Lang2LTL-2: Grounding Spatiotemporal Navigation Commands Using Large Language and Vision-Language Models

Jason Xinyu Liu

AAAI 2024 Fall Symposium UR-RAD

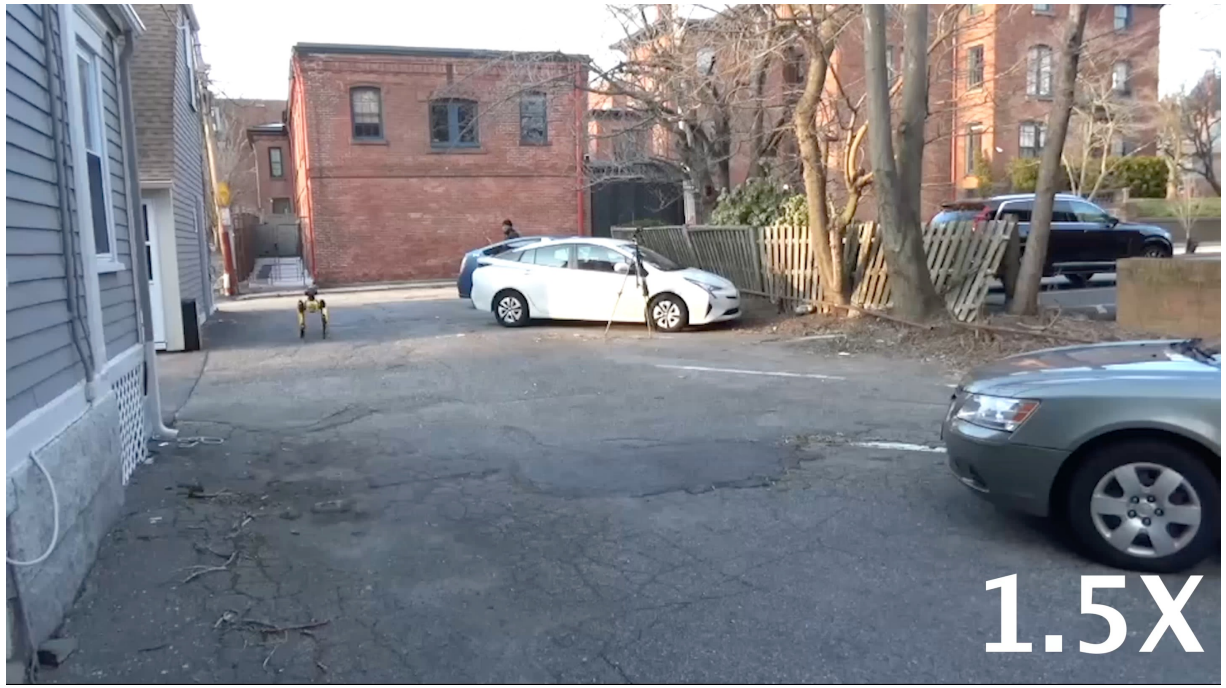
November 8st, 2024

https://jasonxyliu.github.io

https://spatiotemporal-ground.github.io

Grounding Spatiotemporal Language

go to the white car near the dumpster
exactly three times, in addition avoid stairs
in front of the apartment



walk to the chair in front of the bookshelf
but only after the kitchen counter

Grounding Language to Structure

Input

- Spatiotemporal navigation command
- Multimodal semantic map: text + images

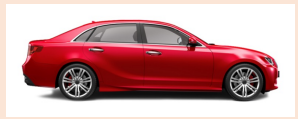
Output

- LTL formula whose propositions are grounded to real-world landmarks

Lang2LTL-2: A Modular Grounding System

Input Command:
Visit the vegan restaurant west of the bakery, only then go to the red car to the right of the bakery

Semantic Database:
plant_city: {amenity: restaurant, diet: vegan}
wildflour: {amenity, bakery}
car_0:

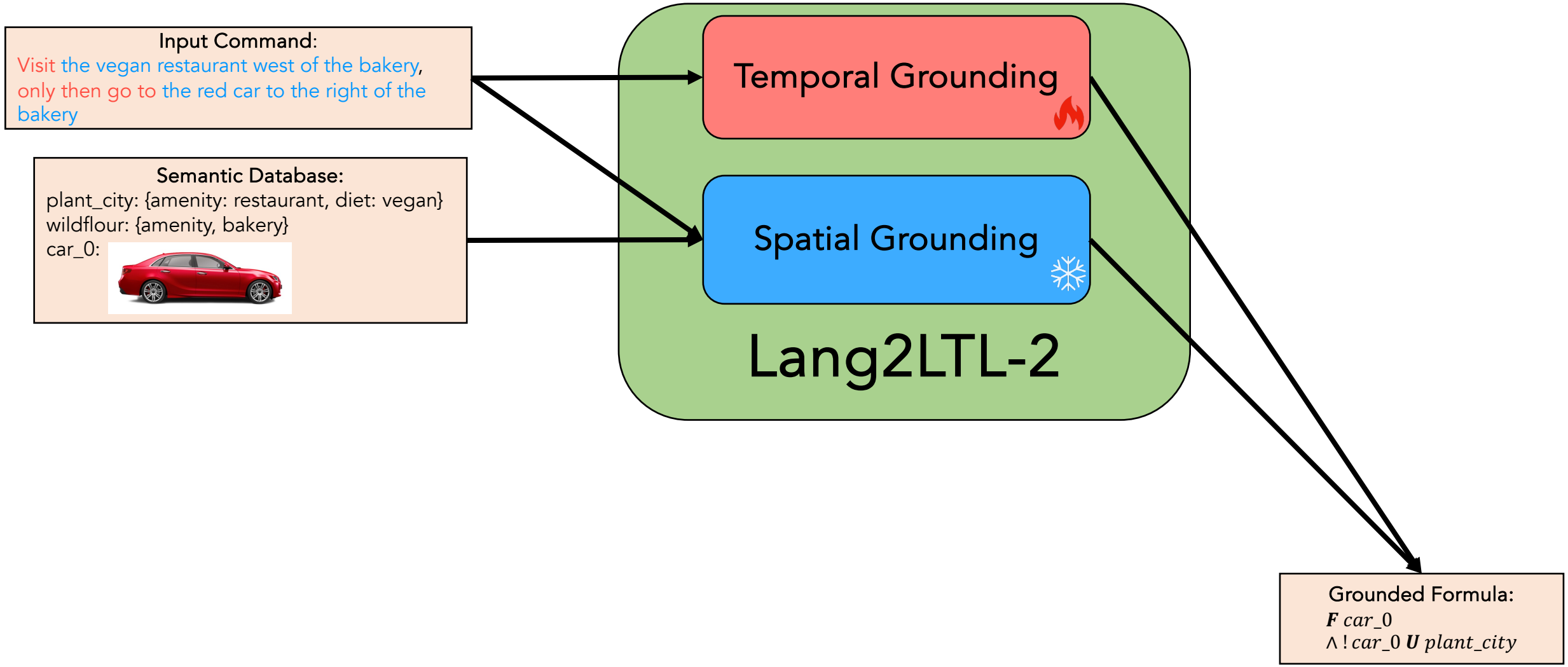


Lang2LTL-2

Grounded Formula:

