Towards Semantic-Aware Aerial Object Retrieval: A VLM-Augmented Behavior Tree Approach

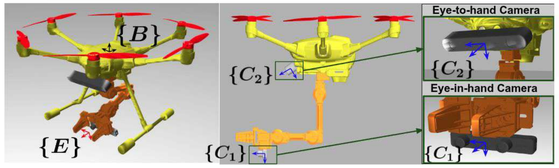
Mohammad Soltanshah, Abolfazl Eskandarpour, Kamal Gupta, Mehran Mehrandezh

SFU

Robotics Algorithms & Motion Planning Lab

Motivation & Problem

- **Aerial manipulator:** Aerial vehicle + Robotic arm
- **Autonomous high-level decision-making** for:
  - Environmental monitoring, inspection, and object retrieval
- **Behavior Trees (BTs)** are modular, reusable, and reactive but assume deterministic conditions.
- **Vision-Language Models (VLMs)** provide semantic understanding but are uncertain for direct robotic decisions.
- **Challenge: How can we safely use VLM semantic reasoning inside BTs?**



Phases for Autonomous Aerial Object Retrieval Mission:

Phase	Description	Phase	Description
0	Initialize Mission	4	Grasp Execution
1	Target Approach with Eye-to-hand	5	Lift and Transport
2	Target Approach with Eye-in-hand	6	Object Placement
3	Pre-Grasp Alignment	7	Return To Home

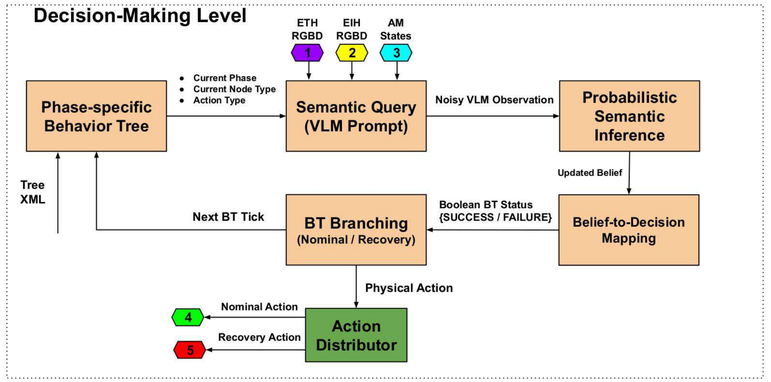
Related Work & Research Gap

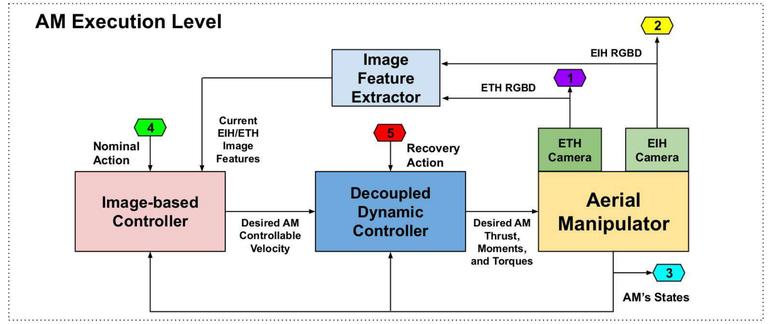
Research Gap: No prior framework combines VLM semantic reasoning with belief-space execution in Behavior Trees.

In this work, VLM outputs are treated as noisy observations rather than deterministic facts.

Existing Approaches	Limitations
BTs [1,2]	Deterministic conditions
VLM-based [3]	Uncertain and uncalibrated outputs
BT+LLM/VLM systems [4,5]	Do not explicitly model semantic uncertainty
Proposed	Explicit semantic uncertainty modeling

Proposed Framework





BT Branching:

- VLM condition nodes, VLM action nodes
- Bayesian belief update
- Multimodal inputs (ETH image, EIH image, robot state)
- Belief-space decision making

Probabilistic Semantic Reasoning:

Semantic State

- Visibility
- Safe camera switching

$$\mathbf{s}_t = [s_t^{(1)}, \ldots, s_t^{(K)}] \in \mathcal{S}$$

$$a_t \in \mathcal{A} = \mathcal{A}_{\text{phys}} \cup \mathcal{A}_{\text{recovery}}$$

$$\mathbf{z}_t = (\mathcal{I}_t, \mathbf{x}_t, \phi_t) \quad \text{multimodal input}$$

Confusion Models

- Constant FP/FN
- Input-conditioned FP/FN

$$C_{\mathbf{z}_t}^{(k)} = \begin{bmatrix} p(o_t^{(k)}=0 \mid s_t^{(k)}=0, \mathbf{z}_t) & p(o_t^{(k)}=1 \mid s_t^{(k)}=0, \mathbf{z}_t) \\ p(o_t^{(k)}=0 \mid s_t^{(k)}=1, \mathbf{z}_t) & p(o_t^{(k)}=1 \mid s_t^{(k)}=1, \mathbf{z}_t) \end{bmatrix} = \begin{bmatrix} 1-\epsilon_{\text{FP}}^{(k)}(\mathbf{z}_t) & \epsilon_{\text{FP}}^{(k)}(\mathbf{z}_t) \\ \epsilon_{\text{FN}}^{(k)}(\mathbf{z}_t) & 1-\epsilon_{\text{FN}}^{(k)}(\mathbf{z}_t) \end{bmatrix}$$

