Demonstrating DVS: Dynamic Virtual-Real Simulation Platform for Mobile Robotic Tasks

Zijie Zheng, Zeshun Li*, Yunpeng Wang*, Qinghongbing Xie[2], Long Zeng†

***Abstract*—With the development of Embodied AI, robotic research has increasingly focused on complex tasks. Existing simulation platforms, however, are often limited to idealized environments, simple task scenarios and lack data interoperability. This restricts task decomposition and multi-task learning. Additionally, current Simulation Platforms face challenges in dynamic pedestrian modeling, scene editability, and synchronization between virtual and real assets. These limitations hinder real-world robot deployment and feedback. To address these challenges, we propose DVS (Dynamic Virtual-Real Simulation Platform), a platform for dynamic virtual-real synchronization in mobile robotic tasks. DVS integrates a random pedestrian behavior modeling plugin and large-scale, customizable indoor scenes for generating annotated training datasets. It features a optical motion capture system, synchronizing object poses and coordinates between virtual and real worlds to support dynamic task benchmarking. Experimental validation shows that DVS supports tasks such as pedestrian trajectory prediction, robot path planning, and robotic arm grasping, with potential for both simulation and real-world deployment. In this way, DVS represents more than just a versatile robotic platform; it paves the way for research in human intervention in robot execution tasks and real-time feedback algorithms in virtual-real fusion environments.**

I. Intended Demonstration

During the conference, attendees will have the opportunity to experience the DVS platform firsthand through a live, interactive demonstration. The session will begin with a comprehensive walkthrough of the platform's key features, including pedestrian behavior simulation, large-scale customizable environments, and real-time optical motion capture. Attendees can interact with the system using multi-device support (PC and VR), utilizing tools such as a mouse, keyboard, or VR controllers to explore the platform's capabilities.

Participants will be able to modify dynamic pedestrian parameters, build custom environments, and collect experimental data in various formats. Additionally, the demonstration will showcase the real-time synchronization between physical robots and their virtual counterparts, highlighting the platform's ability to bridge virtual and physical environments seamlessly. This hands-on experience will provide a deep understanding of how the platform supports dynamic human-robot interaction and navigation tasks.

II. Introduction

Robots are becoming increasingly capable with advances in perception, decision-making, and execution technologies. These improvements have expanded their potential applications in industrial manufacturing[20][2], smart homes[21][6], and other fields[41][4][9]. The transition from rule-based operations to end-to-end learning has enabled robots to tackle more complex tasks. However, achieving high efficiency in real-world scenarios requires a complete workflow: virtual data collection, training, and real-world deployment. Existing simulation platforms often fail to effectively support this closed-loop research due to their functional limitations.

Data collection in robotics typically relies on two approaches: collecting real-world data with physical robots or using virtual agents in simulated environments. Systems like Mobile Aloha aim to reduce the cost of real-world data collection. However, they still require significant hardware investment and expert labor. Simulators provide task-specific modeling tools for focused research. In contrast, simulation platforms offer a broader framework for multi-task and complex scenario investigations, enabling faster iteration. Platforms such as Habitat[31][34][29], iGibson[16][32][39], and Arena[12][11] have facilitated data collection and algorithm training. Yet, their scope remains narrow. Habitat, while extended to support HITL (Human-in-the-Loop)[38][24] and HRC (Human-Robot Collaboration)[1][22], focuses mainly on navigation tasks. iGibson enhances data richness and realism through interactive environments but lacks support for dynamic scenarios. Arena specializes in navigation, while tasks like grasping rely on external simulators such as PyBullet[5] or MuJoCo[35].

Existing simulation platforms are often task-specific or designed for static environments. They struggle with complex long-horizon tasks that require environmental understanding and cross-domain collaboration. For instance, completing a task like *retrieving a bottle from the fridge and placing it on a desk* involves navigation, manipulation, and environment interaction. Current methods decompose such tasks into subtasks across multiple simulators, increasing workload and reducing coherence. Furthermore, most platforms lack models for dynamic scenarios, such as pedestrian behaviors. They also fail to integrate real-world feedback, exacerbating the sim-to-real gap and resulting in significant performance drops during deployment.