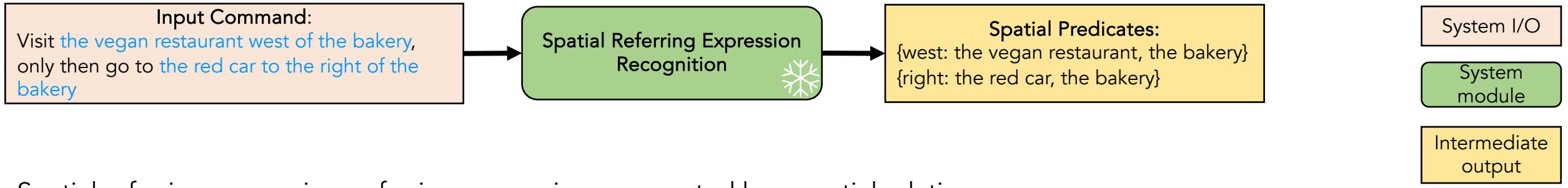
$$\boldsymbol{F}\ car\_0 \wedge !\, car\_0\ \boldsymbol{U}\ plant\_city$$

Lang2LTL-2: A Modular Grounding System



Lang2LTL-2: A Modular Grounding System

Lang2LTL-2: A Modular Grounding System

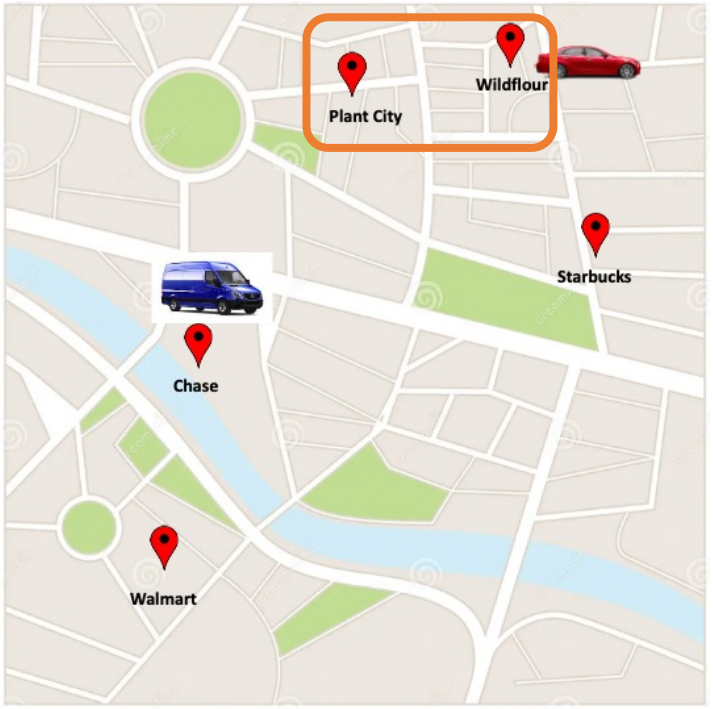


Spatial referring expression: referring expressions connected by a spatial relation

Lang2LTL-2: A Modular Grounding System

Spatial referring expression: referring expressions connected by a spatial relation

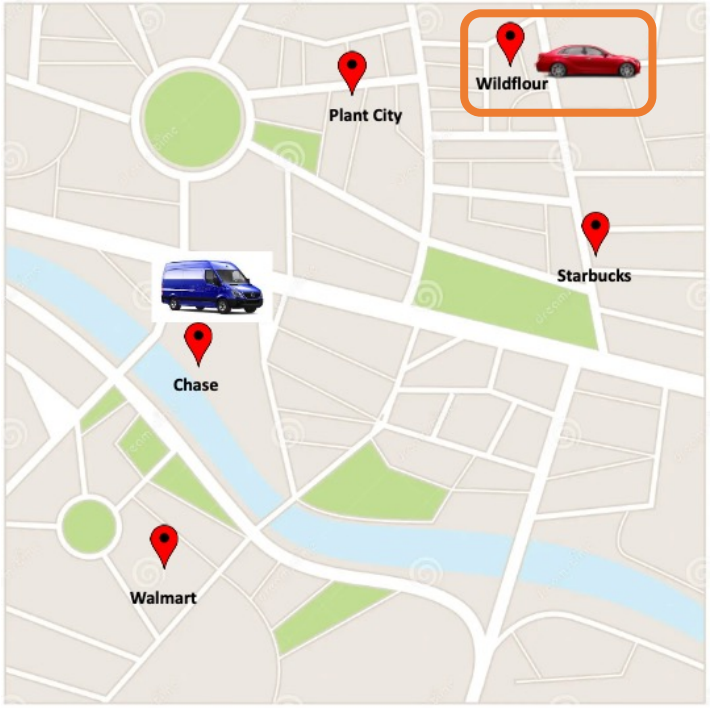
- the vegan restaurant west of the bakery



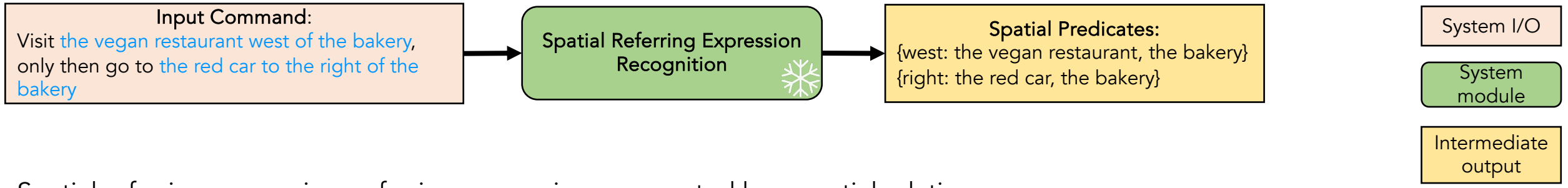
Lang2LTL-2: A Modular Grounding System

Spatial referring expression: referring expressions connected by a spatial relation

- the vegan restaurant west of the bakery
- the red car to the right of the bakery



Lang2LTL-2: A Modular Grounding System

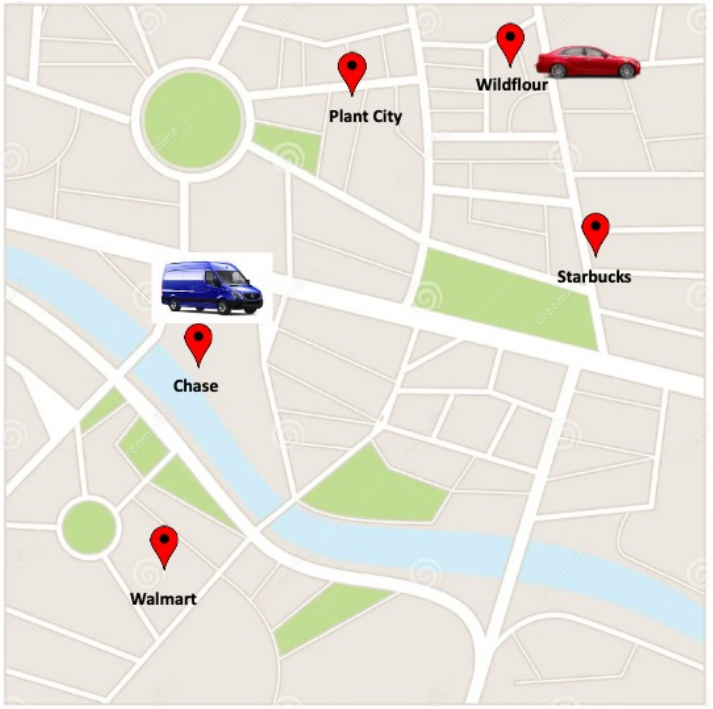


Spatial referring expression: referring expressions connected by a spatial relation

