Your VLM/MLLM can solve your math homework, but can it spot a fire in the middle of a street?

Meet

FlySearch

a benchmark for active, goal-driven exploration,

evaluating:

3D spatial reasoning,

environment understanding,

visual context awareness,

and more!

See a demo at:

flysearch.gmum.net

FlySearch: Exploring how vision-language models explore

Adam Pardyl[1,2] Dominik Matuszek[1,2] Mateusz Przebieracz[2] Marek Cygan[3,4] Bartosz Zieliński[2] Maciej Wołczyk[1]

[1]IDEAS NCBR [2]Jagiellonian University [3]University of Warsaw [4]Nomagic

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation: exploring real open world

The real world is messy and unstructured.

Standard vision models struggle to generalize beyond simple scenarios.

Uncovering critical information requires active, goal-driven exploration.

It's not enough to simply look for the target, nor is it feasible to map everything in sight.

Vision-language models provide great zero-shot performance.

However, their abilities remain limited and largely untested in real-world scenarios.

Idea: evaluate VLM exploration skills in 3D open world scenarios

End goal: create a general exploration model.

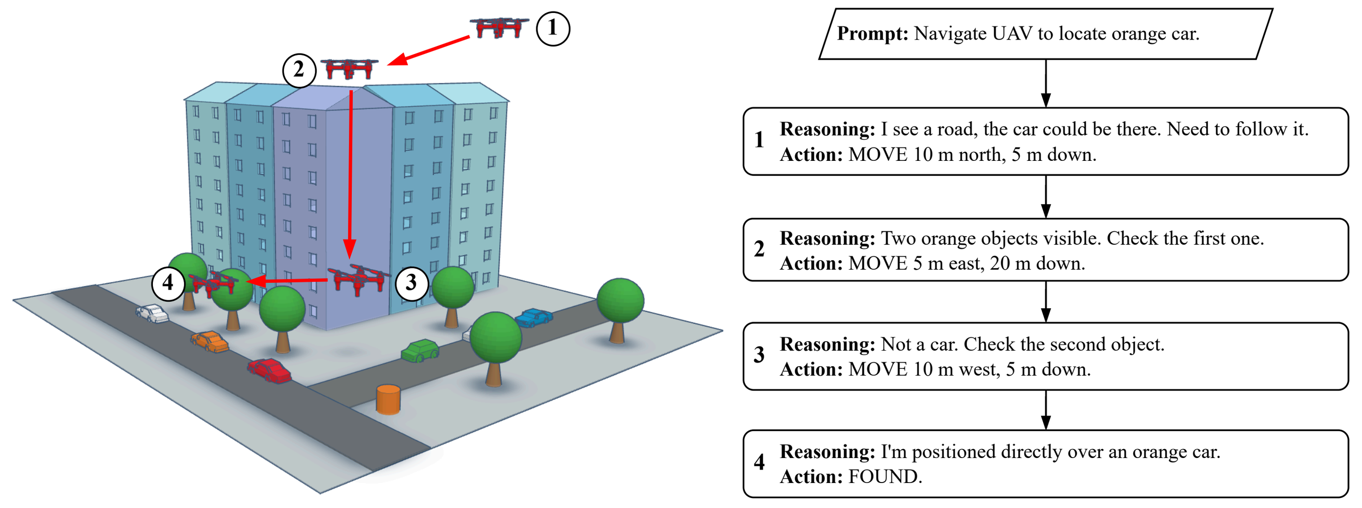
- Navigate in a real 3D open world.
- No fine-tuning or external help required.
- Understand the (visual) context.

How to evaluate such a method?

- Simulation based evaluation.
- Vision-language reasoning-based challenges.
- Tasks requiring context awareness.

What we need: human-like intuition in navigation

Going beyond simple object detection and mapping tasks.



Solution: our new VLM/MLLM benchmark

FlySearch: A benchmark for active, goal-driven exploration. Easy to solve for humans, hard for state-of-the-art VLMs / MLLMs.

