Extended Abstract

Motivation Much of the prior work on policy transfer in reinforcement learning focuses on adapting to different task objectives or rules across similar-sized environments. However, a comparatively underexplored axis of generalization lies in spatial scalability—specifically, whether agents trained on smaller grids can adapt to larger ones under the same task logic. This work explores a setting where the goal and reward structure are fixed, but the grid size—and thus the agent's state distribution and planning complexity—changes substantially. We aim to understand whether effective policy transfer is possible under this form of isolated spatial scaling.

Method We propose a curriculum-guided transfer learning approach to encourage spatial generalization in grid-based reinforcement learning tasks. Agents first follow a curriculum-guided training path across progressively scaling yet still small-sized environments, allowing them to learn a curriculum policy, incrementally build scalable behaviors and value estimates. In the following stage, agent earns improvement in performance when the curriculum policy is transitioned to more complex spatial domains—either through zero-shot deployment of the pretrained policy or through fine-tuning with continued training in the target large-sized environment. We apply this approach to value-based deep reinforcement learning methods—specifically DQN and Dueling DQN—and compare zero-shot and fine-tuning transfer modes against from-scratch training in the large-sized environment.

Implementation We implement the method using MiniHack and evaluate it on three procedurally generated tasks—Basic Navigation, Trap Avoidance, and Maze Traversal—each instantiated at 5×5, 9×9, and 15×15 grid sizes. Observations include symbolic features such as `blstats`, `chars`, and `glyphs`, among others. The input state is formed by combining a 27-dimensional `blstats` vector with a 5×5 local `chars` window, resulting in a 52-dimensional input. Rewards are shaped with sparse proximity bonuses, penalties for steps, traps and wall collisions, and a goal-reaching bonus. Agents are trained on the two smaller environments (5x5 and 9x9) and evaluated on the 15×15 grid via scratch training, zero-shot transfer, or fine-tuning.

Results Curriculum-guided transfer significantly improves policy quality across multiple axes. On the 15×15 Basic Navigation task, fine-tuned agents achieve 100% success with reduced episode lengths and smoother training curves. In Trap Avoidance task, curriculum transfer enables Dueling DQN to improve its success rate from 0.05 (scratch) to 0.32 (fine-tuned), along with a large reward gain. The Maze Traversal task further demonstrates the generalizability of learned priors: both DQN and Dueling DQN fine-tuned agents outperform their scratch counterparts, achieving better convergence and shorter trajectories, suggesting that the method generalizes well across value-based RL models. Qualitative trajectory visualizations reveal that transferred agents follow more coherent paths and exhibit fewer exploratory detours, particularly near high-risk regions.

Discussion While curriculum-guided transfer enhances performance and sample efficiency, several challenges remain open for future exploration. The current curriculum is manually designed based only on grid size; future work may explore more adaptive or task-informed curriculum schedules. Additionally, loss spikes during training suggest that temporal stability can still be improved. Regularization methods, memory-aware exploration, or better credit assignment techniques may further enhance robustness. Finally, the effectiveness of this approach in continuous control or vision-based settings remains an exciting direction for future work.

Conclusion This work introduces and validates a curriculum-based strategy for spatial transfer in DRL. By fixing task semantics and gradually expanding grid sizes, we demonstrate that curriculum-based transfer enables agents to learn reusable behaviors that scale effectively. The approach yields improvements in success rate, convergence speed, and trajectory quality across multiple tasks and architectures. Our results suggest that curriculum-guided training offers a simple yet powerful mechanism for developing RL agents that generalize across spatial complexity, and may serve as a foundation for future extensions involving dynamic curricula, hierarchical methods, and planning-based strategies.

