EmbodiedSplat: Personalized Real-to-Sim-to-Real Navigation with Gaussian Splats from a Mobile Device

Gunjan Chhablani[†], Xiaomeng Ye [†], Rynaa Grover [†], Muhammad Zubair Irshad [§], Zsolt Kira [†]
[†] Georgia Institute of Technology [§] Toyota Research Institute

Abstract

Sim-to-real transfer and personalization remains a core challenge in Embodied AI due to a trade-off between synthetic environments lacking realism and costly real-world captures. We present EmbodiedSplat, a method that personalizes policy training by reconstructing deployment environments using a mobile device and 3D Gaussian Splatting, enabling efficient fine-tuning in realistic scenes via Habitat-Sim. Our analysis of training strategies and reconstruction techniques shows that EmbodiedSplat achieves significant gains—improving real-world ImageNav success by 20–40% over pre-trained policies in an out-of-domain scene—and exhibits strong sim-to-real correlation (0.87–0.97). Code and data will be made public.

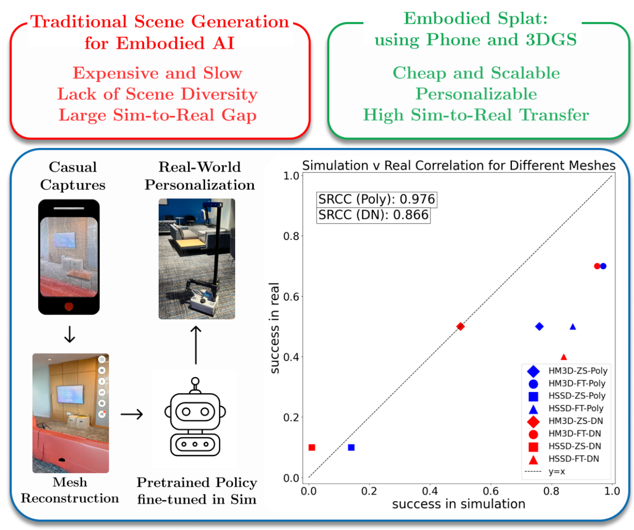

Figure 1. **Overview of EmbodiedSplat:** Mobile phone captures generate 3D Gaussian Splatting meshes for simulation training, enabling agents to transfer effectively to the real world with strong sim-to-real correlation across mesh types.

1. Introduction

Recent advances in Embodied AI have shown strong performance in simulation [7, 8, 15, 22, 23], but sim-to-real transfer remains a major challenge due to limited simulation fidelity and accessibility [10]. Synthetic environments like HSSD [12] often lack real-world complexity, while datasets such as Matterport3D [3] and HM3D [19] rely on expensive hardware and labor-intensive pipelines, limiting scalability and adaptation to diverse deployment settings.

To address this, we propose a framework that leverages open-source 3D Gaussian Splatting [11] (compared with Polycam [17]) to quickly capture deployment scenes using consumer-grade devices and integrate them into Habitat-Sim [18]. This enables training in realistic simulations, improving sim-to-real transfer. Our method combines smartphone accessibility with recent advances in depth-aware 3D representations, supporting rapid policy adaptation.

We evaluate our framework in an out-of-distribution university scene, analyzing reconstruction pipelines and training strategies. Real-world experiments show significant performance gains in image-goal navigation - 20-40% absolute success rate (SR) compared to pre-trained policies on HM3D and HSSD.

While there has been work on using 3D representations for robotics [2, 4, 9, 13, 14, 16, 26], to the best of our knowledge, we are the first to explore a solution towards real-world transfer and out-of-domain personalization for image-goal navigation using 3DGS. Through this work, we attempt to democratize high-quality scene capture and policy training, making it easier to build personalized agents.

2. Methodology

The overall pipeline for bridging and integrating a real-world scene with Habitat-Sim [18] is shown in Fig. 2.

Scene Capture: Our real-world scene is a community lounge set in a university environment. We use a manually-held iPhone 13 Pro Max to record the iPhone RGB-D data using the Polycam application [17, 20]. These captures are processed using Nerfstudio [24] to sample $\sim$1000 frames.

Mesh Reconstruction: We use DN-Splatter [25] as our method of choice for its superior performance on mesh reconstruction, with Metric3D-V2 [6] as our normal encoder.