High-fidelity outdoor urban and natural environments.

3D photorealistic, procedurally generated simulation (UE5).

Designed for VLM/MLLM evaluation.

Simple controls: free-flying camera.

Objective given in natural language.

A human can solve it without guidance.

Two photorealistic, procedurally generated environments

Three standardized difficulty levels

FS-1:

- The object is within line of sight from the starting position (though it may be far away).
- The object is described by text (e.g. "a red sports car").

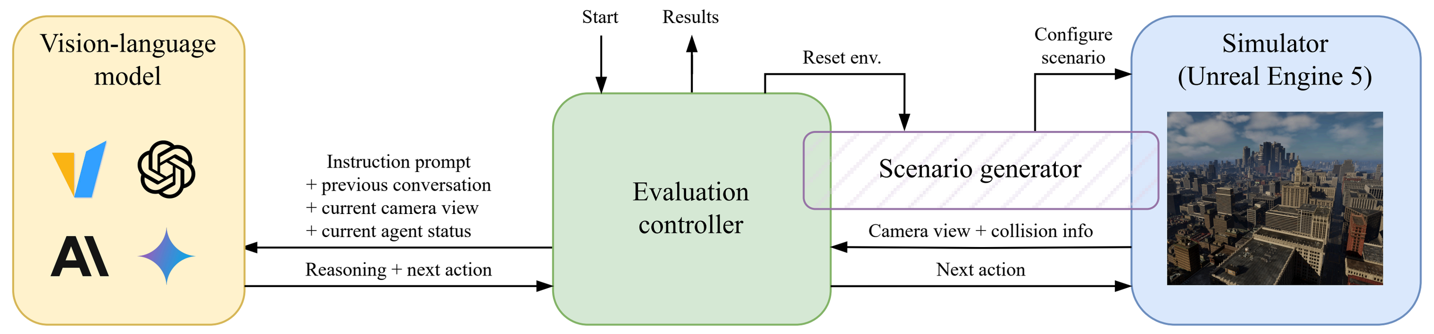
FS-Anomaly-1:

- The target object is an easy-to-spot anomaly (e.g. a flying saucer/UFO).
- The object description is *not* given to the model; the task is to look for an anomaly.

FS-2:

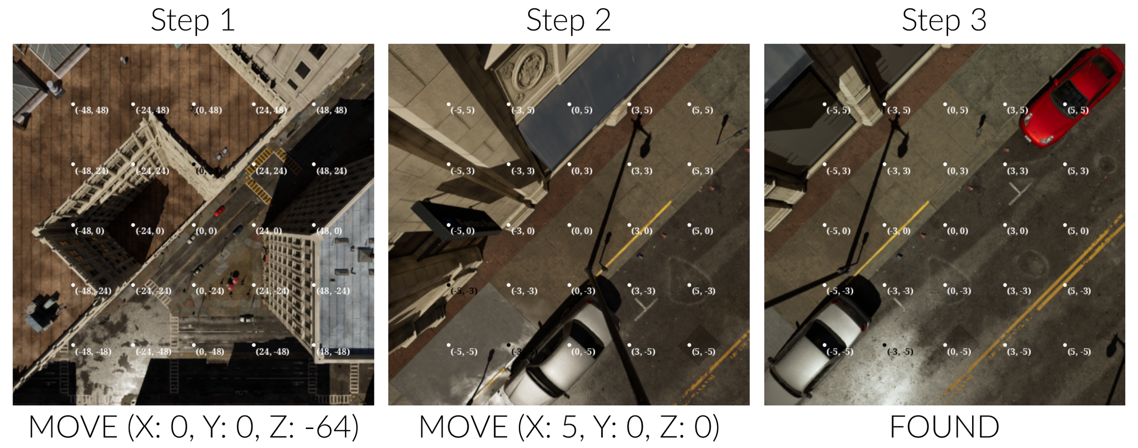
- The object can be hidden behind obstacles.
- Larger search area.
- The object is described by text and an image of a similar object is provided.

