In-Context Learning of Sequential Decision-Making Tasks

Roberta Raileanu

GenPlan Workshop NeurIPS 2023

UCL DARK LAB

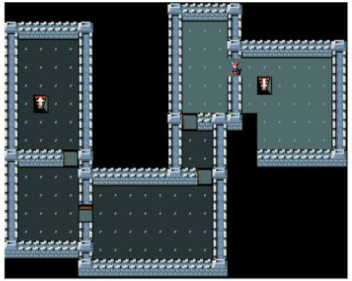
Meta

Can transformers in-context learn sequential decision-making tasks?

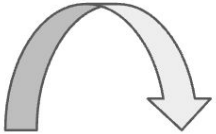
Yes!

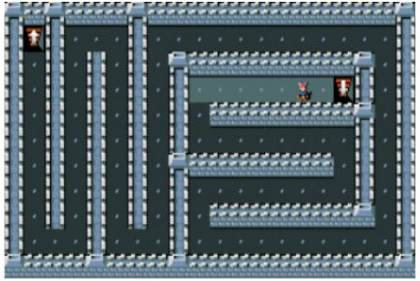
… and No

Our goal: learn new sequential decision-making tasks from handful of demonstrations without any weight updates or environment interactions

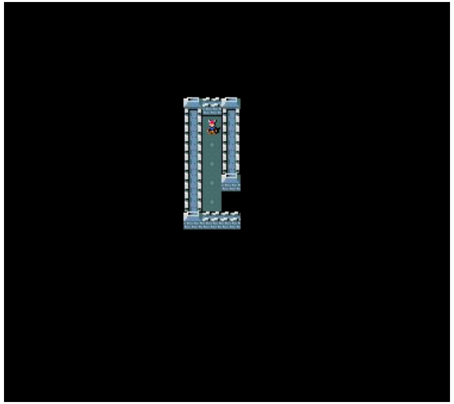


Train Task





Test Task



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

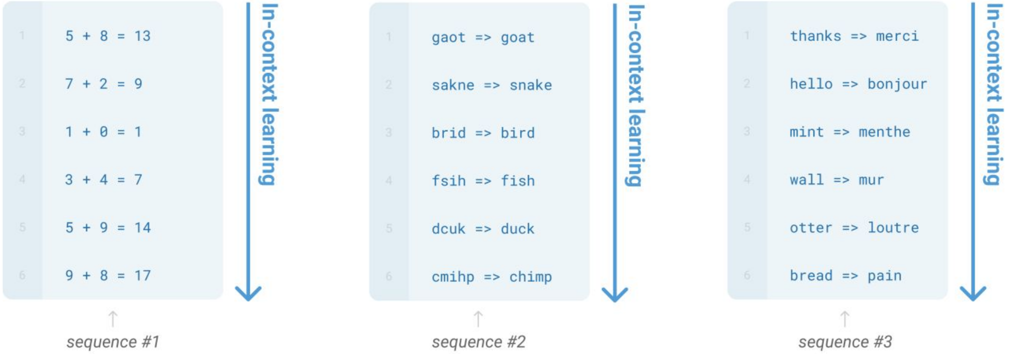
OpenAI

Data Distributional Properties Drive Emergent In-Context Learning in Transformers

Stephanie C.Y. Chan, DeepMind
Adam Santoro, DeepMind
Andrew K. Lampinen, DeepMind
Jane X. Wang, DeepMind
Aaditya K. Singh, University College London
Pierre H. Richemond, DeepMind
James L. McClelland, DeepMind, Stanford University
Felix Hill

MetaICL: Learning to Learn In Context

Sewon Min[1,2] Mike Lewis[2] Luke Zettlemoyer[1,2] Hannaneh Hajishirzi[1,3]
[1]University of Washington [2]Meta AI [3]Allen Institute for AI
{sewon,lsz,hannaneh}@cs.washington.edu mikelewis@fb.com



General-Purpose In-Context Learning by Meta-Learning Transformers

Louis Kirsch[1 2], James Harrison[1], Jascha Sohl-Dickstein[1], Luke Metz[1]
[1]Google Research, Brain Team [2]The Swiss AI Lab IDSIA, USI, SUPSI
louis@idsia.ch, {jamesharrison,jaschasd,lmetz}@google.com

Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

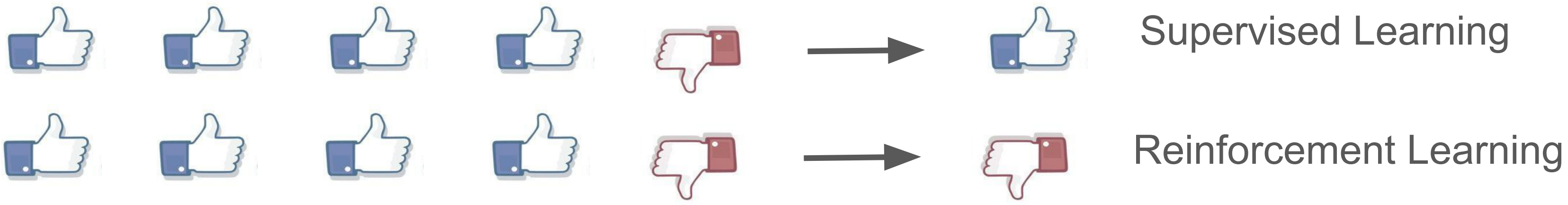
Sewon Min[1,2] Xinxi Lyu[1] Ari Holtzman[1] Mikel Artetxe[2]
Mike Lewis[2] Hannaneh Hajishirzi[1,3] Luke Zettlemoyer[1,2]
[1]University of Washington [2]Meta AI [3]Allen Institute for AI
{sewon,alrope,ahai,hannaneh,lsz}@cs.washington.edu
{artetxe,mikelewis}@meta.com

Unique Challenges of Sequential Decision-Making

- **Distributional Drift** due to the *stochasticity in the environment or the agent's policy*

  → agent needs to generalize to new, potentially out-of-distribution states

- **Unforgiving Environment** with *unrecoverable states* where *a single wrong action can be fatal*
  → agent needs to robustly imitate the expert policy with a high-degree of accuracy



Unlike in supervised or self-supervised learning, taking the right action *most of the time* is **not enough!**

Generalization to New Sequential Decision Making Tasks with In-Context Learning

Sharath Chandra Raparthy[1,*], **Eric Hambro**[1], **Robert Kirk**[1,2], **Mikael Henaff**[1,†], **Roberta Raileanu**[1,2,†]

[1]FAIR at Meta, [2]UCL
[†]Joint advising

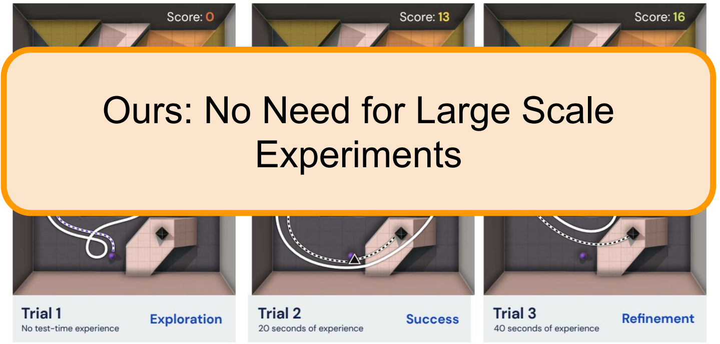
Training autonomous agents that can learn new tasks from only a handful of demonstrations is a long-standing problem in machine learning. Recently, transformers have been shown to learn new language or vision tasks without any weight updates from only a few examples, also referred to as in-context learning. However, the sequential decision making setting poses additional challenges having a lower tolerance for errors since the environment's stochasticity or the agent's actions can lead to unseen, and sometimes unrecoverable, states. In this paper, we use an illustrative example to show that naively applying transformers to sequential decision making problems does not enable in-context learning of new tasks. We then demonstrate how training on sequences of trajectories with certain distributional properties leads to in-context learning of new sequential decision making tasks. We investigate different design choices and find that larger model and dataset sizes, as well as more task diversity, environment stochasticity, and trajectory burstiness, all result in better in-context learning of new out-of-distribution tasks. By training on large diverse offline datasets, our model is able to learn new MiniHack and Procgen tasks without any weight updates from just a handful of demonstrations.

