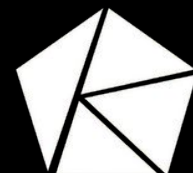


IMAV 2026 - Strasbourg

VLN on the Fly: An Onboard Vision-Language Navigation Stack for Aerial Robots

Marco S. Tayar, Felipe Tommaselli, Gianluca Capezutto, Pedro Antonio Rabelo Saraiva, Pedro H. V. de Freitas, Lucas Kido, Guilherme Sonogo, Ricardo V. Godoy, Marcelo Becker

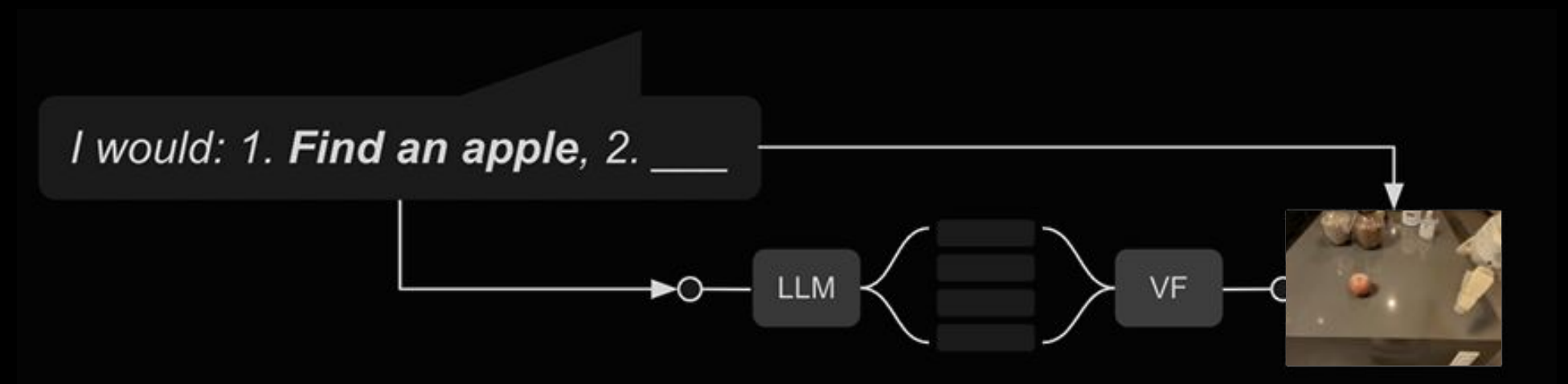


MOBILE ROBOTICS
GROUP

How to turn multimodal LLM reasoning into actions for the robot?



[1] Brown et al., NeurIPS, 2020.



[2] Ahn et al., CoRL, 2022.

VLAAs addresses directly those needs. But a single end-to-end policy often makes failures hard to localize and safety constraints harder to enforce.

Fused policy

Instruction and observations
produce an action

Failure attribution becomes difficult

Explicit handoffs

Image region	3D goal
Planned path	Flown motion

Each stage leaves information that we
can inspect

Say it, and it flies onboard

MOTIVATION

Why onboard aerial VLN remains difficult

End-to-end aerial policies

- Compact policy interface
- Limited failure attribution
- Safety checks remain hard to inspect

Modular onboard stack

- Observable intermediate goals
- Reusable geometric safety guards
- Modules can change without full retraining

