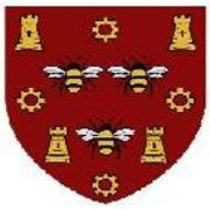

University of Greater Manchester

School of Art and Creative Technologies

**University of Greater Manchester**

# Geometry-Aware Reinforcement Learning for Visual Navigation Agents

Friday Uzochukwu Nkume BSc. (Hons)

Dr. Shukla Shivang

**A Research Paper Submitted in the Partial Fulfilment for the Award of MSc degree in Cloud and Network Security**

**12TH January 2026.**

**University of Greater Manchester, Deane Road, Bolton. BL3 5AB**

# DECLARATION

This is to certify that this thesis titled the “Geometry-Aware Reinforcement Learning for Visual Navigation Agents” is entirely original with no submissions to other educational institutions and the sources utilized in this study have been properly cited and referenced.

# ACKNOWLEDGEMENT

I would like to give thanks to Almighty God because of His guidance, wisdom, and grace during the period of this study.

I would like to thank my supervisor, Dr. Shivang Shukla, because of his priceless assistance, advice, and motivation.

I also wish to thank his Excellency, Rt. Hon. Ogbonnaya Nwifuru, the Governor of the State of Ebonyi, and to the Ebonyi State Scholarship Board for their open-handed generosity and undying devotion to excellence in education.

I also owe a lot of gratitude to the University of Greater Manchester that offered a

conducive environment and resources that facilitated this study.

And lastly, I thank my dearest wife, Ezinne Nkume, for her patient love, enduring love, and encouragement in this process.

# ABSTRACT

Embodied visual navigation poses challenging problems to the reinforcement learning (RL) agents, especially when transferring from familiar training environments to unknown layouts. This work examines whether lightweight geometric inductive priors can enhance sample efficiency, stability, and generalisation in PPO-based navigation in AI2-THOR kitchens. It compared a standard PPO agent with a geometry-aware one that is based on embeddings of low-dimensional geometric state via MultiInputPolicy through a controlled comparative approach. The two models were trained under the same hyperparameters and curriculum timelines. These include single-scene learning on FP1 and multi-scene training of FP1-FP5, and then an out-of-distribution evaluation of FP6-FP8. The geometry-aware PPO showed greater convergence endpoint performance in FP1, with 82.5% success against 62.5% success of the baseline. Geometric priors produced more consistent generalisation in FP1-FP5 with 41.6% success at weighted curricula. The baseline required approximately 38% more training budgets to achieve similar levels of success. There was a small (8.0% vs 4.7%) but positive effect of geometry on out-of-distribution performance, demonstrating some resistance to structural novelty. The results indicate that geometric priors increase stability in optimisation and cross-scene transfer. Catastrophic forgetting is also minimized, but serious difficulties persist in the case of strong distribution shift.

**Keyword: Visual navigation; Geometry priors; Reinforcement learning**

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| Abbreviation | Meaning |
|---|---|
| AI | Artificial Intelligence |
| RL | Reinforcement Learning |
| PPO | Proximal Policy Optimisation |
| CNN | Convolutional Neural Network |
| MLP | Multilayer Perceptron |
| SE(2) | Special Euclidean Group in 2D (rotation + translation symmetry) |
| OOD | Out-of-Distribution |
| FP | FloorPlan (AI2-THOR scene identifier) |
| SR | Success Rate |
| SPL | Success weighted by Path Length |
| RGB | Red-Green-Blue visual input format |
| CNN Encoder | Visual feature extraction network |
| Latent z | Shared latent representation |
| $\pi(a\|o)$ | Policy function (probability of action given observation) |
| V(o) | Value function (state-value estimate) |
| FiLM | Feature-wise Linear Modulation |
| GPU | Graphics Processing Unit |
| SB3 | Stable-Baselines3 |
| RL Policy Head | Action-selection module |
| Value Head | Critic network output |
| VecNorm | Vector Normalisation wrapper |
| Curriculum RL | Progressive multi-scene training schedule |
| Deterministic Evaluation | Evaluation without policy sampling |

# CHAPTER 1

# INTRODUCTION

## Background of Study

Embodied visual navigation deals with the ability of an autonomous agent to perceive, reason, and act in complex three-dimensional environments to achieve spatial goals. Long-horizon decision making under partial observability is required in navigation, which is quite different and more difficult than the normal static perception tasks. Success in navigation depends not only on recognising visual cues but also on preserving coherent spatial representations over time. Here, agents are subjected to continuous integration of sensory input and action outcomes while battling with constraints from occlusions, clutter, and dynamic viewpoints.

The availability of high-fidelity simulation platforms chiefly drives recent advances in embodied artificial intelligence. Environments such as AI2-THOR (Kolve *et al.,* 2017), Gibson (Xia *et al.,* 2018), Matterport3D (Chang *et al.,* 2017), and Habitat (Savva *et al.,* 2019) offer photorealistic indoor scenes, physics-based interaction, and reproducible benchmarks that allow reinforcement learning (RL) agents to be trained and evaluated at scale. It is now feasible to study navigation under controlled conditions, with these platforms in place. This facilitates systematic comparison of algorithms, training regimes, and architectures.

