Making Universal Policies Universal

Extended Abstract

Niklas Hoepner*
University of Amsterdam
Amsterdam, Netherlands
n.r.hopner@uva.nl

David Kuric*
University of Amsterdam
Amsterdam, Netherlands
d.kuric@uva.nl

Herke van Hoof
University of Amsterdam
Amsterdam, Netherlands
h.c.vanhoof@uva.nl

ABSTRACT

The development of a generalist agent capable of solving a wide range of sequential decision-making tasks remains a significant challenge. We address this problem in a cross-agent setup where agents share the same observation space but differ in their action spaces. Our approach builds on the universal policy framework, which decouples policy learning into two stages: a diffusion-based planner that generates observation sequences and an inverse dynamics model that assigns actions to these plans. We propose a method for training the planner on a joint dataset composed of trajectories from all agents. This method offers the benefit of positive transfer by pooling data from different agents, while the primary challenge lies in adapting shared plans to each agent's unique constraints. We evaluate our approach on the BabyAI environment, covering tasks of varying complexity, and demonstrate positive transfer across agents. Additionally, we examine the planner's ability to generalise to unseen agents and show that our method outperforms traditional imitation learning approaches[1].

KEYWORDS

Cross-Agent Learning; Diffusion Models; Generalist Agent

ACM Reference Format:

Niklas Hoepner, David Kuric, and Herke van Hoof. 2025. Making Universal Policies Universal: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Developing a generalist agent capable of addressing diverse sequential decision-making tasks remains a significant challenge [10, 13]. Solving this problem would eliminate the need for task-specific engineering and retraining while enabling positive transfer between tasks. A common ground for many tasks lies in image-based observations, which are prevalent in gameplay [5], robotics [12], and web interfaces [2]. Especially in robotics, learning across different embodiments with differing action and observation space has gathered interest as it allows to train on large mixture datasets and leverage positive transfer for more robust control policies [3, 8, 12]. Creating policies that can be trained on mixture datasets of different agents is an essential step towards a generalist agent.

Universal policies [4] use text-guided video generation to train policies. This involves a two-step process: first, a diffusion model translates task descriptions into observation sequences; second, an inverse dynamics model maps these sequences to actions. This approach enables pretraining on vast instruction-video datasets [7, 11]. While universal policies have been used for individual agents, their potential to handle multiple agents with shared planners and agent-specific inverse dynamics models remains unexplored.

Our study examines this problem in a cross-agent setting where multiple agents share a common observation space but differ in their action spaces. Each agent has limited instruction-trajectory data, insufficient for training robust, agent-specific policies through imitation learning. We extend the universal policy framework to learn a policy applicable across agents that can be trained on the joint dataset obtained by pooling the agent-specific data. The main challenge is ensuring the diffusion-based planner accommodates varying agent capabilities. Without conditioning, the planner risks generating sequences incompatible with an agent's type, leading to errors. However, leveraging combined data provides an opportunity for positive transfer, exposing the planner to a broader set of examples and potentially improving performance across agents. We explore methods to condition the planner on agent-specific information and evaluate its generalisation to unseen agents.

2 UNIVERSAL CROSS AGENT POLICIES

In our setup each agent $n \in N$ has a dataset D_n of M_n instruction-trajectory pairs: $D_n = \{(c_i, x_{1:t_i}, a_{i:t_i})\}_{i=1}^{M_n}$, where $c_i \in C$ is the instruction, $a_{i:t_i}$ is the action sequence, and $x_{1:t_i} \in X^{t_i}$ is the observation sequence. We consider the case where all agents share the same observation space. The datasets D_n are pooled into a mixed dataset $D = \{(c_i, x_{1:t_i}, a_{i:t_i}, n_i)\}_{i=1}^{M}$, where $n_i \in N$ is the agent ID and $M = \sum_{n=1}^{N} M_n$ is the total number of trajectories. The goal is to train a conditional observation sequence generator $p(\cdot|x_0, c, k)$ on the mixture dataset D, leading to a **U**niveral **C**ross **A**gent **P**olicy (UCAP). Instead of generating full sequences $x_{1:t_i}$, we sample random windows of size 4, using the first timestep as the starting observation x_0. The model $p(\cdot|x_0, c, k)$ plans the next three timesteps for agent k following instruction c.