Date: December 8, 2023

Correspondence: Sharath Chandra Raparthy at sharathraparthy@meta.com

Human-Timescale Adaptation in an Open-Ended Task Space

Adaptive Agents Team[1]
[1]DeepMind



Ours: No Need for Large Scale Experiments

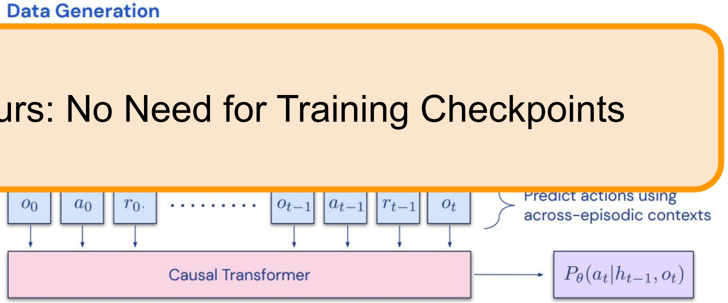
Learning few-shot imitation as cultural transmission

Avishkar Bhoopchand[1], Bethanie Brownfield[1], Adrian Collister[1], Agustin Dal Lago[1], Ashley Edwards[1], Richard Everett[1], Alexandre Fréchette[1], Yanko Gitahy Oliveira[1], Edward Hughes[1], Kory W. Mathewson[1],

Ours: No Need for Vast Computational Resources

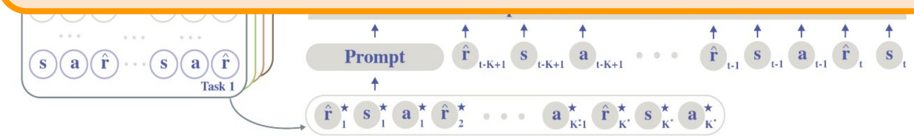
In-context Reinforcement Learning with Algorithm Distillation

Michael Laskin* Luyu Wang* Junhyuk Oh Emilio Parisotto Stephen Spencer
Richie Steigerwald DJ Strouse Steven Hansen Angelos Filos Ethan Brooks
Maxime Gazeau Himanshu Sahni Satinder Singh Volodymyr Mnih
DeepMind



Ours: No Need for Training Checkpoints

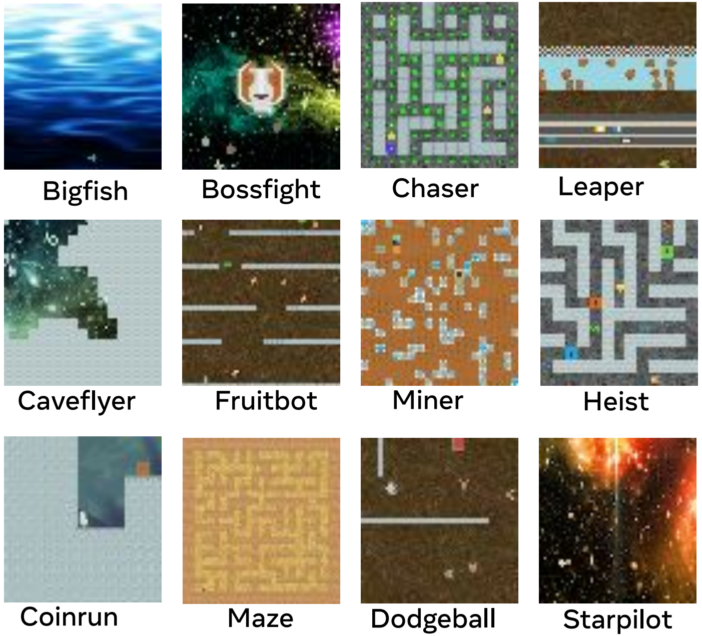
Prompting Decision Transformer for Few-Shot Policy Generalization



Ours: No Need for Rewards
Train and test tasks differ much more

What is In-Context Learning for Sequential Decision-Making?

Train Tasks



In-Context Learning

→

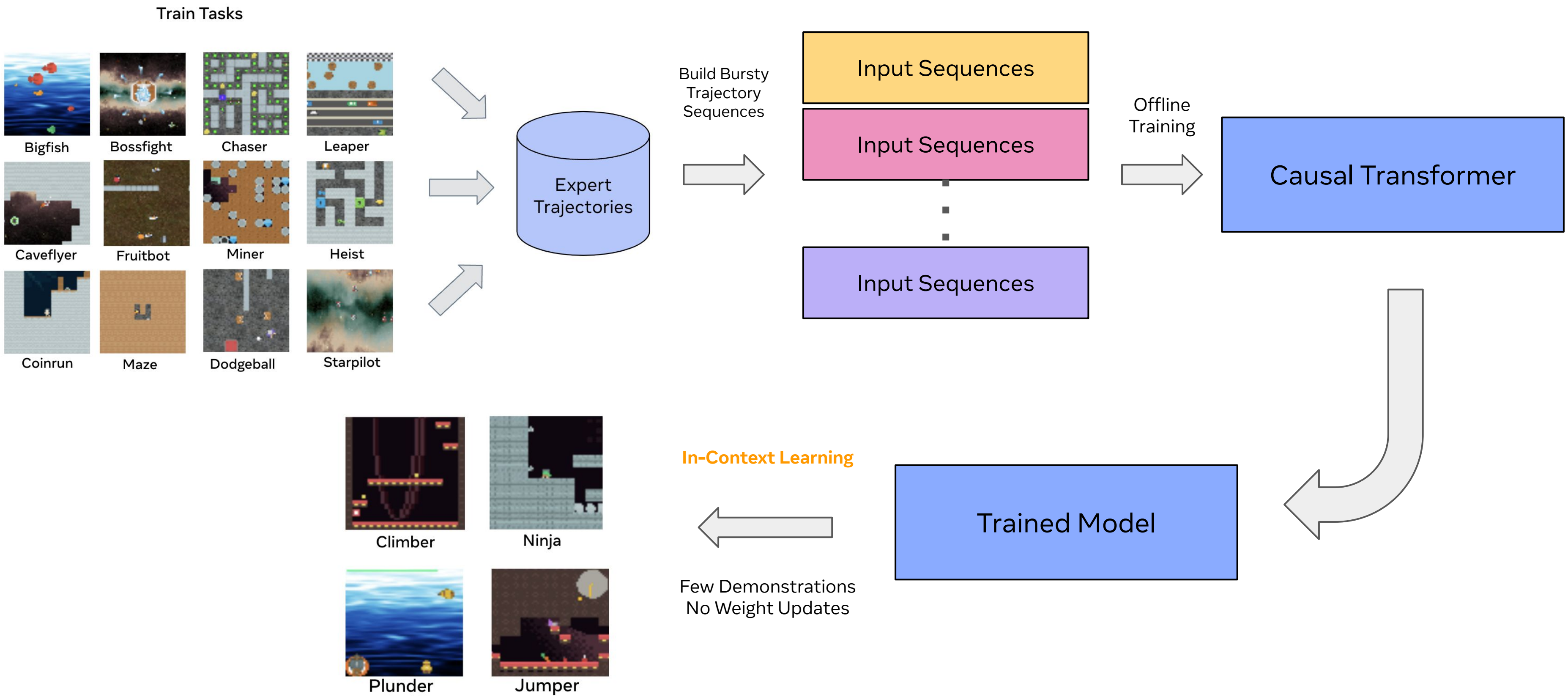
Few Demonstrations
No Weight Updates
No Environment Interactions

Test Tasks

Train and test tasks have *different states, actions, dynamics, and reward functions*