Towards Size-Invariant Policy Learning in Grid Environments via Curriculum-Guided Transfer Reinforcement Learning

Yiling Huang
Department of Electrical Engineering
Stanford University
yilhuang@stanford.edu

Wei Liu
Department of Computer Science
Stanford University
wliu283@stanford.edu

Abstract

Policies trained in small-scale grid environments often fail to generalize to larger domains due to distributional shifts in states, transition dynamics, and sparse rewards. This work explores a curriculum-guided transfer reinforcement learning framework designed to promote spatial generalization by progressively scaling environment size while maintaining fixed game logic. We evaluate Deep Q-Network (DQN) and Dueling DQN agents trained in MiniHack environments across three tasks: basic navigation, trap avoidance, and maze traversal. Agents are progressively trained on 5×5 and 9×9 grids following a curriculum-guided strategy, then transferred to a 15×15 environment via zero-shot or fine-tuned adaptation. Results show that curriculum-based transfer significantly improves success rate, convergence speed, and policy robustness compared to from-scratch training. Our findings highlight environment size as a distinct dimension of complexity and offer curriculum-based pretraining as a promising avenue for developing scalable DRL agents.

1 Introduction

Policies trained for grid-based tasks often fail to generalize when applied to environments of different spatial scales. For example, a warehouse robot (Li et al. (2024)) that learns to navigate in a small testing area may struggle when deployed in a full-sized facility. While the task logic remains unchanged, the larger layout introduces longer planning horizons, more ambiguous observations, and sparser feedback. These challenges stem not from changes in the task itself, but from the increased spatial complexity—posing a fundamental obstacle to policy reuse in deep reinforcement learning. As the environment grows, shifts in the distribution of states and actions can degrade performance, especially when agents are trained from scratch for each new scale.

In standard practice, a new policy is trained from scratch for each grid size. This results in increased training cost and poor reusability of learned behaviors. However, human learners naturally generalize navigation strategies across spatial scales. Inspired by this, we pose the following research question:

> *Can we learn a transferable or reusable policy across spatial scales, enabling efficient and generalizable RL in larger environments?*

To address this, we introduce a curriculum-guided transfer learning framework that progressively trains agents on environments of increasing size. By maintaining fixed task logic and reward structures, we isolate spatial scaling as the key complexity dimension. Our approach consists of two phases: a **curriculum phase** in which agents learn from small to medium grid environments, followed by a **transfer phase**, where we evaluate zero-shot generalization and fine-tuning performance on larger grids.

Stanford CS224R 2025 Final Report

We evaluate this method on three MiniHack-based tasks of increasing difficulty: Basic Navigation, Trap Avoidance, and Maze Traversal. Across all tasks and both DQN variants, we observe that curriculum-trained agents outperform those trained from scratch in terms of success rate, learning efficiency, and behavior stability. Notably, even zero-shot policies—deployed without further training—show consistent improvements over scratch baselines, demonstrating the effectiveness of the learned priors.

This work contributes the following:

- A new framing of environment size as a curriculum dimension in DRL.
- A systematic evaluation of transfer learning strategies across spatial scales under fixed task semantics.
- Empirical evidence that curriculum-guided pretraining improves sample efficiency and robustness of value-based agents in large environments.

In the following sections, we position our work within the existing literature (Section 2), explain the methodology (Section 3), describe the experimental setup (Section 4), present quantitative and qualitative results (Section 5), and conclude with reflections and future directions (Sections 6-8).

2 Related Work

Transfer Learning in Reinforcement Learning Transfer learning in reinforcement learning (RL) aims to leverage knowledge acquired from one task or domain to enhance learning efficiency and generalization in another. Early work, such as Actor-Mimic by Parisotto et al. (2015), demonstrated that a single policy network trained to mimic multiple expert policies across Atari games could generalize across tasks via representation learning. However, such approaches primarily focus on cross-task transfer, where semantics, goals, or action spaces vary. Our work, in contrast, explores intra-task, cross-scale transfer, maintaining constant task logic while varying spatial complexity.

Recent advancements have emphasized architectural generalization and policy distillation. Meta-RL methods like RL^2 by Duan et al. (2016) and PEARL by Rakelly et al. (2019) aim to learn policies that adapt quickly to new tasks. While powerful, these methods often assume task variation and overlook spatial scalability. Our approach is more lightweight and pragmatic: we ask whether pretraining on smaller environments can facilitate learning in larger ones without altering the learning algorithm.