We propose DVS (Dynamic Virtual-Real Simulation Platform), a novel framework designed for multi-task, dynamic, and closed-loop robotic research. DVS addresses these limitations through three key features. First, it supports complex long-horizon tasks with dynamic pedestrian modeling and flexible indoor scene editing. This enables high-fidelity simulation environments for multi-stage operations. Second, it establishes a virtual-real fusion workflow, combining high-

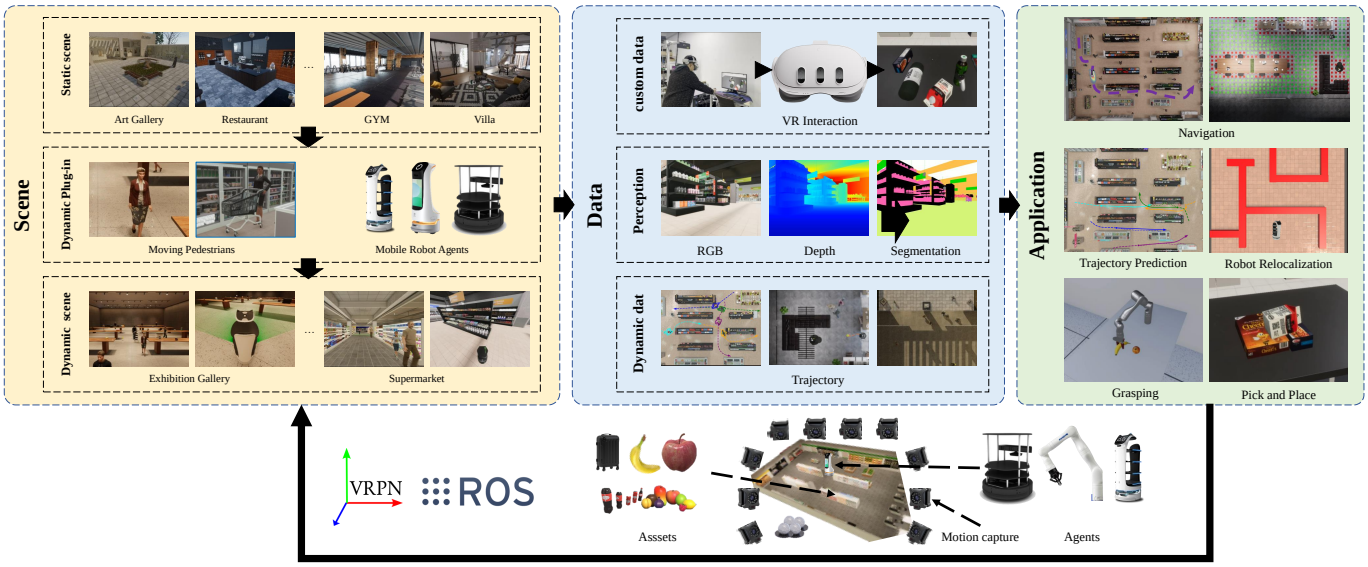

Fig. 1. Overview of DVS platform, which offers a variety of large-scale indoor scene types and dynamic element plugins on the left, enabling users to construct dynamic environments. In the middle, the platform supports various data types that can be generated, such as RGB, depth, and semantic labels. On the right, the data created using this platform can be applied to train robots for tasks such as navigation, trajectory prediction, and grasping. Through a virtual-real fusion feedback mechanism, the platform allows bidirectional mapping of the states of real and virtual agents, enriching the research scenarios.

accuracy optical motion capture and ROS-based communication. This ensures synchronized validation between virtual and physical robots, facilitating optimization based on simulated feedback. Third, it introduces an intervention-based process. Researchers can adjust virtual scenarios in real-time during physical execution, enhancing task flexibility and robustness, and extending HRC research capabilities.

Key contributions of this work are as follows:

- We present a virtual-real fusion simulation platform (DVS) for robotic research, which enables closed-loop sim-to-real transfer validation through virtual-physical synchronization and ROS-based communication. It supports a wide range of tasks.
- We provide dynamic environmental modeling, including pedestrian behavior simulation and flexible scene editing. These capabilities enhance complex task execution through diverse and high-quality data generation.
- We introduce an intervention-enabled workflow. This supports real-time scenario adjustments during physical deployment. The virtual-real synchronization mechanism improves adaptability in dynamic environments, demonstrated through manipulation tasks.

