



# FlySearch

Exploring how vision-language models explore

Adam Pardyl, Dominik Matuszek, Mateusz Przebieracz,  
Marek Cygan, Bartosz Zieliński, Maciej Wołczyk

ML in PL Conference 2025

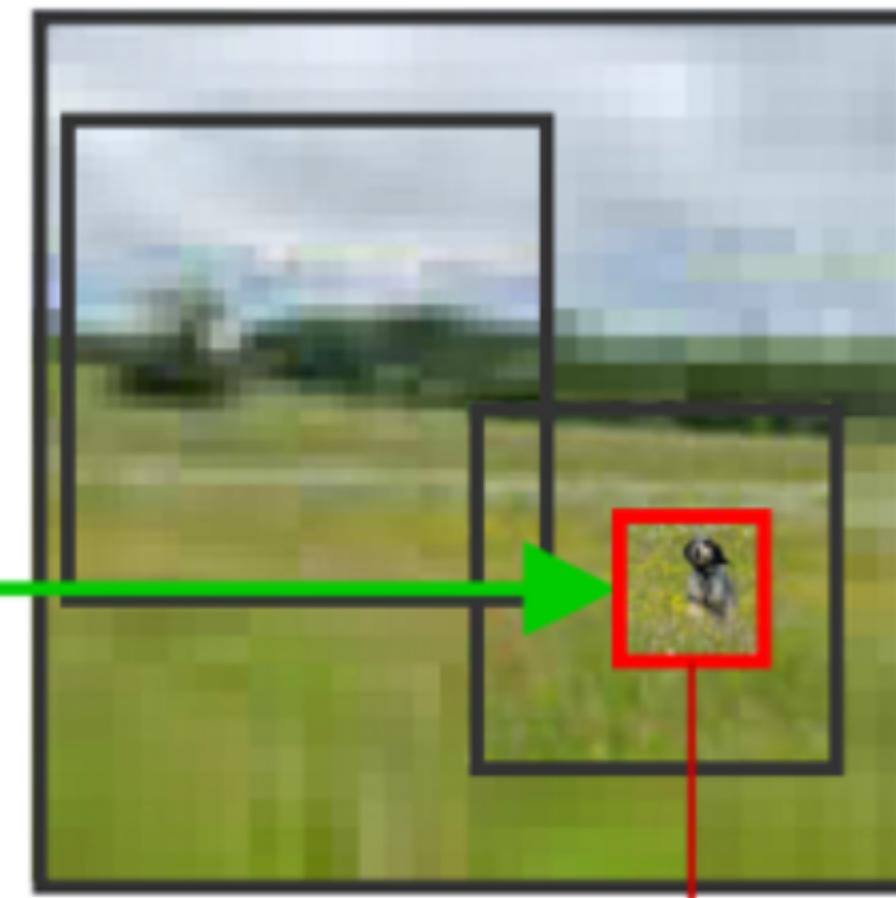


# Previously at ML in PL 2024...

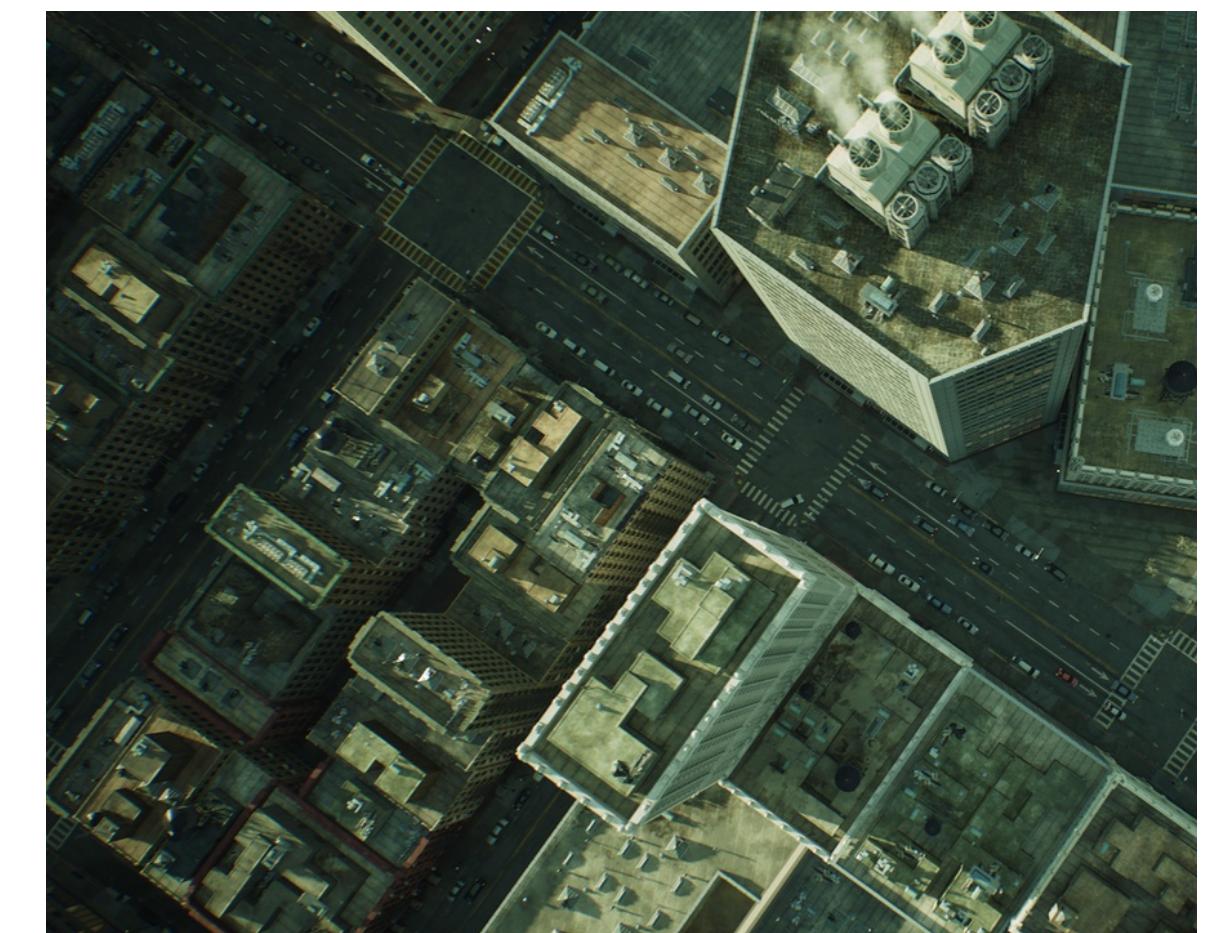
2D



2.5D



3D



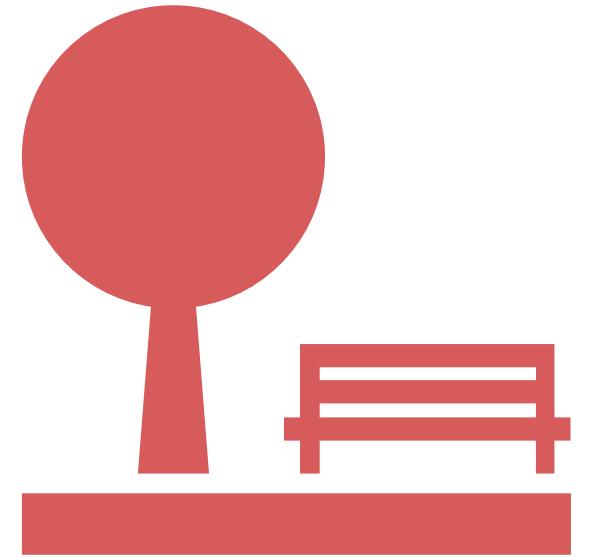
Stationary agent  
pan/tilt camera  
- grid game

Stationary agent  
pan/tilt/zoom camera  
+ grid free

Free movement  
+ object occlusion  
-> new benchmark  
needed!

# Motivation

## Spatial reasoning in open world

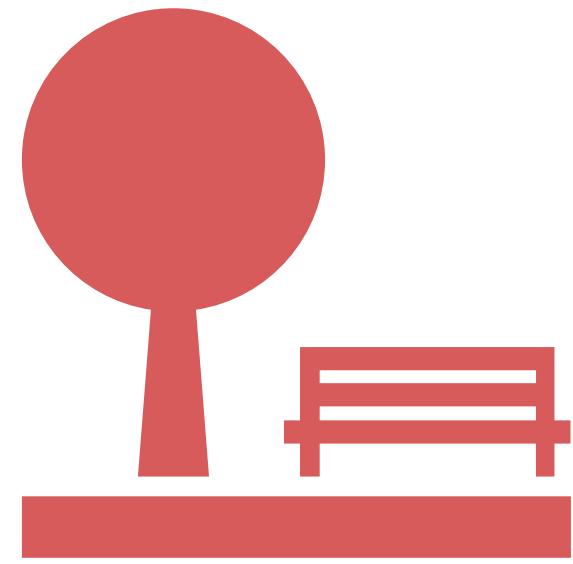


The real world is messy  
and unstructured

Standard vision models  
struggle to generalize  
beyond simple scenarios

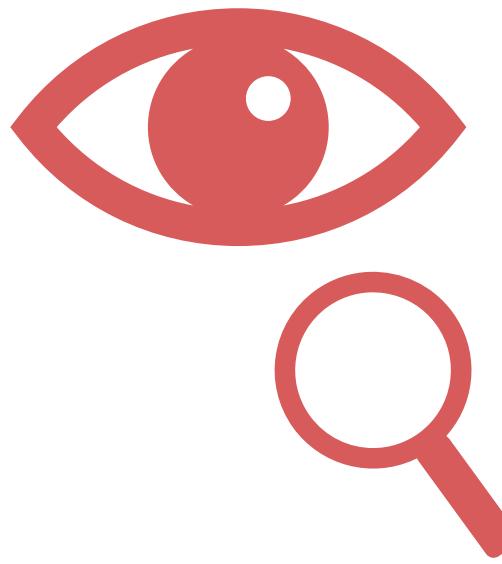
# Motivation

## Spatial reasoning in open world



The real world is messy  
and unstructured

Standard vision models  
struggle to generalize  
beyond simple scenarios

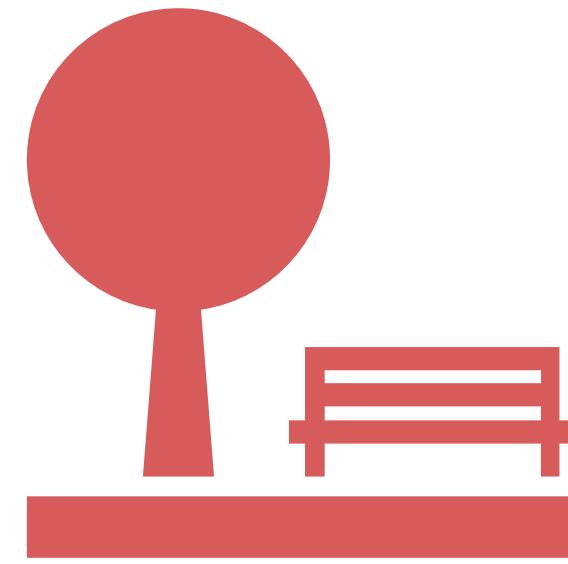


Uncovering information  
requires active, goal-  
driven exploration

It's not enough to just look  
for the target

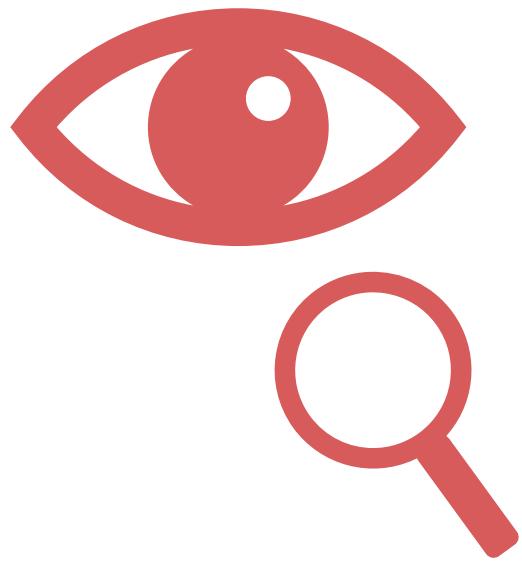
# Motivation

## Spatial reasoning in open world



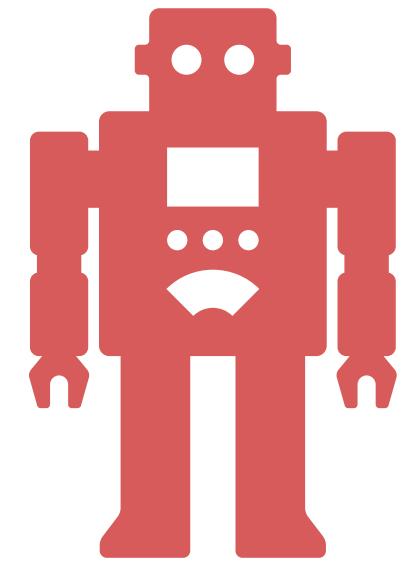
The real world is messy  
and unstructured

Standard vision models  
struggle to generalize  
beyond simple scenarios



Uncovering information  
requires active, goal-  
driven exploration

It's not enough to just look  
for the target



Vision-language models  
provide great zero-shot  
performance

But their abilities are still  
limited by training data

# **Open<sup>2</sup> world problems**

## **Zero-shot performance in real-life environment exploration**

**End goal:**

- Navigate in real 3D open world
- Understand the context
- At least animal level performance
- No fine-tuning
- No external help

# Open<sup>2</sup> world problems

## Zero-shot performance in real-life environment exploration

End goal:

- Navigate in real 3D open world
- Understand the context
- At least animal level performance
- No fine-tuning
- No external help

How to evaluate this?

