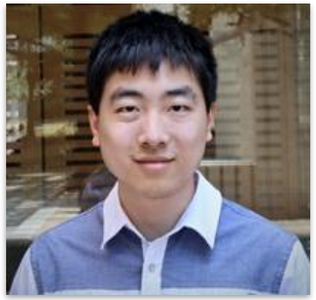
Know Where You're Uncertain When Planning with Multimodal Foundation Models: A Formal Framework

Neel P. Bhatt*

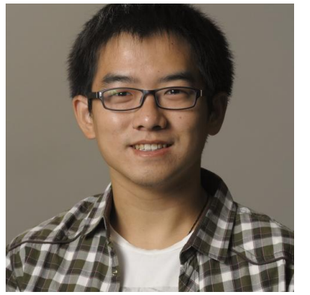


Yunhao Yang*

Rohan Siva

Daniel Milan

Ufuk Topcu



Zhangyang (Atlas) Wang

Project Page

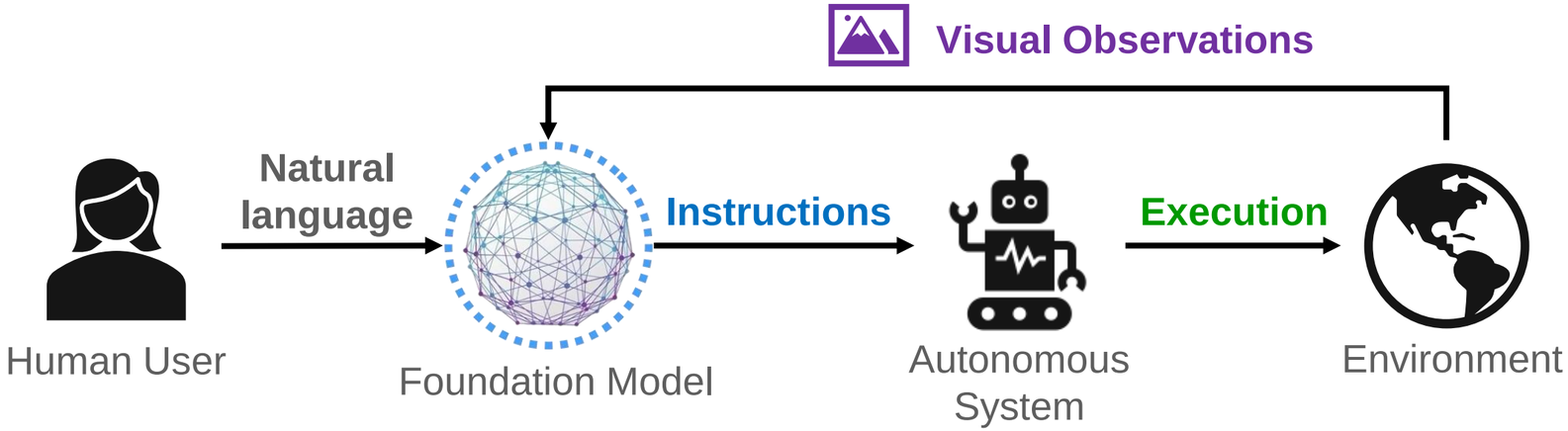
TEXAS
The University of Texas at Austin

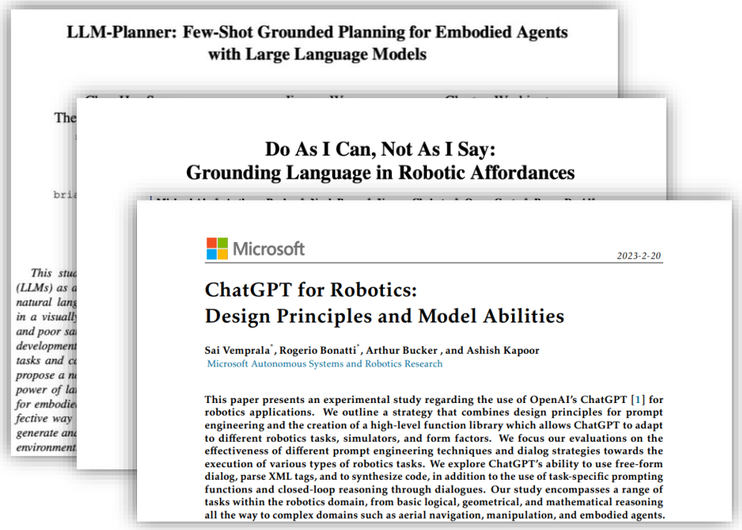
VITA

CENTER FOR aUTonomy

Paper

Multimodal Foundation Models for Plan Generation




LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models

Do As I Can, Not As I Say: Grounding Language in Robotic Affordances

Microsoft 2023-2-20

ChatGPT for Robotics: Design Principles and Model Abilities

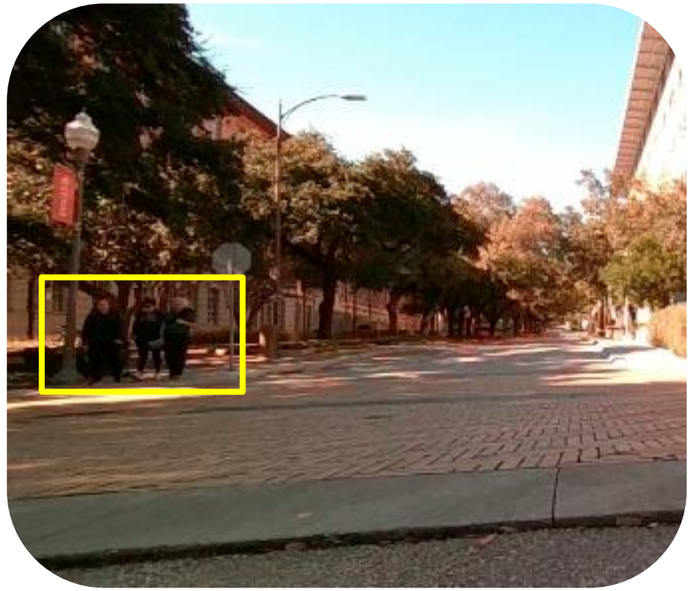
Sai Vemprala*, Rogerio Bonatti*, Arthur Bucker , and Ashish Kapoor
Microsoft Autonomous Systems and Robotics Research

This paper presents an experimental study regarding the use of OpenAI's ChatGPT [1] for robotics applications. We outline a strategy that combines design principles for prompt engineering and the creation of a high-level function library which allows ChatGPT to adapt