III. Related Work

Simulation platforms have become integral to the development and validation of embodied AI algorithms, enabling researchers to train and test robotic systems in controlled environments before deployment in real-world tasks. These platforms have seen significant advancements over the past decade, particularly in the areas of physical modeling, scene realism, and task-specific benchmarks.

The rise of embodied intelligence has driven remarkable progress in robotics and artificial intelligence, particularly for tasks that require agents to interact with and navigate real-world environments. Such tasks—ranging from obstacle avoidance and path planning to human-robot collaboration—demand rigorous testing and training frameworks. In this context, simulation environments have emerged as indispensable tools, offering safe, scalable, and cost-effective platforms for developing and validating embodied AI algorithms. These environments not only enable exploration of high-risk scenarios and faster algorithm iteration but also address the critical challenge of sim-to-real (sim2real) generalization, where models trained in simulation must effectively transfer to real-world robotic systems.

Over the past decade, the development of simulation platforms has been instrumental in advancing embodied intelligence. Platforms such as Gazebo[14], MuJoCo[35], and NVIDIA Isaac Sim[23] have excelled in robotics control and high-precision physical simulation, enabling accurate modeling of robot dynamics and multi-robot systems. Meanwhile, tools like Habitat[29], AI2-THOR[15], and iGibson[16] have prioritized photorealistic environments for navigation and task planning, supporting benchmarks for tasks like object rearrangement, manipulation, and visual question answering. Recent systems such as DialFRED[8] and TEACh[25] have expanded the scope of these benchmarks by integrating natural language dialogue, encouraging richer agent-environment interactions. Despite these advancements, several persistent challenges remain unresolved, hindering the broader applicability of these platforms to dynamic, real-world scenarios.

One key limitation lies in the inability to model dynamic, stochastic scenes that reflect realistic human behaviors and environmental changes. Platforms like Habitat and AI2-THOR, while robust for static or semi-static environments, rely heavily on pre-defined tracks and scripted object interactions, which

TABLE I
COMPARISON OF SIMULATION PLATFORMS. FOR THE SENSOR, S REFERS TO SEMANTIC, L REFERS TO LIDAR

Simulation Platform	Sensors	Dynamic Scenes		VR Interaction	ROS
		Pedestrians	Objects		
Arena[12]	RGB-D, L	✓	×	×	✓
AI2THOR[15]	RGB-D, S	×	×	×	×
Gibson series[16][32]	RGB-D, S, L	×	×	✓	✓
HoME[3]	RGB-D, S	×	×	×	×
Habitat[31][34][29]	RGB-D, S	×	×	✓	×
SAPIEN[40]	RGB-D, S	×	×	✓	×
ThreeDWorld[7]	RGB-D, S	×	×	✓	×
VirtualHome[27]	RGB-D, S	×	×	×	×
DVS(Ours)	**RGB-D, S, L**	✓	✓	✓	✓

constrain their generalizability to real-world, unpredictable conditions. Another challenge is the gap in sim2real generalization. While simulators like iGibson and MuJoCo excel in physical modeling, they often lack the diversity and randomness required to robustly train algorithms for real-world deployment. Moreover, the growing emphasis on human-robot collaboration (HRC)[1] has exposed the limitations of existing platforms, which rarely support real-time interactions such as gesture-based commands, shared workspaces, or natural language dialogue. Systems like HumanTHOR[37] and SEAN[36] have made strides in this direction, but their focus remains on basic social navigation or static collaboration tasks, leaving significant room for improvement. Finally, most existing platforms specialize in either physical modeling or photorealistic simulation but fail to integrate these strengths into a unified framework, creating a critical gap in tools that can comprehensively address the needs of embodied intelligence research.

To address these limitations, we propose a novel virtual-physical integration platform that combines the strengths of high-fidelity physics, dynamic scene modeling, and real-time human-robot collaboration. By introducing stochastic pedestrian behavior modeling—including adjustable avoidance radii, randomized spawning points, and variable motion patterns—our platform supports dynamic and unpredictable environments, enhancing the robustness and generalization of robot algorithms. Additionally, a state-of-the-art optical motion capture system provides sub-millimeter precision data for sim2real transfer, ensuring seamless deployment of simulation-trained models to real-world systems. Real-time human-in-the-loop (HITL) interactions[38], including gesture commands, natural language dialogue, and shared workspace collaboration, further enable realistic HRC experiments. Finally, the integration of annotated synthetic data with real-world motion capture allows simultaneous development and validation across virtual and physical domains, bridging a long-standing gap in embodied AI research.