Say it, and it flies onboard

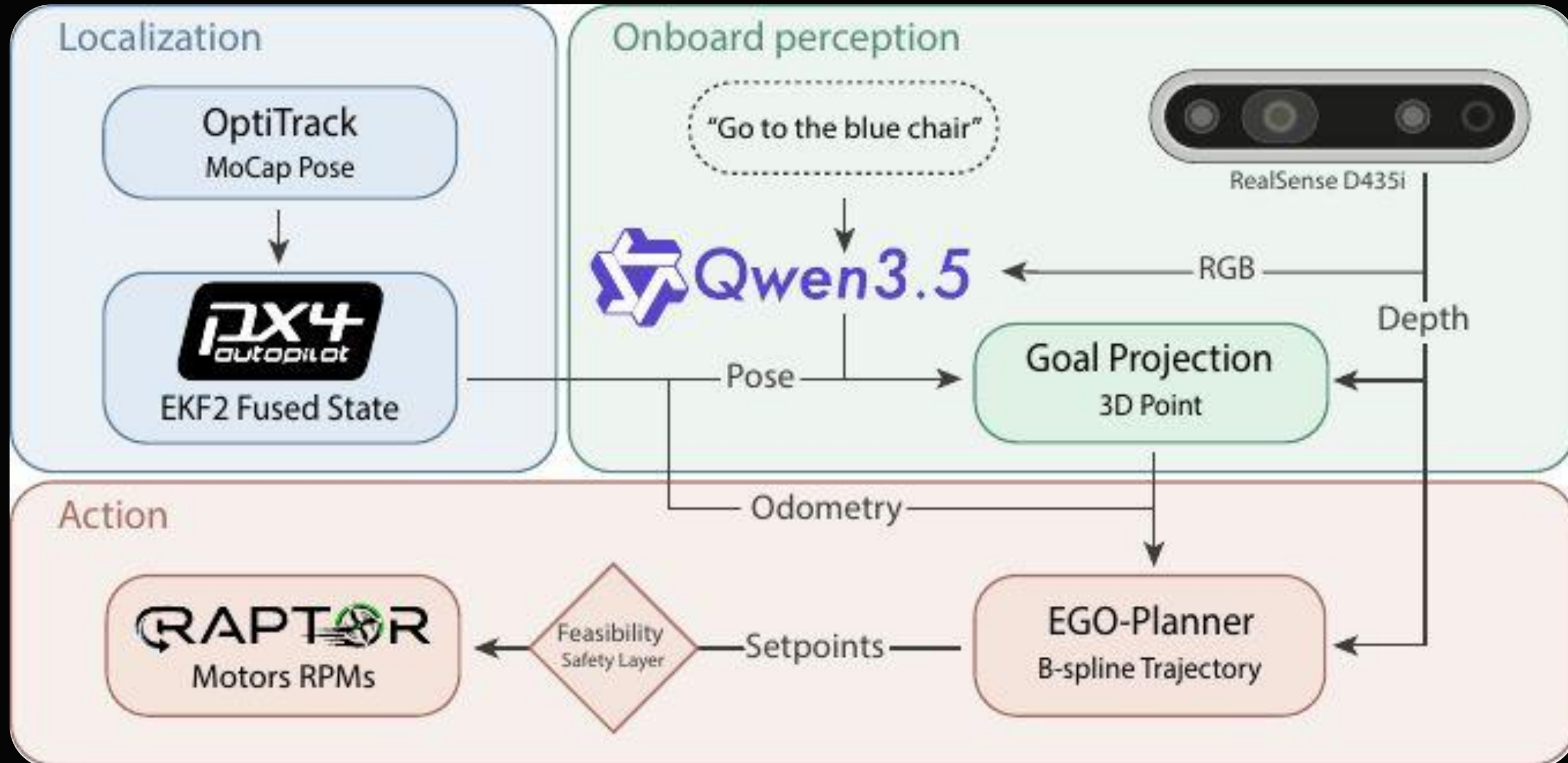
MOTIVATION

Why onboard aerial VLN remains difficult

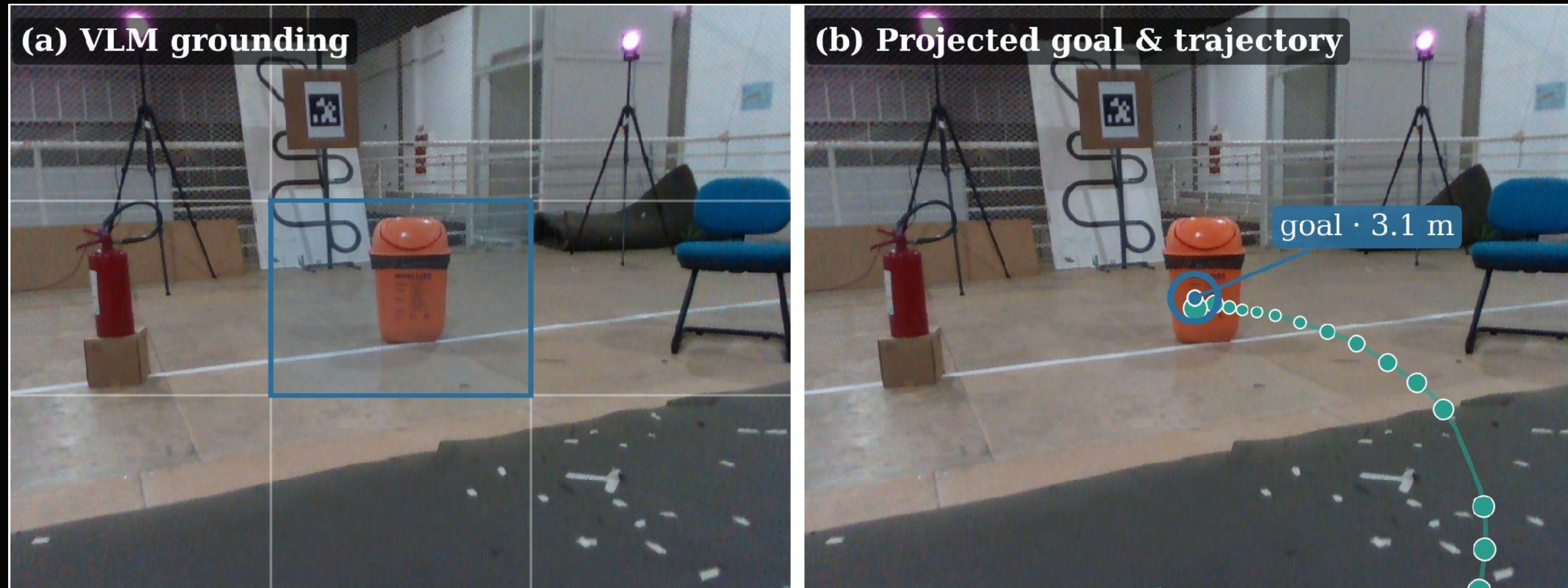
Indoor flight combines semantic and physical constraints

- GPS-denied operation and cluttered geometry
- Natural-language intent must become a metric 3D goal
- Grounding, planning, and control share limited compute
- A single-stage error can cause an unsafe command

One onboard stack, with explicit handoffs



Language grounding becomes a 3D goal



- 1** VLM selects one of nine image regions
- 2** Median valid depth in the chosen cell
- 3** Back-project to a bounded map goal