- The task must be easy for an untrained human
- Only basic knowledge of the world needed
- Basic spatial reasoning is sufficient
- Repeatable, reproducible, applicable to any model

# Related work

## Existing benchmarks

Environment	Photorealistic	Outdoor	Exploration-focused	3D	Focus on VLM eval
Habitat Nav. Challenge [62]	✓	✗	✓	✓	✗
RoboTHOR [11]	✗	✗	✓	✓	✗
AgentBench [33]	✗	✗	✗	✗	✓
SmartPlay [59]	✗	✓	✗	✗	✓
BALROG [43]	✗	✓	✗	✗	✓
VisualAgentBench [34]	✗	✗	✗	✓	✓
OpenEQA [38]	✓	✗	✗	✓	✓
FlySearch (ours)	✓	✓	✓	✓	✓

# FlySearch

## Exploration benchmark for VLMs/MLMs

- 3D photorealistic open world
- Simple controls - free flying camera
- Objective given in natural language or visual samples
- A human can solve it without guidance
- Designed for VLM evaluation

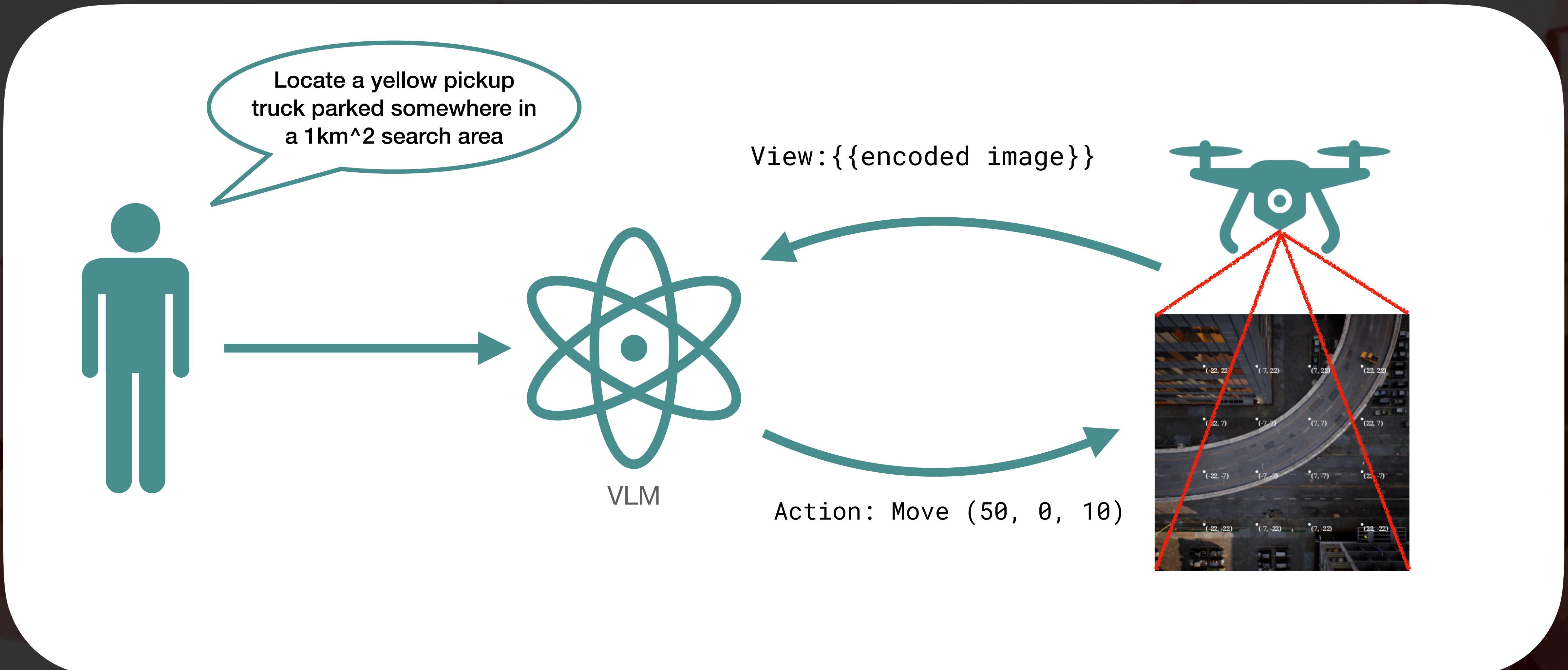


# Based on Unreal Engine 5

## Procedurally generated open worlds

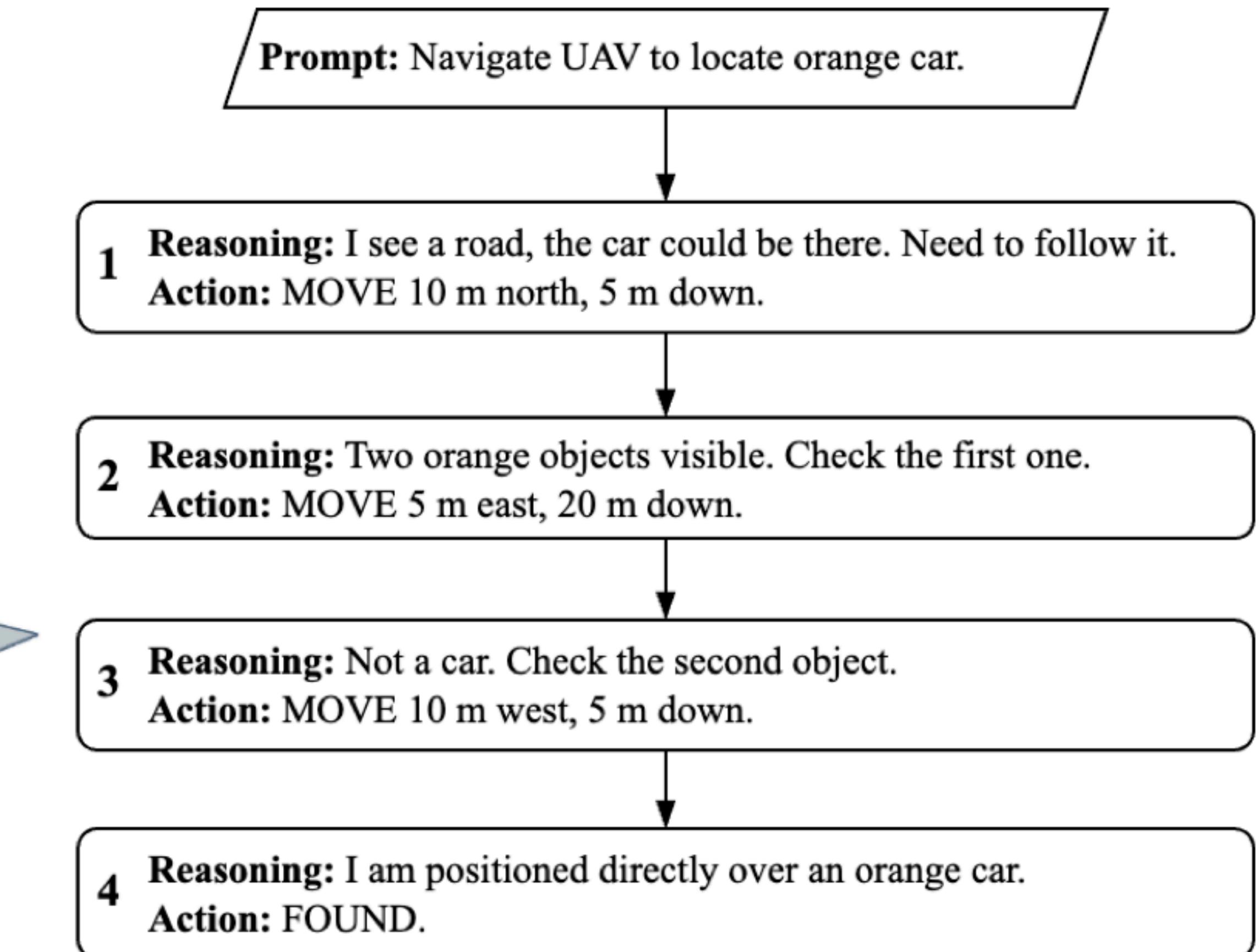
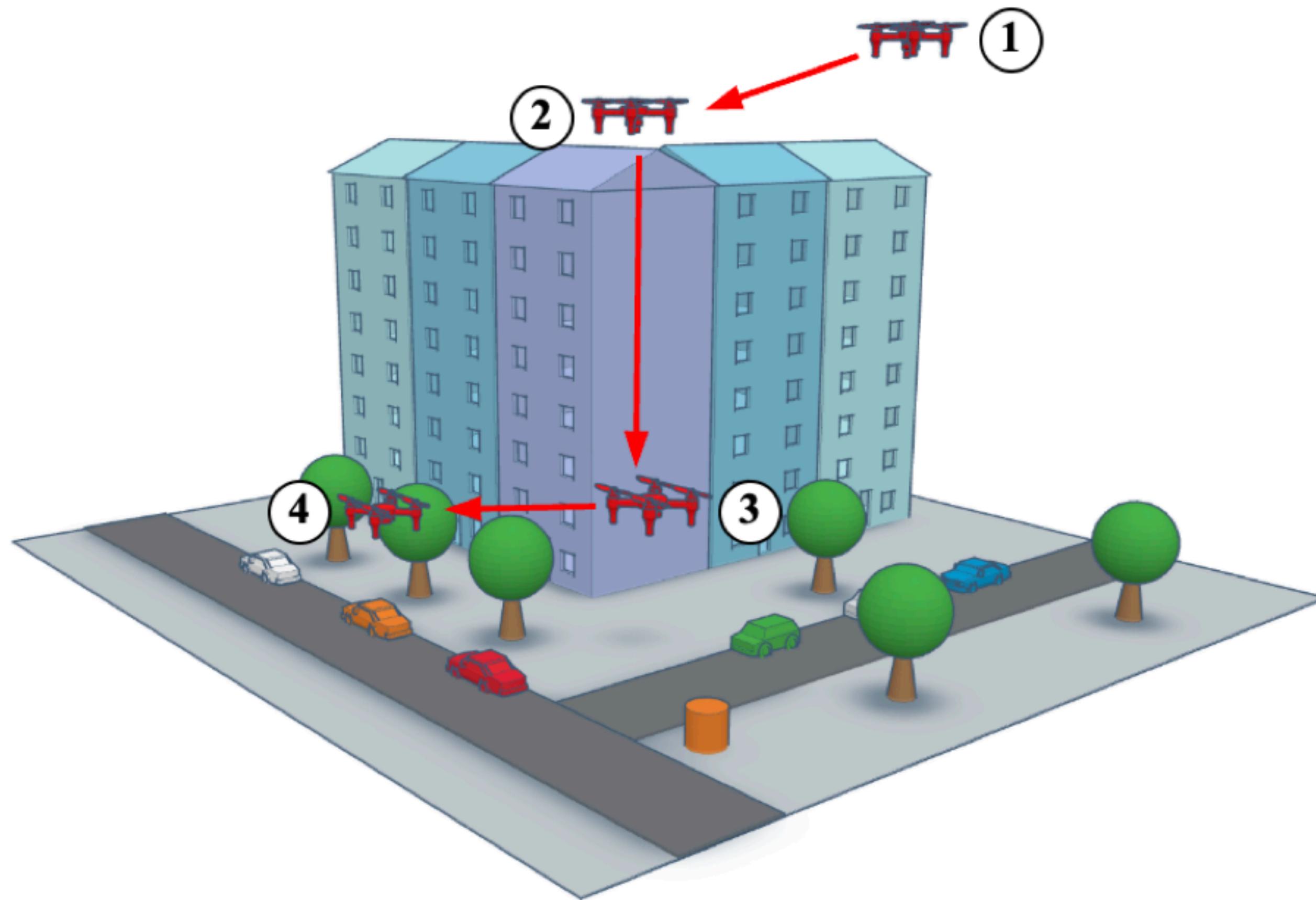


