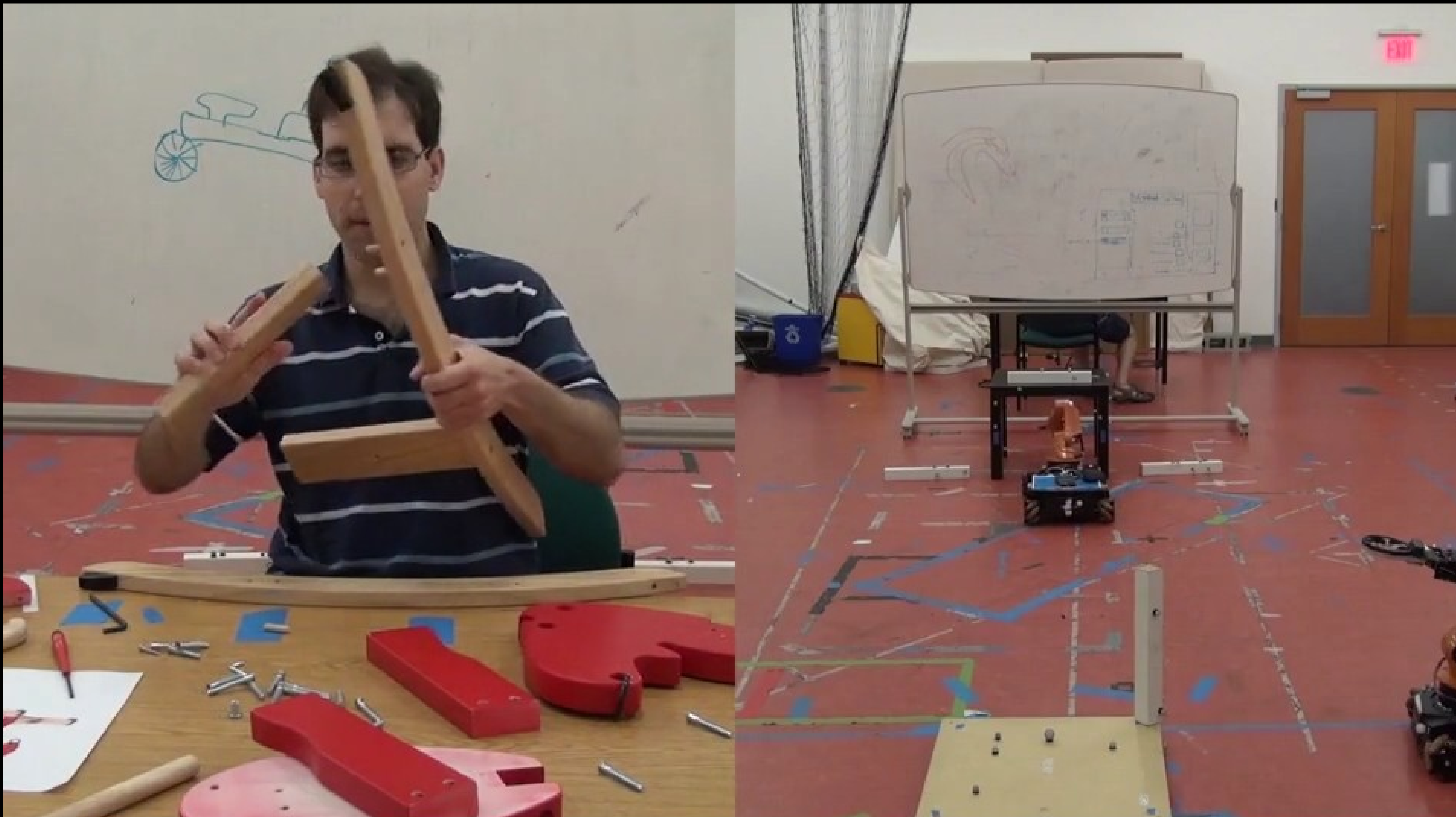
- Does the inverse-semantics method generate requests that are easier to understand than other methods?
  - Corpus-based Evaluation on AMT.
- Does our approach work in an end-to-end-system?
  - User-study in a real-world furniture assembly system.

Evaluation Overview

- Does the inverse-semantics method generate requests that are easier to understand than other methods?
  - Corpus-based Evaluation on AMT.
- Does our approach work in an end-to-end-system?
  - User-study in a real-world furniture assembly system.