One might be tempted to assume that the problem of visual navigation is fully solved with these advancements, but that is unfortunately far from the reality. The challenge shifted from lack of robust training environments to the inability of agents to transfer what is learned in a training environment to a novel scene. Generalisation remains

one of the most persistent and unresolved challenges in embodied navigation. Evidence of generalisation problem abound in empirical studies, with agents trained on a limited set of environments often suffering severe performance degradation when evaluated on unseen scenes. Common causes of degradation include layout variations, object arrangements, or spatial topology. Reported reductions in success rate or navigation efficiency is typically in the range 30-70% when agents are transferred from “seen” to “unseen” environments (Chaplot *et al.,* 2020; Wani *et al.,* 2020; Wijmans *et al.,* 2019). It is also worthy of note that these failures happen even when training budgets are large and optimisation has converged within the training distribution.

Appearance-driven representation is a key factor causing this brittleness. Majority of the contemporary navigation agents use convolutional neural networks (CNNs) with sole operations based on RGB observations. Such architectures excel at visual feature extractions. However, their lack of explicit geometric structure, spatial continuity, and awareness of orientation causes the agents to be brittle and make wrong decisions. Because of this reason, visually similar states observed from different poses or viewpoints may be mapped to similar internal representations, which then leads to state aliasing and unstable policies. This limitation is much common in indoor environments such as AI2-THOR kitchens, where many objects are visually similar but arranged in layouts that differ meaningfully in geometry and connectivity (Anderson *et al.,* 2018). In AI2-THOR kitchens, there are repeating cabinet textures, many rectangular surfaces, and symmetric layouts. Two completely different parts of the kitchen or even two different kitchens in different floor plans may produce images that look almost identical in pixel space. However, their geometry and connectivity differ. One region may allow direct path to the goal; another may be blocked by furniture.

Movement options are then different even when they look the same. The agent cannot learn this structural difference from RGB alone.

In comparison, there are other areas of robotics and embodied learning which take great advantage of geometric structure to improve the efficiency of agents. Research in manipulation, state estimation, and control has shown that imbedding pose information, symmetry constraints, and SE(2)/SE(3)-aware representations can stabilise learning and improve robustness (Simeonov *et al.,* 2023). Here, inductive biases are used to guide learning toward physically meaningful solutions. However, such geometric priors are less common in visual navigation policies. Priority has often been given to end-to-end visual learning over structured representations.

In the quest for at least partial remedy to generalisation failure, researchers have come up with curriculum learning and multi-scene training. Agents are exposed to diverse environments during training. The weight function is tuned according to the diverse scene complexities. These methods aim to reduce overfitting and encourage transferable policies. Although records of improvements in generalisations exist in literature, recent work indicates that naive multi-scene training can introduce new challenges. These include catastrophic forgetting and instability, specifically when scenes differ significantly in geometry (Kadian *et al.,* 2020; Igl *et al.,* 2019). Consequently, diversity alone does not guarantee robust generalisation.

These limitations motivate a re-examination of architectural inductive biases in embodied navigation. In particular, the integration of lightweight geometric signals, such as agent pose into reinforcement learning policies, offers a promising but underexplored direction. This dissertation investigates whether augmenting a standard PPO navigation agent with pose-based geometric priors can improve training

stability, sample efficiency, and generalisation across structurally diverse indoor environments.

**Problem Statement**

Despite steady progress in embodied AI, RL-based navigation agents are still very prone to suffering from generalisation failure. Proximal Policy Optimisation (PPO) and its distributed versions have gained good performance under training settings; however, their performance tends to break down when tested on new layouts (Wani *et al.,* 2020; Wijmans *et al.,* 2019).

Furthermore, while available literature acknowledged curriculum-based and weighted multi-scene training strategies as viable solutions to mitigating generalisation failures, the way these methods interact with policy architecture is very poorly understood. It is not clear whether improved generalisation results mainly from exposure to a variety of environments, or due to property of the architecture, which allow agents to better exploit spatial structure.

The fundamental issue addressed in this dissertation is therefore twofold. First, there is a need to establish whether incorporating pose-based geometric information into a PPO navigation policy improves sample efficiency, training stability, and cross-scene generalisation under identical training conditions. Second, it is necessary to examine how structured multi-scene curricula interact with architectural inductive biases to influence performance and robustness.

## Research Questions

This study is guided by the following research questions:

1. Does a geometry-aware PPO agent yield better single-scene performance as well as training stability compared to a baseline PPO trained with the same conditions?
2. To what extent do pose-based geometric priors improve generalisation from a single training scene (FP1) to FloorPlans (FP2-FP5) with distinct structures?
3. Can curriculum-based and weighted scene sampling overcome catastrophic forgetting and improve overall multi-scene generalisation?
4. Does a geometry-aware PPO agent have superior out-of-distribution robustness on unseen environments (FP6-FP8)?
5. What are the trade-offs between specialised in-scene optimisation and broad generalisation, and how does the incorporation of geometry affect this balance?

Together, these questions establish a comparative study of the role of geometric inductive biases in reinforcement learning-based navigation.

## Significance of the Study

Achieving a reliable navigation in real indoor environments demands agents that are capable of effective working with a large amount of change in geometry, clutter, and appearance. Nevertheless, current RL-based navigation systems require very large training budget, often ranging in billions of frames, to reach a reasonable performance, and even at that, generalisation is not guaranteed (Wijmans et al., 2022). It is both practical and scientific important that more effective and stronger learning strategies should be sought for.

This study is significant for three reasons. The first one is the evaluation of a computationally lightweight mechanism to enhance robustness. This is done through the integration of explicit geometric information in policy learning. Since pose signals are low-dimensional, and easily obtained within numerous robotic systems, these signals offer a realistic direction to improve navigation performance. And this is done without a significant increase of computational cost.

