TRACEBOT

Report on laboratory process action learning by analysing human demonstrations in VR

Deliverable 5.5

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089), 2021-2025

Deliverable Title	D5.5 Report on laboratory process action learning by analysing human demonstrations in VR
Deliverable Lead:	University of Bremen (UOB)
Related Work Package:	WP5: Traceable Semantic Twin: Planning, reasoning, Audit Trail
Related Task(s):	T5.2: Replication of medical lab environments into Traceable Semantic Twin Knowledge Bases for Reasoning T5.3: Traceability-aware process (re-)planning, reasoning and regulatory digital audit trail T5.4: Learning and reasoning about recorded process memories for accessible task-context information T5.5: Computational Parsing and Abstraction of Concrete Traceable Laboratory Automation Actions from Virtual Demonstrations
Author(s):	Prof. Michael Beetz
Dissemination Level:	Public
Due Submission Date:	28/02/2025
Actual Submission:	24/02/2025
Project Number	101017089
Instrument:	Research and innovation action
Start Date of Project:	01.01.2021
Duration:	51 months

Abstract

This report presents our developments in the context of the TraceBot project about AvagentIEASim, a Physics-Enabled Virtual Simulator (PVS) designed to analyze human-like action execution within a physically realistic environment. By incorporating essential physics constraints, such as gravity, friction, collision dynamics, and force interactions, AvagentIEASim enables robotic systems to acquire task-specific physical knowledge, ensuring accurate and reliable motion execution. However, a key challenge lies in automatically generating structured, action-specific motion instructions that seamlessly drive simulations for further experimentation. While human demonstrations serve as a valuable knowledge source, translating them into machine-readable, physics-aware action sequences remains complex due to the limitations of 2D video data in capturing precise object interactions, grasp mechanics, and sequential task flow. To address this, we propose a multimodal semantic representation framework that extracts structured action descriptions from human demonstrations, dynamically generating motion

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

	instructions that adhere to real-world physical constraints. By automating this translation process, AvagentIEASim enhances simulation-based learning and provides a robust foundation for physics-driven robotic task execution, enabling intelligent agents to develop and refine human-like manipulation strategies in a controlled virtual environment.

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

Versioning and Contribution History

Version	Date	Modified by	Modification reason
v.01	13.02.2025	Prof. Michael Beetz (UOB)	Ready for internal revision
v.01r	17.02.2025	Anthony Remazeilles (TECN)	Internal revision
v.02	19.02.2025	Prof. Michael Beetz (UOB)	Revised version ready for submission

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

Table of Contents

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

Executive Summary

Laboratory automation ensures efficiency and accuracy, yet human execution varies due to differences in grasping techniques, applied forces, and object handling methods. These inconsistencies pose challenges for robotic systems, which require precise and repeatable instructions to replicate human actions accurately. While instructional videos provide useful references, traditional 2D video data lacks depth and motion details, making it difficult to extract grasp mechanics, force application, and motion constraints. As a result, converting human demonstrations into machine-readable action sequences remains a significant challenge in robotic automation.

To address this, we develop AvagentIEASim, a Physics-Enabled Virtual Simulator (PVS) that provides a realistic, physics-driven environment for training autonomous agents, including robots and virtual MetaHumans. By integrating physical constraints such as gravity, friction, and collision dynamics, AvagentIEASim enables AI-driven systems to acquire task-specific physical knowledge, ensuring accurate and reliable motion execution. The system further enhances action sequence structuring through a multimodal semantic representation framework, which extracts structured action descriptions from human demonstrations and refines them into machine-executable instructions. To improve temporal coherence and grasping accuracy, we developed a Refinement model that combines Monte Carlo Tree Search (MCTS) with an LLM, ensuring logically structured and physics-aware action sequences that maintain consistency across different manipulation tasks.