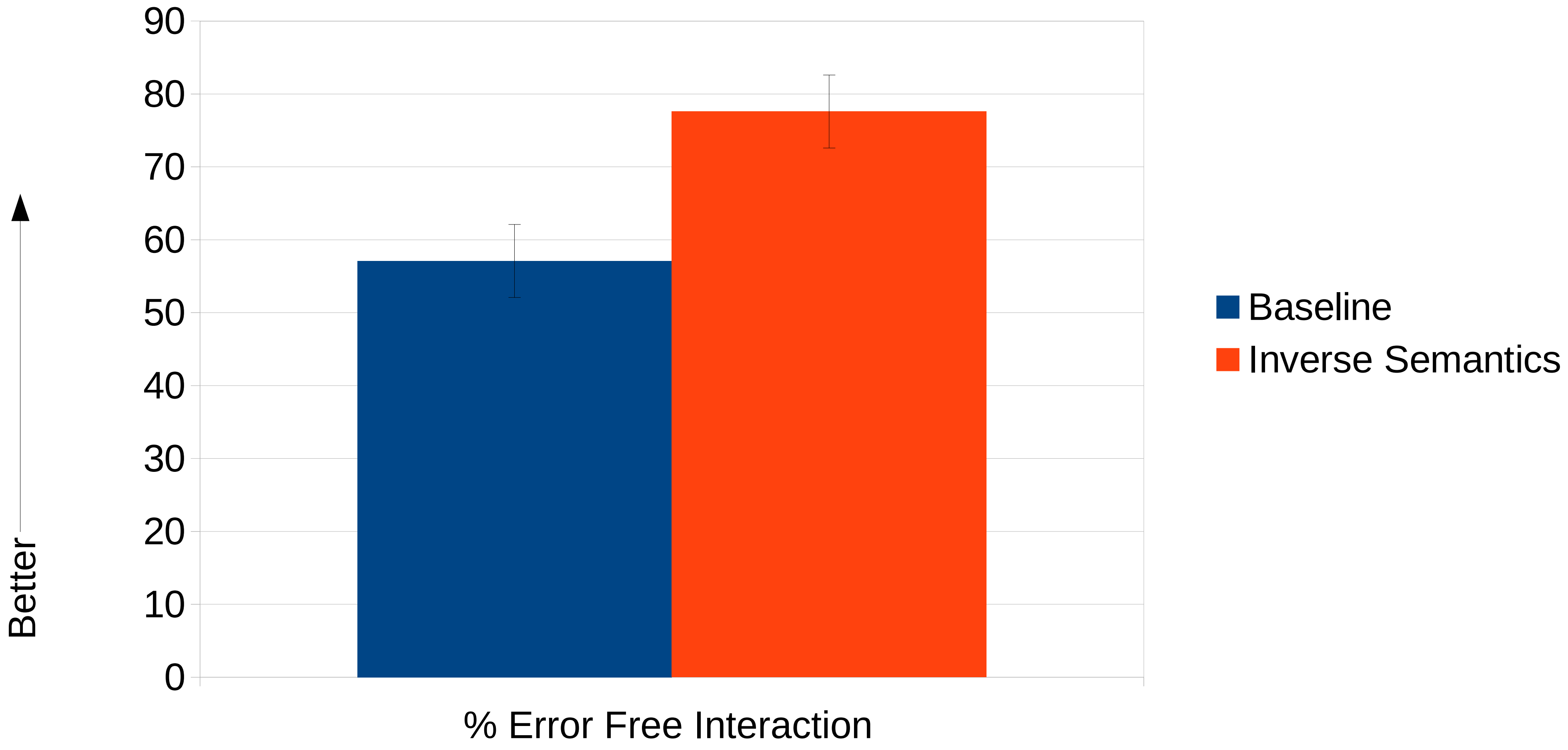
User Study

- Human-robot team assembled Ikea furniture in parallel for 15 minutes.
- Robots asked for help when they encountered failure.
  - Three staged failures (e.g., a part out of reach on the table.)
  - Many unstaged failures (e.g., a part slipped out of the robot's grasp.)
- Human provided whatever help they felt was appropriate.
- Robots continued operating autonomously.