Second, the study gives empirical evidence on the relationship between architecture and curriculum design. One of the most serious barriers of lifelong and multi-scene learning is the risk of catastrophic forgetting. Through the analysis of the interaction between weighted scene sampling and geometric priors, this work gives a deeper insight into how establishing training regimes affects stability and transfer.

Third, the dissertation fills a gap in the literature of embodied navigation. Geometric priors have been demonstrated to be useful in manipulation and control but have limited systematic evaluation in visual navigation tasks. This study adds evidence-based findings on the benefits and limitations of geometry aware policies by carrying out controlled experiments in FP1-P5 and out-of-distribution scenes FP6-FP8.

**Research Contribution to Knowledge**

The contributions of this dissertation are the following:

1. The architecture and training of a reproducible geometry-aware PPO navigation policy with pose embeddings and feature-level fusion in a vision-based framework.
2. An experimental design that makes a fair comparison of the baseline and geometry-aware agents on single scene optimisation, multi-scene generalisation, and out-of-distribution evaluation by design
3. An organised curriculum-based training regimen that used weighted sampling of scenes to overcome catastrophic forgetting in multi-scene learning.
4. Empirical results that pose-based geometric priors result in better training stability, generalisation, and robustness compared to a strong PPO baseline.

## Aim and Objectives of the Research

### 1.1.1 Aim

The main purpose of this study is to create and carry out an empirical testing on a geometry-aware PPO navigation agent to examine whether geometric priors enhance performance, stability, and generalisation in AI2-THOR environments.

### 1.5.2 Objectives

1. To implement a deterministic PPO baseline for point-goal navigation.
2. To extend the baseline with pose-based geometric features using lightweight embedding and fusion.
3. To test generalisation from FP1 to FP2-FP5 in controlled conditions.
4. To develop and evaluate weighted multi-scene curricula to reduce forgetting.
5. To assess out-of-distribution robustness on new environments FP6-FP8.

## Scope of this Study

The experiment is concerned with point-goal navigation problems in AI2-THOR kitchen settings. Only RGB images and low-dimensional pose vectors are used for observations. Depth, semantic segmentation, and explicit memory modules are not considered to eliminate confusion in the architectural effects. Action space is made of conventional discrete navigation actions. The measurement of generalisation occurs in FP1-FP5 while FP6-FP8 are used to test robustness. To maintain methodological focus, all investigations were restricted to PPO-based methods.

## Thesis Organization

The rest of this dissertation is structured in the following way:

Chapter two is a critical review of the literature about embodied navigation, reinforcement learning, geometric priors, and generalisation. Chapter three involves the methodological framework, research design, algorithms, and evaluation protocols. The plan of implementation and reproducibility are outlined in chapter four. In chapter five, the results and analysis of the experiment are documented. Chapter six concludes the study and recommends future research directions.

# CHAPTER 2

# LITERATURE REVIEW

## Foundations of Embodied Visual Navigation

Embodied visual navigation deals with the issue of empowering autonomous agents to sense, think and behave in three-dimensional environments with the help of egocentric sensory input (Zhu *et al.,* 2021) . Observations made by the agent are in general local and perspective dependent. One can correctly thought of navigation as partially observable Markov decision process (POMDP), where successful behaviour would rely on overcoming perceptual aliasing and having an internal representation of spatial structure over a period.

The availability of high-fidelity simulation platforms such as AI2-THOR (Kolve *et al.,* 2017) and Habitat (Savva *et al.,* 2019), has immense contribution to the progress recorded so far in embodied visual navigation domain. These environments provide photorealistic rendering, realistic physics, diverse indoor layouts, and enable reproducibility of experiments at scale. With these simulators, one can carry out systematic evaluation of reinforcement learning agents across various navigation tasks such as point-goal and object-goal navigation, under controlled conditions. The platforms enable agents to be trained and evaluated both in scale and cross-scene scenarios.

These developments notwithstanding, the issue of generalisation has not been resolved yet. One of the recurring themes among several studies is that the performance of navigation agents that are trained in a small number of environments is high in scenes seen but greatly reduces when required to work in unfamiliar layouts. Available literature reports of significant declines in the success of navigation at the

instances of agents interacting with new spatial arrangements (Chaplot *et al.,* 2020) The fall in performance often lies between 30 and 70 percent in the case of cross-scene evaluation (Wani *et al.,* 2020; Wijmans *et al.,* 2019). These findings suggest that much part of the learned policies is dependent on specific visual regularities of a scene as opposed to the ability to develop transferable spatial reasoning ability.

There are two structural characteristics of embodied navigation tasks that are directly associated with this brittleness. First, the geometric variability of indoor environments is very high even in the same semantic category. An example is that kitchens can have similar items but can substantially differ in terms of topology of the room, connectivity, and geometry. Visual-based agents are hence susceptible to overfitting since a perceptually familiar object can be placed in radically different spatial contexts. Second, navigation is sequential in nature: actions change the state and perspective of the agent and mistakes are multiplied on long horizons. Spatial consistency across time is therefore necessary to ensure good performance especially in cluttered or symmetric world where local observations can be ambiguous. Such difficulties indicate one of the major weaknesses of most modern navigation architectures. The use of conventional convolutional neural network encoders is effective for working with egocentric images but does not provide any specific way to capture geometry, orientation, or the continuity of space. Consequently, policies must implicitly learn spatial structure on raw visual input, making them more sensitive to distributional changes as well as more complicated in sample complexities. The literature hence recommends that representations based entirely on appearances are not enough to support strong navigation.