to different robotics tasks, simulators, and form factors. We focus our evaluations on the effectiveness of different prompt engineering techniques and dialog strategies towards the execution of various types of robotics tasks. We explore ChatGPT's ability to use free-form dialog, parse XML tags, and to synthesize code, in addition to the use of task-specific prompting functions and closed-loop reasoning through dialogues. Our study encompasses a range of tasks within the robotics domain, from basic logical, geometrical, and mathematical reasoning all the way to complex domains such as aerial navigation, manipulation, and embodied agents.

Multimodal foundation models offer a **natural interface** for robotic perception and planning by processing **sensory inputs** and **natural language** to generate actionable plans.

Addressing uncertainty in both perception and decision-making remains a critical challenge for ensuring task reliability.

Where Does The Uncertainty Come From?



Turn left at the stop sign

1. Stop
2. Wait for pedestrians
3. Turn left

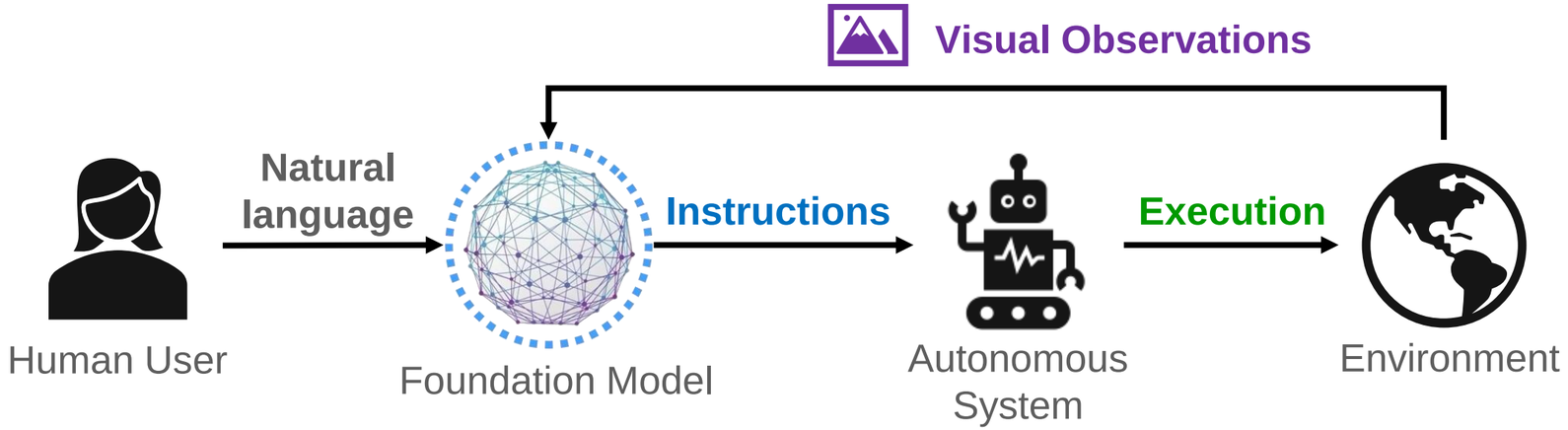
Perception (Visual Obs.)

