Acquiring Rich Models of Objects and Space Through Vision and Natural Language

Matthew Walter
CS & AI Lab, MIT

School of Computer Science
University of Massachusetts, Amherst

February 6, 2013

Robots as Automated Agents

- Advances in:
  - Actuation
  - Planning
  - Control
- Focus:
  - Accuracy
  - Robustness

Courtesy: ABB

Robots as Our Surrogates

- Advances in:
  - Estimation
  - Navigation
  - Planning under uncertainty
- Focus:
  - Accuracy
  - Robustness

[JFR 2008]

RMS Titanic

[IJRR 2006]

Robots as Our Partners


THE BOSTON GLOBE

Business

Science & Innovation

Wheelchairs that listen

PHOTO BY JOHN TLUMACKI/GLOBE STAFF

William Li of MIT (left) and David Hatch at The Boston Home

Now: People Accommodating Robots

Courtesy: AeroVironment

Courtesy: US Army

Where We Need to Be

Where We Need to Be

Representational Divide

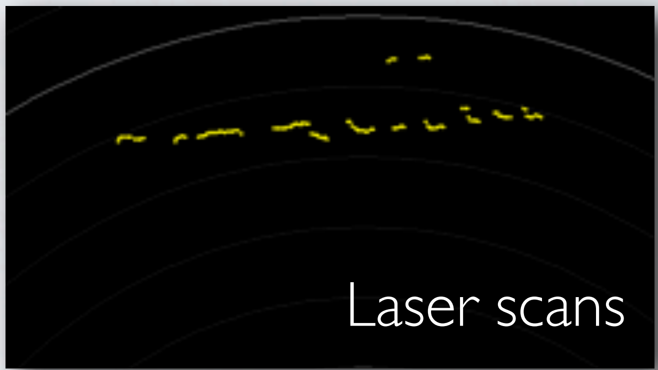
A robot's view of the world is very different from our own

People

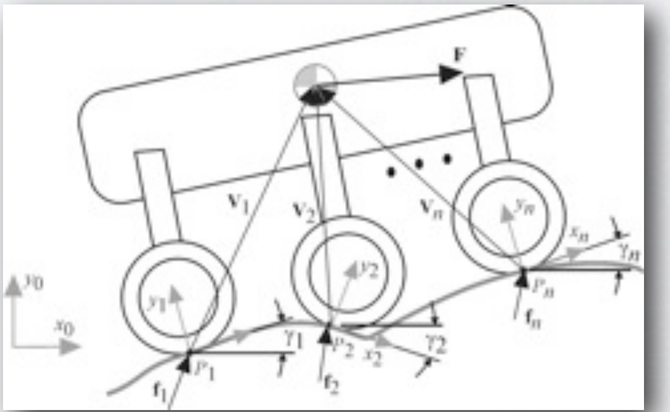
- Objects
- Places
- Actions
- People
- Events

Robots

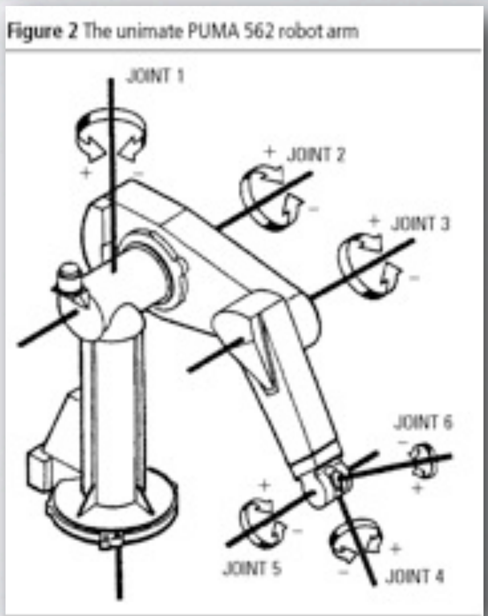
Images



Laser scans



Wheel torques



Joint angles

Matthew Walter

Vision: Shared Situational Awareness

Spatially extended, temporally persistent
model of the robot's surround

- Objects: Identity, properties, relations, actions
- Places: Function, identity, connectivity
- People: Locations, behavior, gestures
- Actions: Means of interacting with the world

Vision: Learning Shared Representations

- Reason over shared knowledge representations
- Acquire situational awareness as they interact with the world
- Learn opportunistically from humans

I. Importance of Situational Awareness

II. Persistent Object Awareness with Vision

III. Semantic Map Learning from Natural Language Descriptions

IV. Future Directions

V. Conclusions

Assistive Mobile Manipulation

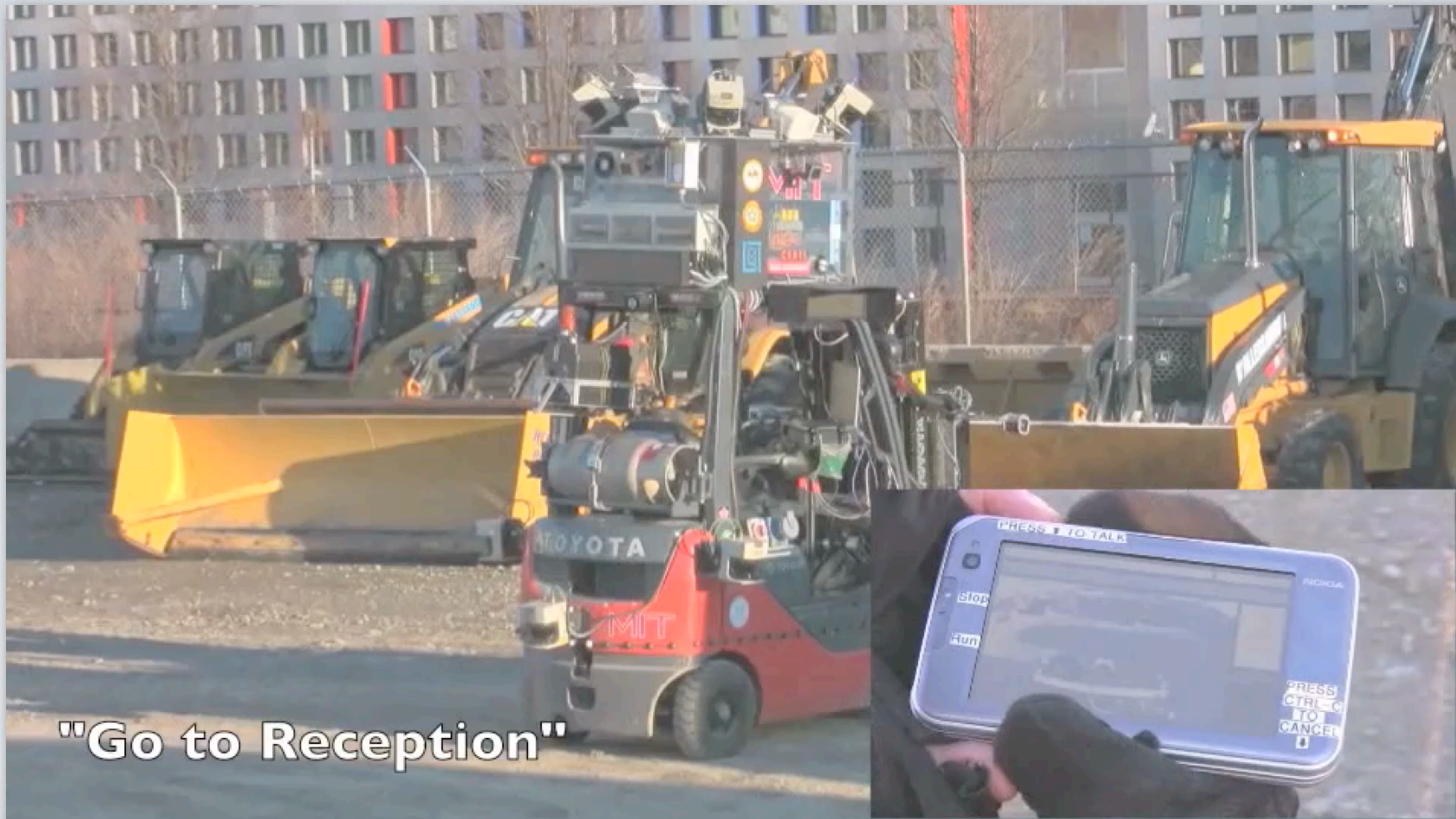
Courtesy: University of Pittsburgh

- Material handling in unstructured environments
- Assisted living for the elderly & disabled

Challenges for Mobile Manipulation

- Unprepared, dynamic environments
  - Coarse localization
  - Uneven terrain
  - Uncontrolled lighting
- Objects unknown a priori
- People everywhere
- Intuitive, human-centered control

Shared Autonomy



Efficient Manipulation via Object Awareness

Courtesy: Kinova Robotics

Object Recognition is Hard!

- Usability requirements:
  - Persistent, reliable detection
  - Efficient object learning
- Challenges:
  - Variable lighting (outdoors)
  - Variable viewpoints
  - Unobserved object relocation
  - Coarse localization

Object Recognition is Hard!

- Usability requirements:
  - Persistent, reliable detection
  - Efficient object learning
- Challenges:
  - Variable lighting (outdoors)
  - Variable viewpoints
  - Unobserved object relocation
  - Coarse localization

Object Recognition is Hard!

- Usability requirements:
  - Persistent, reliable detection
  - Efficient object learning
- Challenges:
  - Variable lighting (outdoors)
  - Variable viewpoints
  - Unobserved object relocation
  - Coarse localization

Object Instance Recognition

	Object category detection [1]	**Visual tracking** [2]	**This work** [3]
Train from one example		✓	✓
Train online		✓	✓
Persistence (hours/days)			✓
Category recognition	✓		
Instance recognition		✓	✓
Real-time performance		✓	✓

[1] Nistér'06, Hoiem'07, Savarese'07
[2] Collins'05, Grabner'08, Kalal'09
[3] CVPRW'10, ISER'10, IJRR'12

One-shot Appearance Learning

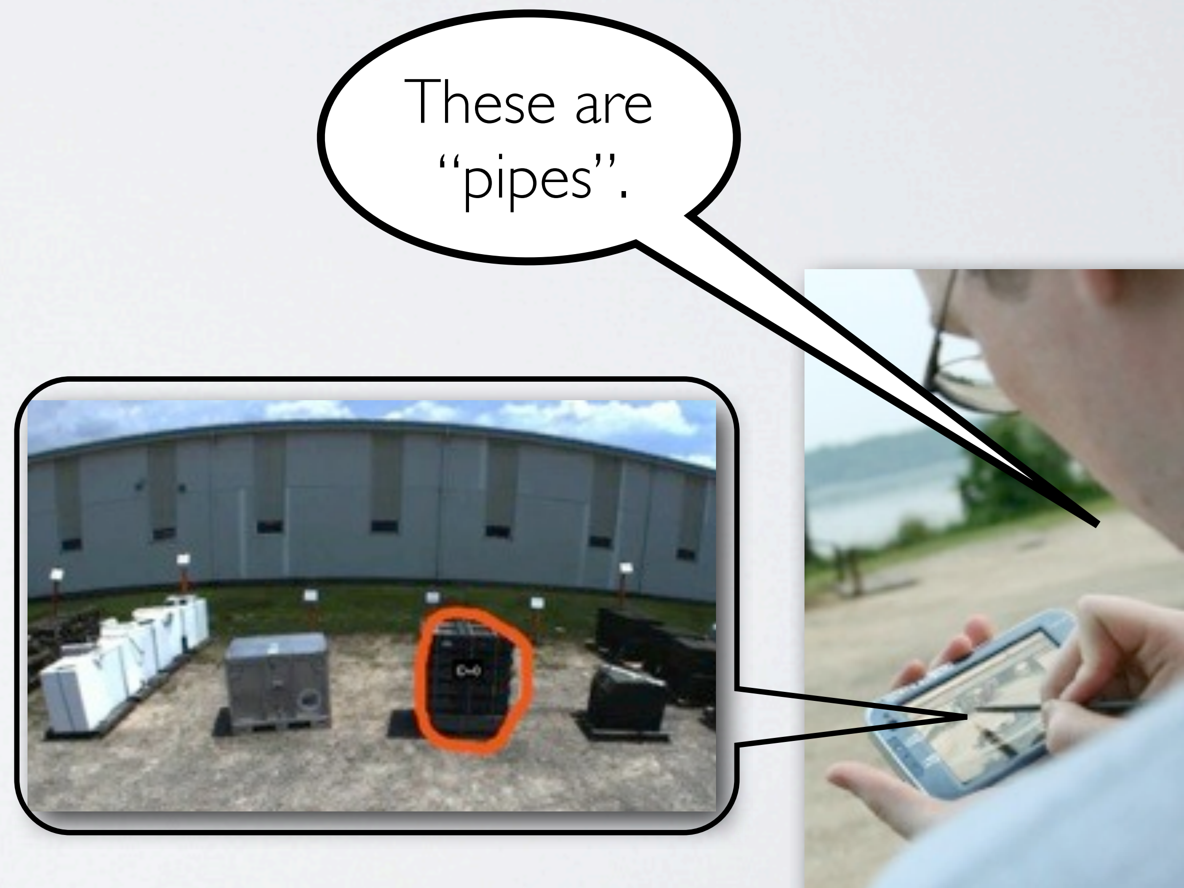
- Key ideas for usable object reacquisition
  - Detect *instances* of the objects used in practice
  - Take advantage of the robot's mobility for learning

One-shot Appearance Learning

- Key ideas for usable object reacquisition
  - Detect *instances* of the objects used in practice
  - Take advantage of the robot's mobility for learning

One-shot Appearance Learning

- User provides a single example of the object’s identity (name & segmentation)
- System bootstraps on user’s example to build an appearance model online
- System takes advantage of robot’s motion to opportunistically capture appearance variations

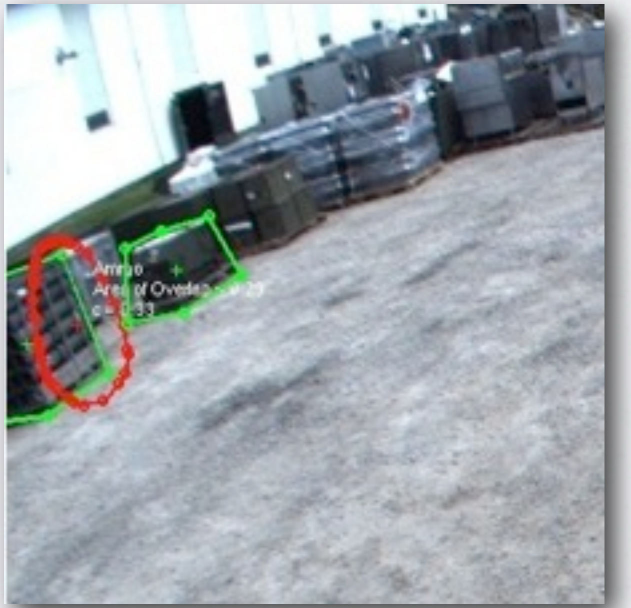


[ISER 2010; IJRR 2012]

One-shot Appearance Learning

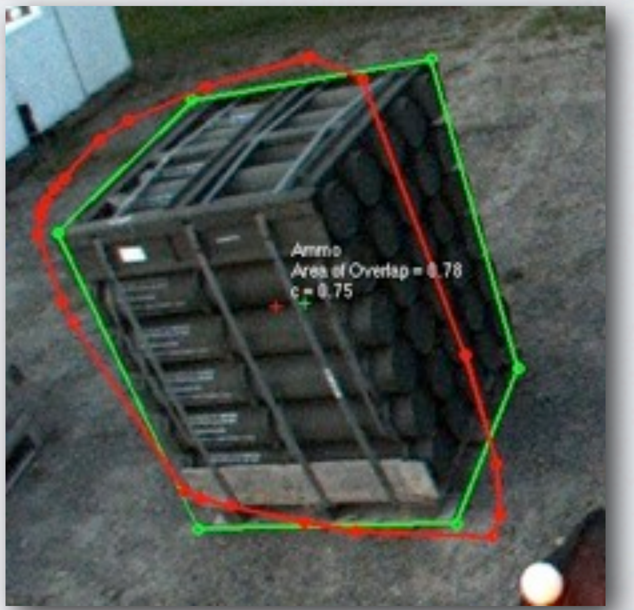
- User provides a single example of the object's identity (name & segmentation)
- System bootstraps on user's example to build an appearance model online
- System takes advantage of robot's motion to opportunistically capture appearance variations

Illumination



Aspect

Relocation



Scale

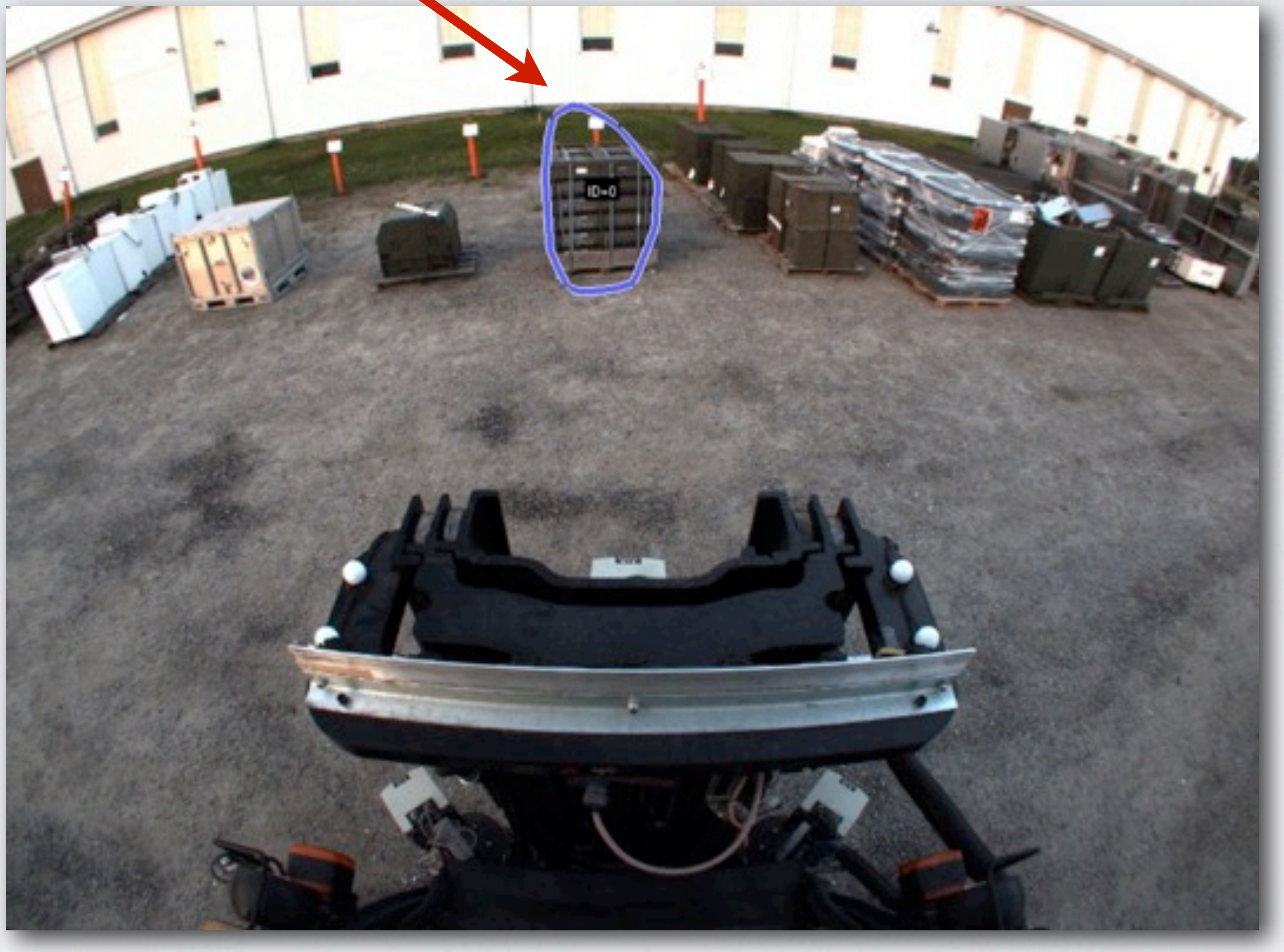
[ISER 2010; IJRR 2012]

Object Reacquisition

Object Reacquisition

Model Instantiation

User circles object in tablet image



Robot's forward-facing camera image

Model Instantiation

SIFT features extracted from initial image

Model Instantiation

View 0 (user gesture)

SIFT features extracted from initial image

Initialize model $\mathcal{M}_i$ to contain single view v_{i1}

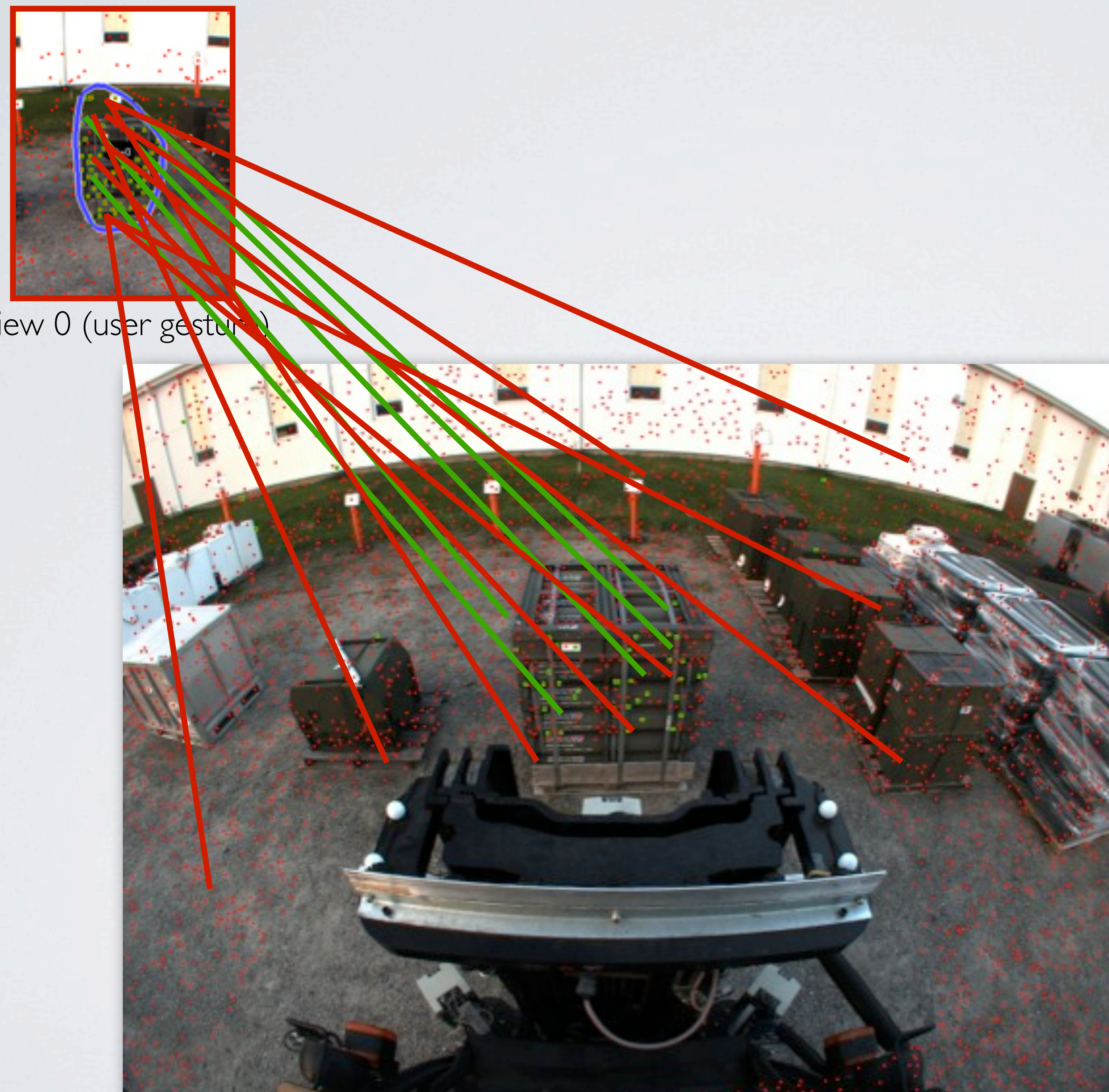
Single-View Matching

View 0 (user gesture)

SIFT features extracted from new image

Extract features and match against all views

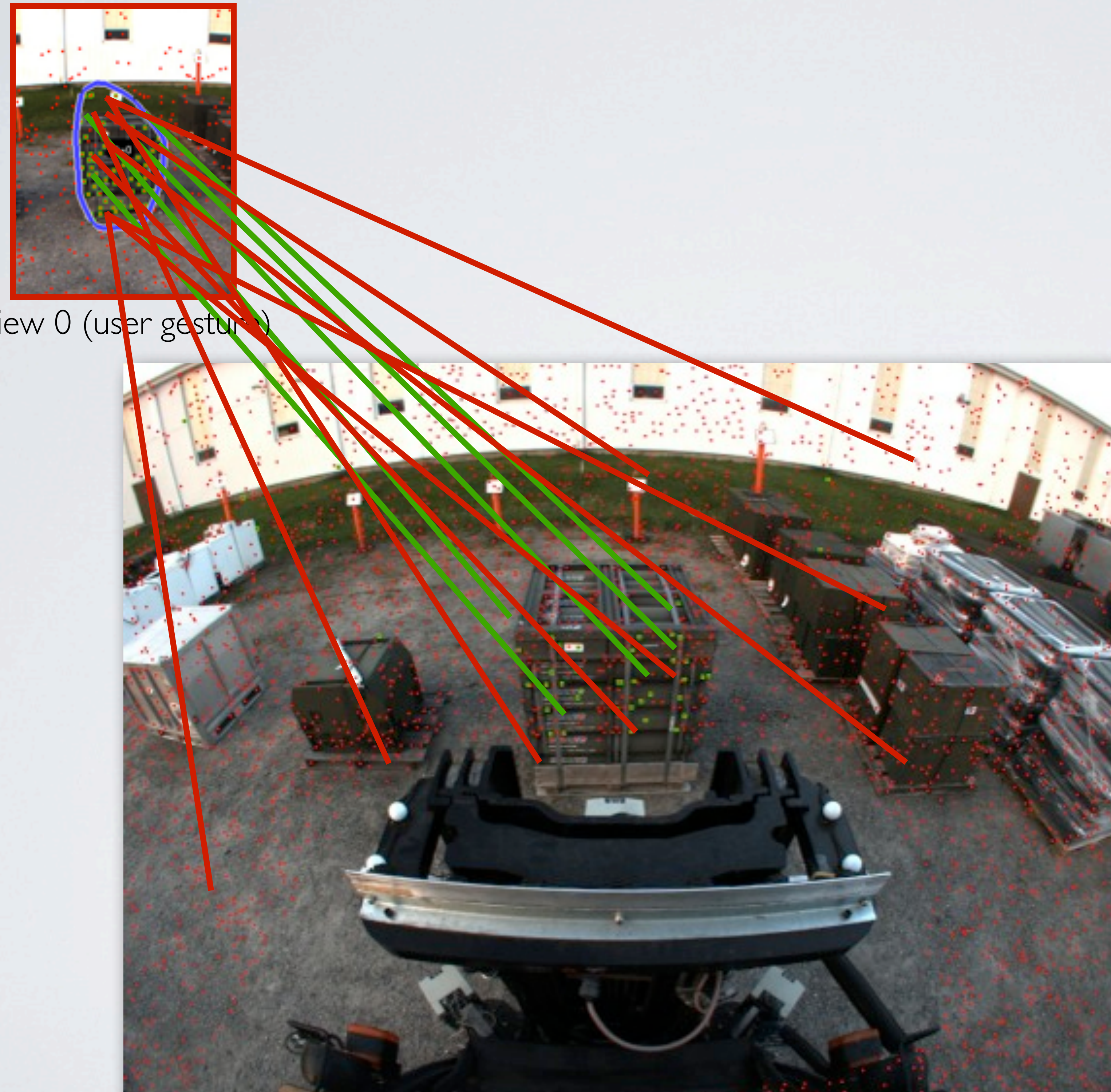
Single-View Matching


View 0 (user gesture)

