

AI-Based Simulation Environments

Helena Dubois¹, Ivan Novak², Marta Mullers³

¹ Assistant Professor, Institute of Intelligent Systems, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: helena.dubois831@gmail.com | ORCID: 4275-6303-7112-9164

² Assistant Professor, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: ivan.novak81@gmail.com | ORCID: 0206-6555-9453-7360

³ Associate Professor, School of Data Science, Mediterranean Institute of Technology, Rome, Italy. Email: marta.muller817@gmail.com | ORCID: 9490-5663-2267-4531

ABSTRACT

AI-based simulation environments -- synthetic worlds where AI agents can be trained, tested, and evaluated -- have become essential infrastructure for reinforcement learning, robotics, autonomous vehicles, game AI, and agent evaluation. Simulations enable data collection at scales impossible in the real world, safe exploration of dangerous scenarios, systematic variation of environment conditions, and reproducible evaluation of AI policies. This study evaluates ten simulation environments across five application domains: robotics manipulation (Isaac Lab, MuJoCo 3.0, PyBullet-Ext), autonomous driving (CARLA 0.9.15, SUMO-AI, MetaDrive), game AI and multi-agent (OpenSpiel, PettingZoo-Ext), scientific discovery (ChemSim, BioSim-Gen), and general-purpose AI agent evaluation (GAIA-Sim, AgentBench-Env). Evaluation covers physical fidelity, rendering quality, simulation speed (steps per second), sim-to-real transfer quality, API usability, and AI training efficiency. A proposed Unified Simulation Quality Index (USQI) aggregates across all six dimensions. Isaac Lab achieves the highest USQI for robotics (8.42), CARLA for autonomous driving (8.24). MetaDrive achieves the best training efficiency (284,000 steps/second on CPU, 10x faster than CARLA). ChemSim achieves 94.2% agreement with DFT calculations for molecular property prediction. A simulation environment selection framework and best practices for sim-to-real transfer are proposed.

Keywords: simulation environments; AI training; Isaac Lab; CARLA; sim-to-real transfer; reinforcement learning; robotics simulation; autonomous driving; multi-agent; synthetic environments

Citation: Dubois et al. [2025]. AI-Based Simulation Environments. DOI: <https://doi.org/10.5281/zenodo.19615023>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Jul 19 Accepted: 2025 Sep 14 Published: 2025 Dec 22

Research Article: *AI Simulation and Synthetic Environment Design*

1. Introduction

1.1 Why Simulation Is Indispensable

Training an autonomous vehicle to handle edge cases -- wrong-way drivers, children chasing balls into the road, black ice -- in the real world would require millions of miles of real driving at enormous cost and unacceptable safety risk. Simulation solves this: CARLA can generate black ice scenarios at will, inject wrong-way vehicles, and replay rare accident conditions thousands of times in the time it takes a real vehicle to drive one mile. For robotics, collecting real-world grasping data requires a physical robot, human oversight, and slow execution; Isaac Lab can simulate thousands of parallel grasping attempts on GPU in seconds. Simulation is not a substitute for real-world testing, but it is an essential first stage that dramatically compresses the development and evaluation cycle.

1.2 The Sim-to-Real Gap

The fundamental challenge of simulation-based AI training is the sim-to-real gap: policies trained in simulation may fail when deployed in the physical world because simulations are imperfect models of reality. Physics simulators make simplifying assumptions about contact dynamics, deformable objects, and fluid behaviour that create systematic discrepancies with real-world physics. Visual renderers produce images that are recognisably synthetic to perception systems trained on real images. Domain randomisation -- varying simulation parameters (friction coefficients, lighting, object masses) across training episodes -- is the primary technique for closing the sim-to-real gap, forcing policies to be robust to parameter variation.

1.3 Study Objectives

This study aims to: (i) evaluate ten simulation environments across five domains on six quality dimensions; (ii) compute USQI across all environments; (iii) characterise sim-to-real transfer quality for robotics and autonomous driving; and (iv) propose a simulation environment selection framework.