Therefore, these findings encourage the investigation of the architectural inductive biases that directly support geometry-conscious reasoning. The inclusion of structural

information consistent with the physical characteristics of navigation such as pose or orientation provides a principled way of minimizing perceptual aliasing and increasing cross-scene generalisation. This motivation is the basis of the current research that explores the possibility of lightweight geometric priors to improve the performance of PPO navigation agents in AI2-THOR kitchen environments.

**Reinforcement Learning for Navigation**

The most current dominant method for embodied visual navigation is reinforcement learning. It enables agents to acquire navigation policies through interaction rather than explicit supervision. The most widely used among available algorithms in embodied AI is the Proximal Policy Optimisation (PPO) (Schulman *et al.,* 2017). This is due to its balance of optimisation stability, scalability, and practical implementation simplicity. Using PPO, navigation is formulated as an on-policy optimisation problem. This enables updating of policies using trajectories collected under the current policy while constraining update magnitude through a clipped surrogate objective. The idea behind this design is to reduce destructive policy updates and eliminate gradient variance. This action is very important because long-horizon navigation tasks are characterised by delayed rewards and sparse success signals.

The practical utility of PPO is further supported by the application of generalised advantage estimation (GAE), which averages temporal credit assignment and entropy regularisation, thereby fostering exploration in training. A combination of these mechanisms makes it possible for PPO to succeed in noisy gradients and unstable reward landscapes, both of which are common in visually rich navigation environments.

Distributed implementations have been the most evident way to demonstrate the scalability of PPO, including DD-PPO (Wijmans *et al.,* 2019). DD-PPO showcases the ability of PPO to be trained on millions to billions of frames and synchronise policy updates while maintaining optimisation stability. With these results, PPO is established as a strong baseline and its extensive use in embodied navigation benchmarks is justified. Innovations in architecture in the context of navigation research are thus often assessed through the change of the policy or observation structure under a PPO framework instead of the replacement of the optimisation algorithm.

Using PPO for navigation tasks does not automatically guarantee robust generalisation, the enumerated strengths notwithstanding. Experimental results provide a consistent record of the high levels of success acquired by PPO based agents in training environments but with significant performance decline on unseen scenes. It is not just a failure mode that is caused by instability in the optimisation, instead, it is a manifestation of the constraints of the underlying policy networks. PPO optimises expected return under the training distribution. Thus, it promotes the exploitation of any consistent correlations that exist in the training environments, even those that are scene specific. Appearance and geometry are likely to mix in learned feature representations, when standard CCNs operating on RGB are used to encode observations. Indeed, (Liu, Suganuma and Okatani, 2024) demonstrate that this entanglement has a strong negative impact on transfer when the agents are introduced to environments sharing the same semantic content but varying in spatial layouts. In this case, the PPO agents tend to converge to brittle policies which can only work well within a small range of the distribution of the trajectories being experienced through training.

Hence, what makes PPO a defensible baseline for controlled experiment is that it is strong and stable, yet its performance depends on the quality of the state representation. This implies that improvements that arise from modification of the observation structure or inductive bias can be interpreted as representational effects, rather than optimisation artefacts.

**Generalisation Challenges in Embodied Navigation**

One of the most intractable issues with embodied visual navigation is generalisation. Although reinforcement learning agents can be highly performing in a training environment, their behaviour tends to deteriorate dramatically when subjected to new spatial layouts, object arrangement or viewpoint configurations. Agents often almost show near random behaviour when evaluated in unseen settings even when their performance is strong in in-distribution (Chen *et al.,* 2023; Tatiya *et al.,* 2022; Zheng *et al.,* 2024). Such failure is not by accident but indicates structural characteristics of embodied tasks and constraints of the inductive biases of current models of navigation. One of the key problems is that a great number of agents make use of memorised action patterns or local visual correlations instead of acquiring transferable spatial representations. Within a rich visual indoors setting, texture, lighting, and object appearance may give a very strong and misleading cue during training. Because of this, policies might learn to associate specific visual patterns to actions without learning which spatial relations they made. Performance collapses as soon as these cues are altered even in the slightest.

Such fragility is especially clear in the contexts where semantic similarity leads to concealment of geometric diversity. A good example is that although there are

common objects, between kitchens, there is a great deal of variability in terms of room topology, corridor connectivity and occlusion structure. Even when the categories of objects are familiar, agents that are trained on a small portion of layouts tend not to transfer learned behaviours to novel configurations. This evidence of failure implies that policies reflect appearance-level regularities and not the structure of the scene.

Multi-scene training has been suggested as a solution to overfitting, in the assumption that diverse environments will induce generalisable representations. Nevertheless, naive multi-scene training often adds to the problem by introducing instability, hence it is not necessarily the solution. Catastrophic forgetting often happens when agents are trained in a succession or in multi-environment settings without the use of explicit structural constraints. The newly acquired behaviours overrule the previously developed ones, leading to oscillating performance across scenes or an inability to maintain competence in previous settings (Tatiya *et al.,* 2022). The implication here is that the introduction of diversity without proper strengthening of representation simply trades overfitting with interference.

Recent evidence also points out that scale alone is not a reliable solution. Even training-free navigation methods based on semantic frontiers still struggle in the presence of large layout changes (Chen *et al.,* 2023). The tendency to be sensitive to structural variation is not removed by scaling in generalist navigation models either (Zheng *et al.,* 2024). These results indicate that the failure of generalisation is not a simple outcome of lack of data volume or exploration strategy. It seems more of significant representational deficiencies, a sort of discrepancy between the task requirements and the model inductive structure.

