VLM as a Virtual Proprioception: Semantic Slip Detection for Sensorless Grippers using Quantized Small-Scale Models

1st Chacharin Lertyosbordin*
King Mongkut's University of Technology Thonburi,
Bangkok, Thailand
*lertyosbordin@outlook.com

2nd Boonyawat Chaithirasakul
Singapore International School Thonburi,
Bangkok, Thailand
ben.bc.student@sisbschool.com

3rd Puvasit Aphisinrungrot
King Mongkut's Institute of Technology Ladkrabang International Demonstration School, Bangkok, Thailand
jaonai1301@gmail.com

4th Sorathon Chungsawanant
International School Bangkok,
Bangkok, Thailand
23077@students.isb.ac.th

5th Nontapat Prayoonthong
International School Bangkok,
Bangkok, Thailand
19600@students.isb.ac.th

2026 7th International Conference on Mechatronics Technology and Intelligent Manufacturing (ICMTIM) | 979-8-3315-6178-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICMTIM69588.2026.11525852

***Abstract* — Reliable manipulation in unstructured environments typically requires expensive force-torque sensors to detect grasping failures. Low-cost robotic arms, relying on open-loop control, lack this proprioceptive feedback, rendering them susceptible to slippage and "air delivery" errors—particularly when handling small objects with dimensional variances. This paper introduces "Virtual Proprioception," a framework that repurposes a quantized Vision-Language Model (VLM), specifically Qwen3-VL-4B, as a semantic tactile sensor. To circumvent the inference latency inherent in large models, we propose a Discrete Keyframe Monitoring strategy that utilizes semantic reasoning to verify object presence at four critical state transitions. The system was validated using a specialized test set of 3D-printed components fabricated with ±2 mm manufacturing tolerances. Experimental results demonstrate that while the baseline open-loop system suffered a 35.0% Critical Failure Rate due to undetected slippage, our proposed framework successfully intercepted all failure instances, reducing the Critical Failure Rate to 0%. These findings confirm that lightweight VLMs can effectively serve as software-defined alternatives to hardware sensors, democratizing high-reliability manipulation for resource-constrained robotics.**

Keywords* — *Vision-Language Models (VLM); Robotic Manipulation; Failure Detection; Low-cost Robotics

I. Introduction

Low-cost robotic manipulators, widely deployed in service and educational sectors, suffer from a fundamental hardware limitation: the absence of proprioceptive force-torque sensors [1]. Unlike industrial counterparts, these systems rely on open-loop control, rendering them susceptible to grasping failures—specifically object slippage during transport—which cannot be detected by internal kinematics alone. While adding external sensors incurs significant hardware costs, recent advancements in Vision-Language Models (VLMs) offer a potential software-defined alternative.

State-of-the-art VLM frameworks, such as Guardian [2] and I-FailSense [3], have demonstrated semantic reasoning capabilities for failure detection. However, these systems predominantly rely on large-scale cloud models or continuous video processing, introducing latency that is prohibitive for closed-loop control on edge devices. The challenge, therefore, is to leverage the semantic understanding of VLMs within the harsh latency and computational constraints of low-cost hardware.

This paper proposes "Virtual Proprioception," a framework that repurposes a quantized small-scale VLM (Qwen3-VL-4B) as a semantic tactile sensor. To circumvent the latency bottleneck of real-time streaming, we introduce a Discrete Keyframe Monitoring strategy. Rather than processing continuous frames, our system queries the VLM only at four critical state transitions: Grasp, Lift, Transport, and Place. This approach enables a resource-constrained manipulator to detect and recover from grasp failures in near real-time without additional sensing hardware.

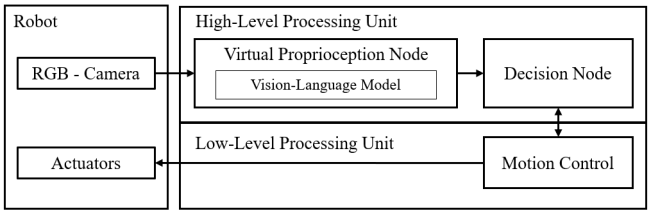

Figure 1. System Design

II. Related Work

A. Learning-Based Grasping without Tactile Feedback

Prior to Foundation Models, grasping reliability relied on geometric heuristics and deep learning. Matl et al. [1] utilized Dex-Net to synthesize policies for suction and parallel-jaw grippers based on depth images. Similarly, Pixel-Reasoning [4] employed EDINet to infer stability from visual features. While effective in structured settings, these methods lack the semantic generalization to handle unseen object variations or "semantic slippage" where an object is held but misaligned.