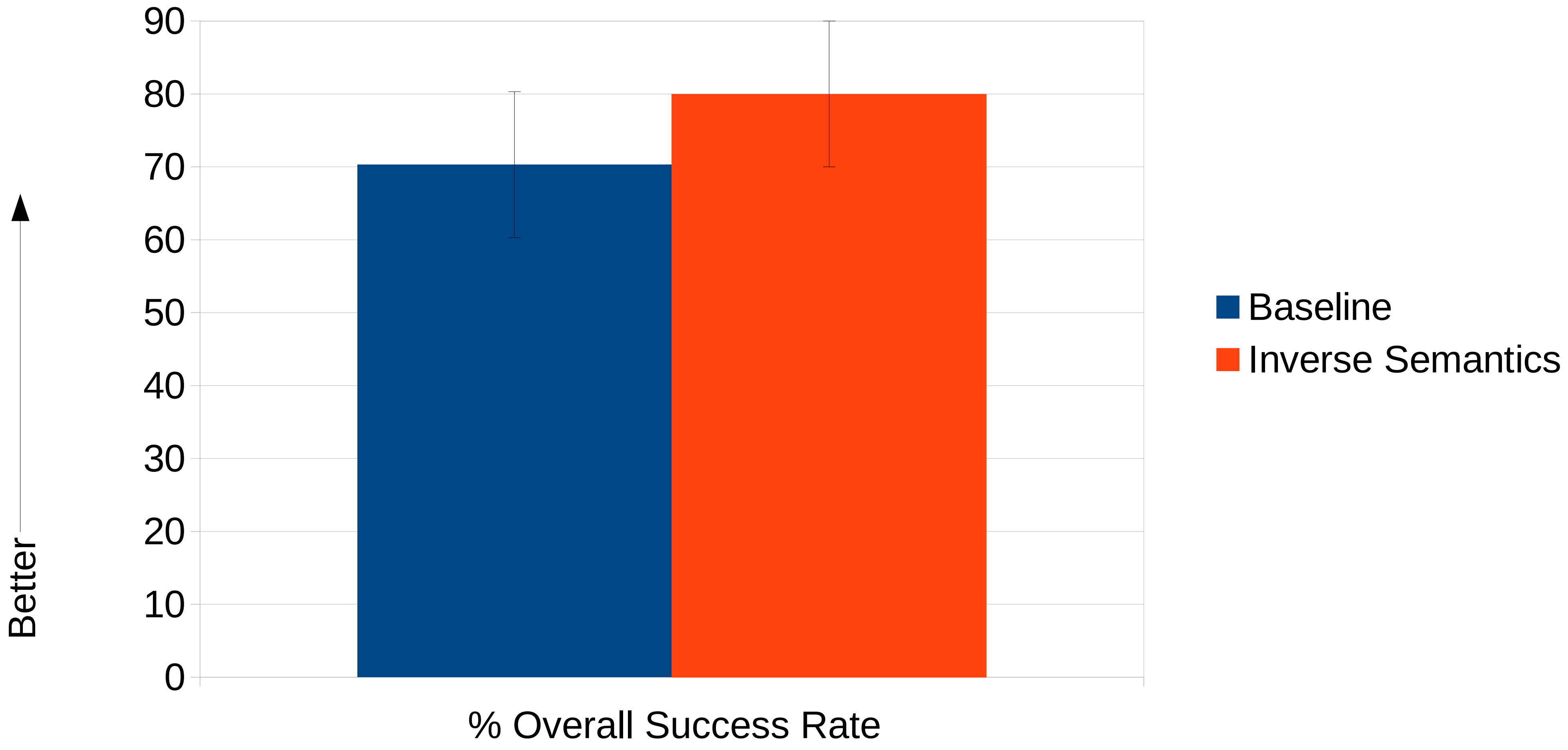

EXIT

User Study Results

Objective Metrics

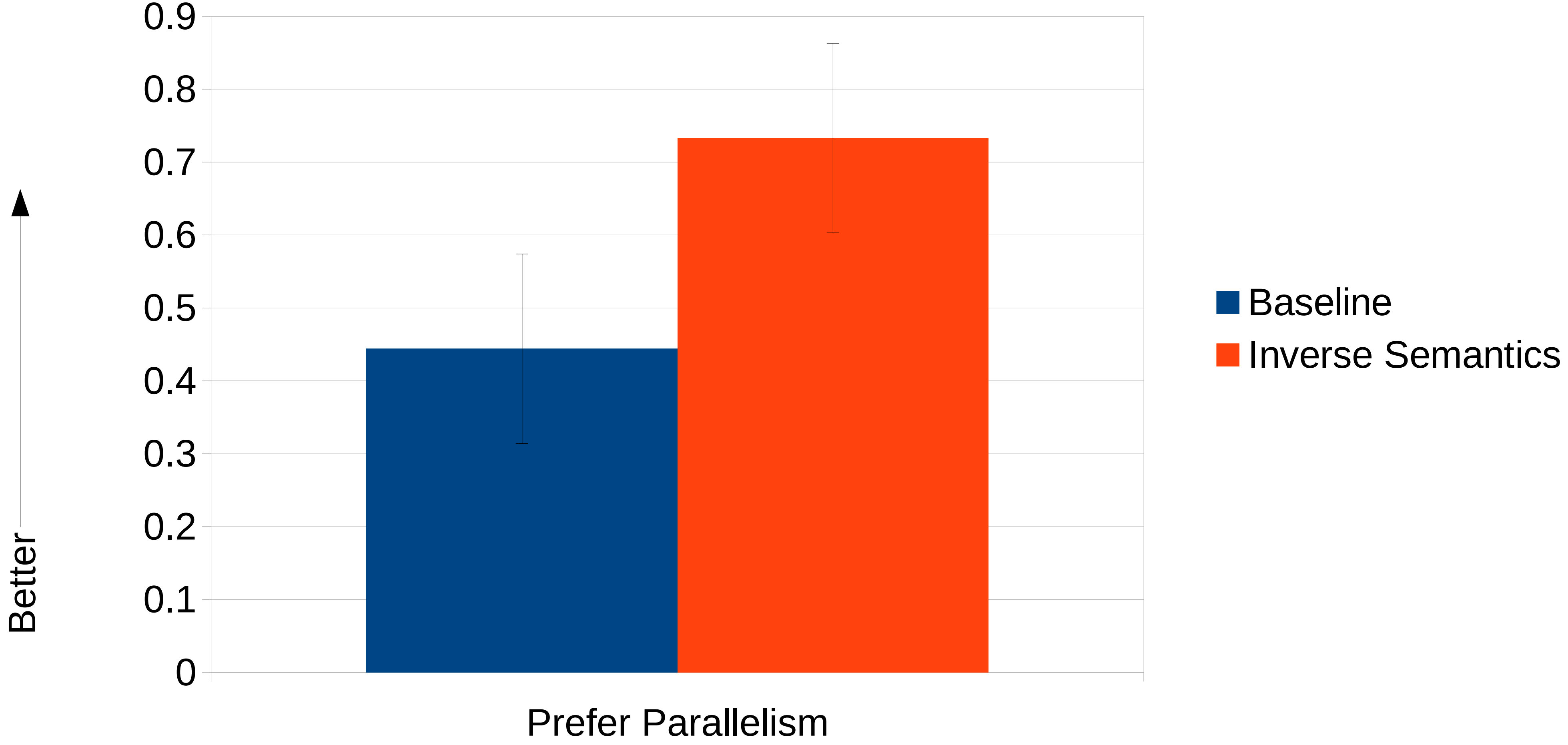


User Study Results
Objective Metrics



User Study Results

Subjective Metrics



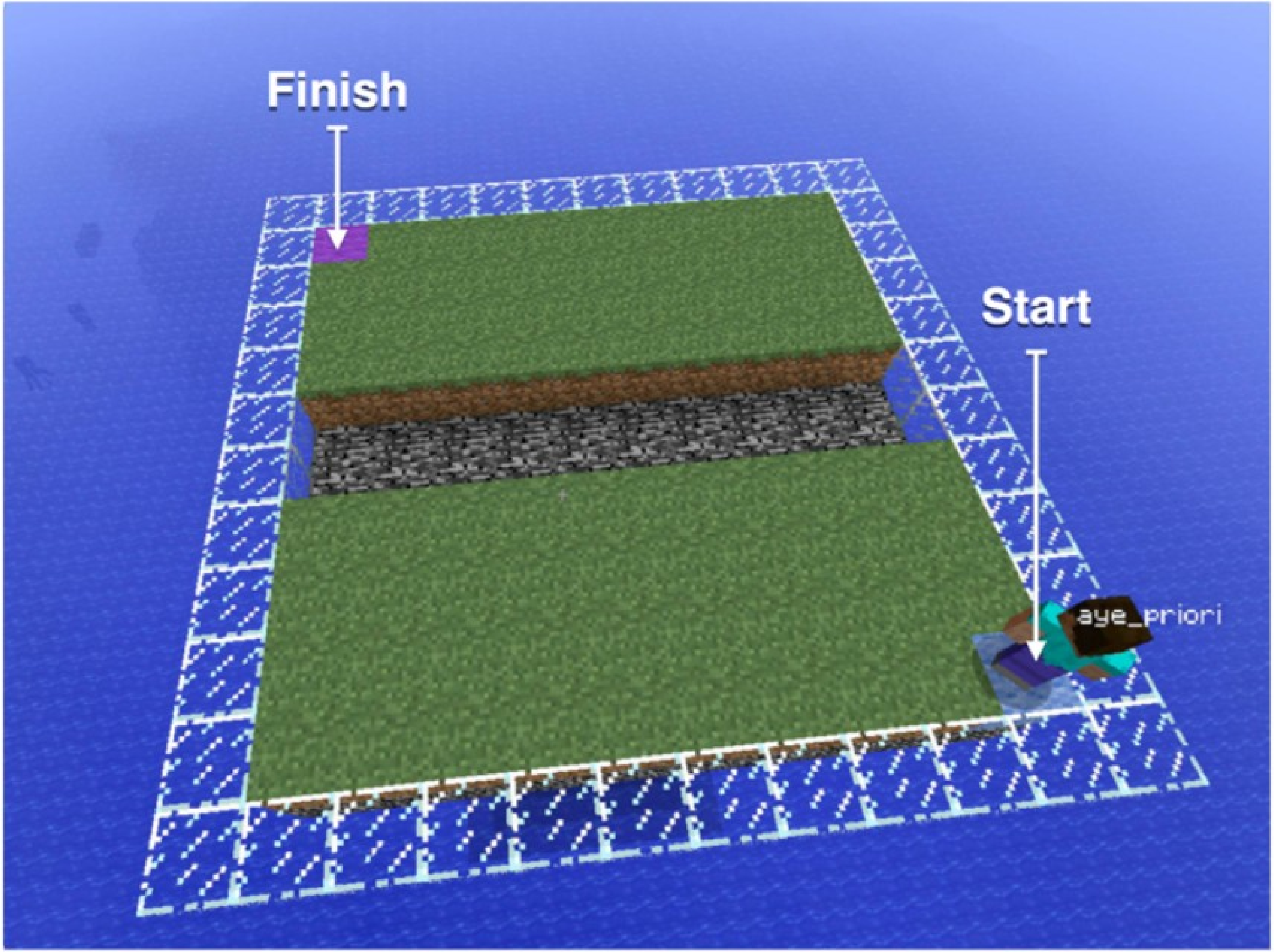
How can robots collaborate with people using natural language?

- Following instructions.
  - “Put the metal crate on the truck.”
- Asking questions.
  - “What does 'the metal crate' refer to?”
- Requesting Help.
  - “Hand me the black leg that is under the table.”