By addressing the critical challenges of dynamic scene modeling, sim2real transfer, and human-robot collaboration, our proposed platform offers a comprehensive solution for advancing embodied intelligence. Its ability to simulate complex, real-world environments and facilitate seamless robot deployment positions it as a transformative tool for future research in navigation, manipulation, and collaboration. A detailed comparison of existing simulation platforms is delineated in Table I.

IV. SYSTEM FRAMEWORK

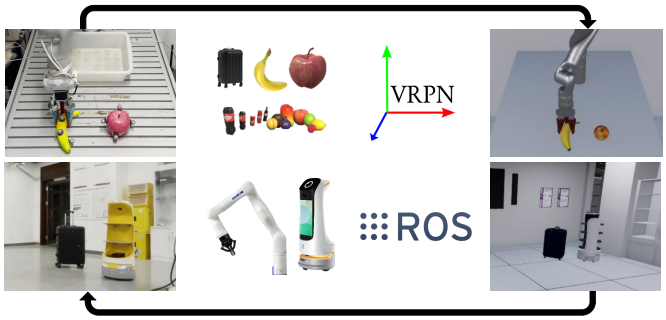

Fig. 2. Virtual-Real Fusion Module

In this section, we describe the key components of our Dynamic Virtual-Real Simulation Platform (DVS), which integrates virtual-real fusion and dynamic scene generation to support advanced robotic research. These two capabilities are designed to address the limitations of existing simulation platforms, enabling more effective training and evaluation of algorithms in real-world conditions. By combining high-fidelity virtual simulation with real-time interactions and dynamic scene modeling, DVS provides a comprehensive environment for testing mobile robots.

A. Virtual-Real Fusion for Seamless Interaction

Virtual-real fusion is a core feature of DVS, enabling precise bidirectional synchronization between the virtual and physical environments. This synchronization is critical for ensuring that algorithms trained in the virtual world can be directly applied to physical robots, thus bridging the sim-to-real gap.

The virtual-real fusion module consists of two primary components: object pose alignment and robot state synchronization. These components work together to ensure that both

the objects and the robots in the simulation environment align accurately with their real-world counterparts.

1) Object Pose Synchronization: Object pose synchronization is a critical feature for bridging the gap between virtual and real environments, enabling accurate interactions between robots and their surroundings in both domains. In DVS, we achieve precise synchronization using a 12-camera motion capture system, which provides real-time tracking with 0.1 mm positional accuracy and 0.1° rotational precision. This allows for high-fidelity pose alignment, essential for ensuring that physical objects, such as robot end-effectors, align accurately with their virtual counterparts in simulation.

The synchronization process begins with extrinsic calibration of the motion capture system. By calibrating the system's extrinsic parameters, we can establish a unified world coordinate system that aligns the virtual and real spaces. This calibration is achieved through the following transformation:

$$T_{\text{virtual}} = R \cdot T_{\text{real}} + t \tag{1}$$

Where:

- R is the rotation matrix derived from the spatial calibration process, defining how the real-world orientation maps to the virtual space.
- T_{real} is the translation vector representing the position of the real-world object.
- t is the translation vector that compensates for any misalignment, ensuring that both spaces share a common origin.

Through this method, the physical object trajectories, such as those of a robot's end-effector, are directly mapped into the virtual environment. This enables precise interaction with virtual objects, improving the realism of simulations and ensuring the accuracy of robotic tasks that require interaction between the real and virtual worlds.

B. Dynamic Scene Generation

The dynamic scene generation module of DVS significantly enhances the realism and complexity of training environments, creating scenarios that more accurately reflect real-world conditions. This module incorporates dynamic pedestrian agents and mobile robotic proxies, both of which are key to simulating the unpredictability and complexity of real-world environments.

1) Dynamic Pedestrian Plug-in: DVS features a pedestrian simulation plugin that introduces human-like agents into the virtual environment. These agents are equipped with variable motion accelerations and socially compliant avoidance behaviors, allowing them to navigate environments with high-density crowd dynamics. The agents' behaviors are modeled to mimic real human interactions, including variable speeds, random movement patterns, and avoidance of obstacles. This makes the platform capable of replicating environments such as crowded supermarkets, busy restaurants, or indoor spaces with dynamic obstacles. The inclusion of pedestrians enhances the realism of the simulation, as mobile robots must navigate and collaborate within environments populated by humans.