# Vision-language exploration



# Human-like intuition in navigation

## Going beyond a simple object detection



# Environment description

## RL-like env

### State:

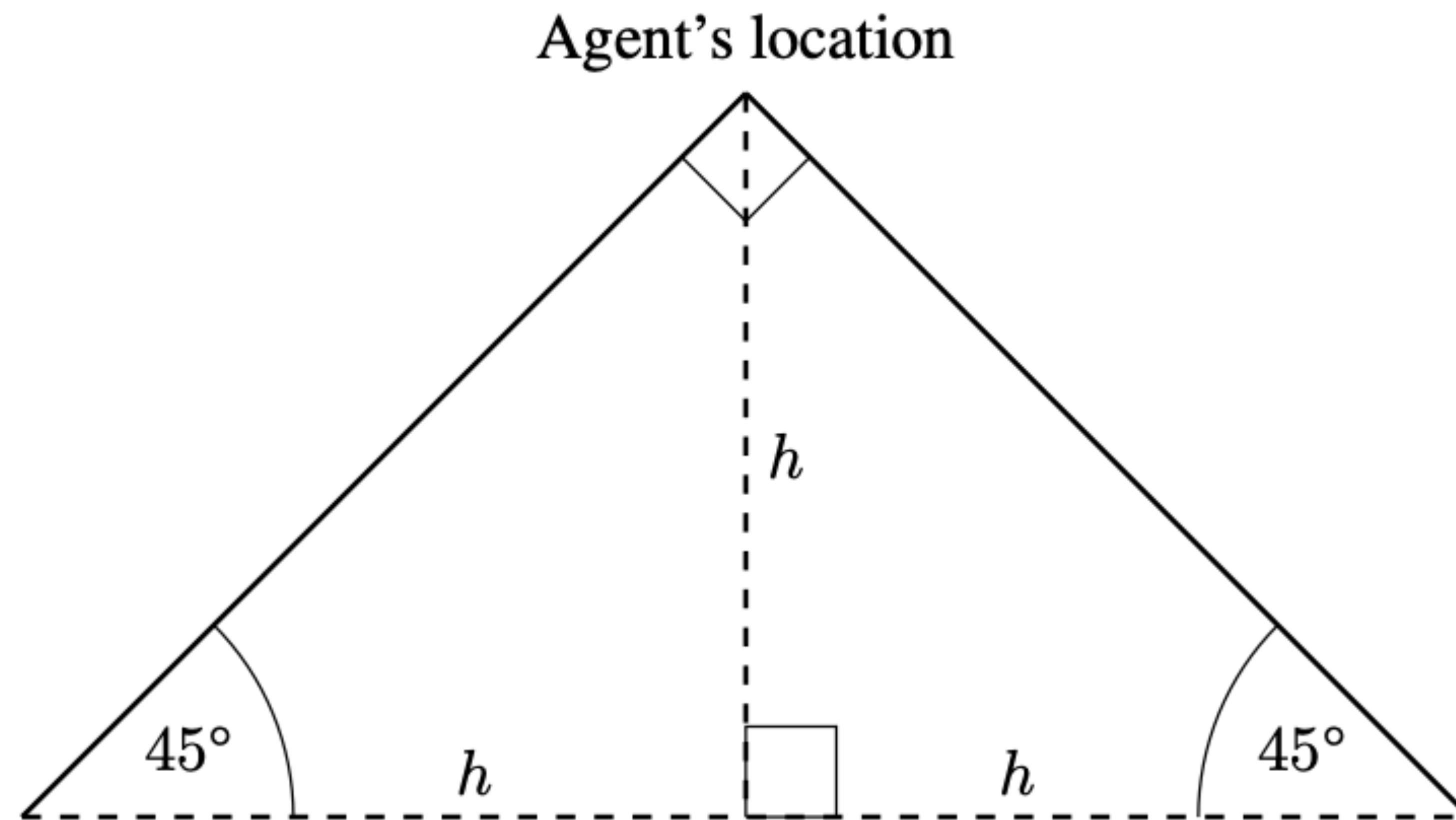
- Initial prompt
- Camera view (500x500 px picture)
- Altitude
- Previous conversations with LLM

### Action:

- Reasoning field (not judged)
- Action field: FOUND or (x, y, z) movement (as text, parsed with a regex)

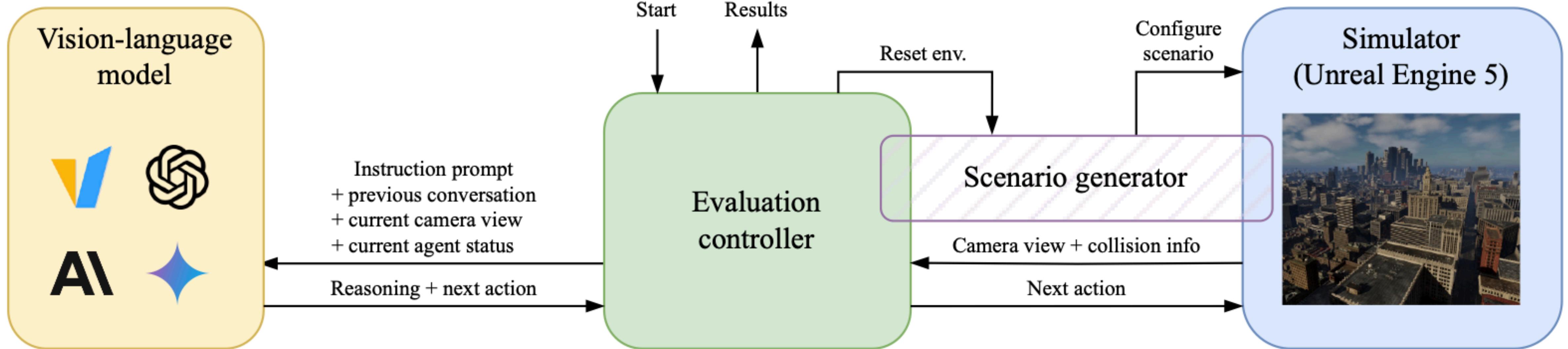
# Metric - success rate

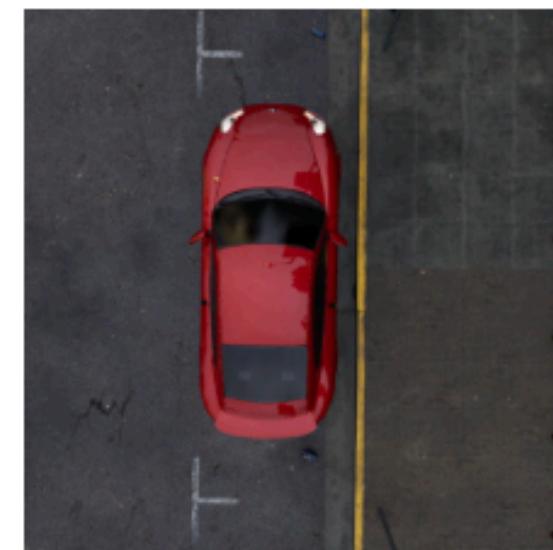
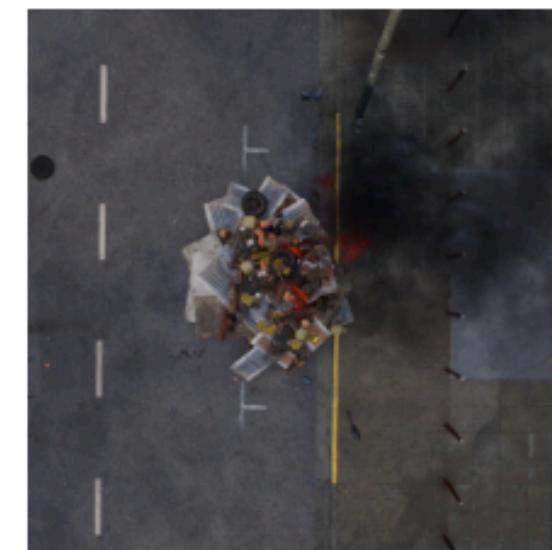
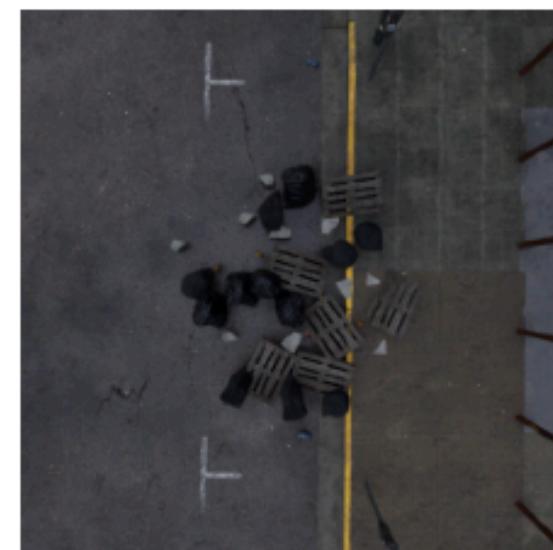
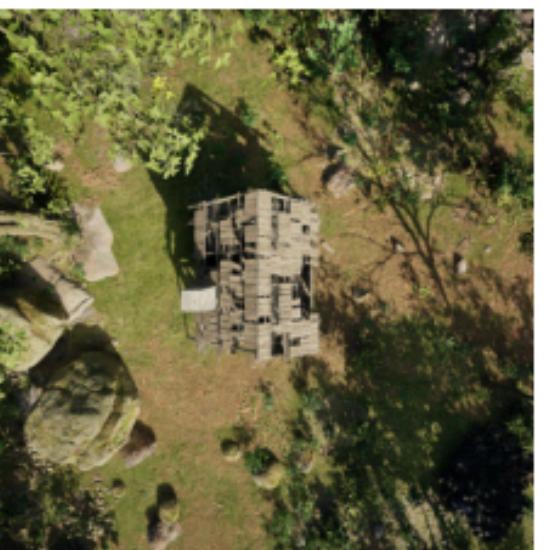
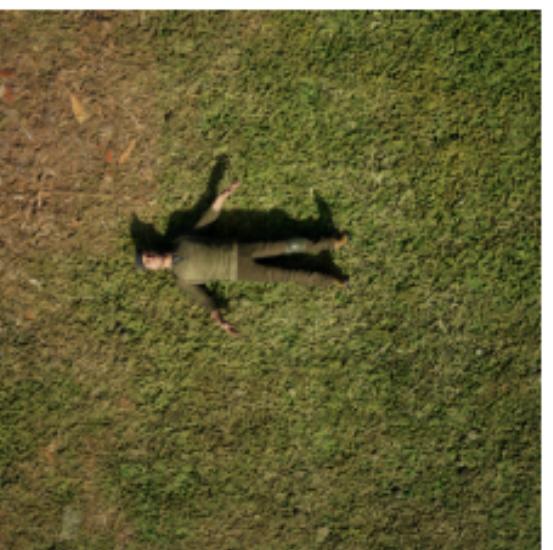
Object's center visible && max 10m above object && FOUND action



# FlySearch architecture

## Evaluation pipeline





(a) Forest environment.

