robot learning freiburg

ZUSE SCHOOL ELIZA

real2gen.cs.uni-freiburg.de

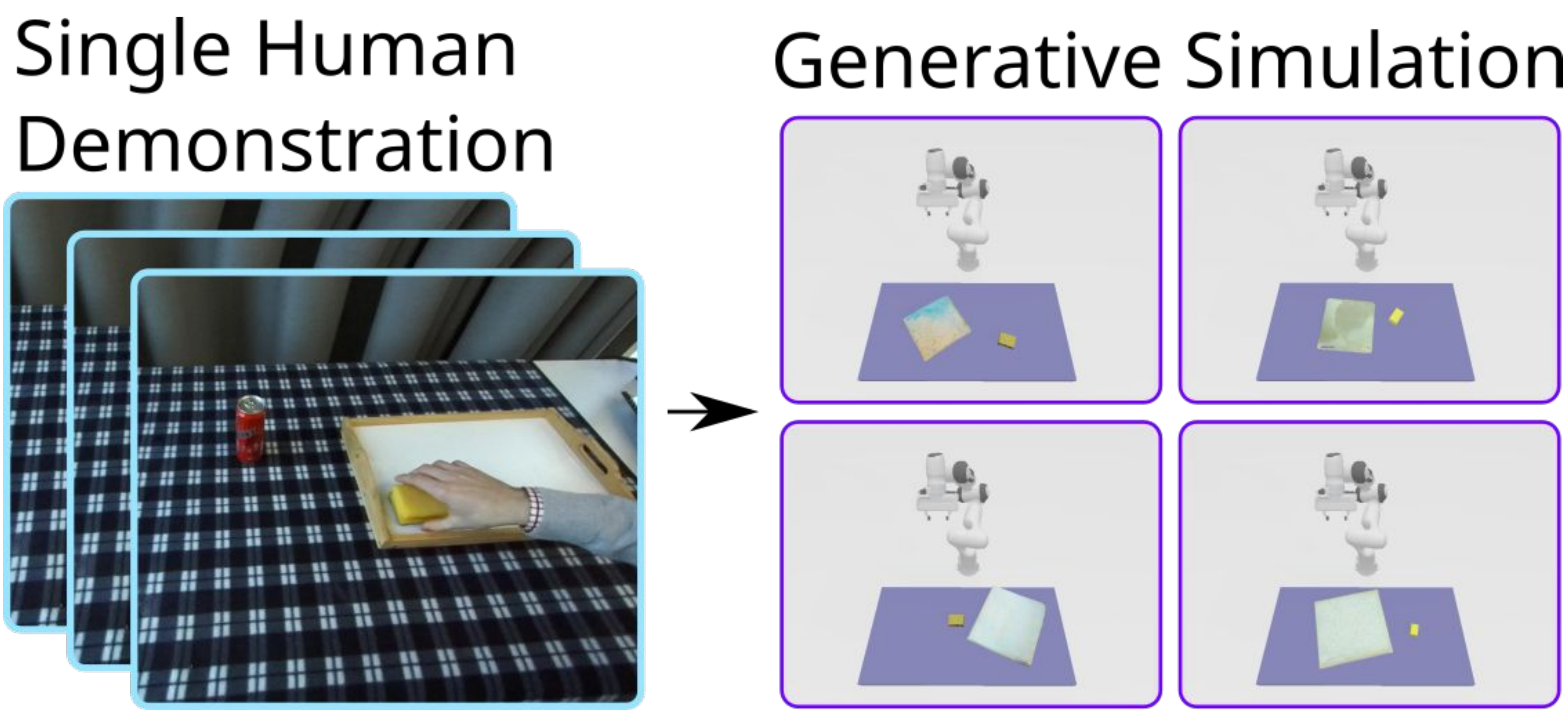
ICRA IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION — ATLANTA 2025

Real2Gen: Imitation Learning from a Single Human Demonstration with Generative Foundational Models

Nick Heppert[1, 2]*, Minh Quang Nguyen[1]*, Abhinav Valada[1]