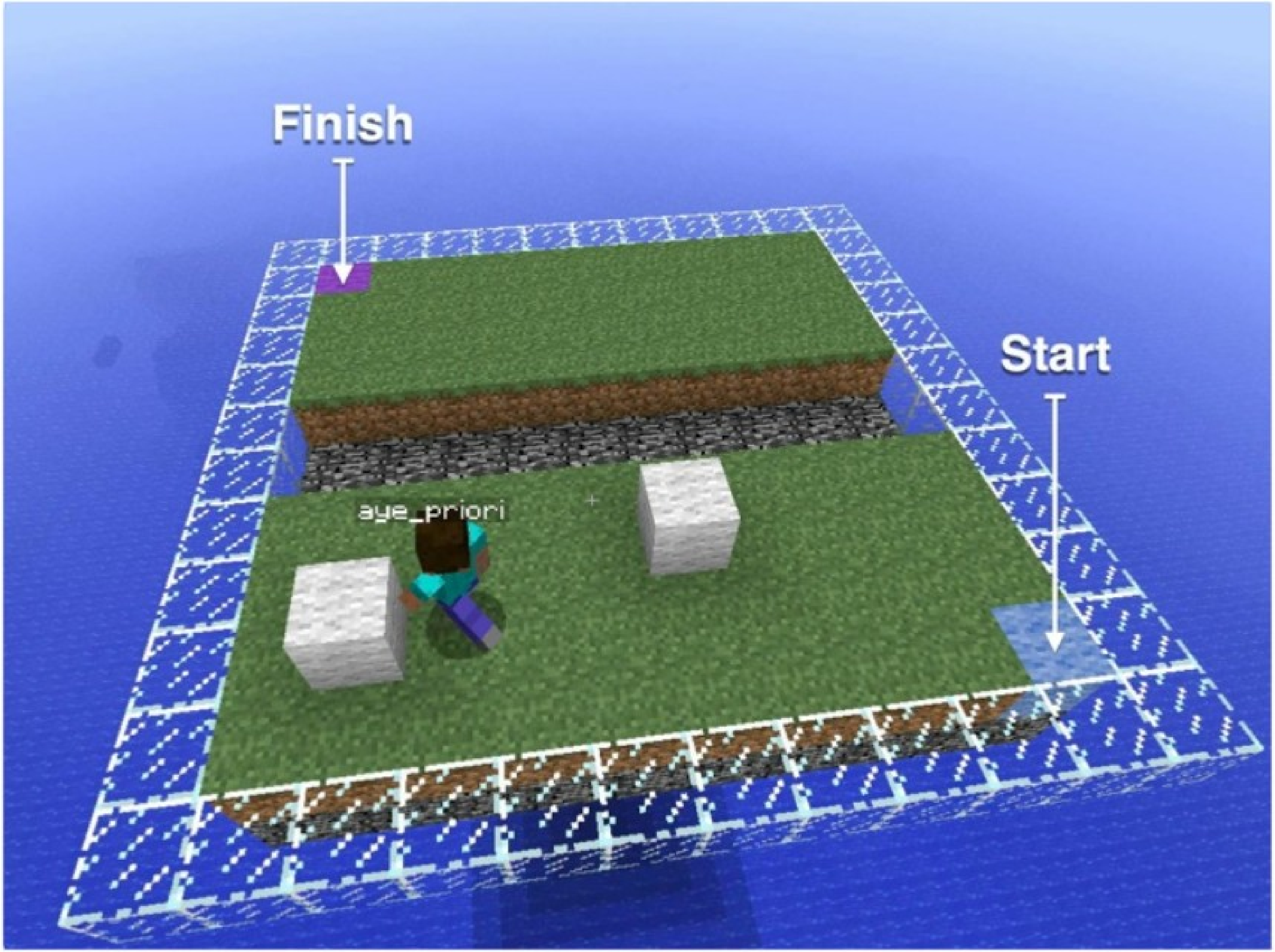
Future Work

- Planning in very large state spaces.
- Grounded dialog.