As part of the TraceBot project, this work focuses on parsing and abstracting laboratory automation actions from virtual demonstrations, allowing AI-driven systems to better understand and execute human-like manipulation tasks. While laboratory workflows are standardized at a procedural level, variations in grasping strategies, force application, and task execution speeds make it difficult for robotic systems to generalize across different demonstrations. These inconsistencies create challenges in ensuring repeatability, accuracy, and adaptability, making human-to-machine translation of actions a crucial aspect of laboratory automation.

AvagentIEASim addresses these challenges by generating structured semantic instructions, capturing detailed manipulation parameters that define user-object interactions. This structured approach enables AI systems to analyze, interpret, and execute tasks with greater precision and contextual awareness, reducing ambiguity in robotic task execution. The simulator provides a physics-aware environment, ensuring that object manipulation is both semantically accurate and physically feasible under real-world constraints. To further enhance robotic adaptability, AvagentIEASim includes an adaptive object grasping mechanism, which adjusts grip stability based on surface friction, object

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

shape, and material properties. Additionally, the bidirectional FK-IK solver enables real-time motion refinement, allowing both robotic and virtual agents to dynamically adjust their movements based on task requirements and environmental conditions.

By automating the generation of structured motion instructions, AvagentIEASim enhances functional verification and task planning, supporting simulation-driven learning, where AI agents refine their understanding of manipulation tasks while ensuring compliance with physical constraints. This approach improves robotic adaptability and execution accuracy, making it easier to integrate learned behaviours into AI-driven planning and reasoning frameworks. Ultimately, AvagentIEASim provides a scalable and flexible solution for advancing intelligent automation, enabling robots to learn, adapt, and perform complex tasks with greater efficiency and precision in laboratory environments.

2 Introduction

This research, developed as part of the TraceBot project, focuses on advancing laboratory automation by addressing challenges in translating human demonstrations into structured, machine-executable instructions. To achieve this, we introduce two key developments: (1) the development of AvagentIEASim, a Physics-Enabled Virtual Simulator (PVS) designed to simulate human-like action execution within a physics-aware environment, and (2) a multimodal semantic representation framework that extracts and refines structured action instructions from human demonstrations. AvagentIEASim integrates real-world physics constraints such as gravity, friction, and collision dynamics to enhance robotic task execution, ensuring that AI-driven agents can accurately learn and perform manipulation tasks. The proposed framework leverages a Refinement Model which combines Monte Carlo Tree Search (MCTS) with an LLM to improve the coherence and logical consistency of action sequences, enabling the generation of structured, machine-executable task-instructions optimized for robotic learning and task planning. The following sections first detail the core features of AvagentIEASim, followed by the semantic refinement process for enhancing robotic task execution and functional verification.

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

3 Physics-Enabled Virtual Simulator - AvagentIEASim

AvagentIEASim is designed to advance robotic manipulation and virtual human (MetaHuman) [3] interactions by integrating real-world physics into a controlled simulation environment (see Figure 1). It incorporates virtual robots such as PR2 [1] and Unitree G1 [2], providing a high-fidelity platform where both robots and MetaHumans can learn, adapt, and interact under physical constraints, including force dynamics, friction properties, and object interactions. By leveraging machine learning techniques, such as 3D human mesh generation (see Figure 2) for motion analysis [4], AvagentIEASim translates human demonstrations into structured action sequences, enabling robots to dynamically acquire and refine task-specific skills through reinforcement learning [5].