[1]University of Freiburg, [2]Zuse School ELIZA, *equal contribution

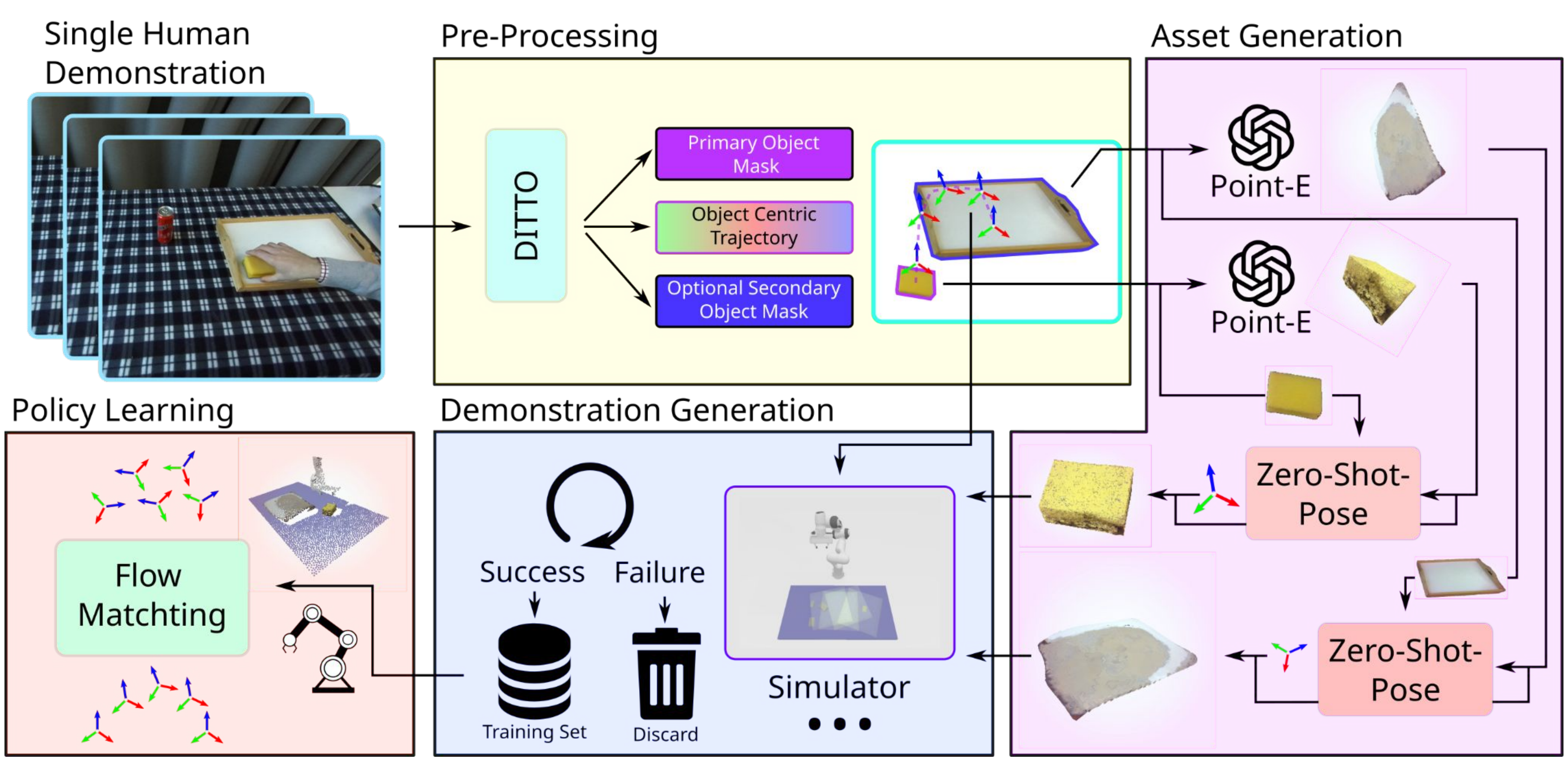
Motivation



- Imitation learning promising paradigm to learn new tasks
- Commonly trained on robot demos, but collecting with tele-operation or kinesthetic teaching is tedious

→ Real2Gen: Transform human demos to robot demos using generative simulation

Real2Gen Overview



1. Use DITTO[1] or ORION[2] to extract object masks + traj.
2. Generate 3D asset using Point-E[3] and align to demo with ZSP[4]
3. Re-generate scenarios and use privileged info to generate demonstrations
4. Train a policy, e.g. flow matching[5]