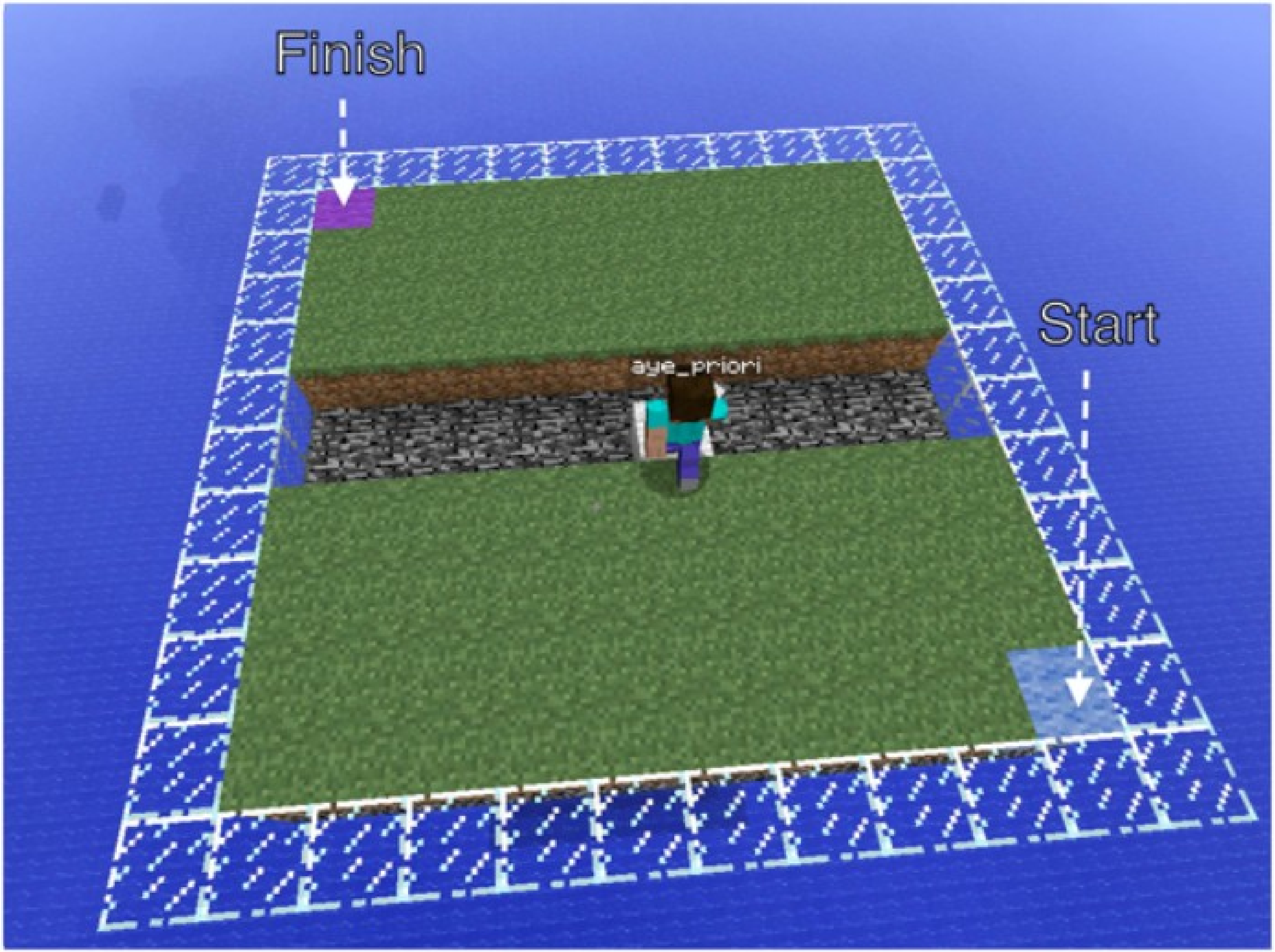
Affordance-Aware Planning



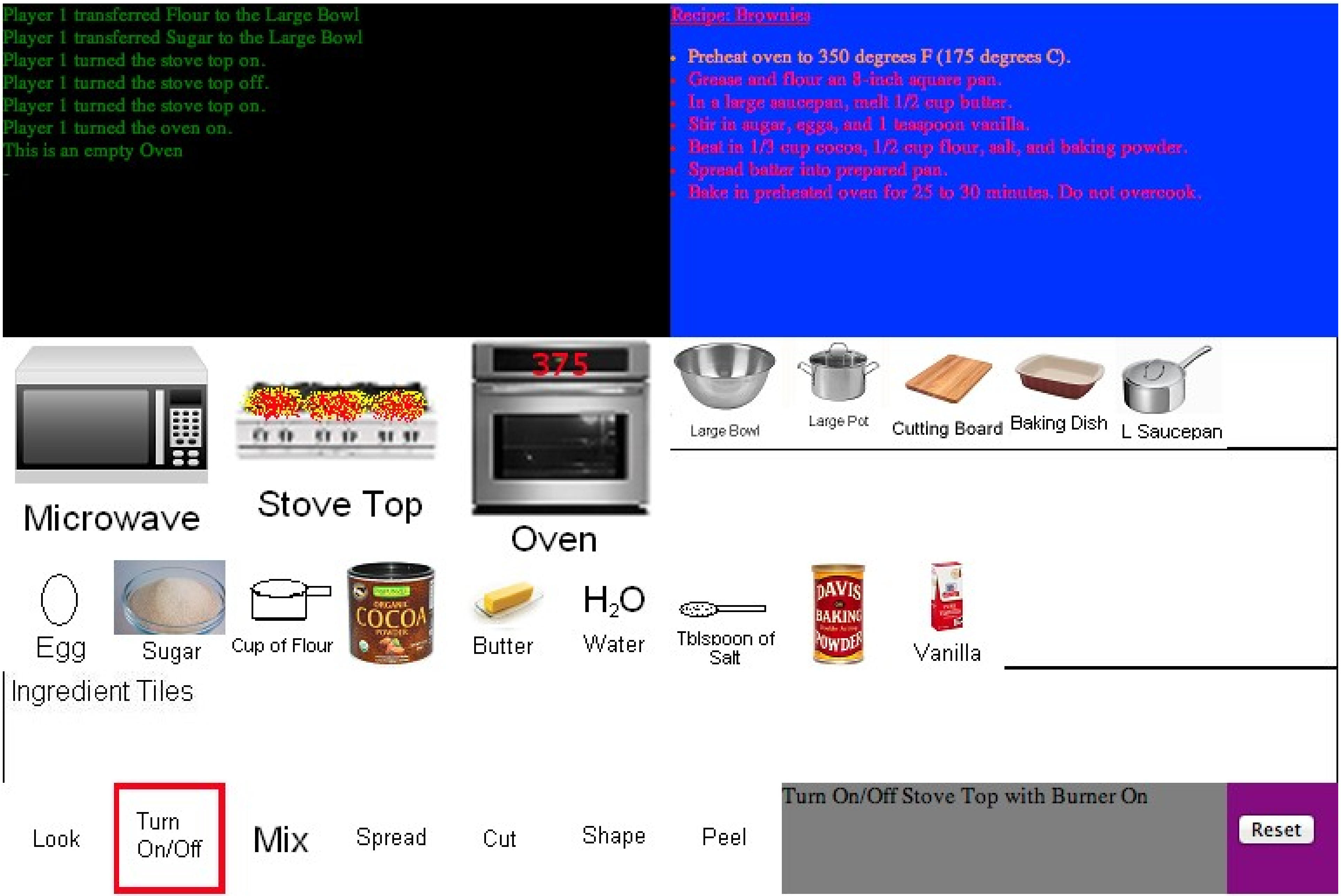
Affordance-Aware Planning



Affordance-Aware Planning



Cooking Game



Multimodal POMDPs for Collaborative Robots

Multimodal POMDPs for Collaborative Robots

- Estimate human's mental state from language, gesture, and perceptual observations.
- Solve POMDPs with very large observation spaces.