**Alternative Modelling Perspectives Beyond Pure RL Policies**

While we have noted that PPO-based pipelines dominate embodied navigation benchmarks, this section acknowledges alternative modelling perspectives. This helps clarify the aspects of generalisation being addressed in this work. An important distinction is between (i) memory-based or continual-learning approaches that stabilise long-horizon behaviour, (ii) policies capable of implicit inference of structure from ego-centric views, (iii) methods that depend on explicit priors (semantic or spatial) to reduce the learning burden.

One of the alternatives are the training-free methods, such as semantic frontier approaches. These models inject structure through heuristics instead of learning a full policy but still show limitations under large layout shift (Chen *et al.,* 2023). The next is the continual-learning method. Work from this wing highlights that multi-scene competence is not achieved by diversity alone, but also by the ability to preserve previously learned features across non-stationary task distributions (Yang *et al.,* 2025). Lastly, knowledge-driven priors demonstrate that injecting structural assumptions can improve robustness, but there is the problem of scene over-specialisation which limits broad transfer (Jin, Wang and Meng, 2024; Tatiya *et al.,* 2022).

The major takeaway from these perspectives is not to justify that a particular approach is better than others. It is to showcase the multi-casual nature of generalisation failures. Some failure modes come from local aliasing and viewpoint ambiguity, which is a representation issue, while others arise from long-horizon partial observability and interference (memory/continual learning issue). This clarification is very crucial for the present study because it narrows the claim. It is the expectation of this work that geometry-aware conditioning should improve representation and optimisation stability,

but not to fully solve memory-dependent failures because explicit memory or mapping mechanisms are not considered in this study.

Table 2.1. below makes these modelling differences explicit and links them to typical failure modes.

**Table 2.1. Navigation modelling methods and typical generalisation failure modes**

| **Approach** | **Core assumption** | **Typical mechanism** | **Strength** | **Common failure mode under shift** | **Why it matters here** |
|---|---|---|---|---|---|
| PPO + RGB CNN | Pixels contain enough signal if trained long enough | CNN encoder + PPO | Strong in-distribution learning | Shortcut learning; state aliasing; OOD collapse | Baseline controlled comparator |
| Geometry-/symmetry-aware RL | Spatial structure should be encoded, not inferred | SE(2)/SE(3) bias; pose conditioning | Improves stability and efficiency | Bias mismatch; does not solve long-horizon memory | Matches dissertation intervention |
| Curriculum/continual learning | Exposure schedule shapes stability and retention | adaptive/weighted sampling; continual strategies | Reduces forgetting | Still limited by representation | Relevant to FP1–FP5 curricula |
| Training-free semantic heuristics | Explicit structure can replace policy learning | semantic frontier logic | Some zero-shot benefits | Fails under major layout/topology shift | Useful contrast to PPO pipelines |

**Geometry and Pose as Inductive Priors**

The inductive priors are critical determinants of the efficiency of learning systems in acquiring transferable representations. In embodied navigation, where agents act in a partially observable environment with viewpoints that are constantly under transformation, lack of explicit geometric structure can severely hurt learning stability and generalisation. Similar observations emanating from different spatial configurations could be projected to similar feature representations, and this would create ambiguity in subsequent decision-making. Geometry-aware representations are proposed to serve as a solution to this problem. The solution exploits equivariance under transformations in either SE(2) or SE(3) so that any modification in agent pose or orientation is associated with predictable modifications in latent representations**.** Differential invariants have been adopted to formalise this principle, with demonstrations on sample complexity reduction and robustness improvement, both achieved by aligning of learned features with the symmetry structure of the environment (Sangalli *et al.,* 2022).

The Pose information is a complementary source of geometric structure. Pose-aware models achieve drastic reduction of viewpoint ambiguity and stabilise perceptual grounding, by imbedding explicit agent orientation and position. This is not a speculation. It has been demonstrated that symmetry-aware neural architectures incorporating geometric constraints exhibit improved navigation performance and stronger generalisation across spatial layouts (Liu, Suganuma and Okatani, 2024). Instead of sole dependence on visual, such models maintain a consistent spatial reference that supports continuity across time and movement.

Judging from a reinforcement learning perspective, geometric priors have direct effect on optimisation dynamics. As distinct physical states can produce similar observations

in visual navigation environments, state aliasing sets in often. The quality of policy-gradient estimates is reduced by this ambiguity. This is because it introduces noise and bias into advantage computation. Geometry-aware representations correct these states by imbedding spatial context. As a result, gradient signal quality is improved and learning is stabilised. Empirical evidence exists that equivariant representations in embodied agents improve planning reliability and reduce sensitivity to environmental variation, specifically in cluttered or repetitive layouts (Brehmer *et al.,* 2023). Inductive biases help a deep learning system to go beyond pattern matching to relational reasoning (Goyal and Bengio, 2022). This in navigation translates to policies that are less reliant on memorised route and more adaptive policies to new spatial arrangements.

Despite these benefits, geometry-aware models are underexploited in visual navigation relative to related fields like manipulation, state estimation and simultaneous localisation and mapping. Most navigation systems still give preference to appearance-based encoders with an implicit belief that the representational shortcomings will be countered by adequate data and exploration. The above reviewed evidence refutes this assumption and indicates that the absence of explicit geometric structure implies that agents do not readily achieve invariant spatial learning.

Collectively, these findings motivate the integration of lightweight geometric and pose-based priors into reinforcement learning architectures for navigation. Although systematic comparisons remain rare, such priors offer a principled mechanism for reducing state ambiguity, improving optimisation stability, and enhancing generalisation across environments.