Qwen3.5-2B, INT4

A safety supervisor gates the VLM output

Before planning

Valid depth in the selected cell

Goal inside the operating volume

Three confident, consistent predictions

During execution

One active goal until completion or failure

Planner checks occupancy and trajectory feasibility

Invalid proposals do not trigger motion

EGO-Planner

Local occupancy map from depth and pose

3D B-spline with collision and motion limits

Replans while the vehicle moves

RAPTOR

Tracks position setpoints with a pretrained foundation policy

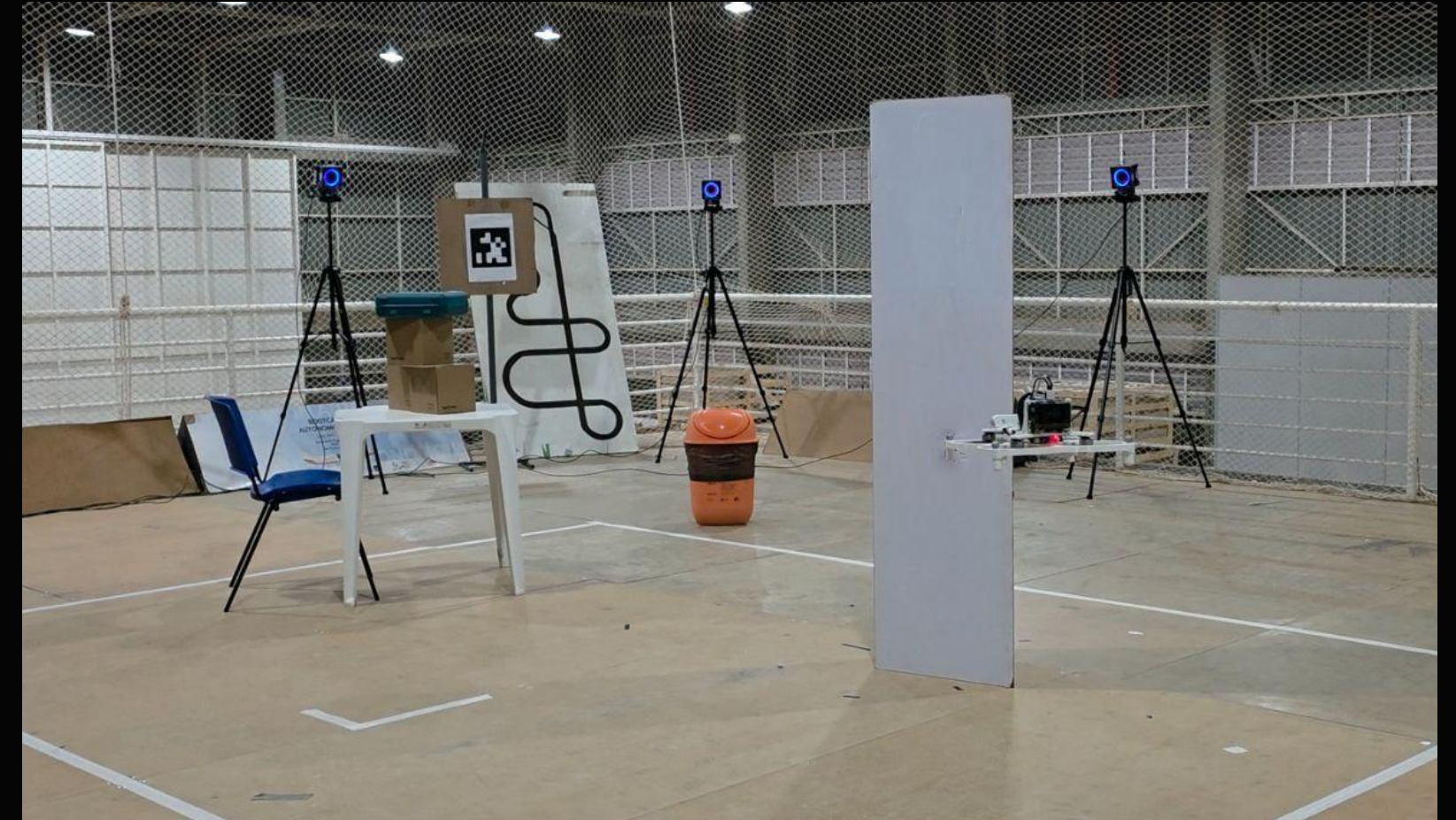
2,084 parameters on the flight controller

