Vision Language Action Models for Humanoid Robotics: Evaluating Capabilities, Limitations, and Future Directions

Jalil Inayat-Hussain
22751096
School of Engineering
The University of Western Australia
Supervisor: Professor Thomas Bräunl

Word Count: 6830 words

November 28, 2025

Declaration of Contribution

My Contribution

I independently designed and developed the comprehensive benchmarking framework, including the five capability dimensions and five-level task complexity hierarchy. I completed the full technical implementation of NVIDIA's Isaac GR00T N1.5 on the Unitree G1 platform, including dataset preparation, three training iterations, deployment architecture, and safety system integration. This represents one of the first documented deployments of GR00T on the G1 platform, as neither NVIDIA nor Unitree had released official deployment code at the time. I conducted all experimental evaluation, including 90 benchmark trials and 30 height sensitivity trials, performed all data collection and analysis, and developed the research conclusions. Professor Thomas Bräunl provided supervisory guidance, laboratory access, and editorial feedback. This work contributed to a collaborative ACRA 2025 submission with co-authors Hongtao Zhang, Joel Smith, and Travis Ryan Ryan, who worked on complementary locomotion and navigation subsystems outside this thesis scope.

Use of AI Tools

I affirm that I have used AI tools solely to improve the quality of written English in this report, in accordance with the permitted uses outlined in course guidelines. These tools were used for proofreading and language enhancement purposes only, and not for generating original content.

In accordance with University Policy, I certify that:
The above information is correct, and the attached work submitted for assessment is my own work and that all material drawn from other sources has been fully acknowledged and referenced.

Student signature: _____ Date: _____

Supervisor confirmation
To the best of my knowledge, the student's contribution outlined above is correct.

Supervisor signature: _____ Date: _____

Project Summary

This research investigates the practical viability of transformer-based Vision–Language–Action (VLA) models as a foundation for embodied intelligence by deploying NVIDIA's Isaac GR00T N1.5 model on the Unitree G1 humanoid platform. A central contribution of this study is the design and implementation of a comprehensive benchmarking framework that evaluates VLA systems across five capability dimensions—perception, language understanding, planning, manipulation, and generalisation—over a hierarchical task structure of increasing complexity.

Through iterative model fine-tuning, system integration, and experimental evaluation across thirty-trial tasks, the study provides the first independent empirical assessment of Isaac GR00T's real-world performance. Results demonstrate strong object recognition and semantic understanding but reveal significant fragility in spatial reasoning, depth estimation, and generalisation to unseen tasks. Even minor deviations from trained configurations, such as a 10 cm change in table height or the introduction of distractor objects, led to complete task failure. These findings highlight a persistent gap between the high performance reported in controlled benchmarks and the inconsistent behaviour observed in real-world deployments.

The research concludes that while transformer-based VLA models represent a major conceptual advance in unifying perception and action sequences, their current embodiment-dependence and environmental sensitivity limit their deployment to structured, tightly calibrated domains.

List of Publications

Submitted for Publication

- Zhang, H., Inayat-Hussain, J., Smith, J., Ryan, T., and Braunl, T. (2025, December 2–5). *A Comprehensive Control Architecture for Humanoid Robots: Integration of Locomotion, Manipulation, and Navigation Subsystems.* 2025 Australasian Conference on Robotics and Automation (ACRA 2025), Perth, Australia. **[Submitted for publication]**.

Nomenclature

VLA Vision-Language-Action

Contents

List of Figures

List of Tables

1 Introduction and Background

1.1 Introduction

Humanoid robots represent the culmination of decades of technological advancement across electrical engineering, computer science, data science, and mechanical engineering [1]. These machines embody humanity's ambitious goal of creating mechanical counterparts that mimic human form and function [2]. While skepticism has historically surrounded the timeline for widespread domestic adoption, recent developments suggest this vision may materialize sooner than previously anticipated.

The commercial landscape reflects this shift. Figure AI, founded by serial entrepreneur Brett Adcock and backed by influential technology leaders including OpenAI and NVIDIA, is reportedly in discussions for new funding at a valuation of $39.5 billion—a remarkable increase from its $6.5 billion valuation just one year prior [3]. This substantial financial commitment signals a paradigm shift in the commercial viability of humanoid robotics.

This project evaluates the practical viability of deploying transformer-based Vision-Language-Action models in humanoid robots by implementing NVIDIA's Isaac GR00T model on the Unitree G1 humanoid platform and developing a benchmarking framework capable of quantitatively assessing both semantic understanding and physical execution in real-world conditions.

1.2 Background

The recent acceleration in humanoid robotics development stems from three critical technological advances: processing power for real-time complex operations [4], improved energy storage with lithium-based battery technologies [5], and transformer neural network architectures facilitating generalist robotics [6], [7]. Recent advancements in embedded computation, power systems, and machine learning have shifted the bottleneck from mechanical to cognitive capability [5].

At the core of this shift is the development of VLA models, which combine visual perception, natural language understanding, and action generation into a unified framework. These models are built by fine-tuning large pretrained language models on multimodal robotic datasets, allowing them to interpret human commands and generate control sequences for physical systems [8], [9]. Rather than merely classifying images or generating text, VLA models produce action tokens that can be mapped to low-level robot control [9].

Recent VLA implementations have adopted one of two dominant architectures. Google's Gemini Robotics uses a hybrid cloud-local design that offloads model inference to remote servers, allowing large models with minimal onboard processing but introducing concerns around latency [10], internet reliability, and data privacy [11]. NVIDIA's Isaac GR00T N1 executes inference locally on the Jetson Thor edge computing platform [12], [13], [14], ensuring responsiveness and operational independence but requiring efficient model compression and power-aware system design [15].

Despite notable progress, current VLA systems lack standardized evaluation frameworks for quantifying real-world performance. Most existing benchmarks focus on perception or language alone, overlooking the full action-perception loop critical to autonomous robotics. DeepMind's ASIMOV benchmark primarily evaluates a model's ability to recognize unsafe scenarios rather than generating appropriate action sequences [16]. Google's ERQA Framework assesses complex reasoning through visual question answering [17], but its multiple-choice format evaluates passive understanding rather than active decision-making and physical execution, limiting its applicability for evaluating complete VLA robotic systems.

This project addresses this limitation by developing a benchmarking framework that evaluates VLA model performance across both cognitive and physical dimensions under real-world constraints.

1.3 Project Objectives

This research is driven by the hypothesis that VLA models provide a viable path forward for humanoid robotics. To systematically investigate this hypothesis, the research will pursue the following objectives:

1. Conduct a comprehensive literature review of VLA models to assess the current state of the art, limitations and challenges, and the overall direction of generalist humanoid robotic models.

2. Deploy and fine-tune NVIDIA's Isaac GR00T N1 VLA model on the Unitree G1 humanoid, documenting technical challenges and implementation solutions.

3. Design and develop a quantitative benchmarking framework that evaluates both semantic understanding and physical execution capabilities of VLA models across diverse scenarios.

4. Assess the model's performance against the benchmarking framework using quantitative metrics.

5. Analyze the experimental results to determine whether VLA models offer a viable path forward for humanoid robotics.

The project's findings will provide a grounded, empirical contribution to the evaluation of generalist robotic systems, offering a repeatable methodology for assessing integrated VLA models in physical robots. The resulting insights are expected to be valuable to researchers, system architects, and industry practitioners involved in the development and deployment of humanoid AI systems.

2 Methodology

2.1 Research Approach

The research methodology is structured around four sequential phases, each building upon the findings of the previous stage to ensure rigorous evaluation and reproducible results. Due to the experimental nature of applying new and lowly documented technology a phased approach allows for iterative refinement of both the technical implementation and evaluation framework.

Phase 1: Literature Review and Foundation

The literature review establishes the theoretical foundation through a systematic examination of existing VLA models, their architectural approaches, and identified limitations. This phase includes a detailed analysis of current VLA benchmarking and evaluation frameworks, revealing significant gaps in standardized assessment methodologies. The full literature review is provided in Appendix A.

Phase 2: System Integration and Model Deployment

Focusing on the practical implementation of NVIDIA's Isaac GR00T N1 model on the Unitree G1 platform, this phase addresses real-world deployment challenges, including hardware integration, software compatibility, dataset preparation, and model fine-tuning. The iterative nature of this phase allows for documentation of technical challenges and solutions providing valuable insights to the broader understanding of VLA model deployment requirements.

Phase 3: Benchmarking Framework Development

A core contribution of this research, this phase involves developing a comprehensive evaluation framework that systematically assesses five key capability dimensions across five levels of task complexity. The framework is specifically designed to evaluate complete VLA systems rather than isolated coponents, addressing current gaps in existing evaluation methodologies.

Phase 4: Framework Validation and Evaluation of NVIDIA Isaac GR00T

Focusing on validating the benchmarking framework and applying it to the GR00T model, this phase includes comprehensive data collection and interpretation of results in context of the original hypothesis.

2.2 Hardware and Software Configuration

2.2.1 Unitree G1 Humanoid Robot

The Unitree G1 is a state-of-the-art humanoid platform featuring advanced locomotion and manipulation capabilities. Key specifications:

- Battery life: 2 hours continuous operation
- Maximum joint torque: 120 N·m
- Weight: 35 kg

Sensor Suite

- **Visual:** Intel RealSense D455 depth camera
- **Proprioceptive:** Joint encoders, 6-axis IMU, force/torque sensors
- **Tactile:** Pressure-sensitive fingertips and palm contact sensors

Manipulation System

The G1 features three-fingered dexterous hands enabling complex object manipulation with multiple degrees of freedom and precise force sensing.

2.2.2 Computing Infrastructure

Onboard Computing

The G1 utilizes dual Jetson Orin NX modules:

- PC1: Manages proprietary software and sensor interfaces
- PC2: Open platform for custom model implementation
- Specifications: 16GB memory, Arm Cortex A78AE CPU

See Appendix B for the complete specifications.

Cloud Computing

Lambda Cloud [18] will provide supplementary GPU resources when needed:

- A100 and H100 instances available
- Optimized for AI workloads with pre-configured ML environments

2.2.3 Software and Datasets

NVIDIA Isaac GR00T N1

NVIDIA's open foundation model for humanoid robot control [19] featuring:

- Multimodal input processing (vision, language, proprioception)
- Cross-embodiment capabilities
- End-to-end task learning

Manipulation Datasets

Comprehensive demonstration data including object grasping, pouring tasks, dual-arm coordination, and common household tasks with synchronized video, joint angles, and sensor readings [20] [21].

Development Tools

- PyTorch 2.0+
- CUDA Toolkit 12.4
- Isaac Sim
- Docker/Kubernetes
- Git/GitHub
- Jupyter Lab

2.3 Benchmarking Framework Design

This section presents a comprehensive benchmarking framework specifically developed to evaluate VLA models in humanoid robotics applications.