SIFT features extracted from new image

Extract features and match against all views

Single-View Matching



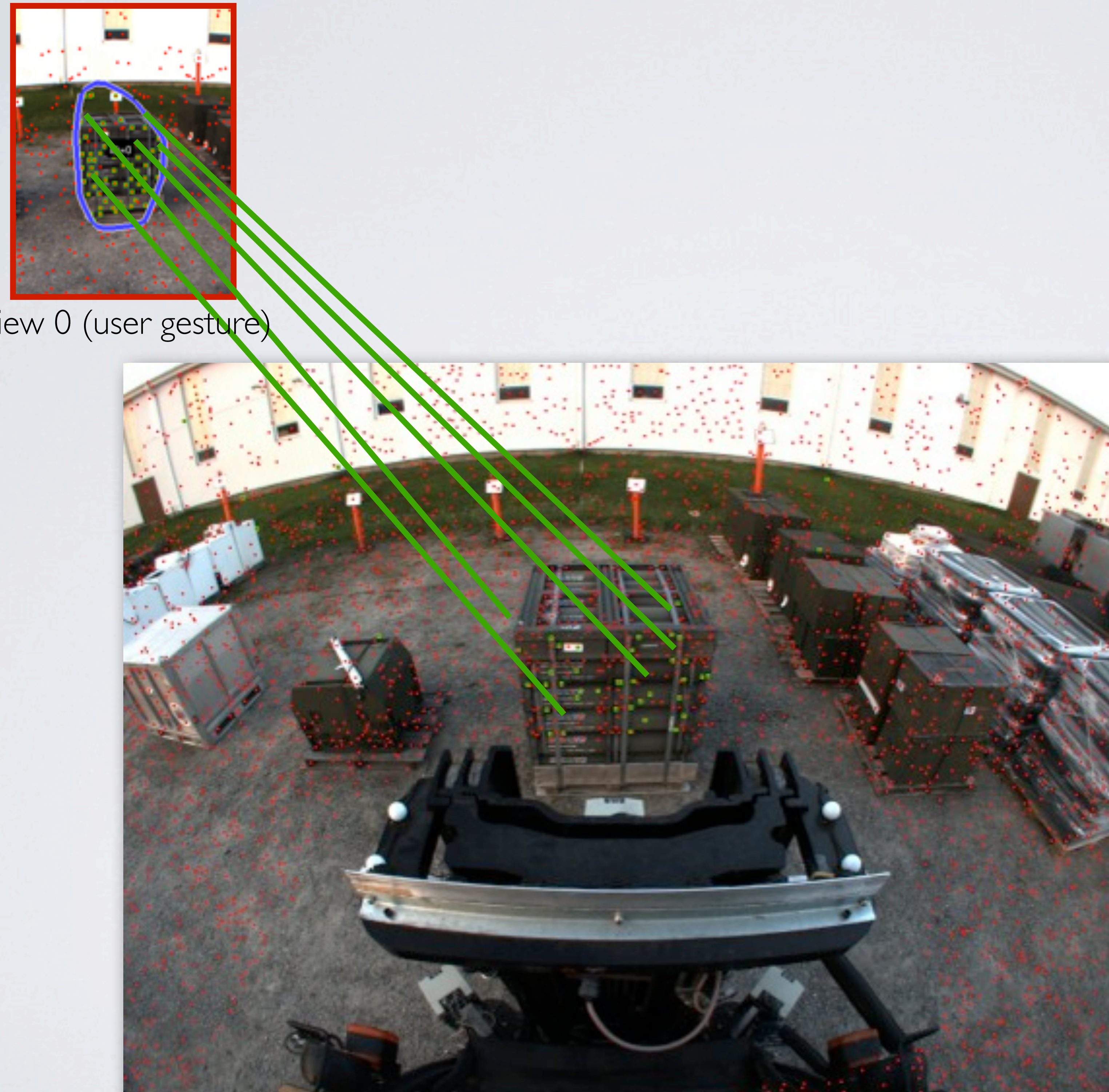
View 0 (user gesture)

SIFT features extracted from new image

RANSAC

1. Sample a subset of pairs
2. Estimate corresponding image-to-image transformation (plane-projective homography)
3. Check consistency with other pairs
4. Repeat if inconsistent

Single-View Matching



View 0 (user gesture)

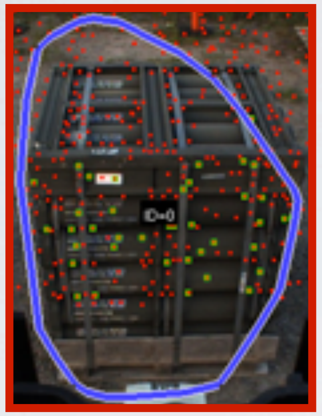
SIFT features extracted from new image

RANSAC

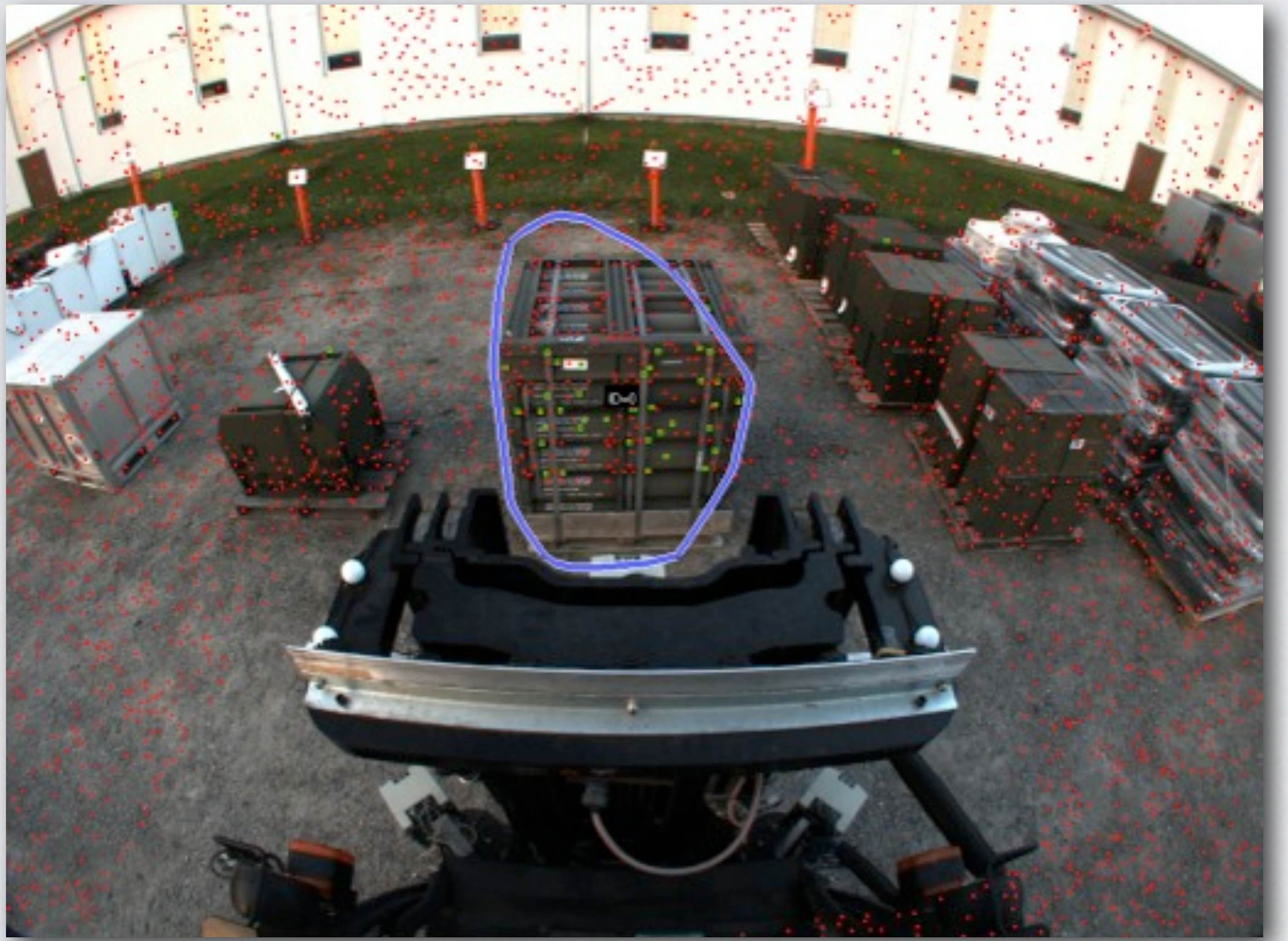
1. Sample a subset of pairs
2. Estimate corresponding image-to-image transformation (plane-projective homography)
3. Check consistency with other pairs
4. Repeat if inconsistent

Model Augmentation

View 0 (user gesture)



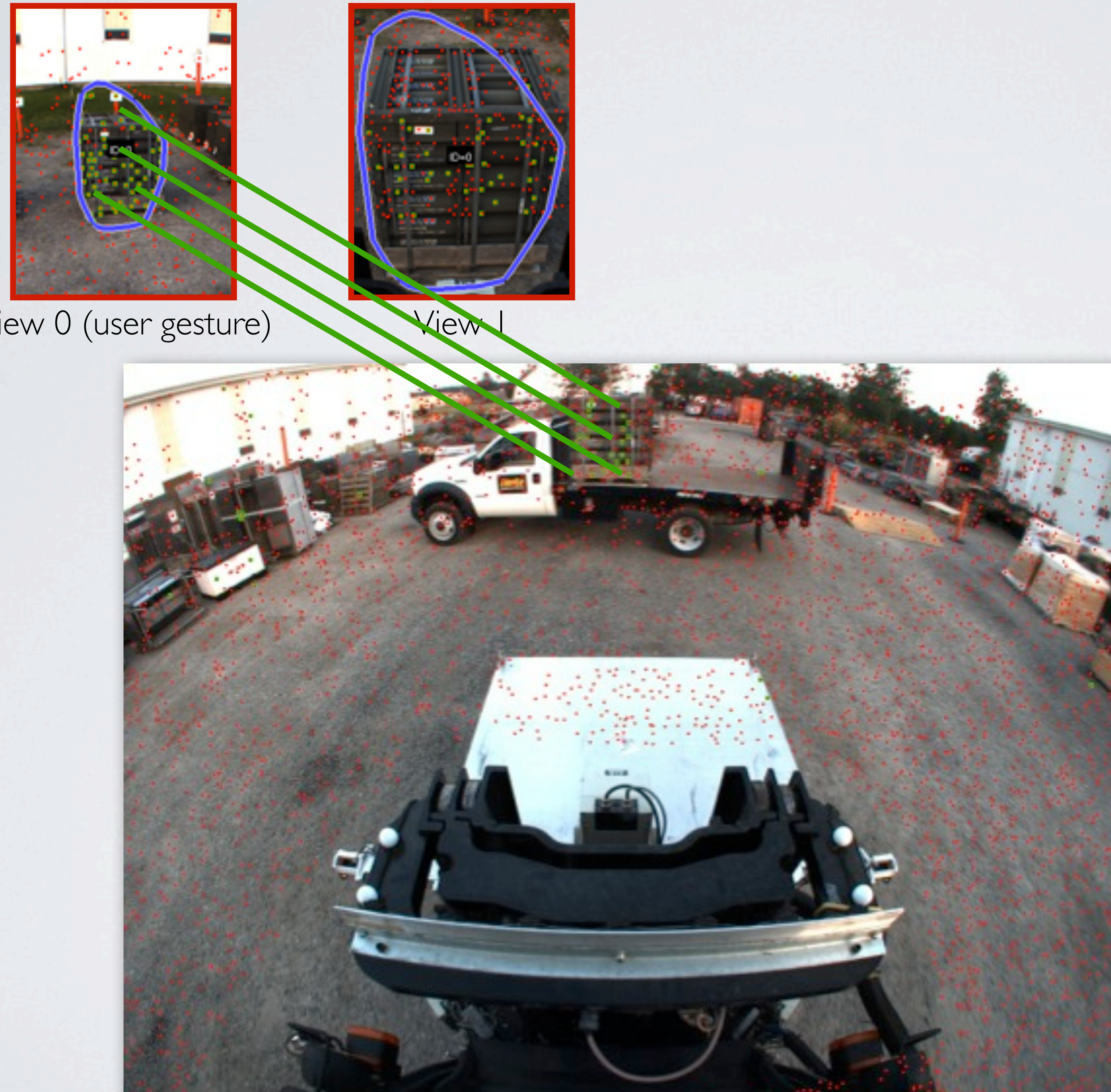
View 1



SIFT features extracted from new image

Generate segmentation and add new view

Model Augmentation



View 0 (user gesture)

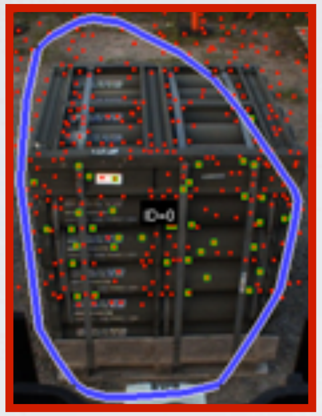
View 1

SIFT features extracted from new image

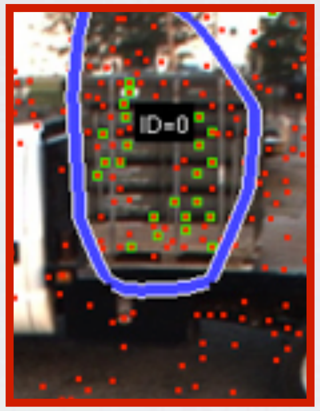
Repeat as object appearance changes

Model Augmentation

View 0 (user gesture)



View 1



View 2

SIFT features extracted from new image

Repeat as object appearance changes

One-shot Appearance Learning

Models opportunistically capture rich appearance variations

View 0 (user gesture)

One-shot Appearance Learning

Models opportunistically capture rich appearance variations

Model 1

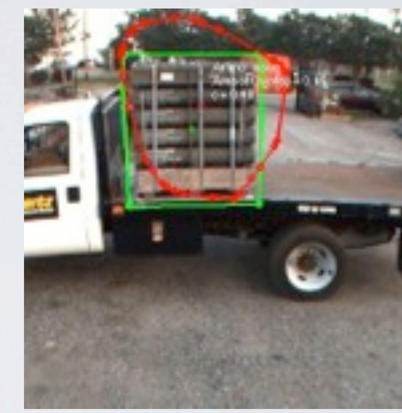

View 0 (user gesture)


View 1


View 2


View 3


View 4

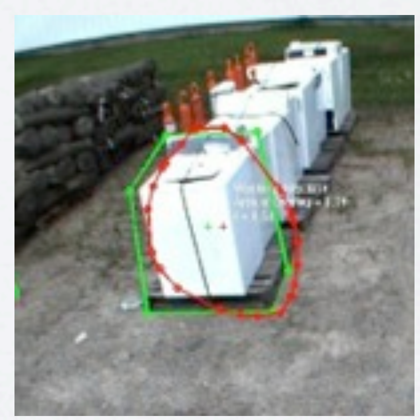

View 5

Model 2


View 0 (user gesture)


View 1


View 2


View 3

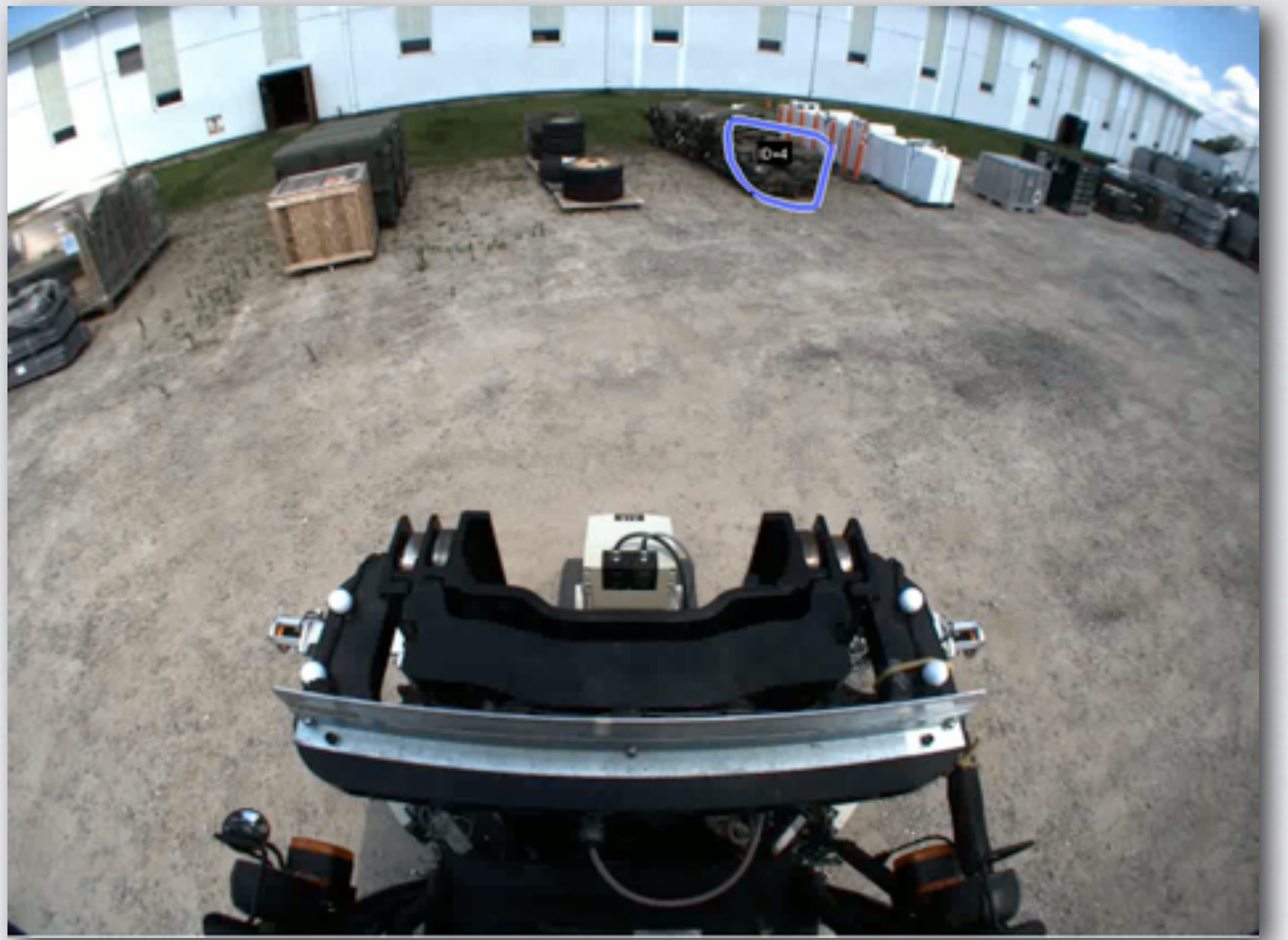

View 4


View 5

Visual Memory Results

- Active, outdoor military warehouse
- Tour and reacquisition separated by hours/days
- Training and detection with different cameras
- Varying conditions

Visual Memory Results



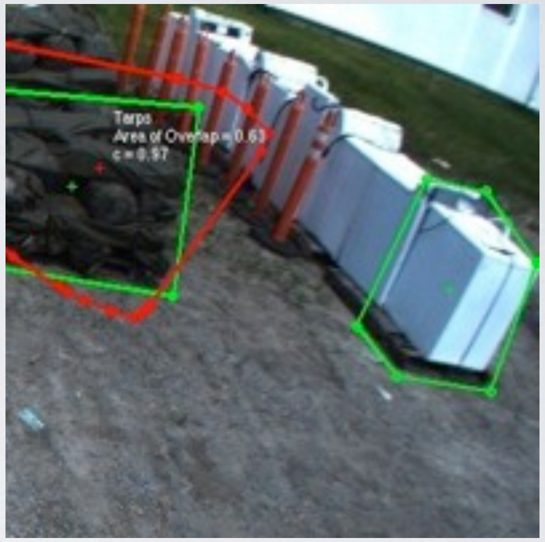
$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Scenario	Train	Test	Delta T	Precision	Recall
1	Afternoon	Afternoon	5 min	94%	54%
2	Evening	Evening	5 min	100%	95%
3	Morning	Evening	14 hours	100%	93%
4	Morning	Evening	10 hours	100%	94%
5	Noon	Evening	7 hours	100%	94%

Visual Memory Results

- Severe saturation
- Motion blur
- Unobserved viewpoints

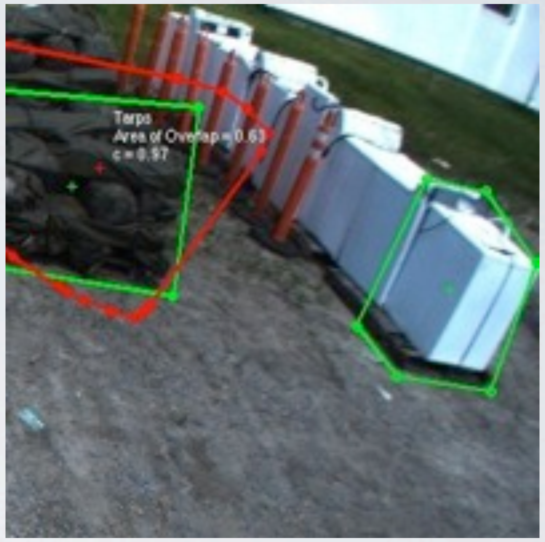


Training example

Saturation

Visual Memory Results

- Severe saturation
- Motion blur
- Unobserved viewpoints



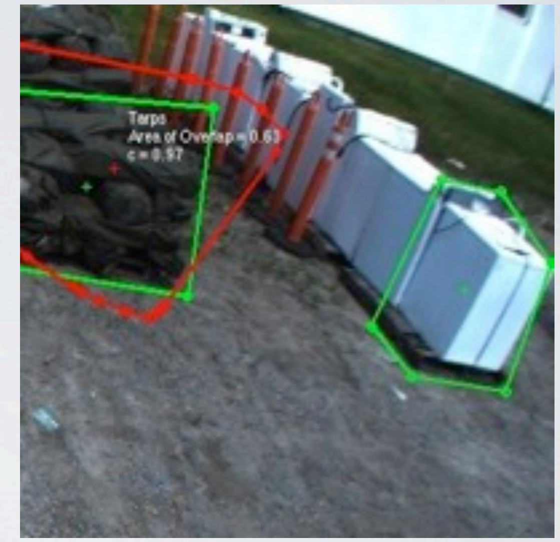
Training example

Saturation

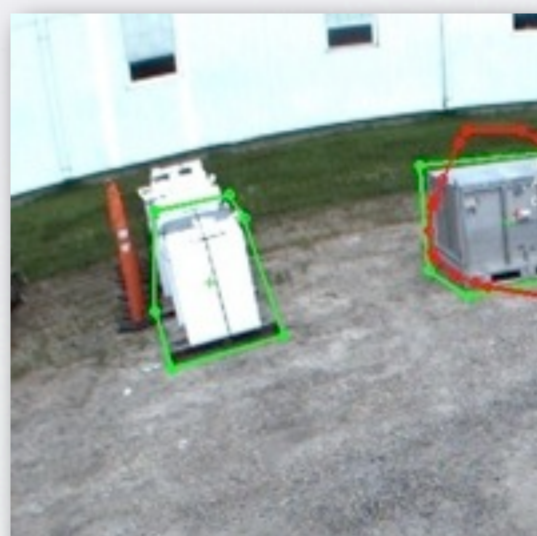
New viewpoint

Visual Memory Results

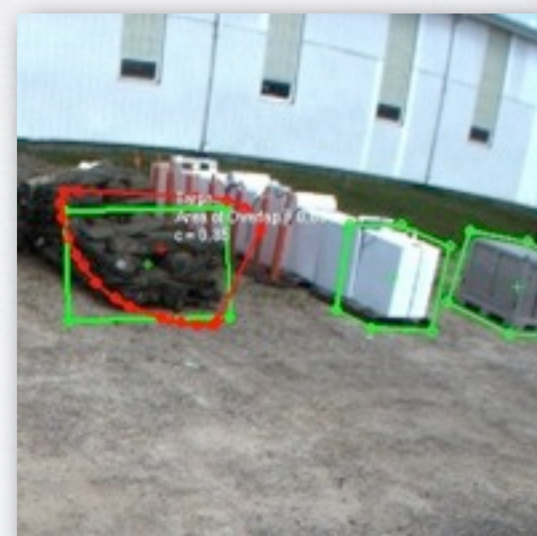
- Severe saturation
- Motion blur
- Unobserved viewpoints



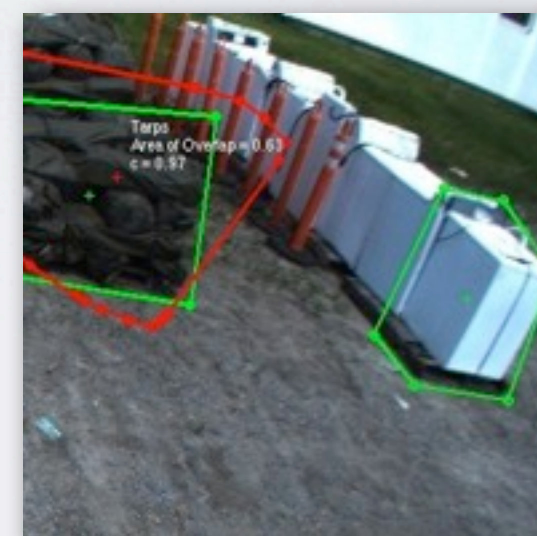
Training example



Saturation



New viewpoint



New viewpoint

Visual Memory Results

Visual Memory Results

Symbol Grounding Problem

Place the lifted tyre pallet, next to another tyre pallet on the trolley.

Lift the tire pallet in the air, then proceed to deposit it to the right of the tire pallet already on the table right in front of you.

Place the pallet of tires on the left side of the trailer.

Please lift the set of six tires up and set them on the trailer, to the right of the set of tires already on it.

lift the tire pallet you are carrying and set it on the truck in front of you

Place the pallet of tires that is on the forklift next to the pallet of tires that is already loaded on the trailer.

Lift tire pallet. Move to unoccupied location on truck. Lower tire pallet. Reverse to starting location. Lower forks. End.

Symbol Grounding Problem

Linguistic elements → Correct referents in the robot's world model

"Grounding"

Place the lifted tyre pallet, next to another tyre pallet on the trolley.

Lift the tire pallet in the air, then proceed to deposit it to the right of the tire pallet already on the table right in front of you.

Place the pallet of tires on the left side of the trailer.

Please lift the set of six tires up and set them on the trailer, to the right of the set of tires already on it.

lift the tire pallet you are carrying and set it on the truck in front of you

Place the pallet of tires that is on the forklift next to the pallet of tires that is already loaded on the trailer.

Lift tire pallet. Move to unoccupied location on truck. Lower tire pallet. Reverse to starting location. Lower forks. End.