Diffusion Model Formulation: Diffusion models perturb data by adding noise to the data and learn to reverse this process to approximate the data distribution. Following the ODE formulation from Karras et al. [6], let $p_{\text{data}}(x)$ be the data distribution and $p(x; \sigma)$ the perturbed distribution with Gaussian noise of standard deviation σ. The probabilistic flow ODE is:

$$dx = -\dot{\sigma}(t)\sigma(t)\nabla_x \log p(x; \sigma(t))dt,$$

*Joint lead authors.
[1]https://github.com/NikeHop/UniversalPolicies/

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

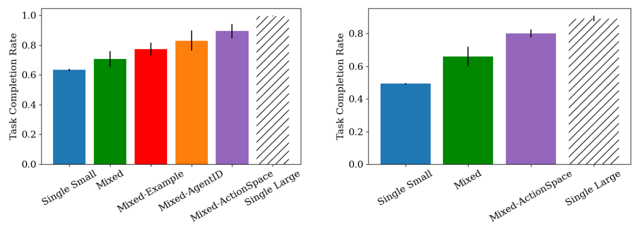

Figure 1: Mean task completion rate of a standard action space agent for various universal policy models in GoToDistractor (left) and GoToDistractorLarge (right). Some variants were not trained on the large environment due to compute limits. Results are averaged over 4 seeds, with error bars showing standard error.

where $x_t \sim p(x_t, \sigma(t))$. The denoising function D_θ is trained via:

$$L(\theta) = \mathbb{E}_{x \sim p_{\text{data}}} \mathbb{E}_{n \sim N(0, \sigma I)} \|D_\theta(x + n; \sigma) - x\|_2^2,$$

with $\nabla_x \log p(x; \sigma) = (D(x; \sigma) - x)/\sigma^2$. At test time, Heun's method is used to generate samples.

Conditioning the Diffusion Model: The denoising network D_θ is conditioned on x_0, instruction c, and agent information k. The instruction is embedded using a T5 variant [9], and x_0 is concatenated along the channel dimension. Classifier-free guidance is not used. Agent conditioning is implemented in three ways:

1. **Agent ID:** A random embedding for each agent type is added to the noise embedding. This does not generalise to unseen agents.
2. **Action Space Representation:** A binary vector $v \in \{0, 1\}^{|A|}$ represents the agent's action space. The vector is embedded and added to the noise embedding.
3. **Example Trajectory:** Example observation sequences illustrating the capability of the agent. Each valid action the agent can take is demonstrated once in the conditioning trajectory.

For each agent $n \in N$, an inverse dynamics model $\text{IVD}_n : X \times X \to A_n$ maps consecutive observations to actions. These models are trained on agent-specific datasets D_n using cross-entropy loss.

3 EXPERIMENTS & RESULTS

The experiments evaluate whether UCAPs exhibit positive transfer, i.e., if training on a pooled dataset leads to higher instruction-following accuracy than single policies trained on agent-specific datasets. We also compare UCAPs to imitation learning baselines adapted to our data setup. Experiments are conducted in the BabyAI environment [1], where agents with varying action spaces (6 in-distribution (ID) agent types for training, 2 out-of distribution (OOD) agent types for testing) navigate gridworlds to objects specified by natural language instructions. We test our method in two BabyAI instances (GoToDistractor, GoToDistractorLarge), differing in size (8x8, 22x22) and number of distractor objects (3, 7).

To test if positive transfer occurs we train a universal policy on the agent specific datasets of different sizes and compare them with UCAP trained on the mixture dataset. The large agent-specific datasets have the same size as the sum of all small agent-specific datasets, so the same size as the mixture datasets. Training on the large agent-specific datasets serves as an upper bound of how much positive transfer we can expect.

Table 1: Average task completion rate of imitation learning (IL) baselines in comparison to UCAP conditioned on an encoding of the action space. Results are averaged over four random seeds and standard errors are in brackets. Bold indicates the best performing model without access to the large agent-specific datasets (SA=Single Agent, CA=Cross Agent).