2.3.1 Framework Design Rationale

Current benchmarking approaches suffer from limitations that compromise their applicability to VLA-based humanoid systems. DeepMind's ASIMOV benchmark focuses primarily on semantic safety recognition rather than action generation, while Google's ERQA framework evaluates passive understanding through multiple-choice questions rather than active physical execution [16], [17]. These approaches fail to assess the integrated performance of vision, language, and action components in dynamic, real-world scenarios.

The proposed framework addresses these limitations by evaluating complete task execution from natural language instruction to physical completion. The framework employs a dual-axis evaluation structure: capability dimensions define what the system must be able to do, while the task complexity hierarchy systematically stresses these capabilities across increasing difficulty levels. Elements of this framework have been submitted for publication at the 2025 Australasian Conference on Robotics and Automation (ACRA) [22].

2.3.2 Capability Dimensions

The framework assesses five fundamental capability dimensions that collectively determine VLA model effectiveness. These dimensions represent orthogonal aspects of system performance that can be independently evaluated yet must function together for successful task execution:

1. **Perception Capabilities**: Environmental understanding through accurate visual scene interpretation, object recognition and property identification, spatial relationship comprehension, and consistent scene understanding across varying conditions. Assessment includes object detection accuracy, depth perception reliability, and scene understanding consistency.

2. **Language Understanding**: Semantic comprehension of task instructions, contextual reasoning about implied requirements, and interpretation of spatial and temporal references. Evaluation examines the ability to parse complex instructions, resolve ambiguous references, and maintain context across multi-step descriptions.

3. **Planning and Reasoning**: Hierarchical task planning, adaptive reasoning when encountering obstacles, and integration of perception feedback into ongoing planning. Assessment includes plan generation efficiency, adaptation to environmental changes, and recovery from execution failures.

4. **Manipulation Capabilities**: Physical interaction through dexterous object manipulation, force control precision, dual-arm coordination, and adaptation to object properties. Evaluation measures task completion accuracy, movement efficiency, and handling of unexpected physical constraints.

5. **Generalisation Performance**: Transfer of learned behaviours to novel scenarios, adaptation to new object categories, and reasoning about unfamiliar configurations. Assessment focuses on zero-shot task performance and behavioural consistency across novel scenarios.

2.3.3 Task Complexity Hierarchy

The framework employs a five-level task hierarchy that systematically increases complexity while maintaining clear evaluation criteria. Each level tests different combinations and intensities of the capability dimensions. Table 2.1 illustrates how each level engages the five capability dimensions with varying intensity.

Table 2.1: Mapping of capability dimensions across task complexity levels. Intensity ratings: Low (L), Medium (M), High (H), Critical (C).

Task Level	**Perception**	**Language**	**Planning**	**Manipulation**	**Generalisation**
Fundamental Operations	M	L	L	M	L
Sequential Coordination	M	M	H	M	L
Environmental Complexity	H	M	M	M	M
Adaptive Reasoning	H	H	H	M	H
Open-Ended Generalisation	H	H	C	M	C

1. **Fundamental Operations**: Basic tasks in controlled environments establish baseline capabilities for core perception-action loops, evaluating perception and manipulation under minimal planning demands.

   Example tasks: "Pick up the red cube," "Pick up the bottle," "Pick up the coffee cup."

   Success criteria are binary with standardised object sets, consistent lighting (320–400 lux) [23], and fixed camera positioning. Tasks involve single-step actions requiring object identification, simple command parsing, and stable grasps.

2. **Sequential Coordination**: Compound tasks requiring multi-step planning assess temporal reasoning integration with execution capabilities, significantly increasing demands on planning and reasoning.

   Example tasks: "Retrieve the mug from the shelf and place it on the table," "Stack the three blocks in order of increasing size," "Pick up the phone and place it next to the keyboard."

   Success requires completing all sub-tasks in logical sequence. The model must decompose instructions into ordered sub-goals, maintain working memory, and coordinate perception updates between sequential steps.

3. **Environmental Complexity**: Tasks with distractor objects and environmental variations stress perception capabilities and selective attention, testing robustness under increased sensory complexity.

   Example tasks: "Build a small tower using only the blue blocks while avoiding the red objects," "Retrieve the phone from the cluttered desk without disturbing other items," "Find and move the green cup from among the scattered objects."

4. **Adaptive Reasoning**: Tasks requiring reasoning about novel constraints evaluate higher-order cognitive capabilities across all dimensions, introducing implicit requirements and contextual reasoning.

   Example tasks: "Retrieve my phone without knocking over the glass of water next to it," "Move the laptop to a safe location away from the edge of the table," "Move these objects to make room for the large package."

5. **Open-Ended Generalisation**: Unstructured tasks requiring creative problem-solving test generalisation and reasoning limits through incomplete specifications with multiple valid solutions.

   Example tasks: "Organize these items in a logical way," "Help me prepare this workspace for the meeting," "Clean up this area and make it presentable."

This hierarchical structure enables systematic diagnosis of VLA model limitations. Failure at lower levels indicates fundamental deficiencies in core capabilities, while failure only at higher levels suggests limitations in reasoning or generalisation rather than basic competence. The explicit mapping between task levels and capability dimensions provides a principled framework for interpreting experimental results and identifying areas requiring improvement.

2.4 Model Training and Development Process

This section outlines the methodology used to adapt and deploy NVIDIA's Isaac GR00T N1 model on the Unitree G1 humanoid platform. The process followed an incremental pipeline designed to ensure compatibility between foundation-model-level capabilities and real-world robotic constraints.

2.4.1 Dataset Preparation

Training data is converted, if not already in the LeRobot format [24], to ensure compatibility with the Isaac GR00T N1 training framework. The LeRobot schema structures multimodal demonstration data—video, joint states, actions, and task descriptions—into synchronized episodes. Each episode represents a complete manipulation sequence from start to completion, enabling consistent temporal alignment between perception and control data.

2.4.2 Fine-Tuning Methodology

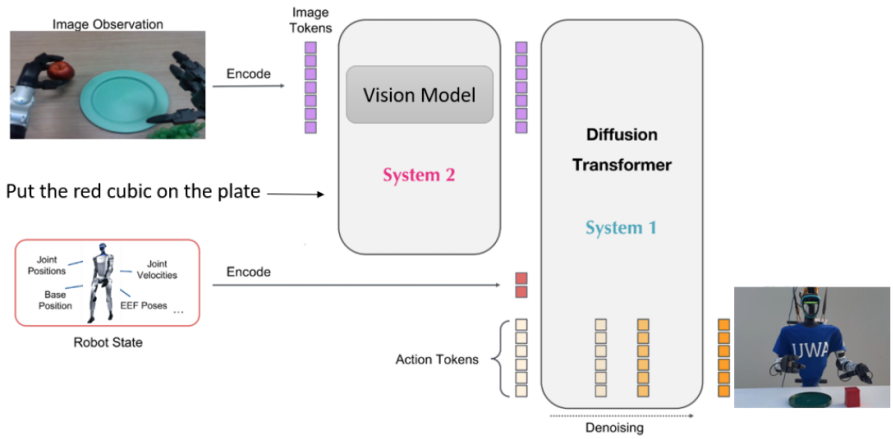


Figure 2.1: Model Architecture from paper submitted for publication at the 2025 Australasian Conference on Robotics and Automation (ACRA)[22].

2.4.3 Model Adaptation and Post-Training

The GR00T-N1-2B model is adapted to the Unitree G1 through a post-training process that preserves foundational vision-language capabilities while specializing action generation for the target embodiment's kinematic and sensory characteristics. The vision-language model backbone remains frozen throughout post-training to preserve language understanding and generalization capabilities, while components responsible for translating high-level understanding into embodiment-specific control signals are trained on Unitree G1 demonstration data [19].

Selective Component Training

Post-training targets two core components bridging general understanding and robot-specific execution. The flow matching transformer head, which functions as the action generation module, is refined to produce smooth and physically valid joint trajectories respecting the Unitree G1's kinematic constraints (see System 1 in Figure 2.1). This component denoises continuous action sequences through

iterative refinement, generating control commands that account for mechanical properties, joint limits, and motion smoothness. The projector layers connecting the vision-language backbone to the action head are trained to align semantic representations with embodiment-specific action and state spaces [19], translating high-level visual and linguistic features into proprioceptive and action representations appropriate for the Unitree G1's control interface.

The frozen vision-language backbone, comprising the visual encoder and language model (see System 2 in Figure 2.1), preserves instruction comprehension, semantic understanding of objects and spatial relationships, and reasoning capabilities developed during pretraining. This avoids catastrophic forgetting while enabling efficient adaptation to new embodiments with limited demonstration data. Post-training uses PyTorch scripts provided by NVIDIA's Isaac GR00T framework, with batch size and training duration adjusted based on computational resources and dataset size.

Training Configuration

Training employs the AdamW optimizer (learning rate 1×10^{-4}, weight decay 1×10^{-5}) with cosine annealing scheduling and early stopping based on validation performance. Training can be performed on RTX 4090 and above.

2.4.4 Deployment Strategy

Computational Architecture

Due to hardware limitations of the dual Jetson Orin NX modules onboard the G1, a hybrid client–server architecture is implemented. The robot acted as a client streaming sensory data to a remote GPU server running the model, which executes inference and returns action sequences in real time. This architecture provides the compute required for the VLA model while maintaining short response times between client and server nodes via ethernet.

2.4.5 Validation and Testing Protocol

Validation is conducted in three progressive stages to ensure system reliability before physical deployment.

Simulation Validation

Initial testing is performed in NVIDIA Isaac Sim, which replicated the G1's kinematics and sensors in a physics-accurate environment. Validation criteria includes:

- Task success rate across pick-and-place tasks.
- Joint limit compliance and motion smoothness.
- Safely returns to default position when model execution is terminated.

Successful completion in simulation is a prerequisite for any physical testing.

Controlled Physical Deployment

Physical validation follows a strict progression:

1. Successful move and hold default position.
2. Successful cancelation of program via G1 remote control.
3. Incremental manipulation with non-fragile objects, using three second sleeps in-between action sequences to ensure safe movement.

Each trial is conducted under supervision with dual-operator safety control, ensuring immediate manual intervention if abnormal behavior occurs.

Benchmarking Framework Integration

After successful simulation and physical deployment, the model will be evaluated using the proposed benchmarking framework. Each task will be executed over thirty independent trials to ensure statistical reliability and to account for variability in sensor perception, actuation precision, and environmental factors. The resulting performance data will be analysed to determine whether the observed outcomes support the underlying hypothesis that VLA models represent a viable path forward for general-purpose humanoid robotics.

3 Implementation

This chapter documents the practical deployment of NVIDIA's Isaac GR00T N1 model on the Unitree G1 humanoid platform. While the previous chapter established the theoretical methodology and experimental design, this chapter addresses the technical challenges encountered during actual system implementation and the solutions developed to overcome these obstacles. The implementation process revealed several critical challenges characteristic of deploying foundation models on physical robotic systems, providing essential insights into the engineering effort and adaptations required to achieve reliable operation in real-world conditions.

