A Multimodal Perception Pipeline for Out-of-Distribution Autonomous Navigation

Zach Goldwyn, Ivan Torriani, Ceyanna Badyal, Saurish Suman, Vinayak Kohli

CSAI Research Team, California Polytechnic University San Luis Obispo

Abstract

Autonomous navigation systems trained on fixed datasets struggle to generalize when deployed in unfamiliar environments. A model trained on high-infrastructure urban settings may fail to correctly interpret a rural tractor, a costumed pedestrian, or an unusual roadside obstacle — not because the model is poorly designed, but because the object or scene lies outside its training distribution. Rather than retraining or replacing existing navigation systems, this work presents a supplementary perception pipeline that can be layered on top of them to handle out-of-distribution (OOD) inputs. The pipeline has three stages: (1) *unknown object detection*, comparing three practical approaches for flagging objects a detector cannot confidently classify; (2) *multimodal scene interpretation*, using vision-language models (VLMs) to produce a natural-language scene description and a binary traversability judgment; and (3) *fiducial-marker-based navigation*, using AprilTag pose estimation to issue directional movement commands gated by the traversability verdict. We evaluate the three detection approaches on a 30-trial known/unknown object protocol, finding that a confidence-thresholded open-vocabulary detector achieves 93.3% accuracy versus 53.3% for a lightweight colour-histogram baseline. We further compare two cloud vision-language APIs on a 9-image traversability task, finding GPT-4.1-nano modestly more accurate than Gemini 2.5 Flash Lite (88.9% vs. 77.8%) at comparable latency (~1.9s). Local, on-device VLMs are shown to be impractical in their current form, requiring 10–30+ minutes per image on CPU versus 2–5 seconds for cloud APIs. Taken together, the three subsystems form a modular, model-agnostic pipeline that can flag perceptual uncertainty and provide contextual scene understanding — capabilities largely absent from standard autonomous navigation stacks operating outside their training domain.

Keywords: Out-of-distribution detection, Open-vocabulary object detection, Vision-language models, Traversability estimation, Fiducial markers, Autonomous navigation

1. INTRODUCTION

Deep learning models used in autonomous navigation are trained on a fixed data distribution, and their performance degrades — often silently — when that distribution is violated at deployment time. A perception stack trained primarily on dense urban traffic may have no representation for a farm vehicle, a person in costume, or debris scattered across a road. In each case the failure is not a defect in the model's training procedure; it is a structural limitation of closed-world learning. As operating environments become less pre-

dictable, this limitation becomes the dominant obstacle to reliable autonomous operation, motivating supplementary perception mechanisms rather than continual retraining of the primary navigation stack.

This paper presents a three-stage pipeline designed to sit alongside an existing navigation system rather than replace it. The pipeline first attempts to detect when an object cannot be confidently classified (Section 4.1), then delegates scene-level reasoning about that uncertainty to a vision-language model that judges whether the scene remains safely traversable (Section 4.2), and finally uses a fiducial marker system to provide the concrete directional control needed to act on that judgment (Section 4.3). Our contributions are:

- A comparative evaluation of four unknown-object detection strategies — a confidence-thresholded open-vocabulary detector, a dual-model class-agnostic/classifier consensus pipeline, a closed-vocabulary baseline, and a lightweight colour-histogram change detector — evaluated under an identical 30-trial protocol.
- A comparative evaluation of six vision-language models (four local, two cloud-hosted) for scene-description-and-traversability judgment, characterizing the practical latency gap between on-device and API-based inference.
- A working fiducial-marker navigation subsystem that converts AprilTag pose estimates into traversability-gated directional commands.
- Property-based validation (Section 4.2) of the shared utilities that glue the multimodal subsystem together, using randomized testing rather than fixed example cases.