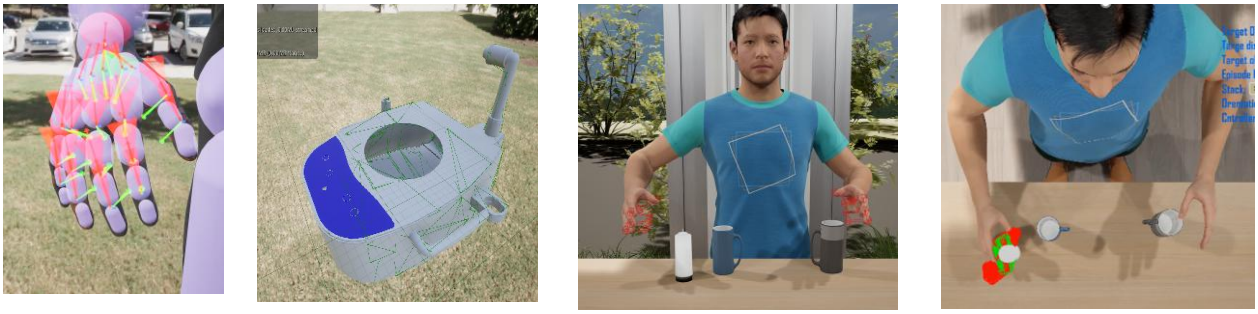

Figure 1: The left two images show the hand and canister-tray collision along with the physics material setup, while the right two images illustrate the collision detection with the canister, where each finger independently detects and responds to collisions.

A key objective of AvagentIEASim is to enhance robotic task planning and execution by offering realistic, physics-based simulations that minimize errors and improve operational efficiency in real-world applications. Through progressive learning mechanisms [6], the simulator enables AI-driven systems to continuously optimize motion strategies for adaptive task execution across diverse environments. Additionally, AvagentIEASim plays a crucial role in bridging the gap between simulated training and real-world robotic deployment by allowing systems to validate and refine motion planning strategies before real-world implementation. The integration of physics-aware control models ensures that robots not only simulate actions accurately but also incorporate real-world constraints such as object mass, material properties, and environmental forces. As a result, AvagentIEASim serves as a versatile research platform that facilitates the development of intelligent, physics-aware robotic systems, enabling more effective interaction, skill acquisition, and real-world task execution.

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

3.1 FK-IK Bidirectional Solver

AvagentIEASim incorporates a real-time kinematic controller that seamlessly integrates Forward Kinematics (FK) and Inverse Kinematics (IK) within the Control Rig, enabling precise motion control for both robots and MetaHumans [7]. Its bidirectional FK-IK solver dynamically switches between automated learning from human demonstrations and real-time user interactions (see Figure 2). In FK mode, motion is driven by real-time or pre-recorded 3D motion tracking, effectively replicating human actions within a physics-aware environment. In IK mode, users can manipulate movements using a joystick interface or a hand-pointed trajectory planner, refining task-specific motions such as grasping and adjusting hand posture. To ensure natural synchronization of body and finger movements, we developed a joint orientation-based mapping system that converts 3D joint