3.1 Deployment Architecture

3.1.1 Client–Server Infrastructure

The system was implemented using a distributed client–server architecture to support rapid experimentation while maintaining the capability for onboard execution, as outlined in the methodology. During deployment, the Unitree G1's dual Jetson Orin NX modules were found to provide sufficient computational resources for local Isaac GR00T N1 inference. However, the external server configuration was retained during development to enable faster iteration, debugging, and testing.

Client System (Unitree G1)

- Handles sensor data acquisition and action execution.
- Captures RGB images at 20 FPS.
- Transmits observations to the inference server.

Server System (RTX 4080 GPU):

- Receives client observations.
- Executes model inference.
- Returns joint position predictions for the next sixteen timesteps.

Observations followed the LeRobot format specification:

```
{
    "video.cam_right_high": camera_image,
    "state.left_arm": left_arm_state,
    "state.right_arm": right_arm_state,
```

```
    "state.left_hand": left_hand_state,
    "state.right_hand": right_hand_state,
    "annotation.human.task_description": [
        "place the red cube on the green plate"
    ],
}
```

Communication between the Unitree G1 and the external server was handled via ZeroMQ for inference data exchange, while Cyclone DDS managed local joint-state control on the physical robot.

3.1.2 Safety Mechanisms

Multiple independent safety mechanisms were integrated to ensure secure operation during experimental testing.

Manual Emergency Stop: Integration with the Unitree G1's handheld controller enables the operator to trigger an immediate safety shutdown via the *select* button.

Automated Safety Protocols: In the event of a software failure or communication loss, the system automatically executes a safety routine: returning to a neutral position followed by damping mode activation.

Multiple Intervention Points: System termination can be initiated either from the robot or the external control PC, ensuring that operator control is maintained regardless of network status.

3.2 Model Training and Development

Model development followed an iterative experimental approach, reflecting the need for retraining to refine stability and task performance. This section documents the evolution of the training process, the challenges encountered, and the modifications made to improve deployment robustness.

3.2.1 Training Infrastructure and Initial Pipeline Setup

A fine-tuning pipeline was established for the Unitree G1 dataset, as described in the methodology. Initial attempts to train locally on an RTX 4080 revealed computational constraints, with GPU memory capacity exceeded even after the batch size was reduced to sixteen. To overcome these limitations, model training was migrated to Lambda Cloud A100 instances, which provided sufficient VRAM and throughput for stable convergence.

This migration enabled the first successful fine-tuning run with the following configuration:

- Dataset: Unitree G1 Block Stacking Dataset
- Batch size: 32
- Steps: 10,000
- Final loss: 0.013

A set of evaluation plots for Model A, showing predictions across a single trajectory, is provided in Appendix C (Figure C.1a).

Training results are shown in Figure 3.1

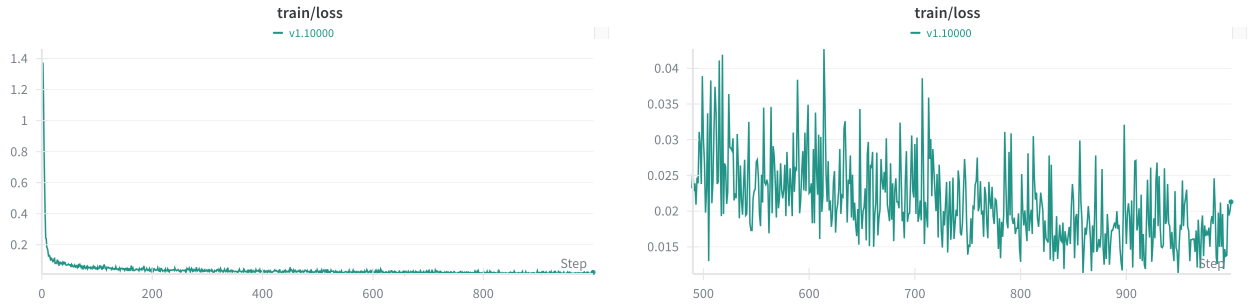
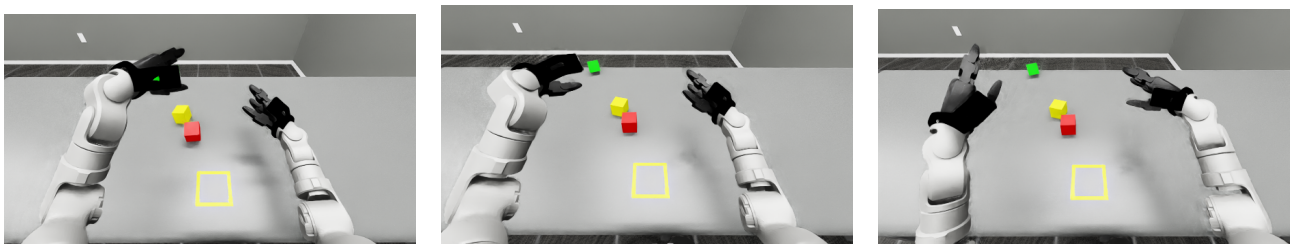


Figure 3.1: Loss curve and convergence for the first training attempt.

3.2.2 Simulation Validation and Initial Deployment



(a) Erratic motion example 1. (b) Erratic motion example 2. (c) Erratic motion example 3.

Figure 3.2: Examples of unstable robot behaviour observed during early deployments.

The first trained model was deployed in the IsaacLab simulation environment. Validation revealed significant behavioral issues:

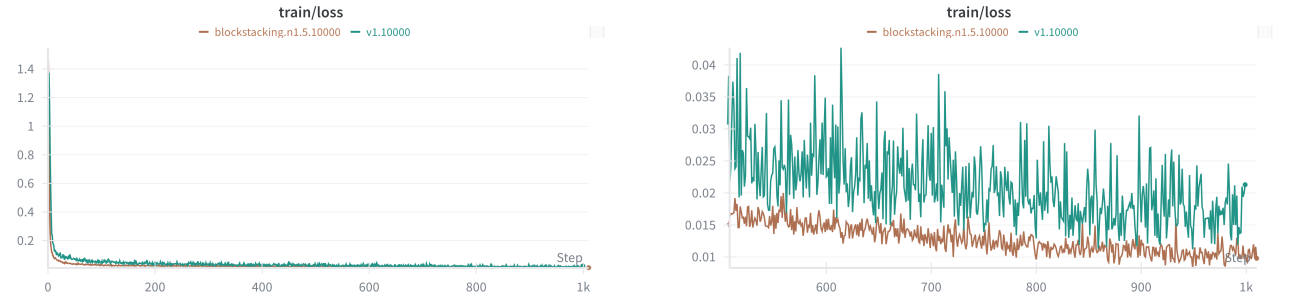
- Erratic arm movements
- Random hand actuation
- Failure to execute coherent manipulation sequences

At this stage, it was unclear whether the problem was due to the deployment code or insufficient training.

3.2.3 Second Model Training

Shortly after the initial model training, NVIDIA released Isaac GR00T N1.5, which introduced architectural improvements. Retraining with identical parameters yielded:

- Batch size: 64
- Final loss: 0.0086
- Improved convergence stability
- Smoother simulated motions



(a) Training loss (full scale).

(b) Training loss (zoomed view).

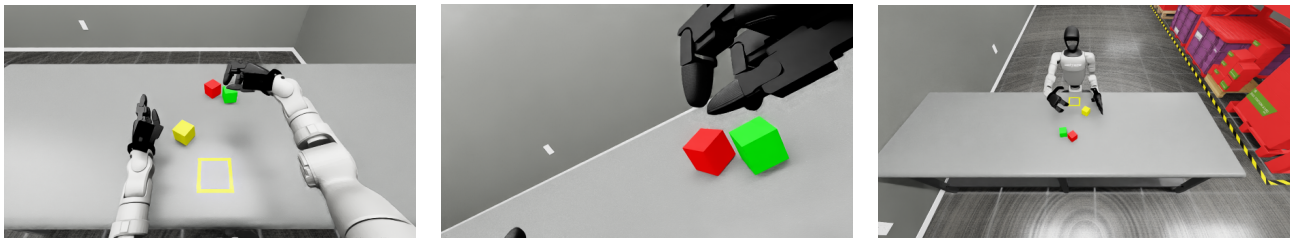
Figure 3.3: Comparison of training loss curves for the initial and the second fine tuning run.

Detailed evaluation results for Model B on the full dataset are included in Appendix C (Figure C.1b).

However, a new issue emerged: depth perception errors. The robot frequently grasped 10–15 cm above the cube. Investigation revealed discrepancies between the training dataset camera and the simulated IsaacLab camera:

- Training data: zoomed-in perspective
- Simulation: wider field of view

To mitigate this, the video modality scale parameter was adjusted. A scaling factor of 0.47 reduced the error, though grasps still occurred about 5 cm above the cube.



(a) Head camera view.

(b) Hand perspective.

(c) Front view.

Figure 3.4: Illustration of depth perception errors: the robot consistently grasped above the target cube across multiple viewpoints.

3.2.4 Deployment to the Physical Robot

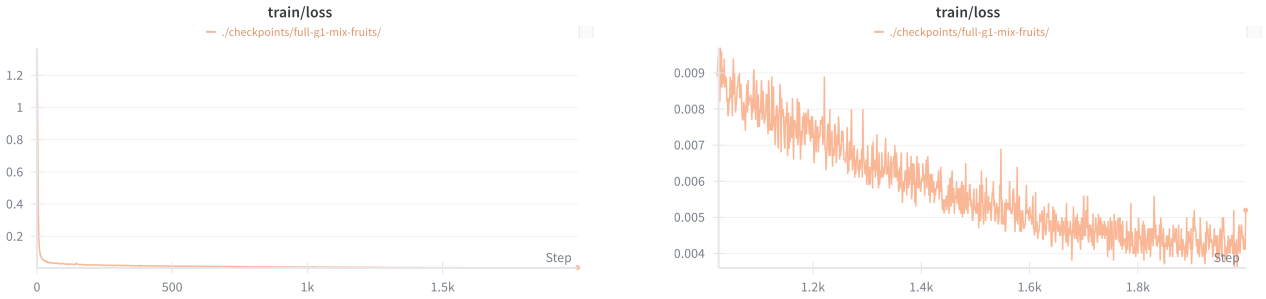
Validating that the model behaviour was safe in simulation, it was then deployed to the Unitree G1 robot to see if the results differed. Performance mirrored the simulation: movements were smoother, but depth perception issues persisted.

3.2.5 Final Model Training

To further improve performance, training was scaled up using the *PhysicalAI-Robotics-GR00T-Teleop-G1* dataset. This dataset contains 1,000 teleoperation trajectories of real Unitree G1 data, where the robot must pick the correct fruit (apple, pear, starfruit, grape) and place it on a plate based on a language prompt. The setup used:

- Batch size: 256 (per NVIDIA recommendations)
- GPU: H100 (upgraded from A100)
- Steps: 20,000
- Final loss: 0.0052 (best so far)

Appendix C presents the complete sequence-level evaluation plots for Model C (Figure C.1c).



