2026 NDIA MICHIGAN CHAPTER
GROUND VEHICLE SYSTEMS ENGINEERING
AND TECHNOLOGY SYMPOSIUM
***AUTONOMY AND ROBOTICS* TECHNICAL SESSION**
AUGUST 11-13, 2026 - NOVI, MICHIGAN

HYBRIDNAV: A HYBRID APPROACH TO ZERO-SHOT OBJECT DETECTION WITH ADAPTIVE EXPLORATION FOR AUTONOMOUS RECONNAISSANCE

Hemanth Indurthi[1], Eric Martinson, PhD[1]

[1]Lawrence Technological University, Southfield, MI

ABSTRACT

Autonomous reconnaissance in unknown or contested environments demands robust perception systems capable of identifying diverse objects without prior training data. This paper presents HybridNAV, a hybrid framework that combines multiple foundation models with an adaptive navigation system for zero-shot object detection and autonomous exploration. Unlike monolithic detection models, HybridNAV's multi-model fusion achieves balanced precision (0.60) and recall (0.58) with a macro F-score of 0.59, representing a 24% improvement over single-model baselines. The adaptive navigation system reduces scan time by 25% and path length by 30% compared to static waypoint approaches, while operating in real-time at 3.2 Hz with sub-100 msec latency on resource-constrained hardware. All processing is performed locally on the robotic platform, eliminating reliance on external communication infrastructure—a critical requirement for operations in communication-denied environments. We evaluate HybridNAV in both simulated indoor scenes and physical robot trials, demonstrating its effectiveness for intelligence gathering in unknown environments.

Citation: H. Indurthi, E. Martinson, "HybridNAV: A Hybrid Approach to Zero-Shot Object Detection with Adaptive Exploration for Autonomous Reconnaissance," In *Proceedings of the Ground Vehicle Systems Engineering and Technology Symposium* (GVSETS), NDIA, Novi, MI, Aug. 11-13, 2026.

1. INTRODUCTION

Autonomous ground vehicles deployed for reconnaissance and intelligence gathering must navigate unfamiliar environments and identify objects of interest without prior knowledge of the operational space. Unlike controlled industrial settings, reconnaissance environments are characterized by unpredictable layouts, diverse and potentially unfamiliar objects, variable lighting, and unreliable communication infrastructure limiting cloud-based processing. With these challenges, traditional object detection approaches that rely on supervised learning from extensive labeled datasets are impractical. Instead, what is needed are zero-

shot detection methods that require no hand labeled data to quickly switch targets when new intelligence and/or security threats become known.

Foundation models like Contrastive Language Image Pre-training (CLIP) and Visual Language Models (VLMs) address the challenges of zero-shot detection by leveraging large-scale pretraining on image-text pairs to recognize novel object categories from natural language descriptions alone. But their detection performance with arbitrary objects can vary wildly. Thankfully, individual VLMs exhibit complementary strengths and weaknesses. High-precision models like Grounding DINO [1] produce accurate localizations but may miss targets in cluttered scenes. High-recall models like CLIPSeg [2] cast a wider net but generate more false positives. Recent monolithic models such as SAM v3 have demonstrated impressive general-purpose segmentation capabilities, yet they can struggle with esoteric or domain-specific objects and do not inherently support spatial relationship reasoning—capabilities critical for reconnaissance tasks such as "identify the object near the doorway" or "locate unusual items on the floor."

This paper presents HybridNAV, a hybrid framework that fuses multiple foundation models with density-aware spatial reasoning and an adaptive navigation system. The key research question we address is: when does a hybrid multi-model pipeline provide advantages over monolithic detection systems for autonomous reconnaissance? Our contributions include: (1) a hybrid detection pipeline that fuses Grounding DINO, SAM v2, CLIPSeg, and DBSCAN clustering for robust zero-shot detection; (2) an adaptive navigation system using a priority-based FSM that dynamically balances object-centric investigation with frontier-based exploration; (3) communication-independent operation with all processing performed locally on resource-constrained hardware. All methods are evaluated comprehensively in both simulated benchmark and physical robot trials.