Table 2.2 links these priors to the specific failure modes emphasised in this work.

**Table 2.2. Inductive bias mapped to dominant failure modes in embodied navigation**

| Failure mode | Why it occurs in PPO + RGB | Bias/mechanism that targets it | Representative evidence in cited work | Expected implication for this study |
|---|---|---|---|---|
| State aliasing | Similar images can correspond to different physical states | Pose conditioning; equivariant features | Sangalli et al. (2022); Liu et al. (2024) | Faster convergence; improved stability |
| Cross-scene brittleness | Reliance on scene-specific correlations | Symmetry-aware representations | Liu et al. (2024); Wijmans et al. (2019) | Modest transfer gains |
| Catastrophic forgetting | Interference under non-stationary multi-scene training | Weighted curricula; continual strategies | Narvekar et al. (2020); Romac et al. (2021); Yang et al. (2025) | Geometry may help, but curricula still needed |
| OOD collapse | Topology and viewpoint distribution shift | Combined structure + robust training | Anderson et al. (2018); Savva et al. (2019); Cobbe et al. (2020) | Geometry may improve robustness slightly, but limits remain |

**Curriculum Learning and Multi-Scene Training**

Curriculum learning provides a systematic perspective to deal with training instability and improve convergence reliability. This approach has been formalised as a mechanism for guiding exploration, and reducing optimisation variance (Narvekar *et al.,* 2020) . Curriculum learning is of particular importance in embodied navigation because scenes differ in geometry, clutter, and navigational constraints. One of the typical baseline strategies is a homogeneous sampling of the scenes. This approach, though intuitively attractive has been found to be ineffective in PPO-based navigation within environments with high structural diversity.

The limitation is overcome by weighted curricula, which specifically regulates exposure to the scenes. Instead of focusing on all the scenes in uniform manner, sampling probabilities are modified to focus more on environments where the performance is poor or where the learning progress is slowest. Adaptive curricula have been used to stabilize training, by adapting the task difficulty to the developing capabilities of the agent (Romac *et al.,* 2021). In the same manner, curriculum-based training has been demonstrated to enhance convergence and robustness in navigation assignments by making sure early stages of learning are not overwhelmed by challenging environments (Xue *et al.,* 2022). This approach is further extended and formulated as a gradual domain adaptation, where distribution alignment reduces transfer failure (Huang *et al.,* 2022).

A critical unresolved question is how curriculum interacts with representation. The operation of curriculum is at the level of experienced scheduling. It has nothing to do with representational adequacy. If the representation is ambiguous, like in the event when appearance entangled, then curriculum can stabilise training without producing genuine transfer. On the other hand, if geometric priors reduce aliasing and stabilise

learning signals, curricula may become more effective because they are applied to a representation that can exploit structural regularities. The literature provides limited controlled evidence on this interaction. By extension, this motivates the present research design: comparing vision-only PPO and geometry-aware PPO under matched curricula and evaluation.

**Out-of-Distribution Robustness**

The main criterion in the assessment of embodied navigation systems is out-of-distribution (OOD) robustness. This tests the ability of an agent to ascertain whether it learned transferable spatial abstraction rather than memorised correlations. Here, the agents are evaluated in environments that differ from training scenes in geometry, layout topology, and object arrangement, while preserving the same task definition. In both Habitat and AI2-THOR, benchmark experiments always indicate that good in-distribution performance does not correlate with reliable behaviour in unseen settings (Anderson *et al.,* 2018; Savva *et al.,* 2019). In many cases, agents which perform well on observed environments have almost random behaviour on unobserved layouts.

This generalisation gap is caused by a few factors. First, a great number of navigation policies are based on the appearance-related information learned in training settings. The cues fail to work reliably when the texture statistics, lighting conditions or the placement of object changes, even slightly, making the performance deteriorate at a fast rate. Agents trained on a fixed set of indoor scenes have been shown to often overfit to scene-specific visual regularities and fail to adapt to new spatial configurations (Anderson et al., 2018). It was further highlighted that generalisation

failures persist even when agents are trained across multiple environments, a pure indication that exposure alone is insufficient to induce robust spatial reasoning (Savva et al., 2019).

Second, OOD fragility is exacerbated by the fact that the standard convolutional encoders do not exhibit spatial invariance. CNN-based policies process egocentric observations without making any explicit realization of global orientation, relative position or spatial continuity. The direct consequence is state aliasing as noted earlier. In conditions of OOD, the ambiguity is further exaggerated with agents having new configurations that violate learnt correlations between appearance and action outcomes.

These observations are supported by the wider body of reinforcement learning literature. RL agents trained in procedurally generated tasks tend to learn surface-level patterns instead of general strategies, which leads to inability to transfer to the unobserved cases(Cobbe *et al.,* 2020). Another study list catastrophic forgetting and representational interference as consistent impediments to continuous and generalised learning (Parisi *et al.,* 2019).

Proposed potential remedies for these problems include structured training strategies, and architectural inductive biases. While these proposals sound convincing, empirical evidence linking these mechanisms to improved OOD robustness in embodied navigation remains sparse.

OOD assessment is, therefore, not just a useful benchmark but a diagnostic instrument of critical importance in determining whether agent navigators develop transferable spatial abstractions. The severe failure mode scenarios in previous studies point to the necessity of controlled experimental designs specifically testing

robustness to structural novelty. This is the reason as to why FP6-FP8 assessments have been incorporated into the current study, which furnishes a principled foundation to consider the impact of architectural decisions and training plans on generalisation outside of the training distribution.