Symbol Grounding Problem

"Put the tire pallet on the truck"

- Objects
- Spatial relations
- Actions
- Places

- Object library
- Transformations, relative positions
- Paths, motion primitives, torques
- Positions, orientations

Symbol Grounding Problem

"Put the $\boxed{\text{tire pallet}}$ on the $\boxed{\text{truck}}$"

- Objects
- Spatial relations
- Actions
- Places

- Object library
- Transformations, relative positions
- Paths, motion primitives, torques
- Positions, orientations

Symbol Grounding Problem

"Put the tire pallet on the truck"

- Objects
- Spatial relations
- Actions
- Places

- Object library
- Transformations, relative positions
- Paths, motion primitives, torques
- Positions, orientations

Symbol Grounding Problem

"Put the tire pallet $\boxed{\text{on}}$ the truck"

- Objects
- Spatial relations
- Actions
- Places

- Object library
- Transformations, relative positions
- Paths, motion primitives, torques
- Positions, orientations

Symbol Grounding Problem

"Put the tire pallet on the truck"

- Objects
- Spatial relations
- Actions
- Places

- Object library
- Transformations, relative positions
- Paths, motion primitives, torques
- Positions, orientations

Symbol Grounding Problem

"Put the tire pallet on the truck"

- Objects
- Spatial relations
- Actions
- Places

- Object library
- Transformations, relative positions
- Paths, motion primitives, torques
- Positions, orientations

Grounding Natural Language Speech

(collaboration with S. Tellex, T. Kollar, S. Teller, & N. Roy)

$$\boxed{\arg\max_{\text{groundings}} p\left(\text{groundings} \mid \text{language}\right)}$$

objects, actions, relations, places

"Drive to the tire pallet"

Grounding Natural Language Speech

(collaboration with S. Tellex, T. Kollar, S. Teller, & N. Roy)

"To the tire pallet"

$$\arg \max_{\gamma \in \mathcal{X}} p\left(\gamma | \lambda\right)$$

Grounding Natural Language Speech

(collaboration with S. Tellex, T. Kollar, S. Teller, & N. Roy)

"To the tire pallet"

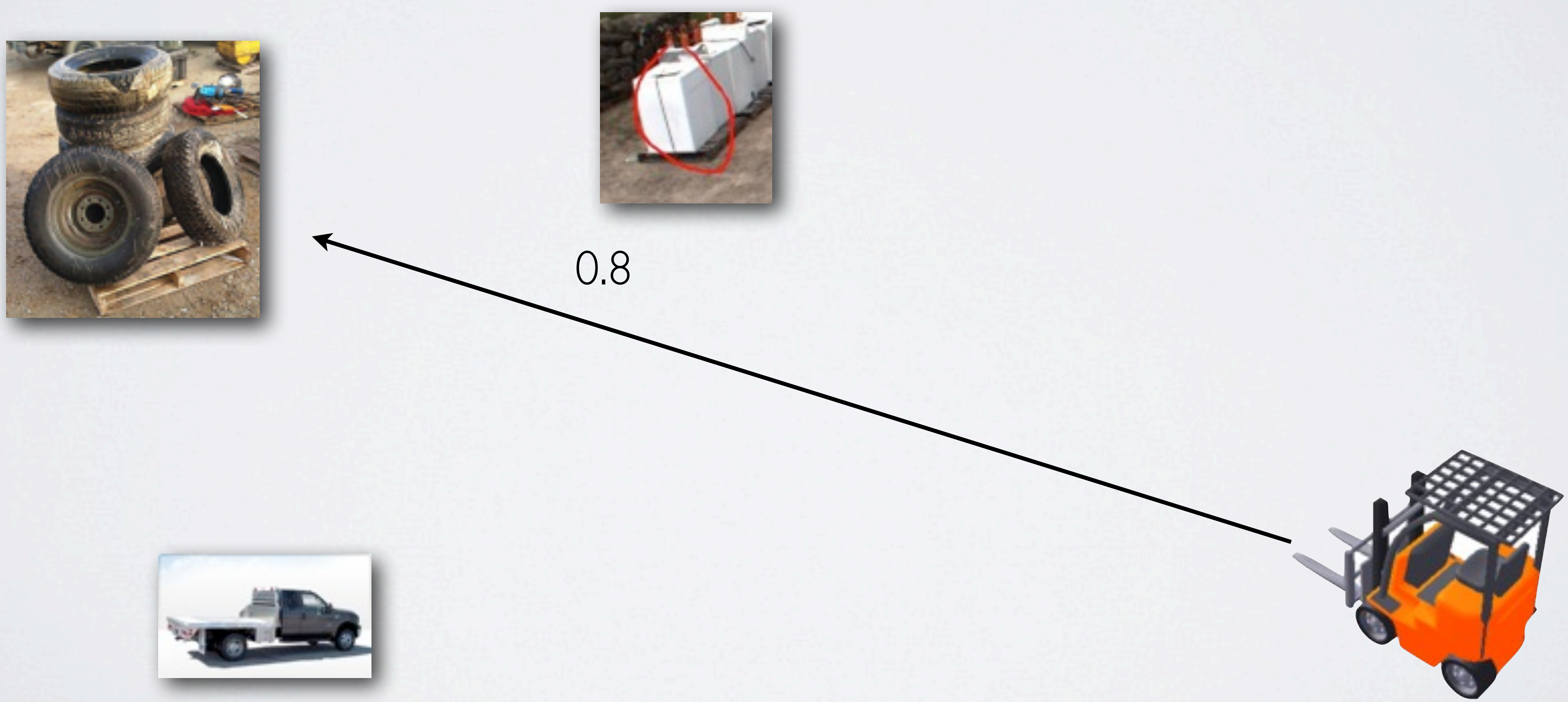
$$\arg\max_{\gamma \in \mathcal{X}} p\left(\gamma | \lambda\right)$$

Grounding Natural Language Speech

(collaboration with S. Tellex, T. Kollar, S. Teller, & N. Roy)

"To the tire pallet"

$$\arg\max_{\gamma \in \mathcal{X}} p\left(\gamma | \lambda\right)$$

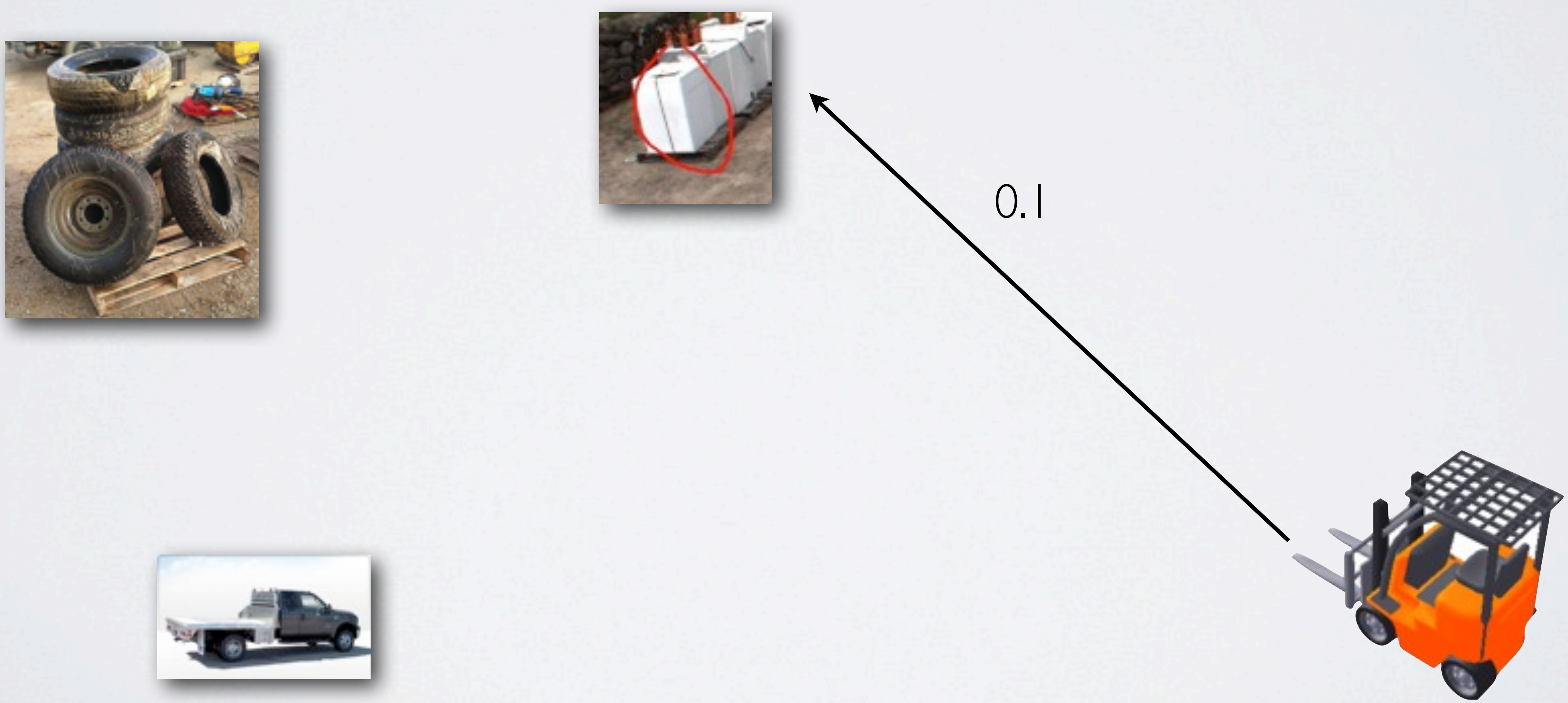


Grounding Natural Language Speech

(collaboration with S. Tellex, T. Kollar, S. Teller, & N. Roy)

"To the tire pallet"

$$\arg\max_{\gamma \in \mathcal{X}} p\left(\gamma | \lambda\right)$$

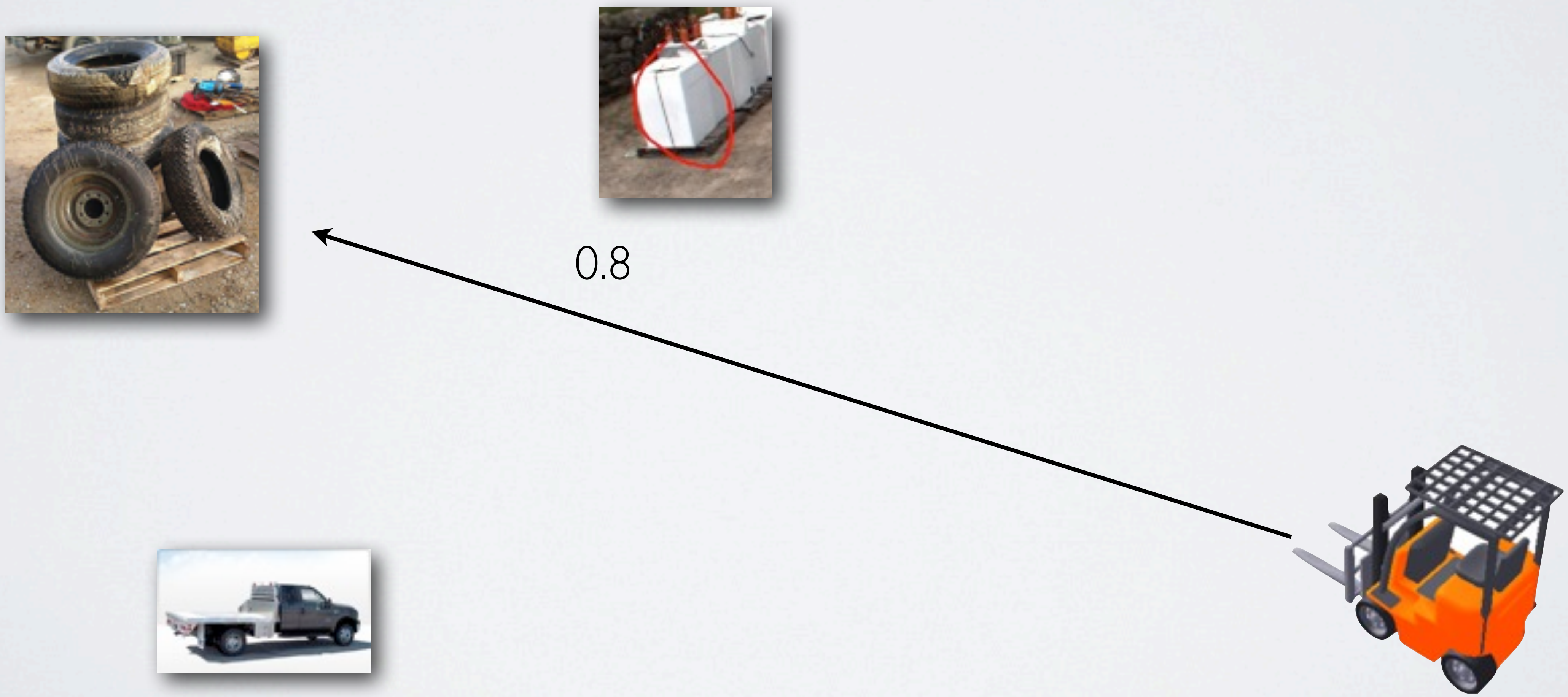


Grounding Natural Language Speech

(collaboration with S. Tellex, T. Kollar, S. Teller, & N. Roy)

"To the tire pallet"

$$\arg\max_{\gamma \in \mathcal{X}} p\left(\gamma | \lambda\right)$$



Grounding Natural Language Speech

$$\boxed{\arg\max_{\text{groundings}} p\,(\textbf{groundings}|\textbf{language})}$$

objects, actions, relations, places

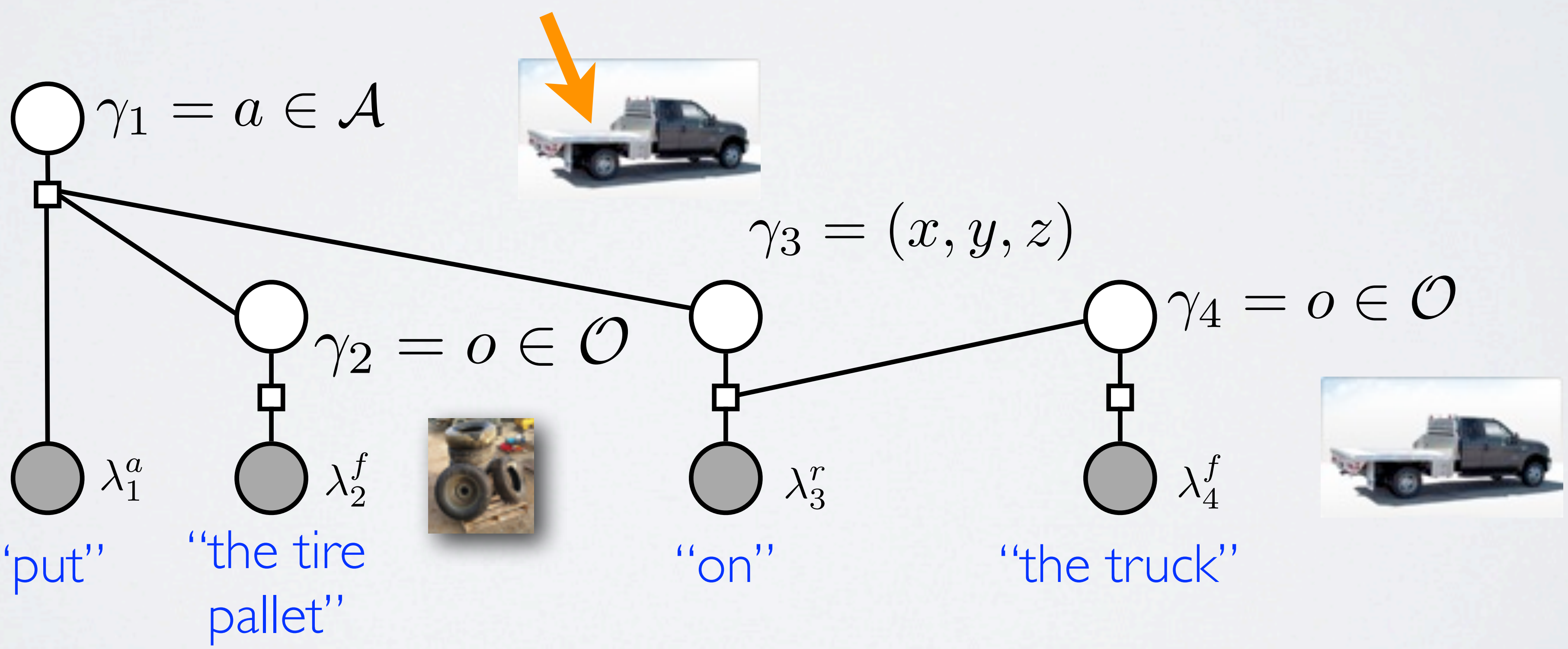
"Put the tire pallet on the truck"

[AAAI 2011; AI Magazine 2011]

Grounding Natural Language Speech

"Put the tire pallet on the truck"

$$\arg\max_{\Gamma} \; (\gamma_1, \gamma_2, \gamma_3, \gamma_4 | \lambda)$$

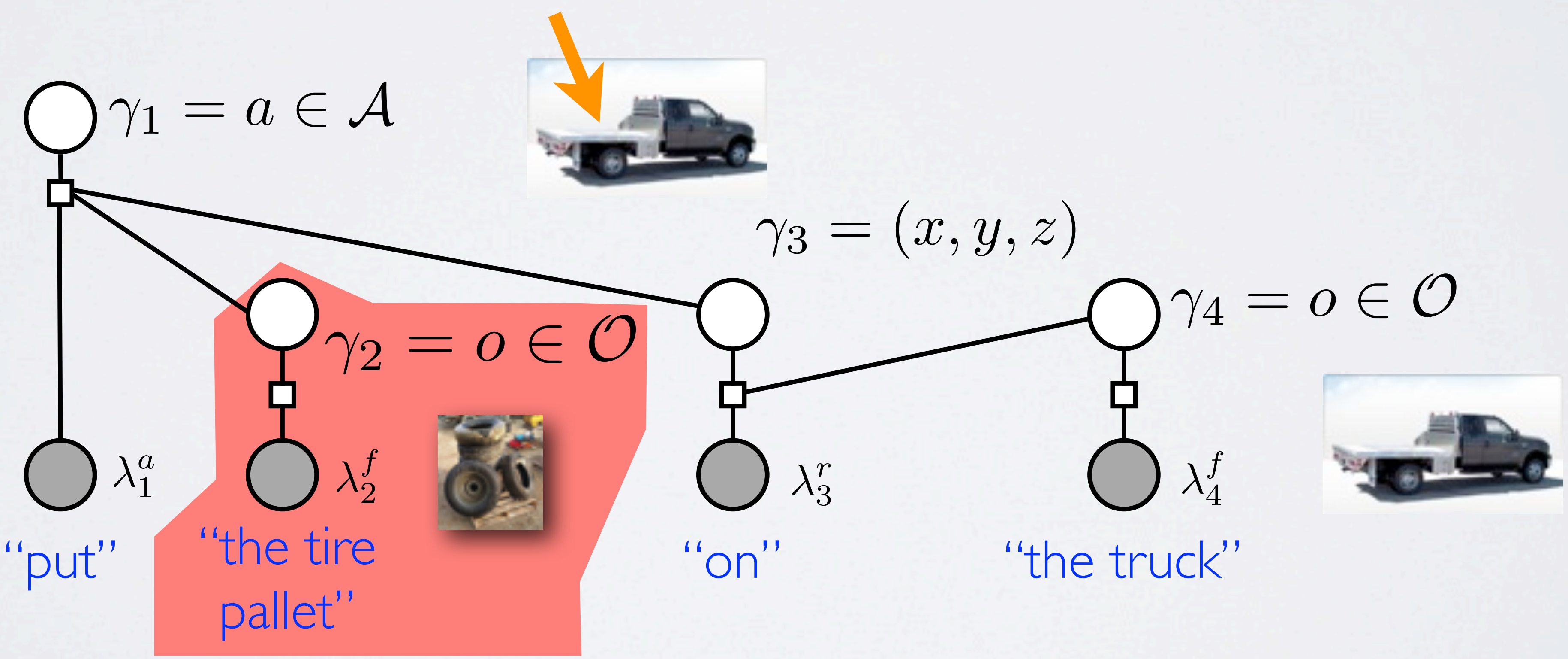


[AAAI 2011; AI Magazine 2011]

Grounding Natural Language Speech

"Put the tire pallet on the truck"

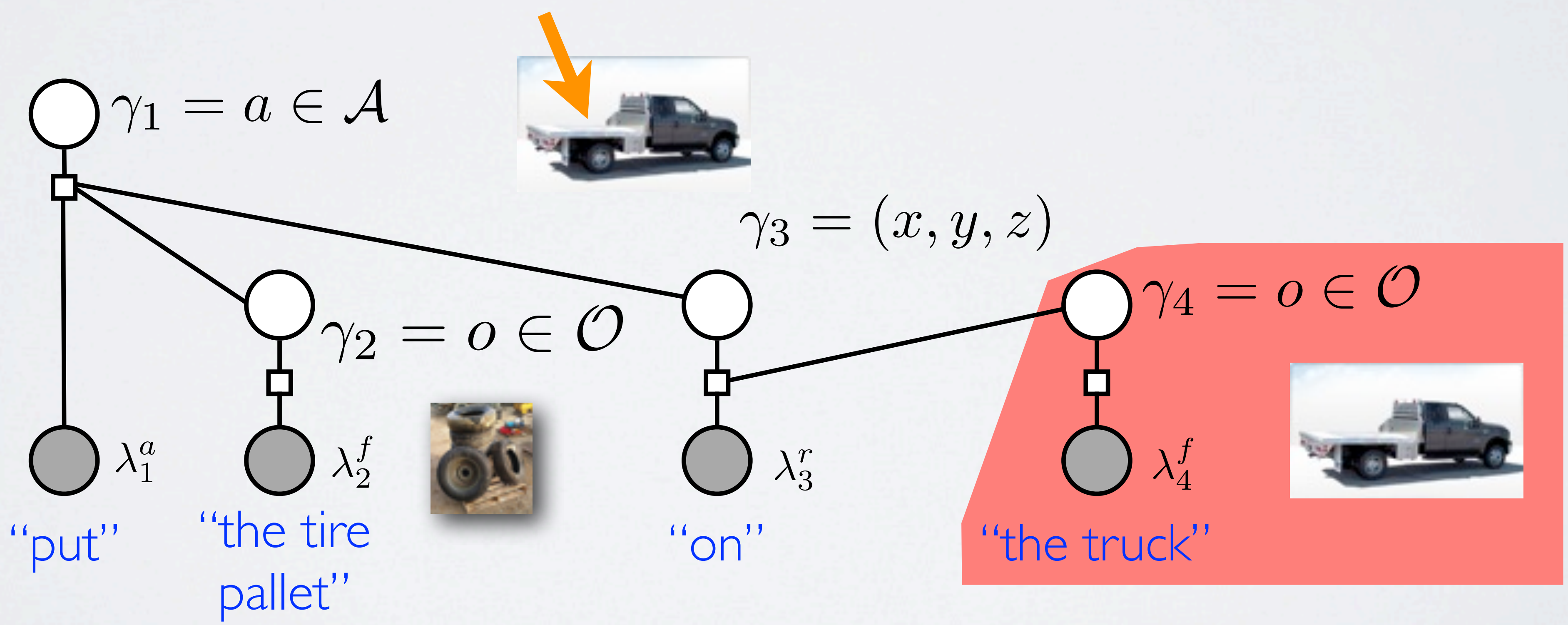
$$\arg\max_{\Gamma} \; (\gamma_1, \gamma_2, \gamma_3, \gamma_4 | \lambda)$$



Grounding Natural Language Speech

"Put the tire pallet on the truck"

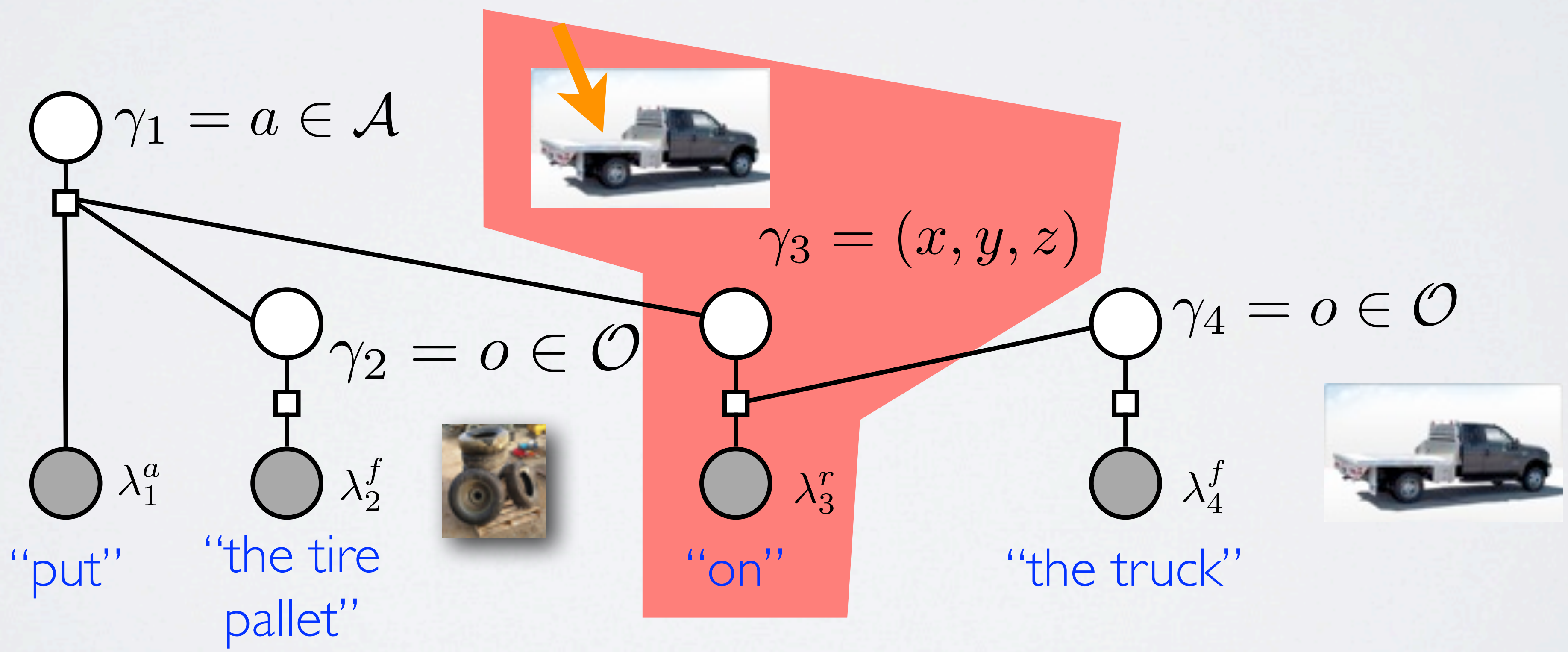
$$\arg\max_{\Gamma} \; (\gamma_1, \gamma_2, \gamma_3, \gamma_4 | \lambda)$$



Grounding Natural Language Speech

"Put the tire pallet on the truck"

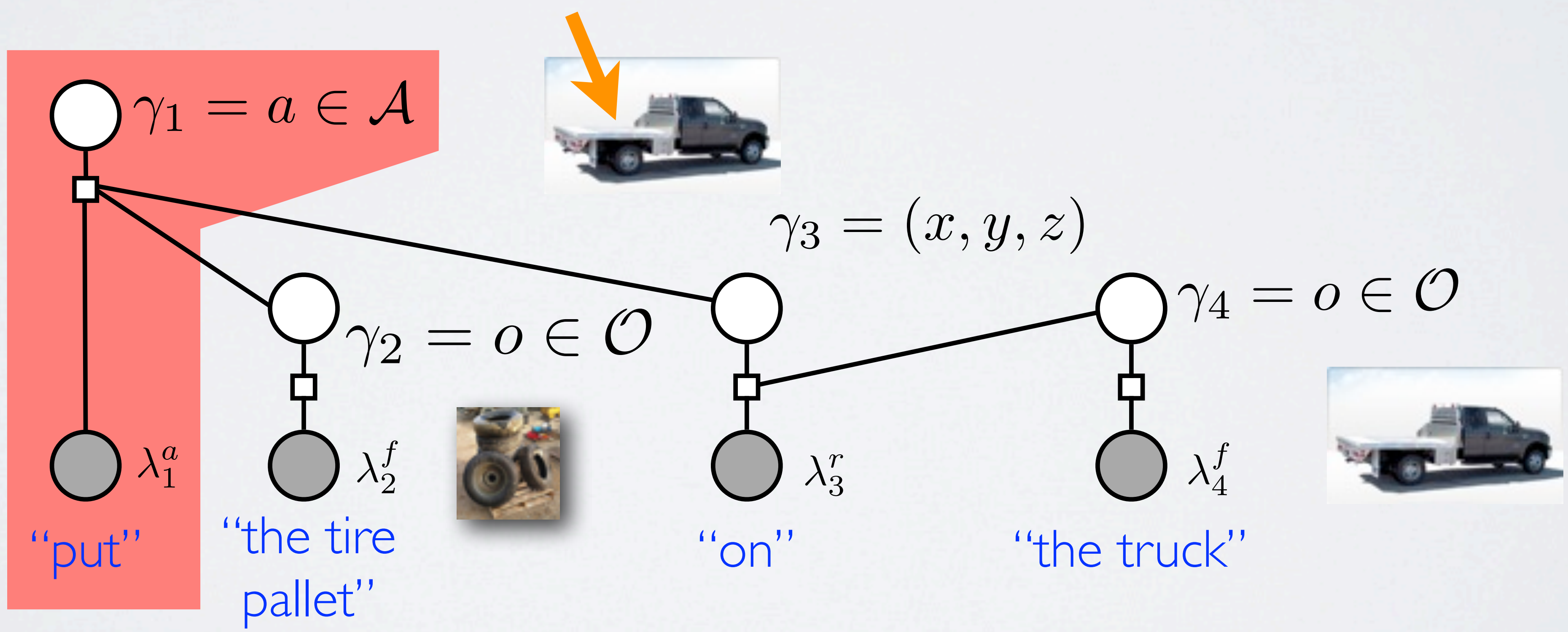
$$\arg\max_{\Gamma} \; (\gamma_1, \gamma_2, \gamma_3, \gamma_4 | \lambda)$$



Grounding Natural Language Speech

"Put the tire pallet on the truck"

$$\arg\max_{\Gamma} \; (\gamma_1, \gamma_2, \gamma_3, \gamma_4 | \lambda)$$



I. Importance of Situational Awareness

II. Persistent Object Awareness with Vision

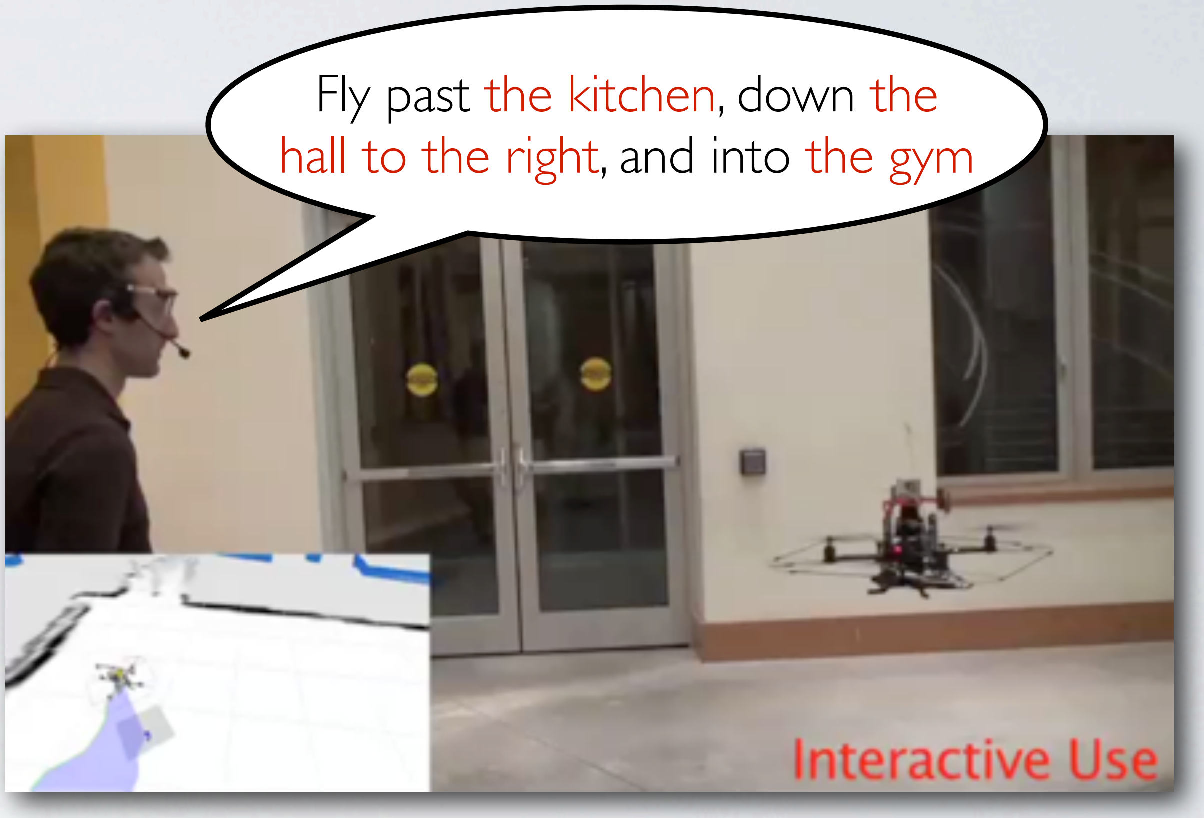
III. Semantic Map Learning from Natural Language Descriptions