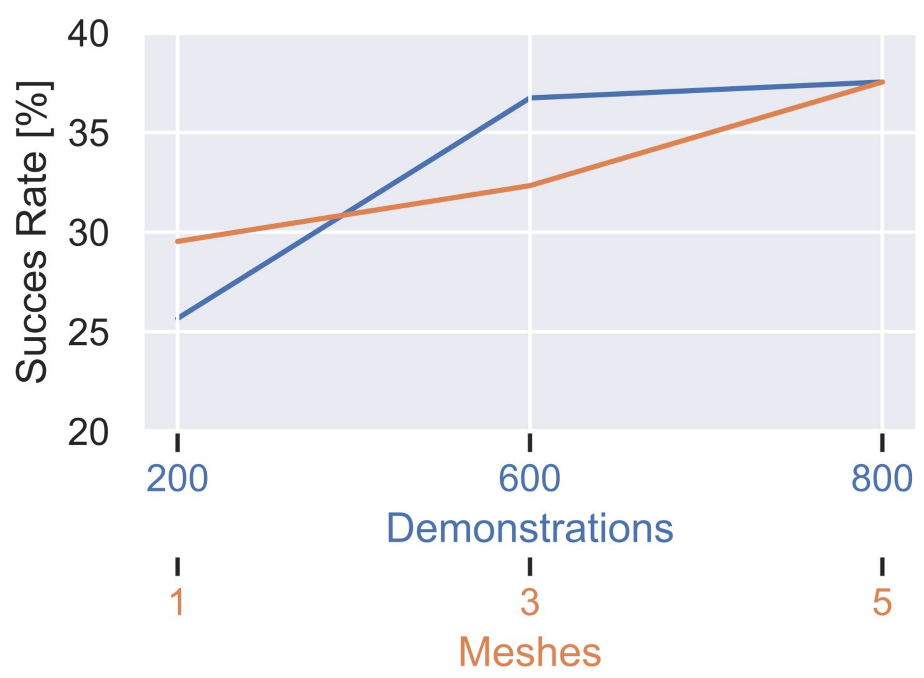
Experiment – Policy Learning

Policy Learning Results

Method	Sponge on Tray [%] (↑)	Coke on Tray [%] (↑)	Paperroll upright [%] (↑)	Mean SR [%] (↑)
DITTO[1]	$6.3^{\pm 2.1}$	$26.0^{\pm 3.6}$	$0.3^{\pm 0.5}$	$10.9^{\pm 2.1}$
DITTO[1] w/ ZSP[4]	$4.3^{\pm 1.2}$	$19.7^{\pm 3.8}$	$0.7^{\pm 0.6}$	$8.2^{\pm 1.8}$
Real2Gen (ours)	$41.3^{\pm 4.5}$	$46.3^{\pm 6.4}$	$25.0^{\pm 1.0}$	$37.5^{\pm 3.0}$

- Real2Gen shows robustness over DITTO

Ablation Study



- Performance gain diminishes with more demonstrations
- Mesh amount are more relevant

Acknowledgement: This work was funded by the Carl Zeiss Foundation with the ReScaLe project. Nick Heppert is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

Experiment – Mesh Generation

- Compare human effort for mesh retrieval against querying a database using tags
- Apply matching and manually verify
- Real2Gen provides almost 3x more available meshes

Mesh Source	Available Meshes	100 Mesh Pre-Selection	Matching Successful and Task Relevant
Point-E[3] (ours)	∞	Random	**54%**
Objaverse[6]	690	Most viewed*	19%
		Random*	18%

**if less than a 100 meshes are available we use all*

References

[1] N. Heppert, M. Argus, T. Welschehold, T. Brox, and A. Valada, "Ditto: Demonstration imitation by trajectory transformation," in Proc. IEEE Int. Conf. on Intel. Rob. and Syst. IEEE, 2024, pp. 7565–7572.

[2] Y. Zhu, A. Lim, P. Stone, and Y. Zhu, "Vision-based manipulation from single human video with open-world object graphs," arXiv preprint arXiv:2405.20321, 2024.

[3] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen, "Point-e: A system for generating 3d point clouds from complex prompts," arXiv preprint arXiv:2212.08751, 2022.

[4] W. Goodwin, S. Vaze, I. Havoutis, and I. Posner, "Zero-shot category-level object pose estimation," in Proc. Springer Eur. Conf. Comput. Vis. Springer, 2022, pp. 516–532.

[5] E. Chisari, N. Heppert, M. Argus, T. Welschehold, T. Brox, and A. Valada, "Learning robotic manipulation policies from point clouds with conditional flow matching," Proc. Conf. on Rob. Learn., 2024.

[6] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2023, pp. 13 142–13 153.