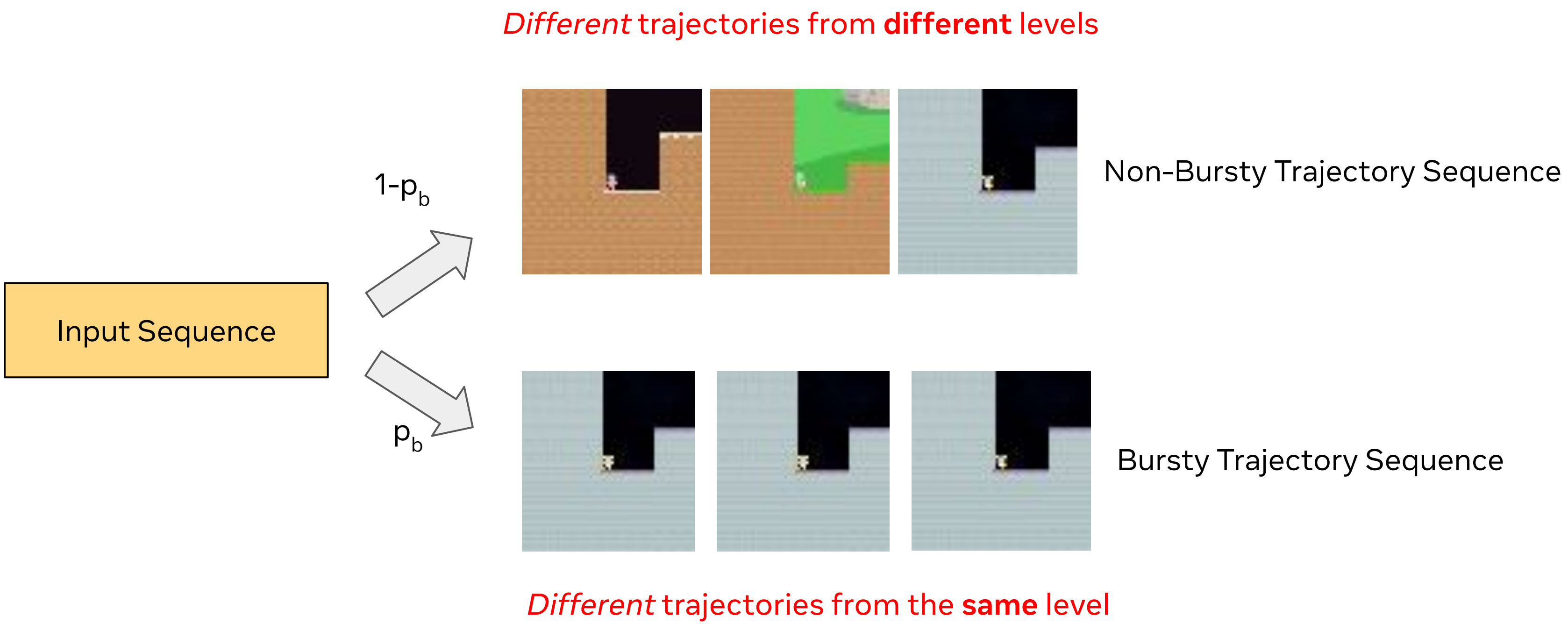
Raparthy et al. 2023

Training Pipeline



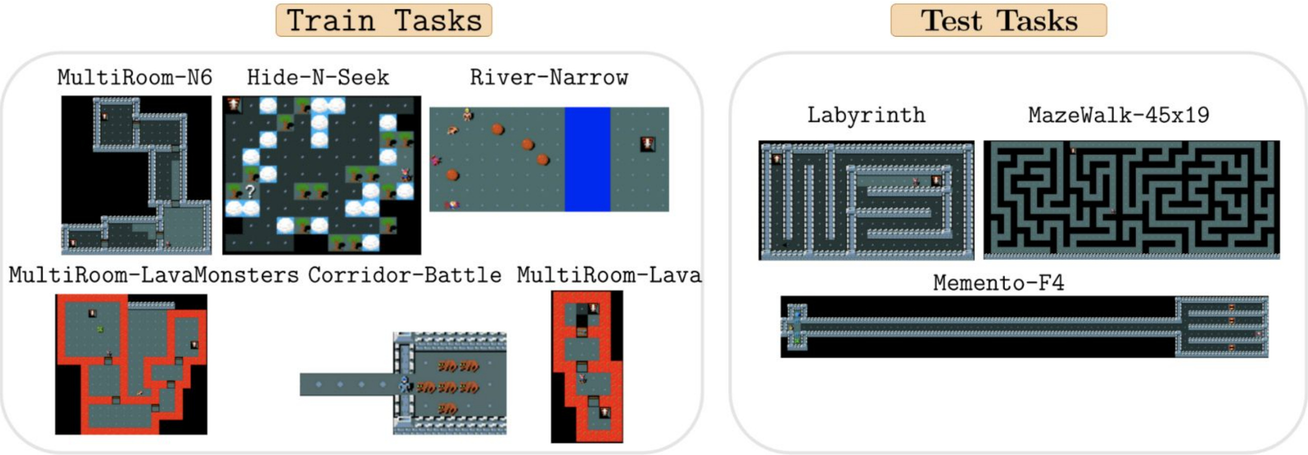
Raparthy et al. 2023

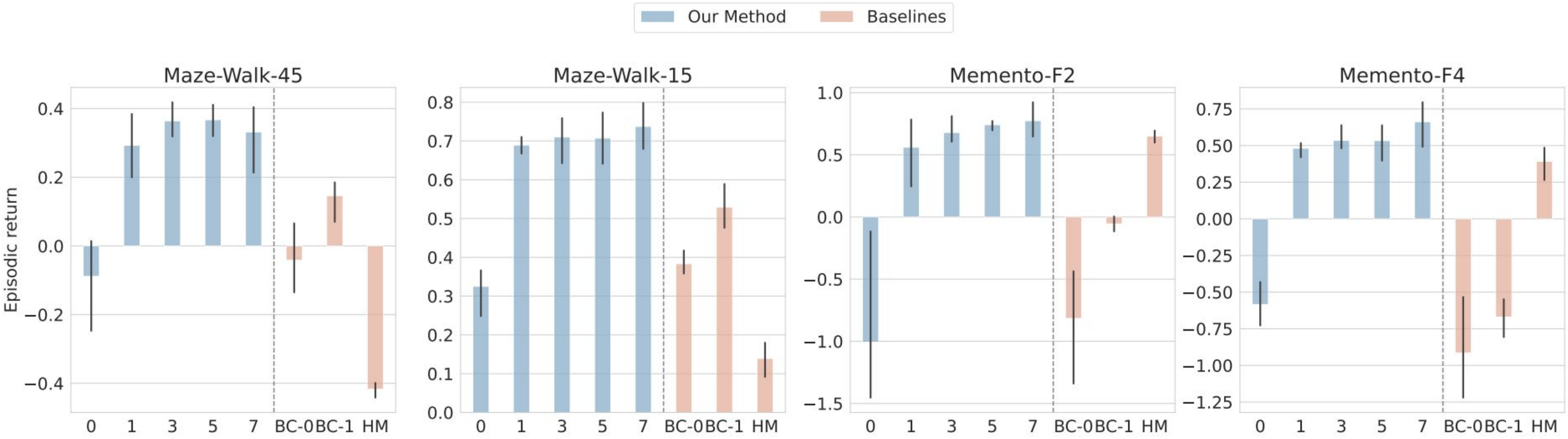
Trajectory Burstiness



Raparthy et al. 2023

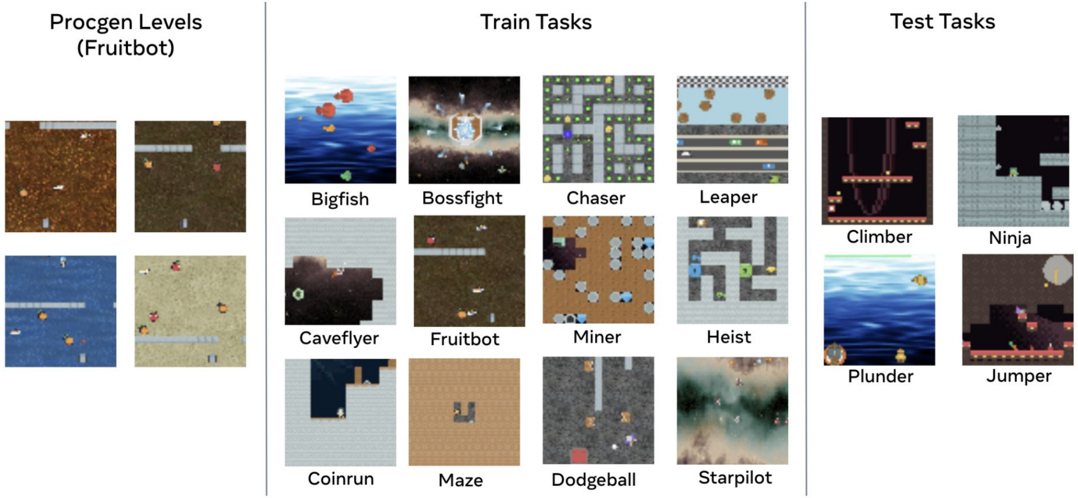
MiniHack Results

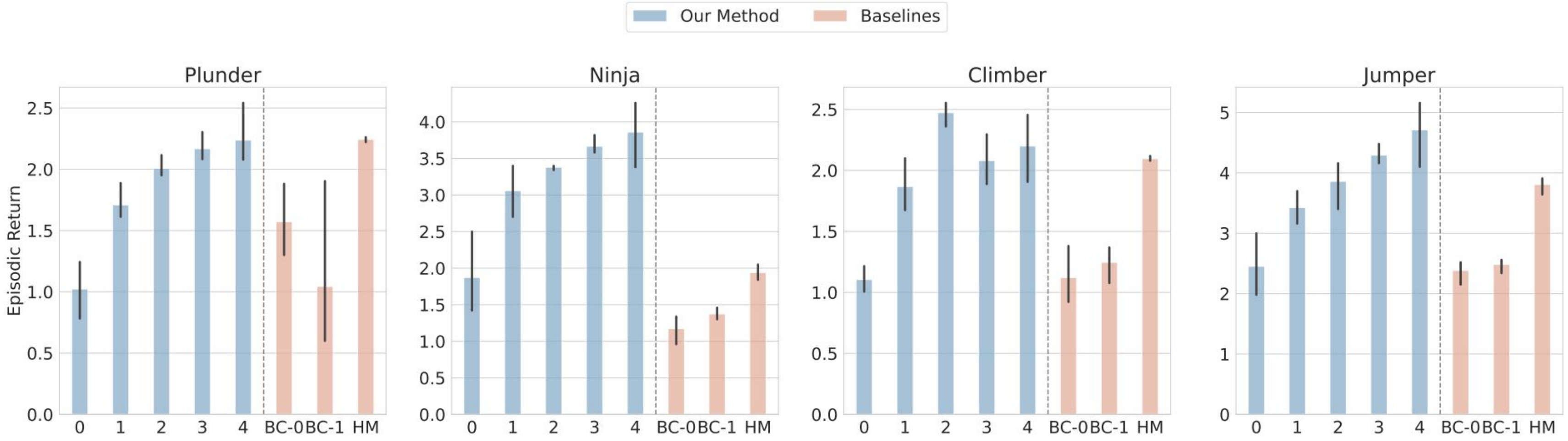




Raparthy et al. 2023

Procgen Results



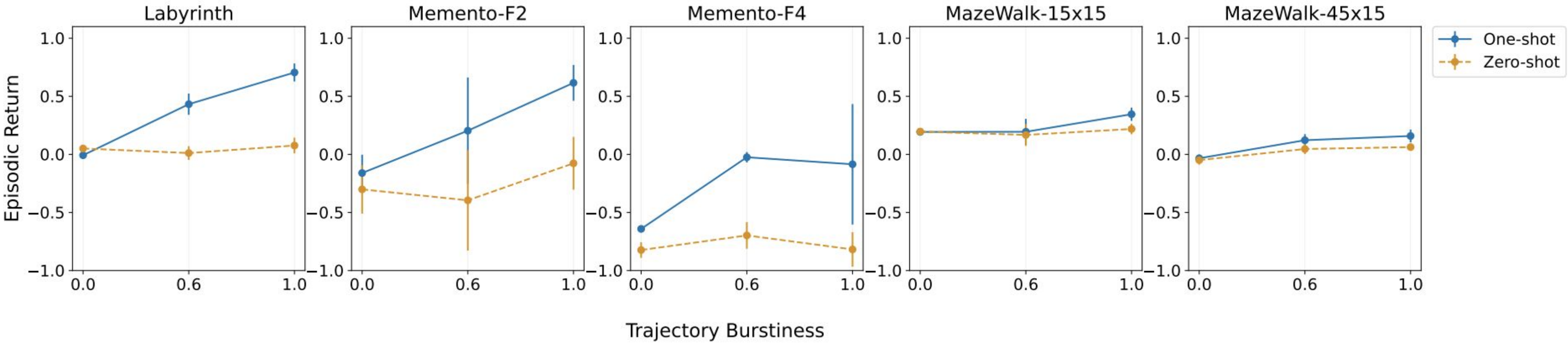


Raparthy et al. 2023

What factors affect ICL and how?