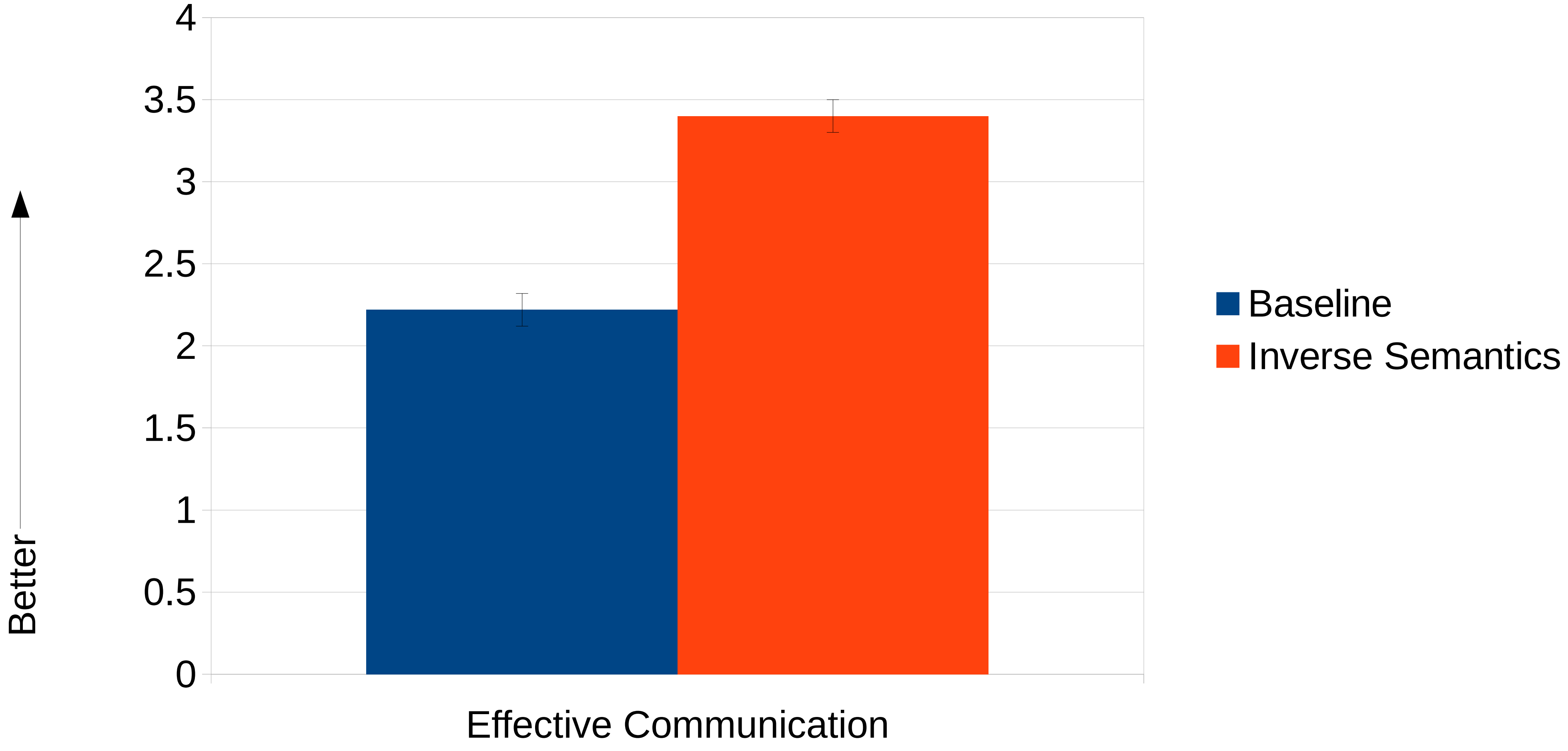
Contributions

- Defined a Bayesian algorithm for generating natural language requests for help.
- Demonstrated that people can understand robotic help requests compared to baselines using a corpus-based evaluation.
- Assessed strengths and limitations in an end-to-end system with a real-world user study.

Stefanie Tellex, Ross Knepper
Adrian Li, Daniela Rus, Nicholas Roy

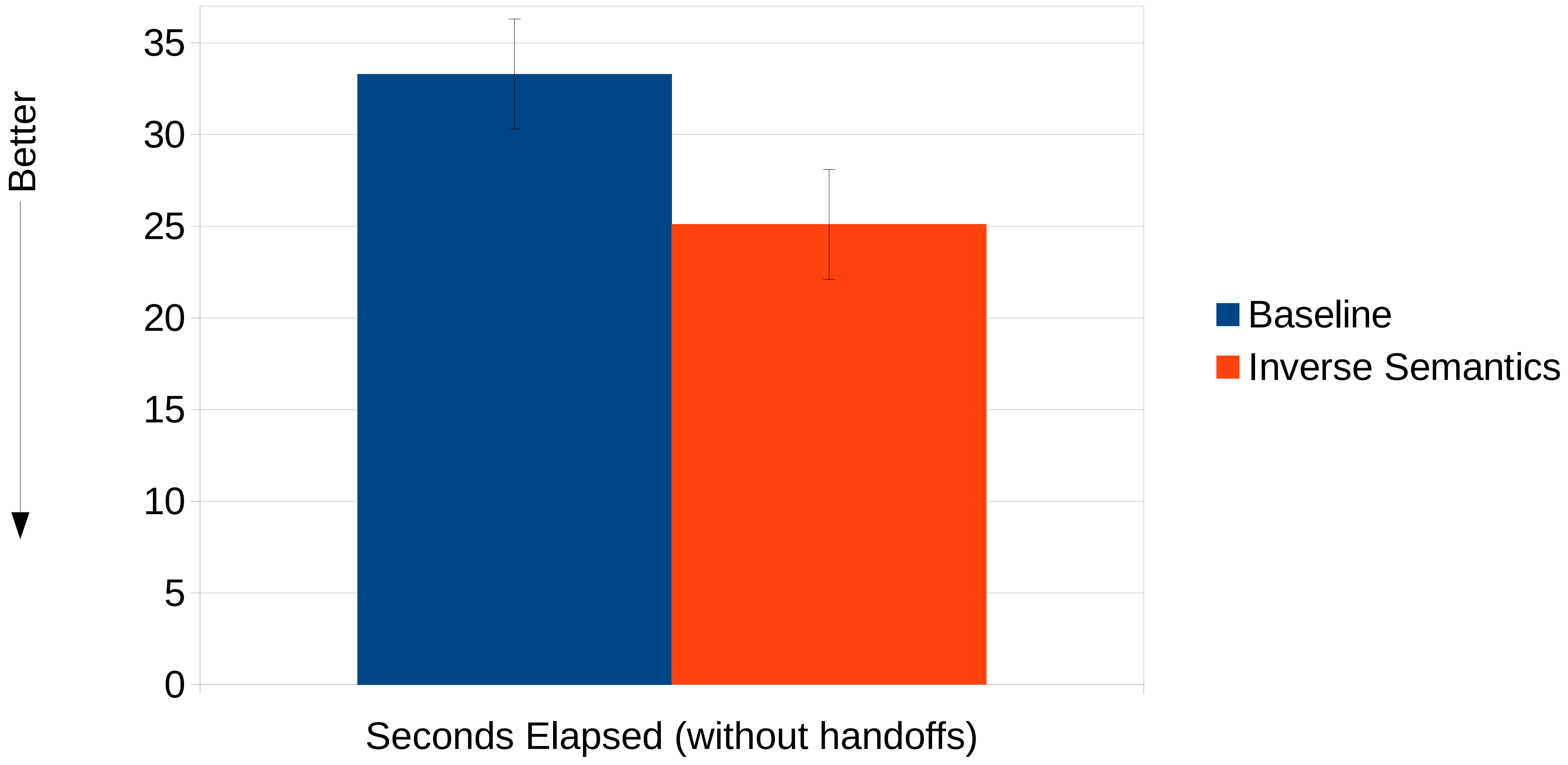
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