Curriculum Learning in Reinforcement Learning Curriculum learning introduces tasks in a meaningful sequence to ease the agent's learning process. Narvekar et al. (2020) proposed a formal framework for curriculum design, arguing that progression in task difficulty aids policy convergence and robustness. Most curriculum strategies, however, vary reward structures, goal complexity, or auxiliary objectives. Few works consider spatial curriculum as a standalone complexity axis.

Some recent studies such as Tervo (2022) have employed increasing maze difficulty (e.g., more walls, longer paths), but the underlying grid size remains fixed. We explicitly treat grid size as the basis for curriculum-based transfer learning, keeping all other factors constant, and examine how scaling the environment alone impacts policy learning and transfer.

Generalization in Grid-Based Environments MiniGrid by Chevalier-Boisvert et al. (2023) and MiniHack by Samvelyan et al. (2021) provide procedurally generated environments for RL agents to test generalization. Prior work often explores generalization through randomization of obstacles, goal locations, or instructions. For example, some studies assess the impact of architectural priors like convolutional neural networks by Cobbe et al. (2018) or relational modules by Igl et al. (2019) in generalizing across tasks.

While these works offer valuable insights into task-level generalization, they rarely focus on scale-based transfer under fixed logic. In contrast, our work contributes an analysis of how spatial expansion—without semantic change—affects policy transfer and learning efficiency.

3 Method

We propose a curriculum-based transfer learning framework for scalable policy learning. This approach, which we refer to as *Train-Small, Transfer-Large*, is motivated by the insight that agents can acquire useful priors in small environments—where learning is more efficient—that can either directly generalize to larger settings or at least serve as strong initializations for fine-tuning. The framework is designed for tasks where the underlying logic remains constant while spatial complexity scales.

3.1 Framework Overview

The method assumes a family of environments that share consistent action and reward structures but vary in size, layout, and trajectory complexity. Such conditions are common in navigation problems, where expanding spatial scale introduces sparser rewards, longer horizons, and increased uncertainty—factors that often hinder direct training in large environments.

To address this, we adopt a two-phase pipeline:

- **Curriculum Phase:** Agents are first trained across multiple smaller grid sizes of increasing spatial complexity using value-based methods (DQN, Dueling DQN). These settings offer denser feedback and shorter planning horizons, facilitating the learning of transferable skills such as goal-seeking, exploration heuristics and trap avoidance. The curriculum over environment sizes can be manually defined or adaptively selected by the agent, enabling flexible transfer across spatial scales.
- **Transfer Phase:** The policy trained during the curriculum phase is applied in a larger environment through two modes:
  - *Zero-Shot Transfer:* The pretrained policy is directly deployed in the large environment without further training. This setting tests the generalization capacity of the learned representations and behavioral priors.
  - *Fine-Tuning:* The pretrained policy is used as initialization and further trained in the large environment. This evaluates the sample efficiency and adaptability of transferred knowledge.

