PhyGS: Physically-Grounded Controllable Scene Generation

Aparajito Saha[1] Zhen Hao Gan[1] Jinjia Guo[1] Jacob Skwirsk[1] Jeremy Acheampong[1]
Anton Arapin[1] Chahyon Ku[1] Yue Hu[1] Nima Fazeli[1]
Bernadette Bucher[1]

[1]University of Michigan, Ann Arbor

Figure 1. **PhyGS** is a simulation framework that enables the controllable generation of physically-grounded and visually diverse building-scale 3D scenes for simulating mobile manipulation tasks. The above image showcases several scenes generated by PhyGS and rendered in IsaacSim; we additionally provide a hardware abstraction layer for the Boston Dynamics Spot with Arm quadruped robot for validating low-level physical parameters and robot control in simulated scenes.

Abstract

The pursuit of generalist robot policies requires evaluation spanning variations of visual environments and physical interactions intractable for a robot to execute in the real world. We introduce PhyGS, a simulation framework designed for the controllable generation of photorealistic, building-scale indoor environments with high-fidelity low-level physics that bridges the gap between procedural scene synthesis and robotic control. By extending Infinigen-Indoors within the IsaacSim ecosystem, PhysGS transforms static visual backdrops into fully interactable environments through automated object articulation, the assignment of rigid-body physical properties, and the integration of ray-traced lighting. Unlike recent agent-based generative tools that rely on stochastic language models, PhyGS employs a rule-based approach to ensure the controllability required for rigorous benchmarking. Finally, we provide a hardware abstraction layer via a unified API and USD model for the Boston Dynamics Spot robot, demonstrating a seamless transition between simulated evaluation and physical deployment to help identify critical performance gaps in state-of-the-art generalist robotic agents.

1. Introduction

Advancements in policy learning together with large scale models for natural language processing have sparked many large-scale efforts toward building algorithms for generalist robots. Generalist robots receive arbitrary natural language commands and execute the described tasks on a robot platform, most commonly a mobile manipulator, though these algorithms are intended to be hardware agnostic. An important question currently facing the robot learning community is: "How well do these generalist algorithms work and where are the performance gaps?" Answering this question is nontrivial because it is intractable to evaluate the vast scope of visually grounded, natural-language-specified tasks across the variation of physical control realizations in the physical world.

Generation of large scale simulation environments has promise in addressing this question. Procedural methods provide controlled generation of large scale targeted testing environments, and simulation enables parallelization of large scale evaluation. For visual and spatial variation, procedural scene generation frameworks automate creation of large scale photorealistic environments [3, 9, 15, 16]. For variation in physical control, physics-driven simulation platforms facilitate the training of robust motor policies and the benchmarking of reinforcement learning algorithms across diverse robotic morphologies [11, 12, 17, 18]. However, current state-of-the-art methods for robot autonomy vary drastically across these axes of low-level physical control and high-level semantic reasoning in addition to significant variation between navigation of large spatial environments and manipulation on a tabletop.

This work provides a simulation framework for controllable generation of photorealistic indoor scenes with high fidelity, low level physics simulation. Recent simulation tools have worked on combining agent-based photorealistic generation with low level physics simulators [14, 21]. However, these approaches do not enable the controllability of procedural scene generation techniques due to their dependence on stochastic language models to interpret the descriptions of scenes rather than directly enabling rule-based generation. While both our approach and agent-based alternatives can generate diverse training data, the controllability of our simulator is critical for developing systematic evaluation benchmarks in simulation.

Contributions. In this work, we extend well-adopted simulation tools to create a generative framework for robotics training and evaluation with high fidelity simulated control in building-scale photorealistic environments. This simulation tool directly extends the integration of Infinigen-Indoors [16], a popular procedural scene generation framework, in IsaacSim [12]. While robot models were previously controllable within imported Infinigen scenes, we make the scenes themselves physically interactable by modifying the scene generation process to create the scene with articulation for appropriate objects. We then augment the import process to IsaacSim to add rigid body physical properties (such as mass and friction), include meshes for computing collisions, replace Blender [1] lighting with IsaacSim lighting to retain photorealism, and enable physics within the scene. To support systematic scene variation, we create a separate process to automate population of additional objects within the scene from Objaverse-XL [4]. Finally, we develop a hardware abstraction layer to demonstrate interaction between our simulator and the real world, through a designed USD model and API for a Boston Dynamics Spot robot with an arm matching the hardware and official SDK on the physical robot.