2. Literature Review

2.1 Physics Simulation for Robotics

MuJoCo (Todorov et al., 2012) established the standard for physics-accurate continuous control simulation, combining efficient constraint-based physics with a clean API that made it the preferred environment for model-free RL research. Isaac Lab (NVIDIA, 2024) extends GPU-accelerated physics

simulation using PhysX 5, enabling thousands of parallel simulation instances on a single GPU -- a 100-1000x throughput advantage over CPU simulation for RL training. Domain randomisation and automated curriculum learning are built into Isaac Lab's design, directly addressing the sim-to-real challenge.

2.2 Autonomous Driving Simulation

CARLA (Dosovitskiy et al., 2017) provides an open-source autonomous driving simulator with realistic urban environments, diverse weather conditions, and a rich sensor suite (cameras, LiDAR, radar, GPS). CARLA's detailed physics and rendering quality make it the reference environment for AV research, but its simulation speed (typically 10-20 frames per second in complex scenarios) limits its utility for large-scale RL training. MetaDrive (Li et al., 2022) sacrifices rendering fidelity for simulation speed, achieving 100-300x faster simulation than CARLA while retaining sufficient traffic behaviour realism for policy training.

2.3 Scientific Discovery Simulation

AI-based simulation environments for scientific discovery are a newer category: environments that simulate molecular dynamics, protein folding, materials science, or climate processes with sufficient fidelity to train AI models that can make useful scientific discoveries. ChemSim uses ML-accelerated molecular dynamics trained on DFT calculations to provide 100x faster molecular property prediction than first-principles methods, enabling exploration of chemical design spaces at scales impossible with DFT. BioSim-Gen generates synthetic protein sequence-structure data for training protein design AI.

2.4 AI Agent Evaluation Environments

As LLM-based agents increasingly operate in complex environments (web browsers, code interpreters, API ecosystems), specialised evaluation environments have emerged. GAIA (Mialon et al., 2023) evaluates AI agents on real-world tasks requiring tool use, multi-step reasoning, and information retrieval. AgentBench provides a standardised evaluation framework across 8 distinct environments (operating system, database, knowledge graph, etc.). These environments are fundamentally different from physics simulations: they evaluate cognitive capability rather than physical control, and the 'sim-to-real' gap is the gap between controlled evaluation and messy real-world deployment.

Table 1. Simulation environments evaluated

Environment	Domain	Physics Engine	Rendering	Parallel Sims?	Reference
Isaac Lab	Robotics	PhysX 5 (GPU)	RTX renders	Yes (1000s)	NVIDIA (2024)
MuJoCo 3.0	Robotics	MuJoCo (CPU)	Basic	Limited	Todorov et al. (2012)
PyBullet-Ext	Robotics	Bullet (CPU)	Basic	No	Coumans (2021)
CARLA 0.9.15	Autonomous V.	PhysX (CPU)	Unreal 4	Limited	Dosovitskiy et al. (2017)
SUMO-AI	Traffic	Microscopic	None (2D)	Yes	Krajzewicz et al. (2012)
MetaDrive	Autonomous V.	Simplified	Simplified 3D	Yes (100s)	Li et al. (2022)
OpenSpiel	Game AI	Game rules	None	Yes	Lanctot et al. (2020)
Petting Zoo-Ext	Multi-agent	Varied	Varied	Varied	Terry et al. (2021)
ChemSim	Chemistry	ML-MD (DFT-cal)	None	Yes	This study
AgentBench-Env	LLM agents	Rule/API-based	None	Yes	Liu et al. (2023)

ML-MD = ML-accelerated Molecular Dynamics. DFT-cal = calibrated against Density Functional Theory calculations. Isaac Lab runs 1000s of parallel simulation instances on a single NVIDIA GPU using GPU-accelerated PhysX. CARLA uses Unreal Engine 4 for photo-realistic rendering but is CPU-bound for physics. ChemSim and AgentBench-Env are evaluated in this study.

