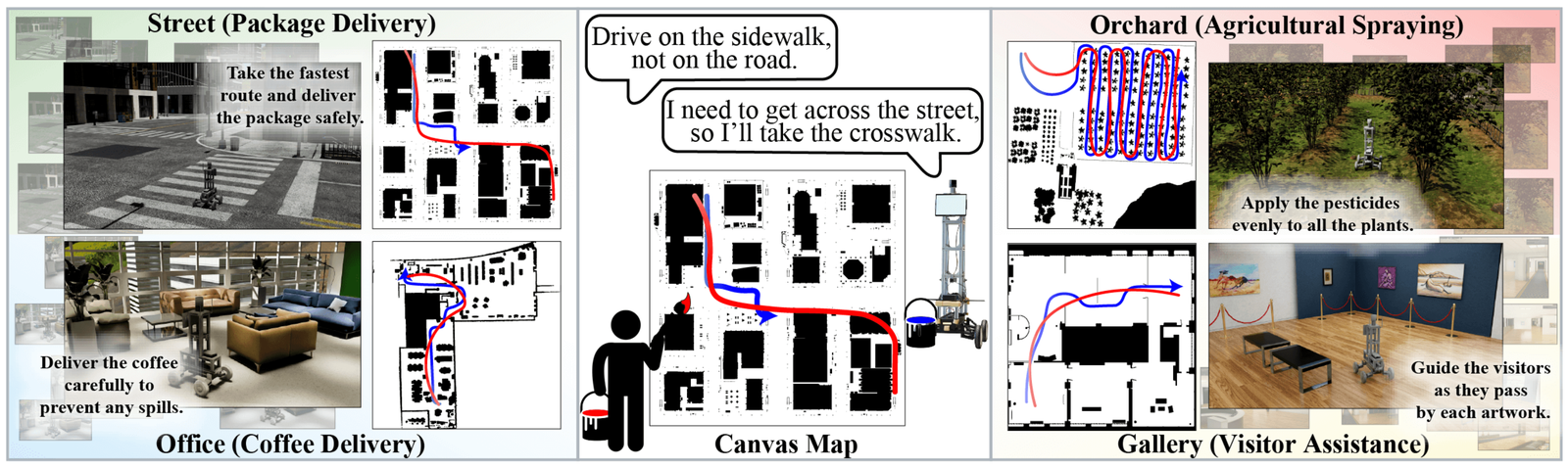
CANVAS: Commonsense-Aware Navigation System for Intuitive Human-Robot Interaction

Suhwan Choi[1], Yongjun Cho[1], Minchan Kim[1], Jaeyoon Jung[1], Myunchul Joe[1], Park Yu Been[1], Minseo Kim[2], Sungwoong Kim[2], Sungjae Lee[2], Whiseong Park[1], Jiwan Chung[2], Youngjae Yu[2]

[1]MAUM AI Inc, [2]Yonsei University

Humans often navigate using abstract yet simple instructions. How can we enable robots to do the same?



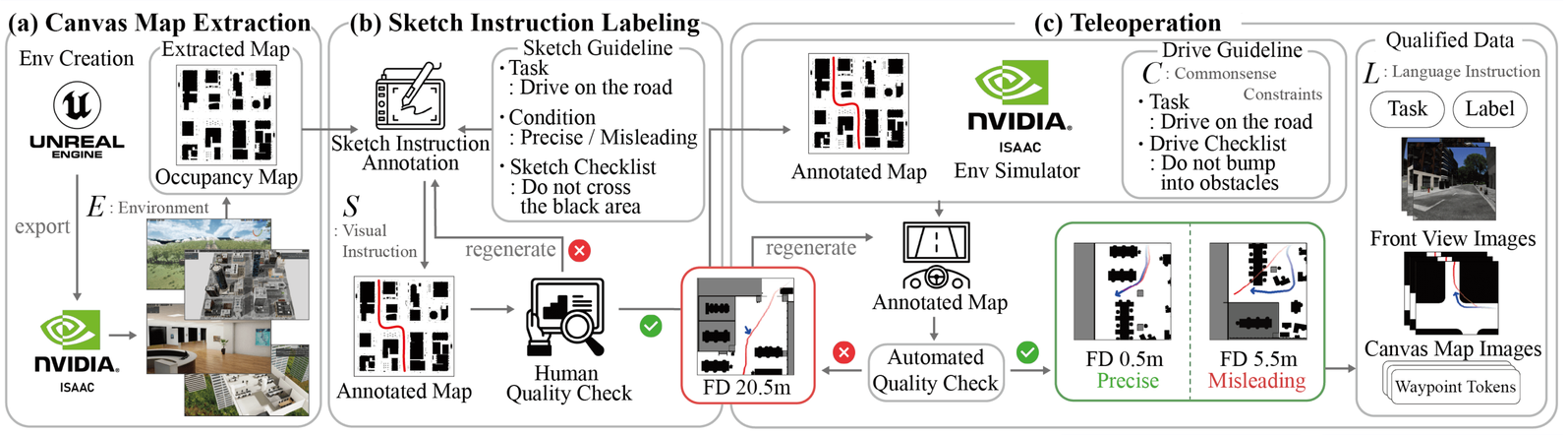
🖼️ **Image** ❌
Sensitive to environmental changes