2. Related Work

Indoor Scene Generation Indoor scene synthesis for simulation largely falls into three paradigms: rule-based, generative-model-based, and agentic. Procedural methods, such as ProcTHOR [3], Infinigen-Indoors [16], and Scene Synthesizer [5], employ hand-crafted rules or constraint-based sampling to generate indoor scenes. Because rules dictate the placement and appearance of assets, the environments are inherently stable and allow for fine-grain control over size and material selection. A key limitation in these engines is their inability to generate open-vocabulary scenes; extending these methods for novel environments requires redefining the constraint rules for the underlying solvers, which can be time consuming. Generative approaches such as ATISS [13] and PhyScene [23] leverage human-generated 3D data [2, 6] to learn spatial priors, resulting in realistic floor plans; however, they are often restricted in object-level generation and placement and perform poorly on large multi-room generation tasks. More recently, Large Language Model (LLM) driven agents and diffusion models have been utilized to capture complex scene distributions and provide open-vocabulary natural language controllability. Methods like Holodeck [24], ARCHITECT [19], SceneWeaver [25], and SAGE [21] employ agentic frameworks or image inpainting to iteratively design and populate indoor scenes. While these scenes offer diverse visual and physical qualities, the inherent stochasticity of the underlying agents prevents fine-grained control over scene characteristics and prevents such methods for being used for evaluation benchmarks which emphasize standardization of the environment.

Articulated Objects in Simulation Creating interactable simulation environments requires high-quality articulated assets. Existing datasets like PartNet-Mobility [22] and GAPartNet [7] have been widely used, but they often suffer from limited visual realism or insufficient physical fidelity of dynamic joints. To address this, ArtVIP [8] introduces a suite of professionally crafted articulated objects with fine-tuned dynamic parameters to match real-world physics; however, this approach is time consuming when scaling up to include diverse appearances and geometries. Other approaches focus on scalable procedural generation; for instance, Infinigen-Articulated [9] enables the creation of diverse, procedurally generated articulated assets with varying parameters, although they cannot retain articulated properties in physical simulations. Alternatively, methods like DRAWER [20] reconstruct interactable digital twins by extracting geometric and articulation properties directly