(a) Training loss (full scale).

(b) Training loss (zoomed view).

Figure 3.5: Loss curve and convergence for final large-scale training.

Deployment results:

- In simulation: improved stability, but depth issue persisted ($\sim$5 cm offset).
- On physical robot: similar depth problems observed.

Upon inspection of the training images, it was discovered that training data lighting was clearer than the laboratory environment, where low lighting caused a rolling frequency distortion in the video feed, see Figure 3.6a.

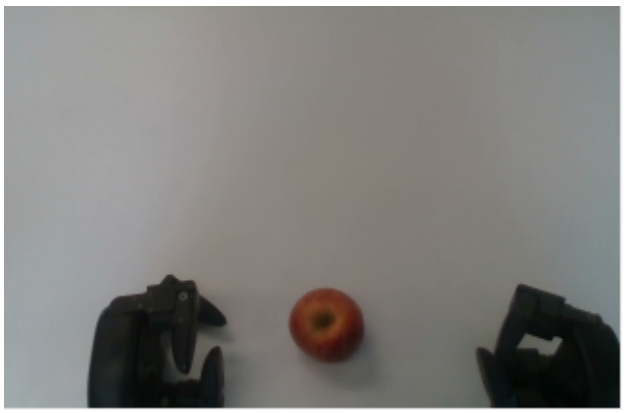
To mitigate this:

- The robot was tested outdoors under natural lighting $\rightarrow$ performance improved, though depth offset remained.
- The table height was raised in the lab $\rightarrow$ performance improved further.

With these adjustments, the robot successfully picked up the red cube and placed it on the plate both outdoors and indoors. Figure 3.7 demonstrates successful task execution using the fine-tuned Isaac GR00T N1 model on the Unitree G1 platform. The sequence shows the robot responding to the natural language instruction "pick up the cube and place it on the green plate" through integrated VLA processing.

3.3 Initial Model Performance Assessment

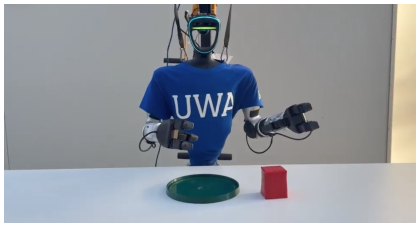
Following successful deployment, an initial performance assessment was conducted using fundamental manipulation tasks from Level 1 of the task complexity hierarchy under suboptimal laboratory conditions (250 lux lighting, 75 cm table height, 20 fps camera). Five object types were tested: red, black, and white cubes (7 cm $\times$ 7 cm $\times$ 7 cm), a ceramic coffee cup, and an apple (approximately 5 cm diameter).



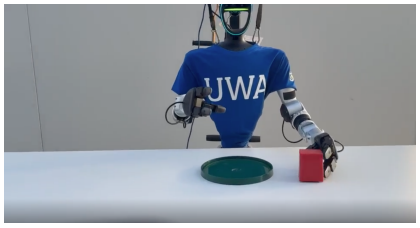
(a) Unitree G1 camera perspective during evaluation.

(b) NVIDIA training dataset camera perspective.

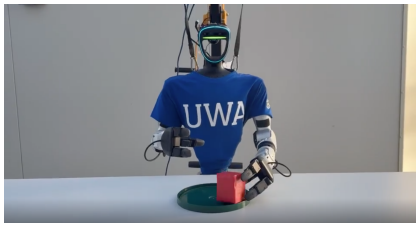
Figure 3.6: Visual perspective comparison deployment platform and training data collection.



(a) Initial scene analysis - Robot identifies target cube and destination plate



(b) Manipulation execution - Robot grasps cube with dexterous three-fingered hand



(c) Task completion - Robot places cube on designated plate

Figure 3.7: Isaac GR00T N1 VLA model executing pick-and-place task on Unitree G1 humanoid robot. The sequence demonstrates integrated VLA capabilities from scene understanding through manipulation to task completion.

3.3.1 Experimental Results

Table 3.1: Task 1.1: Pick up the black cube and hold it.

Trial	Success (Y/N)	Time (s)	Notes
1	Y	28	Picked up cube but dropped it; failed to maintain hold
2	N	–	Fingers jammed (hardware issue)
3	Y	34	Successful execution
4	N	–	Grasping failure
5	N	–	Grasping failure

The model achieved 40% success on cuboid objects, exclusively using the left arm and consistently executing pick-and-place sequences rather than hold instructions, suggesting overfitting to training data patterns.

Table 3.2: Task 1.2: Pick up the red apple.

Trial	Success (Y/N)	Time (s)	Notes
1	N	–	Picked up briefly, then dropped it
2	N	–	Approached apple but failed to locate precisely
3	N	–	Grasping failure
4	N	–	Grasping failure
5	N	–	Did not attempt with right arm

Table 3.3: Task 1.3: Grasp the coffee cup.

Trial	Success (Y/N)	Time (s)	Notes
1	N	–	Knocked cup over
2	N	–	Knocked cup over
3	N	–	Knocked cup over
4	N	–	Knocked cup over
5	N	–	Knocked cup over

Non-cuboid objects resulted in complete failure. The apple task showed the model reliably detected and approached objects but failed to account for spherical geometry, while the coffee cup task demonstrated inability to adapt to cylindrical geometry. The assessment revealed critical limitations: exclusive left arm use suggesting asymmetric training data, robust object detection paired with poor manipulation execution (grasp misalignment of 2 to 5 cm), and no adaptive behavior when initial strategies failed.

3.4 Summary of Implementation Challenges

The deployment process revealed several critical challenges. Initial training on local RTX 4080 hardware proved infeasible due to memory constraints, necessitating migration to cloud infrastructure. Three training iterations were conducted: the first achieved convergence (loss: 0.013) but exhibited severe behavioral instabilities; the second using Isaac GR00T N1.5 improved convergence (loss: 0.0086) but revealed systematic depth perception errors (10 to 15 cm offset); the final iteration using the PhysicalAI-Robotics-GR00T-Teleop-G1 dataset achieved the lowest loss (0.0052) but depth issues persisted.

Investigation revealed camera field of view discrepancies between training data and simulation environment. Adjusting the video modality scale parameter to 0.47 reduced depth offset to approximately 5 cm. Training data lighting conditions were superior to the laboratory environment, where low illumination caused rolling frequency distortion. Testing under natural outdoor lighting and raising table height improved performance, enabling successful pick-and-place execution. These findings demonstrated that reliable embodied performance requires careful alignment between training and deployment conditions.

4 Results

Due to the implementation challenges documented in Chapter 3, the formal benchmark evaluation was conducted under optimized conditions, including modifications to the lighting and table height. The hardware of the hand was also replaced to correct for thumb alignment issues that had caused unstable motion patterns and intermittent jitter during initial testing. This mechanical correction restored proper thumb rotation and produced marked improvements in motion smoothness and coordination. Lighting was increased to 1800 lux to eliminate the rolling frequency distortion observed under laboratory conditions, and table height was raised to 75cm giving the model the best chance for success. All subsequent performance evaluations were conducted using the upgraded hand configuration and optimal environmental setup, presenting systematic evaluation results using the benchmarking framework established in Chapter 2.

4.1 Isaac GR00T N1.5 Performance Evaluation

4.1.1 Baseline Performance Assessment

Table 4.1: Level 1 Baseline Performance – Grasp the Red Cube

Hand Configuration	**Trials**	**Success Rate**	**Mean Time (s)***	**Median Time (s)**	**Range (s)**
Left Hand	1–15	86.7% (13/15)	8.24 ± 1.66	8.03	6.06–11.31
Right Hand	16–30	100% (15/15)	5.57 ± 1.32	5.42	3.98–9.85
Overall	**1–30**	**93.3% (28/30)**	$\mathbf{6.81 \pm 2.00}$	**6.38**	**3.98–11.31**

*Mean times calculated from successful trials only.

Task: "Grasp the red cube."

Conditions: Standard lighting (1800 lux); Table height: 75 cm.

See Appendix D.1 for full trial-by-trial results.

The baseline evaluation demonstrated competent low-level manipulation performance, achieving a high overall grasp success rate of 93.3%. However, notable asymmetries were observed between the left and right hands. The right hand achieved perfect task completion (100%) with faster and more consistent execution times, whereas the left hand achieved 86.7% and displayed greater temporal variability. This disparity is likely attributable to dataset bias introduced during teleoperation-based data collection, where right-arm demonstrations were more frequent in the training dataset.

Additionally, both hands exhibited inconsistent thumb rotation patterns during grasping—an artefact likely originating from limitations in the teleoperation capture setup. Training data were recorded using head-mounted devices such as the Meta Quest or Apple Vision Pro, which provide only partial

hand pose fidelity. As a result, thumb movements were often misaligned or constrained, leading to unnatural and geometry-insensitive grasp configurations. Consequently, the model tended to apply a fixed grasp pattern regardless of object shape or orientation.

It was also observed that successful grasp execution was highly dependent on the object's spatial position relative to the robot. The cube needed to be placed within a narrow range for the model to perform successfully. When positioned too close to the torso or beyond the robot's natural reach, the arms exhibited restricted motion and were unable to adjust their position sufficiently to execute a grasp. This limitation likely reflects range constraints and positional biases present in the training dataset, highlighting the model's restricted operational workspace and limited generalisation to unseen spatial configurations.

While these findings confirm the model's ability to execute stable and repeatable grasping actions under controlled conditions, they also reveal its sensitivity to dataset biases and its limited adaptability to novel object geometries and spatial placements.

4.1.2 Task Complexity Effects

Table 4.2: Level 2 Performance – Pick Up Red Cube and Place on Green Plate

Hand Configuration	**Trials**	**Success Rate**	**Mean Time (s)***	**Median Time (s)**	**Range (s)**
Right Hand	1–15	73.3% (11/15)	10.64 ± 2.80	10.11	6.60–15.65
Left Hand	16–30	13.3% (2/15)	8.76 ± 0.11	8.87	8.65–8.87
Overall	**1–30**	**43.3% (13/30)**	$\mathbf{10.35 \pm 2.66}$	**9.30**	**6.60–15.65**

*Mean times calculated from successful trials only.

Task: "Pick up the red cube and place it on the green plate."

Conditions: Standard lighting (1800 lux); Table height: 75 cm.

See Appendix D.2 for full trial-by-trial results.

The transition to a two-step manipulation task resulted in a marked decline in overall performance, with success rates dropping to 43.3%. The right hand maintained moderate reliability (73.3%), whereas the left hand's success rate decreased substantially to 13.3%. Analysis of failure modes revealed distinct patterns between the two configurations: right-hand failures were primarily related to placement precision, often resulting in the cube resting near the edge of the plate or displacing the plate entirely. In contrast, left-hand failures were dominated by premature release events occurring immediately after a successful grasp.

