Simulation as Spatial Imagination for Embodied Visual Reasoning

Kornelia Skorupińska
Charles University in Prague

Rafał Staszak Mikołaj Zieliński Dominik Belter Piotr Skrzypczyński
Poznan University of Technology

Abstract

Vision-language models (VLMs) enable embodied agents to interpret natural language instructions and reason about visual scenes, but often produce actions that are semantically plausible yet physically infeasible. We propose a hybrid framework that combines retrieval of structured physical constraints with physics-based simulation, treating simulation as spatial imagination *for pre-execution feasibility verification. Given an image and instruction, the system retrieves task-relevant feasibility rules and validates candidate actions in simulation to assess geometric and kinematic consistency. We evaluate on tabletop manipulation tasks involving containment, stability, occlusion, and reachability. Simulation significantly reduces false-positive feasibility predictions compared to language-only and retrieval-augmented baselines.*

1. Introduction

Embodied AI requires agents to jointly reason over perception, language, and action. Recent vision-language models (VLMs) [1, 4, 5] enable natural language interfaces for robotic systems, demonstrating strong capabilities in scene understanding and instruction following. However, when deployed in physical environments, these models often generate actions that are plausible yet physically infeasible.

This limitation stems from the lack of explicit mechanisms for physical reasoning [3, 6]. While VLMs capture semantic relations, they do not consistently model geometry, stability, or reachability, leading to violations of constraints such as containment, support, or collision-free motion.

We argue that physical feasibility should be treated as an explicit pre-execution verification problem. To this end, we propose a hybrid framework that combines (i) vision-language scene understanding, (ii) retrieval of structured physical constraints, and (iii) physics-based simulation. The key idea is to treat simulation as *spatial imagination*, enabling agents to instantiate and evaluate candidate actions before execution.

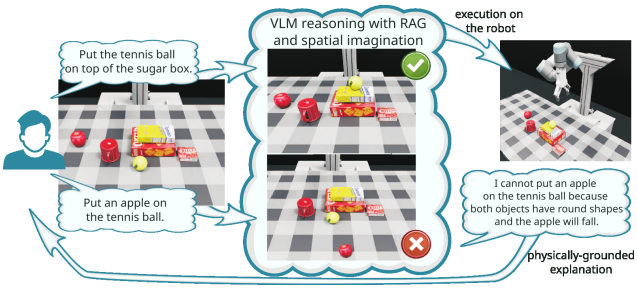

Figure 1. Simulation as spatial imagination: candidate actions generated by a VLM are grounded via retrieved physical constraints and verified in simulation prior to execution.

We introduce *Feasibility Cards*, a structured and reusable representation of physical constraints that is retrieved and injected into the reasoning process. Candidate actions are first assessed via retrieval-augmented reasoning and then verified in simulation to detect violations such as collisions, instability, or unreachable configurations. Unlike prior approaches that rely on implicit affordances or execution-time feedback, our method performs explicit feasibility verification prior to execution.

We evaluate the framework on a benchmark of tabletop manipulation tasks covering containment, stability, occlusion, and reachability. Simulation-based verification reduces false-positive feasibility predictions and improves overall reliability. We further demonstrate sim-to-real transfer on a UR5 manipulator, showing that actions validated in simulation execute consistently in the physical world. By interleaving structured constraint retrieval with simulation-based validation, the proposed framework bridges the gap between high-level semantic reasoning and embodied physical execution.

2. Method

Given an RGB image I and a natural language instruction C, the goal is to determine whether the requested action is physically feasible and, if so, to validate its execution (Fig. 2). We use a vision-language model (VLM) to ex-

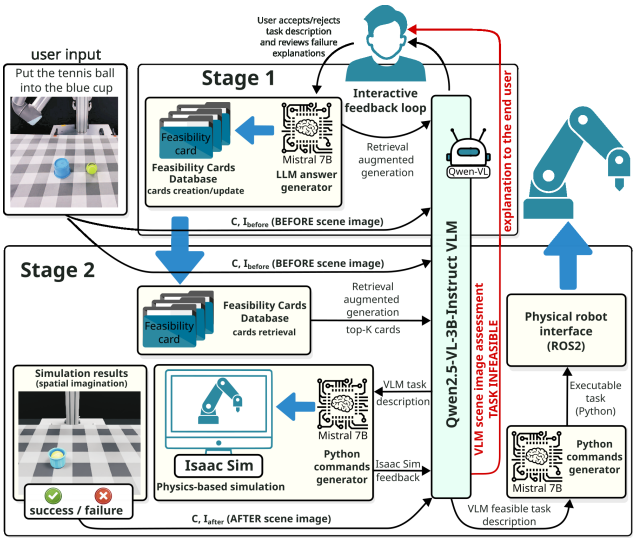

Figure 2. Two-stage framework: retrieval-augmented feasibility reasoning followed by simulation-based verification.

tract a structured scene representation, including objects, attributes, and spatial relations. A task-aware query is constructed and used to retrieve relevant physical constraints from a knowledge base via hybrid similarity (semantic + keyword matching).