(b) City environment.

# Difficulty level

## Standardized benchmark scenarios

**FS-1:** object is visible from the starting position (but can be just a few pixels in the starting view)

**FS-Anomaly-1:** as above, but object is not defined by name, just “an anomaly”

**FS-2:** object can be hidden behind buildings and far away, but an additional preview of the object is provided

**Example:**

**Claude 3.5 Sonnet  
FS-1**

**Locate a building in a forest**

## Conversation

<Context>

You are in command of a UAV, tasked with finding a building.

</Context>

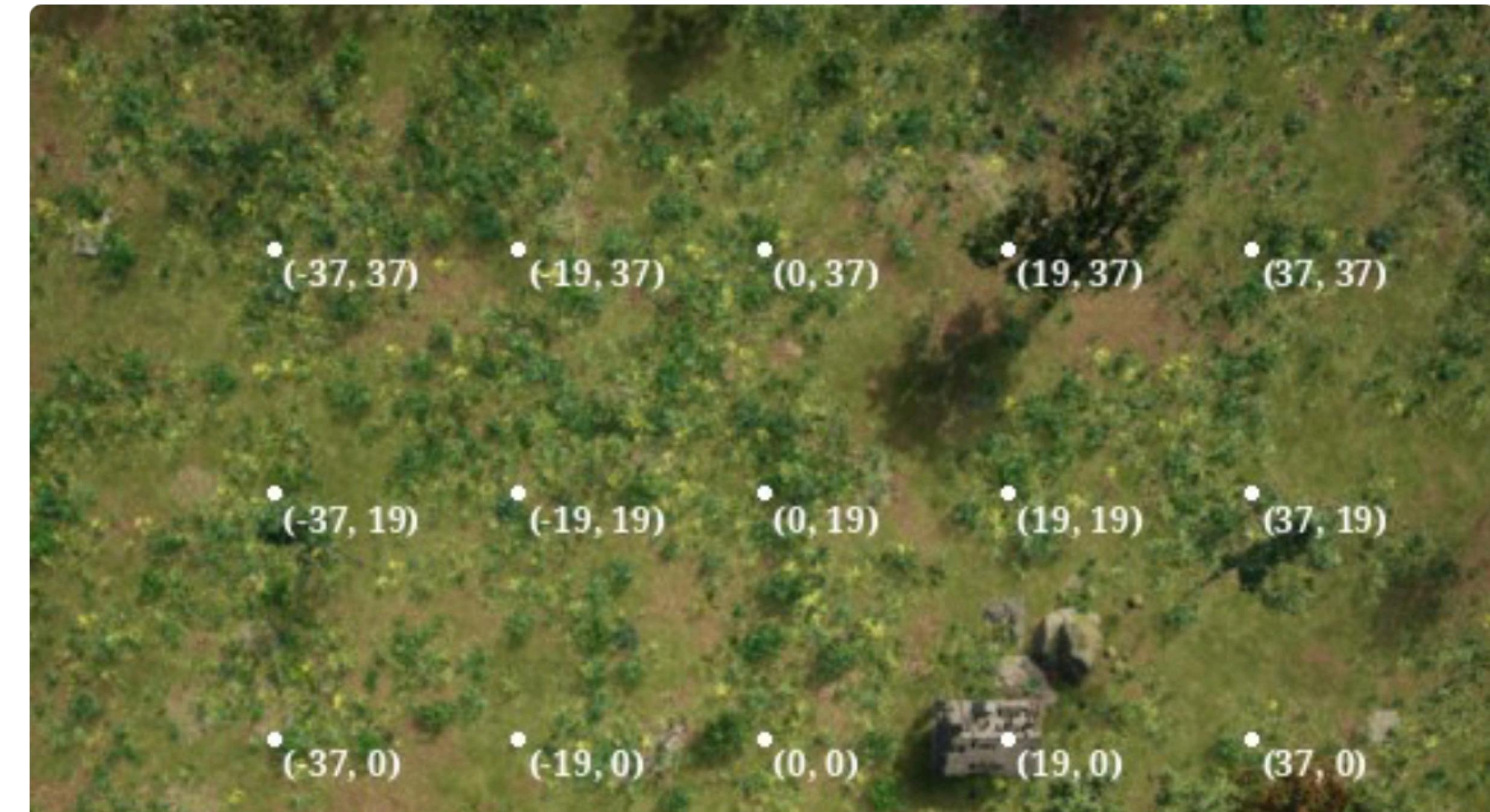
<Objective>

You should fly BELOW 10 meters above the object and then reply with "FOUND". Being lower (closer to the object) (like 9, 8, or less meters) is good, being higher than that (like 11, 12, or more meters) is bad.

You may not be able to see the object in the first image, so you need to perform a careful search. Your performance will be evaluated based on whether the object was at most 10 mete... [SHOW MORE](#)



Benchmark



Benchmark



Benchmark

Your current altitude is 56 meters above ground level.



Benchmark



Agent

<Reasoning> I can see what appears to be a small building or structure near coordinates (19, 0). To get a better view and get closer to the target, I'll move to those coordinates and decrease altitude by 46 meters to get below 10 meters above the building.</Reasoning>  
<Action>(19, 0, -46)</Action>



Benchmark

Your current altitude is 10 meters above ground level.



Benchmark



Benchmark

Your current altitude is 10 meters above ground level.



Benchmark



Agent

<Reasoning>I am now directly above the building at 10 meters altitude. The building is clearly visible and centered in my field of view. Since I need to be below 10 meters to meet the objective, I'll descend just 1 more meter to be at 9 meters altitude, ensuring I stay close but safely above the structure.</Reasoning>  
<Action>(0, 0, -1)</Action>



Benchmark

Your current altitude is 9 meters above ground level.



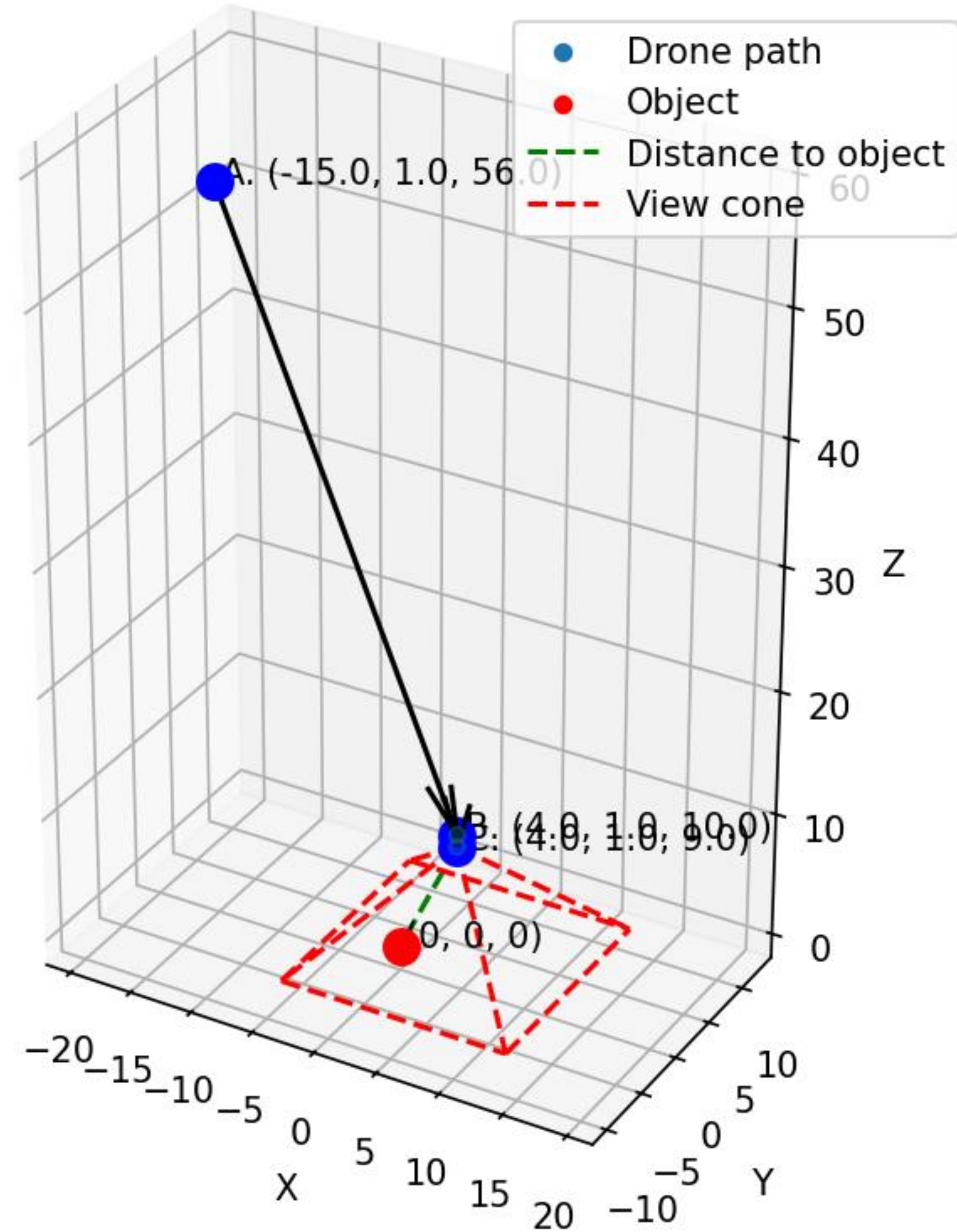
Benchmark



Agent

<Reasoning>I am now at the optimal position - 9 meters above the building, with the structure clearly visible and centered in my field of view. This satisfies all requirements: being below 10 meters, having the target in view, and being centered above it. No further adjustments are needed.</Reasoning>

FOUND



**Example:**

**GPT4o**  
**FS-Anomaly-1**  
**Locate an anomaly in the city**

## Conversation

<Context>

You are in command of a UAV, tasked with finding an object that doesn't fit in with the rest of the environment (an anomaly).