Model	GoToDistractor-Env	
	ID-Agents	OOD-Agents
IL - SA - Small	0.504(0.006)	0.514(0.018)
IL - CA - Union	0.812(0.005)	0.026(0.002)
IL - CA - Union Finetuned	0.803(0.029)	0.7028(0.031)
IL - CA - AH	0.801(0.018)	0.016(0.005)
IL - CA - AH + Finetuned	0.811(0.037)	0.742(0.044)
UCAP	**0.892(0.053)**	0.541(0.034)
UCAP - Finetuned	0.872(0.046)	**0.904(0.039)**
IL - SA - Large	0.953(0.006)	0.944(0.001)

Figure 1 shows that UCAP exhibits positive transfer, as training on the mixture dataset outperforms training a universal policy only on the small agent-specific dataset in case of the standard action space. Optimal performance for UCAP is achieved when conditioning on the action space representation. The mean performance of the universal policies trained on the small agent-specific datasets is 0.672 ± 0.003 averaged over 4 random seeds and all agent types compared to 0.892 ± 0.053 achieved by UCAP when conditioned on the action space encoding in the GoToDistractor environment.

We compare universal policies to imitation learning (IL) baselines adapted for cross-agent datasets, using a convolutional stack followed by an MLP to predict actions from expert demonstrations. Baselines include standard IL, IL with a union of action spaces, and IL with agent-specific MLP heads but a shared convolutional stack. We additionally compare finetuned versions of both IL baselines and universal cross-agent policies. Both IL variants trained on the cross-agent dataset show positive transfer, but none generalise to OOD agents without fine-tuning (see Table 1). UCAP outperforms IL baselines, in both settings with and without finetuning.

4 CONCLUSION

We showed that a diffusion-based planner operating in the shared observation space of agents, combined with agent-specific inverse dynamics models, effectively learns a universal policy for all agents. UCAP outperforms agent-specific policies and imitation learning baselines. Future work should explore scaling the approach to larger datasets with potentially heterogenous observation spaces [8, 10].

ACKNOWLEDGMENTS

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-6630.

REFERENCES

[1] Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. 2019. BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=rJeXCo0cYX

[2] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samual Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2Web: Towards a Generalist Agent for the Web. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/5950bf290a1570ea401bf98882128160-Abstract-Datasets_and_Benchmarks.html

[3] Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. 2024. Scaling Cross-Embodied Learning: One Policy for Manipulation, Navigation, Locomotion and Aviation. *CoRR* abs/2408.11812 (2024). https://doi.org/10.48550/ARXIV.2408.11812 arXiv:2408.11812

[4] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. 2023. Learning Universal Policies via Text-Guided Video Generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/1d5b9233ad716a43be5c0d3023cb82d0-Abstract-Conference.html

[5] William H. Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. 2019. MineRL: A Large-Scale Dataset of Minecraft Demonstrations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 2442–2448. https://doi.org/10.24963/ijcai.2019/339

[6] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html

[7] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2630–2640. https://doi.org/10.1109/ICCV.2019.00272

[8] Abby O'Neill, Abdul Rehman, and Abhiram Maddukuri et. al. 2024. Open X-Embodiment: Robotic Learning Datasets and RT-X Models : Open X-Embodiment Collaboration. In *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*. IEEE, 6892–6903. https://doi.org/10.1109/ICRA57147.2024.10611477

[9] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. https://jmlr.org/papers/v21/20-074.html

[10] Scott E. Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. 2022. A Generalist Agent. *Trans. Mach. Learn. Res.* 2022 (2022). https://openreview.net/forum?id=1ikK0kHjvj

[11] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *CoRR* abs/2111.02114 (2021). arXiv:2111.02114 https://arxiv.org/abs/2111.02114

[12] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. 2024. Octo: An Open-Source Generalist Robot Policy. *CoRR* abs/2405.12213 (2024). https://doi.org/10.48550/ARXIV.2405.12213 arXiv:2405.12213

[13] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong T. Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA (Proceedings of Machine Learning Research, Vol. 229)*, Jie Tan, Marc Toussaint, and Kourosh Darvish (Eds.). PMLR, 2165–2183. https://proceedings.mlr.press/v229/zitkovich23a.html