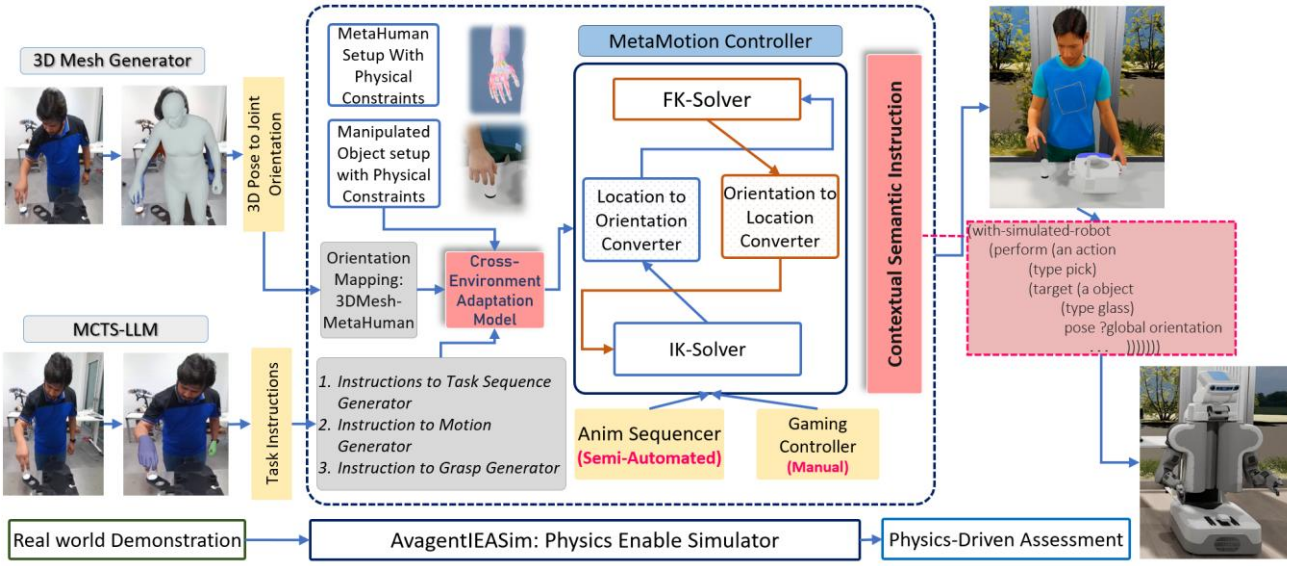

Figure 2: The input-output flow of AvagentIEASim

orientations from SMPLX-based (Skinned Multi-Person Linear model eXpressive) [14] models [4] into precise MetaHuman motions in FK mode. Additionally, we designed an instruction-to-control generation function that processes MCTS-LLM generated instructions to drive the simulations. This process operates in three key (see Figure 2) steps: (1) Task Sequence Generation, synchronizing task and sub-task sequences to execute an action; (2) Motion Generation in IK Mode, producing movement for each step based on task constraints; and (3) Grasp Generation, regulating force dynamics and fine-tuned finger motions for precise object interactions. By integrating these components, AvagentIEASim significantly enhances the simulation of complex grasping behaviors, closely mirroring human hand interactions and improving motion planning in AI-driven agents. This high-

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

fidelity kinematic control system provides a robust and adaptable framework for developing intelligent, physics-aware robotic systems capable of executing intricate manipulation tasks with precision in a physics-enabled simulation environment.

3.2 Adaptive Object Grasping

To ensure physically accurate and dynamic grasping, physics materials have been applied to both MetaHumans and interactive objects within the PVS environment (Figure 1). These materials, fine-tuned using Unreal Engine's physics framework [7], require meticulous parameter optimization based on environmental conditions, object category, and the object's inherent physical properties and dynamics. Through extensive experimentation, we systematically configured key parameters, friction, restitution, mass, and damping to enable adaptive object manipulation, where forces and resistances closely mirror real-world physics. Friction coefficients regulate the sliding resistance between fingers and objects, ensuring a firm but natural grip, while collision presets define object responses upon contact. Additionally, mass properties influence the force needed for lifting and holding, and gravity settings dictate object behavior upon release. These tuneable parameters enhance the learning process, refining adaptive grasping strategies and improving robotic interaction within the simulated environment. By incorporating realistic force interactions, precise finger joint configurations, and soft tissue reflection mechanics, AvagentIEASim ensures that virtual grasping behaviors closely resemble human-object interactions, ultimately improving robotic learning efficiency and enhancing real-world applicability.

To gain deeper insights into action-specific tasks within this physics-enabled environment, an automated, instruction-driven approach is essential for efficient knowledge acquisition, reducing reliance on manual control. By automating the extraction of structured semantic instructions, the system can generate machine-readable action sequences, improving task execution and decision-making. The next section explores how activity video data can be transformed into semantic descriptions, enabling robotic systems to interpret and execute tasks with greater precision and contextual awareness.

4 Activity Video to Semantic Description

Learning from video demonstrations of skilful daily tasks is a valuable knowledge source; however, converting this information into a structured, machine-readable format presents significant challenges. A major limitation arises from the low-dimensional nature of 2D data, which restricts the accurate decoding of object contact timing, action sequence flow, and grasp execution mechanics.