- Image Artifacts
- Lighting
- Occlusion
- ……

Decision (Text Generation)

- Prompt-image inconsistency
- Failure to capture critical image information
- Failure to incorporate safety rules
- ……

Multimodal Foundation Models for Plan Generation



The Central Question

How can we design an automated and reliable framework that enables **uncertainty quantification** and **targeted interventions** for robust perception and planning using multimodal foundation models?

Limitations of Existing Works

1) Provide an aggregate "black block" estimate of uncertainty, lacking insight into whether uncertainty originates from perception or decision-making flaws.

2) Obscure root cause of performance issues which hinders targeted improvements and leads to ↑ queries and ↓ performance.

3) Require human-labelling for calibration (not scalable).

Existing Works

Human

Place the bowl in the microwave, please.

Robot

Which one, plastic or metal?

Human

The plastic one, please.



Environment Context

There is a microwave, a landfill bin, a recycling bin, and a compost bin.

Robot Observations

Observations: I see a metal bowl and a plastic bowl on the counter.

LLM Next Step Prediction with Confidence

Possible next steps:
0.44 - Put plastic bowl in microwave.
0.41 - Put metal bowl in microwave.
0.03 - Put metal bowl in landfill bin
0.08 - Put plastic bowl in recycling bin.

Prediction Set from Conformal Prediction

Conformal prediction threshold: 0.21
Steps with scores above threshold:
0.44 - Put plastic bowl in microwave.
0.41 - Put metal bowl in microwave.

Trigger Human Help

Prediction size 2 > 1 → ask for help.

LLM Generates Question

Question: Which one, plastic or metal?

Robots that ask for help: Uncertainty alignment for large language model planners, CoRL 2023

Contributions

- We present a novel framework to **disentangle and quantify** the inherent source of uncertainty in multimodal foundation models into:
  - **Perception uncertainty** associated with the model's visual processing capabilities and
  - **Decision uncertainty** linked to its ability to generate actionable plans

- We quantify each source using **novel quantification methods** – conformal prediction and Formal-Methods-Driven Prediction (FMDP), leveraging symbolic representations and formal verification techniques for theoretical guarantees