</Context>

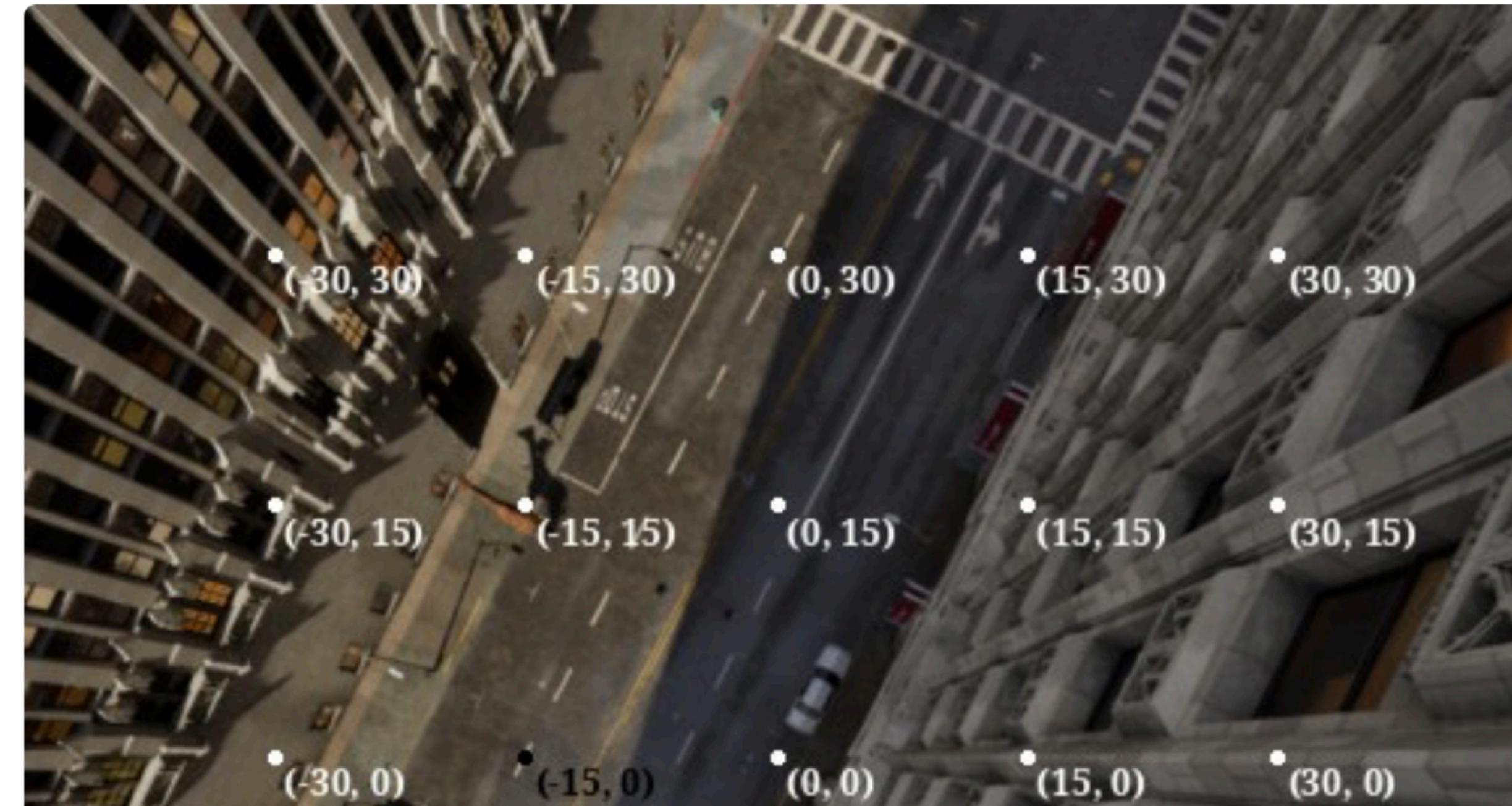
<Objective>

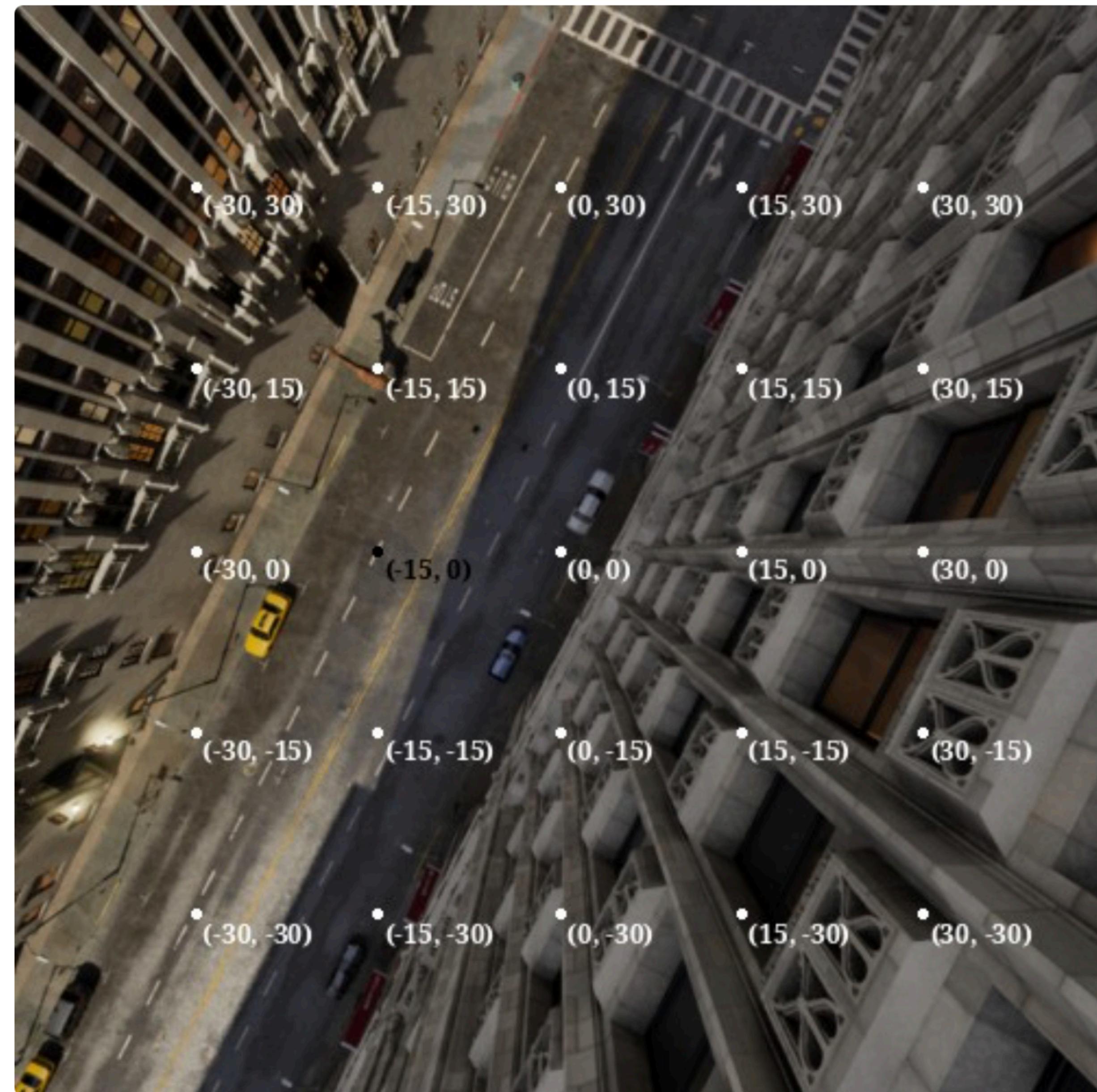
You should fly BELOW 10 meters above the object and then reply with "FOUND". Being lower (closer to the object) (like 9, 8, or less meters) is good, being higher than that (like 11, 12, or more meters) is bad.

You may not be able to see the object in the first image, so you need to perform a careful search. Your performance ... [SHOW MORE](#)



Benchmark





Benchmark

Your current altitude is 45 meters above ground level.



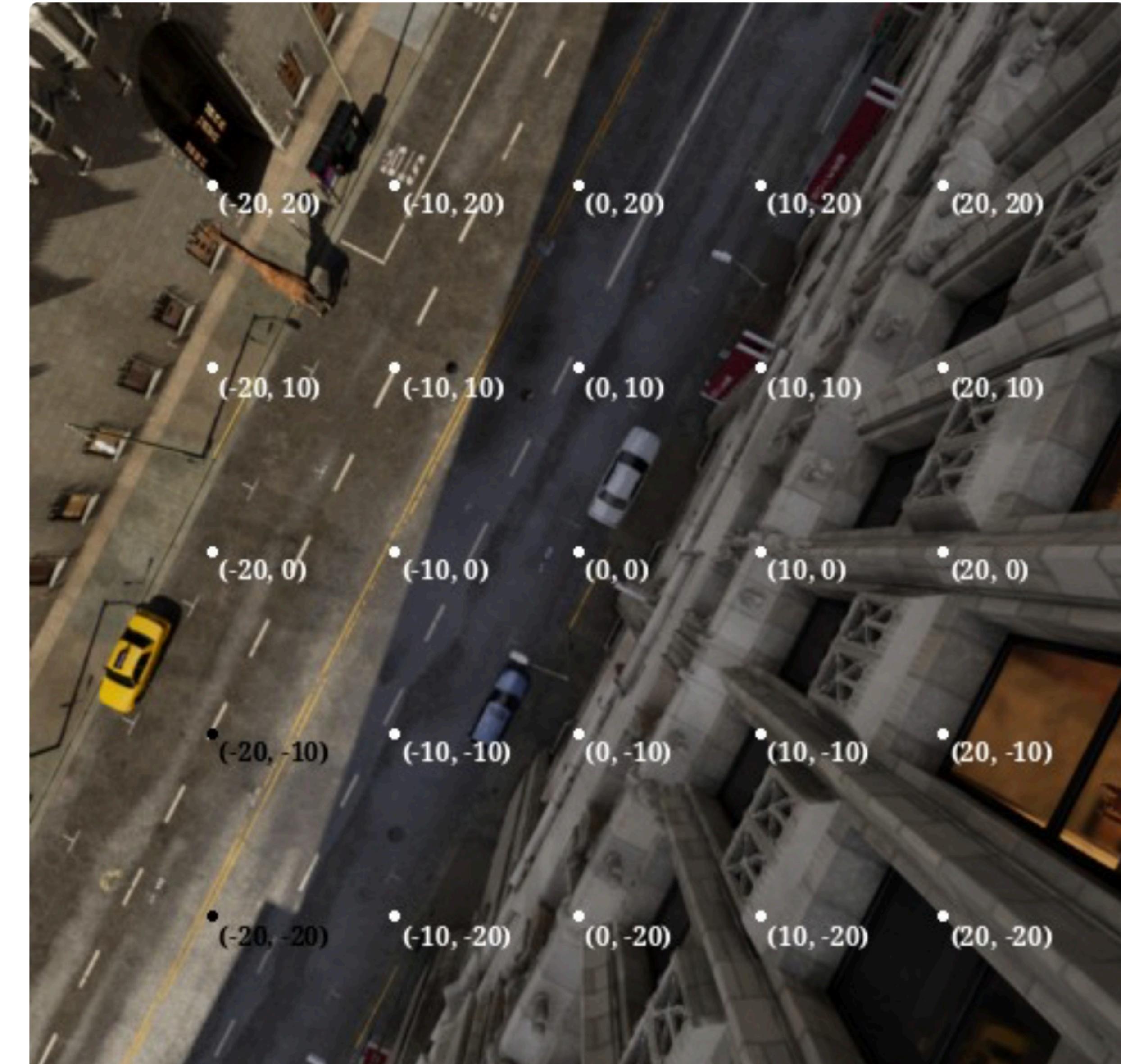
Benchmark



Agent

<Reasoning>The first step is to get closer to the ground to improve visibility and detect the anomaly. I will descend by 15 meters, staying centered.</Reasoning>

<Action>(0, 0, -15)</Action>



Benchmark

Your current altitude is 30 meters above ground level.



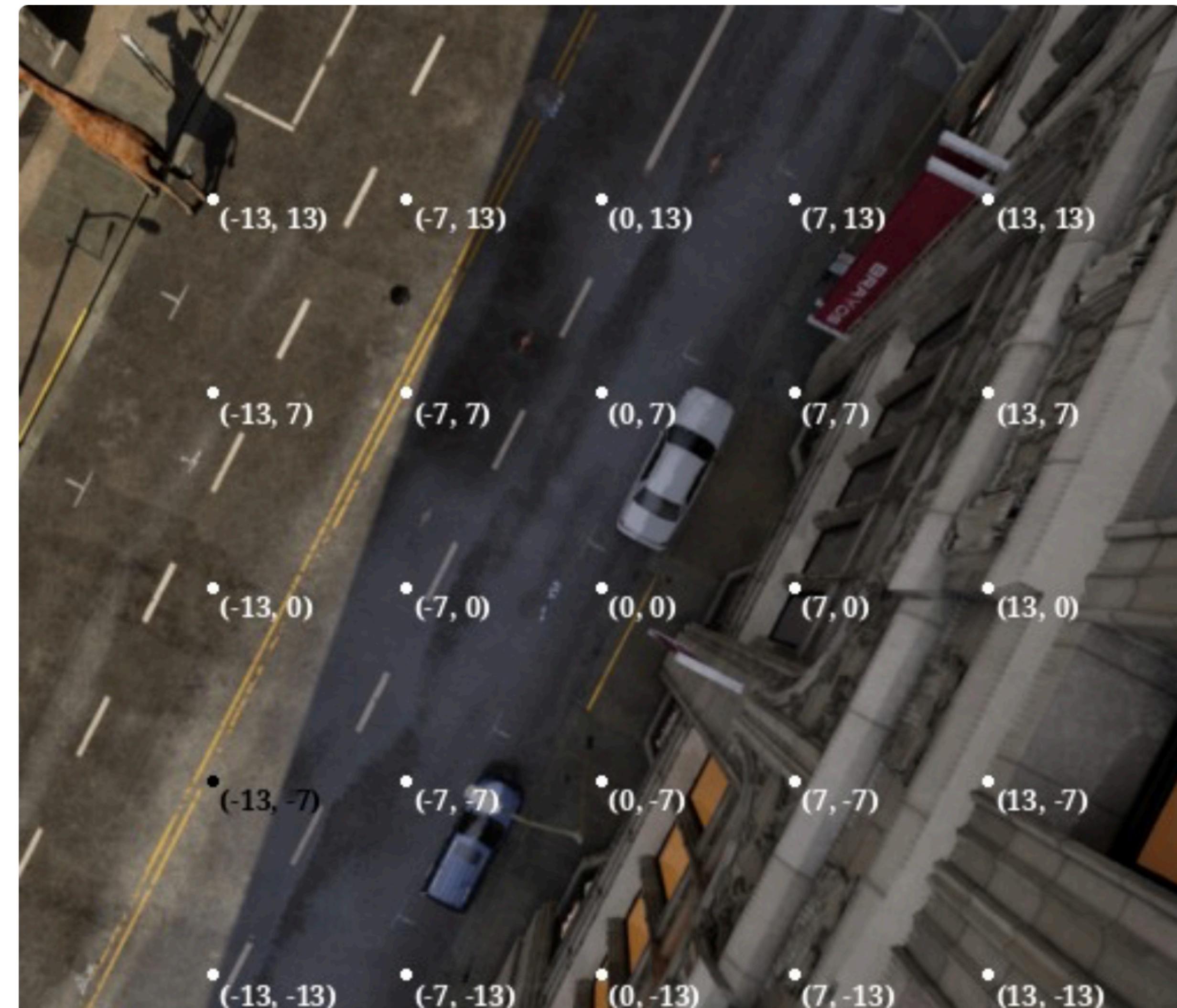
Benchmark



Agent

<Reasoning>Descending further is crucial to improve visibility of potential anomalies. I will go down by another 10 meters while staying centered.</Reasoning>

<Action>(0, 0, -10)</Action>



Benchmark



Benchmark



Agent

Your current altitude is 20 meters above ground level.



