

Mapping Intent: Agentic UAV Route Planning with Semantic Cost Maps

Jeng Wen Joshua Lean¹, Ting-Yu Yen¹, Kang-Chiao Lin², Xu-Peng He², Shih-Hsuan Hung¹, Hung-Kuo Chu¹

¹National Tsing Hua University, Hsinchu, Taiwan ²Industrial Technology Research Institute, Taiwan

joshualean@nthu.edu.tw tingyus995@gmail.com KCling@itri.org.tw

Sepp.Hehr@itri.org.tw hungsh@cs.nthu.edu.tw hkchu@cs.nthu.edu.tw

Abstract— Natural-language UAV missions combine terrain preferences with ordered visits and surveys. We present a vision-language agent that constructs semantic cost maps and route anchors, evaluates A* paths, and revises its decisions through visual feedback. In 27 missions with three repetitions per method, it returns all 81 routes and achieves a mean model-judged checklist violation rate of 14.7%, versus 45.8% for an OVerSeeC implementation. The system integrates semantic terrain modeling, ordered route planning, and visual revision in a unified workflow for compositional aerial missions.

(a) Uniform traversal cost

(b) Semantic cost map

(c) Cost-aware route

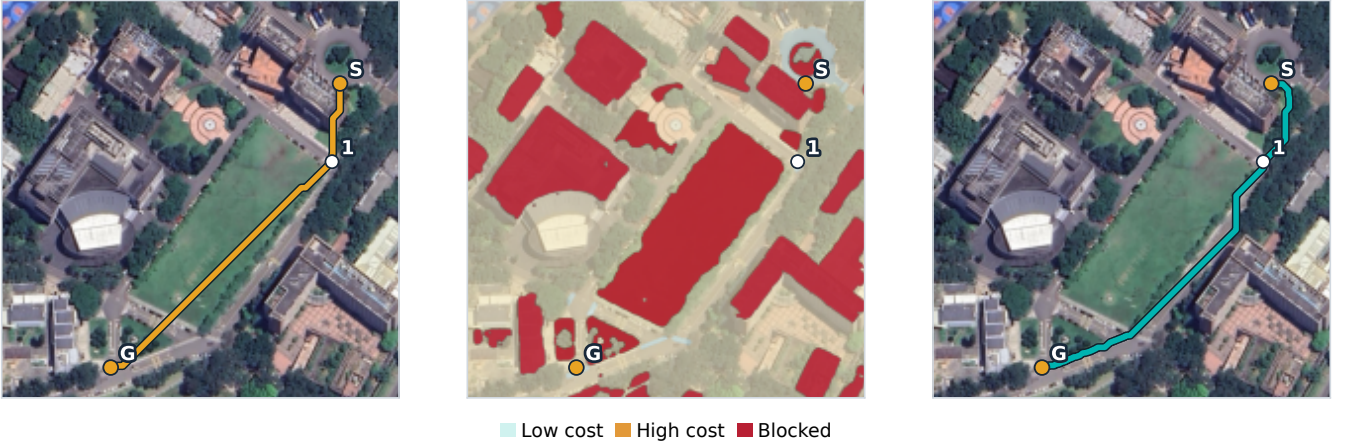


Fig. 1 Semantic costs encode how to travel. The mission requires avoiding roofs and maintained grass, preferring asphalt, and visiting crossing 1. With fixed anchors S, 1, and G, uniform-cost A* crosses the field (a); the recorded semantic map (b) guides A* around its boundary (c). Panel (a) is an illustrative control with no semantic blocks, excluded from benchmark results. Imagery: Google satellite tiles.

I. INTRODUCTION

Aerial missions couple ordered visits and surveys with terrain preferences. OVerSeeC constructs semantic cost maps [1]; UAV-VLPA* combines visual-language grounding, waypoint optimization, and A* [2]. ReAct [3] and Inner Monologue [4] motivate agent interaction through actions and feedback. We study this pattern for aerial route construction: a VLM revises *both* the terrain representation and the ordered anchors after inspecting evaluated paths. Our contribution is an implemented representation-revision workflow and its empirical evaluation on compositional missions.