B. Semantic Failure Detection

The integration of VLMs has shifted focus from geometric to semantic failure analysis. Guardian [2] and I-FailSense [3] represent the state-of-the-art, utilizing VLMs to classify errors (e.g., planning vs. execution failures) by analyzing temporal image sequences. However, these frameworks are designed for post-hoc analysis rather than active control. They prioritize detailed reasoning (explaining why a failure occurred) over the rapid binary verification (checking if the object is held) required for real-time recovery.

C. Efficient Vision-Language Models

Recent efforts to democratize VLMs have led to lightweight architectures. NORA [5] introduces a 3B-parameter model for embodied tasks, and the Qwen-VL series [6] has optimized dense multimodal understanding for smaller parameters. Despite these advances, applications of small VLMs have primarily focused on high-level planning or open-ended instruction following [7], [8]. Our work diverges by applying these efficient models strictly as a low-level feedback mechanism—effectively acting as a virtual sensor within a control loop.

III. Methodology

The proposed "Virtual Proprioception" framework integrates a quantized Vision-Language Model into a closed-loop robotic control system. Unlike traditional visual servoing which relies on high-frequency geometric feedback, our approach utilizes semantic state verification at discrete intervals to compensate for the latency of local VLM inference.

A. System Architecture

The experimental platform comprises a low-cost 4-DOF robotic manipulator driven by RC servo motors, controlled via an Arduino UNO microcontroller (Low-Level Unit). The overall system design and interaction between nodes are illustrated in Figure 1. Visual data is acquired through a generic RGB web-camera (640 x 480 px resolution). The High-Level Processing Unit is a standard laptop Lenovo LOQ 15IRX9 (GPU: RTX 3050) running the VLM on LM Studio. Communication between the high-level interface (Python) and the low-level motion controller (Arduino) is established via a serial interface (UART) at 115200 baud. The Arduino executes a pre-programmed motion sequence continuously, while the decision node serves as a supervisory control interface allowing for interruption commands. The physical hardware setup of the manipulator and sensor is shown in Figure 2.

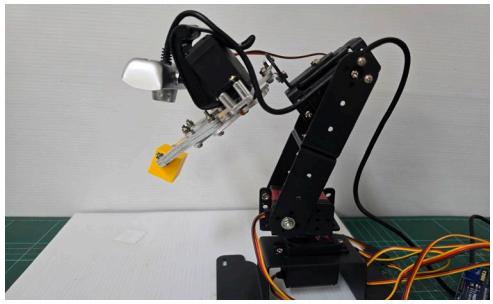

Figure 2. Hardware setup

B. Semantic Perception Engine

We employ Qwen3-VL-4B [6] as the core perception engine for virtual proprioception node. Unlike geometric vision systems that estimate pose or measure gripper aperture, our approach utilizes the VLM's semantic understanding to perform direct object presence verification. The model is tasked with visually confirming whether the specific target object is physically located within the gripper's jaws.

To ensure deterministic outputs, we use the following structured prompt:

"Analyze the gripper camera image. Focus on the gap between the jaws. Is the object physically present? Return ONLY a raw JSON object without Markdown formatting. Schema: {"object_present": true/false, "confidence": number 0-100}."

This approach shifts the burden of validation from mechanical inference (checking motor stall) to semantic recognition, allowing the system to detect failures even if the gripper appears mechanically engaged but is actually empty.

C. Discrete Keyframe Monitoring

To mitigate the inference latency, we abandon continuous monitoring in favor of a discrete keyframe monitoring strategy. The system validates grasp stability only at four critical kinematic singularities:

- Post-Grasp(t_1): Verifies object acquisition following gripper closure.
- Peak-Lift(t_2): Verifies object stability at the peak of the lifting trajectory. (The visual difference between a stable grasp and a slippage event at this state is depicted in Figure 3.)
- Pre-Place(t_3): Verifies object retention following lateral transport and acceleration.
- Post-Place(t_4): Verifies object release and successful deposition at the target location.