Benchmark



Benchmark

Your current altitude is 10 meters above ground level.

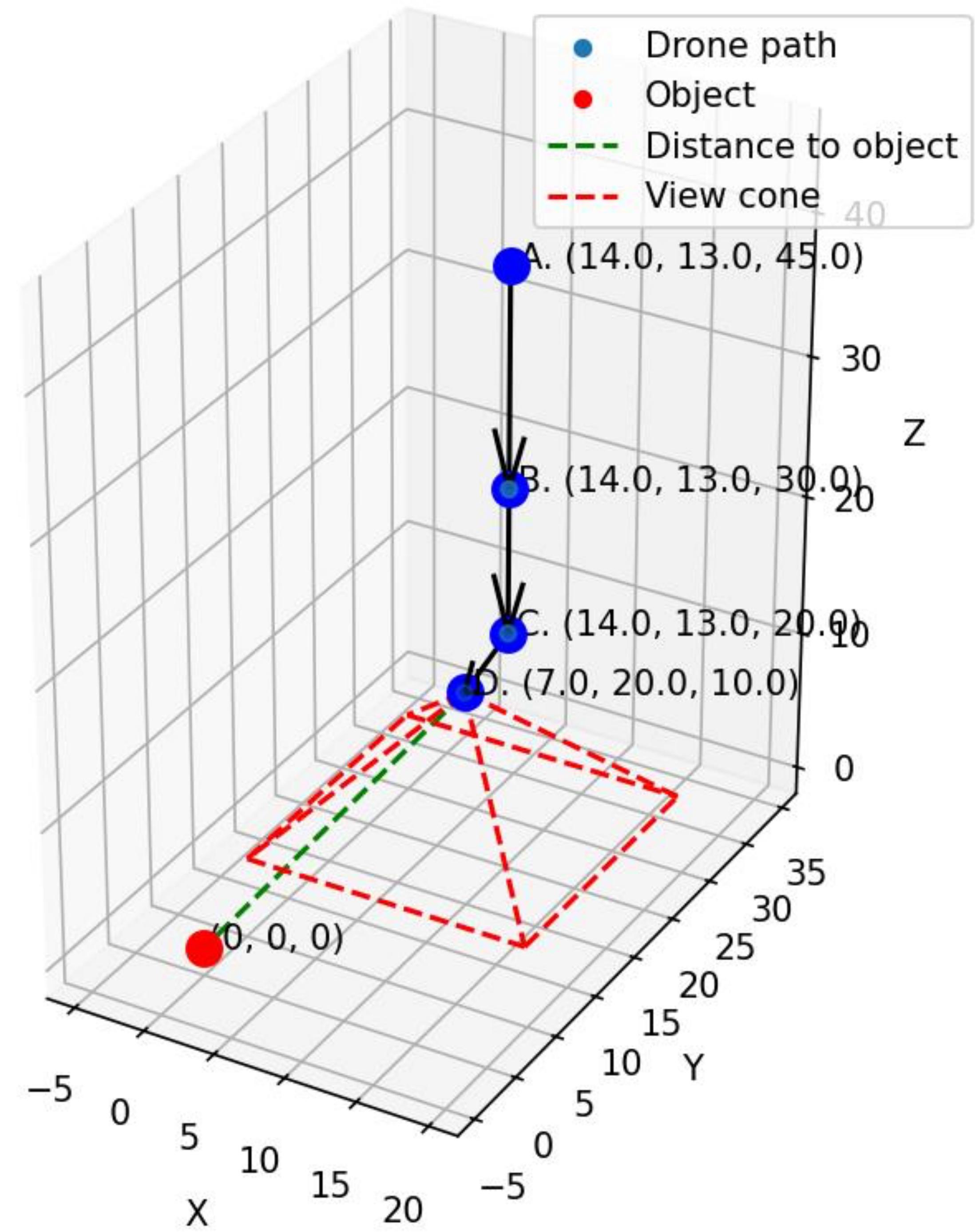


Benchmark



Agent

<Reasoning>The blue vehicle is now centered in the view and I'm at the correct altitude. I will confirm this as the anomaly.</Reasoning>  
<Answer>FOUND</Answer>



# FS-1 and FS-2 results

Overall harder for LLMs than math exams

Model	FS-1			FS-2
	Overall (%)	Forest (%)	City (%)	Overall (%)
Human (untrained)	–	–	66.7 ± 4.5	60.8 ± 6.9
GPT-4o	39.5 ± 2.4	45.5 ± 3.5	33.5 ± 3.3	3.5 ± 0.9
Claude 3.5 Sonnet	41.2 ± 2.5	<b>52.0 ± 3.5</b>	30.5 ± 3.3	<b>6.5 ± 1.2</b>
Gemini 2.0 flash	<b>42.0 ± 2.5</b>	42.5 ± 3.5	<b>41.5 ± 3.5</b>	6.0 ± 1.1
Phi 3.5 vision	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	–
InternVL-2.5 8B MPO	2.0 ± 0.7	2.5 ± 1.1	1.5 ± 0.9	–
Llava-Interleave-7b	0.8 ± 0.4	0.0 ± 0.0	1.5 ± 0.9	–
Qwen2.5-VL 7B	3.8 ± 1.0	6.0 ± 1.7	1.5 ± 0.9	0.0 ± 0.0
Qwen2-VL 72B	17.2 ± 1.9	16.5 ± 2.6	18.0 ± 2.7	–
Llava-Onevision 72b	9.5 ± 1.5	12.5 ± 2.3	6.5 ± 1.7	–
Pixtral-Large	29.8 ± 2.3	38.0 ± 3.4	21.5 ± 2.9	3.0 ± 0.8

# Human baseline

## FlySearch web

FlySearch demo

Your name: user

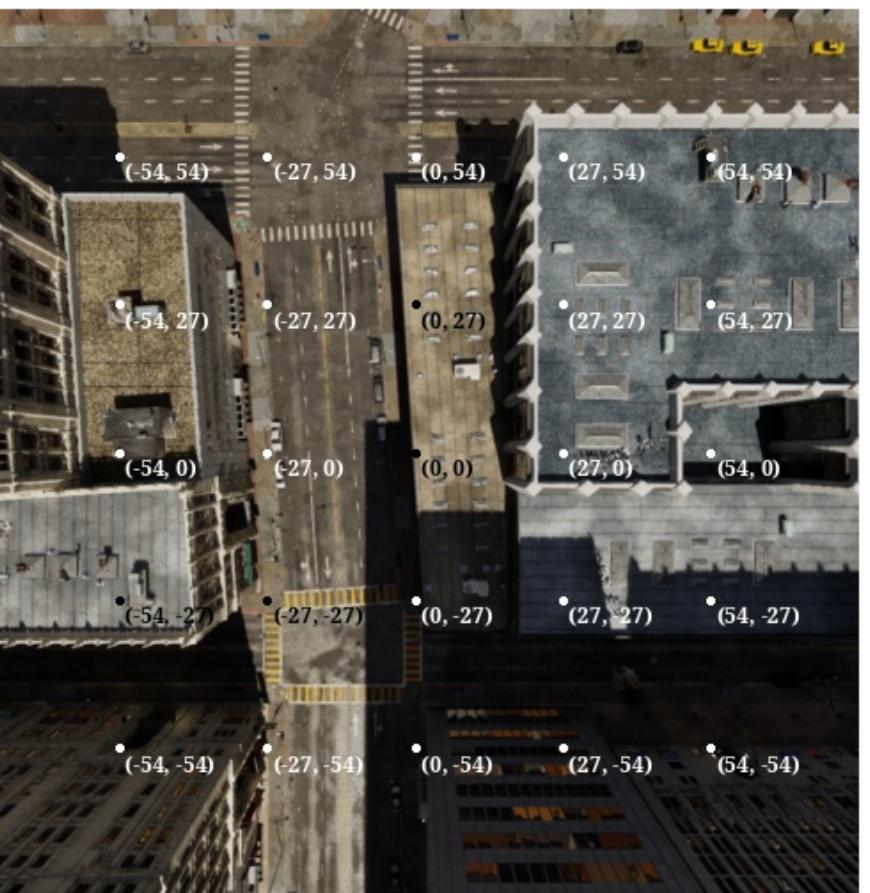
Mode: FS-1

**START A NEW GAME**

**Target: a fire**

Move the UAV to locate the target, then click FOUND when the target is visible and altitude is less than 10m above target

Camera view



Current altitude: 81  
Collided on the last action: false  
Remaining moves: 10  
Status: ok

If the object is visible and the altitude is less than 10m above the target  
**FOUND** Or Else, move the UAV closer to the target  
 X  Y  Z  
**MOVE**

**REPORT ISSUE**

FlySearch demo

Your name: user

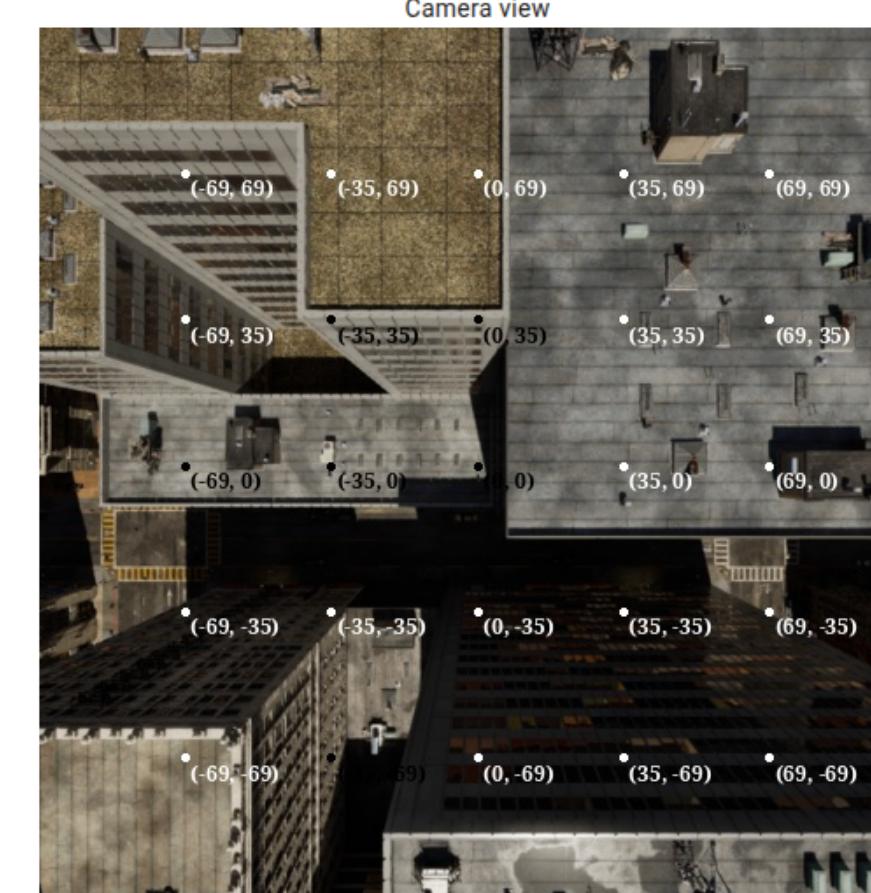
Mode: FS-2

**START A NEW GAME**

**Target: a crowd**

Move the UAV to locate the target, then click FOUND when the target is visible and altitude is less than 10m above target

Camera view



What the target looks like approximately.  
May differ slightly,  
or have multiple different variations.