II. METHOD

Agent state and actions. Given image I and instruction q , the system maintains a mission $\mathcal{M}$, semantic layers $\mathcal{L}_t$, ordered anchors A_t , and evaluated candidates $\mathcal{P}_t$. The VLM selects typed tool calls conditioned on I , q , and prior tool results and images. Seven tools set the mission, segment concepts, configure layers, derive masks, set anchors, evaluate a route, and finalize a candidate. The mission records ordered actions, hard constraints, soft preferences, and entities, with exact instruction spans checked before segmentation. These links make mission elements traceable to the original instruction.

Editable scene representation. SAM 3 [5] (SAM 3.1 checkpoint) supplies confidence masks for agent-selected concepts. Each layer has a threshold, weight $w_k \in [-1, 1]$, hard-block flag, and optional morphology. Set operations and boundary bands support composite regions such as shorelines.

Soft layers $\mathcal{S}$ yield

$$C(x) = \text{clip} \left(0.5 + \sum_{k \in \mathcal{S}} w_k \tilde{M}_k(x), 0.01, 1 \right), \quad (1)$$

where $\tilde{M}_k$ is confidence gated by the configured mask. Negative weights favor traversal; positive weights penalize it. The union of hard masks forms blocked set $\mathcal{B}$. Anchors $A_t = (a_0, \dots, a_m)$ encode visits and loops; supplied endpoints remain fixed.

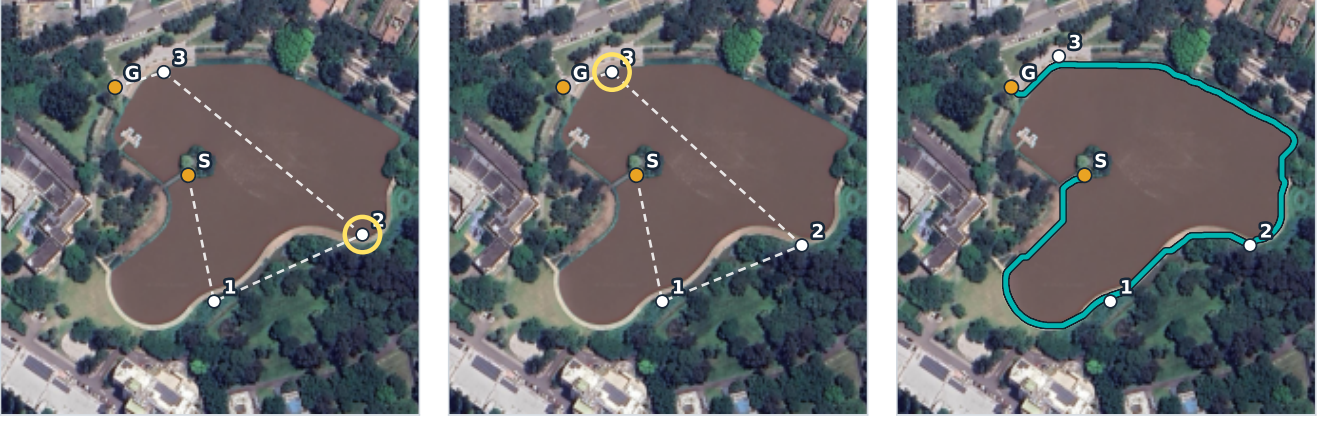
Geometric execution. For each consecutive anchor pair, eight-connected A* minimizes

$$J(P) = \sum_{(x,y) \in P} \|y - x\|_2 [1 + 9C(y)] \quad (2)$$

over grid paths outside $\mathcal{B}$, forbidding diagonal corner cutting. Concatenating these paths enforces the chosen anchor order. Optimality is conditional on the current map and anchors; the VLM determines those representations.

Feedback and termination. An evaluation returns a composite preview of masks, costs, anchors, and route, plus connectivity warnings and the remaining budget. The agent can re-query perception, change layer settings, or move anchors before re-evaluation. The benchmark permits six route evaluations per run. Finalization accepts only an evaluated, connected route with zero hard-mask crossings relative to its configured map. Figure 2 illustrates successive anchor revisions that restore route connectivity.