2. RELATED WORK

The evolution of foundation models has transformed zero-shot object detection by moving beyond fixed-label datasets. CLIP and its derivatives pioneered shared embedding spaces for images and language, enabling open-vocabulary recognition that is essential for unmapped reconnaissance environments. Extensions such as CLIPSeg [1] enable segmentation conditioned on natural language, while transformer-based models including Grounding DINO [2] and SAM v2 [3] improve spatial precision and mask quality. While recent iterations like SAM v3 demonstrate strong general-purpose segmentation, their effectiveness on domain-specific or esoteric objects—common in tactical scenarios—remains an active area of investigation.

New, more robust perception, is only option for improving object detection in autonomous reconnaissance; the vehicle can also optimize how the environment is navigated to maximize information gain. Frontier-based exploration [4] remains a foundational strategy for this task by identifying boundaries between mapped and unmapped regions to ensure systematic coverage. Yet, pure frontier exploration often ignores object-level priorities, a critical drawback when specific targets demand immediate investigation over general area mapping. To bridge this gap, recent research has begun integrating semantic understanding directly into navigation logic, building semantic or probabilistic solutions from open vocabulary object detection solutions like CLIP [5] [6].

This work, HybridNAV, also integrates OVOD into object search, leveraging the complementary strengths of multiple models

through a fusion architecture that balances high-precision localization with high-recall coverage and object centric investigation.

3. SYSTEM OVERVIEW

The perception system (Figure 1) implements open vocabulary object detection (OVOD) by processing synchronized RGB-D frames through two tightly coupled stages: a hybrid detection pipeline, followed by fusion and filtering. Simultaneously, a navigation system is monitoring the current detection state and moving the robot to investigate as appropriate.

3.1. Hybrid Detection Pipeline

First, Grounding DINO [2] receives the color image and a mission-defined prompt set (e.g., target object descriptions), returning high-precision bounding boxes for candidate objects. Each bounding box is enlarged by 10% and fed as a box prompt to SAM v2 [3], which produces pixel-accurate masks that preserve DINO's localization while eliminating boundary artifacts. In parallel, CLIPSeg [1] processes the same frame, yielding an open-vocabulary mask whose high recall compensates for potential DINO misses. The two complementary mask streams—sharp SAM-refined masks and generous CLIP masks—supply evidence for the subsequent fusion stage.

3.2. Spatial-Confidence Fusion and Density-Aware Filtering

The complementary masks from the DINO+SAM and CLIPSeg branches are merged through a three-pass fusion process. First, a spatial index identifies SAM pixels within 0.20 m of CLIPSeg detections. Second, overlapping points receive a confidence boost computed as a weighted average of both models' scores, multiplied by a 1.1 enhancement factor. Third, DBSCAN clustering [7] groups the fused 3D points into object instances, with a density-aware decay mechanism that penalizes sparse clusters while preserving informative evidence. Persistent object tracking assigns stable IDs across frames, enabling multi-view confirmation that reduces false positives over time.

Several alternative detection models were evaluated during development, including YOLO-World, LISA, and PointCLIPv2, but none matched the complementary precision-recall balance of the Grounding DINO + CLIPSeg combination for open-vocabulary reconnaissance tasks [10].