This dynamic pedestrian behavior is essential for training robots on navigation algorithms and human-robot interaction tasks. The agents interact with robots in real-time, allowing researchers to collect diverse data that can be used to refine navigation strategies, path planning, and collaboration algorithms. This is especially important for tasks requiring the robot to adapt to unexpected changes in the environment or human movements.

2) Multi-Robot Plug-in: In addition to pedestrian simulation, DVS supports the integration of multiple mobile robotic agents within the same environment. This capability allows researchers to study multi-robot collaboration and competition in dynamic settings. The simulation of multiple robots operating in close proximity enables the development of cooperative algorithms for tasks such as resource sharing, coordinated navigation, and joint manipulation.

The ability to simulate multi-robot environments in dynamic, cluttered spaces is critical for advancing robotics research. By mimicking real-world challenges such as managing crowded environments or dealing with unexpected obstacles, DVS helps researchers develop more robust algorithms that can handle complex tasks in unpredictable settings.

Together, dynamic pedestrians and multi-robot integration ensure that DVS provides a training environment that closely mirrors real-world operational conditions. These capabilities are essential for developing robots that can navigate complex spaces, collaborate with humans, and adapt to dynamic changes in their environments.

V. Applications of DVS Platform

Our platform supports the full workflow, from data generation to real-world validation. In the previous chapter, we introduced two core modules of our system. This chapter discusses the construction of a large-scale virtual-real fusion dataset and explores the experimental data generation process, along with its application in task training and testing.

A. Data Perception and Generation

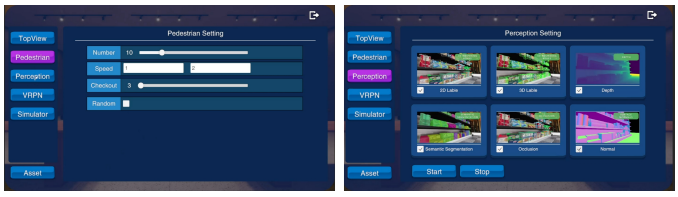

Fig. 3. The interactive interface of the simulation platform: dynamic pedestrian parameters are adjusted on the left, and perception data types are selected on the right.

In robotics research, virtual environments provide clear task representations, enabling agents to perform tasks in controlled settings. Data generation is a core feature of simulation platforms. As shown in Figure 3, our platform facilitates the generation and processing of various data formats, including RGB images, depth maps, 2D/3D bounding boxes, semantic and instance segmentation, and trajectory data, all via a user-friendly interface. These data types support foundational tasks

and enable complex research scenarios. For instance, Liao et al. [18] propose generating activities based on environmental context, Puig et al. [28] investigate human-robot social perception and collaboration, and Li et al. [17] integrate language models to assist robots in decision-making, thereby broadening the scope of research in virtual environments.

To improve data quality and usability, we optimize the data generation process by ensuring smooth camera trajectories and precise depth alignment. Bezier curves are employed to plan smooth camera motion, reducing abrupt changes at trajectory corners, which enhances frame-to-frame feature matching and point cloud reconstruction. Depth data is aligned with RGB timestamps to ensure precise synchronization, which is crucial for multi-sensor fusion and complex scene modeling. These optimizations ensure a smooth transition from static to dynamic environments, providing robust support for advancing robotics research, even as task complexity increases.

B. Robotic Tasks Learning

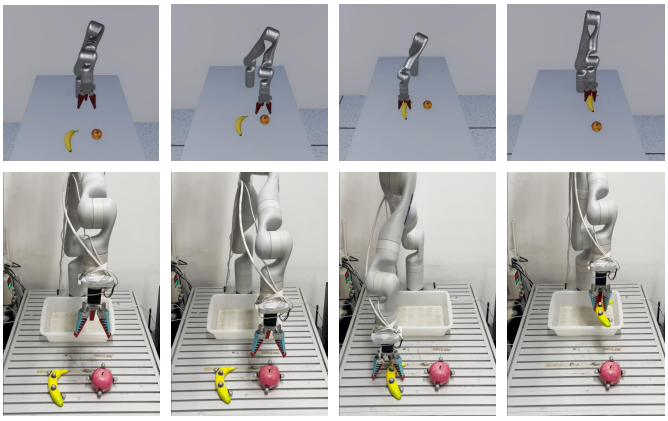

Fig. 4. The robotic arm is interrupted while executing Prompt A and is requested to execute Prompt B. The first row shows the robotic arm in the virtual platform, and the second row shows the real robotic arm.