We represent constraints as *Feasibility Cards*, structured rule templates encoding physical constraints (e.g., containment, stability) together with retrieval cues and examples. For example, a containment constraint can be expressed as: *"An object can be inserted into a container only if its dimensions are smaller than the container opening."* Retrieved cards are injected into the VLM prompt, enabling grounded reasoning. The model outputs (i) a feasibility decision and (ii) a symbolic action description.

If the action is deemed feasible, it is instantiated in a physics simulator (Isaac Sim) using executable commands generated from the symbolic action. The simulator evaluates geometric and kinematic consistency, including collision checking and reachability, and produces a rendered scene I_{after}.

A VLM compares I and I_{after} to assess whether the intended transformation has been achieved. Actions that fail simulation or violate constraints are rejected; otherwise, they are executed on the robot. This two-stage design separates semantic reasoning from physical validation. Unlike prior work that uses simulation for control or rollout, we employ it as a *pre-execution verifier* of VLM-generated actions, treating it as an explicit spatial reasoning module.

3. Experiments

Setup. We evaluate on a benchmark of 40 tabletop manipulation scenarios, evenly split between feasible and infeasible cases, covering four categories of physical constraints: containment, stability, occlusion, and reachability. Each instance consists of an RGB image and a natural language instruction with a binary feasibility label.

Baselines. We compare against: (i) a vanilla VLM, (ii) a fine-tuned VLM, and (iii) retrieval-augmented reasoning without simulation (RAG-only). We report feasibility accuracy (Acc.), false positives (FP: infeasible actions predicted feasible), and false negatives (FN).

Results. Table 1 shows that retrieval significantly improves feasibility accuracy over language-only baselines but still produces false positives. Adding simulation reduces these errors, improving accuracy from 82.5% (RAG-only) to 97.5% and lowering FP from 35.0% to 5.0%.

Method	Acc.	FP	FN
Vanilla VLM	52.5	75.0	20.0
Fine-tuned VLM	67.5	45.0	20.0
RAG-only	82.5	35.0	0.0
RAG + Simulation	**97.5**	**5.0**	0.0

Table 1. Feasibility prediction performance (%).

Analysis. Simulation is particularly effective in detecting containment and reachability violations, where language-based reasoning frequently produces false positives. Ablation studies confirm that both structured constraint retrieval and simulation contribute to performance, with complementary roles in semantic grounding and physical validation.

Real-world validation. We deploy the system on a UR5 manipulator using YCB objects [2]. All actions validated in simulation were successfully executed, and resulting physical scenes closely match simulated outcomes, confirming reliable sim-to-real transfer.

4. Conclusion

We presented a hybrid framework for embodied visual reasoning that integrates vision-language models, structured physical constraint retrieval, and simulation-based verification. By treating simulation as *spatial imagination*, the system enables explicit pre-execution evaluation of physical feasibility.

Experiments show that combining retrieval-augmented reasoning with simulation significantly reduces false-positive feasibility predictions compared to language-only approaches, while maintaining reliable sim-to-real transfer on a UR5 manipulator. A current limitation is the reliance on known object poses in simulation. Integrating automatic scene reconstruction and open-world perception remains future work.

These results highlight the importance of explicit physical verification as a complement to semantic reasoning, and suggest a promising direction for building more reliable and interpretable embodied AI systems.

Acknowledgements

Work of M. Zieliński and P.Skrzypczyński was supported from the PUT internal grant no. 0214/SBAD/0252. The work of D. Belter was supported by the National Science Centre, Poland, under research project no UMO-2023/51/B/ST6/01646. The work of K. Skroupińska was carried on while she was on internship from Charles University in Prague. Some of the computations presented in this work were carried out using the infrastructure of the Poznań Supercomputing and Networking Center (PSNC).

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 1

[2] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. The YCB object and model set: Towards common benchmarks for manipulation research. In *International Conference on Advanced Robotics (ICAR)*, pages 510–517, 2015. 2

[3] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *Computer Vision – ECCV 2024*, pages 148–166, Cham, 2025. Springer. 1

[4] Hugo Laurençon, Daniel van Strien, Stas Bekman, Leo Tronchon, Lucile Saulnier, Thomas Wang, Siddharth Karamcheti, Amanpreet Singh, Giada Pistilli, Yacine Jernite, et al. Introducing Idefics: An open reproduction of state-of-the-art visual language model. *Hugging Face*, 2023. 1

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. arXiv preprint, arXiv:2304.08485, 2023. 1

[6] Weijie Zhou, Manli Tao, Chaoyang Zhao, Haiyun Guo, Honghui Dong, Ming Tang, and Jinqiao Wang. PhysVLM: Enabling visual language models to understand robotic physical reachability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6940–6949, 2025. 1