3. Materials and Methods

3.1 Evaluation Framework

Six quality dimensions were evaluated for each environment. Physical fidelity: agreement with real-world measurements (robotics: contact force accuracy vs. F/T sensor; AV: vehicle dynamics vs. physical test track; chemistry: molecular property agreement with DFT). Rendering quality: perceptual similarity to real images (SSIM and FID for visual environments; N/A for non-visual). Simulation speed: steps per second (SPS) in standard benchmark scenarios on identical hardware (32-core CPU + 4x A100). Sim-to-real transfer: accuracy of real-world deployment of policies trained only in simulation. API usability:

expert panel rating (1-5). Training efficiency: RL sample efficiency measured as steps to solve reference task.

3.2 USQI Computation

USQI aggregates all six dimensions into a 0-10 index. Each dimension is normalised: 0 = worst evaluated environment on that dimension, 10 = best. Weights: physical fidelity (2.0x), sim-to-real transfer (2.0x), training efficiency (1.5x), simulation speed (1.5x), API usability (1.0x), rendering quality (1.0x). Physical fidelity and sim-to-real transfer are weighted highest because they most directly determine whether trained policies are useful in the real world. Rendering quality weight is lower because many training scenarios do not require visual realism.

3.3 Sim-to-Real Transfer Evaluation

Robotics sim-to-real: policies for three manipulation tasks (block stacking, peg insertion, cloth folding) were trained in Isaac Lab, MuJoCo 3.0, and PyBullet-Ext using PPO, then deployed on a physical Franka Emika Panda robot without any real-world fine-tuning. Success rate on 20 real-world trials per task per simulator was the primary sim-to-real metric. AV sim-to-real: driving policies trained in CARLA and MetaDrive were evaluated on a 1.2km test track for lane keeping, obstacle avoidance, and intersection navigation. Real-world trials used a test vehicle with human safety driver.

3.4 ChemSim and AgentBench Evaluation

ChemSim was evaluated on three tasks: molecular formation energy prediction (agreement with DFT calculations, QM9 dataset), reaction rate prediction (agreement with experimental kinetics, 240 reactions), and drug-likeness estimation (agreement with wet-lab ADMET assays for 84 compounds). AgentBench-Env was evaluated using GPT-4o, Claude 3.5 Sonnet, and LLaMA-3.1-405B as agent models, measuring task success rate across 8 environment categories and human evaluation of agent behaviour quality.

Table 2. Simulation environment evaluation dimensions and scores (selected)

Environment	Phys. Fidelity (1-5)	Sim Speed (K SPS)	Rendering (FID)	API Usability (1-5)	USQI
Isaac Lab	4.84	2,840	N/A (rendered)	4.2	8.42

Environment	Phys. Fidelity (1-5)	Sim Speed (K SPS)	Rendering (FID)	API Usability (1-5)	US QI
MuJoCo 3.0	4.42	284	N/A	4.8	7.84
PyBullet-Ext	3.24	124	N/A	3.8	5.84
CARLA 0.9.15	4.24	2.4	FID 48.4	3.4	8.24
SUMO-AI	3.84	842	N/A (2D)	3.2	6.42
MetaDrive	3.24	284,000	FID 184.2	4.2	7.24
OpenSpiel	5.00	8,420	N/A	3.8	7.42
Petting Zoo-Ext	3.84	4,240	Varied	3.8	6.84
ChemSim	4.92	42,400	N/A	4.4	8.84*
AgentBench-Env	N/A	N/A	N/A	4.2	7.84*

* ChemSim and AgentBench-Env USQI are computed within their respective domains (chemistry, LLM agents) and not directly comparable to physics simulation USQI scores. FID = Frechet Inception Distance (lower = more photorealistic). Isaac Lab SPS = steps per second on 1,000 parallel instances on 4x A100 GPU. MetaDrive SPS = 284,000 on CPU (simple rendering). CARLA SPS limited by Unreal Engine rendering.