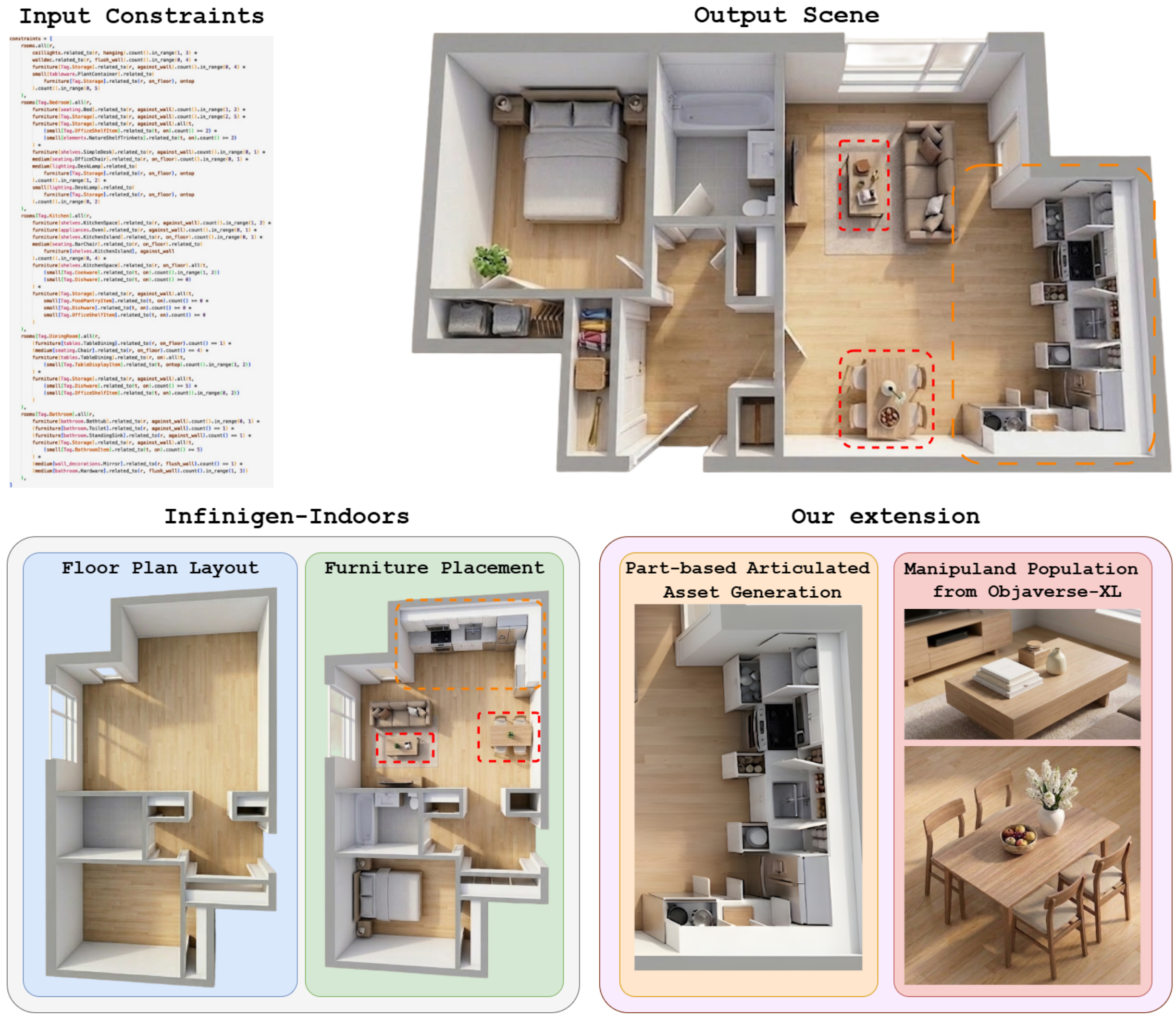

Figure 2. An end-to-end overview of the PhyGS generation process in Blender. Given a user-defined set of constraints for an indoor scene, PhyGS leverages Infinigen-Indoors for designing a valid floor plan and placing large furniture. We then extend the Infinigen engine to decompose assets that require articulation and add necessary joints, before populating the entire scene with labeled assets from Objaverse-XL based on semantic relevance and free volume.

from videos of static real-world scenes, which can enable low-level robot policy training in simulation but cannot be easily extended to larger scenes with visual diversity.

3. Simulation

Embodied agents for mobile manipulation require exposure to diverse scenarios that are often intractable to collect via real-world experiments. While synthetic data offers a scalable alternative, it frequently suffers from sim-to-real gaps in perception or physics. PhyGS addresses this by modifying photorealistic indoor scene generation to enable interaction within a high fidelity physics simulator.

To train and evaluate generalist robot algorithms, a simulator must satisfy four core requirements: 1) extensibility for diverse environment generation, 2) photorealistic rendering, 3) physical interactability, and 4) continuous robotic control. To meet these requirements, we introduce PhyGS to bridge photorealistic procedural scene generation in Infinigen [15] with robotic control in IsaacSim [12] and IsaacLab [11]. PhyGS automates the population of objects from Objaverse-XL [4] within these generated scenes.