2. RELATED WORK

2.1 Open-Vocabulary and Out-of-Distribution Object Detection

Standard object detectors are trained on a closed set of classes and have no mechanism to express uncertainty about inputs outside that set. Yang et al. [11] formalize this gap under the umbrella of *generalized out-of-distribution detection*, unifying outlier, anomaly, and open-set recognition problems. Open-vocabulary detectors such as YOLO-World [1] partially address this by grounding detection in a region-text contrastive objective rather than a fixed class head, allowing a single model to be queried against an arbitrary text vocabulary at inference time. Minderer et al.'s OWL-ViT [4] takes a related approach, transferring a contrastively pretrained image-text encoder to open-vocabulary localization by attaching lightweight detection heads directly to its output tokens; OWL-ViT is the architecture underlying the class-agnostic NanoOWL detector used in our dual-model consensus approach (Section 4.1).

2.2 Fiducial-Marker Navigation

Visual fiducial systems provide a reliable, low-cost alternative to full scene reconstruction when a known landmark is sufficient for localization. Olson's AprilTag [2] introduced a 2D bar-code-style marker with a lexicode-based encoding that guarantees a minimum Hamming distance between tags under rotation, enabling robust 6-DOF pose recovery from a single image. Wang and Olson's AprilTag 2 [3] substantially improved detection rate, false positive rate, and computational cost over the original system. Our navigation subsystem (Section 4.3) builds directly on this line of work, using AprilTag pose estimates as the sole localization signal for directional control.

2.3 Vision-Language Models for Scene Understanding

Recent vision-language models (VLMs) connect a visual encoder to a large language model, enabling natural-language reasoning over image content. LLaVA [5] demonstrated that instruction-tuning a vision encoder and LLM pair on GPT-4-generated multimodal instruction data yields strong general-purpose visual chat ability. Qwen2-VL [6] and InternVL [7] extend this family with native support for variable input resolution

and large-scale vision-language alignment, respectively, and are representative of the open, locally deployable VLMs evaluated in Section 4.2. Moondream2 [12] represents the opposite end of the size spectrum, a sub-2B-parameter open VLM designed to run efficiently on constrained hardware. On the cloud API side, Google's Gemini family [8] and OpenAI's GPT-4o family [9] provide multimodal inference as a hosted service, trading local compute and data control for substantially lower latency — a trade-off we quantify directly in Section 6.2.

2.4 Property-Based Testing

Property-based testing, introduced by Claessen and Hughes' QuickCheck [10], validates code against declared properties over randomly generated inputs rather than a fixed list of example cases. We apply this methodology (via the Python library Hypothesis) to the shared parsing and formatting utilities used across all multimodal-interpretation backends, described in Section 4.2.

3. SYSTEM OVERVIEW

The full pipeline (Fig. 1) runs continuously on a live camera stream. Each captured frame is passed through unknown-object detection; a detection that persists for at least a fixed number of consecutive frames — a persistence gate that filters transient detector flicker — is treated as a confirmed unknown and triggers a multimodal traversability query. Independently of detection state, every frame is also passed through AprilTag pose estimation. Forward motion is gated jointly by the most recent traversability verdict and the tag's estimated distance: the agent may only move forward when the scene has been judged traversable *and* the tag is not already within its stopping distance: otherwise it stops. The annotated frame, including the current detection state, traversability verdict, and movement command, is rendered to the display in real time.