1 Eastern anchor blocks connection 2 North anchor still blocks connection 3 Connected route after anchor repair



Dashed: requested anchor sequence (no complete route). Solid: evaluated A* route. Red: blocked water mask.

Fig. 2 Visual feedback guides anchor revision. The mission requests a shoreline survey through the south, east, and north in order (instruction abridged). Moving the eastern anchor, then the northern anchor, connects the route on evaluation three without changing the layers. S/G mark endpoints. Imagery: Google satellite tiles.

III. EVALUATION

Protocol. Three locations, three image scales, and three difficulty levels yield 27 tasks. Three repetitions per method produce 243 runs. All methods receive the same aerial images, instructions, and fixed endpoints.

Ours and the local OverSeeC implementation use GPT-5.6 Luna with medium reasoning effort. Our UAV-VLPA* implementation uses Molmo-7B to extract waypoints, which are connected directly. Evaluation uses reviewed reference routes and mission-specific requirement checklists.

Metrics. Checklist violation rate (CVR) includes hard and soft requirements. Terrain regret (RRPI) averages $r(x) - 1$ over sampled path pixels, where rank 1 is preferred. Normalized mean nearest-path distance (D) is computed in the predicted-to-reference direction by averaging nearest-path distances and dividing by the image diagonal. Failures receive $CVR = 1$ and no geometric score.

TABLE I
PRELIMINARY COMPARISON. MEANS OVER 81 RUNS PER METHOD;
GEOMETRIC SCORES OMIT FAILED RUNS.

Method	Output $\uparrow$	CVR $\downarrow$	RRPI $\downarrow$	D $\downarrow$
OverSeeC	70/81	45.8%	2.035	.0535
UAV-VLPA*	81/81	40.3%	2.262	.0515
Ours	81/81	14.7%	1.328	.0206

Output counts returned routes, not full mission success. CVR uses all 81 runs; RRPI and D use 70 for OverSeeC and 81 for the other methods.

Findings. Ours returns routes in all runs and its mean checklist violation is 31.1 percentage points lower than OverSeeC’s (Table I). Routes are longer on average: 221.9 m versus 142.7 m for OverSeeC and 164.0 m for UAV-VLPA*. Length alone cannot establish success when missions require visits and surveys.

Twenty-five of our 81 runs use multiple evaluations (mean: 1.49). These traces illustrate how visual feedback supports anchor revision and recovery of route connectivity (Fig. 2).

IV. DISCUSSION AND OUTLOOK

The results support further study of jointly editable semantic maps and ordered anchors for aerial missions. The evaluation covers three locations, heterogeneous baselines, and model-assisted references and judging. Its scope is static 2D route planning; geometric metrics complement checklist-based assessment of mission requirements. Future work will isolate revision’s contribution through a matched single-pass ablation, expand independent human assessment, and evaluate physical flight execution.

ACKNOWLEDGMENT

This work was conducted in collaboration with the Industrial Technology Research Institute (ITRI). The project was funded in part by the National Science and Technology Council of Taiwan (114-2221-E-007-114-MY3, 113-2221-E-007-102-MY3, 114-2221-E-007-115-MY3).

REFERENCES

- [1] R. Rana et al., “OverSeeC: Open-Vocabulary Costmap Generation from Satellite Images and Natural Language,” arXiv:2602.18606, 2026.
- [2] O. Sautenkov et al., “UAV-VLPA*: A Vision-Language-Path-Action System for Optimal Route Generation on a Large Scales,” arXiv:2503.02454, 2025.
- [3] S. Yao et al., “ReAct: Synergizing Reasoning and Acting in Language Models,” in *ICLR*, 2023.
- [4] W. Huang et al., “Inner Monologue: Embodied Reasoning through Planning with Language Models,” arXiv:2207.05608, 2022.
- [5] N. Carion et al., “SAM 3: Segment Anything with Concepts,” arXiv:2511.16719, 2025.