3.1. Procedural Scene Generation

We utilize *Infinigen-Indoors* [16] as our procedural engine due to its use of mathematical rules to create infinite variations of 3D scenes within the OpenUSD framework. We specifically employ *Infinigen* to construct underlying archi-

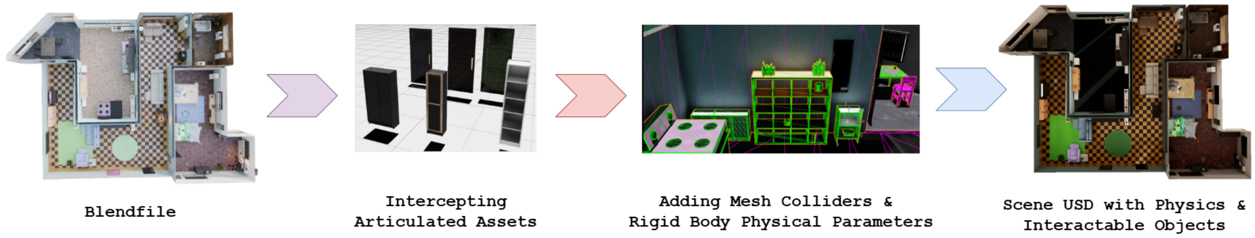

Figure 3. An overview of the scene export process for PhyGS. Infinigen-Indoors leverages Blender for procedural generation, but the default export process to 3D simulation environments sacrifices articulations, collision meshes and lighting. We design a custom exporter that intercepts articulated assets, adding joint parameters, populates all assets with mesh colliders and rigid body physical properties such as mass and friction, and corrects the lighting scheme to retain textural and visual fidelity.

tectural elements and large furniture—components that are difficult to retrieve from standard datasets—while bypassing smaller decorative objects to maintain visual realism. Scenes are generated via configuration files for asset details and constraint files for spatial relations, ensuring that object arrangements respect physical viability (e.g., ensuring free space for accessibility and preventing mesh overlaps).

3.2. Automated Articulation

Interactive search tasks require agents to reason about physical obstructions, such as opening a cabinet to confirm the presence of a target object. While existing simulators like BEHAVIOR-1K [10] provide articulation, they often rely on hand-generated assets that limit scale and introduce procedural bias. In PhyGS, we intercept assets that require articulation and add necessary joints during scene generation in Blender.

A significant technical challenge arises during the export from Blender to USD: the standard baking process collapses articulated joints into static meshes. To overcome this, we developed a post-processing interceptor for the *Infinigen* exporter. Rather than manually redefining constraints for every scene, our system intercepts the parameters of articulated assets (doors, cabinets, and drawers) during the export phase as visualized in Figure 3. We preserve their joint geometry, material properties, and spatial relations, reinjecting this information into the final USD scene. This allows PhyGS to support robotic tasks requiring interaction with articulated objects without manual annotation or compromising the procedural pipeline.

3.3. Semantic Object Population

For a robot to develop hierarchical reasoning, it must encounter objects placed in semantically and spatially meaningful locations. While *Infinigen* provides a base layout, its internal object library lacks the diversity required for robust generalist training. We populate the procedural generated base scenes with assets from Objaverse-XL [4] via an automated process, selecting objects based on relevance to indoor environments and photorealistic fidelity as visualized in Figure 2.

To ensure these objects are interactable for mobile manipulation, we assign them rigid-body physical properties and collision meshes within IsaacSim as visualized in Figure 3. While mass and friction can be randomized, we establish a baseline by fixing the static friction coefficient at 0.8 and setting the mass of interactable target objects to 1.0 kg, well within the 7.0 kg payload capacity of the Spot Arm. Unlike the standard footprint-based placement in *Infinigen*, our automated population method accounts for volumetric availability, allowing for more complex arrangements and diverse grasp-pose selection scenarios.

3.4. Hardware Abstraction and Spot USD

The final component of PhyGS is a high-fidelity model of the Boston Dynamics Spot equipped with a manipulation arm as a unified USD. To facilitate hardware-agnostic algorithm development and sim-to-real transfer, we developed an API that mirrors the official Boston Dynamics SDK. This abstraction of physical control enables policies developed in IsaacLab to be deployed on hardware with minimal modification as visualized in Figure 1.