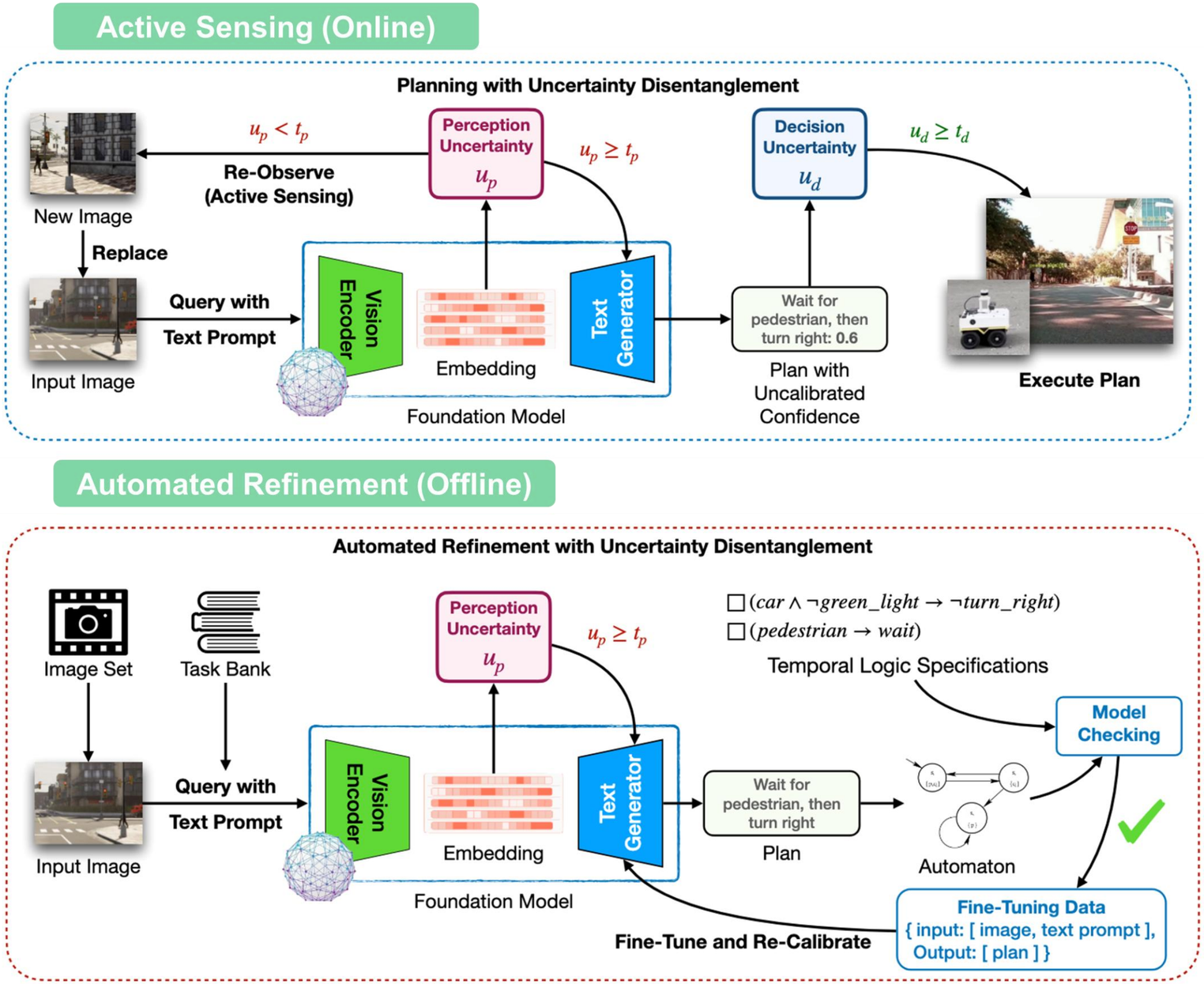
- We implement a **two-part improvement strategy** via targeted interventions: active sensing and automated model refinement.

- Empirical validation in real-world and simulated robotic tasks demonstrate that our framework **reduces variability by up to 40%** and enhances **task success rates by 5% compared to baselines**.

A Brief Outline

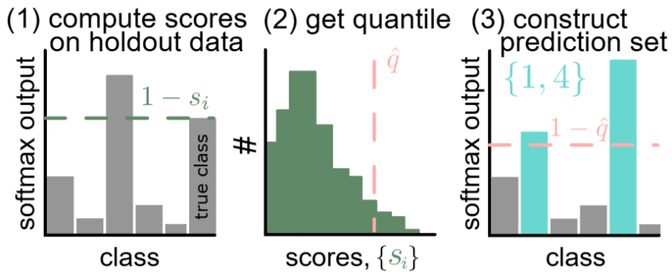
- Overview of the framework
- Perception and decision uncertainty
  - Quantifying perception uncertainty: conformal prediction
  - Decision uncertainty: formal-methods-driven prediction (FMDP)
- Targeted interventions to reduce uncertainty
  - Efficient online inference via active sensing
  - Automated fine-tuning with probabilistic guarantees
- Experimental results
- Takeaways