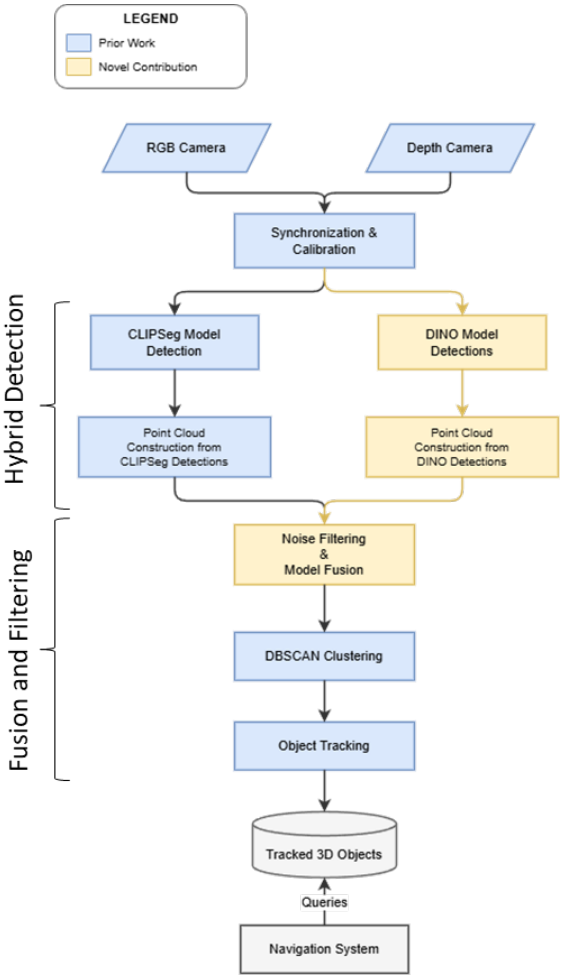

Figure 1. HybridNAV perception pipeline. Blue modules indicate prior work components; yellow modules indicate novel contributions. Parallel CLIPSeg and Grounding DINO branches produce complementary detections that are fused, clustered, and tracked before being passed to the navigation system.

3.3. Adaptive Navigation System

Because target priorities may shift during a mission as new intelligence becomes available, HybridNAV's navigation is governed by a finite state machine (FSM) whose decisions are driven by two live data products: the detected-objects database produced by the perception stack and the occupancy grid maintained by SLAM. Figure 2 illustrates the FSM and its transition conditions. Mission-defined target lists can be updated in the field without retraining, enabling rapid re-tasking in response to evolving intelligence.

The FSM comprises six behavioral states organized in a strict priority hierarchy. Object-driven investigation always takes precedence over frontier-based exploration, which in turn outranks waypoint coverage. This hierarchy ensures that high-value targets are addressed immediately while still guaranteeing systematic area coverage.

In the **Scan** state, the robot evaluates all entries in the detected-objects database using a scoring function:

$$score = \frac{confidence_{max}}{n+1}$$

Where $confidence_{max}$ is the maximum confidence among all points in the cluster and n is the number of images that could see the object cluster. This scoring function automatically favors confident yet under-observed objects.

If an object is detected, the FSM transitions to **Approach**, where the robot navigates toward the target. Upon arriving within a predefined distance, a full 360° pan-tilt scan is performed to capture comprehensive multi-view data for object confirmation. If the detection remains within range but the viewpoint count is still below three, the system transitions to **Circle**, where the robot orbits the target at a fixed radius, performing tilt-only scans at each stop point. These tilt-only scans are more efficient than full sweeps since the object's approximate bearing is already known.

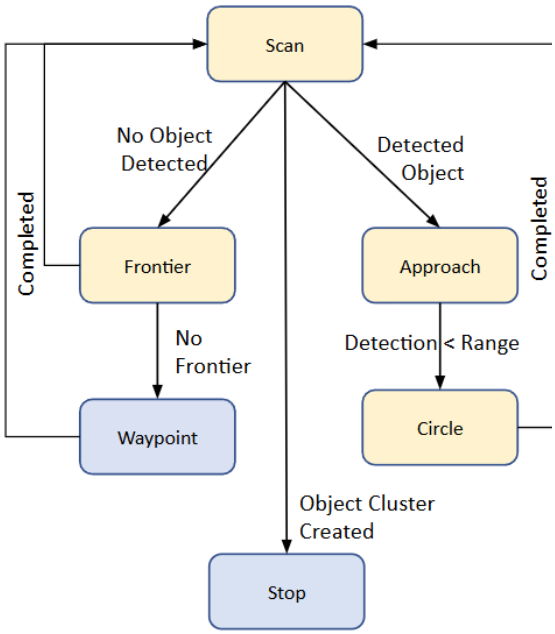

Figure 2. HybridNAV navigation finite state machine

Once the viewpoint threshold is met, the FSM returns to **Scan** to evaluate remaining targets. When a confirmed object cluster meeting all success criteria — sufficient confidence, point cloud quality, and viewpoint diversity — is created, the system enters the **Stop** state.

When no object is detected in **Scan**, the FSM transitions to **Frontier** exploration. The robot navigates to the nearest boundary between mapped and unmapped space and performs a full pan-tilt scan upon arrival to detect new objects in the previously unseen area. Once frontier exploration is completed, the FSM returns to **Scan**. If no frontiers remain, the **Waypoint** state activates a static tour sequence to ensure complete area coverage. Completion of the waypoint tour likewise returns the system to **Scan**, where any new detections may trigger object-driven investigation.