4. Discussion

The development of PhyGS represents a significant step toward closing the gap between large-scale procedural scene generation and high-fidelity physical interaction for robotic learning. By integrating the rule-based controllability of Infinigen with the robust physics engine of IsaacSim, we provide a framework that allows for the systematic evaluation of generalist robots in environments that are both photorealistic and physically reactive. While our use of ray-traced lighting within IsaacSim achieves high visual fidelity, the computational overhead of real-time rendering remains a constraint for massive parallelization of RL policies on consumer-grade hardware. We will continue to expand this effort while using our simulator to train and evaluate new methods for robotic cognitive and physical intelligence.

References

[1] Blender Online Community. Blender - a 3d modelling and rendering package. 2

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *International Conference on 3D Vision (3DV)*, 2017. 2

[3] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[4] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. In *arXiv preprint arXiv:2307.05663*, 2023. 2, 3, 4

[5] Clemens Eppner, Adithyavairavan Murali, Caelan Garrett, Rowland O'Flaherty, Tucker Hermans, Wei Yang, and Dieter Fox. scene_synthesizer: A python library for procedural scene generation in robot manipulation. In *Journal of Open Source Software*, 2024. 2

[6] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 2

[7] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[8] Zhao Jin, Zhengping Che, Tao Li, Zhen Zhao, Kun Wu, Yuheng Zhang, et al. Artvip: Articulated digital assets of visual realism, modular interaction, and physical fidelity for robot learning. In *International Conference on Learning Representations (ICLR)*, 2026. 2

[9] Abhishek Joshi, Beining Han, Jack Nugent, Max Gonzalez Saez-Diez, Yiming Zuo, Jonathan Liu, Hongyu Wen, Stamatis Alexandropoulos, Karhan Kayan, Anna Calveri, Tao Sun, Gaowen Liu, Yi Shao, Alexander Raistrick, and Jia Deng. Procedural generation of articulated simulation-ready assets. In *arXiv preprint arXiv:2505.10755*, 2025. 2

[10] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning (CoRL)*, 2023. 4

[11] Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Munoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. In *arXiv preprint arXiv:2511.04831*, 2025. 2, 3

[12] NVIDIA. Isaac Sim. 2, 3

[13] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[14] Nicholas Pfaff, Thomas Cohn, Sergey Zakharov, Rick Cory, and Russ Tedrake. Scenesmith: Agentic generation of simulation-ready indoor scenes. In *arXiv preprint arXiv:2602.09153*, 2026. 2

[15] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[16] Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, et al. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[17] Russ Tedrake and the Drake Development Team. Drake: Model-based design and verification for robotics, 2019. 2

[18] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. 2

[19] Yian Wang, Xiaowen Qiu, Jiageng Liu, Zhehuan Chen, Jiting Cai, Yufei Wang, Tsun-Hsuan Wang, Zhou Xian, and Chuang Gan. Architect: Generating vivid and interactive 3d scenes with hierarchical 2d inpainting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[20] Hongchi Xia, Entong Su, Marius Memmel, Arhan Jain, Raymond Yu, Numfor Mbiziwo-Tiapo, Ali Farhadi, Abhishek Gupta, Shenlong Wang, and Wei-Chiu Ma. Drawer: Digital reconstruction and articulation with environment realism. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[21] Hongchi Xia, Xuan Li, Zhaoshuo Li, Qianli Ma, Jiashu Xu, Ming-Yu Liu, Yin Cui, Tsung-Yi Lin, Wei-Chiu Ma, Shenlong Wang, Shuran Song, and Fangyin Wei. Sage: Scalable agentic 3d scene generation for embodied ai. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 2

[22] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. Sapien: A simulated part-based interactive environment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[23] Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. Physcene: Physically interactable 3d scene synthesis for embodied ai. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[24] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, et al. Holodeck: Language guided generation of 3d embodied ai environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[25] Yandan Yang, Baoxiong Jia, Shujie Zhang, and Siyuan Huang. Sceneweaver: All-in-one 3d scene synthesis with an extensible and self-reflective agent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 2