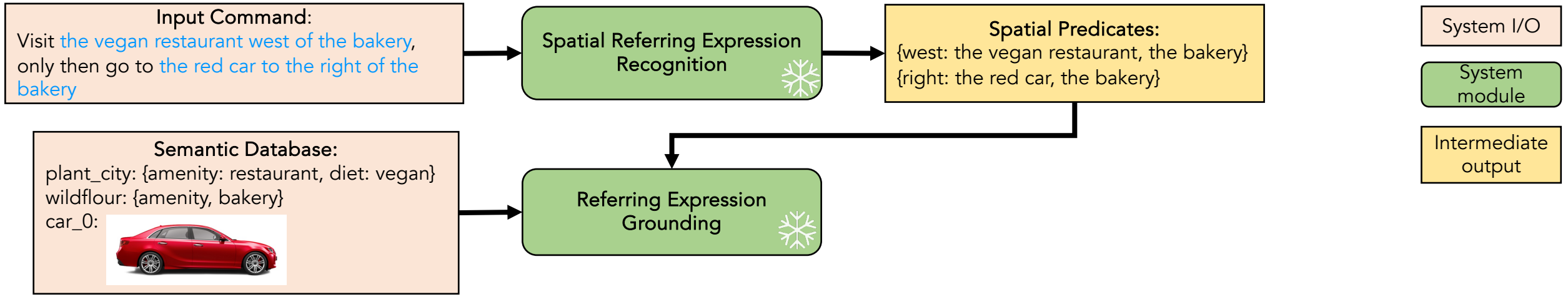
- the vegan restaurant west of the bakery
- the red car to the right of the bakery