IV. Future Directions

V. Conclusions

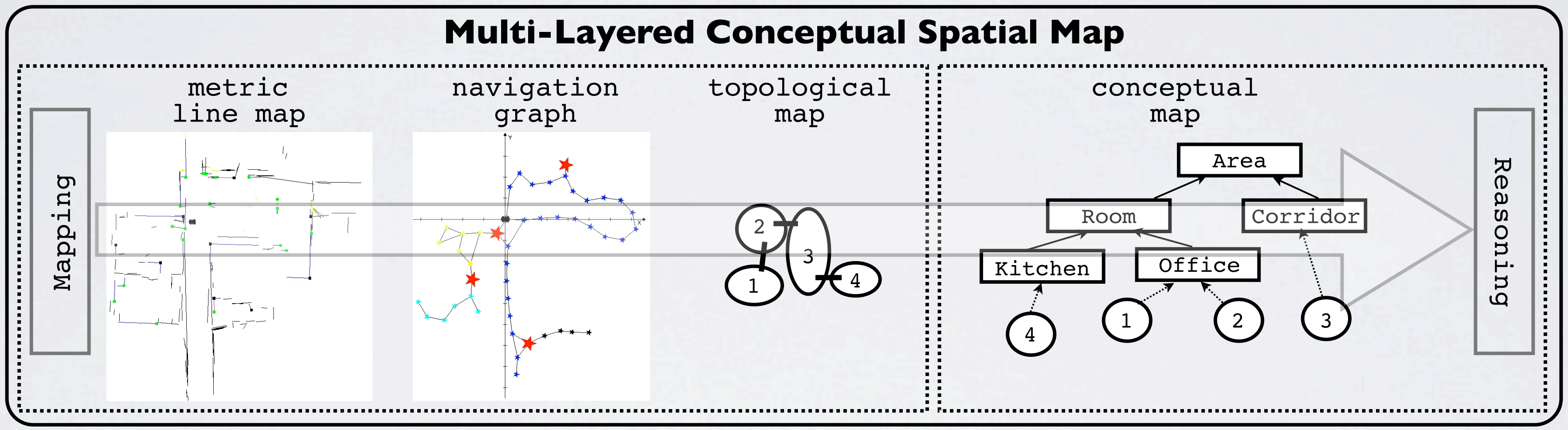
Beyond Objects to Spaces

- Going beyond metric maps
- Human-centric representations of space
  - Spatial relations
  - Semantic attributes (names, use, etc.)
  - Connectivity



State-of-the-Art in Semantic Mapping

- Spatial Semantic Hierarchy (Kuipers 2000)
- Augment SLAM metric/topological SLAM maps with semantic layers



Courtesy: Zender et al. 2008

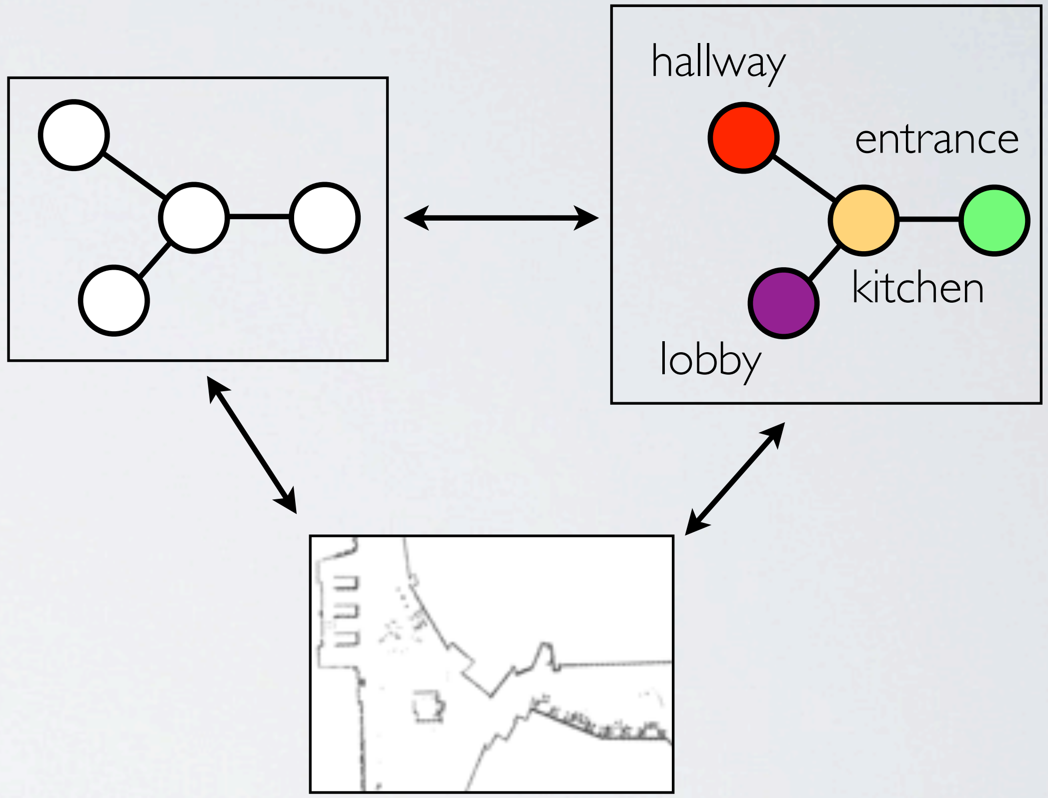
- Infer semantic properties from multiple modalities:
  - Object recognition (Zender et al. 2008; Pronobis et al. 2020)
  - Spoken descriptions and other supervised labels (Diosi et al. 2005; Zender et al. 2008; Pronobis et al. 2020)
  - Place classification (Zender et al. 2008; Pronobis et al. 2020)

Building Semantic Maps with Natural Language

- Learn knowledge representation from narrated tour
- Challenges:
  - People convey high-level concepts but robot perception is low-level
  - Spoken descriptions are ambiguous

Building Semantic Maps with Natural Language

- Solution:
  - Joint metric, topologic, & semantic model supports information fusion
  - Efficient inference strategy
  - Enable layers to influence one another



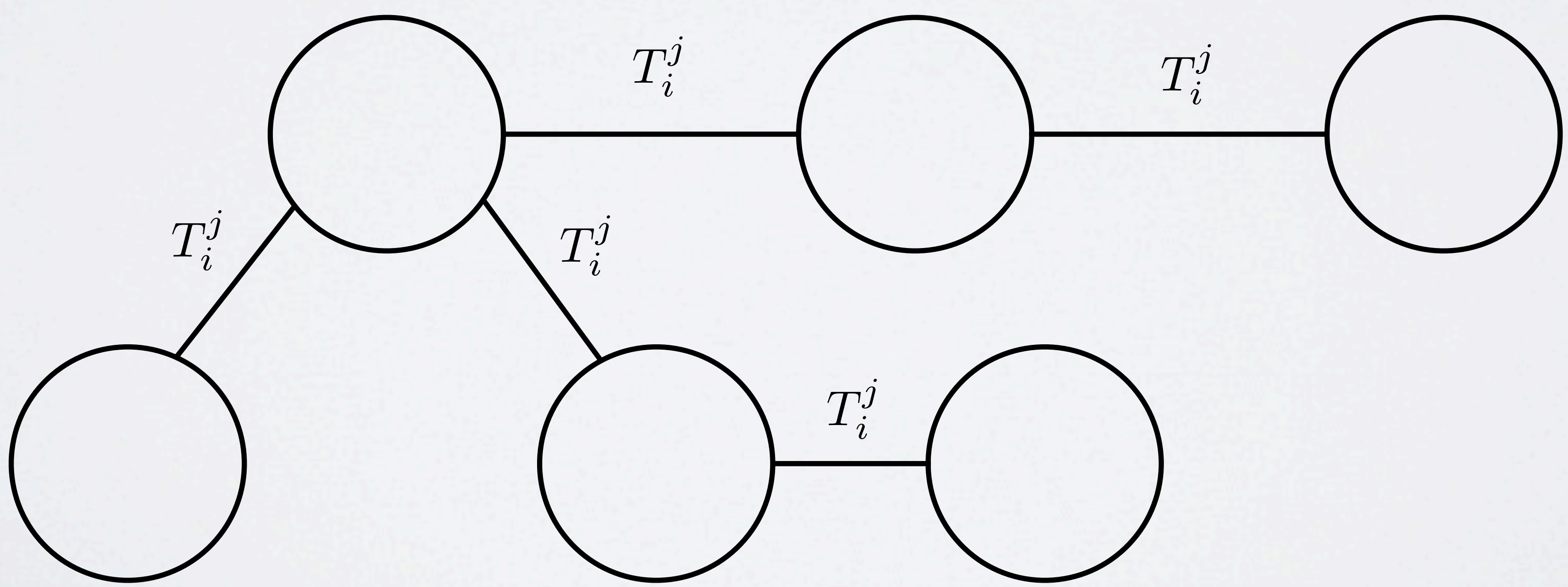
Model: Posterior over Semantic Graphs

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t)$$

Model: Posterior over Semantic Graphs

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t)$$

Topology $G_t = (V_t, E_t)$

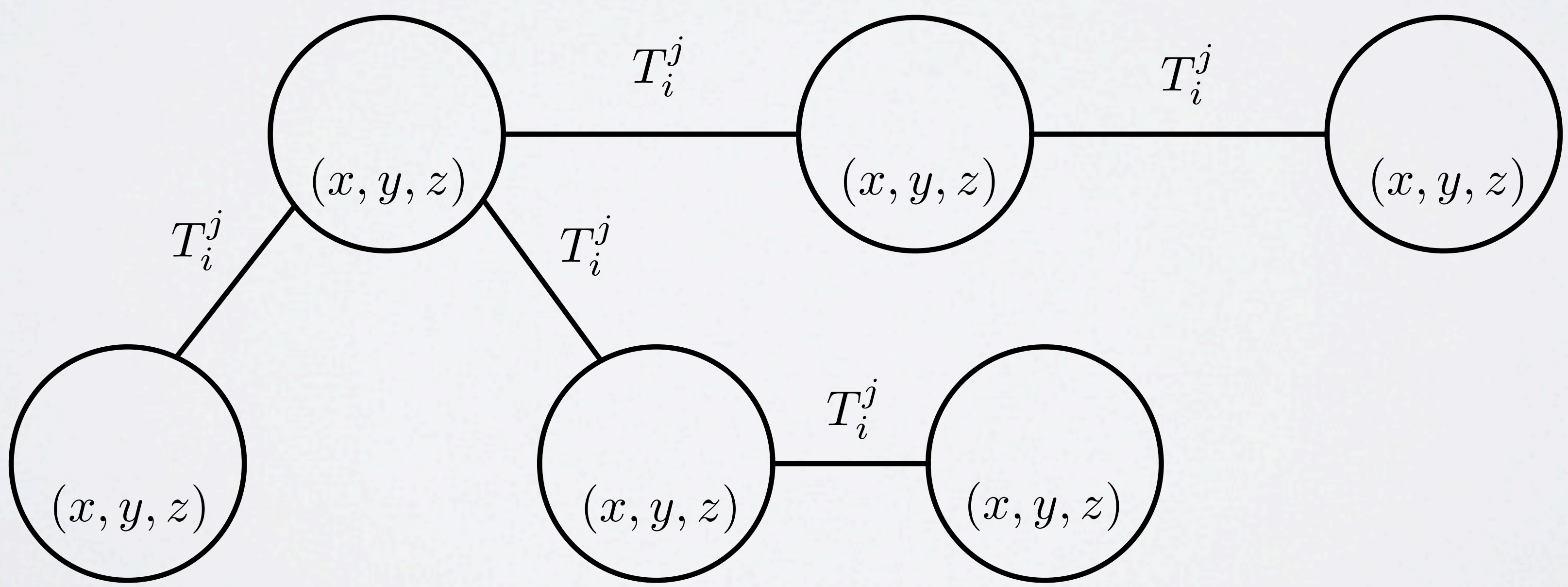


Model: Posterior over Semantic Graphs

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t)$$

Topology $G_t = (V_t, E_t)$

Vertex poses



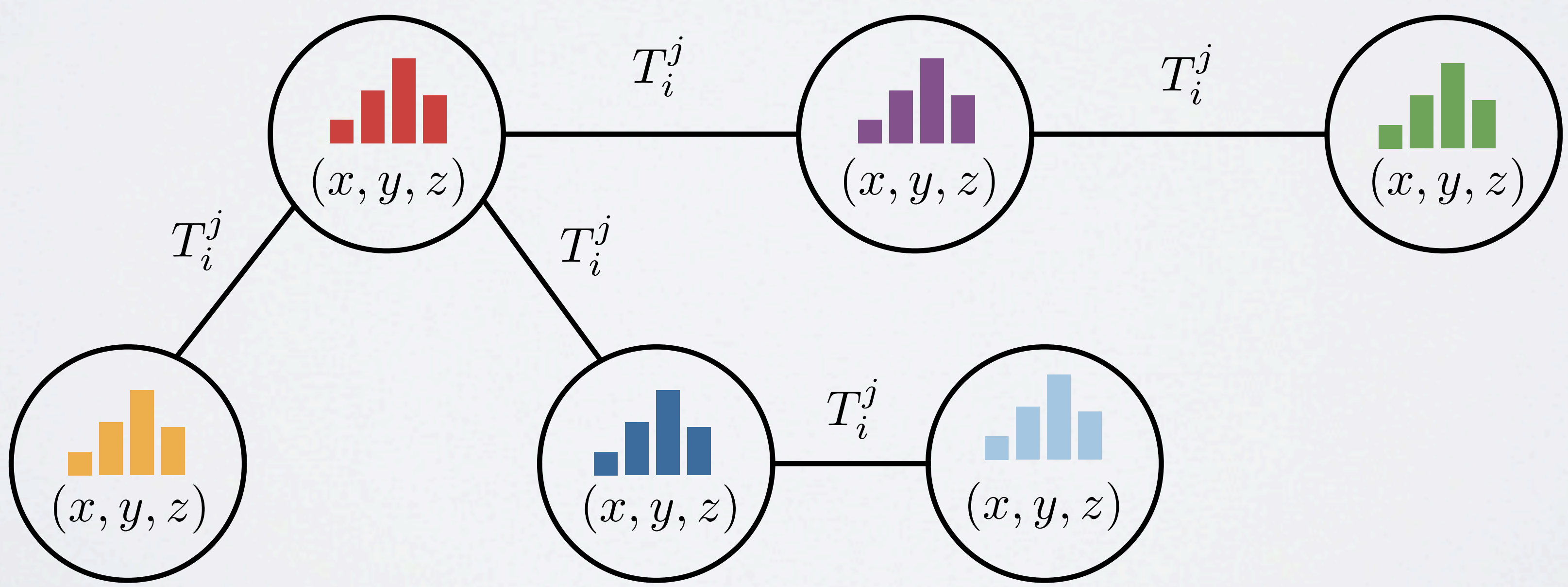
Model: Posterior over Semantic Graphs

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t)$$

Topology $G_t = (V_t, E_t)$

Vertex poses

Semantic labels



Model: Posterior over Semantic Graphs

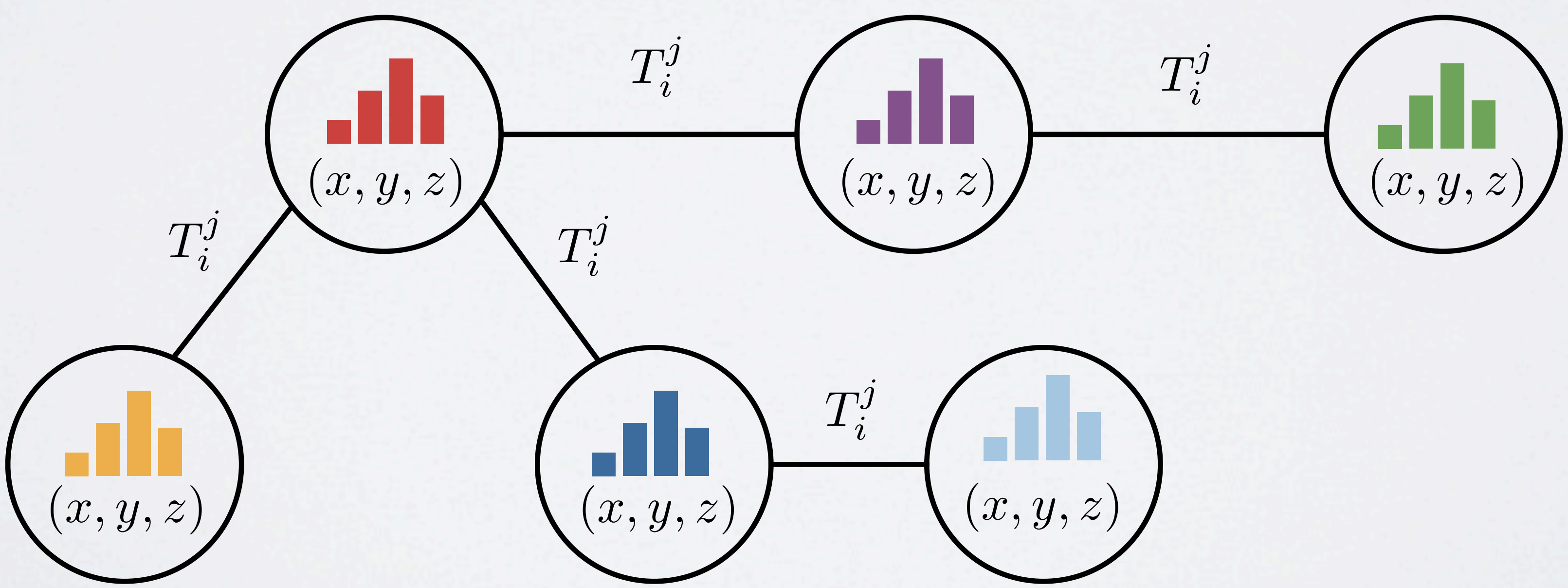
Sensor stream

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t)$$

Topology $G_t = (V_t, E_t)$

Vertex poses

Semantic labels



Model: Posterior over Semantic Graphs

Sensor stream

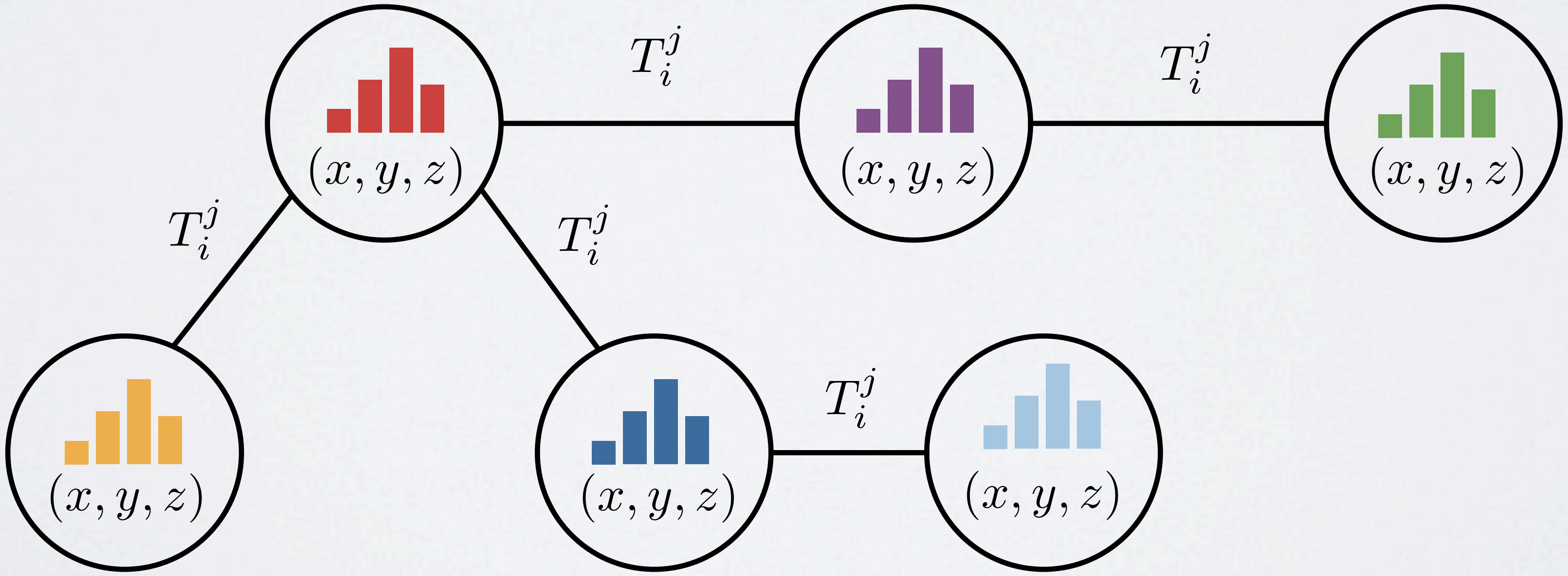
Odometry

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t)$$

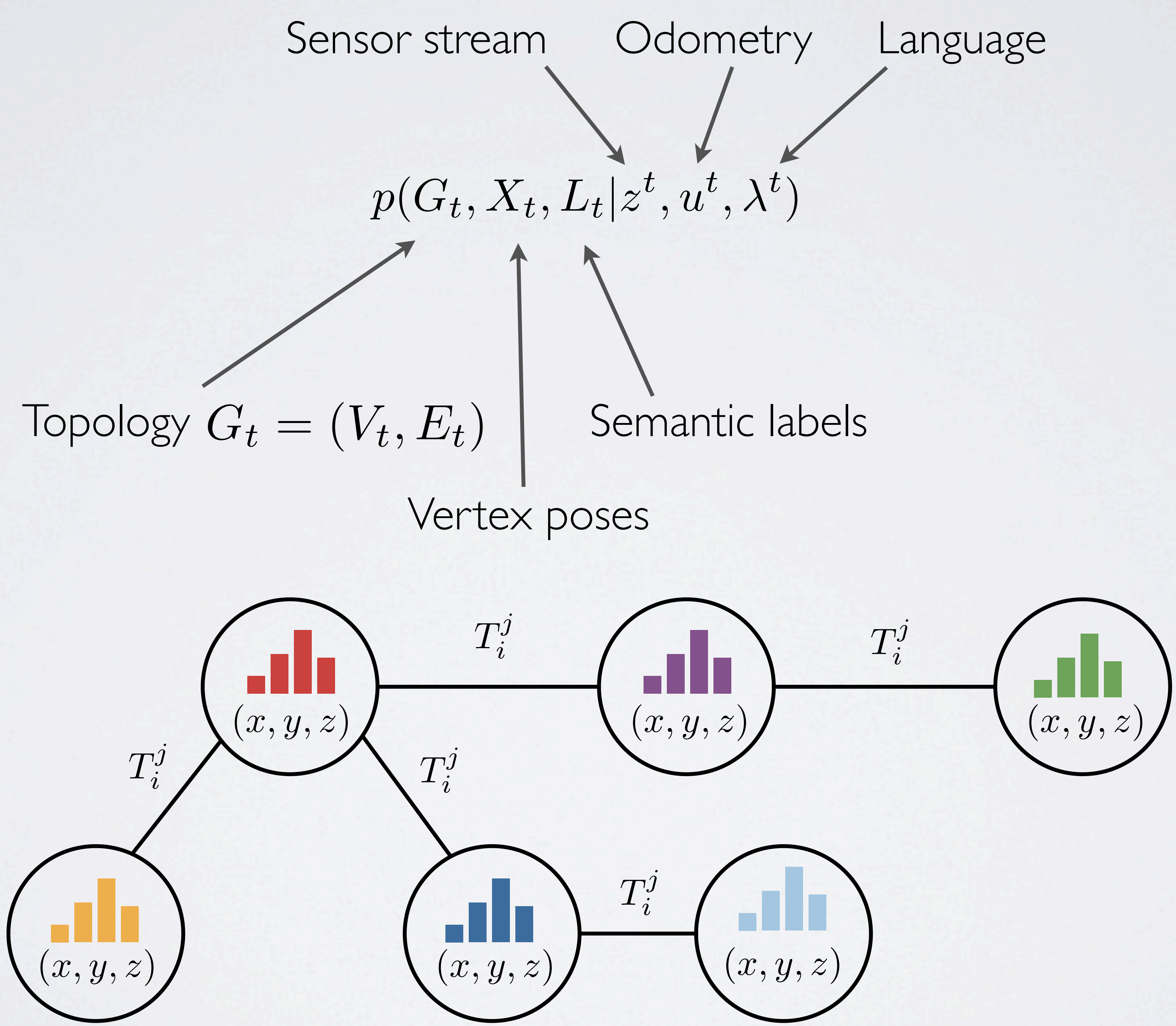
Topology $G_t = (V_t, E_t)$

Semantic labels

Vertex poses



Model: Posterior over Semantic Graphs



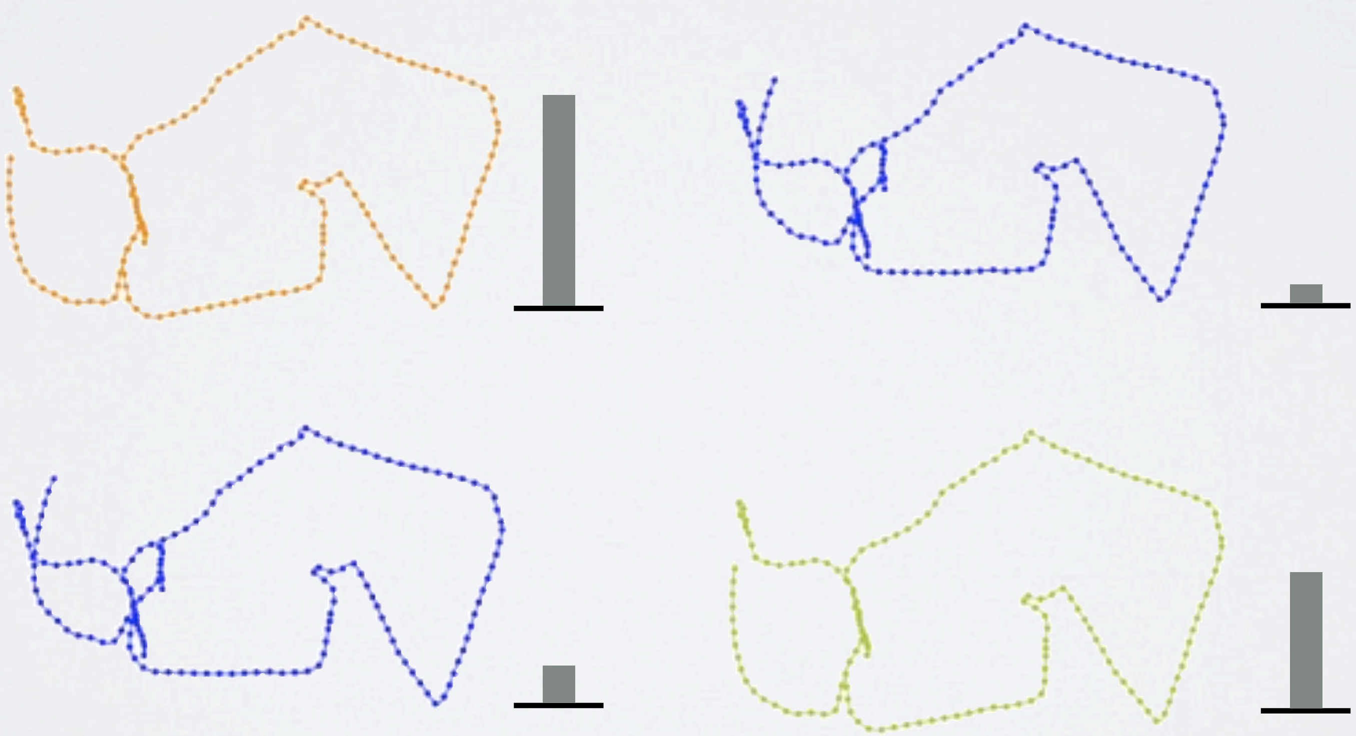
Model: Posterior over Semantic Graphs

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t) = p(L_t | X_t, G_t, z^t, u^t, \lambda^t)\ p(X_t | G_t, z^t, u^t, \lambda^t)\ p(G_t | z^t, u^t, \lambda^t)$$

Model: Posterior over Semantic Graphs

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t) = p(L_t | X_t, G_t, z^t, u^t, \lambda^t)\ p(X_t | G_t, z^t, u^t, \lambda^t)\ \boxed{p(G_t | z^t, u^t, \lambda^t)}$$