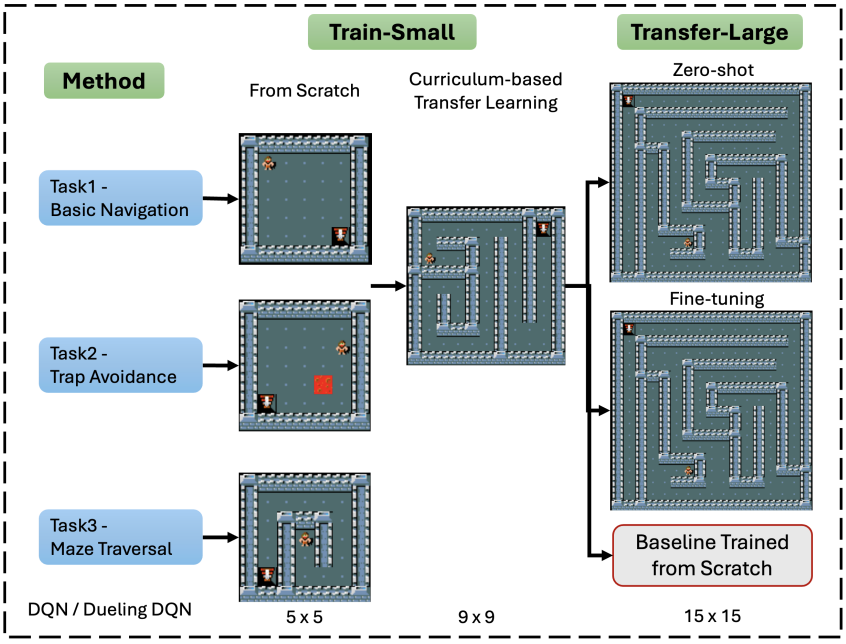


Figure 1: Method and Experimental Setup Overview.

3.2 Design Principles

Our approach is compatible with standard value-based deep RL algorithms. In this work, we adopt both Deep Q-Network (DQN) by Mnih et al. (2013) and its dueling variant dueling DQN by Wang et al. (2016) to demonstrate the method generalizes well across value-based RL models.

DQN. Deep Q-Network (DQN) approximates the optimal action-value function using a neural network and learns it via temporal-difference updates. It incorporates an ϵ-greedy exploration strategy and experience replay, and has served as a foundational algorithm in deep reinforcement learning.

Dueling DQN. Dueling DQN builds on DQN by decomposing the Q-value into two separate estimators: one for the state-value function and another for the advantage of each action. This architecture improves learning stability, especially in environments where many actions result in similar value estimates.

No architectural modifications are required to implement our curriculum-transfer procedure. The method treats spatial scaling as an independent curriculum axis, enabling more efficient and stable policy learning without changing task semantics.

4 Experimental Setup

4.1 Tasks

We evaluate our framework on three procedurally generated MiniHack navigation tasks. All tasks require the agent to reach a staircase, with consistent semantics across grid sizes (5×5, 9×9, 15×15), allowing spatial scaling to be the only complexity factor.

- **Basic Navigation:** The agent starts in the upper-left corner of an empty grid, with the staircase fixed at the bottom-right. The environment is obstacle-free, encouraging direct navigation under sparse rewards.
- **Trap Avoidance:** Agent and goal positions are randomized. Several traps are placed randomly on the grid, penalizing the agent if triggered. The task emphasizes risk-aware exploration and path selection.
- **Maze Traversal:** A solvable maze is procedurally generated with randomized agent and goal positions. The agent must learn to explore efficiently under long horizons and misleading paths.

All task instances are regenerated per episode to ensure generalization across diverse layouts.

4.2 Environment Configuration

We instantiate MiniHack environments under the following protocol:

- **Training:** Agents are trained on 5×5 and 9×9 grids.
- **Evaluation:** Performance is evaluated on 15×15 grids in three settings: zero-shot, fine-tuned, and from-scratch.
- **Observations and State Representation:** At each time step, the agent receives a full observation consisting of symbolic features from MiniHack, including `blstats`, `glyphs`, `chars`, `colors`, and `messages`. From this, we construct a 52-dimensional state representation for learning: a 27-dimensional `blstats` vector and a 5×5 local window of `chars` centered on the agent's position (25 values).
- **Action Space:** Discrete actions: {north, south, east, west, northeast, northwest, southeast, southwest}.
- **Reward Function:**
  - $+10$ for reaching the goal
  - $+3$, $+2$, $+1$ for first-time visits to tiles within Manhattan distances 1, 2, and 3 from the goal

- -0.05 per step
- -0.2 for attempted wall collisions
- -2.0 per time step spent in traps (Trap Avoidance only)

4.3 Baselines and Algorithms

To evaluate the effectiveness of our curriculum-guided transfer approach, we compare three agent configurations:

- **Zero-Shot Transfer:** Policies pretrained on 5×5 and 9×9 grids are directly evaluated in the 15×15 environment without additional training.
- **Fine-Tuned Transfer:** Pretrained policies are further trained in the 15×15 environment, leveraging prior knowledge.
- **Scratch Baseline:** Policies are trained from scratch directly on the 15×15 environment without prior exposure.