Spatial referring expression recognition module

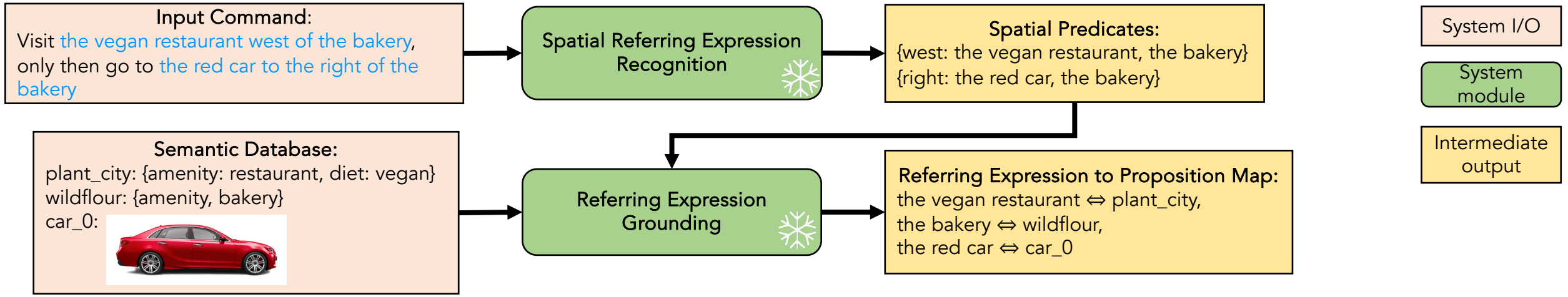
- In-context learning with GPT-4



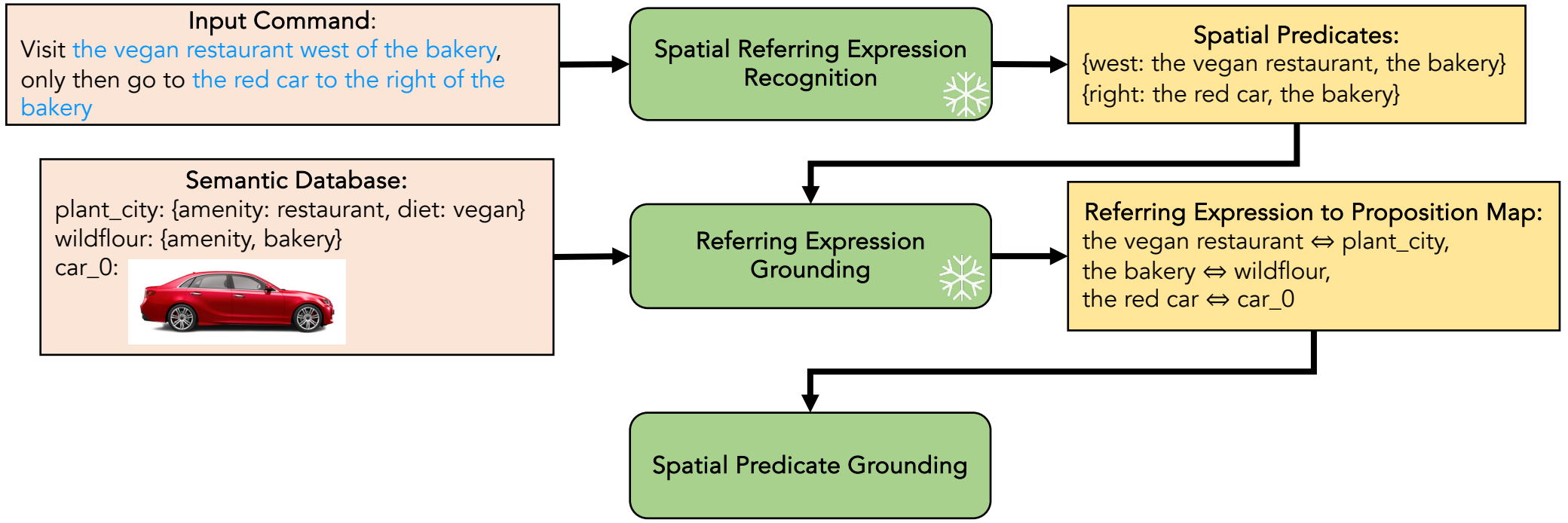
Lang2LTL-2: A Modular Grounding System



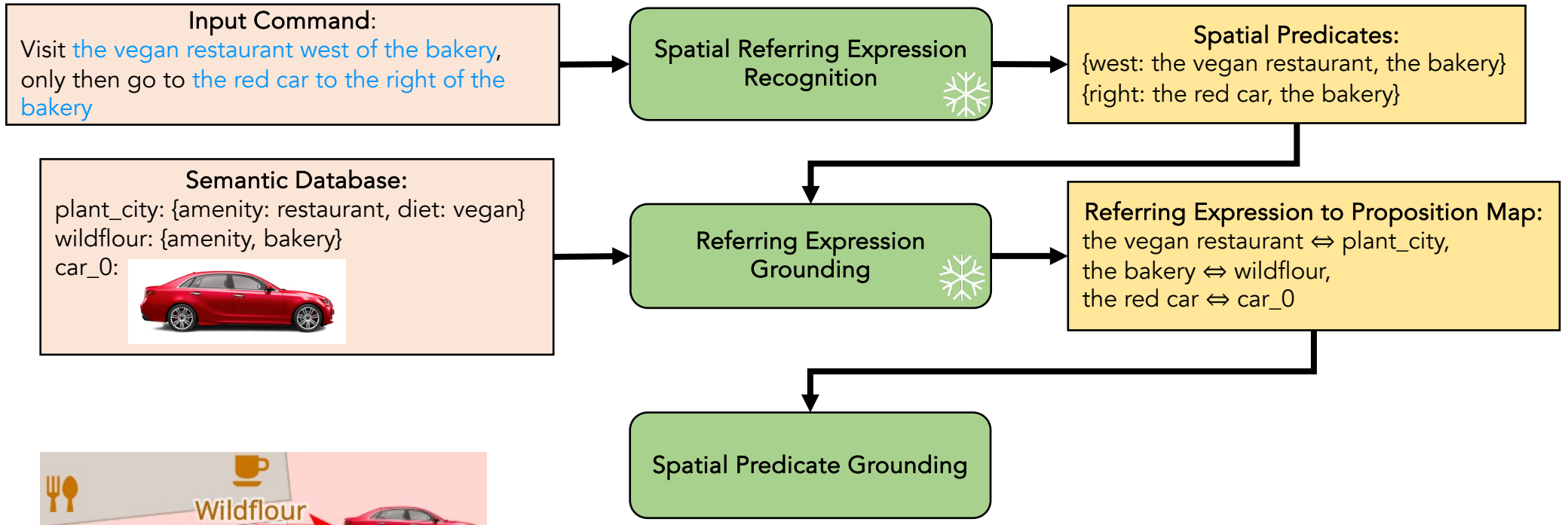
Lang2LTL-2: A Modular Grounding System

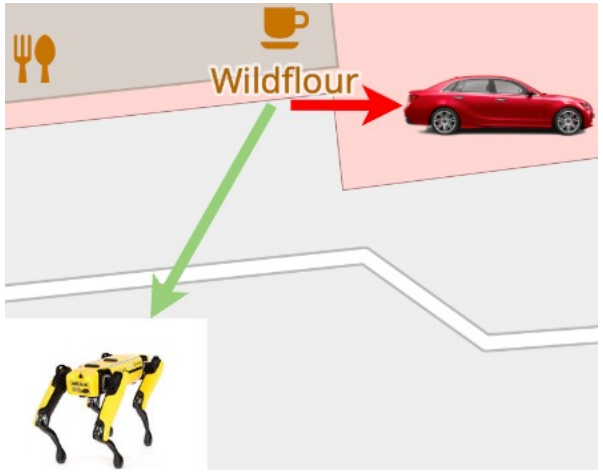


Lang2LTL-2: A Modular Grounding System



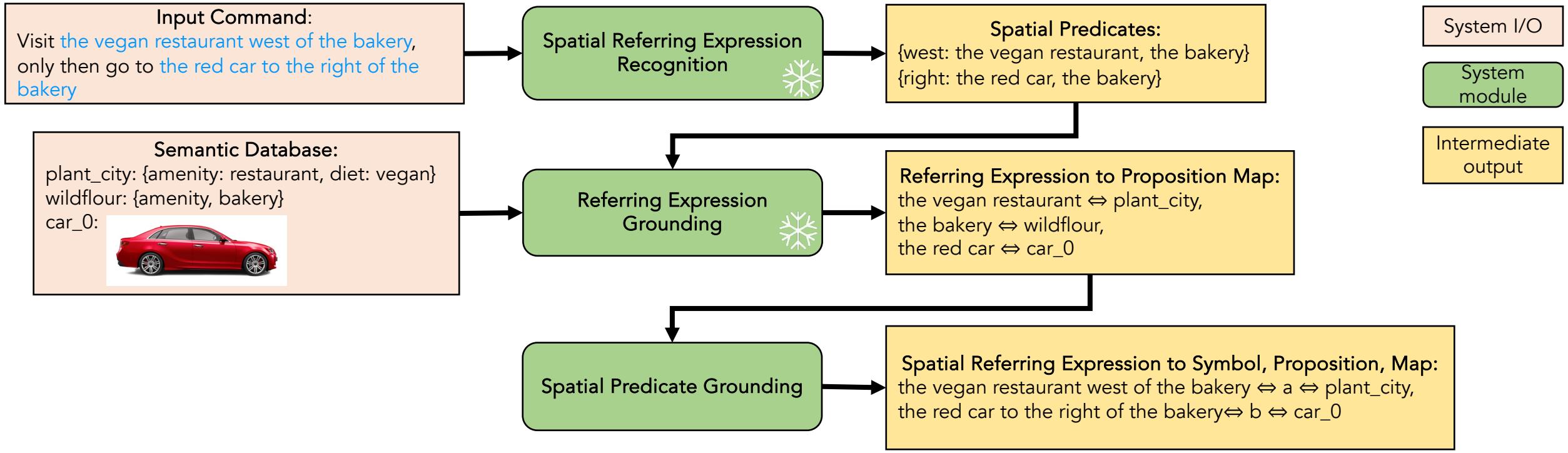
Lang2LTL-2: A Modular Grounding System



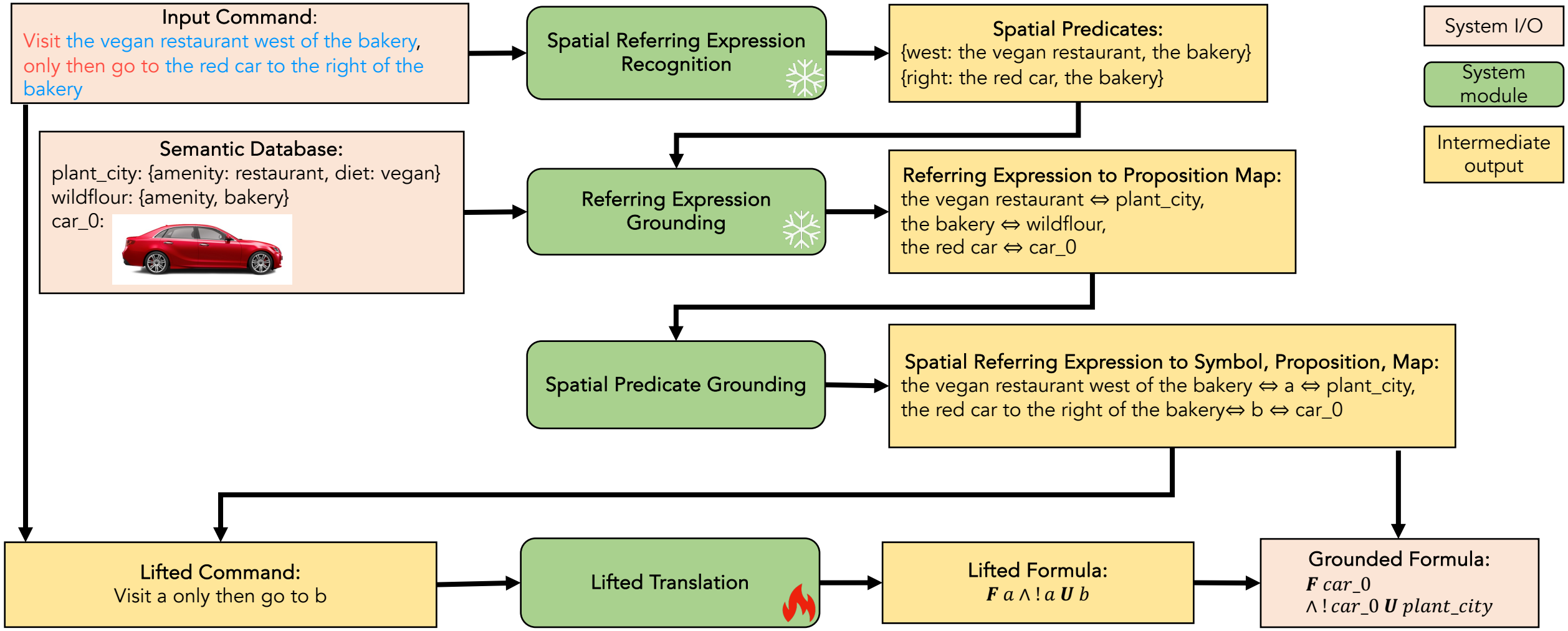


- Spatial referring expression: the red car to the right of the bakery
- Spatial predicate: {right: the red car, the bakery}
- Spatial relation: to the right of

Lang2LTL-2: A Modular Grounding System



Lang2LTL-2: A Modular Grounding System



Lang2LTL-2: Modular Evaluation

Module	Accuracy				
	City 1 (9 landmarks)	*City 2* (34 landmarks)	*City 3* (44 landmarks)	*City 4* (175 landmarks)	*Average*
SRER					
REG					
SPG					
LT					

Lang2LTL-2: Modular Evaluation

Module	Accuracy				
	City 1 (9 landmarks)	*City 2* (34 landmarks)	*City 3* (44 landmarks)	*City 4* (175 landmarks)	*Average*
SRER	$99.45 \pm 0.12\%$	$99.43 \pm 0.26\%$	$99.56 \pm 0.63\%$	$99.39 \pm 0.21\%$	$99.46 \pm 0.34\%$
REG					
SPG					
LT					

Lang2LTL-2: Modular Evaluation

Module		Accuracy				