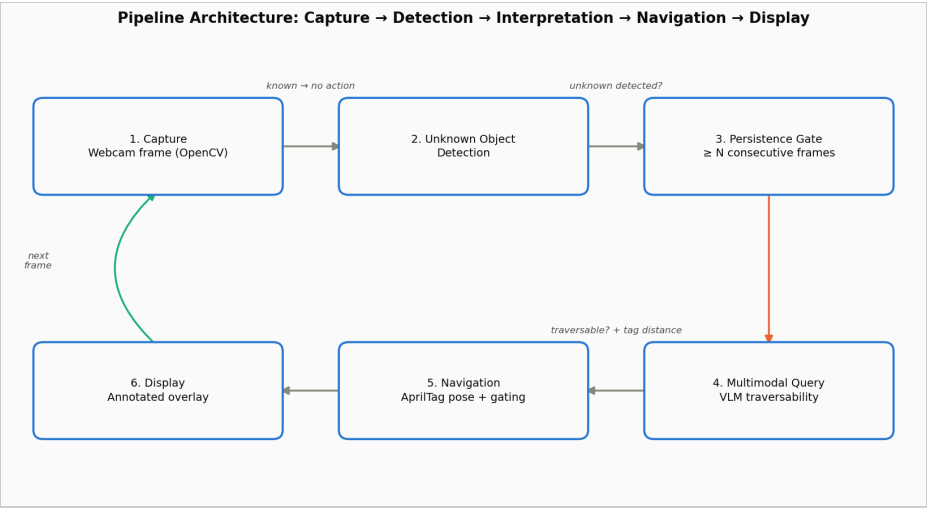

Figure 1. The six-stage pipeline. An unknown detection must persist before it triggers a multimodal query (orange arrow); navigation runs on every frame regardless of detection state and gates forward motion on the traversability verdict and tag distance (bottom row).

4. METHODS

4.1 Unknown Object Detection

Four approaches were implemented and evaluated on live camera input, each addressing the same underlying question — is this object outside the detector's confident operating range? — with a different mecha-

nism and a different computational cost.

Confidence-thresholded YOLO-World. YOLOv8s-worldv2, an open-vocabulary detector [1], is run with agnostic non-maximum suppression and a fixed confidence threshold. Writing the model's confidence for a detected region as c, a detection is labeled with its predicted class when $c \geq 0.50$, and flagged as an unknown object when $c < 0.50$. Confidence acts as a proxy for how well the observed region matches the model's training distribution.

Dual-model consensus (NanoOWL + YOLO-World). A class-agnostic detector, NanoOWL (built on OWL-ViT [4]), answers only whether *something* is present, using a permissive detection threshold (0.08) run every third frame to control latency. YOLO-World is queried in parallel against a fixed 12-class candidate vocabulary. Each NanoOWL detection is matched against YOLO-World's boxes by intersection-over-union, requiring

$$\mathrm{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|} \;\geq\; 0.20$$

for a pair of boxes to be treated as the same object; a matched detection is then classified as **matched** (YOLO-World confidence ≥ 0.35), **weak match** (a lower-confidence match), or **unknown** (no sufficient-overlap match at all). The intuition is that a class-agnostic model can confirm object *presence* even when a closed-candidate classifier cannot confirm object *identity*.

Baseline closed-vocabulary YOLO. A standard YOLOv11n model with no unknown-object logic is included as a baseline: every detection is forced into its closed class set, with no mechanism to express uncertainty about an out-of-distribution object. This baseline illustrates the core limitation motivating the other three approaches.

Colour histogram change detection. As a lightweight, model-free alternative, consecutive frames are converted to an 8×8×8-bin ($N = 512$) BGR colour histogram, and consecutive histograms H_1, H_2 are compared via the Bhattacharyya distance

$$d(H_1, H_2) = \sqrt{1 - \frac{1}{\sqrt{\bar{H}_1 \bar{H}_2 N^2}} \sum_i \sqrt{H_1(i)\, H_2(i)}}$$

where $\bar{H}_k$ is the mean bin value of histogram k. A distance above 0.1 flags a scene-level change. Unlike the other three approaches, this method detects *that* the scene changed, not *what* object is responsible, and requires no pretrained model or GPU.

4.2 Multimodal Scene Interpretation

Once an object is confirmed unknown, the pipeline still needs a decision: can the agent safely continue? We evaluate whether vision-language models can supply that judgment directly from a single image, producing both a natural-language scene description and a binary traversability verdict. Six models were evaluated: four run locally via Hugging Face `transformers` in float16 precision — Qwen2-VL-7B-Instruct [6], LLaVA-1.5-7B [5], Moondream2 [12], and InternVL2-2B [7] — and two accessed via cloud API — GPT-4o-mini [9] and Gemini 2.0 Flash Lite [8]. Local models were queried with two sequential prompts (a description request followed by a yes/no traversability question); the OpenAI backend instead used a single structured prompt requesting a JSON object with description, traversability, and justification fields, reducing latency and API cost. A shared helper module provides consistent traversability-string parsing, output-dict assembly, and output-path construction across all six backends. These utilities were validated with property-based testing [10] (Python Hypothesis, 100 generated examples per property), checking five cor-

rectness properties — for example, that traversability parsing returns true if and only if a model's response begins with "yes," case-insensitive.