**Related Works**

To synthesise the above thematic review, Table 2.3 provides a summary of representative recent work in the fields of reinforcement learning, embodied navigation, geometry-aware modelling, and curriculum learning. The table compares benchmarks, methodological approaches, strengths, and limitations and gives a relevance score based on the direct contribution of each study to the research questions in this dissertation.

**Table 2.3. Recent Publications on RL, Embodied Navigation, Geometry-Aware Models and Curriculum Learning**

| Reference | Category | Benchmark / Env | Method / Model Type | Strength | Limitation | Knowledge Contribution | Relevance (1–5) |
|---|---|---|---|---|---|---|---|
| Schulman et al. (2017) | RL algorithm | Atari, Mujoco | PPO | Stable, simple, strong baseline | Not generalisation-oriented | Establishes PPO as standard | 5 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Wijmans et al. (2019) | Embodied navigation | Habitat | DD-PPO | Near-perfect training success, scalable | Strong overfitting | Data-scale not equal to generalisation | 5 |
| Chaplot et al. (2020) | Embodied navigation | Gibson / MP3D | Semantic exploration plus RL | Leverages spatial semantics | Domain-shift sensitive | Shows structure improves navigation | 4 |
| Savva et al. (2019) | Benchmark | Habitat | Benchmark plus SPL | Standardised metrics and evaluation | Early tasks biased | Benchmark foundations | 5 |
| Liu et al. (2024) | Geometry-aware navigation | Habitat | SE(2)-aware CNN | Better generalisation via symmetry | Complexity | Structure improves navigation | 5 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Liu et al. (2021) | Geometry-aware navigation | Habitat | Early SE(2) model | Good symmetry intuition | Pre-journal | Inductive bias evidence | 4 |
| Sangalli et al. (2022) | Equivariant CNN theory | Synthetic | SE(2) invariants | Strong theory | Not navigation-specific | Formal geometric basis | 4 |
| Brehmer et al. (2023) | Equivariant planning | Grid-world | Equivariant diffusion | Sample-efficient planning | Small-scale tasks | Geometry improves planning | 4 |
| Goyal (2022) | Inductive bias theory | Conceptual | Structural priors | Strong theoretical grounding | No navigation experiments | Bias improves learning | 3 |
| Tatiya et al. (2022) | Scene priors | AV-navigation | Knowledge-driven priors | Improves robustness | Domain-specific | Priors help generalisation | 3 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Chen et al. (2023) | Zero-shot navigation | Habitat | Semantic frontiers | Good zero-shot performance | Needs semantics | Structure plus heuristic | 3 |
| Zheng et al. (2024) | Generalist navigation | Multi-environment | Generalist model | Multi-task capability | Still gaps | Reinforces generalisation issues | 4 |
| Yu et al. (2025) | Continual navigation | Multi-environment | C-NAV | Handles forgetting | Early-stage | Insights on continual navigation | 4 |
| Narvekar et al. (2020) | Curriculum RL survey | Various RL | Curriculum frameworks | Good taxonomy | Not navigation-specific | Shows value of curricula | 4 |
| Romac et al. (2021) | Curriculum RL bench | ICML | TeachMyAgent | Adaptive curricula | Simple environments | Empirical curriculum support | 4 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Xue et al. (2022) | Navigation with curricula | Intralogistics | Mapless navigation plus curriculum | Demonstrates navigation gains | Domain-specific | Curriculum improves robustness | 4 |
| Huang et al. (2022) | Curriculum RL | RL control | Optimal-transport curriculum | Strong theory | Requires task distribution | Domain adaptation logic | 3 |

Table 2.1 demonstrates that there is an evident imbalance in the literature. Core algorithms and benchmarks (Schulman et al., 2017; Savva et al., 2019) remain foundational for embodied navigation evaluation but offer limited mechanisms for generalisation. The fact that larger versions of PPO do not overcome structural brittleness (Wijmans et al., 2019) supports the idea that scale does not guarantee robust navigation

Curriculum-based and geometry-aware research offer promising but disjointed insights. Symmetry-aware models can improve transfer (Liu et al., 2024; Sangalli et al., 2022). Curriculum learning frameworks can stabilise training (Narvekar et al., 2020; Romac et al., 2021). A critical gap in research is scarcity of work isolate geometry effects within PPO under unified curricula and scene/unseen evaluation. This gap is a direct motivation behind the controlled design adopted in this dissertation.

**Debates and Gaps in Current Research**

The reviewed literature demonstrates that there are still debates that surround how embedded navigation agents would be designed to realise robust generalisation. There is a tension between scale-oriented methods, which emphasises large data and minimal architectural conditioning, and structure-oriented solutions, which hold that explicit inductive biases based on geometry, symmetry, and spatial queries are more important. Although large-scale reinforcement learning systems like DD-PPO achieve impressive in-distribution performance when trained on large amounts of data (Wijmans et al., 2019), these improvements are not always replicated in the form of distributional shift robustness. This implies that scale is not enough to deal with the structural complexity of embodied navigation.

The first significant gap relates to the fact that there has been minimal incorporation of geometric inductive priors into the PPO-based navigation systems. Though symmetry-aware and pose-aware models hold theoretical and empirical advantages in perception, planning, as well as manipulation (Sangalli et al., 2022; Liu et al., 2024), their application in visual navigation is very minimal. The second gap is that only few studies isolate the causal effect of geometry under the same optimisation conditions. Most previous studies present a variety of confounding variables at once, such as auxiliary tasks, semantic inputs, memory modules, or distorted reward functions.