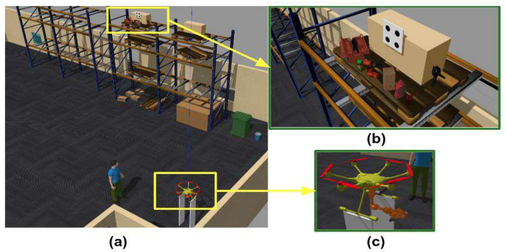
The input-conditioned false-positive and false-negative rates are defined as

$$\epsilon_{\text{FP}}^{(k)}(\mathbf{z}_t) = \bar{\epsilon}_{\text{FP}} \left(1 + \gamma_v^{(k)} \tilde{v}_t - \gamma_H^{(k)} \tilde{H}_t\right)$$

$$\epsilon_{\text{FN}}^{(k)}(\mathbf{z}_t) = \bar{\epsilon}_{\text{FN}} \left(1 + \gamma_v^{(k)} \tilde{v}_t - \gamma_H^{(k)} \tilde{H}_t\right)$$

Experimental Setup

- Simulation in Gazebo warehouse environment
- Aerial manipulator robot
- Eye-to-hand/Eye-in-hand cameras
- LLaVA-1.5-7B
- ROS + Groot



Results

BT variants:

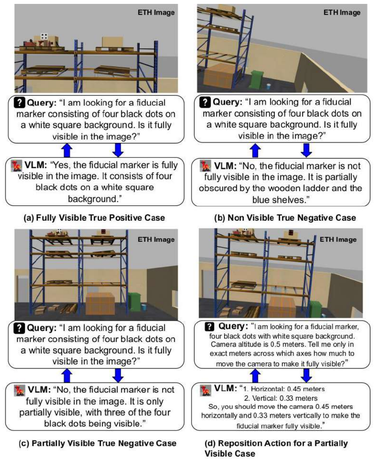
- Deterministic VLM-Augmented BT (DBT)
- Human-in-the-loop-Augmented BT (HIL-BT)
- Belief-space VLM-Augmented BT with fixed FP/FN (VLM-BT-C)
- Belief-space VLM-Augmented BT with Input-Dependent FP/FN (VLM-BT-ID)

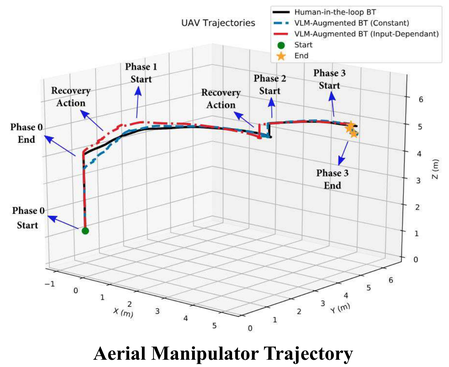
TABLE II: Constant confusion model results for Phase 1 over $N = 318$ image–prompt pairs, and Phase 2 over $N = 296$ image–prompt pairs, including TP/TN and FP/FN counts, computed FP and FN rates, and average VLM execution time.

Phase	TP / TN	FP / FN	$\bar{\epsilon}_{\text{FP}}$ / $\bar{\epsilon}_{\text{FN}}$ FP rate / FN rate	Average VLM Inference Time (s)
Phase 1	176 / 94	39 / 9	0.293 / 0.049	4.67
Phase 2	105 / 165	21 / 5	0.113 / 0.045	5.40

TABLE III: Comparison of Behavior Tree variants over 50 simulated missions.

Phase	Method	Success (%)	Time (s)	Average Recovery Actions
Phase 1	DBT	67	39.1	6.7
	HIL-BT	100	29.3	2.2
	VLM-BT-C	82	41.8	5.4
	VLM-BT-ID	96	35.4	3.1
Phase 2	DBT	71	35.4	3.5
	HIL-BT	100	23.3	1.3
	VLM-BT-C	84	36.7	2.6
	VLM-BT-ID	92	33.6	2.2





Aerial Manipulator Trajectory

- **29% improvement over DBT (Phase 1)**
- **Reduced recovery actions**
- **Reduced mission time**

References

[1] A. Calvo et al., "Mission planning and execution in heterogeneous teams of aerial robots supporting power line inspection operations," ICUAS, 2022.
[2] B. M. Rocamora et al., "A behavior tree approach for battery-aware inspection of large structures using drones," ICUAS), 2024.
[3] T. Wang et al., "Vision-language model-based human-guided mobile robot navigation in an unstructured environment for human-centric smart manufacturing," Engineering, 2025.
[4] H. Zhou et al., "Llm-bt: Performing robotic adaptive tasks based on large language models and behavior trees," in 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 16 655–16 661.
[5] N. Wake et al., "Vlm-driven behavior tree for context-aware task planning," arXiv preprint arXiv:2501.03968, 2025.