User Study Results
Subjective Metrics



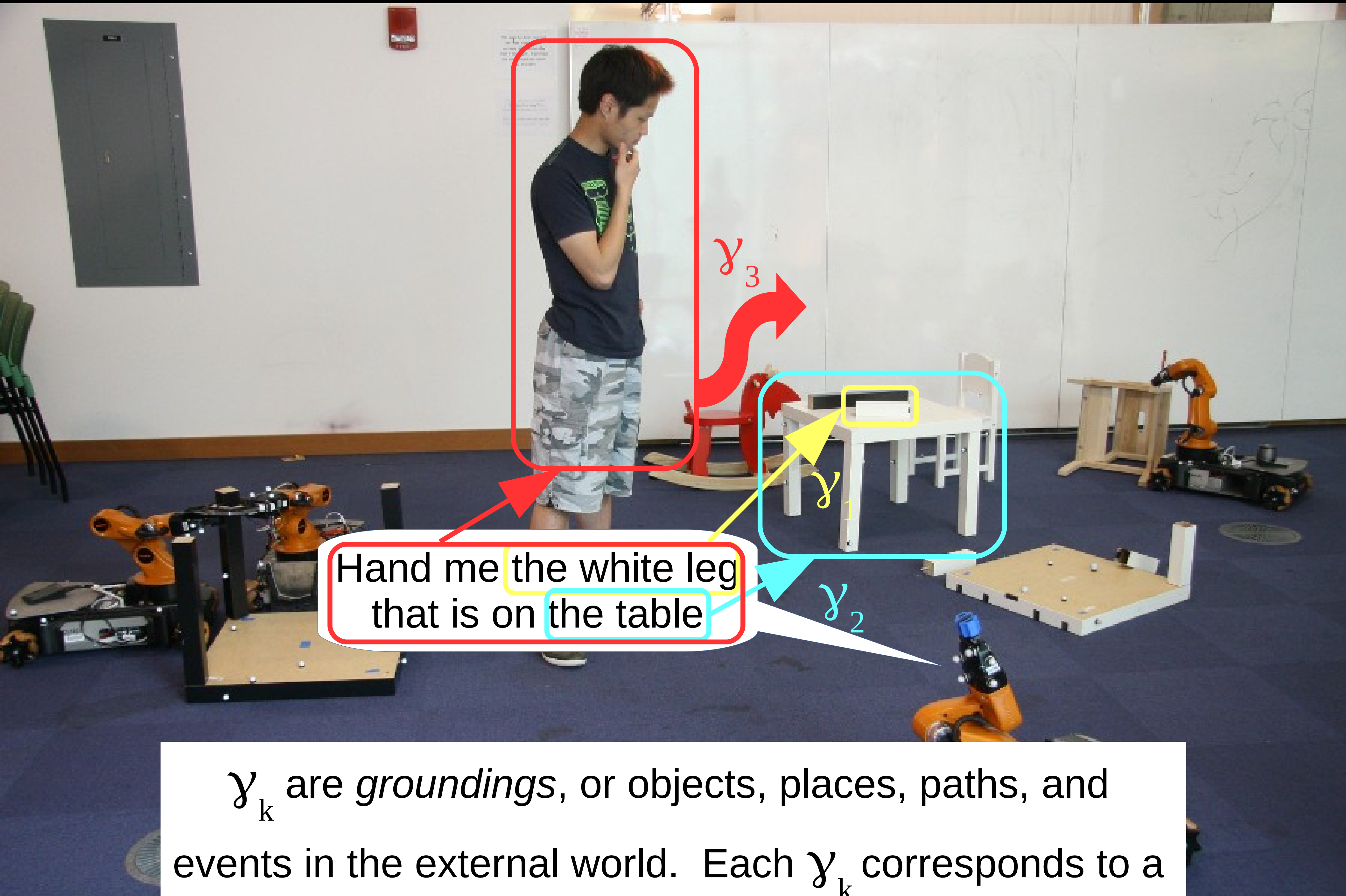
User Study Results

Objective Metrics



Training

- Collect parallel corpus of language paired with groundings.



γ_k are *groundings*, or objects, places, paths, and events in the external world. Each γ_k corresponds to a constituent phrase in the language input.

Corpus-Based Evaluation

- Each user responded to 20 help requests.
- Five algorithms for generating requests.
  - “Help me.”
  - Templates
  - Approach 1
  - Approach 2
  - Handwritten requests.
- Total of 900 trials.

Limitations of Corpus-Based Evaluation

- Ambiguity between “near” and “under.”
- Canned set of 5 choices.
- No indication of how language generation acts in context of the system.

User Comments

“Help me”
“I think if the robot was clearer or I saw it assemble the desk before, I would know more about what it was asking me.”

“Did not really feel like 'working together as a team'”

Inverse Semantics
“More fun than working alone.”

“There was a sense of being much more productive than I would have been on my own.”