1) virtual-real intervention Grasping: A key weakness of learned policies in robotic manipulation is that their success rate in task execution is low when deployed in practice. In heterogeneous deployments using only pretrained weights, the success rate of robots performing tasks across different models tends to approach zero. Even when data collection for specific tasks is done using the robot being deployed and fine-tuned, the success rate of task execution is still only around 90%, making it difficult to apply in the industry. However, due to the characteristics of our platform, which includes virtual-real mapping and benchmark alignment between the virtual environment and the real world, and the fact that the robot has a ROS communication interface, we can supervise and intervene in the robot's tasks in the real world through the platform to improve the success rate of task execution. We set up experimental conditions based on the common manipulation task of grasping. As shown in Figure 44, in order to reflect the characteristics of our platform supervision and intervention, we provide the gripper with wrong instructions at the beginning of the experiment, and interrupt the task and provide new tasks based on virtual scenes through the platform when the gripper is performing the task. We utilized a seven-degree-of-freedom Kinova Gen3 robotic arm to collect nearly a hundred grasping data points on a planar surface. The data was then fine-tuned on the pre-trained models released by OpenVLA-7B[13] and RDT-1B[19], enabling our robotic arm to achieve a high success rate in performing tasks in specific scenarios. At the beginning of the experiment, we provided the robotic arm with prompts to grasp an apple and a banana, and midway through task execution, we interrupted the task on the platform and assigned a new task.The experiment demonstrated that our platform effectively intervened in the robotic arm's task execution. The experimental results are shown in the tableII. Prompts: A: "Pick up the apple"; B: "Pick up the banana."

TABLE II
TASK SUCCESS RATES BY MODULE AND PROMPT ORDER

Module	Prompt Order		Success Rate (%)	
	First	Second	First Task	Second Task
OpenVLA-7B	A	B	0.0	100.0
	B	A	0.0	90.0
RDT-1B	A	B	0.0	80.0
	B	A	0.0	90.0

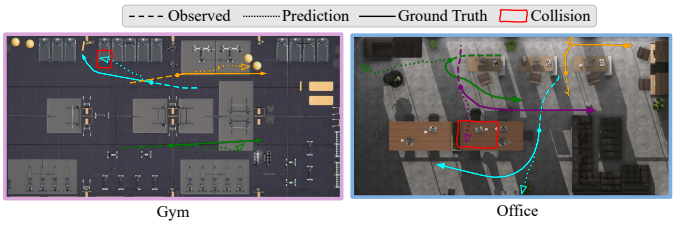

Fig. 5. Visualization of pedestrian trajectory prediction, where each color represents a different pedestrian. The accuracy of the prediction is higher when the predicted trajectory (short dashed line) closely aligns with the ground truth (GT, solid line). In environments with dense static obstacles, such as indoors, the predicted future trajectory may result in collisions (red rectangular box).

2) Dynamic Indoor Pedestrian Trajectory Prediction: Pedestrian trajectory prediction aims to forecast future trajectories based on observed trajectories, while considering complex interactions and environmental layouts. It serves as a crucial connection between the perception system and the planning system.

Three trajectory prediction algorithms, i.e. STGAT [10], Trajectron++ [30] and TUTR [33], are tested on our synthetic indoor scenes (Gym, Office and Supermarket) as well as the official public outdoor dataset (ETH [26]).We use ADE (Average Displacement Error) and FDE (Final Displacement Error) as evaluation metrics, where lower ADE and FDE values indicate better performance. The experimental results are depicted in Table III. Additionally, to analyze pedestrian movement patterns and collision avoidance strategies, we selected two dense indoor scenes (Gym and Office) and visualized the predicted trajectories in Fig. 5.