Burstiness, model size, data size, task diversity, stochasticity

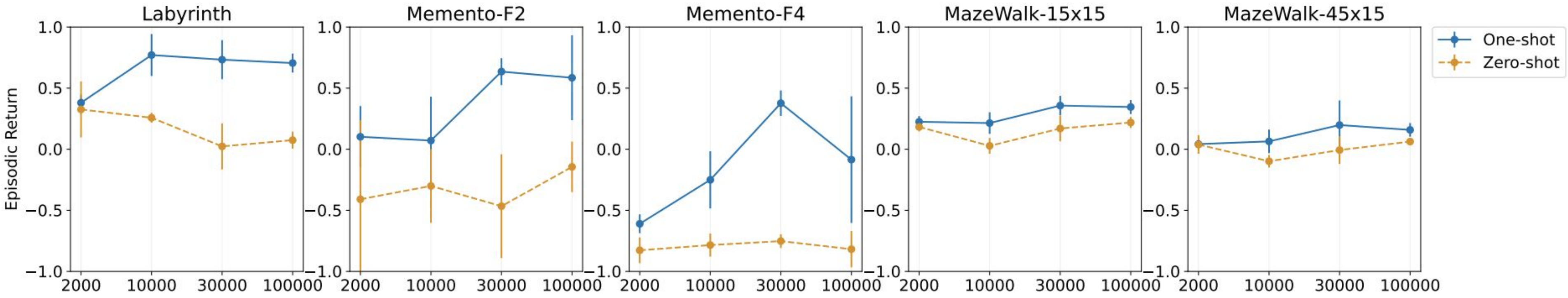
Trajectory Burstiness



In-context learning consistently improves as we increase trajectory burstiness

Raparthy et al. 2023

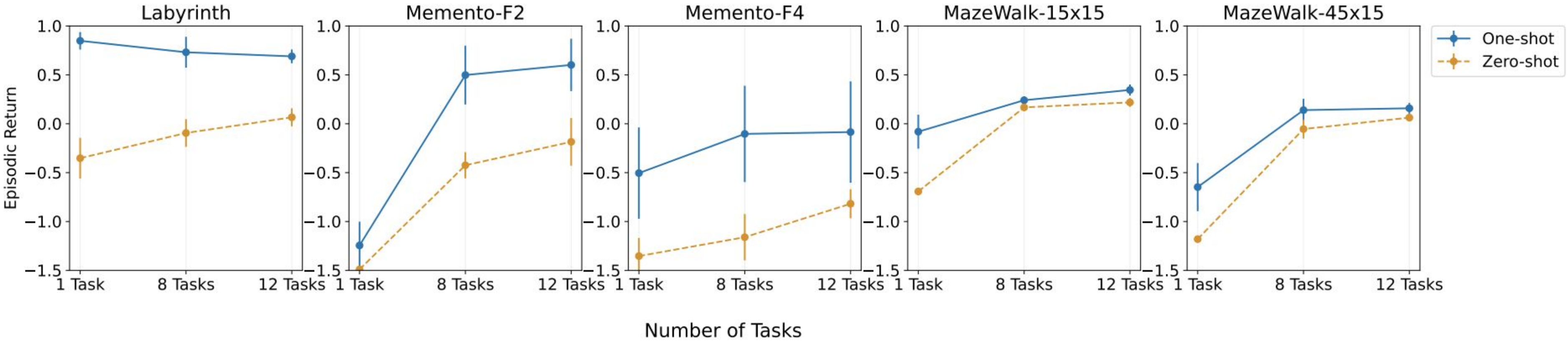
Dataset Size



In-context learning consistently improves as we increase the dataset size

Raparthy et al. 2023

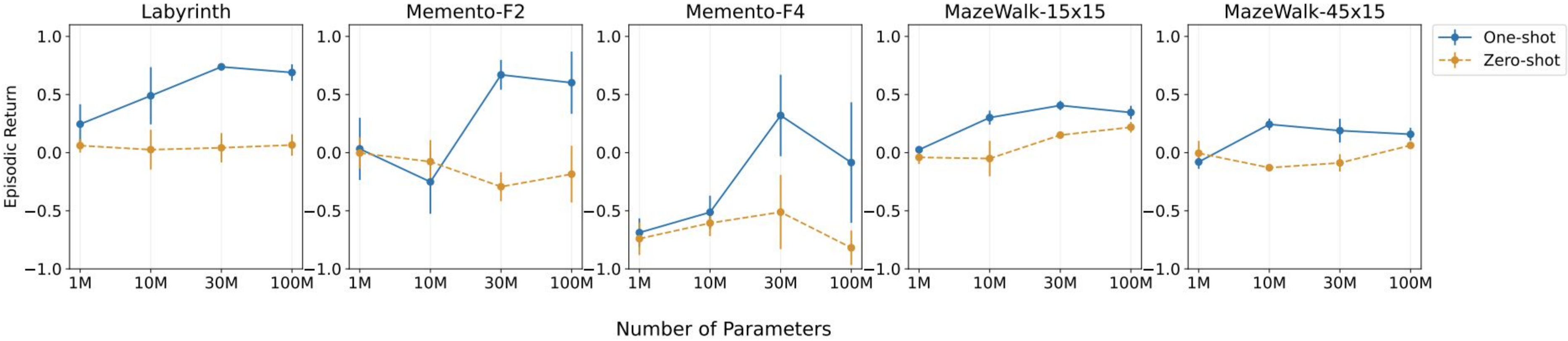
Task Diversity



In-context learning improves with task diversity until it plateaus (here at 12 tasks)