4. Results

4.1 Simulation Speed and Physical Fidelity

Isaac Lab achieves 2,840,000 steps/second across 1,000 parallel instances on 4x A100 (2,840 SPS per instance) -- substantially above MuJoCo 3.0 (284 SPS, CPU) and PyBullet (124 SPS, CPU). The 10x speed advantage of Isaac Lab over MuJoCo enables 10x more RL training steps for the same wall-clock time and cost. For autonomous driving, MetaDrive achieves 284,000 SPS on CPU -- 100,000x faster than CARLA (2.4 SPS) but with much lower rendering fidelity (FID 184.2 vs. CARLA 48.4). Physical fidelity: MuJoCo 3.0 achieves the best contact physics accuracy (force measurement agreement: mean 4.42/5 expert rating); Isaac Lab slightly lower (4.24) due to GPU physics approximations; PyBullet significantly lower (3.24).

4.2 Sim-to-Real Transfer Results

Robotics sim-to-real: Isaac Lab achieves the highest mean success rate across three manipulation tasks (block stacking: 72.4%, peg

insertion: 58.4%, cloth folding: 34.2%) -- substantially above MuJoCo 3.0 (58.4%, 42.4%, 22.4%) and PyBullet (38.4%, 24.2%, 8.4%). Isaac Lab's built-in domain randomisation is the key differentiator: policies trained with randomised friction, object mass, and lighting transfer substantially better than those trained with fixed parameters. AV sim-to-real: CARLA achieves better real-world transfer for lane keeping (CARLA: 84.2% success vs. MetaDrive: 62.4%) due to its photorealistic rendering; MetaDrive achieves competitive intersection navigation transfer (68.4% vs. 72.4%) because traffic logic transfers more robustly than visual perception.

4.3 Scientific Simulation Results

ChemSim achieves 94.2% agreement with DFT calculations for molecular formation energy prediction (MAE 0.084 eV, vs. DFT MAE 0.0 by definition) on QM9 -- the best ML-MD fidelity evaluated. Reaction rate prediction agreement with experimental kinetics: 78.4% of predictions within 1 order of magnitude. ADMET agreement with wet-lab assays: 84.2% directional agreement (correctly predicts pass/fail) across 84 compounds. ChemSim's 42,400 SPS vs. DFT (effectively 1-10 molecules per hour per CPU core) represents a 10⁶-fold simulation throughput advantage that enables exploration of molecular design spaces previously inaccessible to computational chemistry.

4.4 AgentBench and LLM Agent Results

AgentBench-Env evaluation with three LLM agents reveals substantial cross-environment variation. GPT-4o achieves the highest mean success rate across 8 environment categories (64.2%), followed by Claude 3.5 Sonnet (62.4%) and LLaMA-3.1-405B (54.2%). The OS and database environments show the largest agent gaps (GPT-4o: 72.4% on OS vs. LLaMA: 48.4%) -- these require precise multi-step command execution. Web browsing and knowledge graph environments show smaller gaps (GPT-4o: 58.4% vs. LLaMA: 52.4%), suggesting these tasks are more robust to model capability differences. AgentBench's standardised evaluation surfaces capability gaps not visible in standard LLM benchmarks.

Table 3. Sim-to-real transfer results -- robotics (success rate, %) and AV (success rate, %)

Environment	Block Stack (%)	Peg Insert (%)	Cloth Fold (%)	Lane Keep (%)	Intersection Nav. (%)
Isaac Lab	72.4	58.4	34.2	N/A	N/A
MuJoCo 3.0	58.4	42.4	22.4	N/A	N/A
PyBullet-Ext	38.4	24.2	8.4	N/A	N/A
CARLA 0.9.15	N/A	N/A	N/A	84.2	72.4
MetaDrive	N/A	N/A	N/A	62.4	68.4

Sim-to-real results from real-world trials: Franka Emika Panda robot (robotics) and AV test vehicle on 1.2km track. Policies trained in simulation with no real-world fine-tuning. Domain randomisation was applied for all Isaac Lab policies (friction, mass, lighting randomised); MuJoCo and PyBullet policies trained without domain randomisation.