Sample-based representation



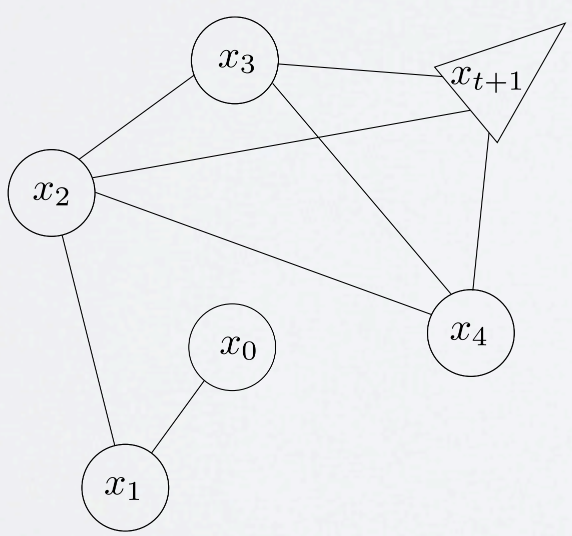
Model: Posterior over Semantic Graphs

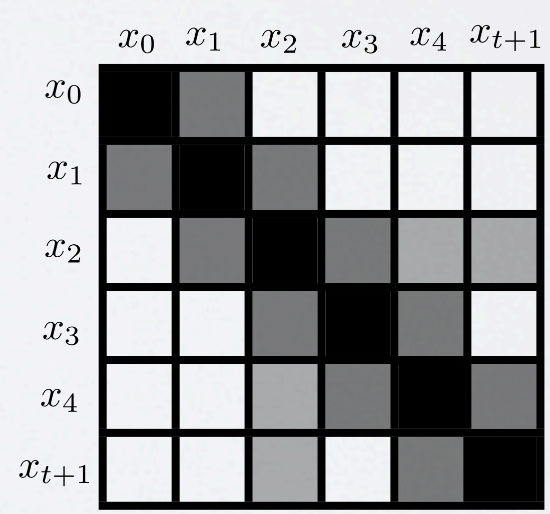
$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t) = p(L_t | X_t, G_t, z^t, u^t, \lambda^t) \; \boxed{p(X_t | G_t, z^t, u^t, \lambda^t)} \; p(G_t | z^t, u^t, \lambda^t)$$

Gaussian (information form)

Sample-based representation

$$p(X_t | G_t, z^t, u^t, \lambda^t) = \mathcal{N}^{-1}(X_t; \Sigma_t^{-1}, \eta_t)$$





Model: Posterior over Semantic Graphs

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t) = \boxed{p(L_t | X_t, G_t, z^t, u^t, \lambda^t)} \; p(X_t | G_t, z^t, u^t, \lambda^t) \; p(G_t | z^t, u^t, \lambda^t)$$

Dirichlet

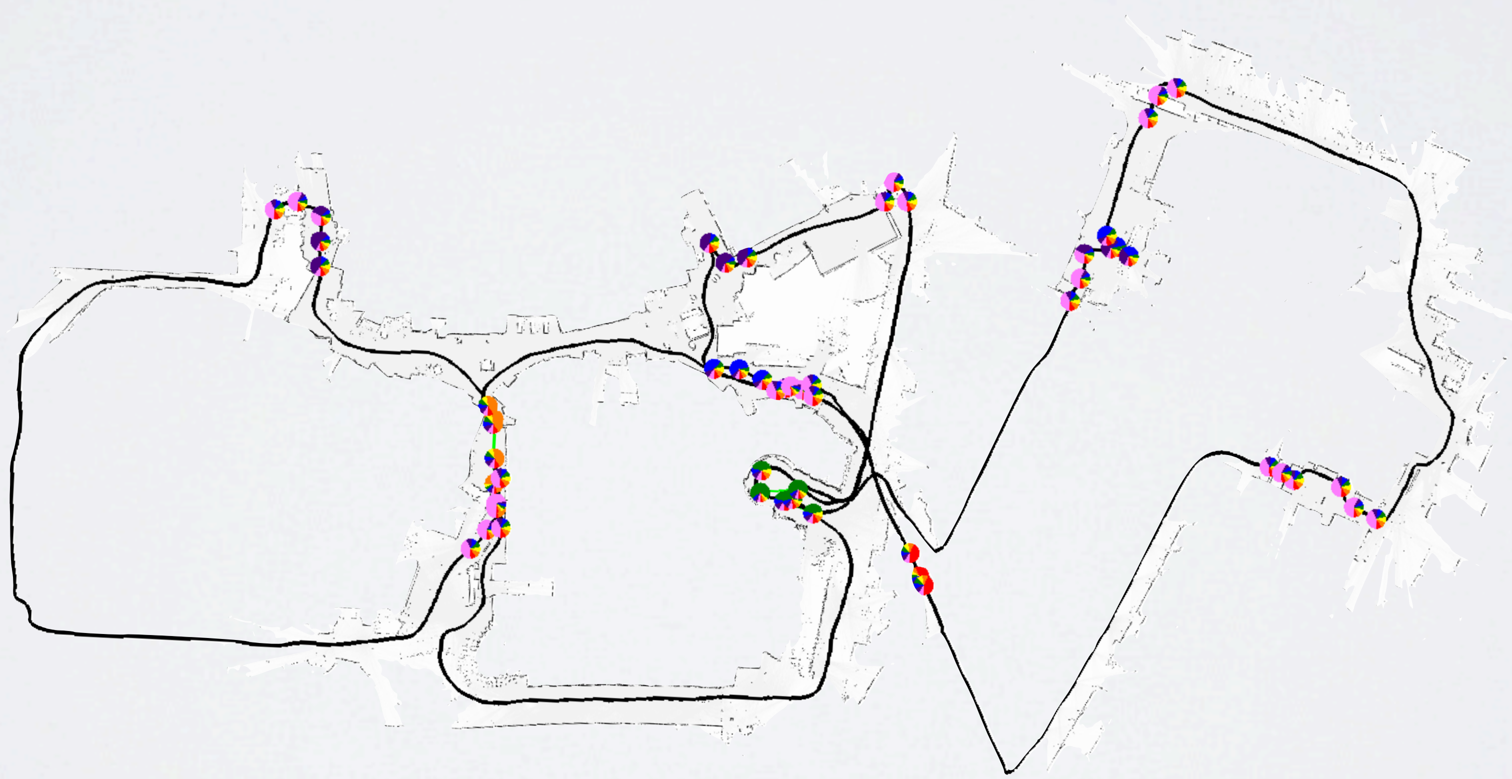
Gaussian
(information form)

Sample-based
representation

Model: Posterior over Semantic Graphs

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t) = p(L_t | X_t, G_t, z^t, u^t, \lambda^t)\ p(X_t | G_t, z^t, u^t, \lambda^t)\ p(G_t | z^t, u^t, \lambda^t)$$

Dirichlet | Gaussian (information form) | Sample-based representation



Matthew Walter

Rao-Blackwellized Particle Filter

Input: $P_{t-1} = \left\{G_{t-1}^{(i)}, X_{t-1}^{(i)}, L_{t-1}^{(i)} w_{t-1}^{(i)}\right\} \quad (u_t, z_t, \lambda_t)$

for each particle i

1 Propose modifications to topology based on metric and semantic maps

2 Perform Bayesian update of Gaussian

3 Update Dirichlet over labels based on language

4 Update weights based on metric observations

Return: $P_t^{(i)} = \left\{G_t^{(i)}, X_t^{(i)}, L_t^{(i)} w_t^{(i)}\right\}$

Rao-Blackwellized Particle Filter

Input: $P_{t-1} = \left\{ G_{t-1}^{(i)}, X_{t-1}^{(i)}, L_{t-1}^{(i)} w_{t-1}^{(i)} \right\} \quad (u_t, z_t, \lambda_t)$

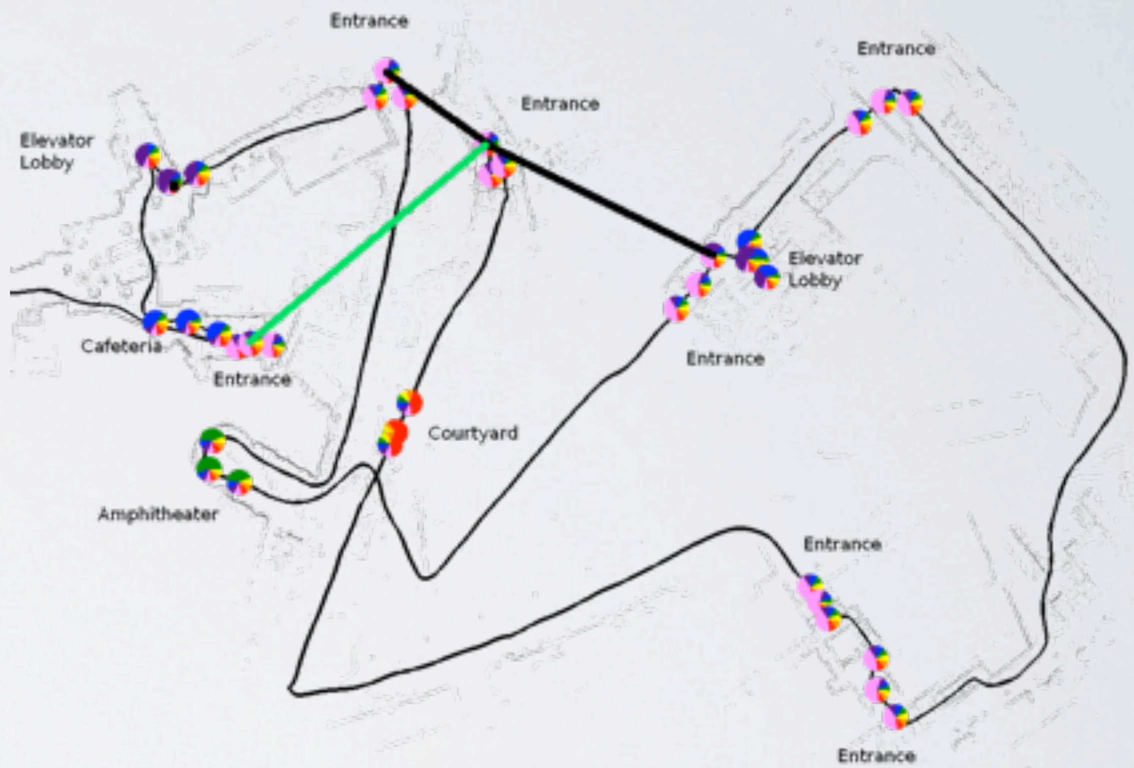
for each particle i

1 Propose modifications to topology based on metric and semantic maps

2 Perform Bayesian update of Gaussian

3 Update Dirichlet over labels based on language

4 Update weights based on metric observations



Return: $P_t^{(i)} = \left\{ G_t^{(i)}, X_t^{(i)}, L_t^{(i)} w_t^{(i)} \right\}$

Rao-Blackwellized Particle Filter

Input: $P_{t-1} = \left\{ G_{t-1}^{(i)}, X_{t-1}^{(i)}, L_{t-1}^{(i)} w_{t-1}^{(i)} \right\} \quad (u_t, z_t, \lambda_t)$

for each particle i

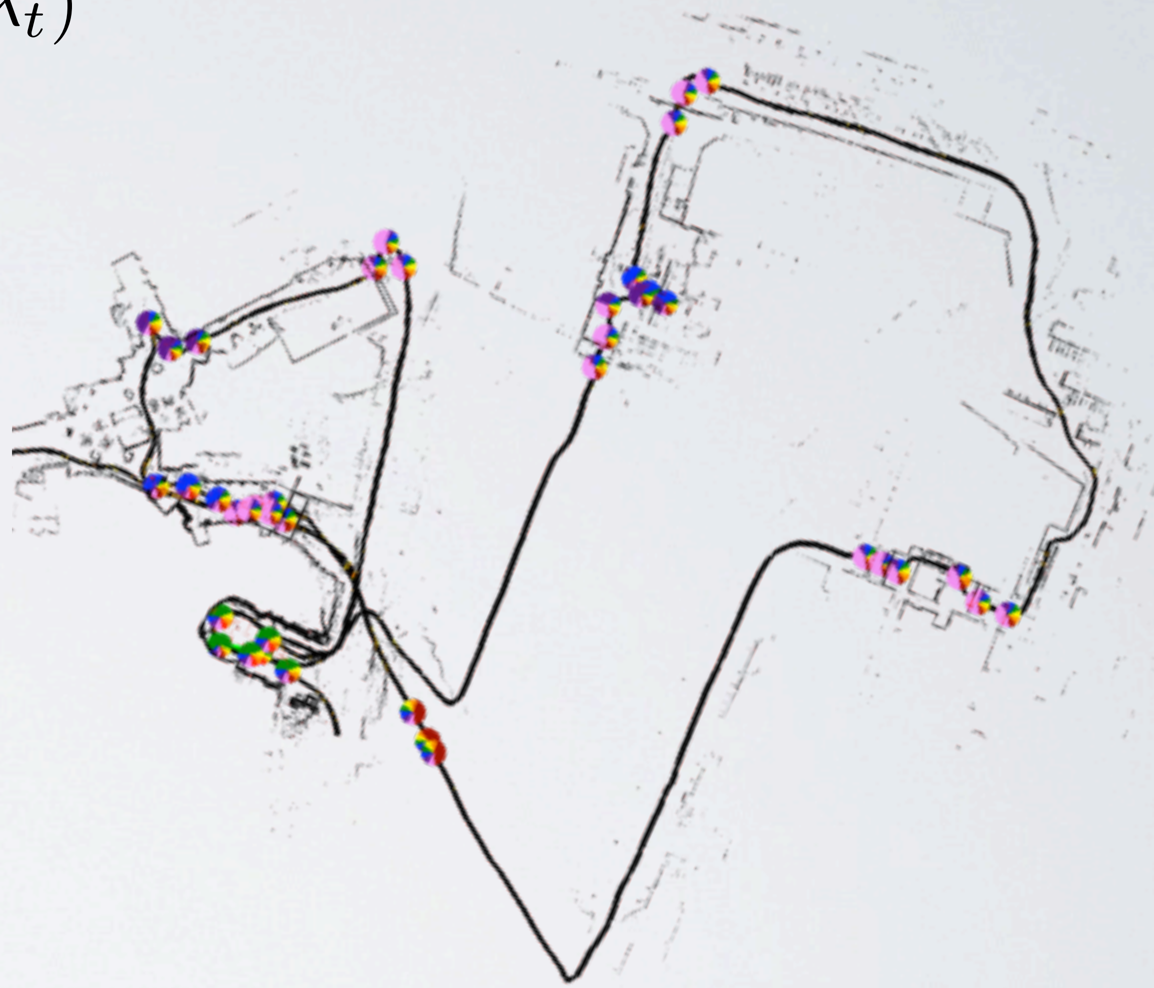
1 Propose modifications to topology based on metric and semantic maps

2 Perform Bayesian update of Gaussian

3 Update Dirichlet over labels based on language

4 Update weights based on metric observations

Return: $P_t^{(i)} = \left\{ G_t^{(i)}, X_t^{(i)}, L_t^{(i)} w_t^{(i)} \right\}$



Rao-Blackwellized Particle Filter

Input: $P_{t-1} = \left\{ G_{t-1}^{(i)}, X_{t-1}^{(i)}, L_{t-1}^{(i)} w_{t-1}^{(i)} \right\} \quad (u_t, z_t, \lambda_t)$

for each particle i

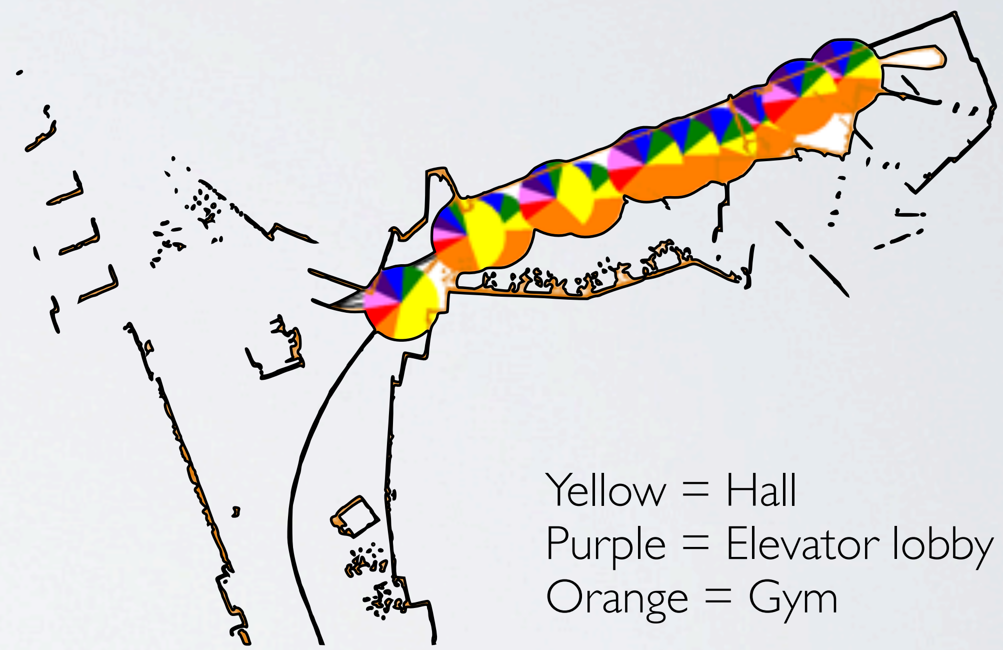
1 Propose modifications to topology based on metric and semantic maps

2 Perform Bayesian update of Gaussian

3 Update Dirichlet over labels based on language

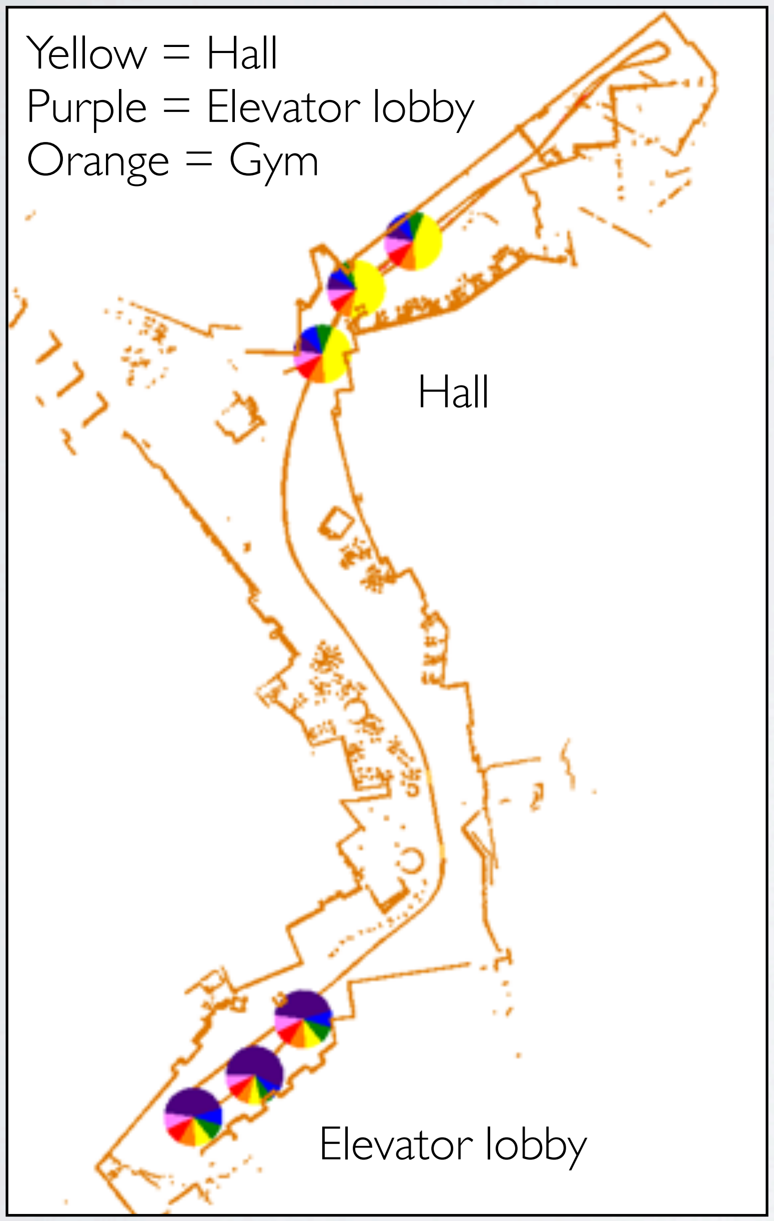
4 Update weights based on metric observations

Return: $P_t^{(i)} = \left\{ G_t^{(i)}, X_t^{(i)}, L_t^{(i)} w_t^{(i)} \right\}$



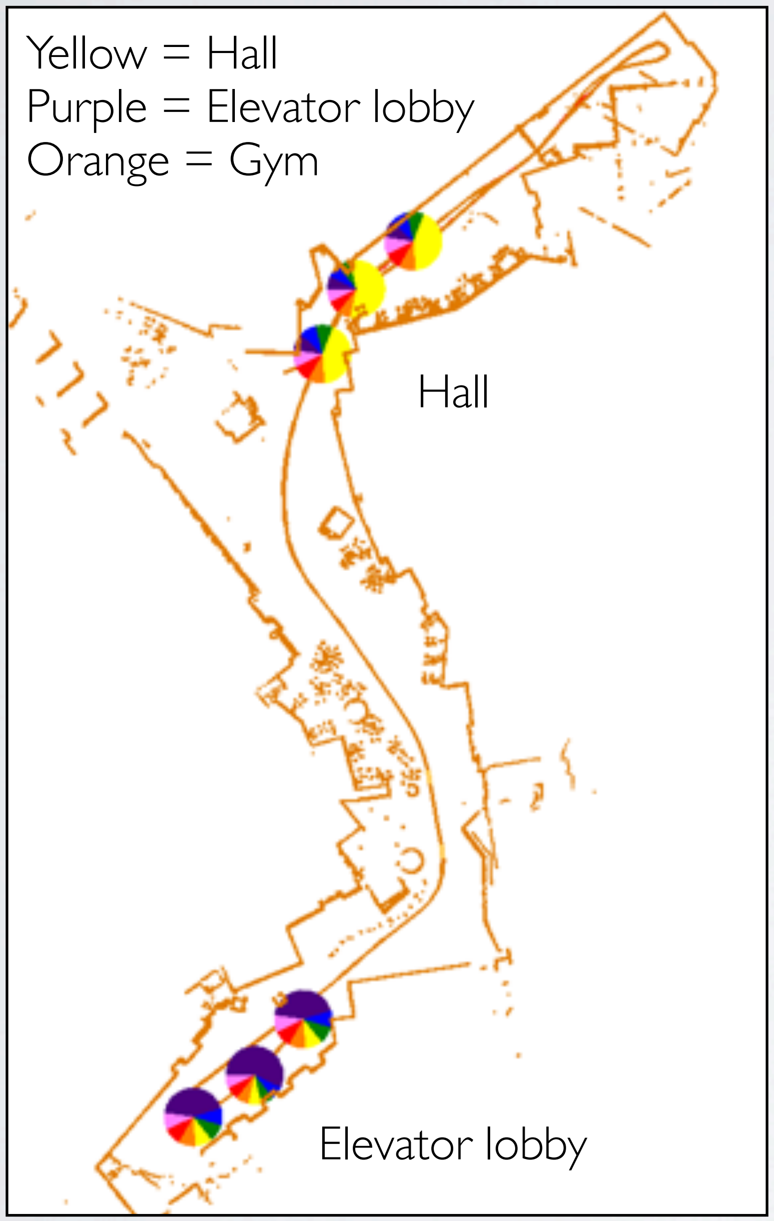
Incorporating Natural Language Descriptions

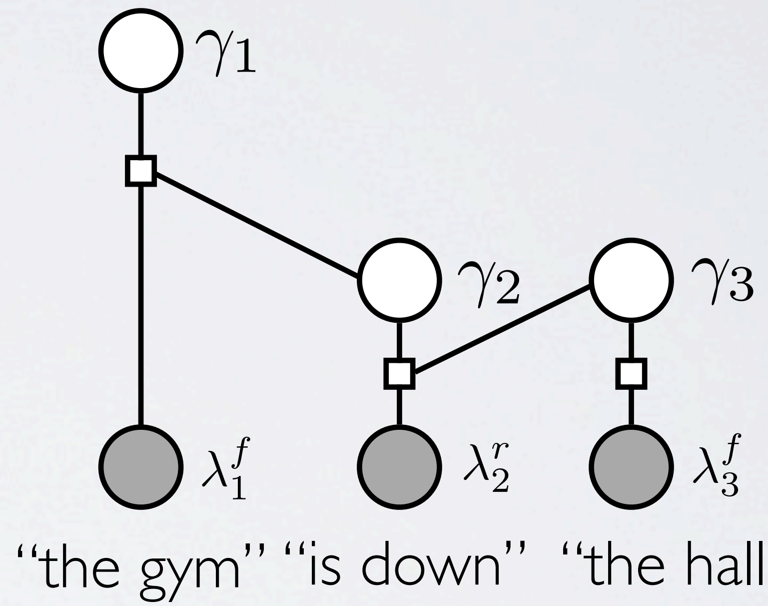
"the gym is down the hall"



Incorporating Natural Language Descriptions

"the gym is down the hall"

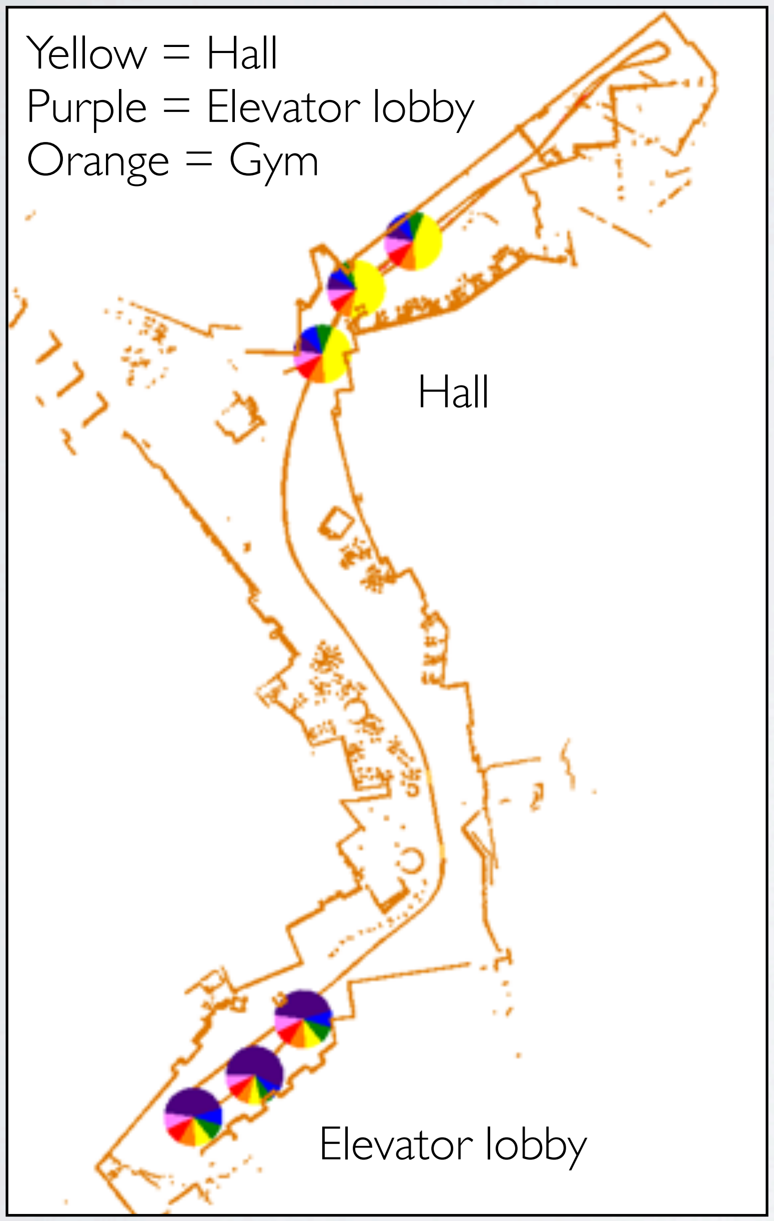


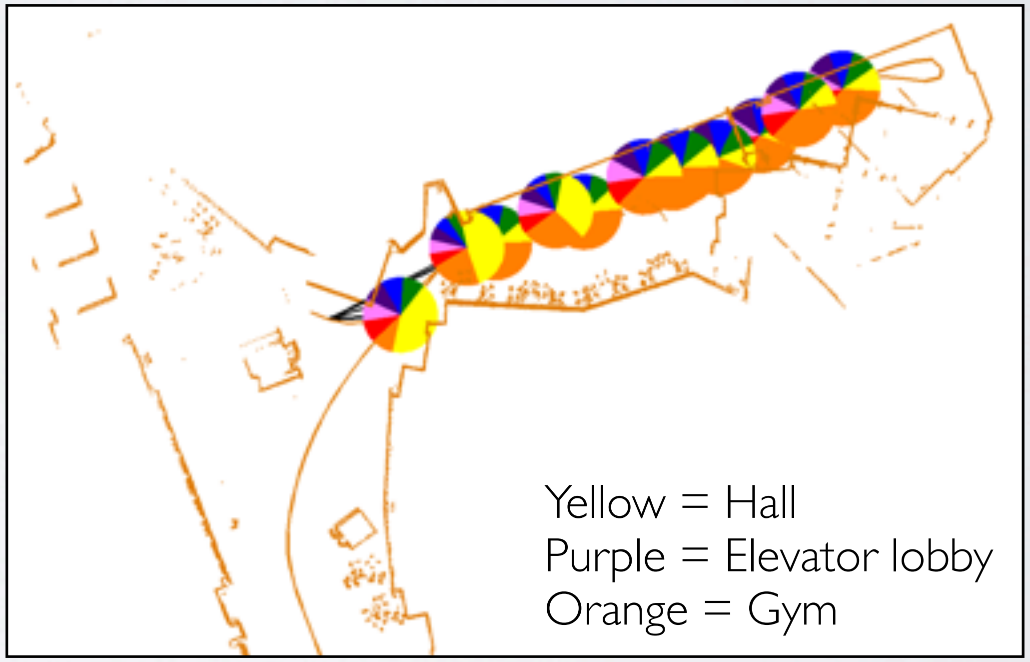


$$p(L_t^{(i)} | L_{t-1}^{(i)}, G_t^{(i)}, X_t^{(i)}, \lambda_t) = \sum_{\gamma} p(L_t^{(i)} | \gamma, L_{t-1}^{(i)}, \lambda_t) \times p(\gamma | L_{t-1}^{(i)}, G_t^{(i)}, X_t^{(i)}, \lambda_t)$$