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

Additionally, 2D representations lack higher-level contextual information, making it difficult to capture the complexities of human-object interactions, particularly in understanding temporal activities, grasp types, orientations, and object-activity relationships: all of which are crucial for semantic understanding. Overcoming these limitations requires extended dimensional representations and the extraction of semantic meaning in a structured, sequential manner, enabling machine learning systems to interpret and reason about complex human actions more effectively.

Recent progress on Vision-Language Models (VLMs) [8], [9], [10] have significantly advanced video understanding, human activity recognition, and robotics by generating textual descriptions from visual data. However, they struggle with temporal coherence, motion continuity, and structured action representation, particularly in sequential video analysis. When processing a sequence of N frames, a temporal action state definition is required to maintain a coherent action flow. A key limitation of VLMs is frame-wise text generation, where each frame is processed independently, disregarding its relationship to preceding or succeeding frames. This results in fragmented and inconsistent narratives, making it difficult to reconstruct step-by-step event sequences. Furthermore, VLMs tend to produce generic and repetitive descriptions, often failing to recognize causal dependencies, such as the necessity of picking before holding an object. This deficiency in structured temporal reasoning limits their effectiveness in high-precision applications, such as robotic task planning based on video-based activity analysis. Thus, an enhanced methodology is required to align VLM-generated descriptions with logical action sequences, ensuring temporal consistency and contextual accuracy while reducing inconsistencies.

To overcome these limitations, we propose a context refined model that integrates Monte Carlo Tree Search (MCTS) [11] [12] with an LLM(Phi-3.5-mini-instruct) [10] to enhance the refinement of action sequences generated by VLM (Phi-3.5-Vision-Instruct) [10]. MCTS is a heuristic search algorithm that systematically explores and evaluates possible decision paths to identify the optimal solution. For our approach, we define the four key steps as follows:

Selection: The best node is chosen based on prior evaluations, ensuring logical continuity and coherence in the refinement process.
Expansion: New child nodes with alternative text refinements are generated, incorporating grasp detection, hand-object distance, and prior segment dependencies to maintain contextual consistency.
Simulation: Multiple refinement pathways are explored using LLM-generated candidate descriptions, with scoring functions evaluating coherence, interaction constraints, and semantic accuracy.

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

Backpropagation: Accumulated rewards update node values, refining the search to retain the most plausible, logically structured, and semantically accurate action sequences.

Our approach begins by dividing N frames into I segments $(I<N)$ where each segment represents the temporal action. In the first stage, each segment is processed by the VLM, generating segment-wise descriptions T_i. To ensure coherence, we introduce the MCTS-LLM Refinement Model (see Figure 3).