These results suggest that the newly installed hand hardware introduced minor calibration inconsistencies relative to the model's learned grasping parameters. Small variations in joint offsets—particularly in thumb alignment—appear to have caused misinterpretation of grasp feedback, triggering unintentional object release. This behaviour underscores the fragility of current VLA control policies when subjected to embodiment changes, highlighting the importance of maintaining strict calibration consistency between hardware and training environments.

Furthermore, task success was highly dependent on the spatial position of the target plate. The model consistently attempted to place the cube in a fixed relative position, likely reflecting spatial biases present in the teleoperation training data. When the plate was positioned outside this expected

range, placement accuracy degraded significantly, further demonstrating the model's limited adaptability to unseen spatial configurations.

4.1.3 Environmental Object Complexity

Table 4.3: Level 3 Performance – Pick Up Red Apple and Place on Green Plate (with Distractors)

Hand Configuration	Trials	Success Rate	Mean Time (s)*	Median Time (s)	Range (s)
Right Hand	1–15	53.3% (8/15)	10.32 ± 1.15	10.26	8.64–12.86
Left Hand	16–30	20.0% (3/15)	10.52 ± 1.10	10.30	8.98–12.80
Overall	**1–30**	**36.7% (11/30)**	$\mathbf{10.38 \pm 1.14}$	**10.26**	**8.64–12.86**

*Mean times calculated from successful trials only.

Task: "Pick up the red apple and place it on the green plate."

Conditions: Standard lighting (1800 lux); Table height: 75 cm; Distractors: blue/red tape, black/white cubes.

See Appendix D.3 for full trial-by-trial results.

4.1.4 Task Complexity Level 3: Environmental Complexity

Performance on Level 3 declined sharply relative to the simpler cube-based tasks, with an overall success rate of 36.7%. The right hand achieved a moderate 53.3% success rate, demonstrating partial task competency, whereas the left hand completed only 20% of trials. Failures were primarily associated with grasp geometry mismatches—most notably, thumb-rotation and finger-closure errors when interacting with the apple's curved surface.

The introduction of distractor objects further reduced reliability, causing navigation conflicts in approximately 10% of trials. These disturbances highlighted the model's limited spatial filtering and weak visual attention mechanisms. Although the model successfully identified target objects based on colour and geometry, the robot frequently collided with nearby distractors, often knocking over adjacent items or catching its elbows on obstacles. Such behaviour indicates insufficient awareness of surrounding spatial context and a lack of collision-avoidance reasoning in the action generation process.

Overall, the Level 3 results reinforce the fragility of VLA-based control when transitioning from structured, idealised tasks to naturalistic environments. The model's reliance on specific visual and spatial priors—tuned to simplified laboratory conditions—significantly constrained its ability to generalise to realistic manipulation contexts.

4.1.5 Task Complexity Levels 4–5: Exclusion Rationale

Levels 4 and 5 of the benchmarking hierarchy—*Adaptive Reasoning* and *Open-Ended Generalisation*—were excluded from experimental evaluation. This decision was based on the model's inconsistent and environment-dependent performance across the lower-complexity tasks (Levels 1–3). The Isaac GR00T N1.5 model exhibited reliable execution only within narrowly constrained scenarios and showed clear degradation in accuracy as task complexity and environmental variability increased.

Under these conditions, advancing to higher-order reasoning tasks would likely have produced data dominated by compounding low-level control and perception errors rather than meaningful evidence of abstract reasoning capability. Since Levels 4 and 5 require contextual inference, multi-step

decision-making, and dynamic adaptation to novel goals, successful completion presupposes stable perception, grasping, and manipulation performance at foundational levels—criteria that the model did not consistently meet.

As a result, the research emphasis was deliberately redirected toward analysing the model's environmental sensitivity and depth-perception limitations rather than escalating task complexity. This pivot enabled a more systematic investigation into one of the core challenges observed throughout testing: the model's difficulty in accurately perceiving depth and maintaining spatial awareness in cluttered environments.

4.1.6 Environmental Sensitivity Analysis

Table 4.4: Height Sensitivity Analysis – Level 2 Task at Reduced Table Height

Configuration	**Table Height**	**Success Rate**	**Primary Failure Mode**
Baseline	75 cm	43.3% (13/30)	Variable
Reduced Height	65 cm	0% (0/30)	Grasping too high above object

Task: "Place the red cube on the green plate."
Conditions: Standard lighting (1800 lux); Table height: 65 cm.
See Appendix D.4 for full trial-by-trial results.

Performance deteriorated sharply when the table height was reduced by 10 cm from the training condition, with success rates collapsing from 43.3% to 0%. Nearly all failed attempts involved consistent depth overestimation, where the robot grasped well above the true object position. This finding demonstrates that even small geometric deviations from the training configuration significantly disrupt the model's learned spatial representations.

5 Discussion

5.1 Sources of Model Fragility

5.1.1 Hardware Configuration Dependencies

The GR00T N1.5 model demonstrated significant dependence on the exact mechanical configuration of the robot from the training data. Small variations in joint calibration or finger alignment produced large behavioural discrepancies, confirming that VLA policies remain highly embodiment-specific. The replacement of the Dex3-1 hands improved stability, yet highlighted how even hardware upgrades require re-alignment with learned control parameters to maintain performance.

5.1.2 Spatial Reasoning and Environmental Adaptation

Performance was further constrained by the model's inability to adapt to geometric or environmental variation. Reducing table height by just 10 cm resulted in total task failure, while the addition of clutter decreased success rates by more than 80%. These results imply that the model's internal spatial embeddings and depth estimation mechanisms are strongly conditioned on its training environment, lacking the adaptability necessary for deployment in variable real-world settings.

5.2 Comparison with NVIDIA's Results

In comparison, with the release of the GR00T-N1.5 model, NVIDIA reports performance benchmarks based on deployments using the Unitree G1 platform. According to their documentation, GR00T-N1.5, fine-tuned on 1,000 teleoperation episodes (the same dataset used in this study), achieved success rates of 98.8% on known objects and 84.2% on five novel objects in structured pick-and-place tasks [25].

When these results are compared to the findings of this research, several notable differences emerge. While the model demonstrates strong perceptual capabilities, accurate object recognition and scene understanding-its physical execution proves highly sensitive to even minor variations in object geometry or hand calibrations. For instance, a 10 cm change in table height resulted in consistent task failure across all 30 trials, and slight misalignments in finger positioning when changing hands, produced erratic and unstable movements. NVIDIA's reported results, while notable, should therefore be understood as reflecting an upper bound of performance—achievable only under tightly controlled conditions that closely match the original training setup.

5.3 Comparison with Related Work

Furthermore, when reviewing current Vision–Language–Action (VLA) literature, a clear divide emerges between research that proposes new model architectures and research that evaluates their reliability. Architecture-focused papers often present highly optimistic results, emphasizing generalization, emergent capabilities, and semantic understanding. For instance, Google's RT-2 model [9] reports strong zero-shot generalization by combining web-scale vision-language training with robotic control—similar to NVIDIA's GR00T. Likewise, Octo, an open-source generalist robot policy trained on over 800,000 trajectories [26], demonstrated promising cross-robot transfer through fine-tuning, akin to the approach taken in this research. However, these studies often downplay physical reliability, overfitting to training environments, and the real-world deployment challenges highlighted in this work.

In contrast, benchmarking and evaluation studies paint a more critical picture. Guruprasad et al. [27] introduced the MultiNet v0.1 benchmark, which evaluates state-of-the-art Vision-Language models like GPT-4o and VLA models such as OpenVLA and JAT across twenty robotic manipulation datasets from the Open X Embodiment corpus. Their findings reveal significant variation in model performance across tasks, robot embodiments, and environmental conditions—with no model demonstrating consistent robustness. They also report high sensitivity to action-space representation and environmental factors, mirroring the patterns observed in this research.

Taken together, these trends expose a gap between architectural optimism and empirical robustness. Many models excel only under benchmark conditions that mirror their training setups, leading to inflated claims of generalization. Under systematic evaluation or real-world perturbations, their weaknesses become apparent: unstable execution, embodiment sensitivity, and brittleness to minor environmental changes. The findings of this study align with this latter view, showing that even state-of-the-art models such as GR00T-N1.5 operate reliably only within narrow tolerances and fail abruptly when those conditions shift.

5.4 Link to Research Hypothesis

The findings discussed in this chapter directly address the central hypothesis of this thesis: whether Vision–Language–Action (VLA) models offer a viable path forward for humanoid robotics. The results indicate that while models such as GR00T-N1.5 demonstrate strong perceptual understanding and task recognition, their physical execution remains brittle under even modest real-world variation. Across both experimental and comparative analyses, VLA systems consistently exhibit the same pattern—semantic competence paired with embodiment fragility.

Therefore, the evidence suggests that current VLA models represent a promising conceptual framework, but not yet a fully viable solution for reliable humanoid control. Their potential lies in the integration of perception, language, and reasoning, but achieving robust physical behaviour will require advances in calibration, adaptive control, and environment-invariant learning.

5.5 Research Limitations

Several methodological and technical constraints influence the scope and generalizability of these findings.

5.5.1 Environmental and Hardware Constraints

The experimental evaluation deviated from initial methodological specifications. While the framework design specified lighting levels of 320 to 400 lux based on workplace standards [23], formal benchmark evaluation utilized 1800 lux illumination to eliminate rolling frequency distortion, limiting direct comparability with systems evaluated under standard conditions.

The replacement of Dex3-1 hand assemblies during the research timeline introduced calibration uncertainties. Premature object releases during Level 2 tasks suggest minor joint offset variations between original and replacement hands may have caused degraded model performance, representing a confounding variable that complicates isolation of model performance from embodiment calibration effects.

5.5.2 Experimental Design Limitations

Evaluation was restricted to the first three task complexity levels due to inconsistent foundational manipulation performance, resulting in incomplete framework validation. Statistical rigor was constrained by practical trial limitations, with baseline evaluation consisting of 30 trials yielding 15 observations per hand configuration.

The evaluation focused exclusively on Isaac GR00T N1.5 deployed on a single Unitree G1 platform. This narrow scope limits generalizability to other VLA architectures or alternative humanoid platforms, as the observed fragility may represent characteristics specific to GR00T N1.5 rather than fundamental limitations of VLA approaches generally.

5.5.3 Training Data and Deployment Alignment

The research utilized pre-existing training datasets rather than custom demonstration data tailored to experimental conditions. The PhysicalAI-Robotics-GR00T-Teleop-G1 dataset reflected environmental conditions and hardware configurations differing from the evaluation setup in camera perspective, lighting characteristics, table height, and hand calibration. These configuration differences likely contributed to performance degradation under environmental perturbations. Additionally, teleoperation-based data collection introduced systematic biases, with human demonstrators exhibiting clear hand preference patterns that likely explain observed performance disparities between hands.