Raparthy et al. 2023

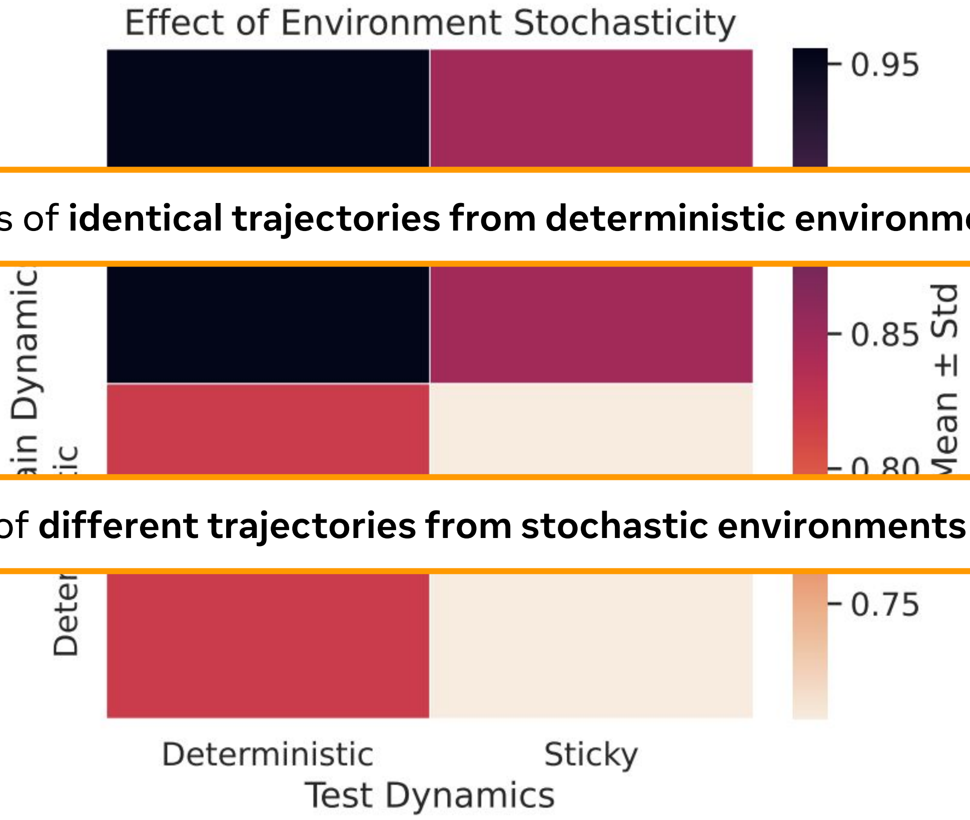
Model Size



In-context learning improves with the model size until it plateaus (here at 30M parameters)

Raparthy et al. 2023

Environment Stochasticity



Training on sequences of **identical trajectories from deterministic environments → brittle policies**

Training on sequences of **different trajectories from stochastic environments → more robust policies**

Raparthy et al. 2023

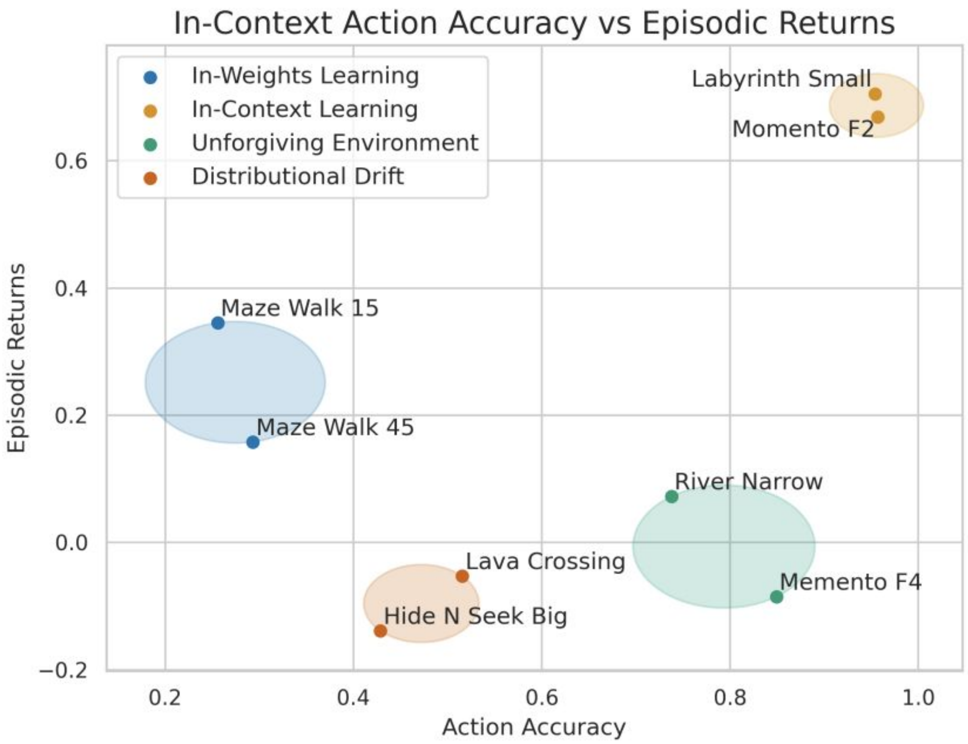
Investigating Failure Modes

1. **In-Context Learning**
   a. High action accuracy and high return

2. **In-Weights Learning**
   a. Training set contains MazeWalk 9 so there is some in-weights generalization to MazeWalk 15 and 45

3. **Unforgiving Environments**
   a. Even if the model imitates the correct actions from its context most of the time, a single mistake is enough for the agent to receive 0 reward (even late in the episode)

4. **Distributional Drift**
   a. The agent drifts away from the context early in the episode and cannot recover since it is OOD for both train and context states



Raparthy et al. 2023

Can Transformers In-Context Learn Sequential Decision-Making Tasks?

Yes, to some extent!

Transformers can in-context learn new sequential decision-making tasks if trained on *bursty sequences of stochastic trajectories, using a large enough model size, and a large and diverse enough dataset.*

But not always!

They still **struggle in vast, stochastic or unforgiving environments** with *distributional drift or unrecoverable states.*

Now What?

- Scale the number of in-context demonstrations to improve performance on “distributional drift” environments
- Uncertainty measures for conservative risk-averse policies in “unforgiving” environments
- Scale to more expressive and open-ended task spaces
- We need an *error correction mechanism to learn what to do in new states and what actions not to take*
- Online learning to the rescue! Combine offline and online learning

How well does offline learning generalize relative to online learning?

The Generalization Gap in Offline Reinforcement Learning

Ishita Mediratta[1,†], **Qingfei You**[1,†], **Minqi Jiang**[1,2], **Roberta Raileanu**[1,2]

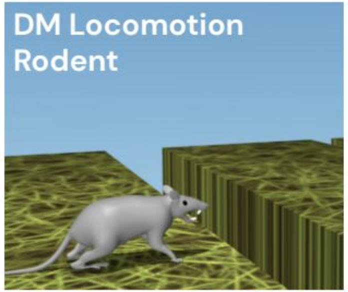
[1]FAIR at Meta, [2]UCL

[†]Joint first author

Despite recent progress in offline learning, these methods are still trained and tested on the same environment. In this paper, we compare the generalization abilities of widely used online and offline learning methods such as online reinforcement learning (RL), offline RL, sequence modeling, and behavioral cloning. Our experiments show that offline learning algorithms perform worse on new environments than online learning ones. We also introduce the first benchmark for evaluating generalization in offline learning, collecting datasets of varying sizes and skill-levels from Procgen (2D video games) and WebShop (e-commerce websites). The datasets contain trajectories for a limited number of game levels or natural language instructions and at test time, the agent has to generalize to new levels or instructions. Our experiments reveal that existing offline learning algorithms struggle to match the performance of online RL on both train and test environments. Behavioral cloning is a strong baseline, outperforming state-of-the-art offline RL and sequence modeling approaches when trained on data from multiple environments and tested on new ones. Finally, we find that increasing the diversity of the data, rather than its size, improves performance on new environments for all offline learning algorithms. Our study demonstrates the limited generalization of current offline learning algorithms highlighting the need for more research in this area.