💬 **Language** ❌
Limited for detailed navigation

🖌️ **Sketch** ✅
Robust & contains enough detail

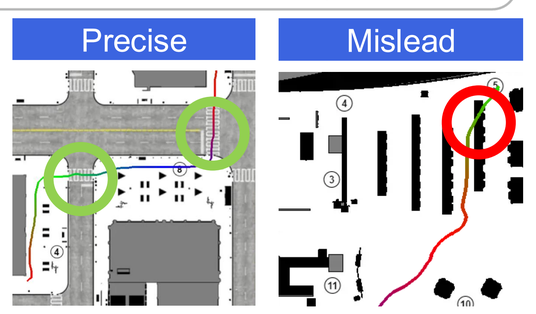
But how can we handle with **noisy** and **imprecise** instructions? 😈

COMMAND dataset

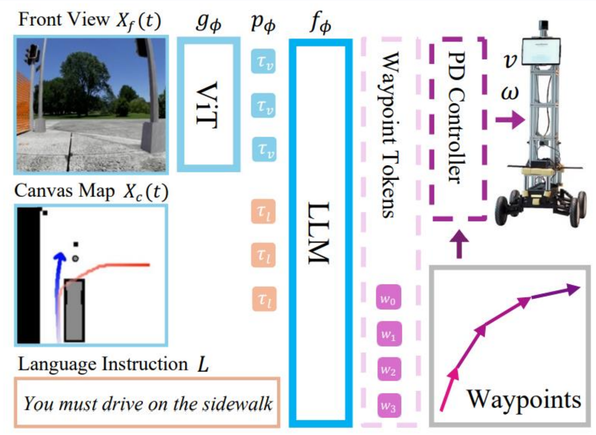


COMMAND dataset enables robot to use commonsense understanding to transform abstract or noisy human instructions into optimal trajectories.

- All Annotated map **collected by human**, spanning over **48 hours** and **219 km**
  - Data worker draw sketch trajectories and teleoperate the robot.
- **Precise trajectory** is the most efficient route while **misleading trajectory** includes noises like as passing through a wall or obstacle.
- Use **Fréchet distance** to check records automatically.
  - Fréchet distance is distance between curves used for inspection and labeling.

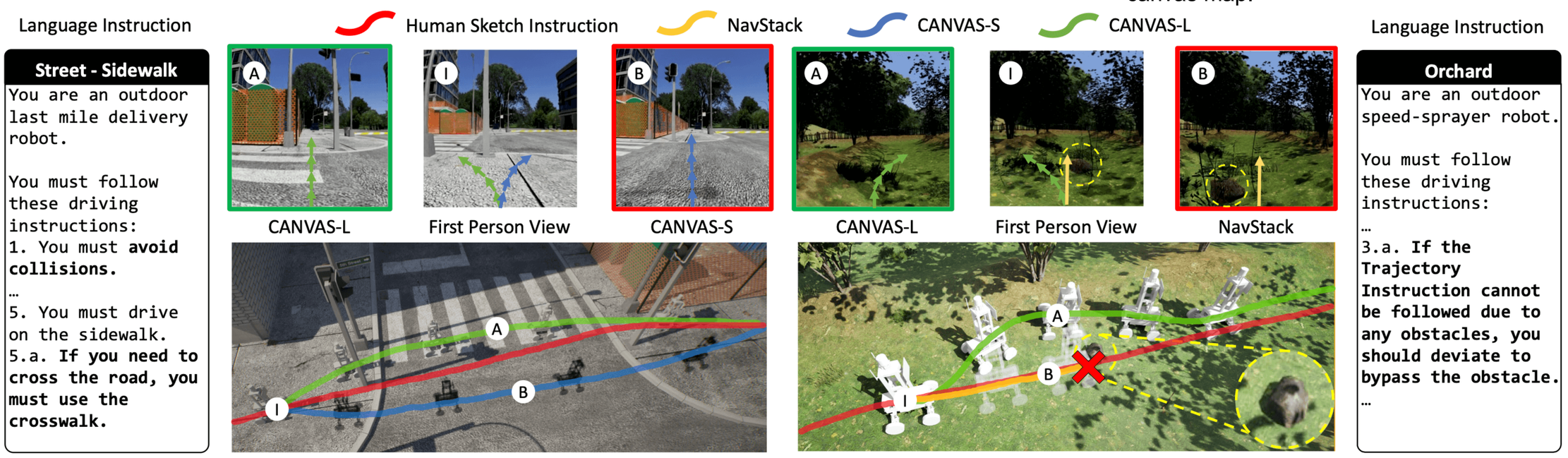


CANVAS model



CANVAS is built upon VLM architecture to achieve **common-sense aware**.