5.6 Future Work

Building on the findings of this research, several directions for continued investigation are recommended to advance the reliability and adaptability of VLA models for humanoid robotics:

1. **Training on Custom and Diverse Data:** Future experiments should re-train the model on datasets that extend beyond structured pick-and-place episodes to include deliberate environmental variations—such as differing table heights, cluttered workspaces, and variable lighting conditions. Incorporating such diversity would enable the model to develop more robust spatial and perceptual representations, improving adaptability to non-ideal and dynamic environments. It would also be valuable to assess how performance changes when the model is both trained and deployed on the same physical robot, reducing embodiment mismatches.

2. **Incorporating Depth and Multi-View Perception:** One of the major limitations identified in this study was the model's unreliable depth perception during grasping. Future work should explore architectural adjustments that explicitly integrate depth information, potentially through stereo vision systems or additional wrist-mounted cameras. These enhancements could improve the model's understanding of object geometry and distance, leading to more stable and precise manipulation.

3. **Evaluating Alternative VLA Architectures:** Applying the developed benchmarking framework to alternative VLA and multimodal models would help determine whether the observed fragility is specific to GR00T-N1.5 or general across architectures. Comparative evaluation could reveal common weaknesses and guide the development of more robust cross-embodiment learning strategies.

4. **Standardised Benchmarking Framework Development:** Expanding the current benchmarking framework into an open-source, community-driven standard would promote reproducibility and enable meaningful cross-institutional comparison. Establishing shared protocols for data collection, evaluation metrics, and environmental perturbation testing would significantly accelerate progress toward reliable humanoid autonomy.

6 Conclusion

6.1 Summary of Findings

This research set out to evaluate the practical viability of VLA models for humanoid robotics through the deployment of NVIDIA's Isaac GR00T N1.5 on the Unitree G1 platform. The project addressed two primary objectives: the implementation of a transformer-based generalist model on a physical humanoid system, and the development of a benchmarking framework capable of quantitatively assessing model performance across perception, language understanding, and manipulation domains.

Experimental results demonstrated that GR00T N1.5 exhibits strong perceptual recognition and semantic understanding, aligning with trends observed in comparable VLA models such as RT-2 [9], OpenVLA [8], and Octo [26]. However, its real-world execution was constrained by embodiment-specific dependencies, limited spatial generalisation, and high sensitivity to environmental variation. The model achieved reliable performance only within narrowly defined operational configurations—such as consistent object positioning and fixed table height—while even modest deviations in geometry or lighting led to severe performance degradation.

These results highlight a significant disparity between industry-reported success rates (e.g., NVIDIA's 98.8% post-training success on Unitree G1 [25]) and the empirical reliability observed under real-world experimental conditions. The discrepancy underscores the necessity for rigorous, standardised benchmarking frameworks that test beyond idealised laboratory scenarios.

6.2 Research Contributions

The project's contributions can be summarised as follows:

1. **Benchmarking Framework:** A comprehensive evaluation framework was designed to assess humanoid VLA models across multiple capability dimensions—perception, language understanding, planning, manipulation, and generalisation. This provides a structured methodology for future performance assessment under controlled but progressively complex conditions.

2. **Unitree G1 Deployment Code:** Although not originally intended as a core research outcome, one of the most significant contributions of this project was the successful deployment of NVIDIA's GR00T model onto the Unitree G1 humanoid platform. At the time of writing, neither NVIDIA nor Unitree has released official deployment code for running VLA models on the G1. The implementation developed in this research therefore represents one of the first publicly demonstrated end-to-end deployments of GR00T on this platform. This code has since been adopted by multiple researchers across both academic institutions and industry, providing a practical foundation for further experimentation and benchmarking of humanoid VLA systems.

3. **Empirical Evaluation of GR00T N1.5:** The implementation and testing of Isaac GR00T N1.5 on a Unitree G1 humanoid yielded empirical evidence on the system's strengths and limitations, providing one of the first independent validations of NVIDIA's claimed model capabilities.

4. **Identification of Critical Limitations:** The findings isolate key bottlenecks in current VLA systems—namely embodiment dependence, spatial fragility, short-horizon reasoning, and reliance on dataset alignment—which together constrain scalability to real-world environments.

6.3 Implications for Humanoid Robotics

The outcomes of this study reinforce the broader consensus emerging across robotics research: while transformer-based architectures represent a major step toward unifying perception, language, and action, they are not yet capable of robust, general-purpose autonomy. The observed failures under minor environmental changes demonstrate that current humanoid systems remain highly sensitive to training conditions and mechanical calibration.

For the humanoid robotics industry, these results suggest that near-term commercial deployment of generalist VLA systems will be feasible only within tightly controlled operational domains—such as structured manufacturing or warehouse environments—rather than unstructured domestic settings. Progress toward open-world adaptability will depend on advances in multimodal data diversity, improved cross-embodiment transfer learning, and hybrid architectures that integrate symbolic reasoning with transformer-based learning.

6.4 Concluding Remarks

In conclusion, this research contributes to the growing body of evidence that while VLA models such as Isaac GR00T N1.5 represent a pivotal advance in embodied AI, they remain limited by embodiment dependence and fragile real-world adaptability. The proposed benchmarking framework and empirical findings serve as a foundation for future investigations into robust, general-purpose humanoid intelligence. The path to truly autonomous humanoid systems will likely hinge on bridging the current gap between model intelligence and embodied reliability—an intersection where engineering precision and data-driven learning must evolve together.

References

[1] F. R. Noreils, *Humanoid robots at work: Where are we?* 2024. arXiv: 2404.04249 [cs.CY]. [Online]. Available: https://arxiv.org/abs/2404.04249.

[2] J. Wang, C. Wang, W. Chen, Q. Dou, and W. Chi, "Embracing the future: The rise of humanoid robots and embodied AI," *Intelligence & Robotics*, vol. 4, no. 2, pp. 196–209, 2024, ISSN: 2770-3541. DOI: 10.20517/ir.2024.12. [Online]. Available: https://www.oaepublish.com/articles/ir.2024.12.

[3] Reuters. "Robotics startup Figure AI in talks for new funding at $3.95 billion valuation — Bloomberg," Accessed: 2025-04-18. [Online]. Available: https://www.reuters.com/technology/artificial-intelligence/robotics-startup-figure-ai-talks-new-funding-395-billion-valuation-bloomberg-2025-02-14/.

[4] A. Fijany and A. K. Bejczy, Eds., *Parallel Computation Systems for Robotics: Algorithms and Architectures* (World Scientific Series in Robotics and Automated Systems). Singapore; New Jersey; London; Hong Kong: World Scientific, 1992, vol. 2, Includes bibliographical references, ISBN: 9810206631. [Online]. Available: https://www.worldscientific.com/doi/pdf/10.1142/9789814360210_fmatter.

[5] Boston Dynamics. "Electric: New era for Atlas," Accessed: 2025-04-19. [Online]. Available: https://bostondynamics.com/blog/electric-new-era-for-atlas/.

[6] NVIDIA Corporation. "NVIDIA announces Isaac GR00T N1 — the world's first open humanoid robot foundation model — and simulation frameworks to speed robot development." Press release, Accessed: 2025-04-27. [Online]. Available: https://nvidianews.nvidia.com/news/nvidia-isaac-gr00t-n1-open-humanoid-robot-foundation-model-simulation-frameworks.

[7] VentureBeat. "UC Berkeley's transformer-based robot control system generalizes to unseen environments," Accessed: 2025-04-19. [Online]. Available: https://venturebeat.com/ai/uc-berkeleys-transformer-based-robot-control-system-generalizes-to-unseen-environments/.

[8] M. J. Kim et al., *OpenVLA: An open-source vision-language-action model*, 2024. arXiv: 2406.09246 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2406.09246.

[9] A. Brohan et al., *RT-2: Vision-language-action models transfer web knowledge to robotic control*, 2023. arXiv: 2307.15818 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2307.15818.

[10] D. Uchechukwu, A. Siddique, A. Maksatbek, and I. Afanasyev, "ROS-based integration of smart space and a mobile robot as the Internet of Robotic Things," in *2019 25th Conference of Open Innovations Association (FRUCT)*, IEEE, Nov. 2019, pp. 339–345. DOI: 10.23919/fruct48121.2019.8981532. [Online]. Available: http://dx.doi.org/10.23919/FRUCT48121.2019.8981532.

[11] B. Ulsmaag, J.-C. Lin, and M.-C. Lee, *Investigating the privacy risk of using robot vacuum cleaners in smart environments*, 2024. arXiv: 2407.18433 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2407.18433.

[12] Y. Zhu, L. Fan, and NVIDIA GEAR Team, "NVIDIA Isaac GR00T N1: An open foundation model for humanoid robots," NVIDIA, White Paper, 2025. [Online]. Available: https://d1qx31qr3h6wln.cloudfront.net/publications/GR00T_1_Whitepaper.pdf.

[13] NVIDIA Newsroom. "NVIDIA announces Project GR00T foundation model for humanoid robots and major Isaac robotics platform update," Accessed: 2025-05-03. [Online]. Available: https://nvidianews.nvidia.com/news/foundation-model-isaac-robotics-platform.

[14] NVIDIA Corporation. "NVIDIA Blackwell architecture technical overview," Accessed: 2025-04-26. [Online]. Available: https://resources.nvidia.com/en-us-blackwell-architecture.

[15] S. Saha and L. Xu, *Vision transformers on the edge: A comprehensive survey of model compression and acceleration strategies*, 2025. arXiv: 2503.02891 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2503.02891.

[16] P. Sermanet, A. Majumdar, A. Irpan, D. Kalashnikov, and V. Sindhwani, *Generating robot constitutions & benchmarks for semantic safety*, 2025. arXiv: 2503.08663 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2503.08663.

[17] Gemini Robotics Team et al., *Gemini robotics: Bringing AI into the physical world*, 2025. arXiv: 2503.20020 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2503.20020.

[18] Lambda Labs. "Lambda cloud: GPU cloud for deep learning," Accessed: 2025-04-27. [Online]. Available: https://lambdalabs.com/cloud.

[19] NVIDIA Corporation. "NVIDIA Isaac GR00T N1: Open foundation model for generalized humanoid robot reasoning and skills." GitHub repository. Licensed under Apache License 2.0, Accessed: 2025-04-27. [Online]. Available: https://github.com/NVIDIA/Isaac-GR00T.

[20] Unitree Robotics. "G1 Dual Arm Grasping Dataset." Dataset created using LeRobot. Available on Hugging Face, Accessed: 2025-04-27. [Online]. Available: https://huggingface.co/datasets/unitreerobotics/G1_DualArmGrasping_Dataset.

[21] Unitree Robotics. "Z1 Dual Arm Pour Coffee Dataset." Dataset records joint angle and gripper data for the Z1 dual-arm robot. Available on Hugging Face, Accessed: 2025-04-27. [Online]. Available: https://huggingface.co/datasets/unitreerobotics/Z1_DualArm_PourCoffee_Dataset.