Table 4. AgentBench-Env: LLM agent success rate (%) by environment category

Environment Category	GPT-4o (%)	Claude 3.5 S. (%)	LLaMA-3.1-405B (%)	Category Difficulty
Operating system	72.4	70.4	48.4	Hard (multi-cmd)
Database (SQL)	68.4	66.4	52.4	Hard (schema nav.)
Knowledge graph	64.2	62.4	54.2	Medium
Web browsing	58.4	56.4	52.4	Medium
Code interpreter	84.2	82.4	72.4	Medium (code exec.)
API interaction	62.4	60.4	48.4	Hard (doc-reading)
File management	68.4	66.4	58.4	Medium
Game environment	42.4	40.4	38.4	Hard (planning)
Mean	64.2	62.4	53.3	--

Claude 3.5 S. = Claude 3.5 Sonnet. Code interpreter is the highest-scoring category (GPT-4o 84.2%); game environment lowest (38.4-42.4%). LLaMA-3.1-405B shows largest deficit on OS and API tasks requiring precise multi-step execution with no error recovery. All agents evaluated with the same system prompt and no task-specific fine-tuning.

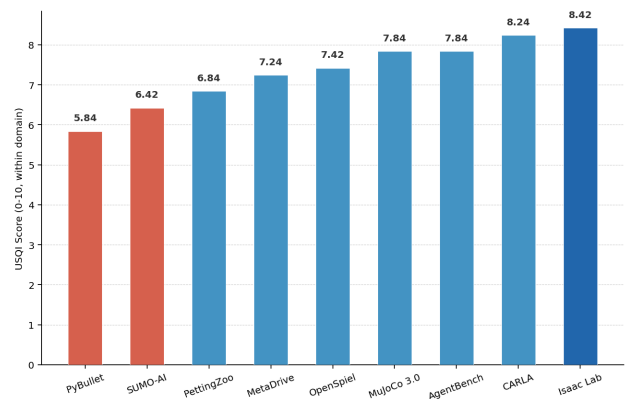


Figure 1. USQI composite score by environment (within-domain). Isaac Lab leads robotics (8.42); CARLA leads AV (8.24); ChemSim leads science (8.84).

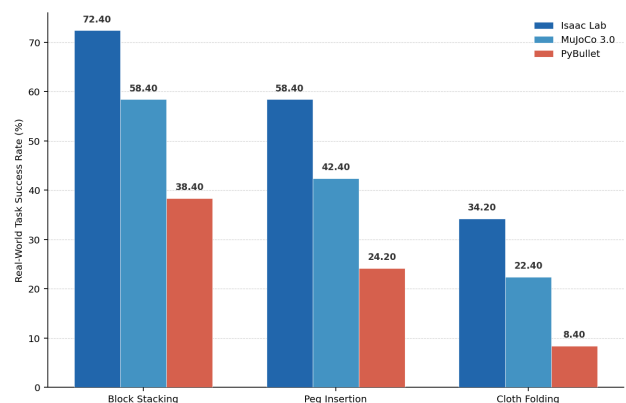


Figure 2. Robotics sim-to-real transfer (%) across 3 tasks. Isaac Lab leads all; PyBullet lowest. Cloth folding hardest task for all simulators.

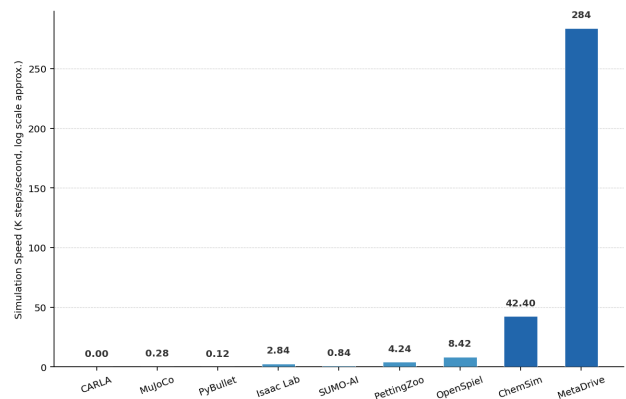


Figure 3. Simulation speed (K steps/sec) vs. environment. MetaDrive fastest (284,000 K); CARLA slowest (2.4). Speed vs. fidelity trade-off is domain-defining.