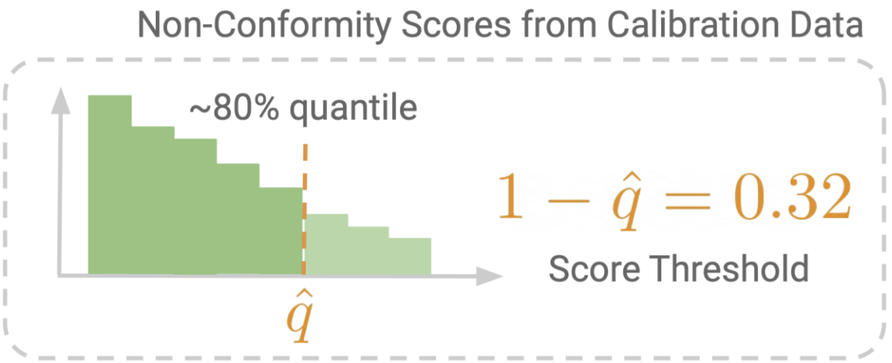
Overview of the Framework



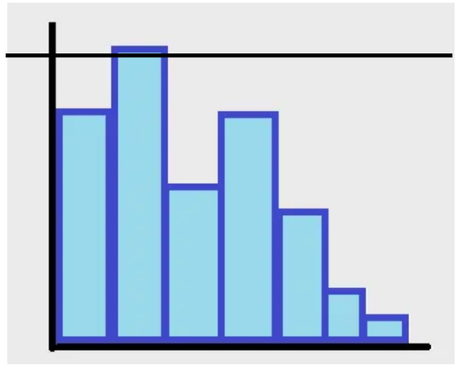
Perception Uncertainty

At least 80% probability of being correct





Confidence score threshold 0.32



At least 80% probability of being correct

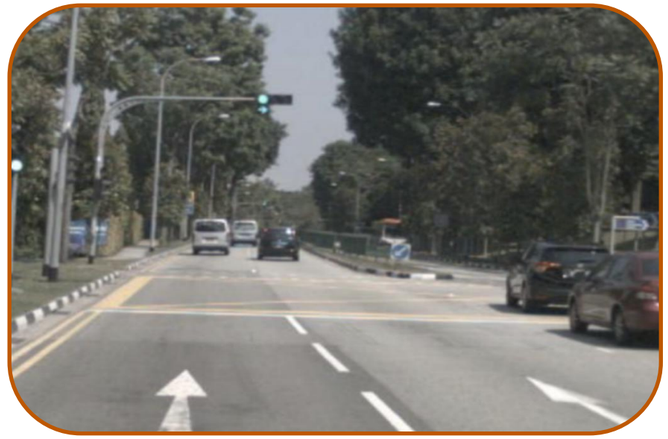
Perception Uncertainty Score

- A theoretical lower bound on the probability of correctly identifying objects in the image

Decision Uncertainty

Decision Uncertainty Score

- Given a set of specifications, expressed in temporal logic,
- A *decision uncertainty score* of a *plan* is a theoretical lower bound probability of the plan satisfying the specifications



Go straight at the traffic light

1. The traffic light is green and there are no pedestrians
2. Move forward

Specifications (in temporal logic):

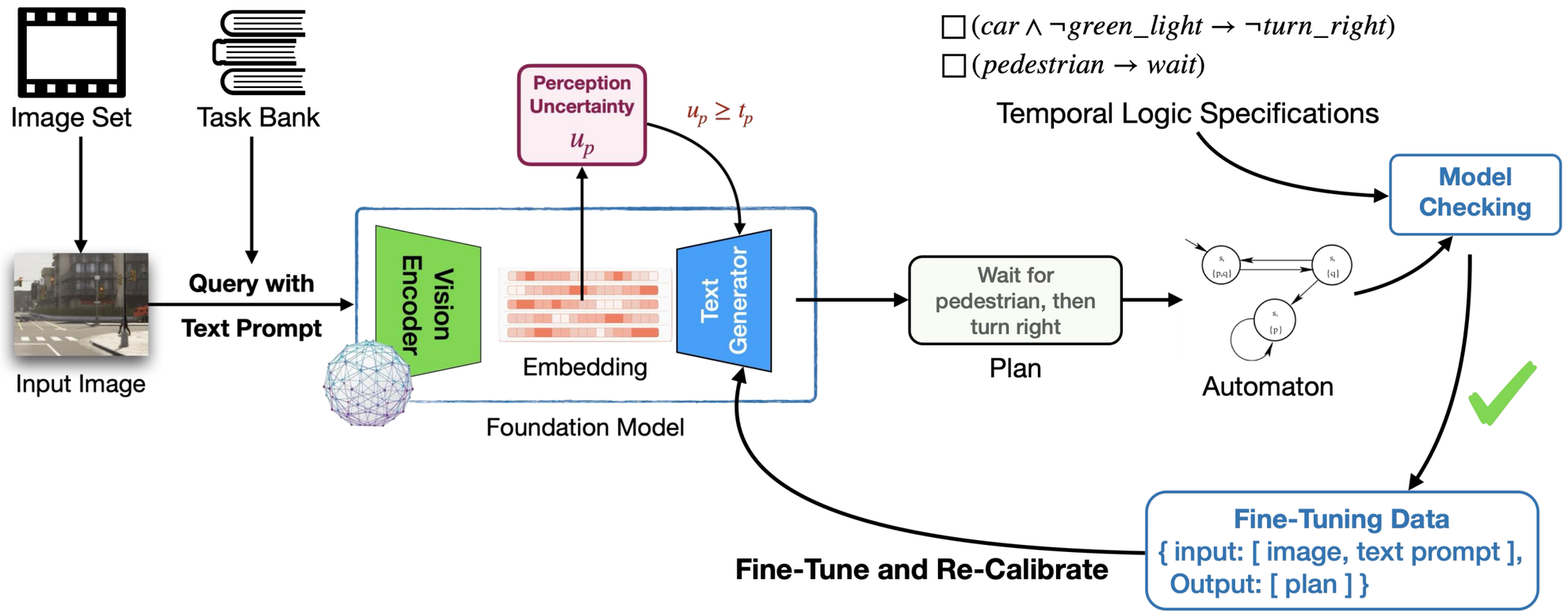
$$\square(\neg \text{green traffic light} \rightarrow \neg \text{go straight}),$$