Motion between all states is governed by a potential-fields controller adapted from prior

work [12] that blends an attractive vector toward the current navigation goal with repulsive vectors derived from LIDAR readings and virtual scan rays synthesized from the occupancy grid. These virtual rays simulate range sensor output from the known map, enabling the robot to proactively avoid obstacles not yet visible to its physical sensors. If translational velocity drops below 0.05 m/s for two seconds during any navigation phase, a random ±10° rotation is injected to escape local minima.

4. EXPERIMENTAL SETUP

4.1. Platform

All experiments were conducted on a Hello Robot Stretch 2 [8] mobile manipulator (Figure 3) equipped with an Intel RealSense D435i RGB-D camera, a Hokuyo URG-04LX LIDAR, and onboard computation (Intel i7-8650U CPU with NVIDIA MX450 GPU). The system runs ROS on Ubuntu, with all perception and navigation modules executing locally without external compute dependencies.

Figure 3. Stretch-2 mobile robot used to evaluate HybridNAV

4.2. Evaluation Regimes

HybridNAV was evaluated in two complementary regimes: (1) an offline benchmark using ten ScanNet [9] indoor reconstruction scenes to isolate detection quality, and (2) real-time trials on the physical Stretch 2 platform in a controlled laboratory environment that exercised the full perception-to-navigation loop.

Detection performance was assessed using precision, recall, F-score, and average cluster density with IoU ≥ 0.50 as the matching threshold. The evaluation follows a greedy one-to-one matching procedure identical to the ScanNet evaluation protocol: predicted masks are sorted by confidence, and each prediction is declared a true positive if its semantic label matches the ground-truth label and the IoU with any unmatched ground-truth mask meets the threshold. Remaining unmatched predictions become false positives, and unmatched ground-truth instances become false negatives.

Navigation efficiency was measured via four metrics: total path length (sum of Euclidean distances between consecutive base poses from odometry), average navigation time (wall-clock duration from search start to Done state or timeout), total scan time per scene (cumulative time during scan actions, excluding driving), and estimated remaining battery percentage at mission end. These metrics capture both the speed and energy cost of reconnaissance operations.

For the offline ScanNet evaluation, performance was computed per class per frame, macro-averaged across classes, and then averaged across all ten scenes. For the real robot trials, the same averaging procedure was applied across ten separate trials in the laboratory environment. Baseline comparisons included CLIPSeg alone, OMDet with SAM v2, and Grounding DINO with SAM v2 (without CLIPSeg fusion), as well as a static waypoint navigation baseline for the navigation experiments.

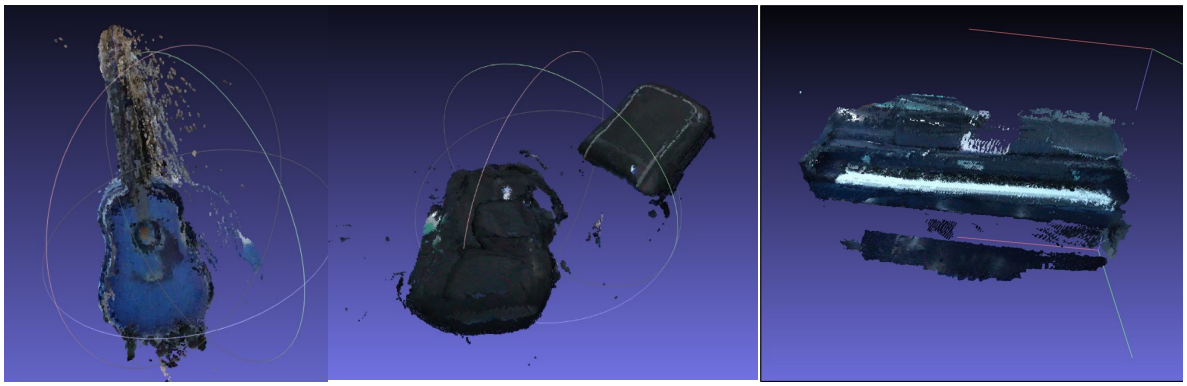

Figure 4. Sample HybridNAV point cloud detections from ScanNet scenes: guitar (left), backpack (center), and upright piano (right). Per-point color reflects detection confidence.