[22] H. Zhang, J. Inayat-Hussain, J. Smith, T. Ryan, and T. Braunl, "A comprehensive control architecture for humanoid robots: Integration of locomotion, manipulation, and navigation subsystems," in *Proceedings of the 2025 Australasian Conference on Robotics and Automation (ACRA)*, Submitted for publication, Perth, Australia, 2025.

[23] WorkSafe Queensland. "Lighting — WorkSafe Queensland guidelines," Accessed: 2025-10-05. [Online]. Available: https://www.worksafe.qld.gov.au/safety-and-prevention/hazards/workplace-hazards/dangers-in-your-workplace/lighting.

[24] R. Cadene, S. Alibert, et al., *Lerobot: State-of-the-art machine learning for real-world robotics*, 2024. [Online]. Available: https://github.com/huggingface/lerobot.

[25] NVIDIA Research, GEAR Lab. "GR00T N1.5: A vision-language-action model for general-purpose humanoid control," Accessed: 2025-10-05. [Online]. Available: https://research.nvidia.com/labs/gear/gr00t-n1_5/.

[26] Octo Model Team et al., *Octo: An open-source generalist robot policy*, 2024. arXiv: 2405.12213 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2405.12213.

[27] P. Guruprasad, H. Sikka, J. Song, Y. Wang, and P. P. Liang, *Benchmarking vision, language, & action models on robotic learning tasks*, 2024. arXiv: 2411.05821 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2411.05821.

[28] C. Caiazza, S. Giordano, V. Luconi, and A. Vecchio, "Edge computing vs centralized cloud: Impact of communication latency on the energy consumption of LTE terminal nodes," *Computer Communications*, vol. 194, pp. 213–225, Oct. 2022, ISSN: 0140-3664. DOI: 10.1016/j.comcom.2022.07.026. [Online]. Available: http://dx.doi.org/10.1016/j.comcom.2022.07.026.

[29] N. Sanghai and N. B. Brown, *Advances in transformers for robotic applications: A review*, 2024. arXiv: 2412.10599 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2412.10599.

[30] Z. Yang, C. Garrett, D. Fox, T. Lozano-Pérez, and L. P. Kaelbling, *Guiding long-horizon task and motion planning with vision language models*, 2024. arXiv: 2410.02193 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2410.02193.

A Literature Review: The Dawn of Humanoid Robotics

A.1 Introduction

Humanoid robots represent the culmination of decades of technological advancement across electrical engineering, computer science, data science, and mechanical engineering [1]. These machines embody humanity's ambitious—and to some, unsettling—goal of creating mechanical counterparts that mimic human form and function [2]. While skepticism persists regarding the timeline for widespread domestic adoption of robots capable of everyday tasks such as folding laundry, organizing dishes, or preparing meals, several companies are accelerating development to make this vision a reality sooner than previously anticipated. Among these ventures, Figure AI, founded by serial entrepreneur Brett Adcock, has emerged as a particularly promising contender. Backed by influential technology leaders including OpenAI and NVIDIA, Figure is reportedly in discussions for new funding at a valuation of $39.5 billion—a remarkable increase from its $6.5 billion valuation just one year prior [3]. This substantial financial commitment from industry titans signals a paradigm shift in the commercial viability of humanoid robotics, suggesting that consumer applications may materialize in the relatively near future. This recent acceleration raises an important question: Why now? The answer lies in three critical technological bottlenecks that have now been substantially addressed. First, the processing power necessary to execute complex robotic operations in real-time has advanced significantly [4]. Second, improvements in energy storage—particularly lithium-based battery technologies—have yielded power sources with the energy density and longevity required for practical mobile robotics [5]. Third, and perhaps most transformative, has been the development of transformer neural network architectures, which are believed to facilitate generalist robotics—an approach strongly endorsed by NVIDIA's CEO Jensen Huang, who declared that "the age of generalist robotics is here" according to recent announcements [6] [7].

A.2 Vision-Language-Action Models: A Paradigm Shift

Transformer models have emerged as the critical enabling technology for robotics, inspiring a new class of models known as vision-language-action (VLA) models. These models leverage foundation large language models pretrained on internet-scale datasets, which equips them with semantic reasoning, problem-solving, and visual interpretation capabilities—making them exceptionally valuable for generalist robotic applications [8]. These foundation models are subsequently fine-tuned on robot datasets to output appropriate actions. By modifying these models to integrate directly with low-level robotic control systems and produce 'action tokens' instead of text tokens, robots can translate natural language commands directly into physical task execution [9].

A.3 Architectural Approaches to Vision Language Action Models

As of writing, several major technology companies have developed their own VLA models, each with different architectural approaches and design philosophies. Two notable examples include Google's Gemini Robotics and NVIDIA's Isaac GR00T N1, which represent different approaches to implementing VLA capabilities in humanoid robots.

A.3.1 Cloud-Based Architecture: Google's Gemini Robotics

Google released their Gemini Robotics foundation models in March 2025, introducing an architecture comprised of a cloud backbone and a local decoder with a claimed 250ms end-to-end latency between raw observations and low-level execution [17]. This hybrid architecture presents both advantages and challenges. On the positive side, offloading the model inference to cloud infrastructure reduces on-device power consumption, allowing battery capacity to be allocated to the mechanical operations rather than processing [28]. However, this approach introduces potential downsides: network latency can impact real-time performance, and any disruption in internet connectivity could severely impair functionality [10]. The cloud-based approach also raises questions about privacy and data security, as sensitive environmental information must be transmitted to external servers for processing [11].

A.3.2 Edge Computing Approach: NVIDIA's Isaac GR00T N1

In contrast to Google's cloud-dependant appproach, NVIDIA has released an open-source VLA model, Isaac GR00T N1, designed for local execution. Their architecture divides processing between two systems: a Vision-Language model for environmental perception and interpretation, paired with an action generation system that translates high level goals into specific movements [12]. To support this edge-computing approach, NVIDIA has developed the Jetson Thor computing platform, built on their Blackwell architecture and featuring 800 TOPS of 8-bit floating point performance [13] [14]. This hardware is specifically engineered to deliver low-latency, high-throughput inference for transformer models directly on the robot, eliminating the need for constant cloud connectivity. The edge computing approach offers several advantages: reduced latency in time-sensitive operations, continued functionality during network outages, and enhanced privacy as data remains on the device. However, it faces challenges in terms of power consumption and computational constraints compared to virtually unlimited cloud resources.

A.4 Current Limitations and Challenges

Despite the promising advancements in VLA models for humanoid robotics, several significant limitations persist:

A.4.1 Edge Computing Constraints

While NVIDIA's Jetson Thor represents a significant advancement in edge computing for robotics, the ability to run transformer models with billions of parameters on battery-powered devices remains challenging [15]. Trade-offs between model size, inference speed, and power consumption continue to constrain deployment options.

A.4.2 Data Collection Bottlenecks

The core challenge in this field is that training robust VLA models requires vast amounts of demonstration data showing language commands paired with appropriate robotic actions in diverse scenarios [29]. This data collection process is extraordinarily expensive and time-consuming and is limited to the 24 hours you have in a day, although, Nvidia Researchers have pioneered methods of synthesizing robot training data from real data collected, allowing multiple angles of the same shot to be produced. [19].

A.4.3 Long-Term Memory Integration

Current VLA models excel in responding to immediate instructions, but struggle with maintaining context over extended interactions or remembering past experiences to inform future actions. Integrating effective long-term memory systems remains an open research challenge, with some researchers proposing hierarchical planning algorithms for long-horizon planning [30]

A.4.4 Generalization Across Embodiments

Models trained on one robotic platform often struggle to transfer learned behaviors to robots with different physical configurations, joint arrangements, or actuation mechanisms. This lack of generalization between embodiments complicates deployment across various hardware platforms [9].

A.5 The Need for Comprehensive Benchmarking

While several benchmarking frameworks exist for evaluating specific capabilities of robotic systems, there remains a significant gap in comprehensive evaluation methodologies for integrated Vision-Language-Action (VLA) models in humanoid robotics. Current benchmarking approaches fall short in several critical dimensions:

A.5.1 DeepMind's ASIMOV Benchmark

DeepMind's ASIMOV benchmark, while innovative in its focus on semantic safety, primarily evaluates a model's ability to recognize unsafe scenarios rather than generating appropriate action sequences [16]. The benchmark lacks assessment of physical embodiment constraints and fails to evaluate performance in dynamic, evolving environments where robots must continually adjust their actions based on real-time feedback.

A.5.2 Google's ERQA Framework

Google's ERQA Framework advances beyond atomic capabilities to assess complex reasoning through visual question answering. However, its multiple-choice format evaluates passive understanding rather than active decision-making and physical execution [17]. It was designed specifically for Visual Language Models rather than complete Visual Language Action systems, limiting its applicability for evaluating full robotic systems.

A.5.3 Research Direction

As discussed, huge financial investments are flowing into humanoid robotics with a strong industry belief that Vision-Language-Action models represent the optimal path forward for generalised humanoid robots. Despite this enthusiasm, fundamental questions remain unanswered: Are these models truly viable for practical humanoid applications? Is insufficient robot training data the primary bottleneck? Or do the models themselves contain inherent architectural limitations?

This research aims to investigate these questions and test the hypothesis that VLA models are the way forward for humanoid robotics. To accomplish this, NVIDIA's Isaac GR00T N1 will be deployed on a Unitree G1 humanoid. However, to properly evaluate the system's performance, we must first address the critical gap in evaluation methodologies for VLA models.

The primary contribution of this work will be the development of a comprehensive benchmarking framework specifically to assess the performance of on-board VLA models. By applying this benchmarking methodology to evaluate the Isaac GR00T N1 implementations on the Unitree G1, we will generate empirical evidence to address the central hypothesis regarding VLA models as a foundation for humanoid robotics. The results will provide critical insights into current capabilities, limitations and potential directions for future development in this rapidly evolving field.