All agents are implemented using PyTorch and trained with either standard DQN or Dueling DQN. The networks consist of two fully connected layers with ReLU activations; in the dueling variant, the final layer splits into separate value and advantage streams.

4.4 Evaluation Metrics

We report the following quantitative metrics:

- **Success Rate:** Fraction of episodes in which the goal is reached.
- **Average Reward:** Total reward accumulated per episode.
- **Episode Length:** Average number of steps before termination.
- **Convergence Speed:** Episodes needed to reach plateau performance.
- **Training Stability:** Variance and smoothness of loss and Q-values.

In addition to metrics, we visualize agent trajectories to capture qualitative behaviors such as trap evasion, wall-following, and inefficient exploration loops.

5 Results

We present quantitative and qualitative results on the 15×15 evaluation grid using both DQN and Dueling DQN. Metrics are tracked throughout training and final evaluation to assess convergence behavior, transfer performance, and algorithmic generality.

5.1 Quantitative Evaluation

To investigate training efficiency and final performance, we illustrate representative learning dynamics from the 15×15 Basic Navigation task. Figure 2 and Figure 3 show the training loss and average Q-values, respectively, across 15,000 episodes for both DQN and Dueling DQN, under scratch and fine-tuning regimes. While we focus on Basic Navigation as a case study, similar convergence patterns were observed in the Trap Avoidance and Maze Traversal tasks.

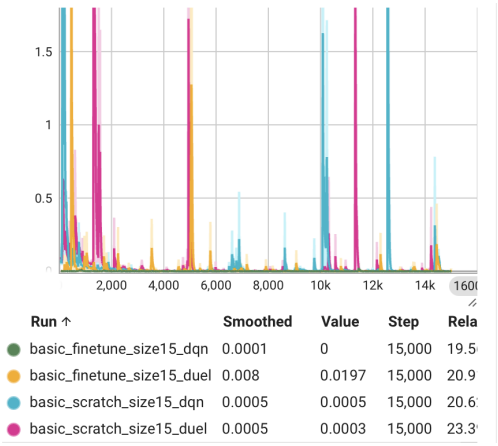


Figure 2: Training loss on Basic Navigation (15×15). Fine-tuned models converge faster and more stably.

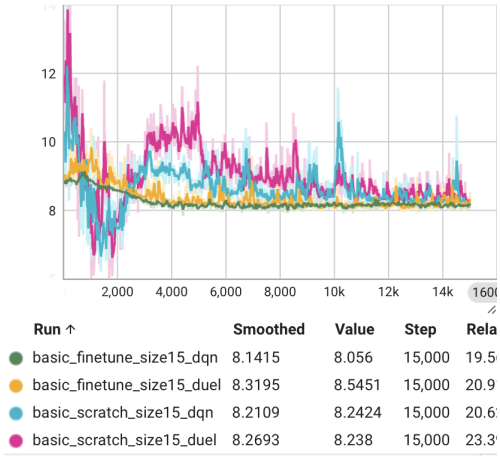


Figure 3: Average Q-value over training. Fine-tuned models maintain consistent value estimates.

We observe that fine-tuned policies consistently converge faster than those trained from scratch. This is most evident in DQN, where the scratch model exhibits high variance and delayed loss reduction, while the fine-tuned variant reaches low loss within a few thousand episodes. Dueling DQN shows more stable dynamics overall, but fine-tuned versions still converge faster and maintain more consistent Q-value trajectories. These results support the claim that curriculum transfer serves as a strong initialization, improving learning efficiency in large-scale settings.

To complement the training analysis, Table 1 summarizes the final evaluation performance across all tasks and methods. Each policy is evaluated on the 15×15 grid using three metrics: success rate, average episode reward, and episode length.