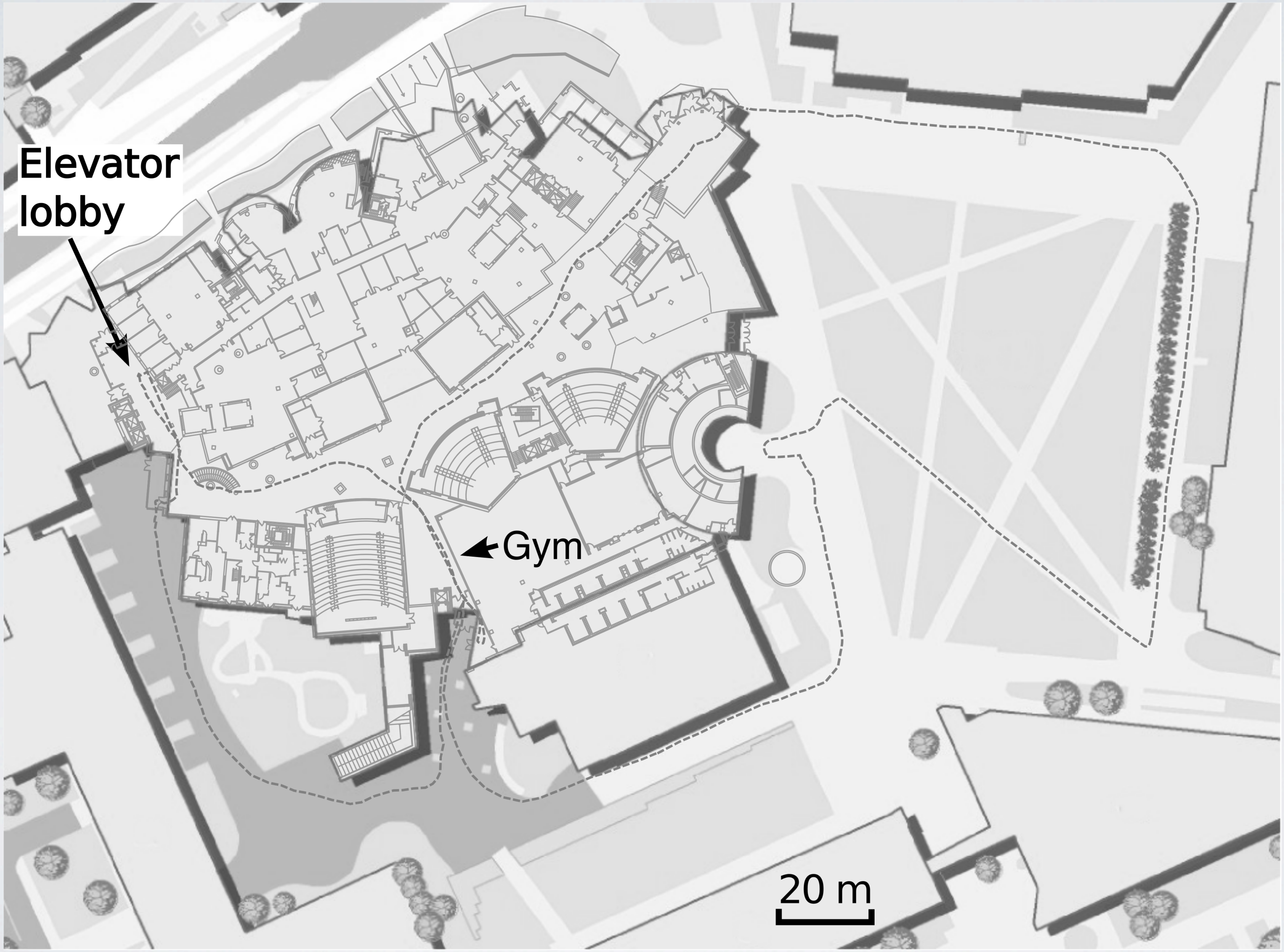
Incorporating Natural Language Descriptions

"the gym is down the hall"