$$\square(\text{stop sign} \rightarrow \Diamond\, \text{stop}),$$

How can we check whether the plan satisfies the logical specifications?

Decision Uncertainty Score = 0.7
“at least 70% probability that the plan satisfies the specifications”

Automated Fine-tuning With Probabilistic Guarantees



Obtaining High-quality Fine-tuning Data Without HIL

Go straight at the intersection

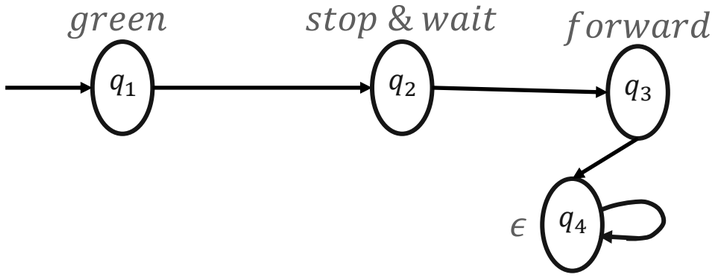
Algorithm 1: Natural Language to Kripke Structures

input textual instruction T, atomic proposition set AP, set Y of observed objects
output $(Q, q_0, \delta, \lambda)$
$Ph = \{Ph_1, Ph_2, ...\}$ = **parse**(T)
Q, δ = $[q_0]$, [] {Define a set of states and transitions. q_0 denotes initial states}
$\lambda(q_0) = Y \cap AP$ {The initial state's label is the observed objects from the image}
for Ph_i in Ph
  Q.append(q_i), δ.append$((q_{i-1}, q_i))$,
  $\lambda(q_i) = \{p \in AP : p \in Ph_i\}$
end for
Q.append(q_{done}), δ.append$((q_{|Ph|}, q_{done}))$,
δ.append$(q_{done}, q_{done}))$, $\lambda(q_{done}) = \emptyset$

Controller Construction

1. The traffic light is green
2. Stop at the intersection
3. Wait for pedestrian
4. Move forward

Confidence 0.25



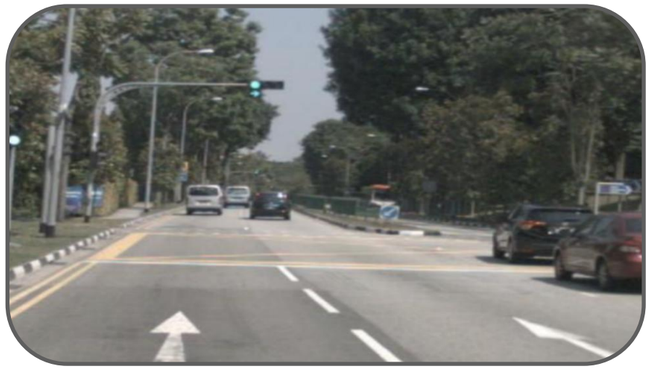
$\square$(¬green traffic light $\rightarrow$ ¬go straight),