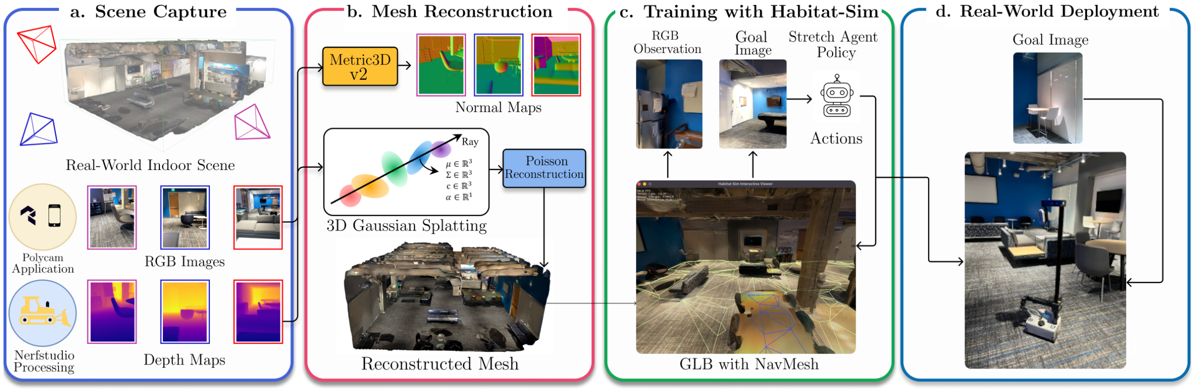

Figure 2. **The EmbodiedSplat Pipeline**: Integrating real-world captures with Habitat-Sim [18]: (a.) Capture scenes with Polycam [17] and extract data using Nerfstudio [24] (b.) Use DN-Splatter [25] to train GS using depth-normal regularization, with normals from Metric3D-V2 [6] (c.) Process the mesh and load into Habitat-Sim (d.) Deploy trained policy in the real world.

It takes approximately 20-30 minutes per capture, and 1-2 hours of training with DN-Splatter [25] to generate this mesh, which is significantly lesser compared to the cost and several hours of capture and processing with Matterport [1] cameras. In addition to the mesh produced, Polycam also provides a mesh with its exported data. We use this mesh for comparison purposes. We fix the orientation of these meshes using Blender, and then load them into Habitat-Sim [18], on which ImageNav episodes are generated. The training and deployment is done following the Habitat [18] pipeline and Home-Robot [21] framework.

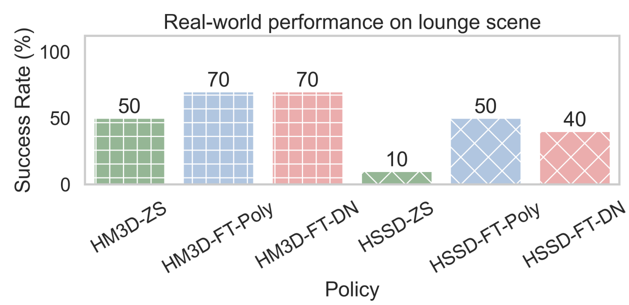

Figure 3. **Real world results of zero-shot and fine-tuned models on `lounge` scene**.

3. Experimental Results

First, we pre-trained policies on the HM3D [19] (83.08% HM3D val SR) and HSSD [12] (63.15% HSSD val SR), trained for 600M and 1200M steps, respectively. Then, we fine-tuned the policies for 20M additional steps with learning rate $2.5e{-}6$ for the LSTM policy and $6e{-}7$ for the visual encoder, following a fine-tuning strategy similar to that of Deitke et al. [5]. We evaluate both zero-shot and fine-tuned policies in the real-world scene on a Stretch robot for 10 episodes each capped at 100 steps. To evaluate success, we record number of steps and the distance to the goal.

Fig. 3 shows that the zero-shot HM3D policy achieves a 50% SR, demonstrating our hypothesized lack of generalization. This is in contrast with the results reported in Silwal et al. [23] showing 90% zero-shot real-world SR. We attribute this discrepancy to the structural and semantic differences between the `lounge` and the apartment-style scenes typically encountered in HM3D. Fine-tuning on the POLYCAM and DN mesh reconstructions of this scene improves performance, with real-world SRs increasing up to 70%. For HSSD, zero-shot performance is significantly lower at 10%, while fine-tuned policies improve SRs to 50% with POLYCAM and 40% with DN mesh. These results highlight the need for realistic captures, especially high-fidelity reconstructions of the deployment environment, to help with improved sim-to-real transfer.

Fig. 1 illustrates the Sim-to-Real Correlation Coefficient (SRCC) [10] between simulation and real-world performance. The observation suggests that improvements in evaluation performance on DN and POLYCAM meshes in simulation translate to improved real-world performance. This demonstrates that our approach can efficiently adapt policies to novel real-world environments.