Task / Method	**Success Rate**	**Avg. Reward**	**Avg. Steps**
Basic Navigation (DQN, Scratch)	0.86	-2.67	55.83
Basic Navigation (DQN, Zero-Shot)	1.00	9.03	15.00
Basic Navigation (DQN, Fine-Tune)	1.00	9.03	15.00
Basic Navigation (Dueling DQN, Scratch)	1.00	8.73	22.14
Basic Navigation (Dueling DQN, Zero-Shot)	1.00	9.05	15.00
Basic Navigation (Dueling DQN, Fine-Tune)	1.00	9.07	14.00
Trap Avoidance (DQN, Scratch)	0.19	-16.48	252.01
Trap Avoidance (DQN, Zero-Shot)	0.26	-45.97	230.81
Trap Avoidance (DQN, Fine-Tune)	0.30	-43.28	220.10
Trap Avoidance (Dueling DQN, Scratch)	0.05	-70.93	285.37
Trap Avoidance (Dueling DQN, Zero-Shot)	0.25	-9.44	258.05
Trap Avoidance (Dueling DQN, Fine-Tune)	0.32	-6.33	226.36
Maze Traversal (DQN, Scratch)	0.04	-33.44	288.11
Maze Traversal (DQN, Zero-Shot)	0.16	-14.01	257.04
Maze Traversal (DQN, Fine-Tune)	0.25	-17.32	238.55
Maze Traversal (Dueling DQN, Scratch)	0.02	-74.54	294.02
Maze Traversal (Dueling DQN, Zero-Shot)	0.15	-15.97	266.08
Maze Traversal (Dueling DQN, Fine-Tune)	0.15	-18.12	256.75

Table 1: Evaluation results on the 15×15 grid across tasks and training settings.

Key Results and Analysis The experimental results provide clear evidence that curriculum-guided transfer accelerates learning and enhances final performance. In the Basic Navigation task, fine-tuned agents achieve perfect success rates (1.00) with nearly minimal steps—15.00 for DQN and 14.00 for

Dueling DQN—while scratch-trained counterparts require significantly more steps (55.83 and 22.14, respectively). Training curves further illustrate this advantage: fine-tuned models stabilize both loss and Q-values within 5,000 episodes, whereas scratch policies converge more slowly and with greater volatility.

These benefits are even more pronounced in harder tasks. In Trap Avoidance, fine-tuned Dueling DQN improves success rate from 0.05 to 0.32 and raises average reward from -70.93 to -6.33, demonstrating a strong ability to overcome sparse and punitive feedback. Similarly, in Maze Traversal, curriculum transfer raises DQN's success rate from 0.04 to 0.25 and reduces episode length by nearly 50 steps. These findings highlight that policies trained in smaller environments can internalize spatial heuristics that scale effectively to more complex domains.

These improvements also generalize across architectures. In Maze Traversal, both DQN and Dueling DQN agents benefit similarly from curriculum pretraining, showing that the underlying idea is robust across value-based reinforcement learning algorithms. This cross-architecture consistency underscores the broader applicability of curriculum-based transfer as a method for enhancing policy learning beyond a single model choice.

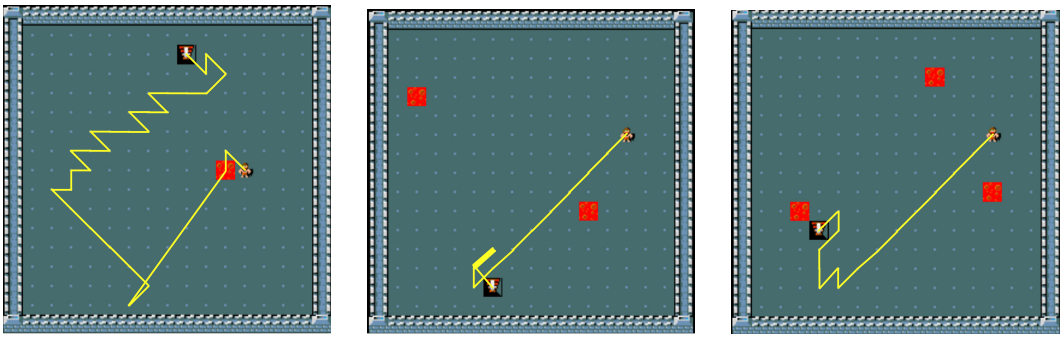
Nonetheless, training remains somewhat unstable. Occasional spikes in loss appear even in fine-tuned runs, indicating that curriculum transfer, while helpful, does not fully address the temporal credit assignment issues in deep RL. Stabilizing training in such environments may require more robust exploration or better regularization.