B NVIDIA Jetson Orin NX

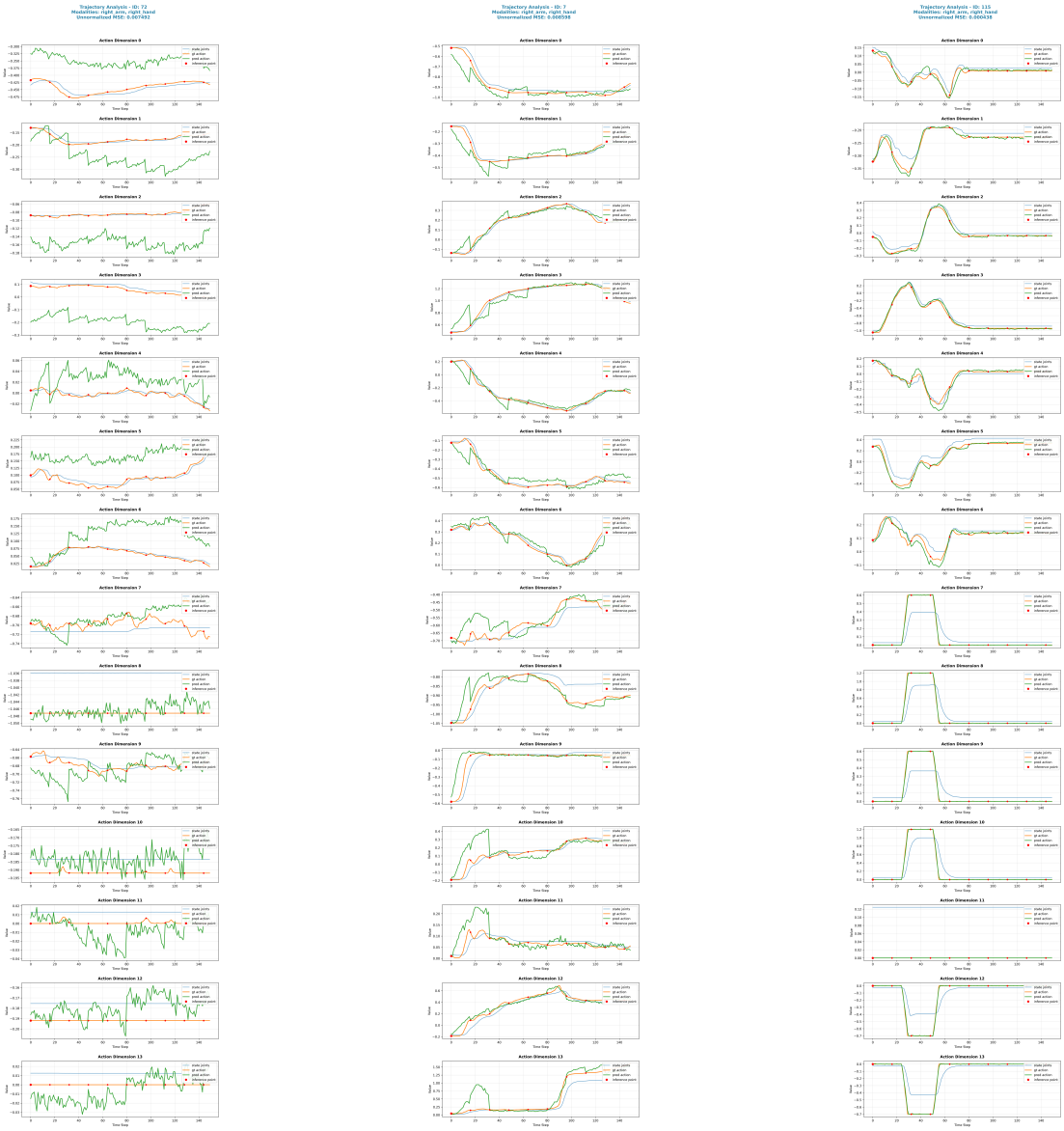
parameter	development computing unit (PC 2)
Model	Jetson Orin NX
CPU	Arm® Cortex®-A78AE
Number of cores	8
Number of threads	8
Max largest rate	2 GHz
graphic memory Memory	16G
Memory	16G
Cache	2MB L2 + 4MB L3
Storage	2T
Intel ® Image Processing Unit	No
GPU	1024 NVIDIA Ampere architecture Gpus with 32 Tensor cores
Maximum dynamic frequency of graphics card	918MHz
Gaussian and Neuro Accelerator	3.0
Intel ® deep learning promotion	Yes
Intel ®Adaptix™ Technology	Yes
Intel ® hyperthreading technology	Yes
Instructions set	64bit
OpenGL	4.6
OpenCL	3.0
DirectX	12.1
IP address	192.168.123.164

Figure B.1: Complete Specifications for Nvidia Jetson Orin NX

C Extended Evaluation Results

C.1 Per-Model Dataset Evaluation

The following figures show detailed evaluation of the three models trained. Each column corresponds to a trajectory, with the plots showing the ground truth, the state and the predicted positions for each time step.



(a) Model A (b) Model B (c) Model C

Figure C.1: Full evaluation plots for all models across dataset sequences.

D Raw Experimental Data

D.1 Level 1 Trial Data

Table D.1: Level 1 Trial Data – Task: Grasp the Red Cube

Trial	Hand	Success (Y/N)	Time (s)	Error Type	Notes
1	Left	Y	8.03	–	Smooth grasp, stable hold
2	Left	N	9.12	Depth error	Grasped too high above cube
3	Left	Y	7.85	–	Proper center alignment
4	Left	Y	8.31	–	Steady grasp and release
5	Left	N	10.24	Reach limit	Cube near max arm extension
6	Left	Y	8.02	–	Stable contact, no slip
7	Left	Y	6.87	–	Smooth closure
8	Left	Y	7.55	–	Consistent grasp angle
9	Left	Y	8.11	–	Minor wrist offset
10	Left	N	–	Vision miss	Target misidentified
11	Left	Y	9.24	–	Slightly delayed closure
12	Left	Y	8.50	–	Corrected approach path
13	Left	Y	8.22	–	Smooth finish
14	Left	N	–	Depth error	Missed cube by 1 cm
15	Left	Y	7.90	–	Stable performance
16	Right	Y	5.90	–	Clean execution
17	Right	Y	4.86	–	Fast completion
18	Right	Y	6.02	–	Stable closure
19	Right	Y	5.78	–	Good alignment
20	Right	Y	5.22	–	Smooth grasp, minimal error
21	Right	Y	6.10	–	No anomalies
22	Right	Y	5.01	–	Quick reaction
23	Right	Y	4.93	–	Consistent accuracy
24	Right	Y	5.32	–	Clean placement
25	Right	Y	6.48	–	Stable grip
26	Right	Y	5.60	–	Correct trajectory
27	Right	Y	4.70	–	Optimal path
28	Right	Y	6.15	–	Smooth grasp
29	Right	Y	5.42	–	Fast response
30	Right	Y	5.10	–	Final confirmation

Conditions: Standard lighting (1800 lux); Table height 75 cm.

D.2 Level 2 Trial Data

Table D.2: Level 2 Trial Data – Task: Pick Up Red Cube and Place on Green Plate

Trial	Hand	Success (Y/N)	Time (s)	Error Type	Notes
1	Right	Y	9.80	–	Accurate placement
2	Right	Y	10.32	–	Slight rotation adjustment
3	Right	N	–	Placement	Cube placed at plate edge
4	Right	Y	11.05	–	Smooth movement
5	Right	Y	8.65	–	Efficient grasp
6	Right	Y	9.30	–	Stable grasp and release
7	Right	N	–	Placement	Knocked plate slightly
8	Right	Y	10.20	–	Consistent execution
9	Right	N	–	Vision	Incorrect depth estimation
10	Right	Y	10.80	–	Slight wrist rotation delay
11	Right	Y	15.65	–	Extended repositioning
12	Right	Y	9.55	–	Smooth control
13	Right	N	–	Placement	Edge placement again
14	Right	Y	6.60	–	Fastest successful trial
15	Right	N	–	Plate contact	Plate moved
16	Left	N	–	Release	Dropped immediately after grasp
17	Left	Y	8.65	–	Corrected grasp
18	Left	N	–	Calibration	Thumb offset
19	Left	N	–	Release	Dropped cube
20	Left	N	–	Release	Object fell mid-transfer
21	Left	N	–	Release	Finger alignment issue
22	Left	Y	8.87	–	Smooth placement
23	Left	N	–	Calibration	Slight offset
24	Left	N	–	Drop	Failed hand closure
25	Left	N	–	Release	Cube fell early
26	Left	N	–	Calibration	Grip angle misaligned
27	Left	N	–	Release	Object fell prematurely
28	Left	N	–	Release	Unstable grasp
29	Left	N	–	Calibration	Incorrect joint offset
30	Left	N	–	Release	Incomplete closure

Conditions: Standard lighting (1800 lux); Table height 75 cm.

D.3 Level 3 Trial Data

Table D.3: Level 3 Trial Data – Pick Up Red Apple and Place on Green Plate (with Distractors)

Trial	Hand	Success (Y/N)	Time (s)	Error Type	Notes
1	Left	Y	10.23	–	Successful grasp and placement
2	Right	N	10.68	Grasping	Fingers misaligned on apple
3	Right	N	10.41	Reach	Grasped too close to body
4	Left	Y	10.21	Drop	Dropped apple near plate
5	Right	N	10.20	Navigation	Failed to move around object
6	Right	N	–	Grasping	Fingers misaligned
7	Right	Y	9.86	–	Clean lift and placement
8	Right	Y	10.32	–	Stable performance
9	Left	N	10.45	Thumb	Thumb not closing properly
10	Right	Y	12.86	–	Extended manipulation time
11	Right	N	8.98	Drop	Dropped apple during lift
12	Right	N	10.26	Thumb	Thumb rotation error
13	Right	Y	10.42	–	Corrected path mid-trial
14	Right	Y	–	–	Successful grasp, no time recorded
15	Left	N	9.86	Hand closure	Incomplete closure
16	Right	Y	10.14	–	Smooth grasp and place
17	Right	N	8.86	Drop	Dropped apple off target
18	Right	N	8.64	Collision	Knocked apple away
19	Right	N	10.30	Grasping	Failed to close on apple
20	Right	N	10.14	Placement	Placed apple at edge of plate
21	Left	N	–	Thumb	Not closing properly
22	Left	N	–	Thumb	Not closing properly
23	Left	N	–	Grasping	Hand not fully closing
24	Left	N	–	Grasping	Incomplete closure
25	Left	N	–	Grasping	Partial closure detected
26	Left	N	–	Grasping	Hand not closing completely
27	Left	N	–	Grasping	Hand not closing completely
28	Left	N	–	Hardware	Thumb stuck mid-motion
29	Left	N	–	Grasping	Incomplete closure
30	Left	N	–	Grasping	Incomplete closure

Conditions: Standard lighting (1800 lux); Table height 75 cm; Distractors: blue/red tape, black/white cubes.

D.4 Height Sensitivity Trial

Table D.4: Height Sensitivity Trial Data – Level 2 Task at 65 cm Table Height

Trial	Hand	Success (Y/N)	Time (s)	Error Type	Notes
1	Left	N	–	Depth	Grasped too high
2	Left	N	–	Depth	Grasped above cube
3	Left	N	–	Vision	Target misaligned
4	Left	N	–	Depth	Missed by several cm
5	Left	Y	11.20	–	Successful correction attempt
6–14	Left	N	–	Depth	Consistent high grasping
15–30	Right	N	–	Depth	All failed due to incorrect depth offset

Task: "Place the red cube on the green plate."

Conditions: Standard lighting (1800 lux); Table height: 65 cm.