Date: December 12, 2023
Correspondence: Ishita Mediratta at ishitamed@meta.com
Code: https://github.com/facebookresearch/gen_dgrl

Existing Offline Datasets and Methods

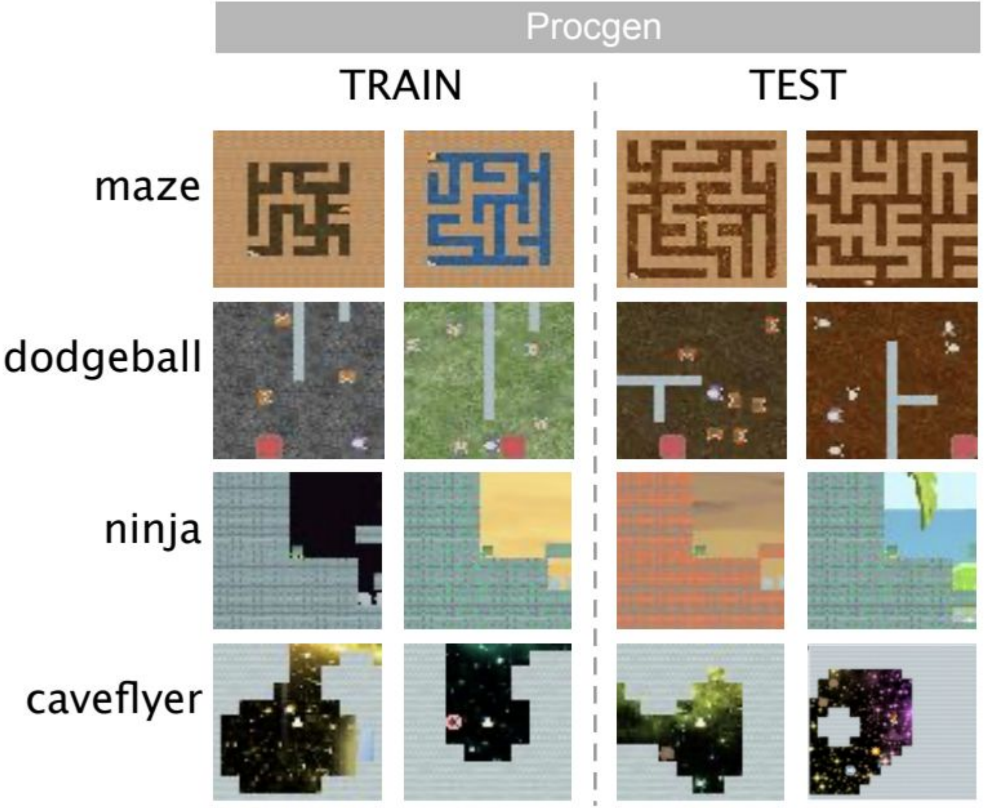
Task domain	 DM Control Suite / Real World RL Suite	 DM Locomotion Humanoid	 DM Locomotion Rodent	 Atari 2600
Action space	continuous	continuous	continuous	discrete
Observation space	state	pixels	pixels	pixels
Exploration difficulty	low to moderate	high	moderate	moderate
Dynamics	deterministic / stochasitic	deterministic	deterministic	stochastic

Singleton Environments

Not suitable for assessing generalization

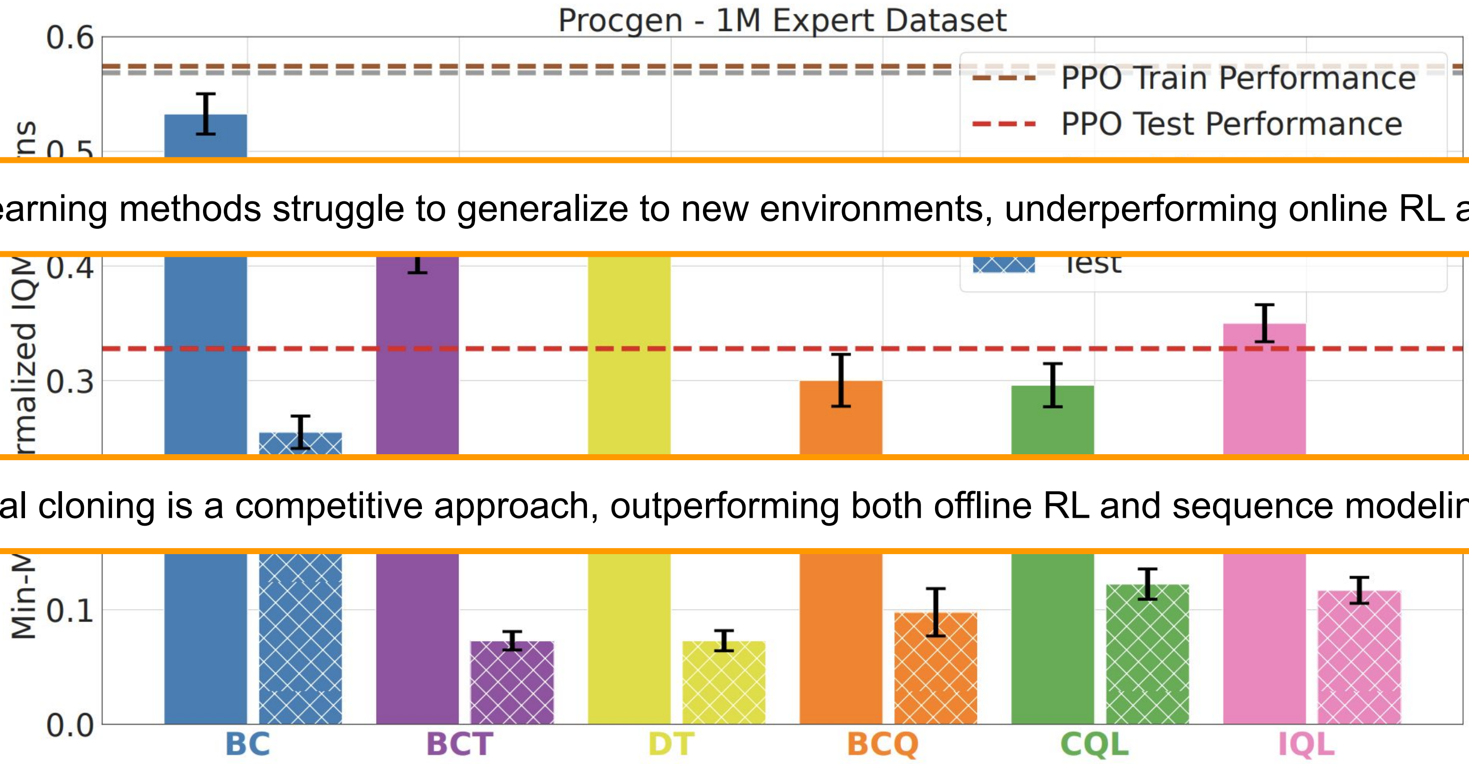
Mediratta et al. 2023

Generalization to New Environments



Mediratta et al. 2023

Generalization to New Environments: Expert Data

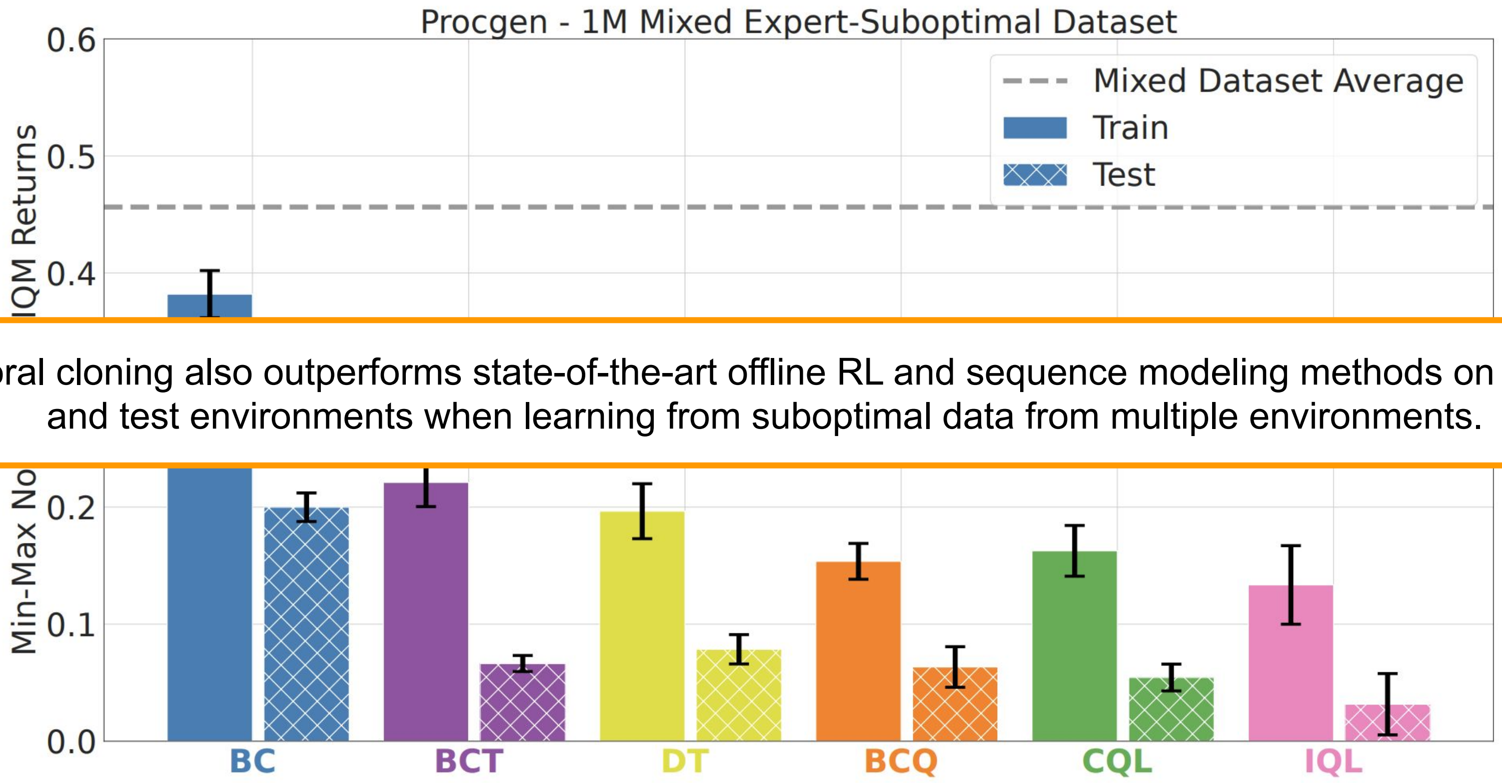


Offline learning methods struggle to generalize to new environments, underperforming online RL at test time