		City 1 (9 landmarks)	*City 2* (34 landmarks)	*City 3* (44 landmarks)	*City 4* (175 landmarks)	*Average*
SRER		$99.45 \pm 0.12\%$	$99.43 \pm 0.26\%$	$99.56 \pm 0.63\%$	$99.39 \pm 0.21\%$	$99.46 \pm 0.34\%$
REG	Top-1	$99.68 \pm 0.72\%$	$97.98 \pm 1.07\%$	$88.74 \pm 2.14\%$	$78.35 \pm 1.97\%$	$91.19 \pm 8.84\%$
	Top-5	$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$	$99.56 \pm 0.24\%$	$99.15 \pm 0.34\%$	$99.68 \pm 0.41\%$
	Top-10	$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$	$99.70 \pm 0.17\%$	$99.98 \pm 0.05\%$	$99.92 \pm 0.15\%$
SPG						
LT						

Lang2LTL-2: Modular Evaluation

Module		Accuracy				
		City 1 (9 landmarks)	*City 2* (34 landmarks)	*City 3* (44 landmarks)	*City 4* (175 landmarks)	*Average*
SRER		$99.45 \pm 0.12\%$	$99.43 \pm 0.26\%$	$99.56 \pm 0.63\%$	$99.39 \pm 0.21\%$	$99.46 \pm 0.34\%$
REG	Top-1	$99.68 \pm 0.72\%$	$97.98 \pm 1.07\%$	$88.74 \pm 2.14\%$	$78.35 \pm 1.97\%$	$91.19 \pm 8.84\%$
	Top-5	$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$	$99.56 \pm 0.24\%$	$99.15 \pm 0.34\%$	$99.68 \pm 0.41\%$
	Top-10	$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$	$99.70 \pm 0.17\%$	$99.98 \pm 0.05\%$	$99.92 \pm 0.15\%$
SPG		$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$	$99.53 \pm 0.33\%$	$99.35 \pm 1.46\%$	$99.72 \pm 0.75\%$
LT						