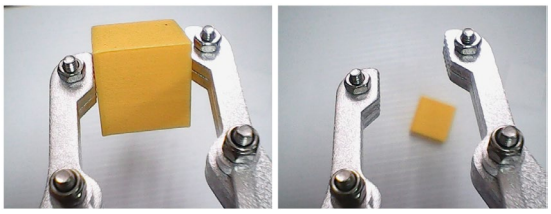

Figure 3. OpenCV snapshot of the Peak-Lift (t_2) state, illustrating the visual distinction between a stable grasp and a slippage event.

D. Closed-Loop Failure Recovery

The Decision Node operates on a confidence-thresholded logic. Upon receiving the JSON payload from the Virtual Proprioception Node, the system evaluates the object_present status. If a failure is detected (e.g., object_present: false) with a confidence score exceeding 80%, the node transmits a "STOP" byte-string to the Low-Level Motion Control. The Arduino then triggers an interrupt routine to halt motion and reset the manipulator to a safe home configuration.

IV. Experiments And Results

To validate the reliability of the proposed framework, we conducted a comparative study on small-scale industrial assembly tasks. The experiment highlights the vulnerability of open-loop systems to "false positive" grasps and demonstrates how visual semantic verification can mitigate this risk.

A. Experimental Setup and Test Objects

The robotic platform utilized for validation comprises a 4-DOF manipulator equipped with a standard RC servo gripper. To rigorously evaluate system performance under realistic manufacturing variances, a specialized test set was fabricated, consisting of 20 distinct 3D-printed samples across 4 colors for each object category. The test objects included rigid cubes with a nominal dimension of 20 x 20 mm and cylinders with a nominal diameter of 20 mm.

Crucially, both object types were produced with a dimensional variance of ±2 mm to simulate geometric irregularities. During the experimental protocol (N=40 trials), these objects were introduced into the workspace via manual placement without fixed jigs, thereby inducing a natural positional uncertainty of approximately ±5 mm relative to the robot's target coordinates. The specifications of the test set are detailed in Table I.

TABLE I. Specifications Of The Fabricated Test Set

Object Category	Color	Available Sizes (mm)	Total Quantity
Cube	Red	18, 19, 20, 21, 22	5
	Green	18, 19, 20, 21, 22	5
	Blue	18, 19, 20, 21, 22	5
	Yellow	18, 19, 20, 21, 22	5
Cylinder	Red	18, 19, 20, 21, 22	5
	Green	18, 19, 20, 21, 22	5
	Blue	18, 19, 20, 21, 22	5
	Yellow	18, 19, 20, 21, 22	5
		Grand Total	40

B. Control Strategies and Metrics

We compared two distinct control strategies under identical conditions:

1) Baseline (Fixed-Position Control): The servo is commanded to a pre-set angle corresponding to a 20 mm jaw aperture. In cases where the object slips or is undersized due to variance, the gripper retains the 20 mm gap without applying holding force. Lacking feedback, the robot blindly completes the transport cycle (Air Delivery).

2) Proposed (Visual Gatekeeper): Immediately following the grasp command, the system captures a keyframe image. The Qwen3-VL-4B model semantically verifies whether the object is physically present within the aperture. Only upon positive confirmation does the robot proceed.

To quantify system reliability, three distinct performance metrics were defined. Task Success (S) is recorded when the object is successfully retained throughout the trajectory and deposited at the designated target. Conversely, Critical Failure (F_{crit}) characterizes the detrimental "blind execution" scenario, where the object is lost due to slippage, yet the controller—lacking feedback—unknowingly proceeds to complete the full transport cycle with an empty gripper.

Finally, Safe Interception (I_{safe}) serves as the primary indicator of the proposed framework's efficacy. This metric accounts for instances where the VLM correctly identifies object absence (returning object_present: false) and triggers an immediate halt, effectively converting a potential critical failure into a managed system reset.

C. Results and Analysis

The quantitative results from the 40 experimental trials are summarized in Table II. The data reveals a significant divergence in operational reliability between the baseline and the proposed framework.