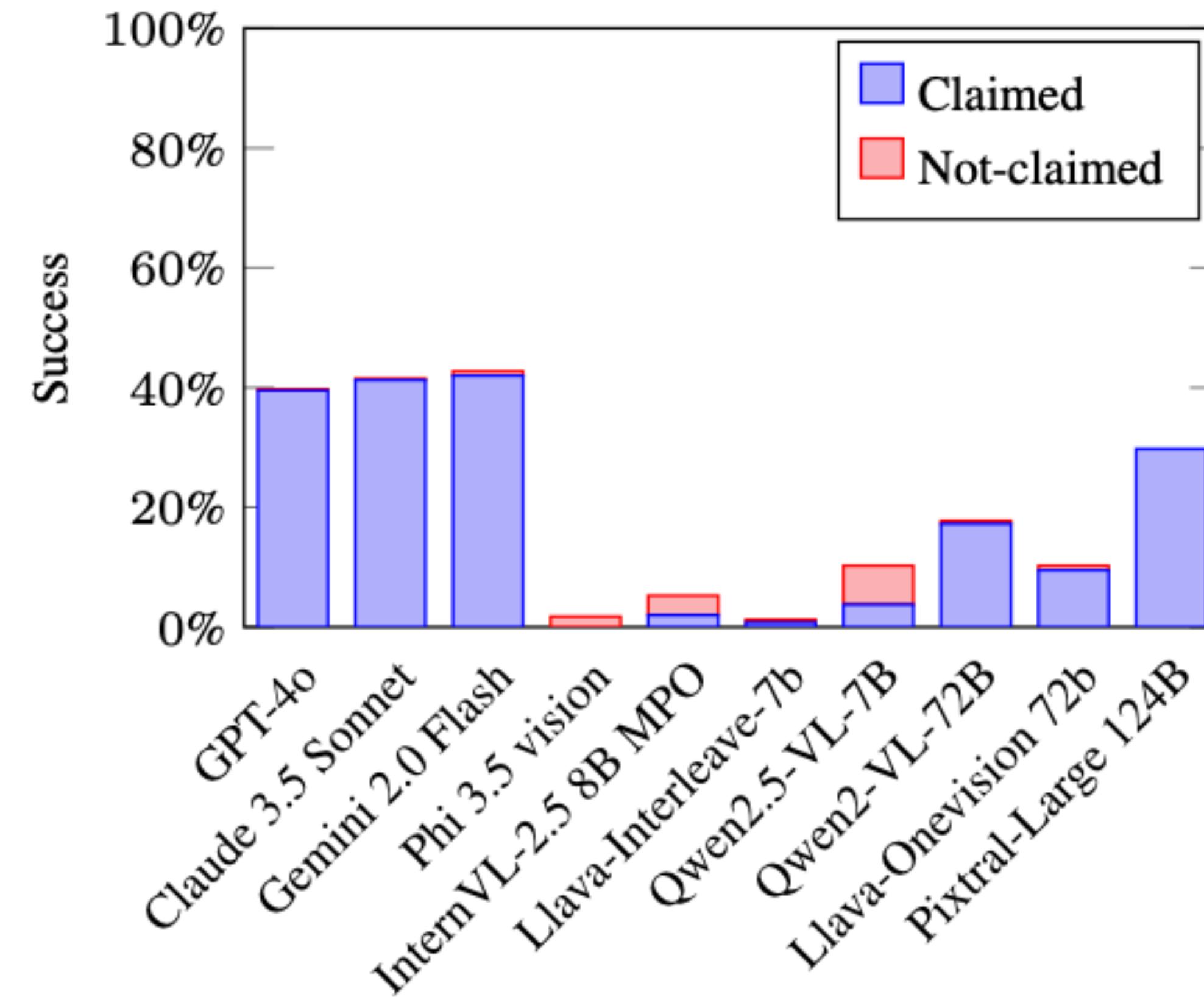
Current altitude: 104  
Collided on the last action: false  
Remaining moves: 20  
Status: ok

If the object is visible and the altitude is less than 10m above the target  
**FOUND** Or Else, move the UAV closer to the target  
 X  Y  Z  
**MOVE**