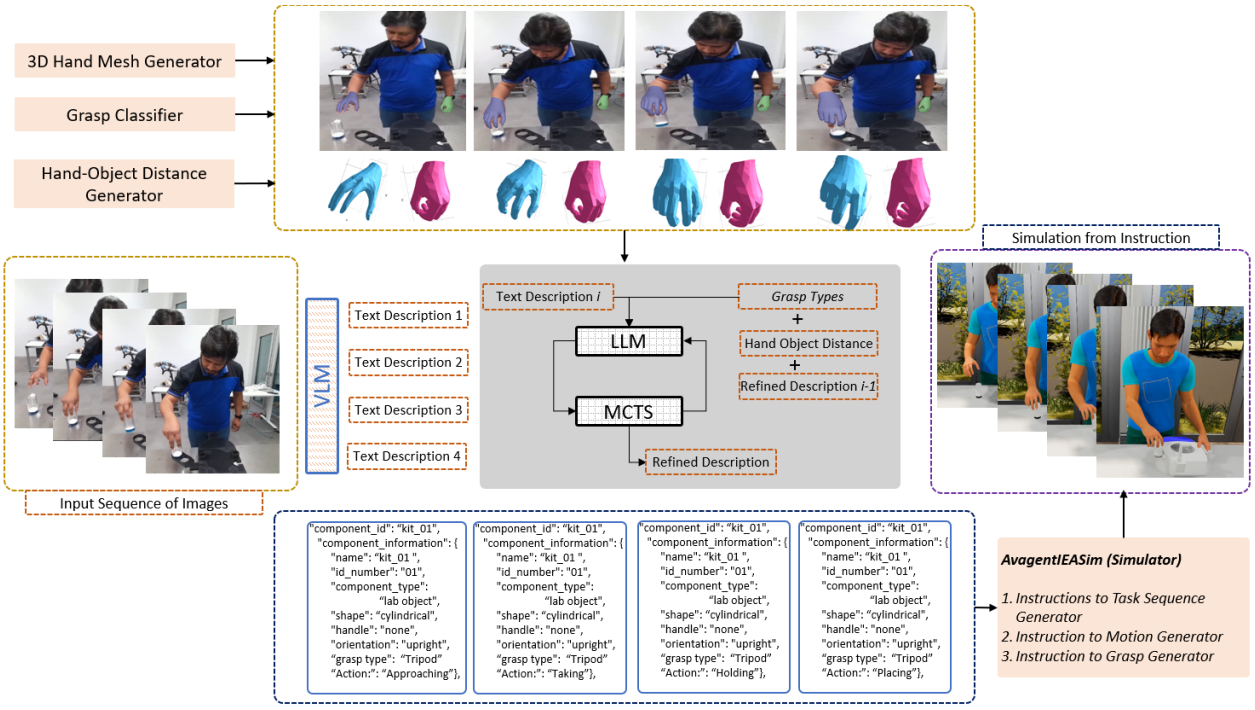

Figure 3. Proposed MCTS-LLM Refinement Model

Before detailing the refinement process, we introduce three key components to enhance the refining accuracy. First, we use a 3D Hand Mesh Model $F_h(X_i)$ to generate a structured hand representation. The model $F_h(X_i)$ helps to understand the grasp type and hand-object distance from its extracted features. Second, we define a linear feedforward regressor network for $d_i = F_{ho}(X_i)$, which calculates the distance (see Table 2) between the hand and the object. The model $F_{ho}(X_i)$ processes features from the *l-th* layer of $F_h(X_i)$ and the targeted object, which is localized using VLM-generated text descriptions. Lastly, we employ a grasp classifier $g_i = F_{gp}(F_{co}^l)$, that determines the grasp type (see Table 2) by analysing the features from specific *l-th* layer of the hand model F_h . The refined text R_i represents the updated action description for S_i after applying MCTS optimization. To ensure the refinement improvements, we use an optimization function $M(T_i, d_i, g_i, R_{<i}, \Phi)$ which integrates

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

MCTS with an LLM Φ, specifically Phi-3.5 [10]. This approach optimizes the sequences of action state by improving the correctness and coherence, following the equation:

$$P(R) = \prod_{i=1}^{I} P(R_i \mid M(T_i, d_i, g_i, R_{<i}, \Phi)) \qquad (1)$$

Where the MCTS-LLM Refinement Model is defined as:

$$M(T_i, d_i, g_i, R_{<i}, \Phi) = arg\ max \sum_{j=1}^{J} P\left(R_i^j \mid T_i, d_i, g_i, R_{<i}, \Phi\right) \qquad (2)$$