5. RESULTS

5.1. Detection Performance

Table 1 summarizes detection performance across the ten ScanNet scenes. Figure 4 highlights some detections from this same evaluation. HybridNAV's fusion achieved the highest F-score (0.59) with balanced precision (0.60) and recall (0.58), surpassing single-model approaches: CLIPSeg (F1 = 0.48), OMDet [10] (F1 = 0.46), and DINO + SAM v2 alone (F1 = 0.52). While OMDet posted the highest individual precision (0.65), it missed many targets, yielding the lowest recall (0.35)—a critical liability for reconnaissance where missed detections are more costly than false alarms.

***Table 1**. Detection performance metrics across models (ScanNet benchmark).*

Model	Prec.	Rec.	F1	Clust. Dens.
CLIPSeg	0.44	0.54	0.48	10.8
OMDet	0.65	0.35	0.46	14.5
DINO+SAM	0.58	0.48	0.52	12.2
HybridNAV	0.60	0.58	0.59	13.9

Stratified analysis by object distance and size revealed that HybridNAV particularly excelled on far (> 1 m) and small (< 0.25 m diagonal) objects—categories most sensitive to viewpoint diversity and most relevant to reconnaissance scenarios where targets may be partially obscured or at a distance. The persistent-ID tracking and multi-view confirmation enabled by the FSM's scan actions proved especially valuable for these challenging cases.

5.2. Navigation Efficiency

Table 2 compares HybridNAV's navigation performance against a static-waypoint baseline across the laboratory trials. HybridNAV reduced mean scan time from 25 s to 20 s (−25%) and total path length by 30%. The priority-based FSM ensured that high-value targets were investigated first, with frontier exploration filling gaps in coverage.

***Table 2**. Navigation efficiency comparison (Real robot trials).*

Metric	Waypoint	HybridNAV
Scan time (s)	25.0	20.0
Path length (rel.)	1.00	0.70
Frame rate (Hz)	N/A	3.2
Latency (ms)	N/A	< 100

In real-time trials, the system consistently identified and confirmed target objects while maintaining thorough area coverage. The FSM's priority hierarchy proved effective: when a high-confidence detection was made during frontier exploration, the robot immediately transitioned to approach and investigate the target before resuming area coverage. This behavior mirrors the desired reconnaissance pattern of "investigate then continue."

Energy efficiency was also favorable. The reduced path length and targeted scan behavior resulted in lower cumulative motor usage compared to exhaustive waypoint traversal, extending the effective mission duration on battery power—a practical consideration for field-deployed reconnaissance platforms.

5.3. System Performance

HybridNAV sustained 3.2 Hz throughput with end-to-end latency below 100 msec on the Stretch 2's onboard i7-8650U and MX450 GPU. CPU and GPU utilization were sampled at 1 Hz throughout operation, remaining within acceptable bounds for extended missions. These results demonstrate the feasibility of running the full multi-model pipeline on resource-constrained platforms without external compute support—a critical requirement for deployment in communication-denied environments.

6. CONCLUSION

A key finding of this work is that the hybrid multi-model approach provides the greatest advantage in scenarios involving: (1) esoteric or domain-specific objects that may fall outside the training distribution of large monolithic models; (2) situations requiring spatial relationship reasoning between detected objects; and (3) environments where multi-view confirmation is essential for reducing false positives. For common, well-represented object categories in controlled settings, monolithic models such as SAM v3 may offer comparable or superior performance with lower architectural complexity. The value proposition of HybridNAV is strongest in the open-ended, unpredictable conditions characteristic of reconnaissance operations.

6.1. Implications for Military Reconnaissance

The ability to specify targets via natural language prompts—rather than retraining a classifier—offers significant operational flexibility for military applications. Mission planners can update the target object list between deployments without requiring machine learning expertise or retraining infrastructure. For example, a reconnaissance mission in an urban environment might target "backpack," "wire bundle," or "electronics on floor" descriptions that can be changed in the field in response to new intelligence.

A communication-independent architecture ensures operation in GPS-denied and communication-contested environments, while the priority-based navigation naturally maps to reconnaissance doctrine: investigate high-priority targets first, then systematically clear remaining areas. The onboard processing eliminates the latency and vulnerability associated with transmitting sensor data to remote servers for analysis.