No retraining for the MIRA airframe



15 open-volume flights

3 referents × 5 trials



6 additional cluttered flights

Same onboard stack, added obstacles

13 of 15 flights reached the named referent

87%

success

Trash bin

5 / 5

Chair

4 / 5

Fire extinguisher

4 / 5

5.72 cm

mean goal error

Two failures stopped short after the target left view

The VLM fits the onboard query schedule

0.79 s

median inference

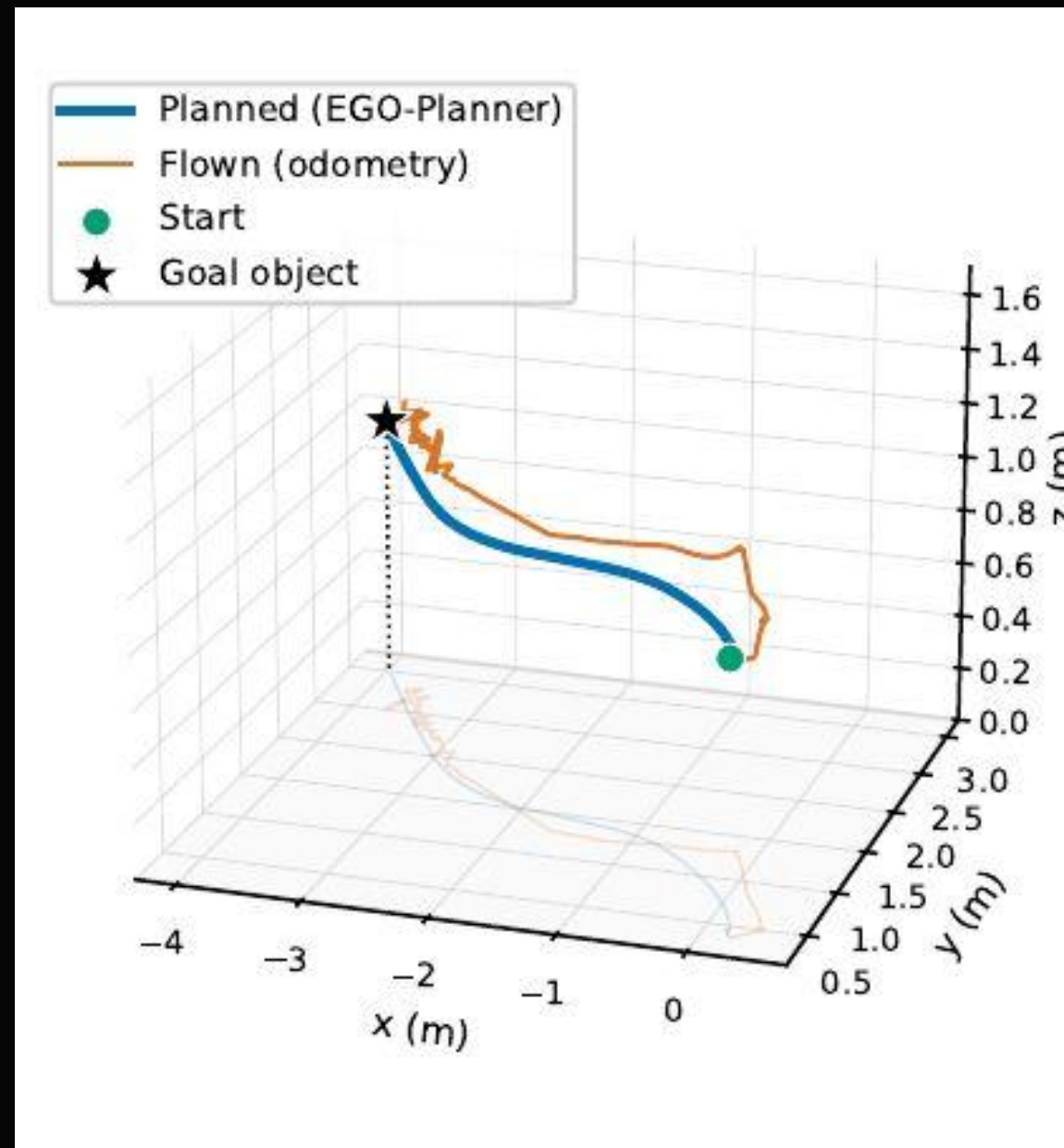
0.5 Hz

configured query rate

0.88 s at the 95th percentile

39.3% average GPU use

Planning and tracking continue between VLM queries



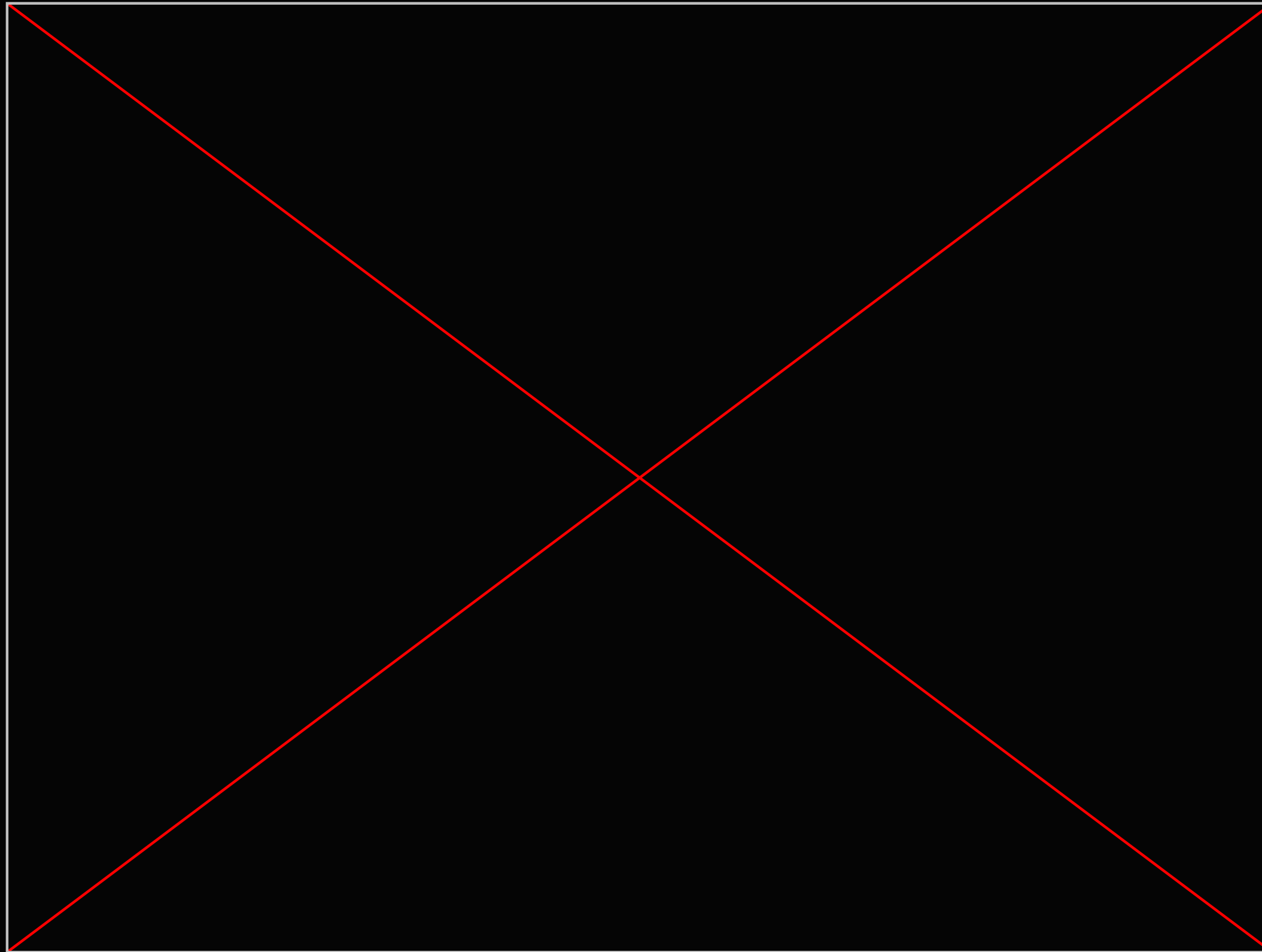
6 trials

3 reached the goal without collision

1 avoided obstacles; stopped at 20.7 cm

2 never received a valid goal

17.6 cm mean tracking error across the four executed runs



CONCLUSIONS

Contributions and Findings

- Fully onboard modular VLN is feasible on a small aerial robot
- **Safety** (i) planning guards can be incorporated into VLA-type of systems
- **Hierarchical** (ii) pipelines can be simplified with the use of foundation control policies
- **Onboard** (iii) is feasible given a model (and quantization) that matches the computational capacity
- Coarse grid, Speed benchmark, Usability, ...

Grounding

A named region
becomes a metric goal

Feasibility

13 / 15 successful flights
in the open volume

Diagnosis

Separate goal rejection
from tracking failure

- Current frame only, no spatial memory
- Accepted goal persists between updates
- Limited depth range and invalid samples
- External odometry and fixed altitude
- One room, known starts, few referents

- Scene memory and target reacquisition
- Asynchronous, confidence-aware goal refresh
- Complementary sensing and uncertainty-aware planning
- Onboard localization and full 3D evaluation
- More scenes, varied starts, longer tasks