4.3 Navigation

Once an unknown object or non-traversable scene has been identified, the agent needs a reliable mechanism to orient itself using a known landmark. This subsystem detects a printed AprilTag [2][3] in the camera frame and recovers its 3D pose using camera intrinsics obtained via a checkerboard calibration procedure (8×6 inner-corner grid, 20 mm squares). Writing the tag's translation vector as (t_x, t_y, t_z), movement commands are derived from fixed thresholds on lateral offset t_x and depth t_z, gated by the traversability verdict from Section 4.2:

$$\text{command} = \begin{cases} \text{Move Left} & t_x < -0.075 \text{ m} \\ \text{Move Right} & t_x > +0.075 \text{ m} \\ \text{Move Forward} & t_z > 0.10 \text{ m} \ \wedge \ \text{traversable} \\ \text{Stop} & \text{otherwise} \end{cases}$$

summarized in Table 1.

Table 1. Movement control logic derived from AprilTag pose.

Condition	Command
Lateral offset < −7.5 cm	Move Left
Lateral offset > +7.5 cm	Move Right
Depth > 10 cm (and scene traversable)	Move Forward
Depth ≤ 10 cm, or scene not traversable	Stop
No tag detected	No Tag Detected

The lateral thresholds define a dead-band within which the tag is considered aligned; conditions are evaluated in this order so that the agent corrects heading before advancing.

5. EXPERIMENTAL SETUP

Detection trials. Each of the three quantitatively evaluated detection approaches (confidence-thresholded YOLO-World, dual-model consensus, and colour histogram change detection) was tested under an identical protocol: five known objects (scissors, cell phone, fork, knife, cup) and five unknown objects (thermometer, lighter, chapstick, keys, inhaler) were each presented to the camera for three trials, for 30 trials per approach. Each trial was scored as a true positive (TP, an unknown object correctly flagged), true negative (TN, a known object correctly left unflagged), false positive (FP, a known object incorrectly flagged), or false negative (FN, an unknown object missed). From these counts we report accuracy, precision, recall, and F1 for each approach:

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

VLM traversability comparison. Two cloud vision-language APIs — Gemini 2.5 Flash Lite and GPT-4.1-nano — were compared on nine test images (test1–test10, excluding test4, which was unavailable), using an identical prompt requesting a structured JSON response containing a scene description, a binary traversabil-

ity judgment, and a one-sentence justification. Wall-clock latency was measured per API call, and traversability judgments were scored against manually assessed ground truth.

6. RESULTS

6.1 Detection Trials

Table 2 and Figure 2 summarize the 30-trial results for each detection approach. The confidence-thresholded YOLO-World detector achieved the highest performance across all four metrics (93.3%), correctly identifying 14 of 15 unknown objects and 14 of 15 known objects. The dual-model consensus approach traded some recall for a lower false-negative rate elsewhere in the confusion matrix, reaching 80.0% accuracy (73.3% recall, 84.6% precision). The colour histogram change detector performed substantially worse (53.3% accuracy, 13.3% recall): because it responds to global frame-level colour change rather than object identity, it frequently failed to register a sufficiently large change when a small object was introduced into an already-populated scene, consistent with its intended role as a coarse pre-filter rather than a standalone detector.

Table 2. Detection-trial confusion matrices (30 trials each: 15 known, 15 unknown).

Approach	TP	TN	FP	FN	Acc.	Prec.	Recall	F1
Confidence-Thresholded YOLO-World	14	14	1	1	93.3%	93.3%	93.3%	93.3%
Dual-Model Consensus (NanoOWL + YOLO-World)	11	13	2	4	80.0%	84.6%	73.3%	78.6%
Colour Histogram Change Detection	2	14	1	13	53.3%	66.7%	13.3%	22.2%