$\square$(stop sign $\rightarrow \Diamond$ stop),

$\square$(green light $\wedge$ ¬pedestrian $\rightarrow \bigcirc$ ¬wait),

Obtaining High-quality Fine-tuning Data Without HIL

Go straight at the intersection



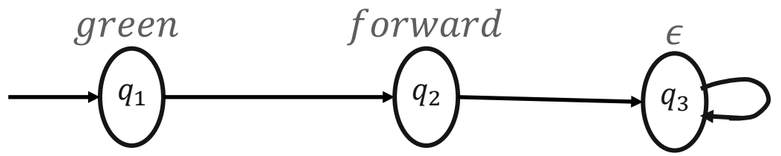
Algorithm 1: Natural Language to Kripke Structures

input textual instruction T, atomic proposition set AP, set Y of observed objects
output $(Q, q_0, \delta, \lambda)$
$Ph = \{Ph_1, Ph_2, ...\}$ = **parse**(T)
Q, δ = $[q_0]$, [] {Define a set of states and transitions. q_0 denotes initial states}
$\lambda(q_0) = Y \cap AP$ {The initial state's label is the observed objects from the image}
for Ph_i in Ph
  Q.append(q_i), δ.append$((q_{i-1}, q_i))$,
  $\lambda(q_i) = \{p \in AP : p \in Ph_i\}$
end for
Q.append(q_{done}), δ.append$((q_{|Ph|}, q_{done}))$,
δ.append$(q_{done}, q_{done}))$, $\lambda(q_{done}) = \emptyset$