Lang2LTL-2: Modular Evaluation

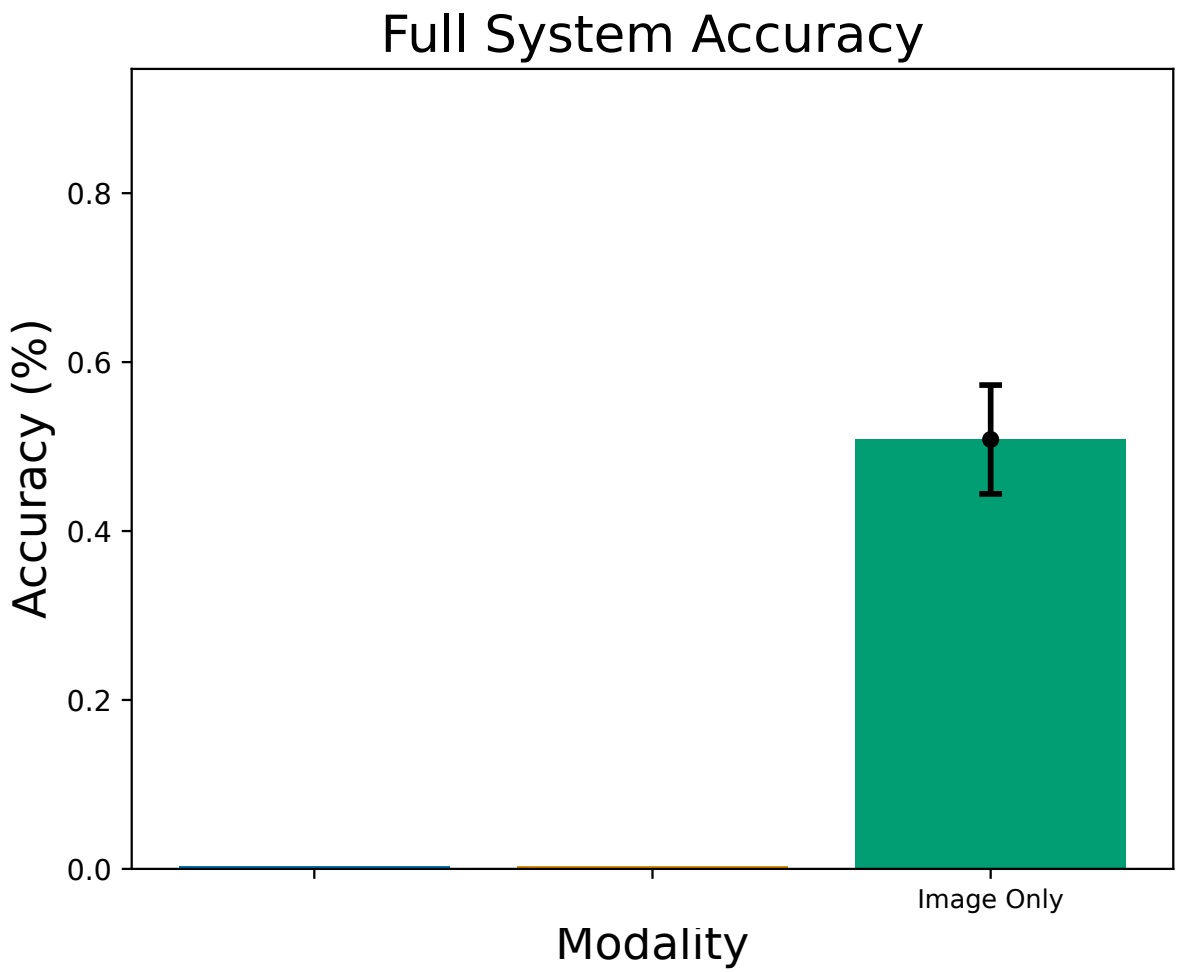
Module		Accuracy				
		City 1 (9 landmarks)	*City 2* (34 landmarks)	*City 3* (44 landmarks)	*City 4* (175 landmarks)	*Average*
SRER		$99.45 \pm 0.12\%$	$99.43 \pm 0.26\%$	$99.56 \pm 0.63\%$	$99.39 \pm 0.21\%$	$99.46 \pm 0.34\%$
REG	Top-1	$99.68 \pm 0.72\%$	$97.98 \pm 1.07\%$	$88.74 \pm 2.14\%$	$78.35 \pm 1.97\%$	$91.19 \pm 8.84\%$
	Top-5	$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$	$99.56 \pm 0.24\%$	$99.15 \pm 0.34\%$	$99.68 \pm 0.41\%$
	Top-10	$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$	$99.70 \pm 0.17\%$	$99.98 \pm 0.05\%$	$99.92 \pm 0.15\%$
SPG		$100.00 \pm 0.00\%$	$100.00 \pm 0.00\%$	$99.53 \pm 0.33\%$	$99.35 \pm 1.46\%$	$99.72 \pm 0.75\%$
LT	Finetuned T5-base	$99.45 \pm 0.00\%$	$99.45 \pm 0.00\%$	$99.45 \pm 0.00\%$	$99.45 \pm 0.00\%$	$99.45 \pm 0.00\%$
	RAG-10	$69.33 \pm 0.25\%$	$70.34 \pm 0.13\%$	$69.65 \pm 0.58\%$	$70.39 \pm 0.84\%$	$69.93 \pm 0.62\%$
	RAG-50	$83.79 \pm 0.06\%$	$83.93 \pm 0.12\%$	$83.75 \pm 0.52\%$	$83.93 \pm 0.65\%$	$83.85 \pm 0.33\%$
	RAG-100	$88.20 \pm 0.58\%$	$88.25 \pm 1.04\%$	$87.79 \pm 0.39\%$	$87.70 \pm 0.13\%$	$87.98 \pm 0.54\%$

Lang2LTL-2: Full System Evaluation

- Average success rates
- 21,780 semantically diverse commands
  - 47 temporal patterns
  - 19 spatial relations
- 4 cities
- 5 seeds

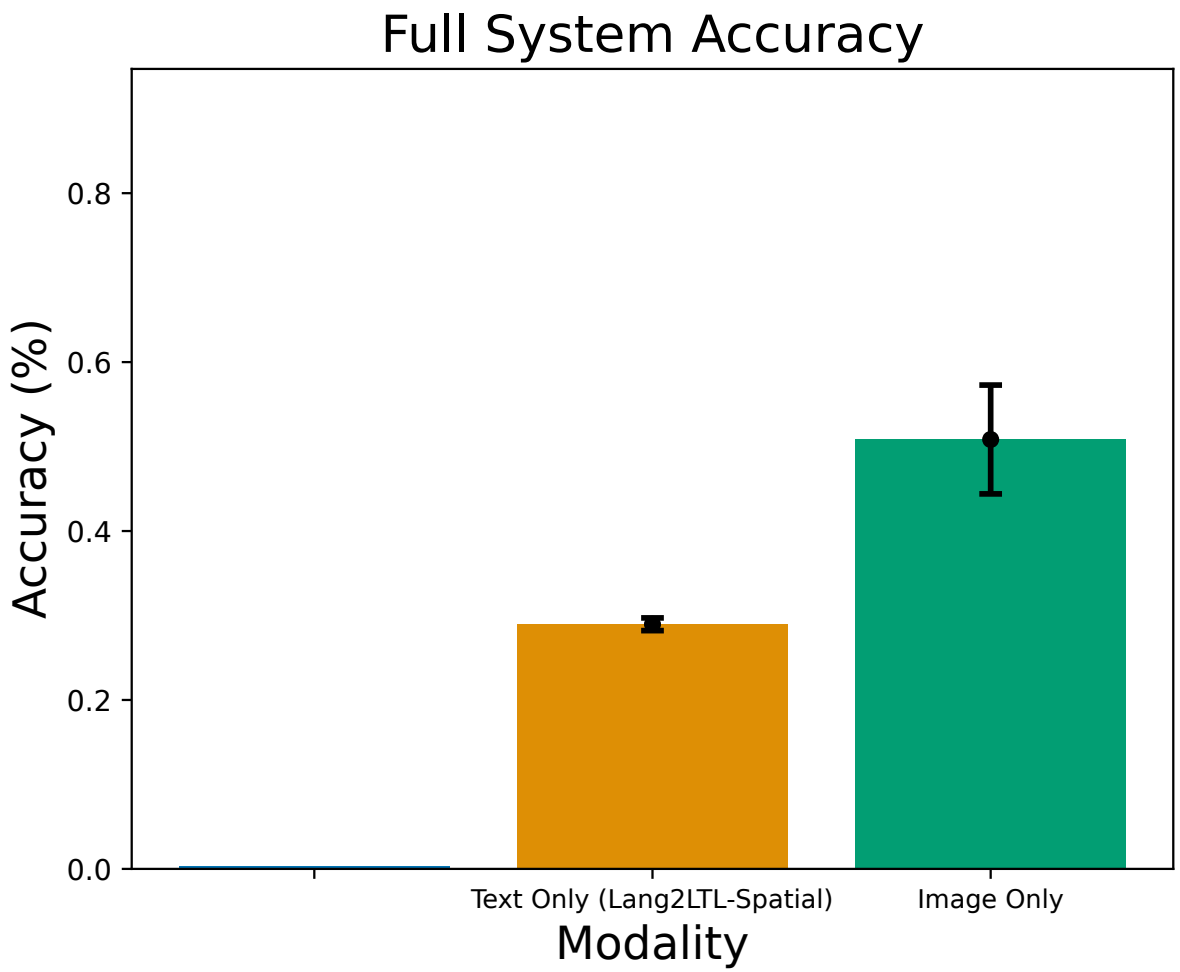
Lang2LTL-2: Full System Evaluation

- Average success rates
- 21,780 semantically diverse commands
  - 47 temporal patterns
  - 19 spatial relations
- 4 cities
- 5 seeds