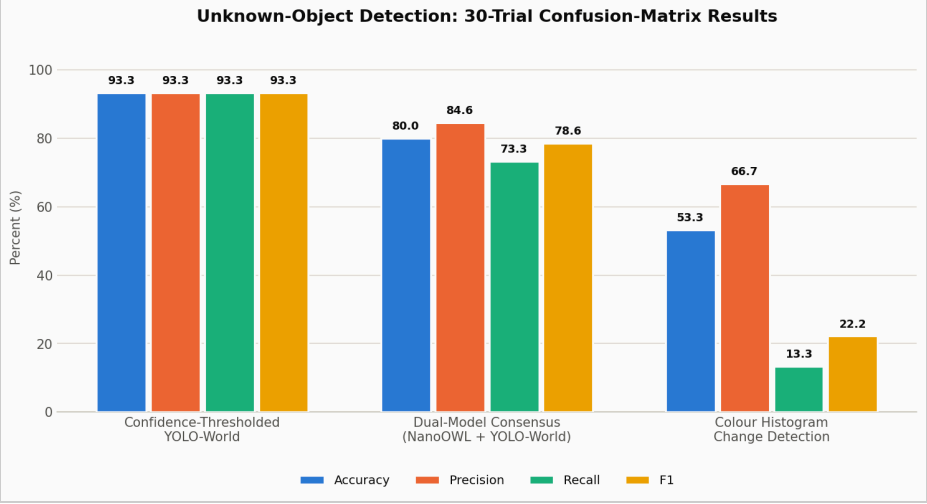

Figure 2. Accuracy, precision, recall, and F1 across the three detection approaches. For navigation safety, recall is the more safety-critical metric, since a missed unknown object (false negative) is more dangerous than a false alarm.

6.2 Multimodal Traversability Comparison

GPT-4.1-nano outperformed Gemini 2.5 Flash Lite on traversability accuracy (88.9%, 8/9, versus 77.8%, 7/9) at near-identical mean latency (1.894s versus 1.944s; Table 3, Fig. 3). Gemini's two errors both involved ambiguous human presence or surface appearance: it judged a scene containing a person in a morph suit next to a novelty pedestrian sign as traversable, failing to treat the pedestrian-adjacent hazard as a reason to stop, and judged a graffiti-covered but structurally sound bridge as non-traversable, over-indexing on visual degradation as a proxy for structural unsafety. GPT-4.1-nano's single error was the opposite failure

mode — over-caution — judging a scene with a roadside hitchhiker and a clear road ahead as non-traversable. Both models performed reliably on unambiguous cases (a fallen tree, a flooded road, livestock on the roadway, an aircraft blocking the road) and struggled only on scenes involving human presence or ambiguous infrastructure condition.

Table 3. Per-image traversability judgment (Gemini 2.5 Flash Lite vs. GPT-4.1-nano) against manually assessed ground truth.

Image	Ground truth	Gemini	OpenAI
test1	false	✗ true	✓ false
test2	false	✓ false	✓ false
test3	true	✓ true	✓ true
test5	false	✓ false	✓ false
test6	false	✓ false	✓ false
test7	true	✗ false	✓ true
test8	false	✓ false	✓ false
test9	true	✓ true	✗ false
test10	false	✓ false	✓ false
Correct	—	**7/9**	**8/9**

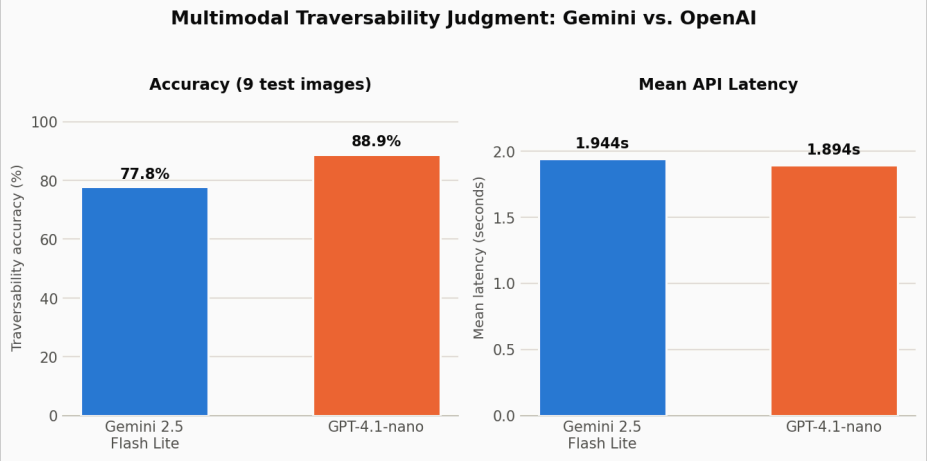

Figure 3. Traversability accuracy and mean API latency, Gemini 2.5 Flash Lite vs. GPT-4.1-nano, across 9 test images.