Behavioral cloning is a competitive approach, outperforming both offline RL and sequence modeling methods

Mediratta et al. 2023

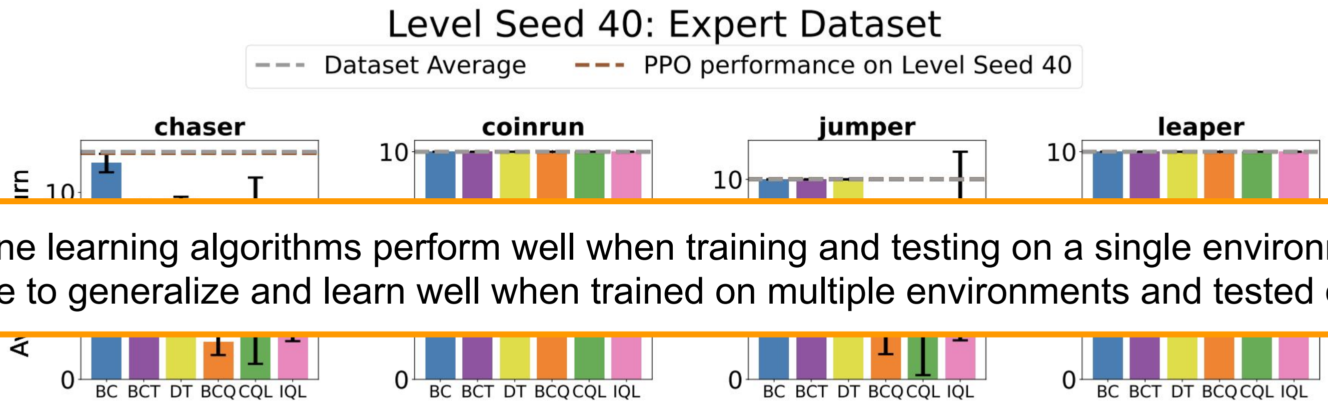
Generalization to New Environments: Mixed Data



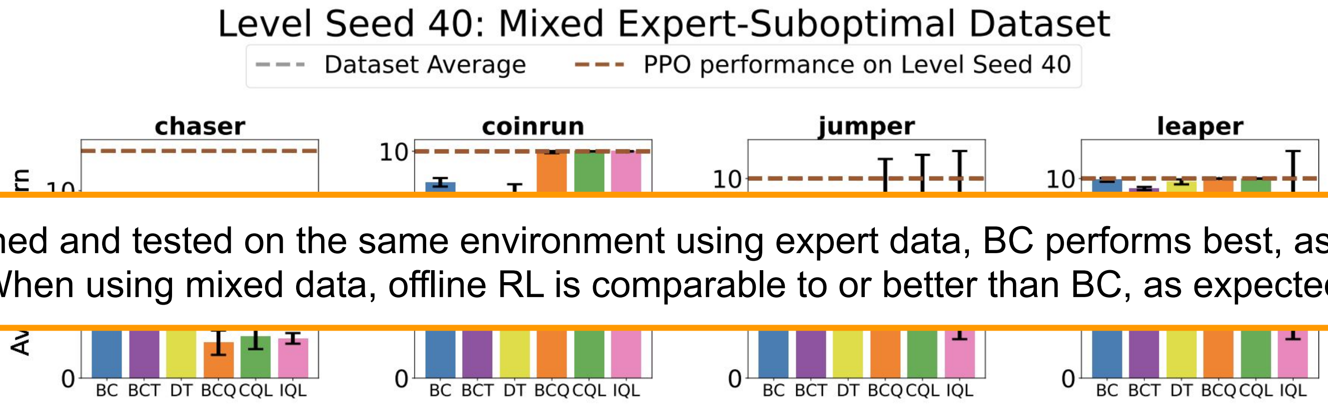
Behavioral cloning also outperforms state-of-the-art offline RL and sequence modeling methods on both train and test environments when learning from suboptimal data from multiple environments.

Mediratta et al. 2023

Training and Testing on a Single Environment



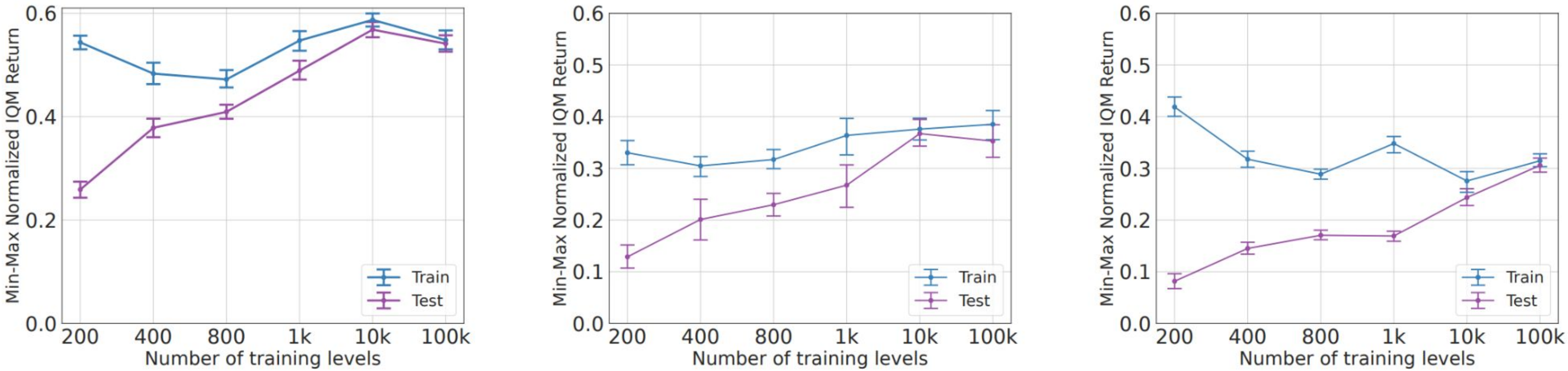
Offline learning algorithms perform well when training and testing on a single environment,
but struggle to generalize and learn well when trained on multiple environments and tested on new ones



When trained and tested on the same environment using expert data, BC performs best, as expected.
When using mixed data, offline RL is comparable to or better than BC, as expected.

Mediratta et al. 2023

The Effect of Data Diversity on Generalization

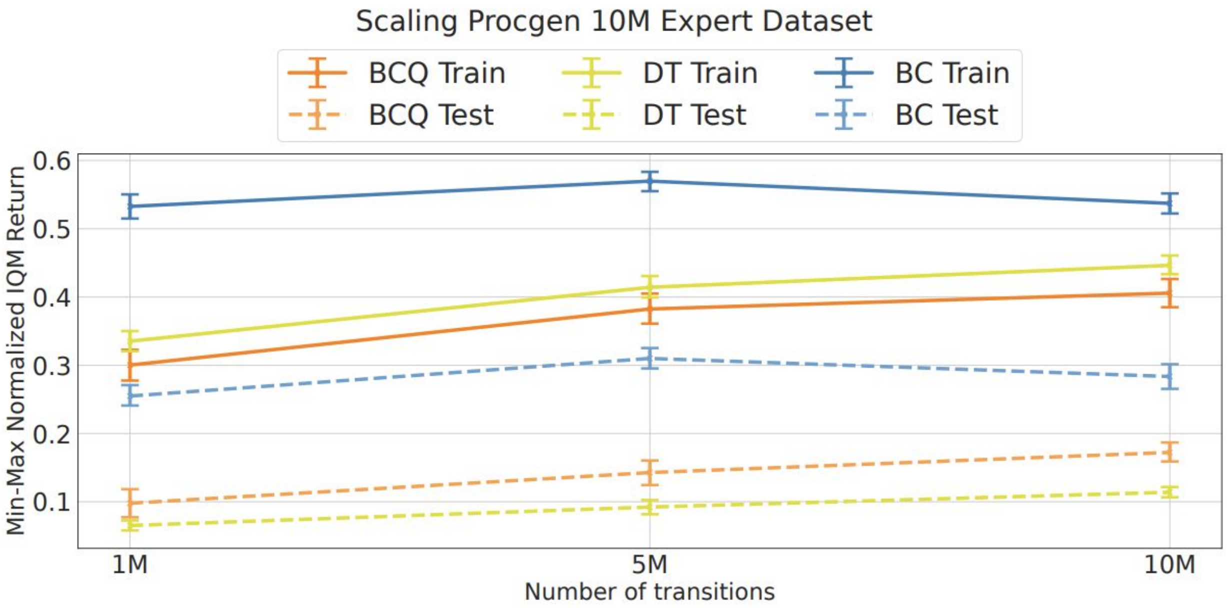


(a) BC **(b)** BCQ **(c)** DT

Increasing the diversity of the dataset
improves performance on new environments for all offline learning algorithms