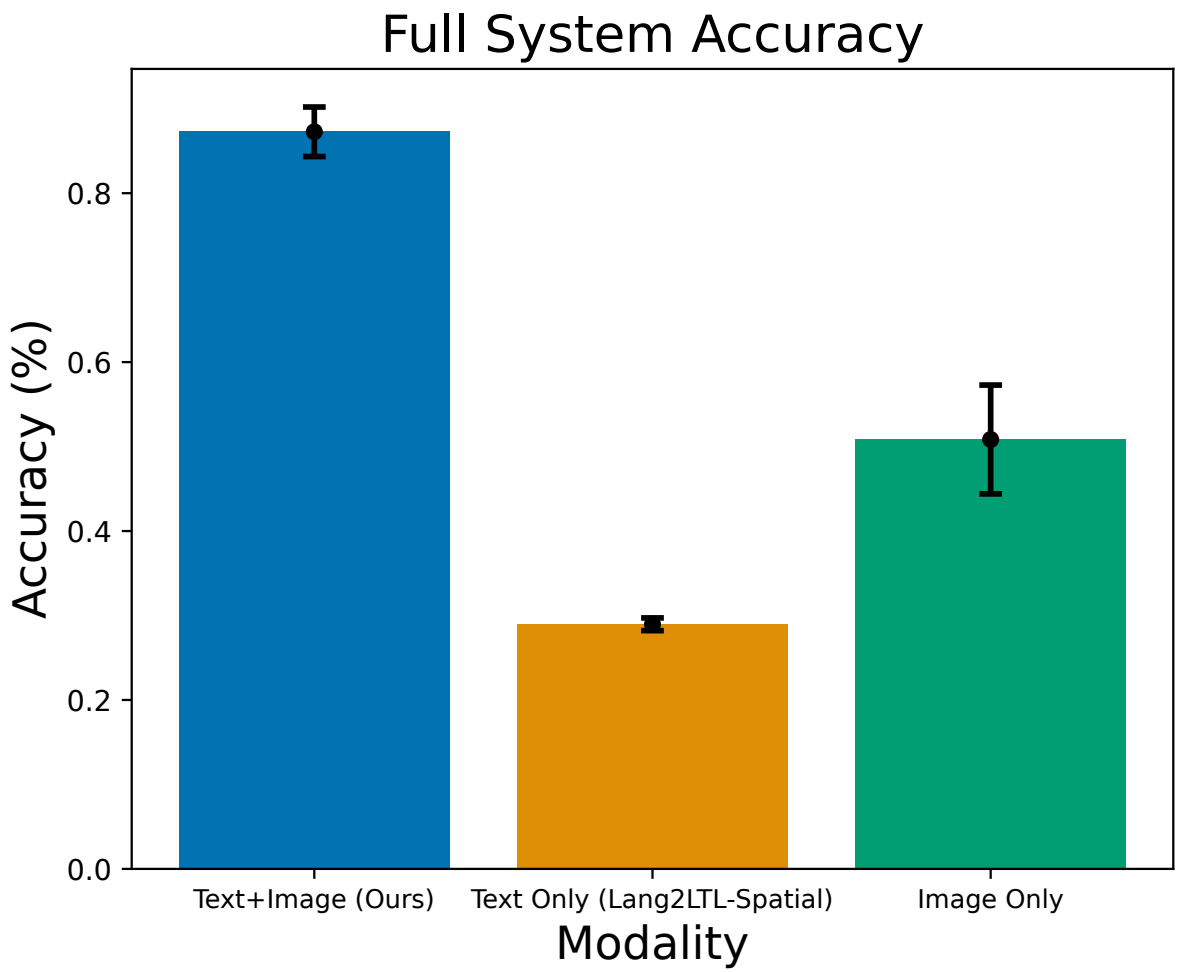
Lang2LTL-2: Full System Evaluation

- Average success rates
- 21,780 semantically diverse commands
  - 47 temporal patterns
  - 19 spatial relations
- 4 cities
- 5 seeds



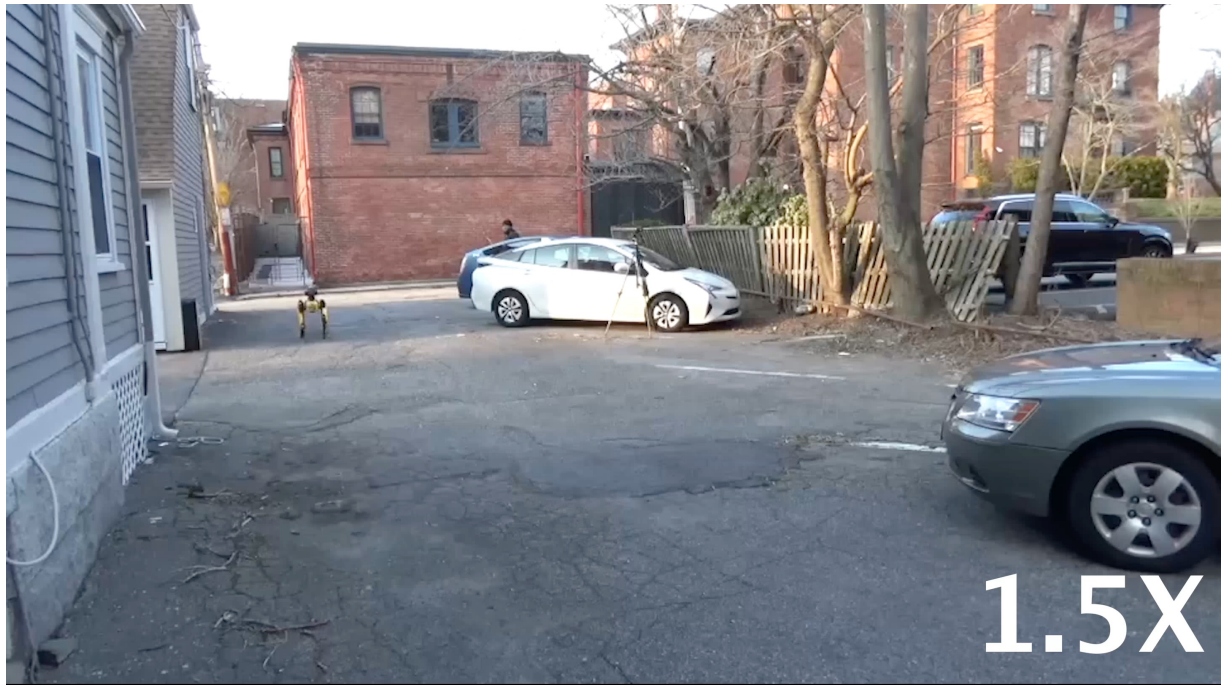
Lang2LTL-2: Full System Evaluation

- Average success rates
- 21,780 semantically diverse commands
  - 47 temporal patterns
  - 19 spatial relations
- 4 cities
- 5 seeds