Overall, all three methods experience a significant performance decrease when applied to indoor scenes compared to the outdoor ETH scene. Specifically, the ADE for STGAT decreases from 0.79 to 1.42 (79.7%) when generalizing from the ETH scene to the Supermarket scene, while the FDE for STGAT decreases from 1.48 to 2.88 (94.5%) in the same scenario. We analyze this performance drop from three perspectives. First, compared to outdoor scenes, narrow indoor spaces are often filled with numerous static obstacles, which can interfere with human trajectory decision-making and lead to collisions. Second, indoor human interactions are more frequent due to communication or obstacles caused by people standing in the way, making predictions more challenging. Third, indoor spaces are generally smaller than outdoor environments, with pedestrian trajectories being less spread out, making predictions more sensitive to small positional changes. If the model was trained on larger, more open outdoor spaces, it may not have learned to adapt to the smaller, more dynamic movements of indoor environments.

The results also underscore the importance of robust spatial-temporal modeling in trajectory prediction tasks. TUTR's transformer-based architecture appears particularly well-suited for capturing intricate interactions over time, leading to its superior performance. Trajectron provides a balance of stability and accuracy but lags behind in highly dynamic environments. Conversely, STGAT's graph-based approach, while effective in simpler scenarios, struggles in complex environments, highlighting its limitations in handling high-dimensional spatial-temporal variability. These findings offer valuable insights for future research, emphasizing the need for models that can generalize effectively across diverse scenarios while maintaining low computational overhead.

TABLE III
EXPERIMENTS ON PEDESTRIAN TRAJECTORY PREDICTION. GYM, OFFICE AND SUPERMARKET ARE OUR SYNTHETIC INDOOR SCENES, WHILE ETH [26] IS THE OFFICIAL PUBLIC OUTDOOR DATASET.

Scene	Method	ADE ↓	FDE ↓
Gym	STGAT	1.39	3.01
	Trajectron++	0.59	1.02
	TUTR	0.70	1.19
Office	STGAT	1.38	2.75
	Trajectron++	0.89	1.60
	TUTR	0.81	1.40
Supermarket	STGAT	1.42	2.88
	Trajectron++	0.96	1.82
	TUTR	0.83	1.50
ETH	STGAT	0.79	1.48
	Trajectron++	0.52	0.97
	TUTR	0.43	0.83

VI. CONCLUSION

We propose a dynamic virtual-Real simulation platform that integrates configurable pedestrian behavior simulation, large-scale indoor environments, optical motion capture, and ROS-based bidirectional virtual-reality communication. The platform introduces two major innovative modules for virtual-reality integration, overcoming current limitations in robotic simulation systems for dynamic scenarios and real-world deployment. Experimental results show that DVS supports navigation and human-robot interaction research, achieving closed-loop performance in real-world missions. Future work will focus on integrating haptic feedback, developing AI-driven intervention strategies, and improving compatibility with industrial robotic arms. This platform creates a new paradigm for closed-loop virtual-reality interaction, advancing human-robot collaboration and dynamic environment adaptation.

REFERENCES

[1] Arash Ajoudani, Andrea Maria Zanchettin, Serena Ivaldi, Alin Albu-Schäffer, Kazuhiro Kosuge, and Oussama Khatib. Progress and prospects of the human–robot collaboration. *Autonomous robots*, 42:957–975, 2018.

[2] Janis Arents and Modris Greitans. Smart industrial robot control trends, challenges and opportunities within manufacturing. *Applied Sciences*, 12(2):937, 2022.

[3] Simon Brodeur, Ethan Perez, Ankesh Anand, Florian Golemo, Luca Celotti, Florian Strub, Jean Rouat, Hugo Larochelle, and Aaron Courville. Home: A household multimodal environment. *arXiv preprint arXiv:1711.11017*, 2017.

[4] Chang Che, Bo Liu, Shulin Li, Jiaxin Huang, and Hao Hu. Deep learning for precise robot position prediction in logistics. *Journal of Theory and Practice of Engineering Science*, 3(10):36–41, 2023.

[5] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning, 2016.

[6] Sanjoy Das, Indrani Das, Rabindra Nath Shaw, and Ankush Ghosh. Advance machine learning and artificial intelligence applications in service robot. In *Artificial Intelligence for Future Generation Robotics*, pages 83–91. Elsevier, 2021.

[7] Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020.

[8] Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4): 10049–10056, 2022.

[9] Jane Holland, Liz Kingston, Conor McCarthy, Eddie Armstrong, Peter O'Dwyer, Fionn Merz, and Mark McConnell. Service robots in the healthcare sector. *Robotics*, 10(1):47, 2021.