7. DISCUSSION AND LIMITATIONS

Across all three subsystems, evaluation was qualitative-to-moderate scale rather than large-benchmark: detection trials cover 30 presentations per approach and the VLM comparison covers 9 images, both researcher-supplied rather than drawn from a standardized dataset. Within unknown-object detection, the manually curated 12-class YOLO-World candidate vocabulary may not generalize across environments, and NanoOWL was run on CPU, considerably slower than the NVIDIA edge hardware it targets; a planned custom class-agnostic model trained on researcher-collected data was designed but not completed in this phase. Within multimodal interpretation, no ground-truth traversability labels exist beyond manual researcher assessment, all four local models were evaluated CPU-only, and traversability judgments are made from sin-

gle static frames with no temporal context, depth, or sensor fusion — the single most consequential finding here is practical rather than qualitative: local VLM inference (10–30+ minutes/image on CPU) is not currently viable for real-time deployment, while cloud APIs (2–5s/image) introduce a hard dependency on network connectivity and third-party availability. Within navigation, the system assumes a single tag per scene, issues only discrete threshold-based commands with no continuous controller, degrades at steep viewing angles or long distances, and has no temporal smoothing across frames. Most significantly, the three subsystems have not yet been integrated into a single running system in this evaluation phase; the pipeline architecture in Figure 1 reflects the intended integration, exercised in demo form (Section 3), but the quantitative results in Section 6 were collected on each subsystem independently.

8. CONCLUSION AND FUTURE WORK

We presented a three-stage supplementary perception pipeline for autonomous navigation in out-of-distribution environments, combining unknown-object detection, multimodal traversability assessment, and fiducial-marker-based navigation into a modular, model-agnostic architecture. Our results show that simple confidence-based detection can be highly effective (93.3% accuracy) when the underlying detector is open-vocabulary, that cloud vision-language APIs are currently the only practical option for real-time scene interpretation, and that AprilTag-based navigation provides a reliable, low-cost mechanism for the directional control needed to act on that interpretation. The most direct next steps are: full integration of the three subsystems into a single running pipeline evaluated end-to-end; quantization (e.g., INT8, GGUF) and hardware acceleration to make local VLM inference viable for real-time, network-independent deployment; and evaluation on a larger, standardized benchmark rather than researcher-curated test sets.

REFERENCES

1. T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "YOLO-World: Real-Time Open-Vocabulary Object Detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2401.17270.
2. E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2011.
3. J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2016.
4. M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, and N. Houlsby, "Simple Open-Vocabulary Object Detection with Vision Transformers," in *Proc. European Conf. Computer Vision (ECCV)*, 2022. arXiv:2205.06230.
5. H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2304.08485.
6. P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge et al., "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," arXiv preprint arXiv:2409.12191, 2024.
7. Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai, "InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2312.14238.
8. Gemini Team, Google DeepMind, "Gemini: A Family of Highly Capable Multimodal Models," arXiv preprint arXiv:2312.11805, 2023.
9. OpenAI, "GPT-4o System Card," 2024. arXiv:2410.21276.
10. K. Claessen and J. Hughes, "QuickCheck: A Lightweight Tool for Random Testing of Haskell Programs," in *Proc. ACM SIGPLAN Int. Conf. Functional Programming (ICFP)*, 2000, pp. 268–279.

11. J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized Out-of-Distribution Detection: A Survey," arXiv preprint arXiv:2110.11334, 2021.

12. vikhyatk, "moondream2" (model card), Hugging Face, 2025. https://huggingface.co/vikhyatk/moondream2