**REPORT ISSUE**

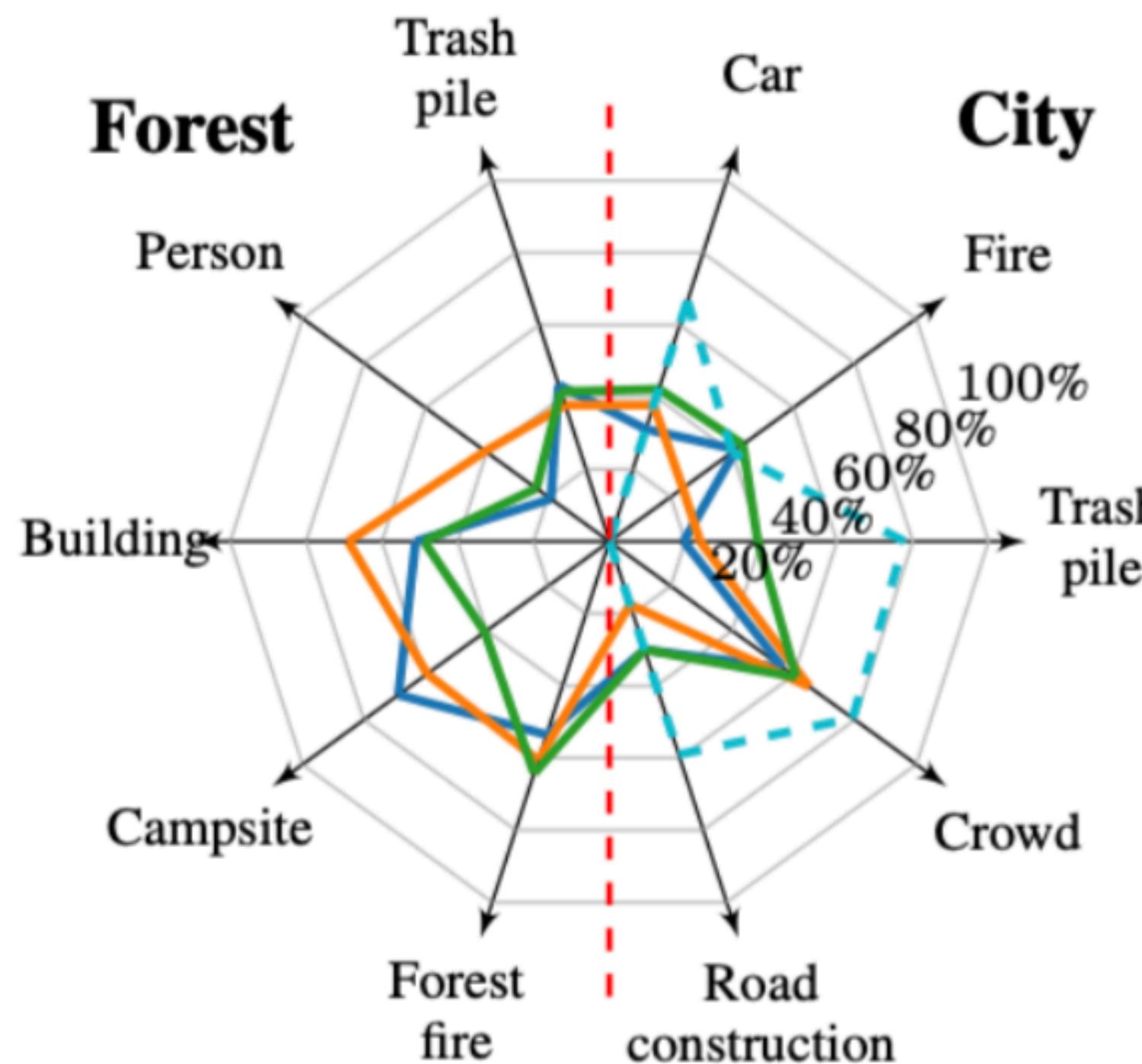
# Small LLMs struggle to format action

## Not claimed successes on FS-1

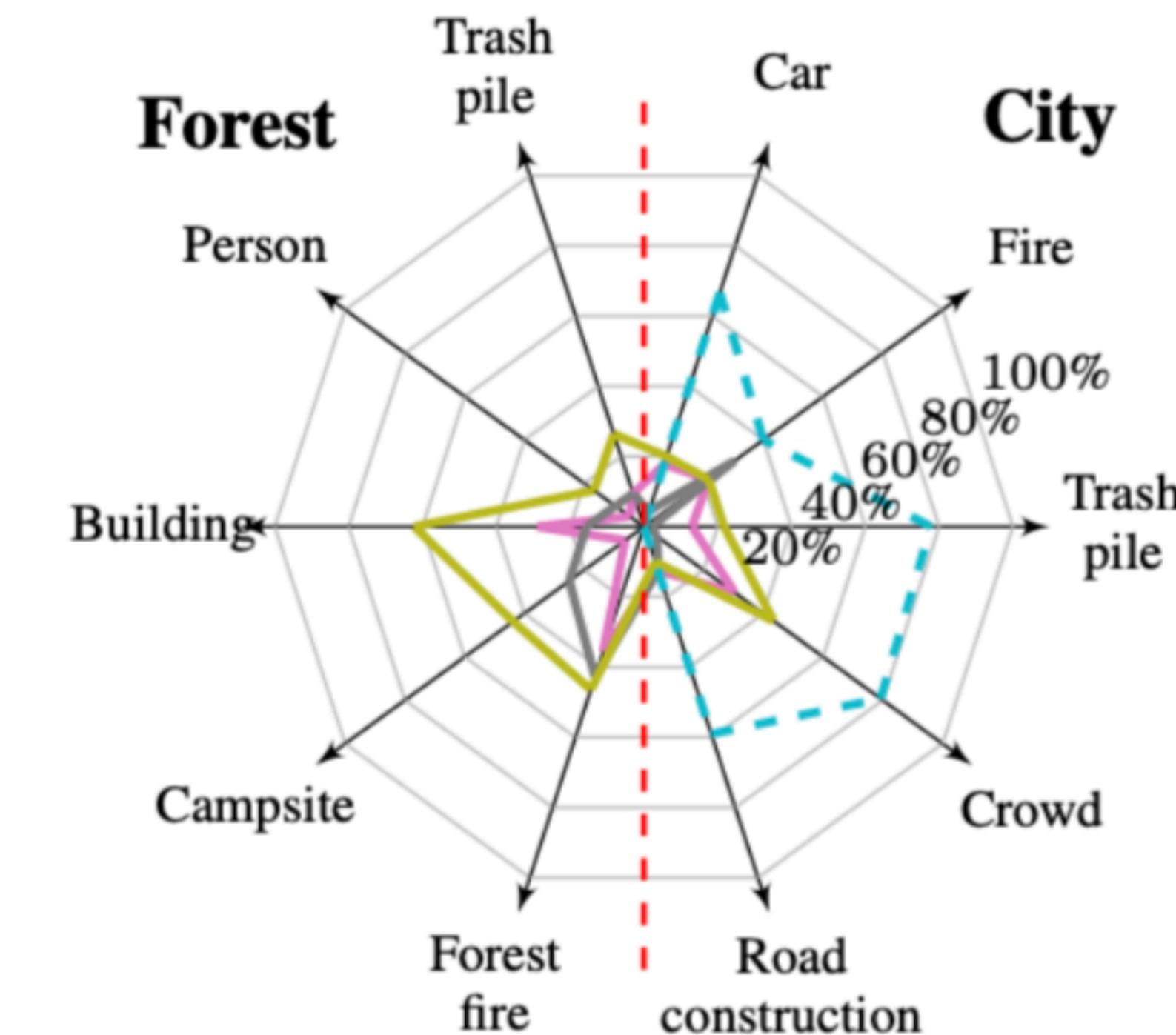


# FS-1 per class results

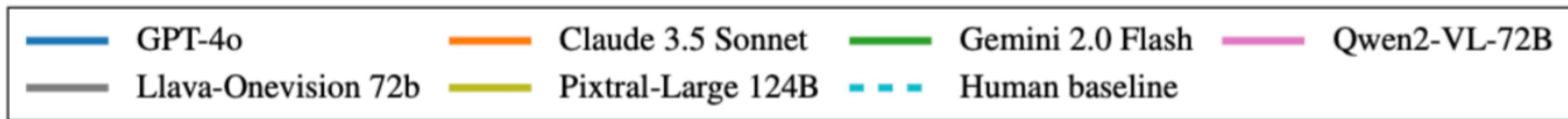
Some objects are harder than others



(a) Proprietary models



(b) Open source models



# FS-anomaly-1 results

Small models cannot locate anomalies

Model	FS-Anomaly-1		
	Overall (%)	Forest (%)	City (%)
GPT-4o	$27.0 \pm 3.1$	$39.0 \pm 4.9$	$15.0 \pm 3.6$
Claude 3.5 Sonnet	$27.5 \pm 3.2$	$37.0 \pm 4.9$	$18.0 \pm 3.9$
Gemini 2.0 flash	<b><math>35.5 \pm 3.4</math></b>	<b><math>46.0 \pm 5.0</math></b>	<b><math>25.0 \pm 4.4</math></b>
Phi 3.5 vision	$0.0 \pm 0.0$	$0.0 \pm 0.0$	$0.0 \pm 0.0$
InternVL-2.5 8B MPO	$3.5 \pm 1.3$	$6.0 \pm 2.4$	$1.0 \pm 1.0$
Llava-Interleave-7b	$0.0 \pm 0.0$	$0.0 \pm 0.0$	$0.0 \pm 0.0$
Qwen2.5-VL 7B	$2.8 \pm 1.2$	$3.7 \pm 2.1$	$2.0 \pm 1.4$
Qwen2-VL-72B	$7.5 \pm 1.9$	$10.0 \pm 3.0$	$5.0 \pm 2.2$
Llava-Onevision 72b	$8.5 \pm 2.0$	$11.0 \pm 3.1$	$6.0 \pm 2.4$
Pixtral-Large	$15.0 \pm 2.5$	$26.0 \pm 4.4$	$4.0 \pm 2.0$

# Ablations

## FS-1 and FS-Anomaly-1

	Setting	City		Forest		Overall	
		Gemini	Pixtral	Gemini	Pixtral	Gemini	Pixtral
FS-1	5 steps limit	34.5%	15.5%	41.5%	33.5%	38.0%	24.5%
	<b>10 steps limit (baseline)</b>	<b>41.5%</b>	<b>21.5%</b>	42.5%	<b>38.0%</b>	<b>42.0%</b>	<b>29.8%</b>
	20 steps limit	33.5%	13.5%	<b>45.5%</b>	36.0%	39.5%	24.8%
FS-Anomaly-1	<b>Searching for an anomaly (baseline)</b>	25.0%	4.0%	46.0%	26.0%	35.5%	15.0%
	Searching for explicit object types	<b>34.0%</b>	<b>7.0%</b>	<b>59.0%</b>	<b>34.0%</b>	<b>46.5%</b>	<b>20.5%</b>

	Setting	City		Forest	
		Gemini	Pixtral	Gemini	Pixtral
FS-1	<b>Baseline</b>	<b>41.5%</b>	<b>21.5%</b>	<b>42.5%</b>	<b>38.0%</b>
	Compass actions	17.5%	21.0%	17.5%	22.0%
	No grid overlay	17.0%	15.5%	31.5%	20.0%

# **Can we improve the result by fine-tuning?**

## **Zero-shot benchmark and fine-tuning**

- Train on Forest, evaluate on City
  - No object overlap, vastly different environments
- LLM online training with a simulator is another level of complication
- Supervised fine-tuning gives negligible improvement on this task

# GRPO fine-tuning on FlySearch

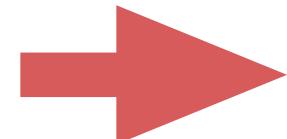
## Group Relative Policy Optimisation

- A dataset of pre-recorded Forest exploration trajectories
- For each trajectory we create a conversation completion problem
- Reward:
  - Reasoning score
  - Action score (move/found)
- GRPO fine-tuning of Qwen 7B with LoRA and frozen vision encoder

# FS-1 and FS-2 results

## Fine tuning

Model	FS-1			FS-2
	Overall (%)	Forest (%)	City (%)	Overall (%)
Human (untrained)	–	–	66.7 ± 4.5	60.8 ± 6.9
GPT-4o	39.5 ± 2.4	45.5 ± 3.5	33.5 ± 3.3	3.5 ± 0.9
Claude 3.5 Sonnet	41.2 ± 2.5	<b>52.0 ± 3.5</b>	30.5 ± 3.3	<b>6.5 ± 1.2</b>
Gemini 2.0 flash	<b>42.0 ± 2.5</b>	42.5 ± 3.5	<b>41.5 ± 3.5</b>	6.0 ± 1.1
Phi 3.5 vision	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	–
InternVL-2.5 8B MPO	2.0 ± 0.7	2.5 ± 1.1	1.5 ± 0.9	–
Llava-Interleave-7b	0.8 ± 0.4	0.0 ± 0.0	1.5 ± 0.9	–
Qwen2.5-VL 7B	3.8 ± 1.0	6.0 ± 1.7	1.5 ± 0.9	0.0 ± 0.0
Qwen2-VL 72B	17.2 ± 1.9	16.5 ± 2.6	18.0 ± 2.7	–
Llava-Onevision 72b	9.5 ± 1.5	12.5 ± 2.3	6.5 ± 1.7	–
Pixtral-Large	29.8 ± 2.3	38.0 ± 3.4	21.5 ± 2.9	3.0 ± 0.8
Qwen2.5-VL 7B, GRPO on Forest	–	57.0 ± 3.5	27.0 ± 3.1	0.0 ± 0.0



[The benchmark](#)   [Getting started](#)   [User guide](#)   [Internals](#)   [Examples](#)

# FlySearch: Exploring how vision-language models explore

Adam Pardyl, Dominik Matuszek, Mateusz Przebieracz, Marek Cygan, Bartosz Zieliński, Maciej Wołczyk

[Paper](#)[Code](#)



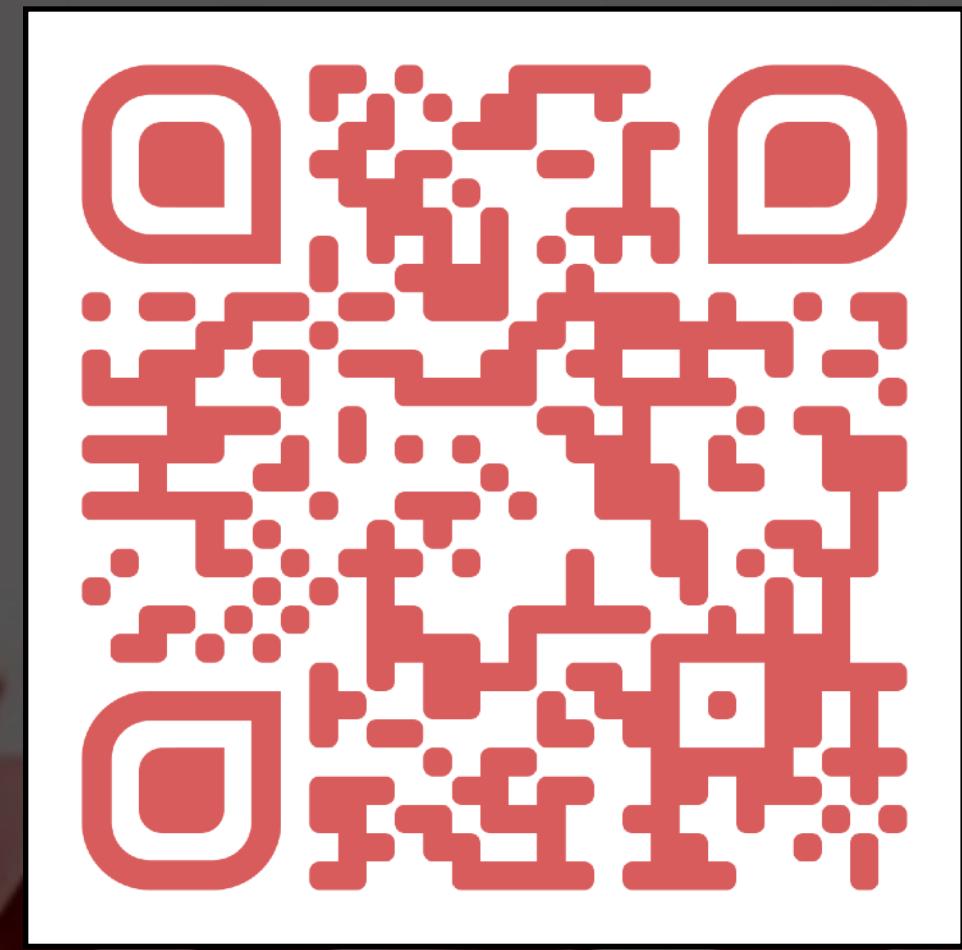
## Leaderboard

Model	FS-1		FS-Anomaly-1	FS-2
	Forest (%)	City (%)	Overall (%)	City (%)
Human (untrained)	--	66.7 ± 4.5	--	60.8 ± 6.9
Gemini 2.0 flash	42.5 ± 3.5	<b>41.5 ± 3.5</b>	<b>35.5 ± 3.4</b>	6.0 ± 1.1
Claude 3.5 Sonnet	<b>52.0 ± 3.5</b>	30.5 ± 3.3	27.5 ± 3.2	<b>6.5 ± 1.2</b>
GPT-4o	45.5 ± 3.5	33.5 ± 3.3	27.0 ± 3.1	3.5 ± 0.9
Pixtral-Large	38.0 ± 3.4	21.5 ± 2.9	15.0 ± 2.5	3.0 ± 0.8
Qwen2-VL 72B	16.5 ± 2.6	18.0 ± 2.7	7.5 ± 1.9	--
Llava-Onevision 72b	12.5 ± 2.3	6.5 ± 1.7	8.5 ± 2.0	--
Qwen2.5-VL 7B	6.0 ± 1.7	1.5 ± 0.9	2.8 ± 1.2	0.0 ± 0.0
InternVL-2.5 8B MPO	2.5 ± 1.1	1.5 ± 0.9	3.5 ± 1.3	--
Llava-Interleave-7b	0.0 ± 0.0	1.5 ± 0.9	0.0 ± 0.0	--
Phi 3.5 vision	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	--

# Acknowledgments

- This research was funded by National Science Centre, Poland (grant no. 2023/50/E/ST6/00469 and Sonata Bis grant no 2024/54/E/ST6/00388), lead by prof. Bartosz Zieliński.
- This paper has been supported by the Horizon Europe Programme (HORIZONCL4-2022-HUMAN-02) under the project "ELIAS: European Lighthouse of AI for Sustainability", GA no. 101120237.
- The research was supported by a grant from the Faculty of Mathematics and Computer Science under the Strategic Programme Excellence Initiative at Jagiellonian University.
- We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2024/017483.
- Some experiments were performed on servers purchased with funds from the Priority Research Area (Artificial Intelligence Computing Center Core Facility) under the Strategic Programme Excellence Initiative at Jagiellonian University.

[flysearch.gmum.net](http://flysearch.gmum.net)



# FlySearch

Exploring how vision-language models explore

Adam Pardyl, Dominik Matuszek, Mateusz Przebieracz,  
Marek Cygan, Bartosz Zieliński, Maciej Wołczyk

ML in PL Conference 2025



group of machine  
**gmum**  
learning research



**IDEAS**  
NCBR



nomagic