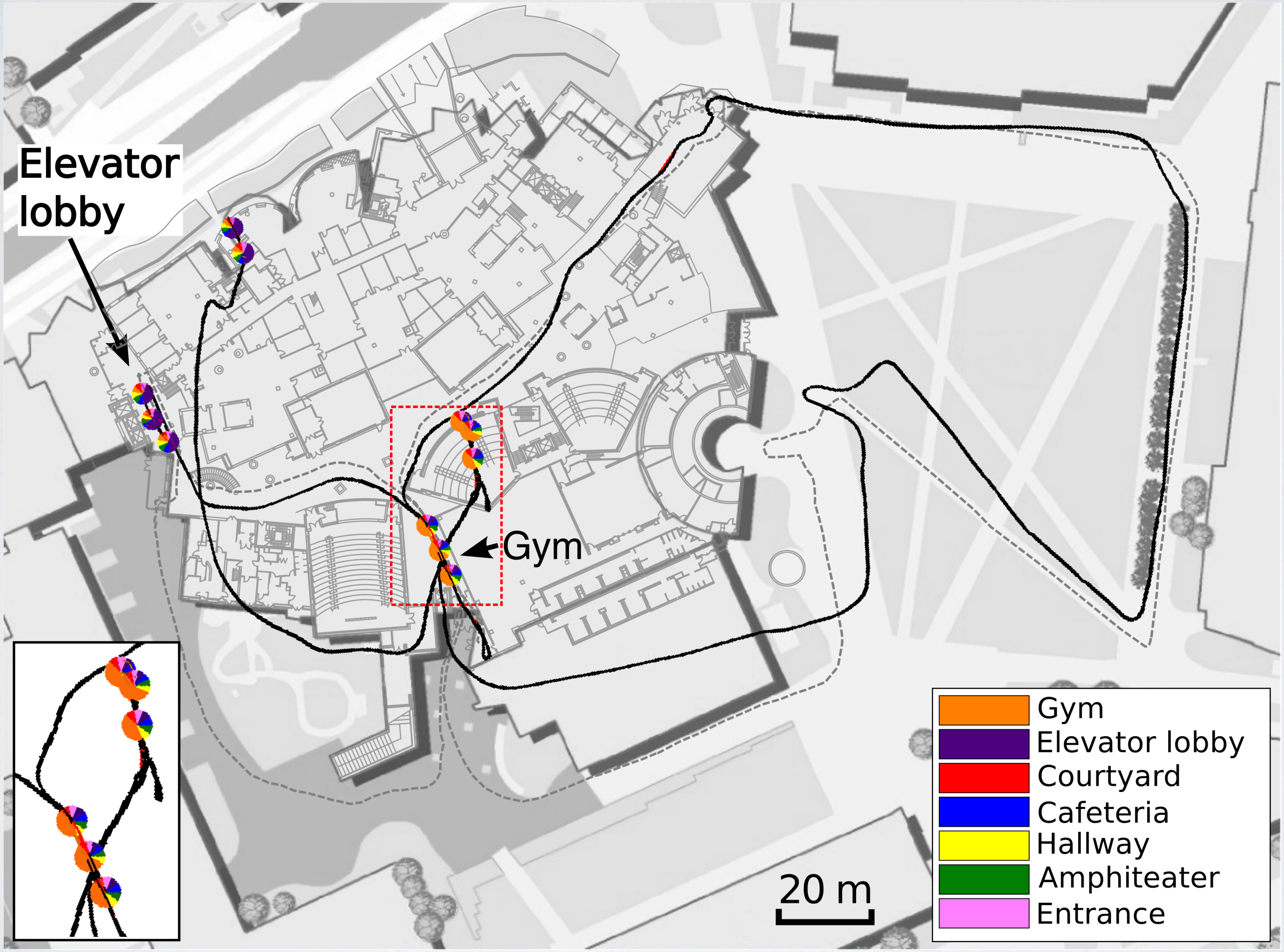




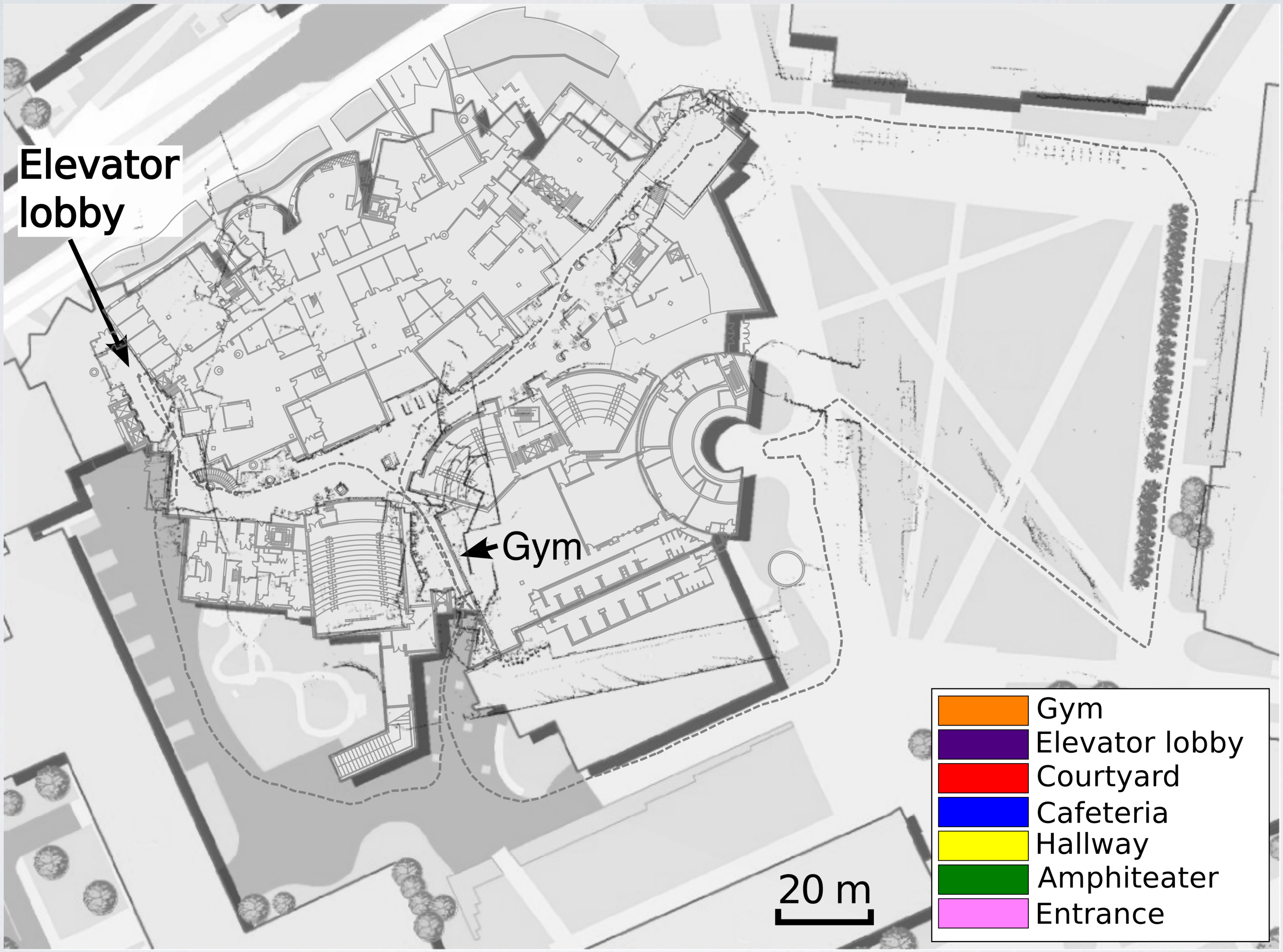
No Language Constraints



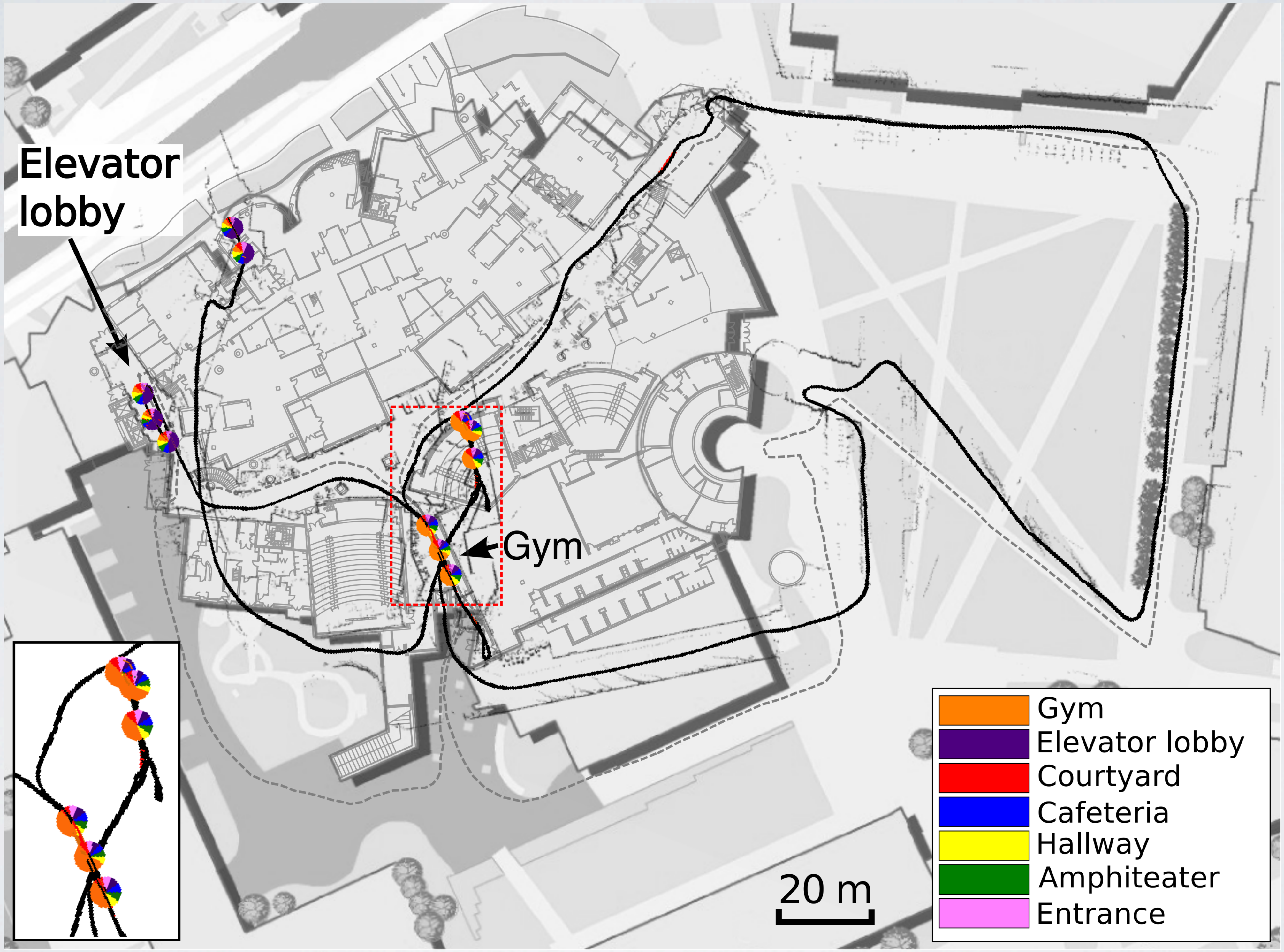
No Language Constraints



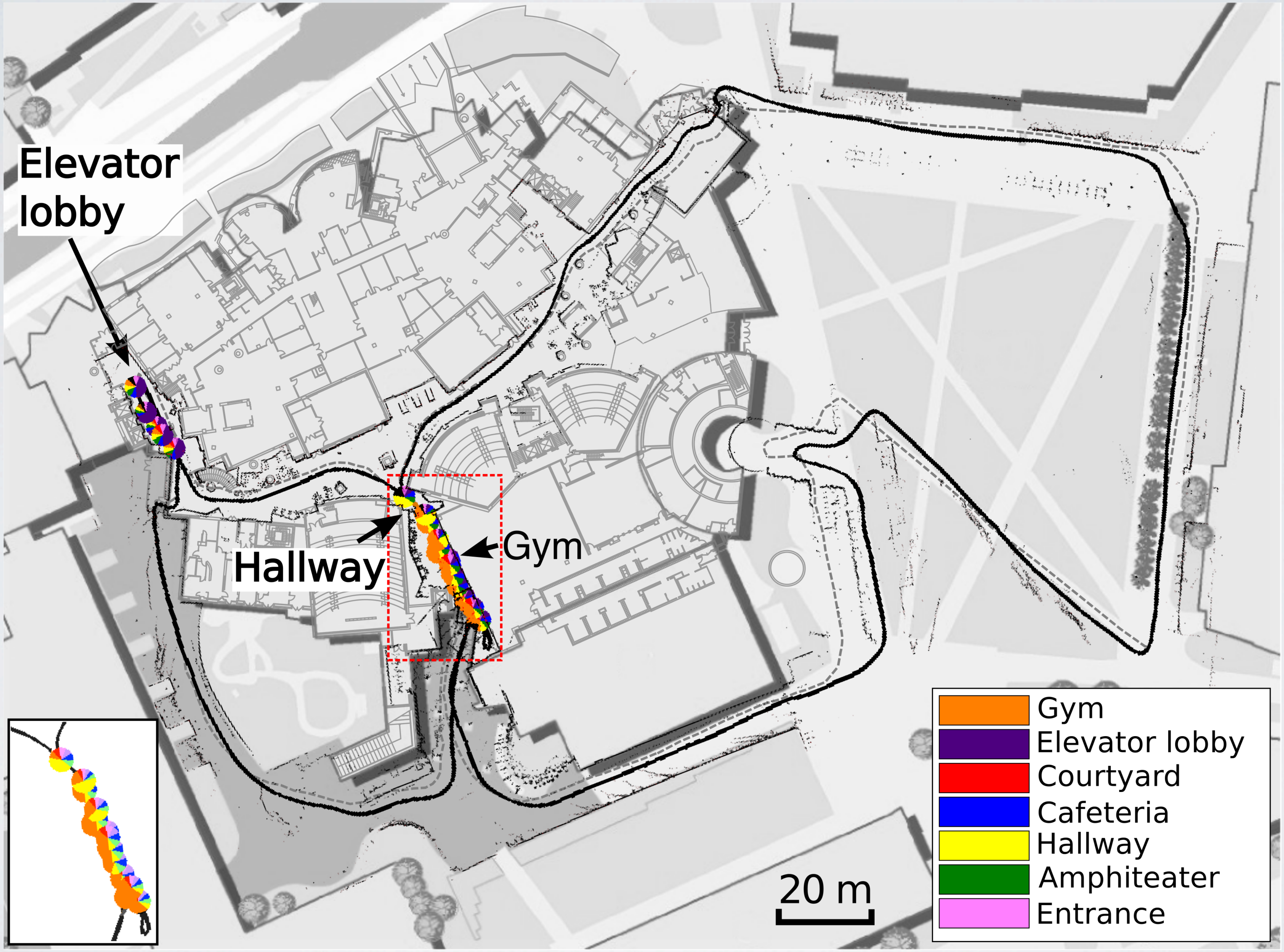
No Language Constraints



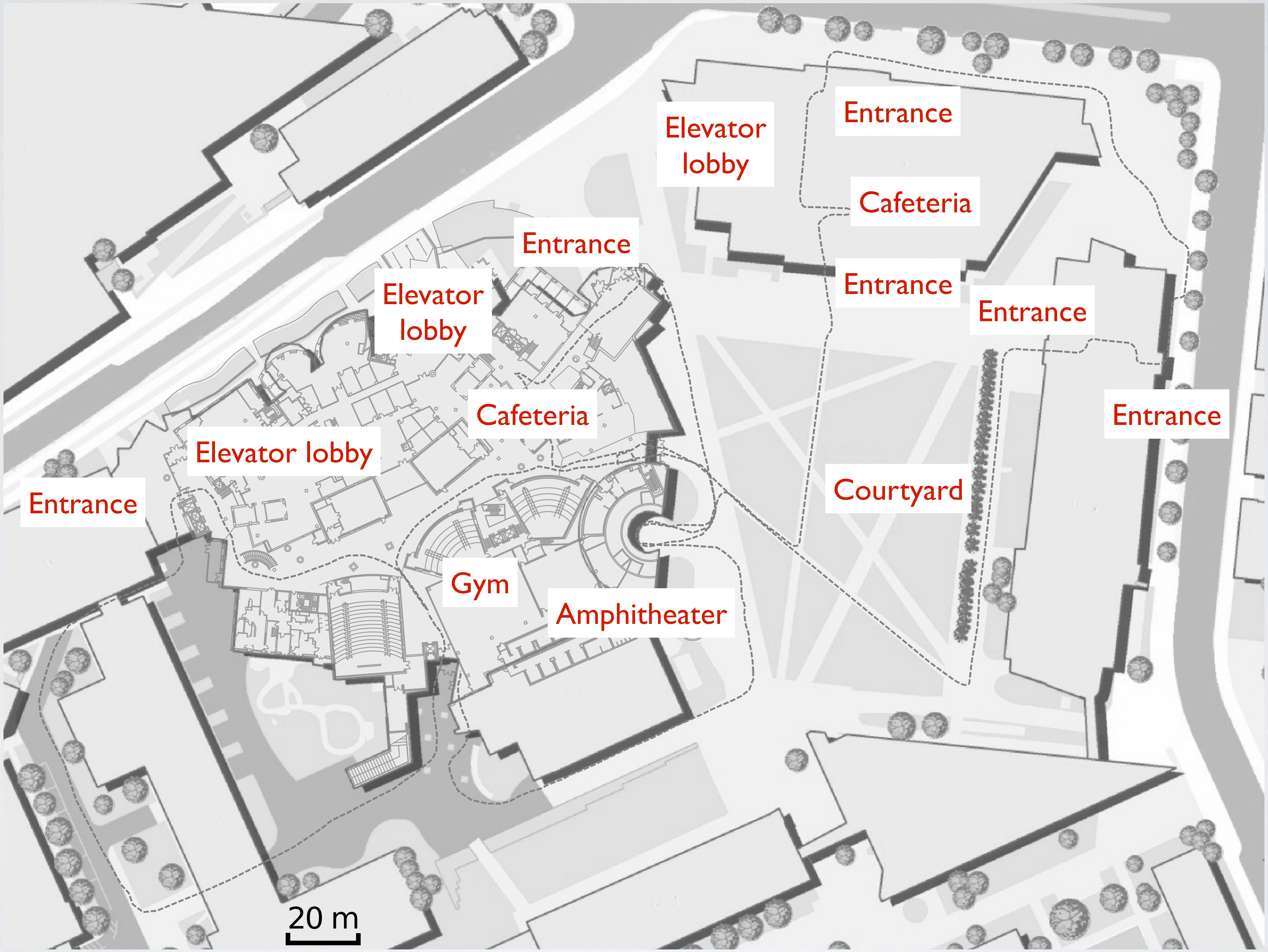
No Language Constraints



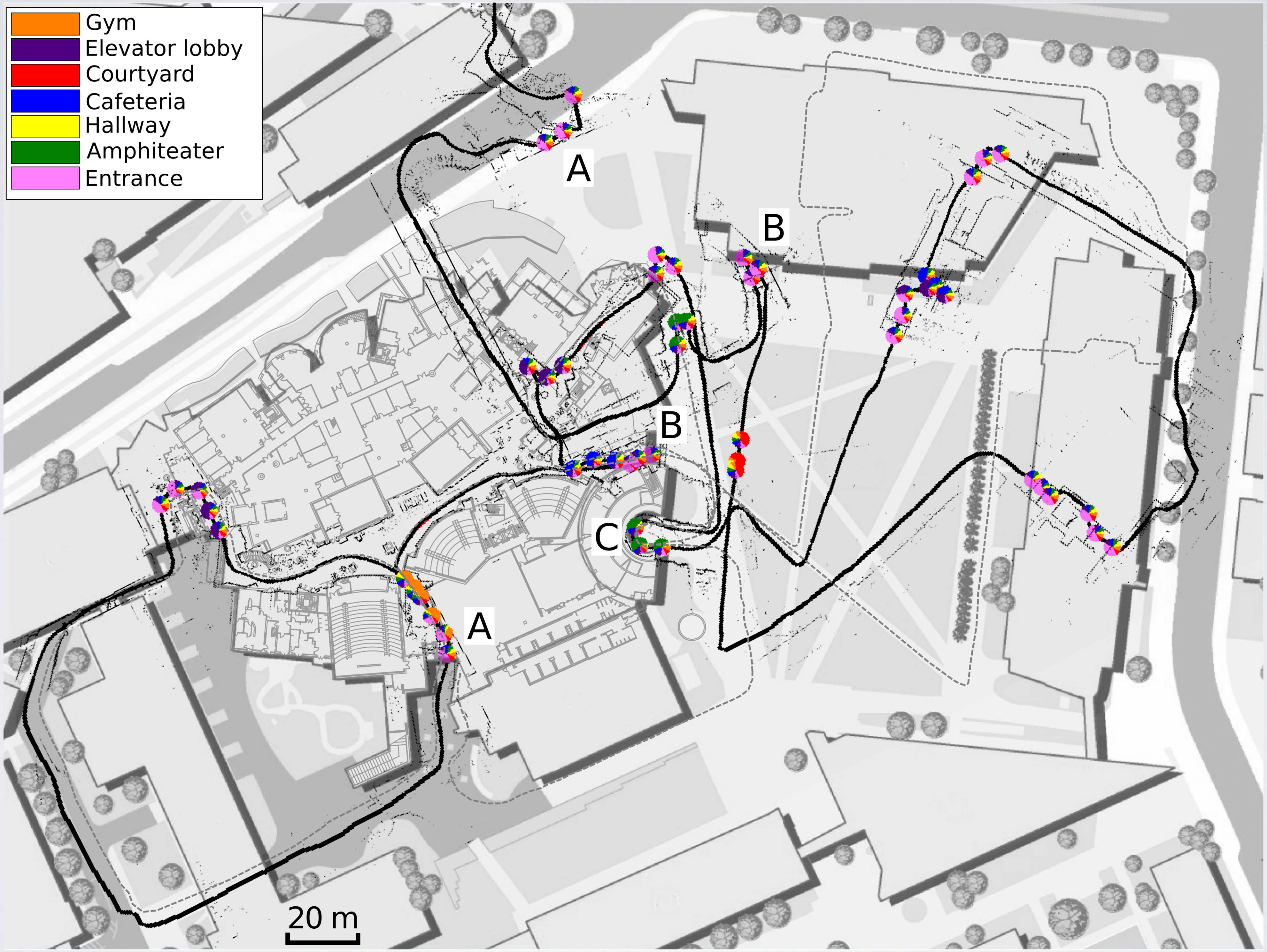
With Language Constraints



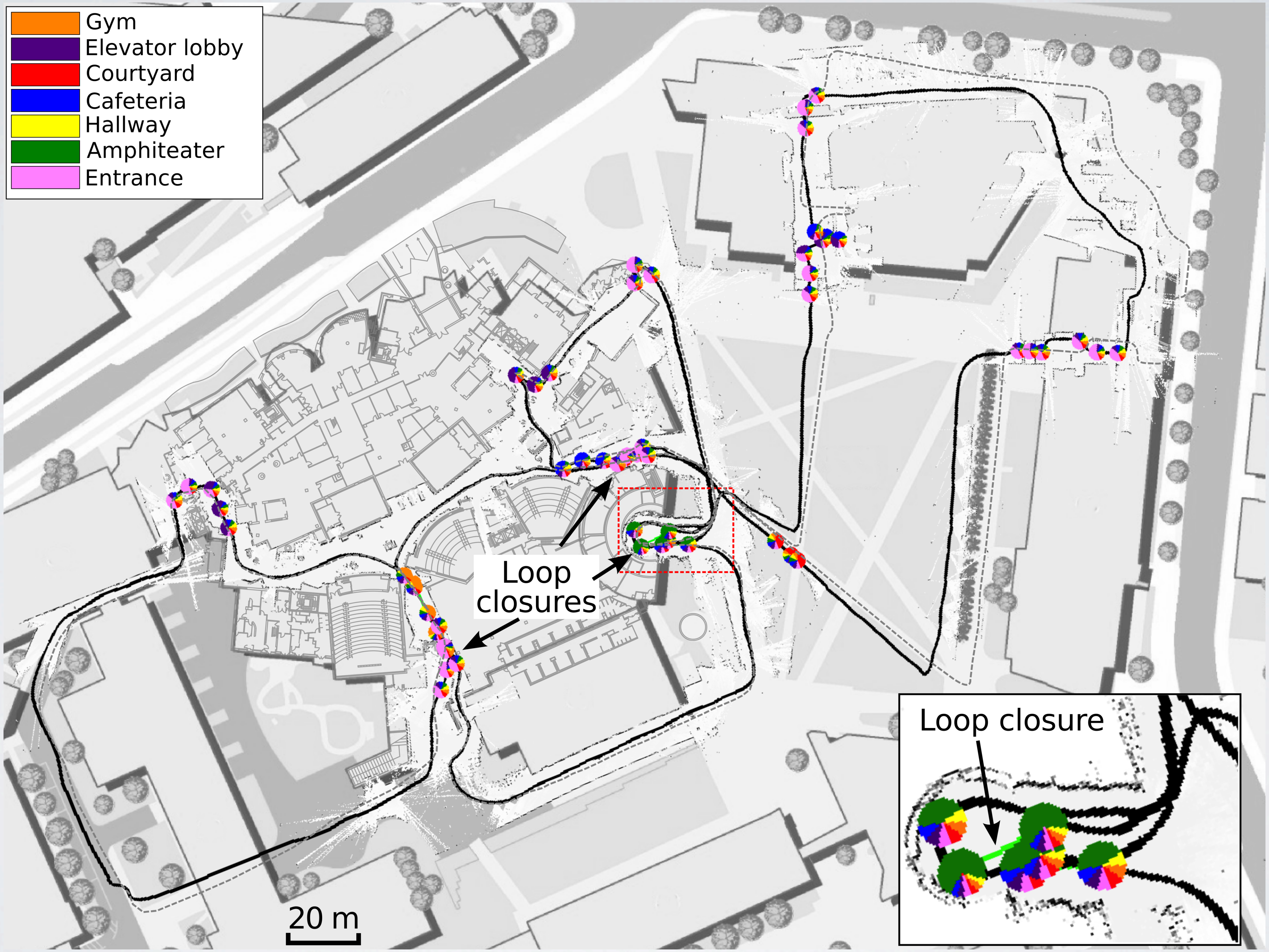
No Language Constraints



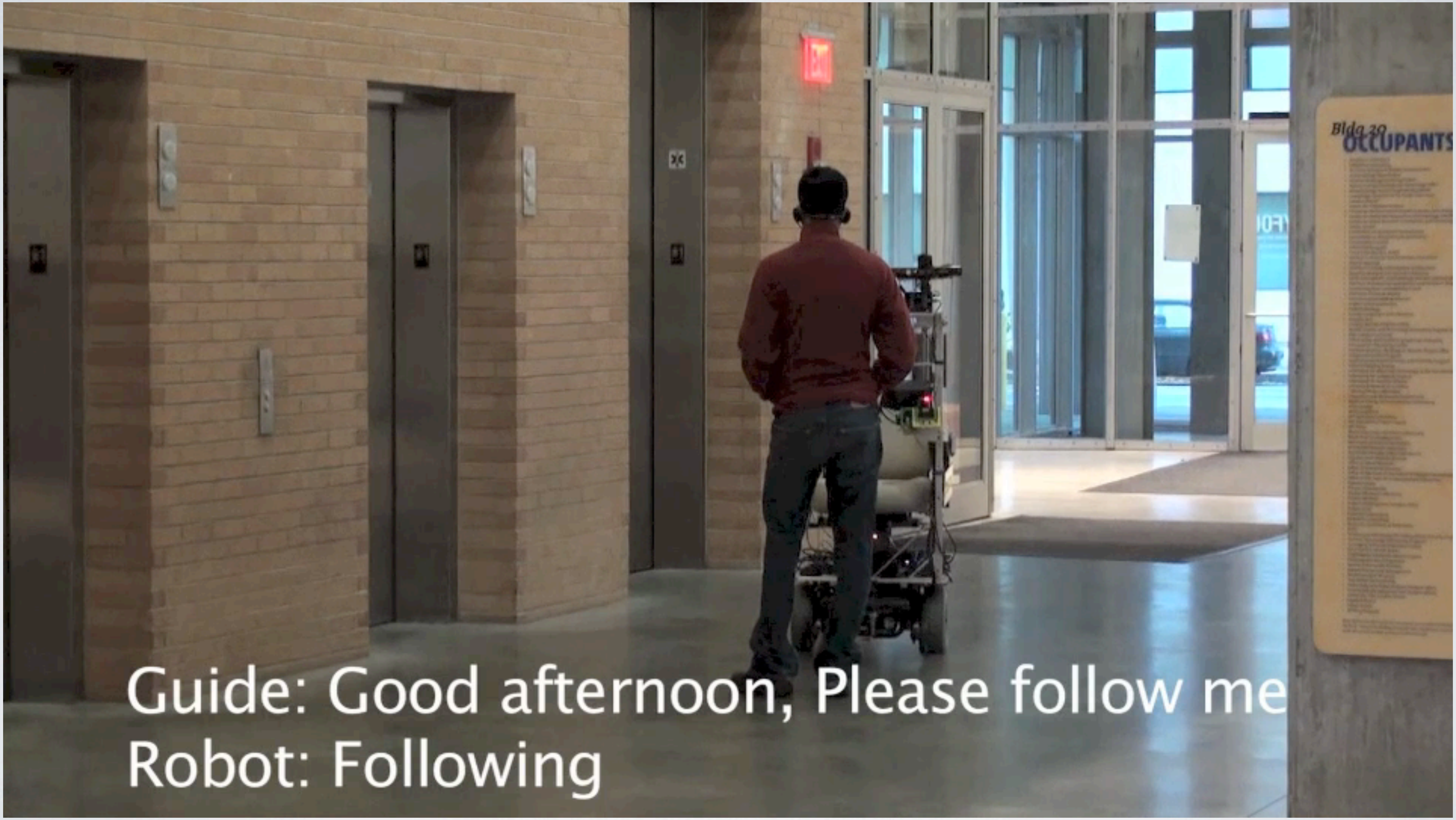
No Language Constraints



With Language Constraints



Preliminary Results - With Language Constraints



I. Importance of Situational Awareness

II. Persistent Object Awareness with Vision

III. Semantic Map Learning from Natural Language Descriptions

IV. Future Directions

V. Conclusions

Enhancing Models of Objects and Space

- Object category recognition
  - Data-driven models
  - Transfer learning
  - Limited supervision via human intervention
  - Efficient retrieval and matching

- New sources of information
  - Objects (e.g., co-occurrence)
  - Vision-based scene classification
  - Higher-level concepts
  - Building topology databases

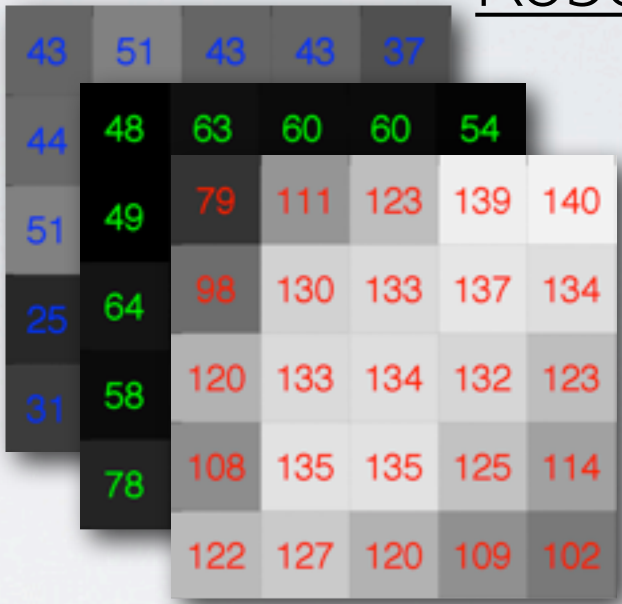
- Exploration-based natural language grounding

Where are We Going?

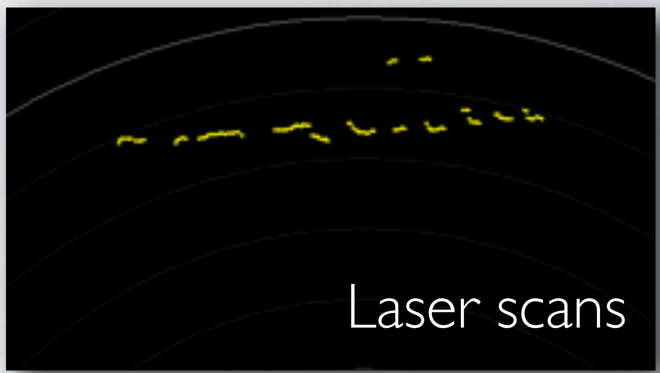
People

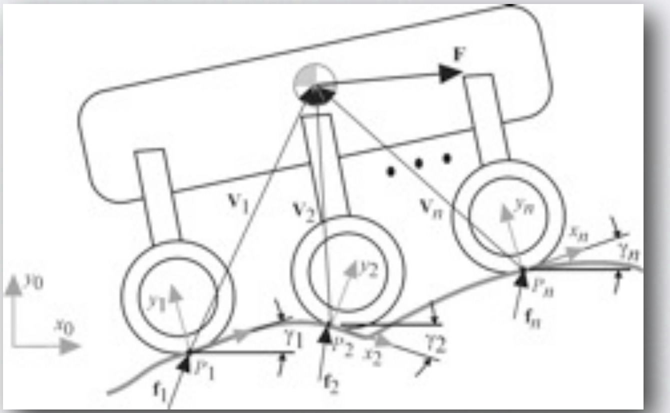
- Objects
- Places
- Actions
- People
- Events

Robots

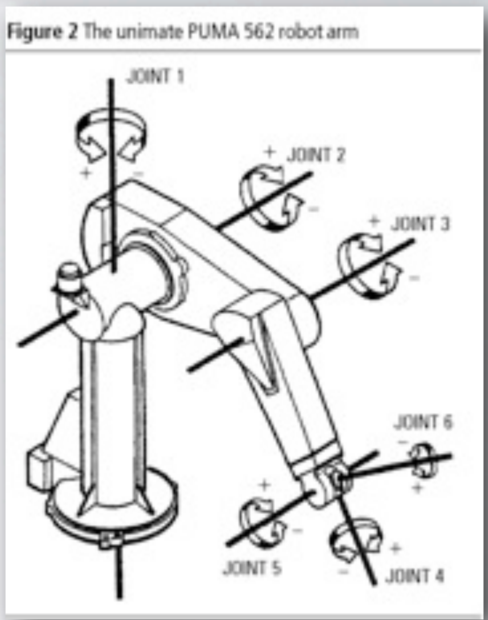


Images





Wheel torques



Joint angles

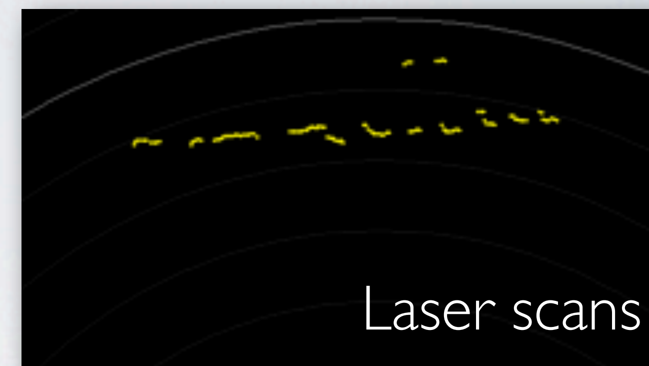
Where are We Going?

People

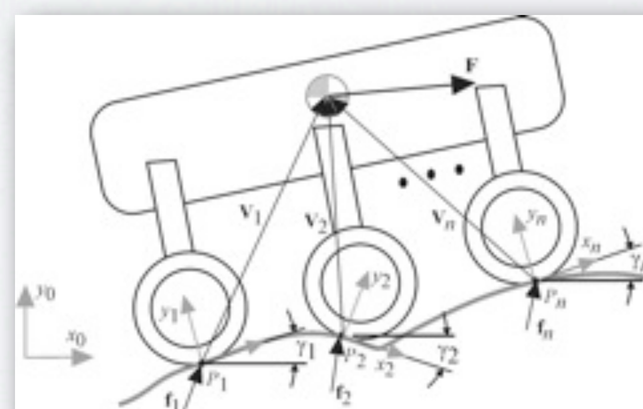
- Objects
- Places
- Actions
- People
- Events

Robots

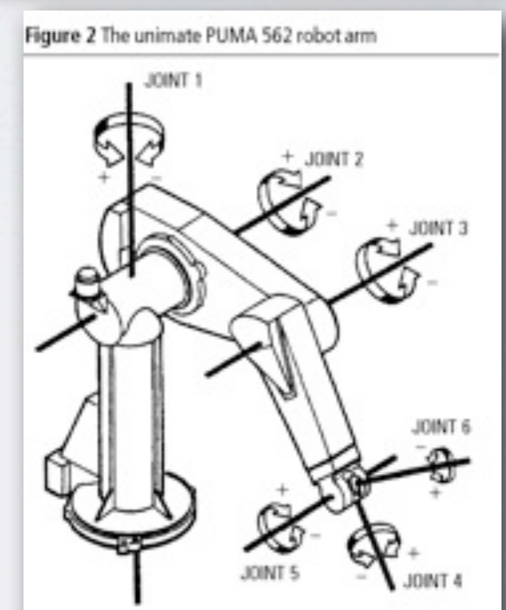
Images



Laser scans



Wheel torques



Joint angles

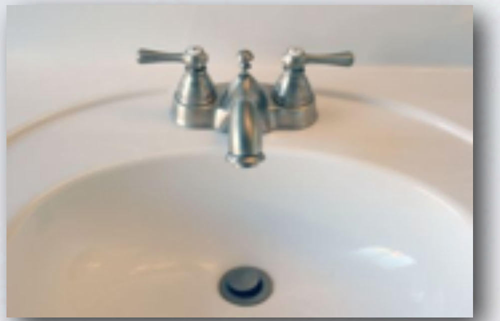
Learning Rich Action Spaces

- Low-level, object-specific actions don't scale
- Long-term planning & inference is intractable

Courtesy: Willow Garage

Learning Rich Action Spaces

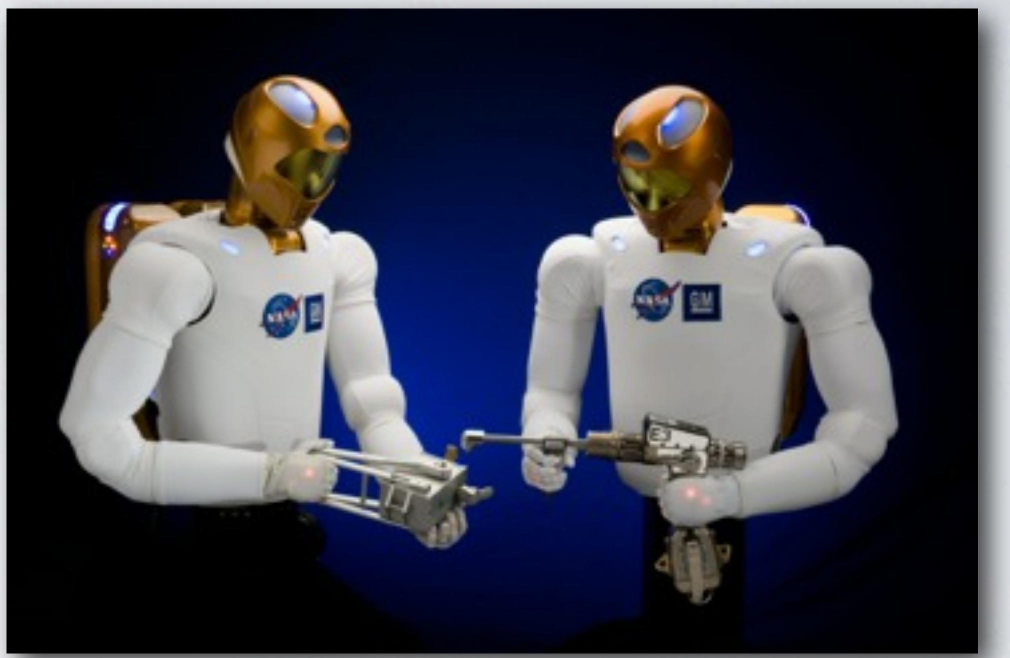
- Low-level, object-specific actions don't scale
- Long-term planning & inference is intractable





Learning Rich Action Spaces

- Robots need higher-level representations
  - Structured state/action space
  - Affordance-based action model
  - Affordances are grounded in perception
- Human-provided information is critical to efficient learning
- Robots must formulate representation based on experience



Learning Rich Action Spaces

- Robots need higher-level representations
  - Structured state/action space
  - Affordance-based action model
  - Affordances are grounded in perception
- Human-provided information is critical to efficient learning
- Robots must formulate representation based on experience

Learning Rich Action Spaces

- Robots need higher-level representations
  - Structured state/action space
  - Affordance-based action model
  - Affordances are grounded in perception
- Human-provided information is critical to efficient learning
- Robots must formulate representation based on experience

Learning via Deliberate Actions

- Robots need higher-level representations
  - Structured state/action space
  - Affordance-based action model
  - Affordances are grounded in perception
- Human-provided information is critical to efficient learning
- Robots must formulate representation based on experience

I. Importance of Situational Awareness

II. Persistent Object Awareness with Vision

III. Semantic Map Learning from Natural Language Descriptions

IV. Future Directions

V. Conclusions

Contributions

- Showed that human-robot collaboration requires intuitive control
- Argued that the key missing capability is situational awareness
- Perception is critical to enabling awareness
- Demonstrated algorithms that opportunistically learn rich models of objects and space from human-provided cues

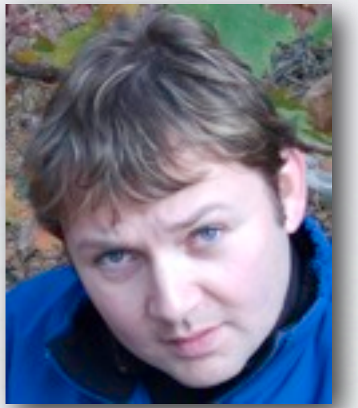
Contributions


Sachi Hemachandra


Stefanie Tellex


Bianca Homberg


Sudeep Pillai


Yuli Friedman


Matthew Antone

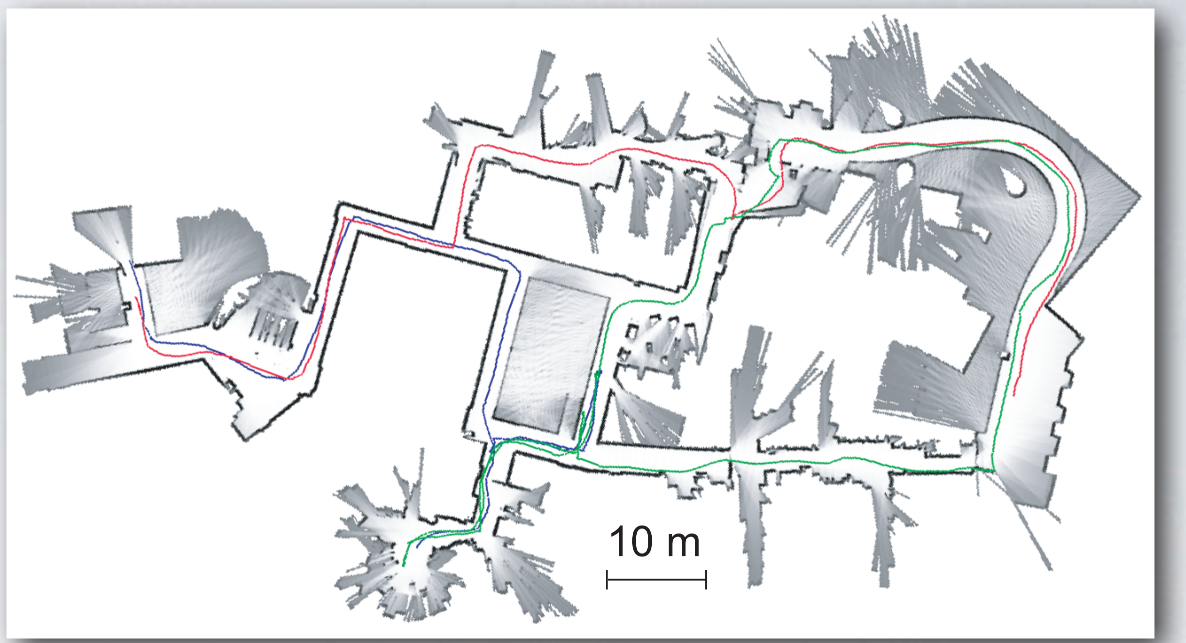

Seth Teller

Contributions

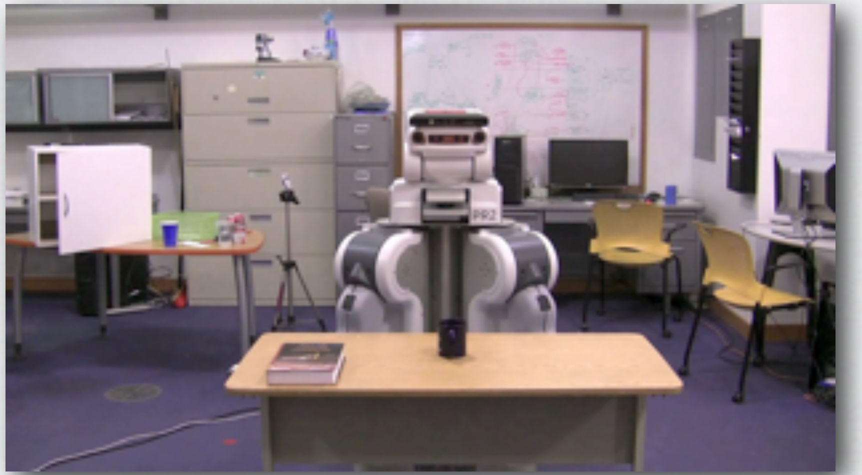
- Showed that human-robot collaboration requires intuitive control
- Argued that the key missing capability is situational awareness
- Perception is critical to enabling awareness
- Demonstrated algorithms that opportunistically learn rich models of objects and space from human-provided cues

Getting There

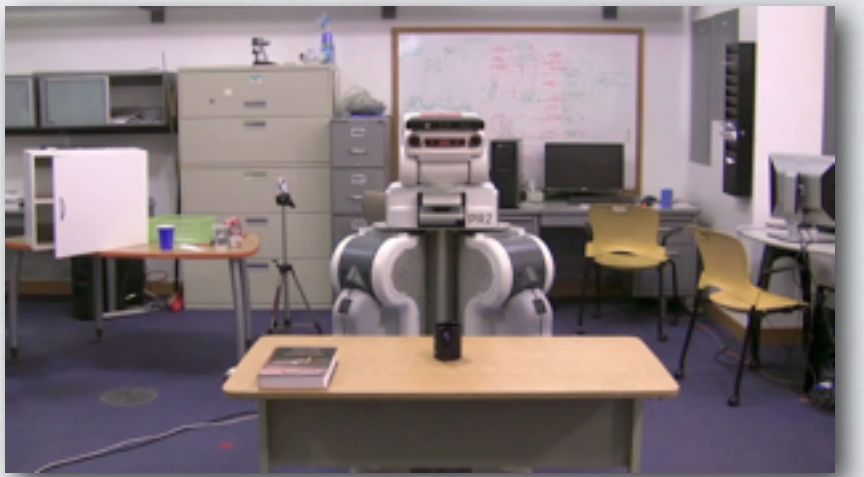
- Autonomous navigation
- Motion planning & control
- Planning under uncertainty
- Manipulation
- Localization & mapping
- Perception
- Efficient control
- Natural interaction
- Trusted autonomy



Navigation using uncalibrated cameras

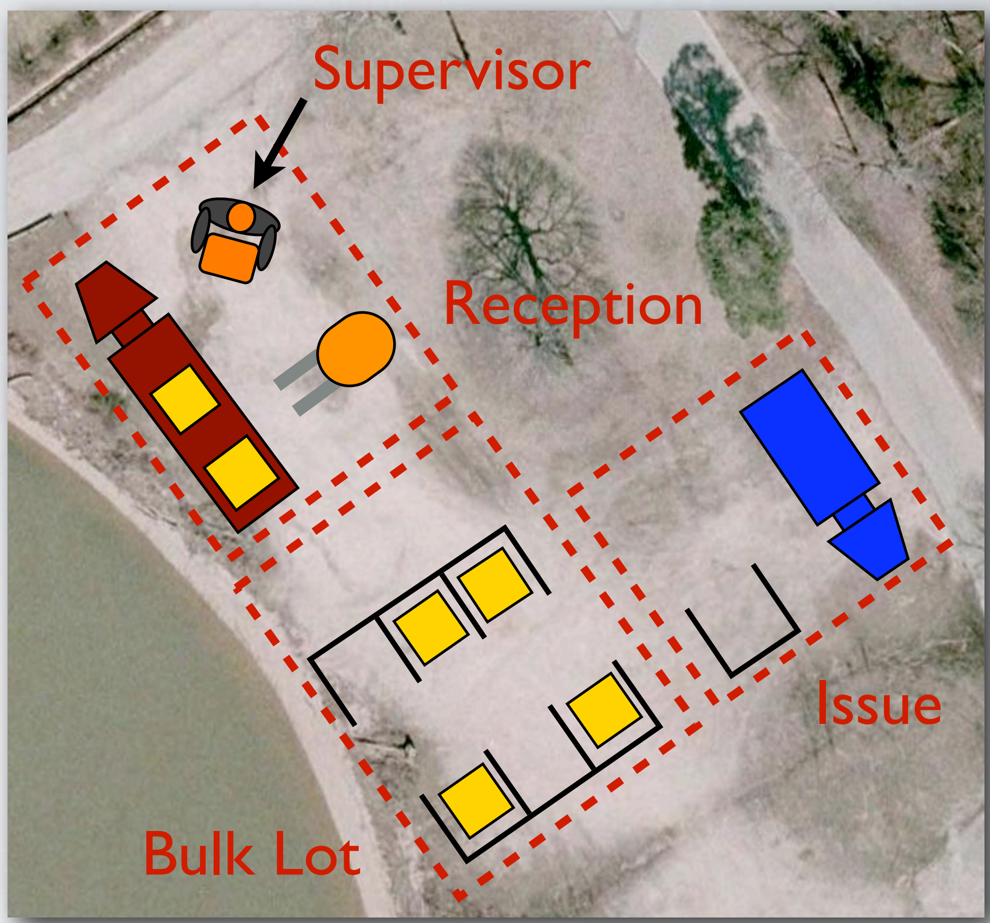


RRT



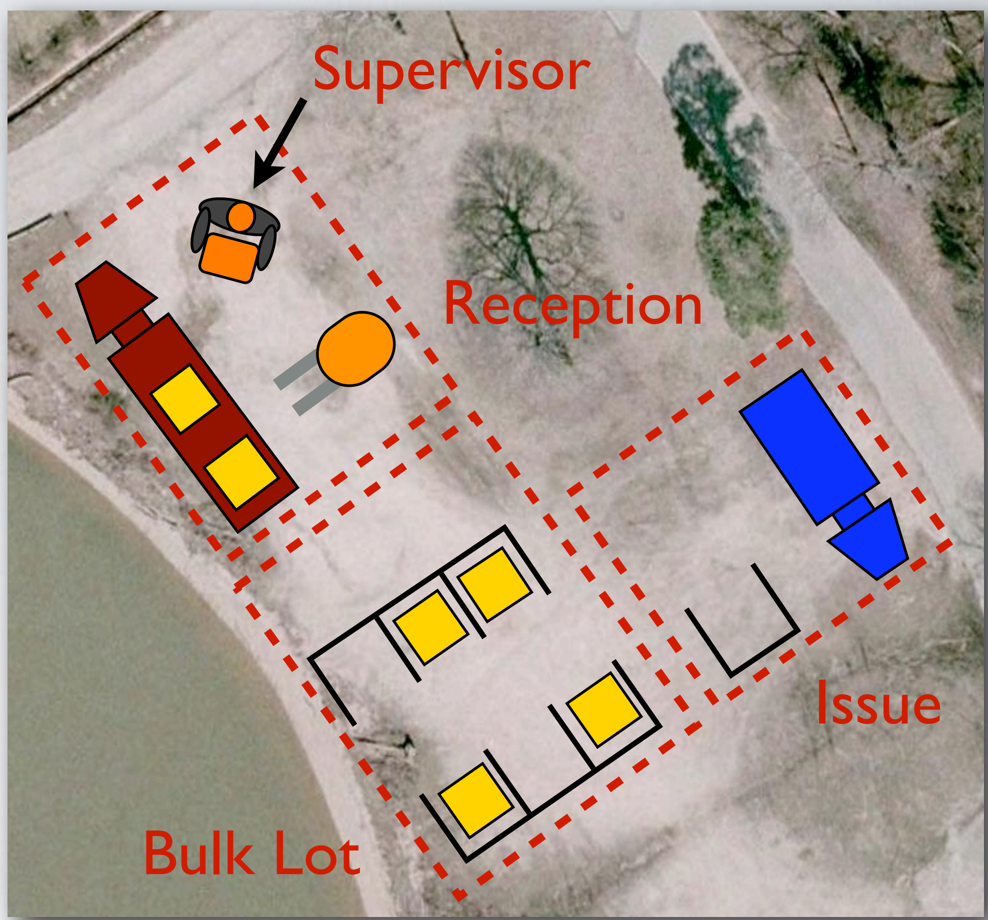
Anytime RRT*: optimal planning

Mobile Manipulation for Logistics



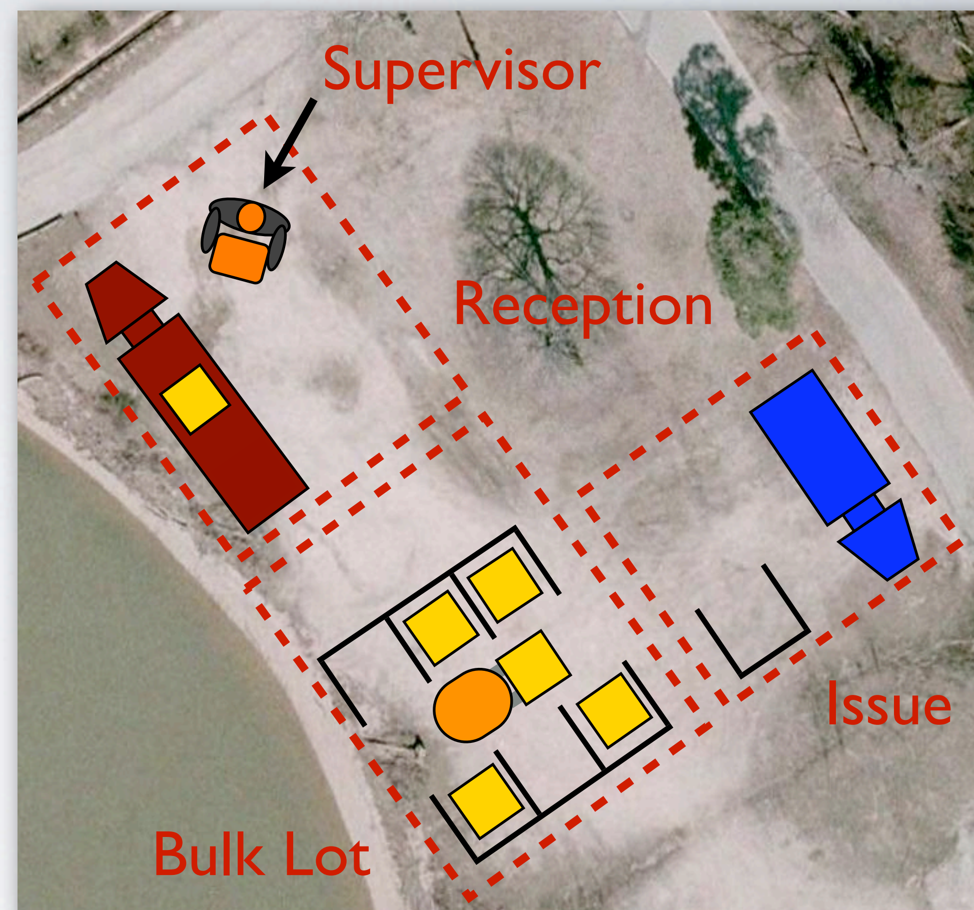
1. Pick up objects off trucks or the ground
2. Transport items to storage locations
3. Load particular objects onto customer trucks or the ground