Controller Construction

1. The traffic light is green
2. Move forward

Confidence 0.75

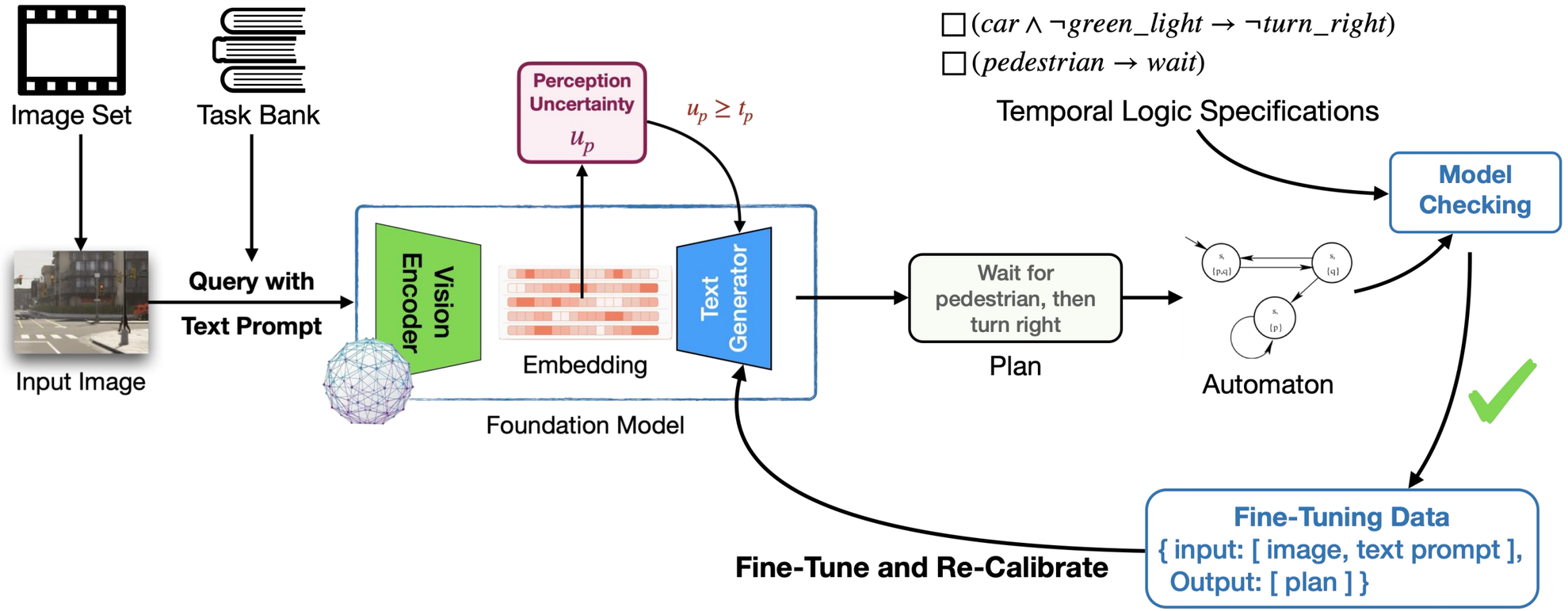


$\square(\neg$green traffic light $\rightarrow \neg$go straight),

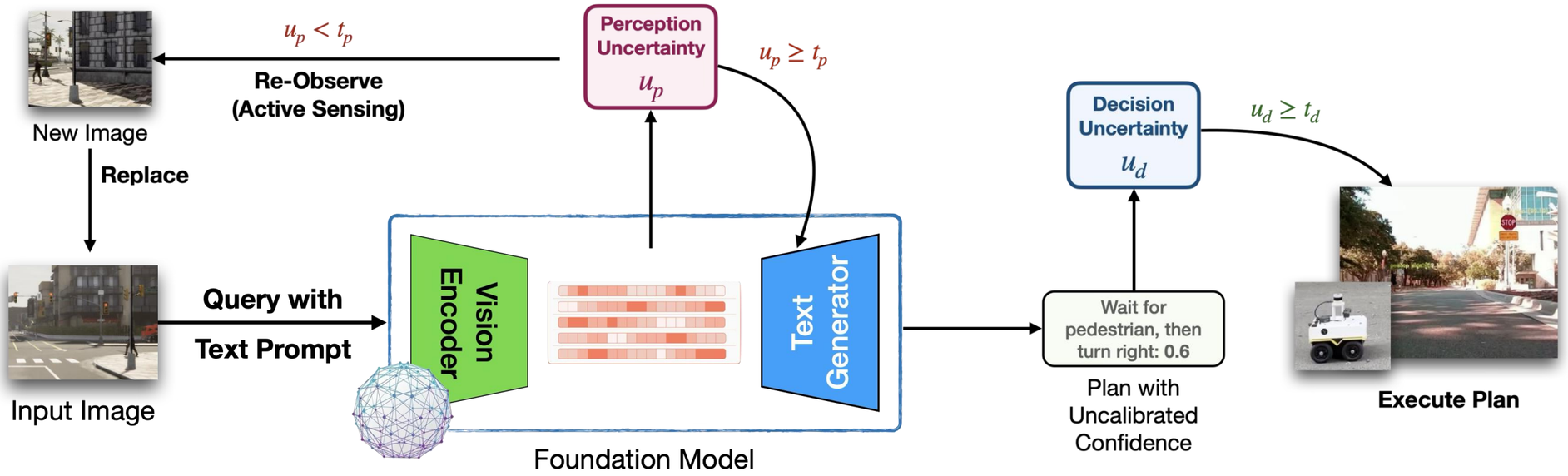
$\square$(stop sign $\rightarrow \Diamond$ stop),

$\square$(green light $\wedge \neg$pedestrian $\rightarrow \bigcirc \neg$wait),

Automated Fine-tuning With Probabilistic Guarantees



Active Sensing at Inference

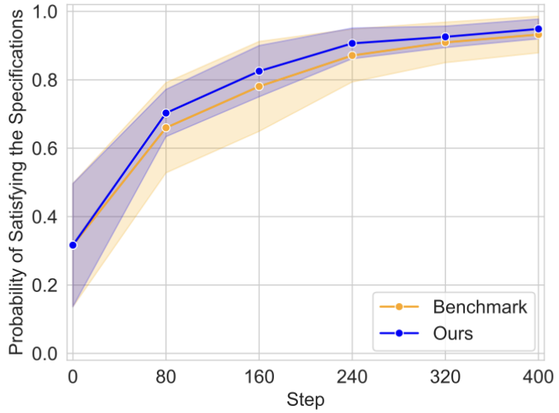


Qualitative Demonstrations

Quantitative Results

Benchmarking

Planning Pipeline	Avg. Percp. Unc. Score	Avg. Dec. Unc. Score	Prob. of Satisfying Spec. (Avg)	Prob. of Satisfying Spec. (SD)
Raw Model w/o AS	0.842	0.279	—	—
Raw Model with AS	**0.936**	0.306	0.316	0.180
Fine-tuned Model (Benchmark) with AS	**0.936**	0.931	0.933	0.048
Fine-tuned Model (Ours) with AS	**0.936**	**0.955**	**0.959**	**0.025**



Key Contributions:

(1) Identification and disentanglement of the source of uncertainty in multimodal foundation models into:
- **Perception uncertainty** associated with the model's visual processing capabilities
- **Decision uncertainty** linked to its ability to generate actionable plans

(2) Uncertainty-guided targeted interventions: scalable model fine-tuning (offline) and active sensing (online)

(3) Reduction of decision variability by up to **40%** with a single re-query and up to **2x** increase in number of specifications satisfied

Know Where You're Uncertain When Planning with Multimodal Foundation Models: A Formal Framework

Neel P. Bhatt*, Yunhao Yang*, Rohan Siva, Daniel Milan, Ufuk Topcu, Zhangyang Wang

Thank you!

Project Page

Paper

TEXAS
The University of Texas at Austin

VITA

CENTER FOR
aUTonomy