5.2 Qualitative Analysis

To better understand the behavioral differences between policies, we visualize representative trajectories from the 15×15 Trap Avoidance task using DQN-based agents under three training regimes: from scratch, zero-shot transfer, and fine-tuning (Figures 4a–4c). While the chosen examples all depict successful episodes, they reveal distinct patterns in navigation efficiency, trap awareness, and action smoothness.

The scratch-trained agent (Figure 4a) exhibits erratic movement and frequent course corrections. Its trajectory features a long, inefficient zig-zag path and passes through a trap zone, suggesting poor spatial planning and limited understanding of environmental hazards. Although it eventually reaches the goal, the behavior reflects a reactive rather than strategic policy, likely resulting from unstable training and insufficient exploration.

In contrast, the zero-shot agent (Figure 4b) exhibits a more directed global strategy, following a relatively straight-line trajectory toward the goal while successfully avoiding all traps. However, close inspection reveals a noticeably thickened path segment near the staircase, indicating repeated back-and-forth movement before entering the final tile. This suggests that although the policy generalizes well in terms of large-scale planning and hazard avoidance, it still struggles with local decision certainty in unfamiliar configurations. The behavior reflects partial transfer of useful priors, with residual hesitation arising from the absence of environment-specific fine-tuning.



(a) (b) (c)

Figure 4: Agent trajectories in the 15×15 Trap Avoidance task using DQN: (a) scratch-trained, (b) zero-shot, (c) fine-tuned. All examples are successful episodes; trap tiles are shown in red.

The fine-tuned policy (Figure 4c) further refines this behavior into a more coherent strategy. The agent moves confidently toward the goal with minimal detours and performs precise, trap-aware adjustments near sensitive areas. Unlike the zero-shot case, no redundant movement occurs near the staircase, suggesting that fine-tuning helps resolve remaining ambiguities in local policy execution. This indicates that curriculum pretraining, when combined with limited environment-specific adaptation, yields policies that are both generalizable and behaviorally stable even in high-risk scenarios.

Overall, these visualizations highlight the practical advantages of curriculum-guided transfer learning: it leads to more efficient exploration, stronger generalization, and greater behavioral stability—all essential traits for scaling reinforcement learning to larger and more complex environments.

6 Discussion

While our results highlight the promise of curriculum-guided transfer learning in grid-based reinforcement learning tasks, several aspects present exciting directions for continued exploration. In our experiments, curriculum was implemented through a manually designed progression of environment sizes. While effective, this form of scaling is only one of many possible ways to structure learning. Future work could explore adaptive curricula that respond to the agent's performance or learning dynamics, tailoring task difficulty in a more fine-grained, data-driven manner.

Another promising avenue relates to learning stability. Despite improvements in convergence and final performance, we observed that all training regimes, including fine-tuning, exhibit occasional instability in loss. This suggests that while curriculum provides a valuable initialization, additional mechanisms—such as uncertainty-aware exploration or better regularization—may be needed to fully stabilize learning in sparse-reward or long-horizon tasks.

Beyond methodology, this project highlighted the importance of careful environment and reward design. Subtle choices in reward shaping, grid layout, and transfer protocol had significant effects on training outcomes. Managing these components, especially across varying environment sizes, required a balance of control and realism—underscoring the practical complexity often encountered in deep RL experimentation.