TABLE II. System Reliability Comparison

Method	Object Type	N	S	F_{crit}	I_{safe}	Critical Failure Rate
Baseline (Open-Loop)	Cube	20	14	6	-	30%
	Cylinder	20	12	8	-	40%
	Average	40	65.0%	14	-	35%
Ours (Proposed)	Cube	20	14	0	6	0%
	Cylinder	20	12	0	8	0%
	Average	40	65%	0	14	0%

1) Baseline Performance: The Baseline method achieved an average task success rate of 65.0%. While functional for majority cases, the residual failure rate (35.0%) represents a severe operational hazard. The Cylinder object, with its rotational instability, proved particularly problematic; 40% of attempts (8 out of 20) resulted in a "False Positive" grasp where the object slipped, yet the robot unknowingly proceeded to deliver empty air.

2) Impact of Virtual Proprioception: The Proposed Framework maintained the same physical success rate (65.0%) as the baseline, which is expected as the VLM does not alter the mechanical grasping capability. However, the crucial contribution lies in the total elimination of Critical Failures. As shown in Table II, the VLM successfully intercepted 100% of the slip incidents (6/6 for Cubes, 8/8 for Cylinders).

3) Zero-Failure Validation: The Critical Failure Rate dropped to 0% across all trials. The Qwen3-VL-4B model demonstrated robust sensitivity in detecting the "empty aperture" state, even under partial occlusion. This result confirms that while the VLM cannot physically prevent slippage, it guarantees that no invalid transport cycle is ever completed, effectively raising the logical reliability of the system to 100% without hardware modifications.

V. Discussion

The results presented in Section IV demonstrate a fundamental decoupling of mechanical grasp stability from operational reliability. While the physical success rate of the low-cost gripper remained capped at 65% due to mechanical limitations, the logical reliability of the system reached 100% through Virtual Proprioception. This section discusses the theoretical implications and boundaries of these findings.

A. Semantic Verification vs. Mechanical Limits

The high Critical Failure Rate (35.0%) observed in the baseline highlights a fundamental flaw in open-loop, fixed-position control. In our setup, the gripper was commanded to a rigid 20 mm aperture. Given the ±2 mm dimensional variance of the 3D-printed test set, artifacts with physical dimensions smaller than the setpoint (e.g., 18-19.9 mm) often failed to generate sufficient clamping force. In these undersized cases, slippage was not merely probable, but physically inevitable due to the lack of frictional engagement against gravity. Traditional geometric approaches [1], [4] often overlook these stochastic physical factors.

Our framework addresses this ambiguity by repurposing the VLM from a passive observer to an active controller. This approach aligns with the "State Recognition" capabilities reviewed by Kawaharazuka et al. [9], but significantly extends their paradigm by embedding the recognition directly into the real-time control loop. By using Qwen3-VL-4B to semantically verify the outcome of the mechanical interaction, our system serves as a robust proxy for stability, proving that semantic verification is more reliable than probabilistic geometric assumptions in unstructured environments.

It should be emphasized that the current experimental validation focuses on a controlled yet diverse set of geometrically varying 3D-printed objects, serving as a proof-of-concept for semantic failure interception. While these artifacts introduce dimensional uncertainty, future evaluations involving objects with broader material properties—such as metallic, transparent, reflective, or deformable surfaces—will be necessary to fully characterize the generalization limits of purely vision-based semantic verification.

B. Efficiency of Discrete Keyframe Monitoring

Recent VLM-based fault detection frameworks, such as Guardian [2] and I-FailSense [3], have set high benchmarks for accuracy by analyzing continuous video streams. However, these methods incur high computational costs and latency, rendering them unsuitable for low-cost, edge-deployed microcontrollers.

Our study proves that continuous monitoring is not strictly necessary for pick-and-place tasks. By identifying critical Kinematic Singularities (Grasp, Lift, Transport, Place), we can reduce the inference load to discrete events without compromising safety. This sparse monitoring strategy allows a quantized model to function effectively on consumer hardware, offering a practical alternative to heavy, cloud-dependent architectures [5], [7].

From a system design perspective, this discrete verification strategy reflects an intentional trade-off between throughput and reliability. Whereas industrial-grade robotic systems often prioritize uninterrupted motion and cycle efficiency through high-bandwidth sensing and continuous perception pipelines, the proposed framework prioritizes safety and failure interception under strict computational constraints. This design choice aligns with the target domain of low-cost and educational robotics, where missed grasp failures are more detrimental than marginal increases in cycle time.