Evaluation pipeline



A detailed description of the evaluation pipeline is available in the paper.

Example exploration trajectory



Example of a successful trajectory in FS-1 performed by GPT-4o. Note the presence of the grid overlay on images, which helps the model to compute the relative position of the object.

Scan the QR code on the left for a video demonstration.

Main results

Model	FS-1			FS-Anomaly-1	FS-2
	Overall (%)	Forest (%)	City (%)	Overall (%)	Overall (%)
Human (untrained)	–	–	66.7 ± 4.5	–	60.8 ± 6.9
GPT-5*	$\mathbf{47.7 \pm 7.3}$	$\mathbf{57.3 \pm 6.9}$	38.1 ± 7.6	$\mathbf{44.5 \pm 4.6}$	5.2 ± 2.1
Gemini 2.0 flash	42.0 ± 2.5	42.5 ± 3.5	$\mathbf{41.5 \pm 3.5}$	35.5 ± 3.4	6.0 ± 1.1
Gemini 2.5 flash*	39.5 ± 7.7	47.4 ± 8.1	31.5 ± 7.4	38.0 ± 4.6	4.9 ± 2.9
Claude 4.5 Sonnet*	39.7 ± 7.0	55.1 ± 7.4	24.3 ± 6.6	35.5 ± 4.4	1.6 ± 1.6
Claude 3.5 Sonnet	41.2 ± 2.5	52.0 ± 3.5	30.5 ± 3.3	27.5 ± 3.2	$\mathbf{6.5 \pm 1.2}$
GPT-4o	39.5 ± 2.4	45.5 ± 3.5	33.5 ± 3.3	27.0 ± 3.1	3.5 ± 0.9
Pixtral-Large	29.8 ± 2.3	38.0 ± 3.4	21.5 ± 2.9	15.0 ± 2.5	3.0 ± 0.8
Qwen2-VL 72B	17.2 ± 1.9	16.5 ± 2.6	18.0 ± 2.7	7.5 ± 1.9	–
Llava-OneVision 72B	9.5 ± 1.5	12.5 ± 2.3	6.5 ± 1.7	8.5 ± 2.0	–
Qwen2.5-VL 7B	3.8 ± 1.0	6.0 ± 1.7	1.5 ± 0.9	2.8 ± 1.2	0.0 ± 0.0
InternVL-2.5 8B MPO	2.0 ± 0.7	2.5 ± 1.1	1.5 ± 0.9	3.5 ± 1.3	–
Llava-Interleave-7B	0.8 ± 0.4	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	0.0 ± 0.0	–

Success rates (± standard errors) of the evaluated models for the FS-1, FS-Anomaly-1 and FS-2 challenges.

* denotes results added after the submission deadline. The most recent leaderboard is available at our website.

Fine-tuning baseline

Model	FS-1		FS-2
	Forest (%)	City (%)	Overall (%)
Qwen2.5-VL 7B	6.0 ± 1.7	1.5 ± 0.9	0.0 ± 0.0
Qwen2.5-VL 7B, GRPO on Forest	57.0 ± 3.5	27.0 ± 3.1	0.0 ± 0.0

Success rates of a model fine-tuned with GRPO on the Forest environment (but not the specific FS-1 scenarios) and evaluated on the City environment.

More experiments and results are available in the paper.

Join the challenge!

If you would like to submit your agent to the FlySearch leaderboard, please check our webpage:

flysearch.gmum.net

or scan the QR code on the left.

We accept submissions of both standard VLMs/MLLMs and agentic frameworks.

Acknowledgments

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. 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). 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.

IDEAS NCBR

JAGIELLONIAN UNIVERSITY IN KRAKÓW

group of machine learning research gmum

UNIVERSITY OF WARSAW

nomagic