- **Language Instructions** are verbal commands which help agent's understanding.
- **Canvas Map** is an occupancy map extracted from simulated environments.
- **Sketch Instruction** is trajectory drawn on canvas map.



Results on simulated environments

Method	Precise SR(↑)	Precise CR(↓)	Precise TDD(↓)	Misleading SR(↑)	Misleading CR(↓)	Misleading TDD(↓)	Total SR(↑)
Seen Environment							
Office							
NavStack	87%	13%	0.846m	0%*	100%*	-	-
CANVAS-S	**100%**	**0%**	**0.730m**	87%	13%	0.843m	93%
CANVAS-L	**100%**	**0%**	0.802m	**100%**	**0%**	**0.753m**	**100%**
Street (Road)							
NavStack	**100%**	**0%**	1.654m	0%*	100%*	-	-
CANVAS-S	**100%**	**0%**	1.189m	**100%**	**0%**	**1.075m**	**100%**
CANVAS-L	97%	3%	**1.117m**	97%	3%	1.236m	97%
Street (Sidewalk)							
NavStack	53%	53%	1.450m	0%*	100%*	-	-
CANVAS-S	60%	40%	1.451m	47%	53%	2.379m	54%
CANVAS-L	**87%**	**13%**	**1.394m**	**53%**	**47%**	**1.839m**	**70%**
Orchard							
NavStack	0%	87%	-	0%*	100%*	-	-
CANVAS-S	**73%**	60%	**1.561m**	**60%**	**33%**	1.448m	**67%**
CANVAS-L	67%	**47%**	1.759m	**60%**	53%	**1.392m**	64%
Unseen Environment							
Gallery							
NavStack	**100%**	**0%**	0.783m	0%*	100%*	-	-
CANVAS-S	87%	13%	**0.773m**	**33%**	**66%**	0.938m	60%
CANVAS-L	**100%**	7%	0.9m	**33%**	**66%**	**0.856m**	**67%**

- Outperforms NavStack in environment with many obstacles
  - In orchard, **ROS NavStack** records **0%**, **CANVAS** achieves **69%**
- Enable to navigate **unseen environments**
- Effectively handles **misleading scenarios!**

Violation rates for commonsense-constraints

Environment	Method	Precise IVR(↓)	Misleading IVR(↓)
Street (Road)	NavStack	7%	100%*
	CANVAS-S	**0%**	**7%**
	CANVAS-L	17%	30%
Street (Sidewalk)	NavStack	7%	100%*
	CANVAS-S	**0%**	26%
	CANVAS-L	**0%**	**13%**

Instruction Violation Type
- Right hand traffic (Road)
- Drive on sidewalk and crosswalk
- …

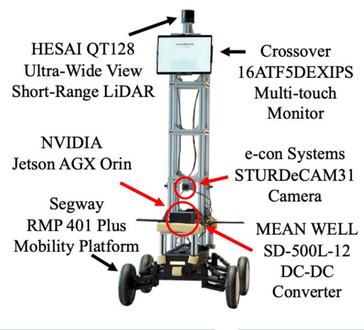
CANVAS know commonsense rules!

$$\text{IVR} = \frac{N_{Instruction\ Violation}}{N}$$

Results on real environments (Office)

Method	Precise SR(↑)	Misleading SR(↑)	Total SR(↑)
NavStack	**100%**	0%*	-
CANVAS-S	77%	**60%**	**69%**
CANVAS-L	93%	33%	63%

- Trained solely on **simulated data**
- Strong **Sim2Real** transfer capabilities
- SLAM with **FAST-LIO2** was used to find the robot's current position



Ablation study on the effect of VLM pre-training

Environment	Method	Precise SR(↑)	Misleading SR(↑)	Total SR(↑)
Seen - Office	CANVAS-L	**100%**	**100%**	**100%**
	w/o Pre-training	**100%**	87%	93%
Unseen - Gallery	CANVAS-L	**100%**	33%	**67%**
	w/o Pre-training	60%	**40%**	50%
Real - Office	CANVAS-L	**93%**	**33%**	**63%**
	w/o Pre-training	73%	**33%**	53%

Project Page