On a broader level, this work reinforces the intuition that leveraging structure across tasks—through well-designed progression—can help scale reinforcement learning to more complex domains. The ability to reuse learned behaviors across tasks is not only computationally efficient, but also aligned with how intelligent systems, both artificial and biological, tend to learn.

7 Conclusion

This work explores curriculum-guided transfer learning as a method for enhancing policy generalization and learning efficiency in value-based reinforcement learning. By progressively training on increasingly complex environments, we enable agents to acquire transferable priors that accelerate convergence and improve final performance across diverse tasks.

Our experiments demonstrate that curriculum-based fine-tuning consistently outperforms training from scratch, both in standard metrics and in policy interpretability. These benefits hold across tasks of varying complexity and across architectures (DQN and Dueling DQN), suggesting that the approach generalizes well.

Looking forward, extending curriculum learning to dynamic, multi-task, or adaptive settings could further improve transferability. We also see potential in combining curriculum methods with recent advances in hierarchical RL, representation learning, and planning-based approaches to better scale reinforcement learning in complex environments.

8 Team Contributions

- **Yiling Huang:** Designed all experimental environments and task configurations, implemented DQN and dueling DQN algorithm, ran experiments for basic navigation task, led poster design and report writing.

- **Wei Liu:** Implemented curriculum and fine-tuning pipelines, ran experiments for trap avoidance and maze traversal tasks, performed trajectory visualization and comparative analysis, co-wrote sections of the final report.
- **Joint Work:** Conceptualized project direction and framing, iterated on environment realism and curriculum strategy, collaboratively reviewed prior work, discussed, revised, and finalized the final report.

Changes from Proposal We ended up making a few meaningful changes from our original plan. Instead of using MiniGrid as proposed, we switched to MiniHack to gain more flexibility in designing environments and customizing observations. While we initially planned to focus only on maze traversal, we added two additional tasks—basic navigation and trap avoidance—to better evaluate generalization across task complexity. For algorithms, we replaced Double DQN with Dueling DQN after early testing showed better training stability. Finally, our original transfer learning idea evolved into a curriculum-guided approach, where policies are trained on progressively larger grids before being applied to the final 15×15 environment.

References

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. 2023. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *Advances in Neural Information Processing Systems* 36 (2023), 73383–73394.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. 2018. Quantifying generalization in reinforcement learning. *arXiv preprint arXiv:1812.02341* (2018).

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. 2016. RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning. *arXiv preprint arXiv:1611.02779* (2016).

Maximilian Igl, Kamil Ciosek, Yingzhen Li, Sebastian Tschiatschek, Cheng Zhang, Sam Devlin, and Katja Hofmann. 2019. Generalization in reinforcement learning with selective noise injection and information bottleneck. In *Advances in Neural Information Processing Systems*, Vol. 32.

Keqin Li, Lipeng Liu, Jiajing Chen, Dezhi Yu, Xiaofan Zhou, Ming Li, Congyu Wang, and Zhao Li. 2024. Research on reinforcement learning based warehouse robot navigation algorithm in complex warehouse layout. In *2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. IEEE, 296–301.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. 2020. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research* 21, 181 (2020), 1–50.

Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. 2015. Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342* (2015).

Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. 2019. Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables. In *International Conference on Machine Learning (ICML)*.

Mikayel Samvelyan, Robert Kirk, Vitaly Kurin, Jack Parker-Holder, Minqi Jiang, Eric Hambro, Fabio Petroni, Heinrich Kuttler, Edward Grefenstette, and Tim Rocktäschel. 2021. MiniHack the Planet: A Sandbox for Open-Ended Reinforcement Learning Research. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. `https://openreview.net/forum?id=skFwlyefkWJ`

Aki Tervo. 2022. *Effects of Curriculum Learning on Maze Exploring DRL Agent Using Unity ML-Agents*. Master's thesis. University of Turku. https://www.utupub.fi/bitstream/10024/154305/1/Gradu%20-%20Aki%20Tervo.pdf

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando De Freitas. 2016. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*. PMLR, 1995–2003.