[10] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings*

of the IEEE/CVF international conference on computer vision, pages 6272–6281, 2019.

[11] Linh Kästner, Reyk Carstens, Lena Nahrwold, Christopher Liebig, Volodymyr Shcherbyna, Subhin Lee, and Jens Lambrecht. Demonstrating arena-web: A web-based development and benchmarking platform for autonomous navigation approaches. In *Robotics: Science and Systems*, 2023.

[12] Linh Kästner, Volodymyir Shcherbyna, Huajian Zeng, Tuan Anh Le, Maximilian Ho-Kyoung Schreff, Halid Osmaev, Nam Truong Tran, Diego Diaz, Jan Golebiowski, Harold Soh, et al. Arena 3.0: Advancing social navigation in collaborative and highly dynamic environments. *arXiv preprint arXiv:2406.00837*, 2024.

[13] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[14] Nathan Koenig and Andrew Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ international conference on intelligent robots and systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 3, pages 2149–2154. Ieee, 2004.

[15] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[16] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*, 2021.

[17] Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, et al. Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems*, 35:31199–31212, 2022.

[18] Yuan-Hong Liao, Xavier Puig, Marko Boben, Antonio Torralba, and Sanja Fidler. Synthesizing environment-aware activities via activity sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6291–6299, 2019.

[19] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

[20] Zhihao Liu, Quan Liu, Wenjun Xu, Lihui Wang, and Zude Zhou. Robot learning towards smart robotic manufacturing: A review. *Robotics and Computer-Integrated Manufacturing*, 77:102360, 2022.

[21] Matteo Luperto, Javier Monroy, Jennifer Renoux, Francesca Lunardini, Nicola Basilico, Maria Bulgheroni, Angelo Cangelosi, Matteo Cesari, Manuel Cid, Aladar Ianes, et al. Integrating social assistive robots, iot, virtual communities and smart objects to assist at-home independently living elders: the movecare project. *International Journal of Social Robotics*, 15(3):517–545, 2023.

[22] Eloise Matheson, Riccardo Minto, Emanuele GG Zampieri, Maurizio Faccio, and Giulio Rosati. Human–robot collaboration in manufacturing applications: A review. *Robotics*, 8(4):100, 2019.

[23] Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 8(6):3740–3747, 2023. doi: 10.1109/LRA.2023.3270034.

[24] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023.

[25] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. TEACh: Task-driven Embodied Agents that Chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025, 2022.

[26] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th international conference on computer vision*, pages 261–268. IEEE, 2009.

[27] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8494–8502, 2018.

[28] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. *arXiv preprint arXiv:2010.09890*, 2020.

[29] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023.

[30] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 683–700. Springer, 2020.

[31] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019.

[32] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, et al. igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7520–7527. IEEE, 2021.

[33] Liushuai Shi, Le Wang, Sanping Zhou, and Gang Hua. Trajectory unified transformer for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9675–9684, 2023.

[34] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in neural information processing systems*, 34:251–266, 2021.

[35] 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.

[36] Nathan Tsoi, Alec Xiang, Peter Yu, Samuel S. Sohn, Greg Schwartz, Subashri Ramesh, Mohamed Hussein, Anjali W. Gupta, Mubbasir Kapadia, and Marynel Vázquez. Sean 2.0: Formalizing and generating social situations for robot navigation. *IEEE Robotics and Automation Letters*, 7(4):11047–11054, 2022. doi: 10.1109/LRA.2022.3196783.

[37] Chenxu Wang, Boyuan Du, Jiaxin Xu, Peiyan Li, Di Guo, and Huaping Liu. Demonstrating humanthor: A simulation platform and benchmark for human-robot collaboration in a shared workspace. *arXiv preprint arXiv:2406.06498*, 2024.

[38] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381, 2022.

[39] Fei Xia, William B Shen, Chengshu Li, Priya Kasimbeg, Micael Edmond Tchapmi, Alexander Toshev, Roberto Martín-Martín, and Silvio Savarese. Interactive gibson benchmark (igibson 0.5): A benchmark for interactive navigation in cluttered environments, 2020.

[40] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020.

[41] Dongbo Xie, Liang Chen, Lichao Liu, Liqing Chen, and Hai Wang. Actuators and sensors for application in agricultural robots: A review. *Machines*, 10(10):913, 2022.