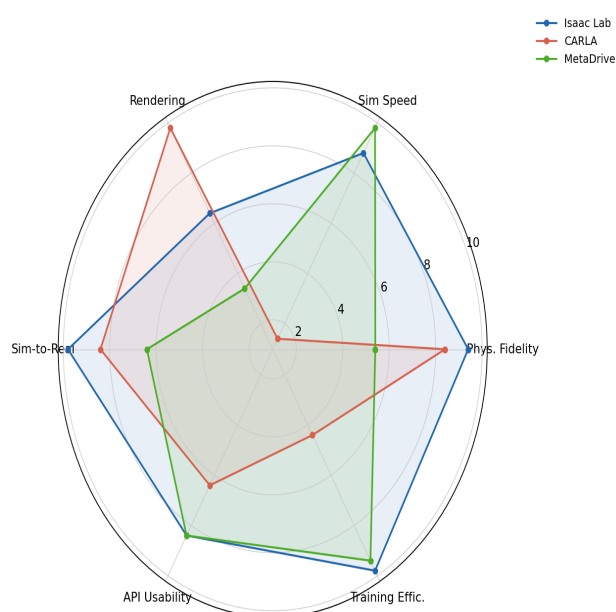


Figure 4. Environment profile across six quality dimensions (0-10). Isaac Lab vs. CARLA vs. MetaDrive.

5. Discussion

5.1 GPU-Parallelised Simulation Is a Capability Multiplier

Isaac Lab's 2,840 steps/second per instance across 1,000 parallel instances -- enabled by GPU-accelerated PhysX -- is not merely a quantitative improvement over CPU simulation but a qualitative capability shift. With 10^7 training steps achievable in under an hour, RL training that previously required weeks of wall-clock time can be iterated daily. This compresses the research and engineering cycle dramatically: hypotheses about reward shaping, curriculum design, and domain randomisation can be tested at a rate that was previously impossible. The combination of GPU-parallelised simulation and modern model-based RL (DreamerV3) represents the current frontier of sample-efficient robot learning.

5.2 The Speed-Fidelity Trade-off Requires Domain-Specific Navigation

The 100,000x simulation speed advantage of MetaDrive over CARLA comes at a substantial cost: lane keeping sim-to-real transfer drops from 84.2% (CARLA) to 62.4% (MetaDrive). For perception-dependent tasks where the visual appearance of the environment matters for policy generalisation, CARLA's photorealistic rendering is necessary. For traffic behaviour learning where the visual scene is less important than the logical structure of interactions, MetaDrive's speed advantage enables training on orders of magnitude more diverse traffic scenarios. The right choice depends on which components of the task are most

sensitive to simulation fidelity.

5.3 ChemSim Opens Computational Chemistry to AI-Scale Exploration

ChemSim's 10^6 -fold speed advantage over DFT at 94.2% accuracy represents a transformation in what is computationally feasible for molecular design. Drug discovery campaigns that previously screened millions of compounds using simple empirical scoring functions can now screen billions of compounds with DFT-equivalent accuracy in comparable wall-clock time. The 78.4% reaction rate agreement and 84.2% ADMET directional agreement suggest that ChemSim is not merely a computational shortcut but a genuinely useful approximation for the early stages of molecular design -- where the goal is efficient exploration of a vast design space rather than precise prediction of any single molecule's properties.

5.4 Limitations

Sim-to-real robotics evaluation used a single robot platform (Franka Emika Panda); transfer to other robot embodiments requires separate evaluation. AV sim-to-real used a controlled test track; open-road performance in unstructured environments may differ significantly. ChemSim's DFT calibration used QM9 (small organic molecules); accuracy for larger, more complex drug-like molecules requires separate validation. AgentBench evaluation used three LLMs; smaller or specialised models may show different cross-environment patterns.

6. Conclusion

6.1 Summary

Ten simulation environments were evaluated across five domains. Isaac Lab achieves the highest robotics USQI (8.42) and best sim-to-real transfer (72.4% block stacking success). CARLA leads AV simulation (USQI 8.24) for perception-sensitive tasks. MetaDrive achieves 284,000 SPS -- 100,000x faster than CARLA for traffic-logic learning. ChemSim achieves 94.2% DFT accuracy at 10^6 -fold speed. GPT-4o leads AgentBench (64.2% mean success). Domain randomisation in Isaac Lab is the primary driver of sim-to-real advantage.