C. Democratizing Reliability via Software-Defined Sensors

While effective, the proposed framework relies heavily on visual cues, making it susceptible to environmental factors such as extreme lighting conditions or total occlusion where the gripper morphology completely hides the object. Additionally, the inference latency of the VLM necessitates a "stop-and-check" motion profile, which reduces the overall cycle time compared to high-speed industrial systems.

Importantly, this sensorless design is a deliberate architectural choice rather than a fundamental constraint. The Virtual Proprioception framework is inherently compatible with multi-modal extensions, and future iterations could incorporate tactile, force, or proprioceptive cues as complementary sensing channels. In such a configuration, semantic visual verification would operate alongside low-level physical sensing to form a lightweight sensor fusion paradigm, improving robustness under extreme visual degradation while preserving the framework's cost-efficiency.

Similarly, advances in NPU-accelerated inference and hardware-aware quantization [10] present a clear pathway toward mitigating the current "stop-and-check" limitation. By offloading VLM inference to dedicated neural accelerators, it may become feasible to perform near-continuous semantic monitoring without explicit motion pauses, thereby closing the performance gap between software-defined sensing and traditional industrial solutions.

In this sense, the proposed framework does not seek to replace industrial-grade tactile sensing systems, but rather to democratize a baseline level of operational reliability. By transforming a brittle open-loop manipulator into a semantically supervised system using only commodity hardware, Virtual Proprioception offers a pragmatic bridge between low-cost robotics and safety-aware manipulation.

VI. Conclusion

This work presented "Virtual Proprioception," a framework retrofitting reliability onto sensorless manipulators via a quantized Qwen3-VL-4B model. Addressing the vulnerability of fixed grippers to dimensional variances, our approach transformed operational safety. While the baseline suffered a 35.0% Critical Failure Rate, our framework intercepted 100% of slippage incidents, reducing the rate to 0%. By substituting brittle mechanical assumptions with semantic visual monitoring, we demonstrate that lightweight VLMs act as effective software-defined alternatives to force sensors, democratizing high-reliability manipulation for low-cost robotics.

References

[1] M. Matl, "Learning robotic grasping policies for suction cups and parallel jaws in simulation," University of California at Berkeley, Tech. Rep. UCB/EECS-2019-119, Aug. 2019.

[2] P. Pacaud, R. Garcia, S. Chen, and C. Schmid, "Guardian: Detecting robotic planning and execution errors with vision-language models," *arXiv preprint arXiv:2512.01946*, Dec. 2025.

[3] C. Grislain, H. Rahimi, O. Sigaud, and M. Chetouani, "I-FailSense: Towards general robotic failure detection with vision-language models," *arXiv preprint arXiv:2509.16072*, Sep. 2025.

[4] C. Shi, C. Miao, X. Zhong, X. Zhong, H. Hu, and Q. Liu, "Pixel-reasoning-based robotics fine grasping for novel objects with deep EDINet structure," *Sensors*, vol. 22, no. 11, art. no. 4283, Jun. 2022.

[5] C.-Y. Hung et al., "NORA: A small open-sourced generalist vision-language action model for embodied tasks," *arXiv preprint arXiv:2504.19854*, Apr. 2025.

[6] Qwen Team, "Qwen3-VL technical report," *arXiv preprint arXiv:2511.21631*, Nov. 2025.

[7] P. Zhi et al., "Closed-loop open-vocabulary mobile manipulation with GPT-4V," *arXiv preprint arXiv:2404.10220*, Mar. 2025.

[8] H. Li et al., "ROBOGROUND: Robotic manipulation with grounded vision-language priors," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[9] K. Kawaharazuka, Y. Obinata, N. Kanazawa, K. Okada, and M. Inaba, "Robotic applications of pre-trained vision-language models to various recognition behaviors," *arXiv preprint arXiv:2303.05674*, Oct. 2023.

[10] W. Chen et al., "AutoNeural: Co-designing vision-language models for NPU inference," *arXiv preprint arXiv:2512.02924*, Dec. 2025.