Mobile Manipulation for Logistics



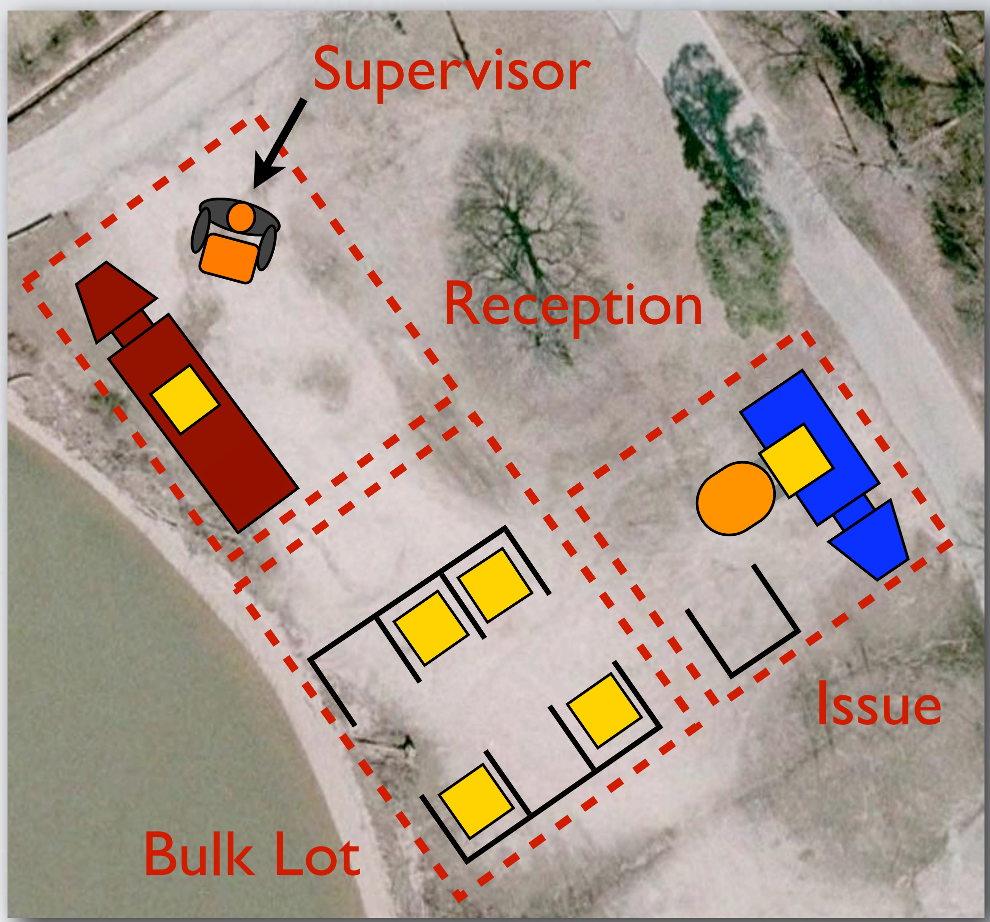
1. Pick up objects off trucks or the ground
2. Transport items to storage locations
3. Load particular objects onto customer trucks or the ground

Mobile Manipulation for Logistics



1. Pick up objects off trucks or the ground
2. Transport items to storage locations
3. Load particular objects onto customer trucks or the ground

Mobile Manipulation for Logistics



1. Pick up objects off trucks or the ground
2. Transport items to storage locations
3. Load particular objects onto customer trucks or the ground

The Platform

Teleoperation?

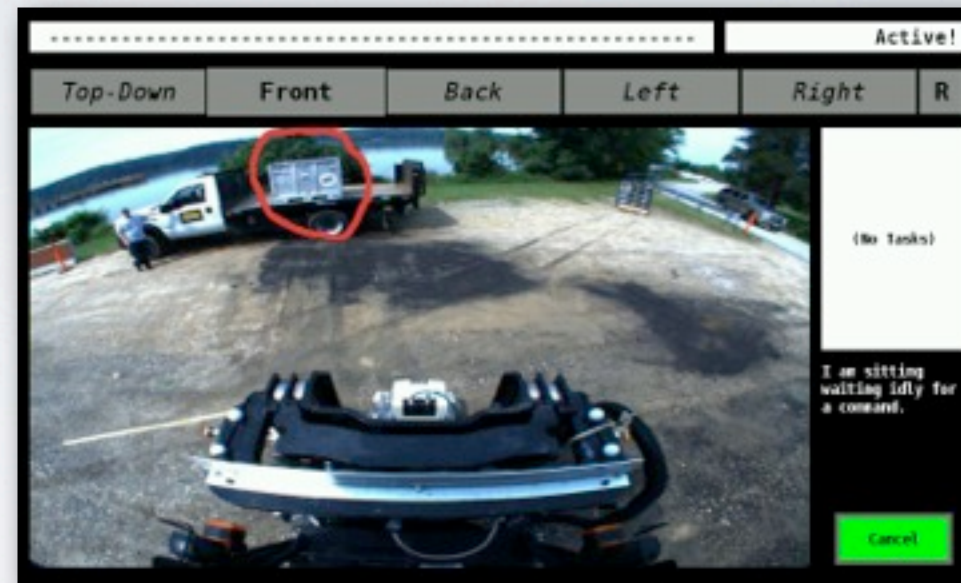
Teleoperation?

Shared Autonomy

- Hierarchical task-level autonomy
  - Reduce tasks into simpler sub-tasks
- Shared situational awareness
- Robot can request help when needed

Command via Shared World Model

- Hand-held tablet interface
  - Robot's eye view with annotated images
  - Onboard speech recognition
  - Interprets pen-based gestures
- Microphones pick up external speech



Grounding Natural Language Speech

(collaboration with S. Tellex, T. Kollar, S. Teller, & N. Roy)

$$\boxed{\arg\max_{\text{groundings}} p\left(\text{groundings} \mid \text{language}\right)}$$

objects, actions, relations, places

"Drive to the tire pallet"

Grounding Natural Language Speech

(collaboration with S. Tellex, T. Kollar, S. Teller, & N. Roy)

"To the tire pallet"

$$\arg \max_{\gamma \in \mathcal{X}} p\left(\gamma | \lambda\right)$$

Grounding Natural Language Speech

(collaboration with S. Tellex, T. Kollar, S. Teller, & N. Roy)

"To the tire pallet"

$$\arg\max_{\gamma \in \mathcal{X}} p\left(\gamma | \lambda\right)$$

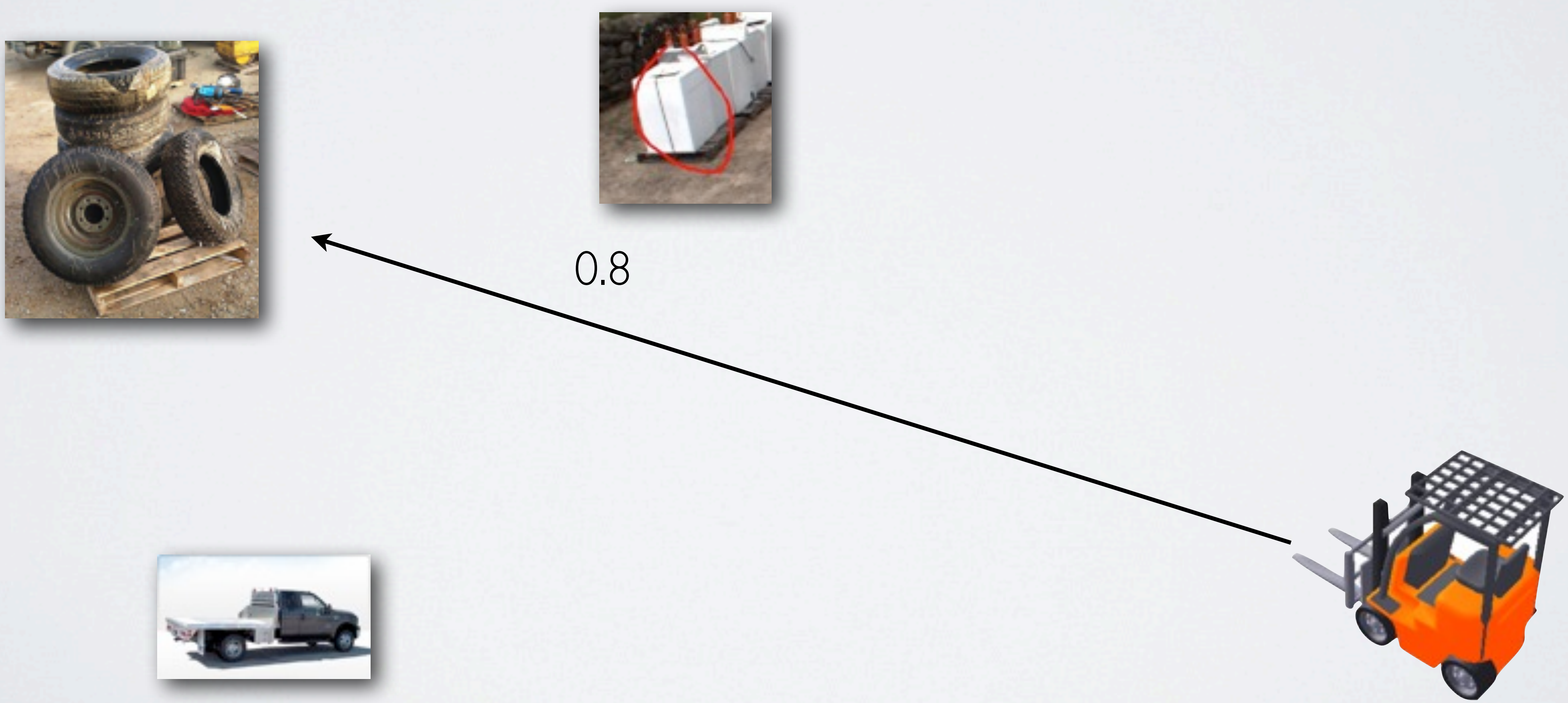
0.1

Grounding Natural Language Speech

(collaboration with S. Tellex, T. Kollar, S. Teller, & N. Roy)

"To the tire pallet"

$$\arg\max_{\gamma \in \mathcal{X}} p\left(\gamma | \lambda\right)$$

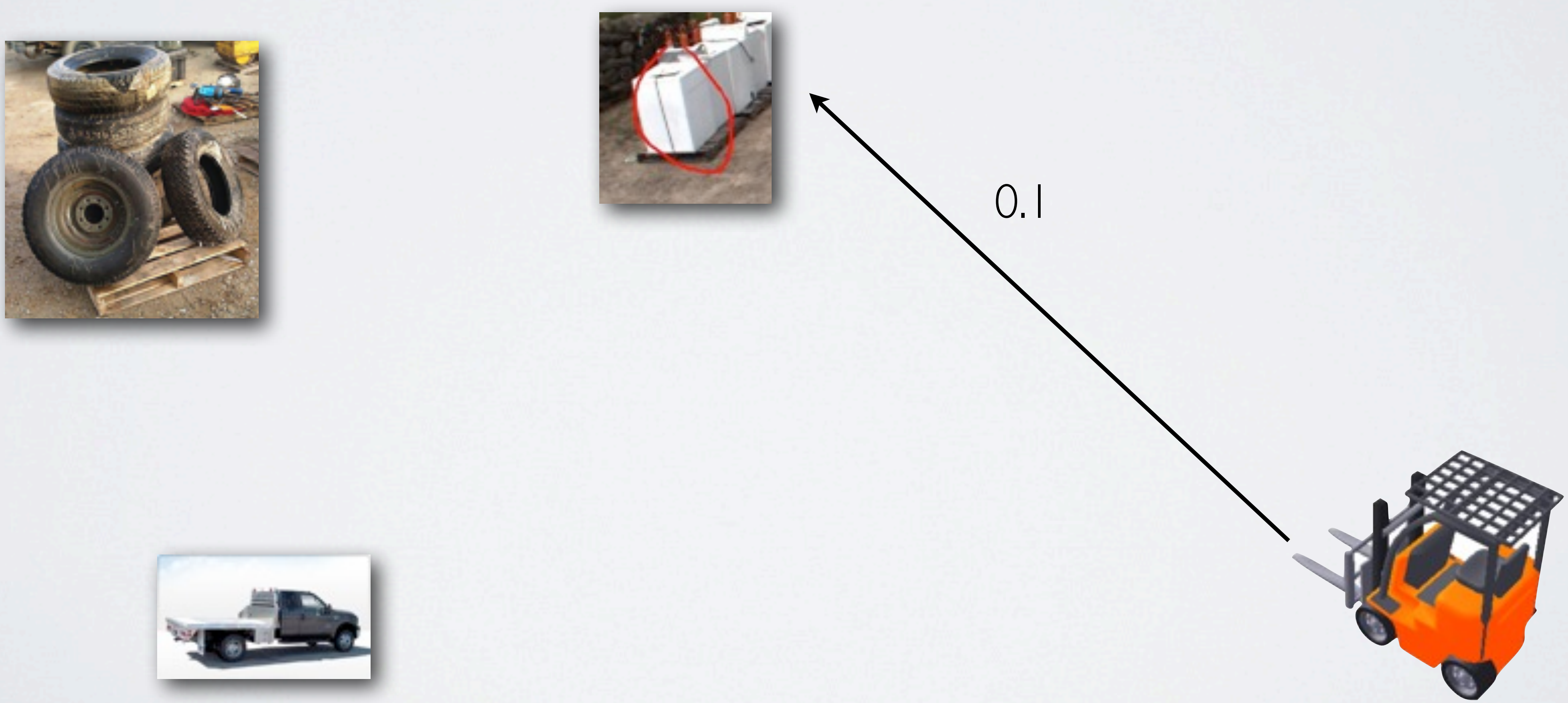


Grounding Natural Language Speech

(collaboration with S. Tellex, T. Kollar, S. Teller, & N. Roy)

"To the tire pallet"

$$\arg\max_{\gamma \in \mathcal{X}} p\left(\gamma | \lambda\right)$$

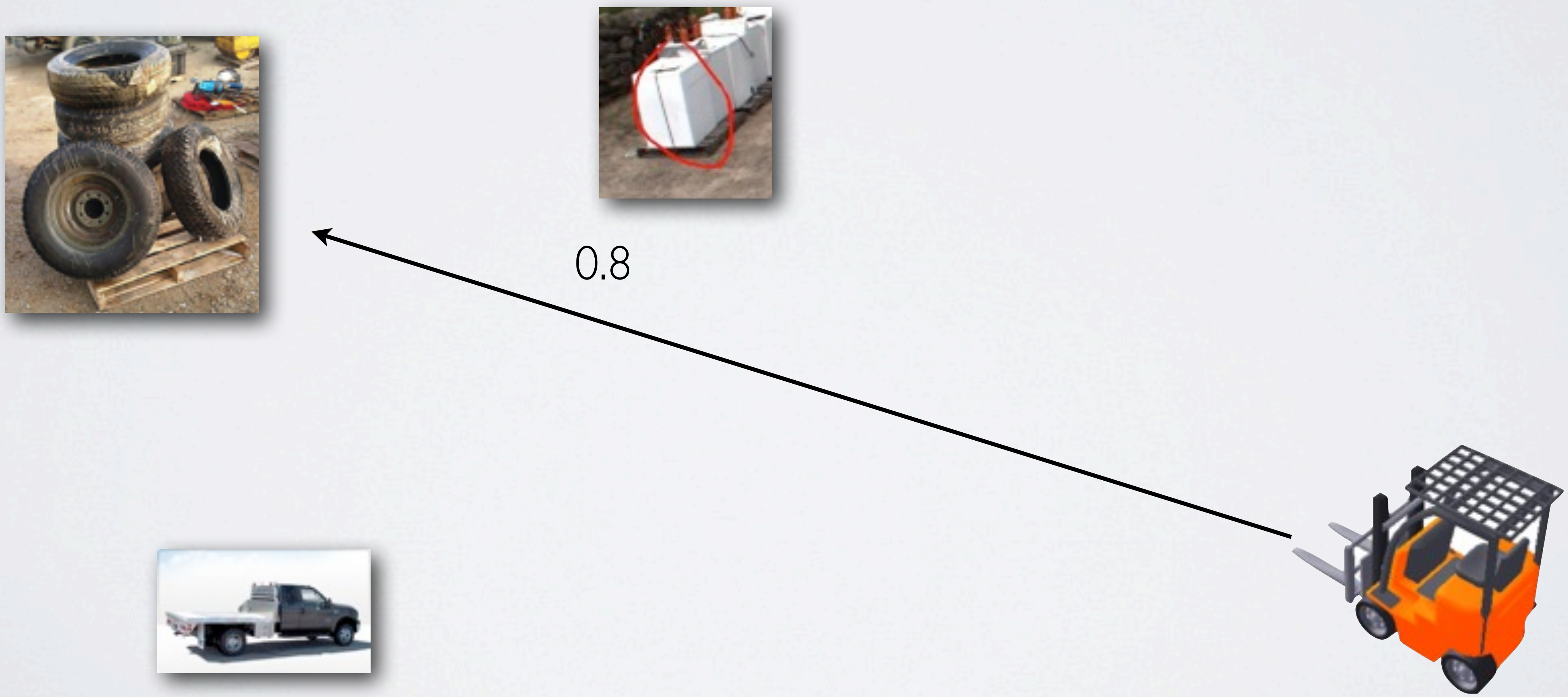


Grounding Natural Language Speech

(collaboration with S. Tellex, T. Kollar, S. Teller, & N. Roy)

"To the tire pallet"

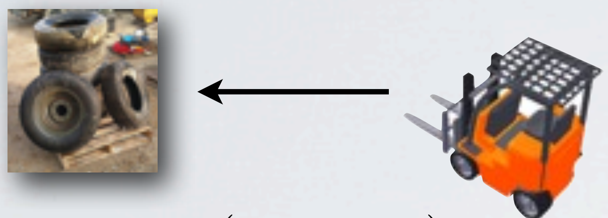
$$\arg\max_{\gamma \in \mathcal{X}} p\left(\gamma | \lambda\right)$$



Learning the Grounding Distributions

Training Set

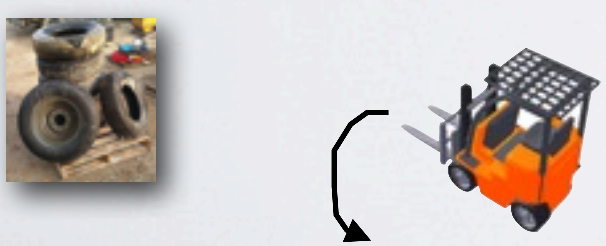
"To the tire pallet"



$$\gamma = (x_1, y_1, z_1)$$

⋮

!("To the tire pallet")

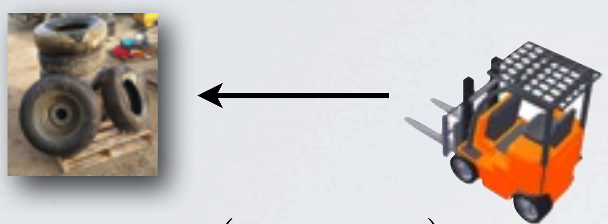


$$\gamma = (x_2, y_2, z_2)$$

Learning the Grounding Distributions

Training Set

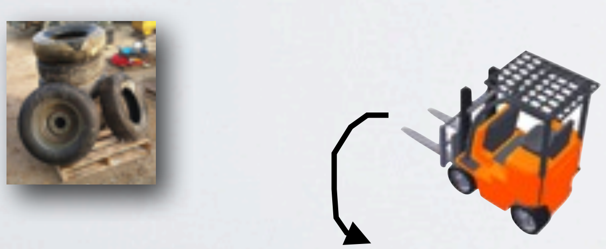
"To the tire pallet"



$\gamma = (x_1, y_1, z_1)$

$\vdots$

!("To the tire pallet")



$\gamma = (x_2, y_2, z_2)$

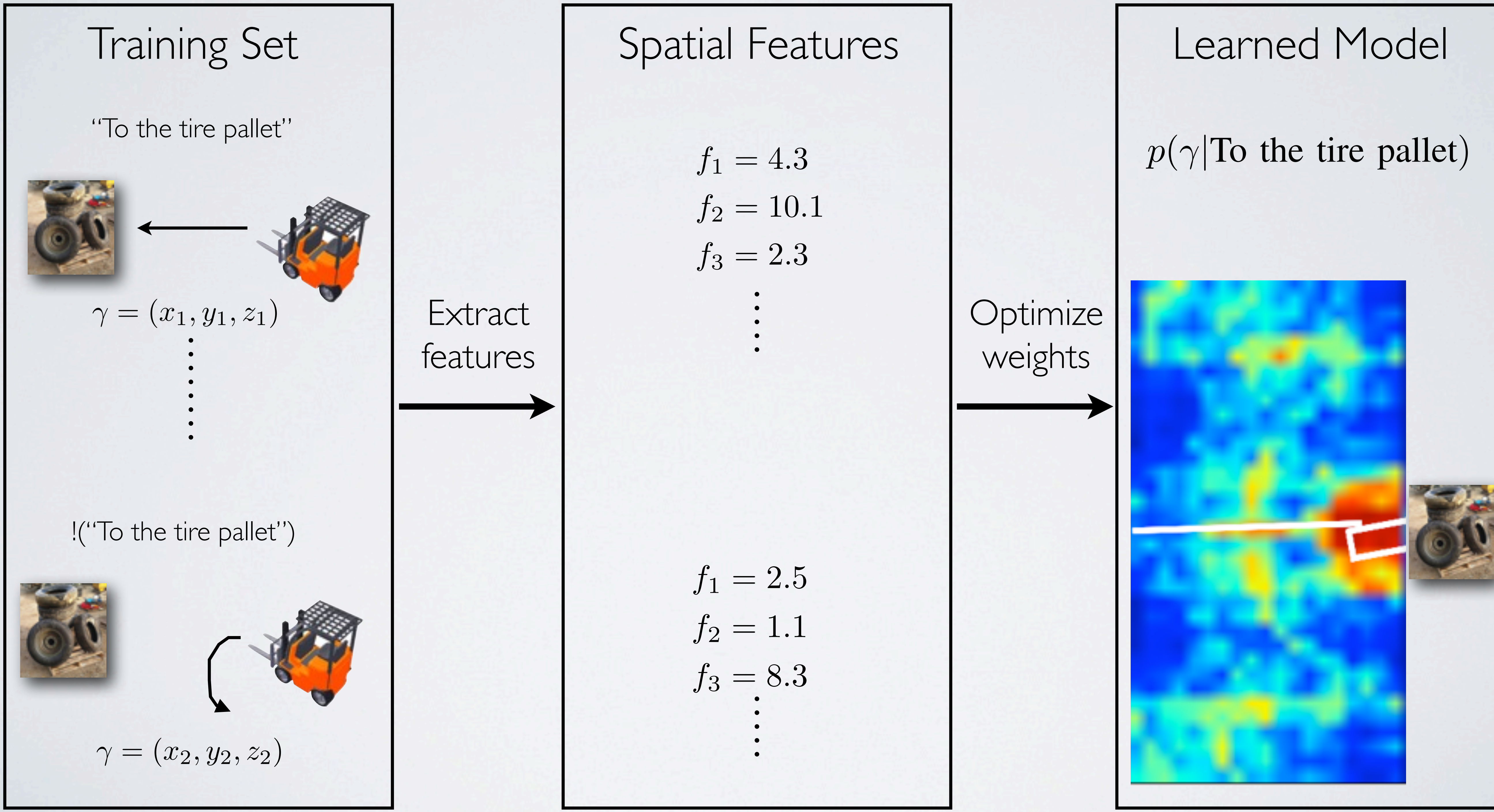
Extract features

Spatial Features

$f_1 = 4.3$
$f_2 = 10.1$
$f_3 = 2.3$
$\vdots$

$f_1 = 2.5$
$f_2 = 1.1$
$f_3 = 8.3$
$\vdots$

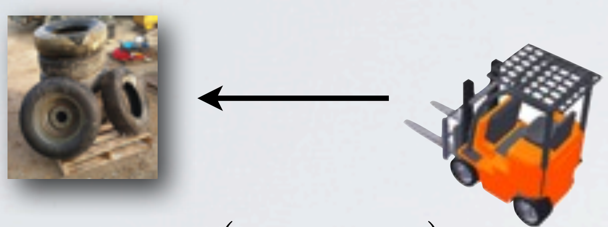
Learning the Grounding Distributions



Learning the Grounding Distributions

Training Set

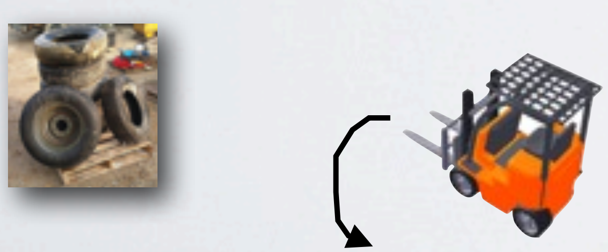
"To the tire pallet"



$\gamma = (x_1, y_1, z_1)$

$\vdots$

!("To the tire pallet")



$\gamma = (x_2, y_2, z_2)$

Extract features →

Spatial Features

$f_1 = 4.3$
$f_2 = 10.1$
$f_3 = 2.3$
$\vdots$

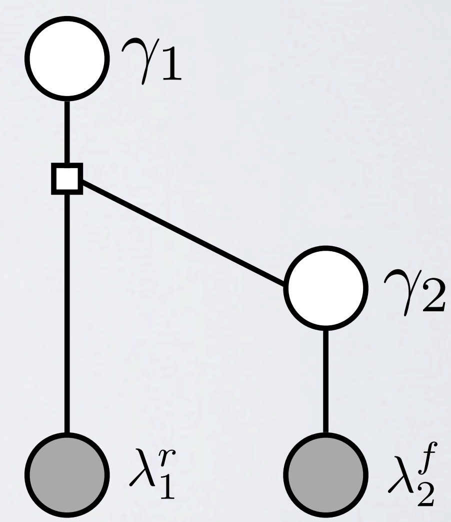
$f_1 = 2.5$
$f_2 = 1.1$
$f_3 = 8.3$
$\vdots$

Optimize weights →

Learned Model

$p(\gamma | \text{To the tire pallet})$

grounding graph



"to" "tire pallet"