Mediratta et al. 2023

The Effect of Data Size on Generalization



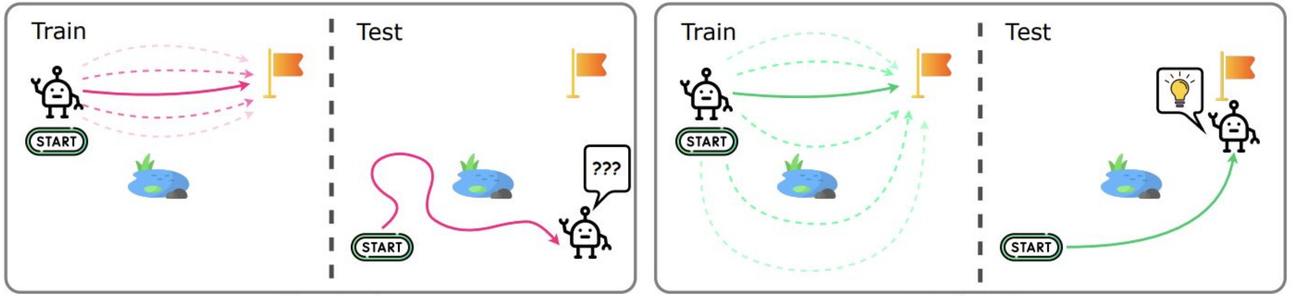
Increasing the size of the dataset without also increasing its diversity doesn't lead to significant improvements on new environments for any of the offline learning algorithms

Mediratta et al. 2023

Online RL > Behavioral Cloning > Offline RL > Sequence Modeling
wrt *zero-shot generalization to new environments*

Why does online generalize better than offline learning?

One hypothesis: the agent learns online from feedback and interaction, collects its own data so it sees a more diverse set of states which can inform what to do in unseen states at test time



Standard Explorative

On the Importance of Exploration for Generalization in Reinforcement Learning

Yiding Jiang*
Carnegie Mellon University
yidingji@cs.cmu.edu

J. Zico Kolter
Carnegie Mellon University
zkolter@cs.cmu.edu

Roberta Raileanu
Meta AI Research
raileanu@meta.com

Explore to Generalize in Zero-Shot RL

Ev Zisselman*, Itai Lavie, Daniel Soudry, Aviv Tamar
Technion – Israel Institute of Technology

Can Transformers In-Context Learn Sequential Decision-Making Tasks?

Yes, to some extent!

Transformers can in-context learn new sequential decision-making tasks if trained on *bursty sequences of stochastic trajectories, using a large enough model size, and a large and diverse enough dataset.*

But not always!

They still **struggle in vast, stochastic or unforgiving environments** where *distributional drift is common or a single bad action can be fatal.*

Thank you!

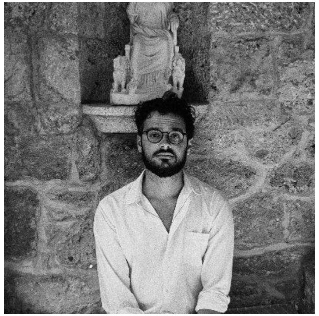
Sharath Raparthy

Ishita Mediratta

Qingfei You

Mikael Henaff

Minqi Jiang



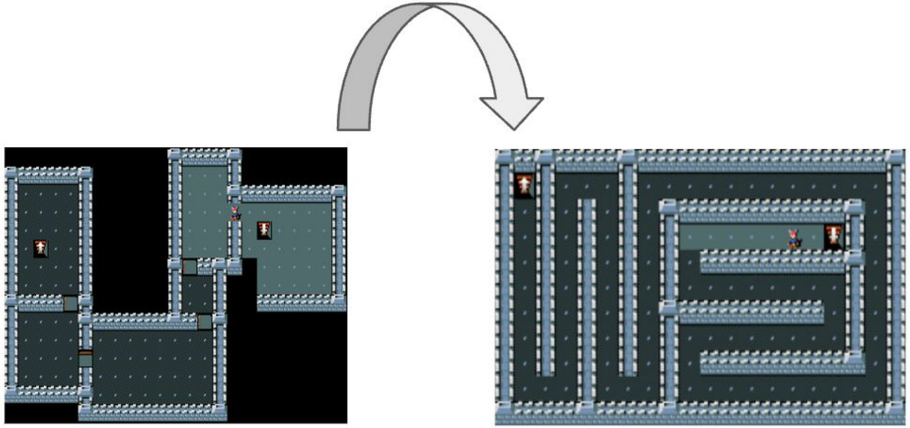
Eric Hambro

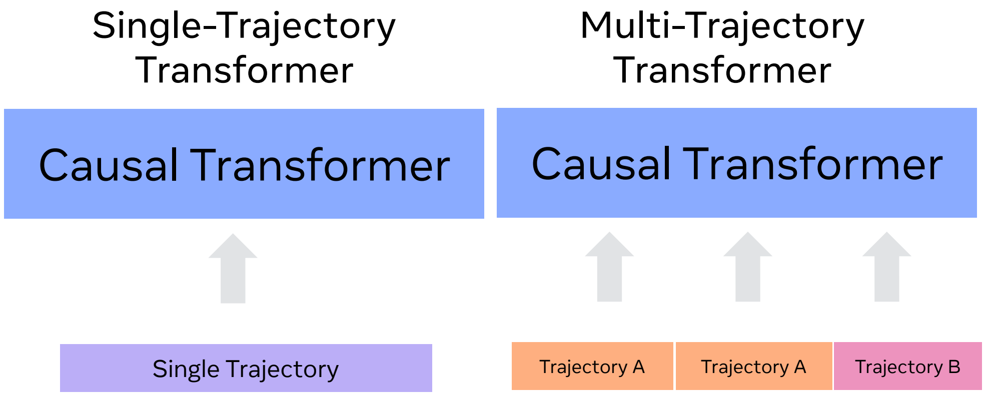
Robert Kirk

Motivational Experiment

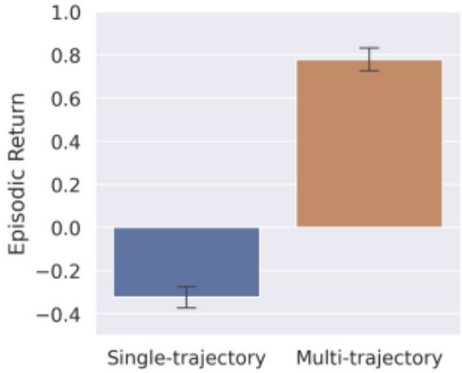
Setup

1. Train on 100k MiniHack-MultiRoom-N6-v0
2. Test on MiniHack-Labyrinth-Small-v0



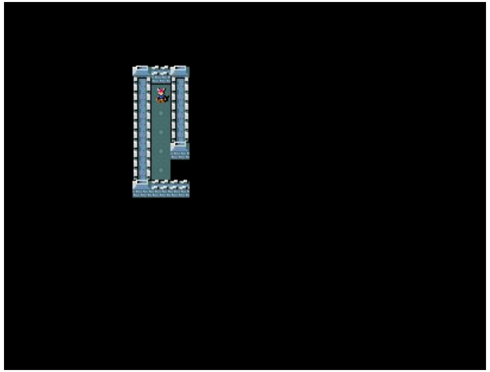


One-Shot Evaluation on Labyrinth

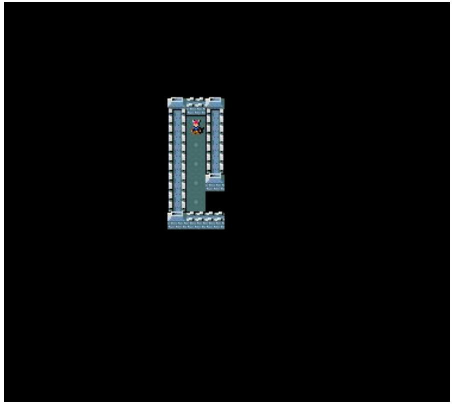


Policy Visualisations

Single-Trajectory Transformer



Multi-Trajectory Transformer



Raparthy et al. 2023