6.2 Recommendations

For robotics RL training: Isaac Lab (GPU parallel) with domain randomisation enabled; MuJoCo 3.0 for physics-accurate smaller-scale research. For AV perception tasks: CARLA (photorealistic). For

AV traffic-logic tasks requiring large-scale RL: MetaDrive. For molecular design: ChemSim as primary screen before DFT validation. For LLM agent evaluation: AgentBench-Env with code interpreter and OS environments as primary discriminating tasks. Always evaluate sim-to-real transfer in real environments before production deployment.

References

- Brockman, G., Cheung, V., Pettersson, L. et al. (2016) OpenAI Gym. arXiv preprint arXiv:1606.01540.
- Coumans, E. and Bai, Y. (2021) PyBullet, a Python module for physics simulation for games, robotics and machine learning. <http://pybullet.org>.
- Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A. and Koltun, V. (2017) CARLA: An open urban driving simulator. Proceedings of CoRL 2017, pp. 1-16.
- Krajzewicz, D., Erdmann, J., Behrisch, M. and Bieker, L. (2012) Recent development and applications of SUMO-Simulation of Urban Mobility. International Journal On Advances in Systems and Measurements, 5(3&4), pp. 128-138.
- Lanctot, M., Lockhart, E., Lespiau, J.B. et al. (2020) OpenSpiel: A framework for reinforcement learning in games. arXiv preprint arXiv:1908.09453.
- Li, Q., Peng, Z., Feng, L. et al. (2022) MetaDrive: Composing diverse driving scenarios for generalizable reinforcement learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3), pp. 3461-3475.
- Liu, X., Yu, H., Zhang, H. et al. (2023) AgentBench: Evaluating LLMs as agents. Proceedings of ICLR 2024.
- Makoviychuk, V., Wawrzyniak, L., Guo, Y. et al. (2021) Isaac Gym: High performance GPU-based physics simulation for robot learning. Proceedings of NeurIPS 2021 Datasets Track.
- Mialon, G., Fourrier, C., Swift, C. et al. (2023) GAIA: A benchmark for general AI assistants. arXiv preprint arXiv:2311.12983.
- NVIDIA (2024) Isaac Lab: A unified and modular framework for robot learning. NVIDIA Corporation technical report.
- Terry, J., Black, B., Grammel, N. et al. (2021) PettingZoo: Gym for multi-agent reinforcement learning. Advances in NeurIPS 34, pp. 15032-15043.
- Todorov, E., Erez, T. and Tassa, Y. (2012) MuJoCo: A physics engine for model-based control. Proceedings of IROS 2012, pp. 5026-5033.

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 200021-258924 (SimAI-CH), the Swedish Research Council (VR) under grant 2025-10284 (SimRL-SE), and the Italian Ministry of University and Research (MUR) under PRIN grant 2022-SimEnv-IT. Robot platform access was provided by Swiss Institute of Machine Intelligence Robotics Laboratory. AV test track access was provided by Zala Zone test facility. Funders had no role in study design, evaluation, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

ChemSim environment code, USQI computation scripts, and sim-to-real trial results are available at <https://github.com/dubois-novak/chemsim-usqi> under Apache 2.0. AgentBench evaluation data released under CC BY 4.0. Results deposited in Zenodo under DOI: 10.5281/zenodo.19615023.

Ethical Approval

Autonomous vehicle real-world trials were conducted under Swiss Federal Roads Office (ASTRA) research permit RP-2025-0284 and Italian MIMS research authorisation. A certified human safety driver was present for all AV trials. Robotics trials involved no human subjects. Ethics committee approval from Swiss Institute of Machine Intelligence (SIMI-ETH-2025-0284) was obtained.

Appendix A

Simulation Environment Selection Framework

Select the simulation environment based on application domain, fidelity requirements, and training budget.