4. Conclusion

In this work, we presented a scalable pipeline for bridging the sim-to-real gap in image navigation using 3D Gaussian Splats and Polycam. Leveraging iPhone-captured scenes, our approach enables efficient policy personalization and high-quality training with minimal effort and cost. This practical framework supports accessible scene collection for large-scale embodied AI research. In future, we aim to extend this approach to more complex tasks, such as rearrangement and mobile manipulation, to further advance real-world applications.

References

[1] Capture, share, and collaborate the built world in immersive 3D — matterport.com. https://matterport.com/. [Accessed 08-03-2025]. 2

[2] Arunkumar Byravan, Jan Humplik, Leonard Hasenclever, Arthur Brussee, Francesco Nori, Tuomas Haarnoja, Ben Moran, Steven Bohez, Fereshteh Sadeghi, Bojan Vujatovic, and Nicolas Heess. Nerf2real: Sim2real transfer of vision-guided bipedal motion skills using neural radiance fields, 2022. 1

[3] 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, 2017. 1

[4] Timothy Chen, Ola Shorinwa, Joseph Bruno, Javier Yu, Weijia Zeng, Keiko Nagami, Philip Dames, and Mac Schwager. Splat-nav: Safe real-time robot navigation in gaussian splatting maps, 2024. 1

[5] Matt Deitke, Rose Hendrix, Luca Weihs, Ali Farhadi, Kiana Ehsani, and Aniruddha Kembhavi. Phone2proc: Bringing robust robots into our chaotic world, 2022. 2

[6] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. 2024. 1, 2

[7] Muhammad Zubair Irshad, Chih-Yao Ma, and Zsolt Kira. Hierarchical cross-modal agent for robotics vision-and-language navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021. 1

[8] Muhammad Zubair Irshad, Niluthpol Chowdhury Mithun, Zachary Seymour, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. Semantically-aware spatio-temporal reasoning agent for vision-and-language navigation in continuous environments. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4065–4071, 2022. 1

[9] Muhammad Zubair Irshad, Mauro Comi, Yen-Chen Lin, Nick Heppert, Abhinav Valada, Rares Ambrus, Zsolt Kira, and Jonathan Tremblay. Neural fields in robotics: A survey, 2024. 1

[10] Abhishek Kadian, Joanne Truong, Aaron Gokaslan, Alexander Clegg, Erik Wijmans, Stefan Lee, Manolis Savva, Sonia Chernova, and Dhruv Batra. Sim2real predictivity: Does evaluation in simulation predict real-world performance? *IEEE Robotics and Automation Letters*, 5(4):6670–6677, 2020. 1, 2

[11] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1

[12] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Schacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. *arXiv preprint arXiv:2306.11290*, 2023. 1, 2

[13] Xiaohan Lei, Min Wang, Wengang Zhou, and Houqiang Li. Gaussnav: Gaussian splatting for visual navigation, 2024. 1

[14] Yulong Li and Deepak Pathak. Object-aware gaussian splatting for robotic manipulation. In *ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation*, 2024. 1

[15] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence?, 2024. 1

[16] Pierre Marza, Laetitia Matignon, Olivier Simonin, Dhruv Batra, Christian Wolf, and Devendra Singh Chaplot. Autonerf: Training implicit scene representations with autonomous agents, 2023. 1

[17] Polycam. Polycam. Accessed: 2024-11-10. 1, 2

[18] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars, and robots. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 2

[19] Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai, 2021. 1, 2

[20] Xuqian Ren, Wenjia Wang, Dingding Cai, Tuuli Tuominen, Juho Kannala, and Esa Rahtu. Mushroom: Multi-sensor hybrid room dataset for joint 3d reconstruction and novel view synthesis, 2024. 1

[21] Meta AI Research. Home robot. https://github.com/facebookresearch/home-robot, 2023. GitHub repository. 2

[22] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research, 2019. 1

[23] Sneha Silwal, Karmesh Yadav, Tingfan Wu, Jay Vakil, Arjun Majumdar, Sergio Arnaud, Claire Chen, Vincent-Pierre Berges, Dhruv Batra, Aravind Rajeswaran, Mrinal Kalakrishnan, Franziska Meier, and Oleksandr Maksymets. What do we learn from a large-scale study of pre-trained visual representations in sim and real environments?, 2024. 1, 2

[24] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 1, 2

[25] Matias Turkulainen, Xuqian Ren, Iaroslav Melekhov, Otto Seiskari, Esa Rahtu, and Juho Kannala. Dn-splatter: Depth and normal priors for gaussian splatting and meshing, 2024. 1, 2

[26] Yuxuan Wu, Lei Pan, Wenhua Wu, Guangming Wang, Yanzi Miao, and Hesheng Wang. Rl-gsbridge: 3d gaussian splatting based real2sim2real method for robotic manipulation learning, 2024. 1