Finally, the modular approach allows individual perception components to be upgraded independently as newer models become available, without redesigning the entire system. This modularity also enables mission-specific tuning—for example, adjusting the confidence threshold to favor recall over precision in high-stakes scenarios, or swapping CLIPSeg for a domain-fine-tuned model to optimize performance for specific operational contexts. The system's demonstrated ability to operate on commercial off-the-shelf hardware (Intel i7 + NVIDIA MX450) suggests feasibility for integration with existing military robotic platforms.

6.2. SUMMARY/CONCLUSIONS

This paper presented HybridNAV, a hybrid framework combining multiple vision-language models with adaptive navigation for zero-shot object detection in unknown

environments. The system achieves a 24% improvement in detection F-score over single-model baselines while reducing scan time by 25% and path length by 30%, all while operating in real-time on resource-constrained hardware without external communication dependencies.

The results demonstrate that hybrid multi-model pipelines offer meaningful advantages for reconnaissance applications, particularly for detecting uncommon objects, leveraging spatial relationships, and confirming detections through multi-view analysis. The priority-based navigation system effectively balances target investigation with area coverage, producing behavior well-suited to reconnaissance doctrine.

Future work will focus on field trials in more diverse and operationally relevant environments - outdoor settings, multi-floor buildings, and degraded visibility conditions - to validate readiness for real-world deployment.

7. REFERENCES

[1] T. Lüddecke and A. S. Ecker, "Image Segmentation Using Text and Image Prompts," in *Conf on Computer Vision and Pattern Recognition*, New Orleans, 2022.

[2] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu and L. Zhang, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *Euro Conf on Computer Vision*, Milan, Italy, 2024.

[3] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. R{\"a}dle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick and P. Dolla, "SAM 2: Segment Anything in Images and Videos," in *13th Int Conf on Learning Representations*, Singapore, 2025.

[4] B. Yamaauchi, "A Frontier-Based Approach for Autonomous Exploration," in *Computational Intelligence in Robotics and Automation*, 1997.

[5] N. Yokoyama, S. Ha, D. Batra, J. Wang and B. Bucher, "VLFM: Vision-Language Frontier Maps for Zero-Shot Semantic Navigation," in *Int Conf on Robotics and Automation (ICRA)*, Yokohama, Japan, 2024.

[6] F. L. Busch, T. Homberger, J. Ortega-Peimbert, Q. Yang and O. Andersson, "One Map to Find Them All: Real-time Open-Vocabulary Mapping for Zero-shot Multi-Object Navigation," in *Int Conf on Robotics and Automation (2025)*, Atlanta, USA, 2025.

[7] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, " A density-based algorithm for discovering clusters in large spatial databases with noise," in *Int Conference on Knowledge Discovery and Data Mining*, Portland, 1996.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Int. Conference on Machine Learning*, Online, 2021.

[9] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser and M. Nießner, "ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes," in *Computer Vision and Pattern Recognition*, Honolulu, USA, 2017.

[10] E. Martinson and H. Indurthi, "Augmenting Open Vocabulary Object Detection with Large

Language Models for Home Service Robots*," in *International Conference on Automation Science and Engineering (CASE)*, Los Angeles, 2025.

[11] Hello Robot, "Stretch Docs," March 2026. [Online]. Available: https://docs.hello-robot.com/latest/..

[12] T. Zhao, P. Liu and K. Lee, "OmDet: Large-scale vision-language multi-dataset pre-training with multimodal detection network," *IET Computer Vision,* vol. 18, no. 5, 2024.

8. CONTACT INFORMATION

Hemanth Venkata Indurthi
hindurthi@ltu.edu

Eric Martinson, Ph.D.
Department of Mathematics and Computer Science
Lawrence Technological University
Southfield, MI 48075
emartinson@ltu.edu

9. ACRONYMS

CLIP – Contrastive Language-Image Pretraining

DBSCAN – Density-Based Spatial Clustering of Applications with Noise

DINO – Distillation with No Labels (Grounding DINO)

FSM – Finite State Machine

IoU – Intersection over Union

RGB-D – Red-Green-Blue plus Depth

ROS – Robot Operating System

SAM – Segment Anything Model

VLM – Vision-Language Model