Where J is the number of MCTS iteration and R_i^j is the Candidate refinement at iteration j. The MCTS-LLM Refinement Model refines VLM-generated action sequences T_i while ensuring logical coherence and semantic accuracy. Given a sequence of segments S_i, each associated with hand-object distance d_i, grasp type g_i, and text description T_i, the goal is to optimize the refined description R_i by leveraging $M(T_i, d_i, g_i, R_{<i}, \Phi)$, where MCTS iteratively explores the best refinement. The MCTS-based optimization S_i is treated as a node in a search tree, where possible text refinements are evaluated based on temporal consistency, logical order, and interaction constraints. The tree expansion process generates alternative descriptions for each segment, informed by grasp detection, hand-object interaction distance, and prior segment dependencies. A reward function assigns scores to each refinement, considering logical coherence, causality, and consistency with detected hand-object interactions. During simulations, MCTS explores multiple refinement pathways by sampling candidate descriptions T_i suggested by the LLM Φ. The backpropagation step then updates the scores, selecting the most probable refinements to ensure correctness. Incorrect or inconsistent descriptions such as misclassified grasp types, missing transitions (e.g., pick before hold), or redundant actions are pruned, ensuring that only the most plausible action sequences are retained (see red highlights in Table 2). By integrating grasp classification g_i and hand-object proximity d_i into the refinement process, the model effectively filters out false-positive VLM outputs, ensuring that the final structured text accurately represents human-object coherent interactions.

Table 1, Generated Instruction Format


```
{"component_id": "Canister_kit_01",
 "component_information": {
      "name": " canister_kit",
      "id_number": "01",
      "component_type": "lab_object",
      "shape": "cylindrical",
      "size": "medium",
      "handle": "none",
      "orientation": "upright",
      "weight": 1},
   "executed_action": {
      "primary_action": "taking",
      "secondary_action": "none"},
   "grasp_descriptor": {
      "grasp_type": "Tri pod",
      "contact_points": "three_fingers",
      "holding_type": "one_handed",
      "hand_orientation": "top_to_bottom"},
   "environmental_factors": {
      "surface_conditions": "flat_surfaces"}}
```


Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

From each segment, we systematically generate a semantic instruction, embedding detailed action-specific knowledge that encompasses object properties, executed actions, grasp descriptors, and environmental constraints. To achieve this, we leverage an LLM, carefully prompting it with a structured template to ensure precise and contextually relevant instruction generation. This structured format, as illustrated in Table 1, enhances the interpretability and usability of action sequences, making them more effective for robotic task learning, physics-driven virtual simulations, and assistive AI applications in human activity recognition.

To further evaluate the effectiveness of our approach, we present experimental outputs from our model, demonstrating how our proposed MCTS-LLM Refinement Model refines raw VLM-generated descriptions into structured, context-aware instructions. Screenshots of the analysed scenario can be seen at the top of Figure 3. As shown in Table 2, our refinement process significantly improves the coherence, accuracy, and logical sequence of action descriptions, ensuring that they align with real-world physical constraints and task dependencies. This refinement leads to a more reliable and interpretable representation of action sequences, ultimately optimizing robotic motion planning and execution in both simulated and real-world environments.

Table 2: Displays the output of the refined text generated by the MCTS-LLM Refinement Model

	VLM Text Description	Hand Object Distance	Grasp Type	Refined Text from MCTS-LLM
1	Segment 1: The person is standing in an office setting, preparing to perform a canister pick-and-place task into a canister-tray.	2.8	None	T1 Initial State: The person is preparing to perform a canister pick-and-place task.
2	Segment 2: The person is standing in an office setting, preparing to interact with a canister placed on a table in front of them.	1.13	None	T2 Approach: The person is moving towards the canister placed on the table.
3	Segment 3: The person is interacting with a canister placed on a table, possibly preparing for a pick-and-place task involving a canister-tray.	0.76	Tripod	T3 Approach: The person reaches towards the canister, adjusting hand position.
4	Segment 4: The person is holding a canister near a canister-tray on a table.	0.3	Tripod	T4 Holding: The person firmly grasps the canister using a tripod grip.
5	Segment 5: The person is holding a canister and appears to be preparing to place it into a canister-tray.	0.27	Tripod	T5 Pick: The person lifts the canister, preparing for placement.
6	Segment 6: The person is holding a canister and appears to be preparing to place it into a canister-tray.	0.45	Spherical (Misclassified)	T6 Transition: The person moves the canister towards the canister-tray with a tripod grasp.
7	Segment 7: The person is holding a canister and placing it into a canister-tray.	0.25	Tripod	T7 Aligning: The person aligns the canister above the canister-tray, preparing for release
8	Segment 8: The person is holding a canister and appears to be preparing to place it into a	0.33	Spherical (Misclassified)	T8 Aligning: The person adjusts the canister over the canister-tray, ensuring correct alignment.