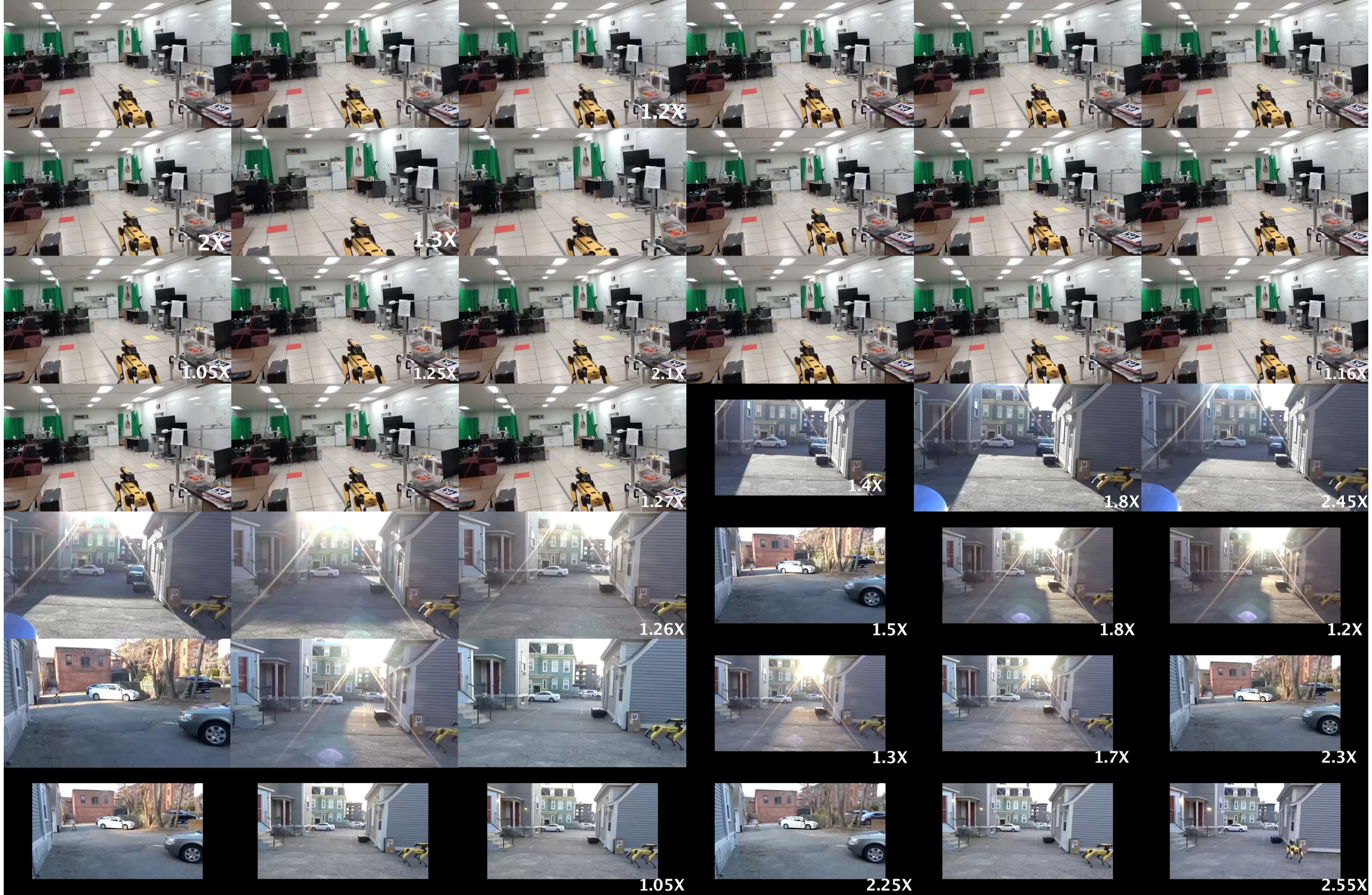


Lang2LTL-2: Robot Demonstration

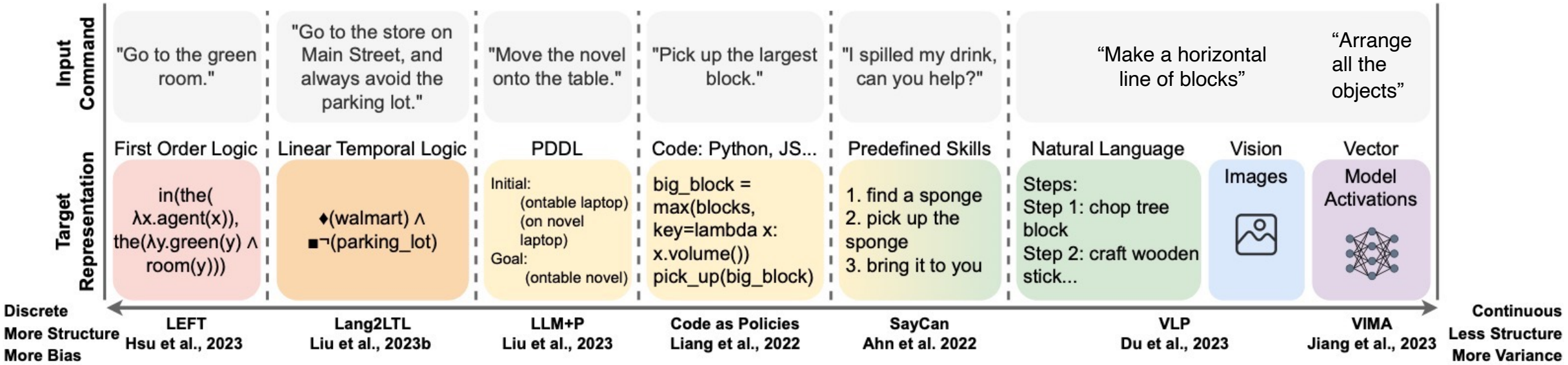
go to the white car near the dumpster
exactly three times, in addition avoid stairs
in front of the apartment



walk to the chair in front of the bookshelf
but only after the kitchen counter


1.2X
2X
1.3X
1.05X
1.25X
2.1X
1.16X
1.27X
1.4X
1.8X
2.45X
1.26X
1.5X
1.8X
1.2X
1.3X
1.7X
2.3X
1.05X
2.25X
2.55X

Robotic Language Grounding



Symbols

- Discrete
- More Structure; More bias
- Unambiguous
- Verifiable
- Interpretable

High-dimensional Embeddings

- Continuous
- Less structure; More variance
- Adaptive

Lang2LTL-2: Grounding Spatiotemporal Language

Contributions

- Grounding in novel environments without retraining on language data: 93.53% success rate
- New benchmark of 21,780 semantically diverse commands: 47 temporal + 19 spatial
- Use multimodal semantic map
- Deployed in indoor and outdoor environments

https://jasonxyliu.github.io

https://spatiotemporal-ground.github.io

A Survey of Robotic Language Grounding