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

	canister-tray. (Incorrect sequence, already placed in step 7)			
9	Segment 9: The person is holding a canister and appears to be preparing to place it into a canister-tray. (Incorrect sequence, already placed in step 7)	0.29	Tripod	T9 Place Aligning: The person stabilizes the canister, ensuring secure positioning before release.
10	Segment 7: The person is holding a canister and placing it into a canister-tray.	0.7	Tripod	T10 Placing: The person gently places the canister into the canister-tray.
11	Segment 11: The person is holding a canister and placing it into a canister-tray.	1.1	None	T11 Completion: The person releases the canister inside the canister-tray, completing the task.

5 Conclusion

In this work, we presented the development of AvagentIEASim, a Physics-Enabled Virtual Simulator designed to bridge the gap between human demonstrations and robotic task execution by integrating structured semantic action representations with real-world physics constraints within the TraceBot use case. Our approach transforms instructional video data into machine-readable instructions, enabling AI-driven agents to interpret and execute human-like actions with greater accuracy. This functionality also enables future research on automatic plan generation and grounding onto different robot platforms.

To refine the VLM-generated action sequences, we implemented an MCTS-LLM Refinement Model, ensuring logical coherence, grasp classification accuracy, and hand-object interaction fidelity. This improves temporal consistency and task-specific reasoning, making the sequences more effective for robotic learning and simulation. AvagentIEASim integrates a semantic instruction pipeline that extracts task-relevant motion details and translates them into executable actions using the Task Sequence Generator, Motion Generator, and Grasp Generator. These modules ensure adherence to physical constraints like force dynamics, object properties, and environmental interactions. By leveraging these capabilities, AvagentIEASim provides a robust framework for refining AI-driven robotic behaviors, and enhancing motion planning, task execution, and real-world deployment.

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)

6 References

[1] Willow Garage PR2 Overview. Available at: https://www.willowgarage.com/pages/pr2/overview

[2] Unitree Robotics – Agile and Intelligent Robots. Available at: https://www.unitree.com/g1

[3] Unreal Engine MetaHuman Framework. Available at: https://www.unrealengine.com/en-US/metahuman

[4] Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y. "One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer", CVPR, 2023.

[5] Levine, S., Finn, C., Darrell, T., Abbeel, P. "End-to-end training of deep visuomotor policies." Journal of Machine Learning Research, 2016.

[6] Bengio, Y., Louradour, J., Collobert, R., Weston, J. "Curriculum learning." Proceedings of the 26th International Conference on Machine Learning, 2009.

[7] Unreal Engine Physics Material Overview. Available at: https://docs.unrealengine.com/4.27/en-US/InteractiveExperiences/Physics/Materials/

[8] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J. "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond." arXiv preprint arXiv:2308.12966, 2023. Code available at: https://github.com/QwenLM/Qwen-VL

[9] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G. "LLaMA: Open and Efficient Foundation Language Models." arXiv preprint arXiv:2302.13971, 2023.

[10] Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., et al. "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone." arXiv preprint arXiv:2404.14219, 2024.

[11] Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., Colton, S. "A Survey of Monte Carlo Tree Search Methods." IEEE Transactions on Computational Intelligence and AI in Games, vol. 4, no. 1, pp. 1-43, 2012.

[12] Guan, X., Zhang, L. L., Liu, Y., Shang, N., Sun, Y., Zhu, Y., Yang, F., & Yang, M. “rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking”, 2025. arXiv. https://arxiv.org/abs/2501.04519

[13] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10975–10985.

Horizon 2020

TraceBot receives funding from the European Union's H2020-EU.2.1.1. INDUSTRIAL LEADERSHIP programme (grant agreement No 101017089)