A third unsolved question is about the relationship between architectural inductive bias and curriculum learning. Curricula can stabilise learning (Narvekar et al., 2020; Romac et al., 2021), but their effectiveness may depend on whether the representation can encode transferable structure.

Lastly, out-of-distribution (OOD) robustness has not been given a comprehensive exploration in a single experimental environment. In embodied benchmarks, OOD robustness is often reported (Savva et al., 2019; Anderson et al., 2018). However, it is rare to find work that perform this analysis in single-scene learning, multi-scene generalisation, and OOD robustness in one, controlled pipeline. This disaggregation prevents the capability of making holistic conclusions regarding design decisions and their effect on the performance of navigation among the regimes.

Together, these gaps are the main driving force behind the research questions that shape this dissertation and justify a methodology that imposes a strict control on the experiment. Chapter 3 thus outlines the methodological framework which is used to isolate representational effects of geometry-aware conditioning within PPO in AI2-THOR.

# CHAPTER 3

# METHODOLOGY

## Introduction

This chapter presents the methodological basis of the study of geometry-aware reinforcement learning for embodied visual navigation. Here, a controlled experimental-comparative methodology is adopted to evaluate whether introducing explicit geometric priors into a Proximal Policy Optimisation (PPO) agent improves learning stability, generalisation across environments, and robustness to unseen layouts. Two agents are considered that are subjected to the same experimental conditions (i) a control PPO agent which only uses visual observations and (ii) a geometry-aware PPO agent enhanced with low-dimensional geometric information based on agent-goal relations.

The results of reinforcement learning (RL) are extremely dependent on experimental design parameters, such as determinism of environment, reward shaping, hyperparameter setting, and protocols. Any little variation in each of these factors may result in significant deviation in performance. This high sensitivity makes attribution of performance differences unreliable if controls are insufficient. This methodology therefore gives precedence to rigorous isolation of variables, reproducibility and deterministic analysis so that any identified variation between agents can be ascribed to architectural inductive bias as opposed to confounding implementation effects.

The methodological design is a direct operationalisation of Research Questions RQ1-RQ5, because it divides the experiments into three consecutive stages: (i) single-

scene learning, (ii) multi-scene generalisation, and (iii) out-of-distribution (OOD) evaluation. Such a design is compatible with best practice in the evaluation of embodied AI. It aids the assessment of both in-distribution performance and generalisation behaviour.

## Research Philosophy and Methodological Logic

### 3.1.1 Philosophical Grounding

The research is based on a pragmatist epistemology, in which knowledge is considered based on empirical efficacy, as opposed to theoretical ideal. It is not aimed at suggesting a generally best navigation architecture, but to find out whether geometry-conscience inductive bias can have quantifiable and reproducible advantages under real-world experimental limitations.

It follows an empirical reinforcement learning research methodology, where hypotheses are confirmed by controlled experimentation and quantitative comparison. The learning outcomes are based on the repeated interaction between the agents, and the environment. The conclusions are based on the observed pattern of the performance across different training and evaluation regimes and not based on closed-form analytical proofs. The methodological philosophy and practical flow are depicted in Figure 1 below.

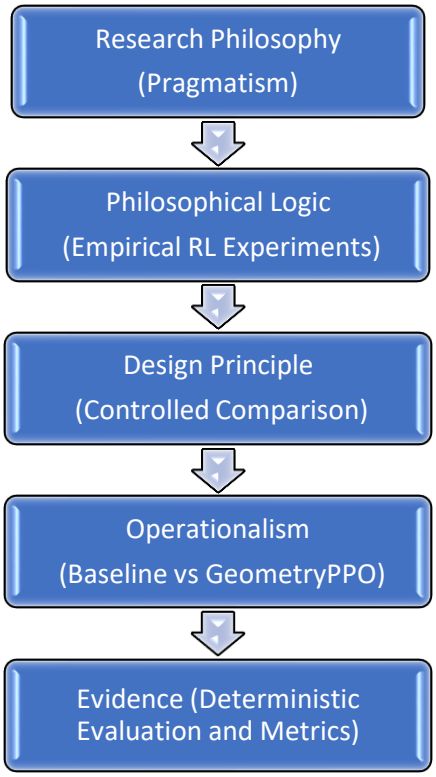


**Figure 1. Methodological Philosophy and Practical Flow**

### 3.1.2 Comparative Experimental Logic

The experimental design used is a comparative one, to isolate the causal effect of geometric priors. The agents are exposed to the same environment, reward schemes, hyperparameters, training plans and evaluation processes. The only manipulated variable is the state of the policy representation, which is the inclusion of low-dimensional geometric information in addition to visual observations. This design is deliberate and aims at ensuring internal validity and allows direct attribution of performance differences to architectural inductive bias.

### 3.1.3 Ethical Considerations and Simulation-Based Justification

All the experiments take place in the AI2-THOR simulator (v5.1.1). The determinism, safety, and cost efficiency of simulation-based experimentation justify its application as it allows large-scale RL experimentation in a manner that does not involve physical danger or uncontrollable environmental variability. Deterministic simulation is specifically critical for PPO-based evaluation, whereby the learning dynamics can be hidden by stochasticity.

The ethical and environmental issues are considered by limiting training budgets, eliminating unproductive experimentation runs, and unjustifiable retraining. The methodology acknowledges the computational cost of RL experimentation and employs effective and goal-oriented training schedules in compliance with responsible AI practice.

**Research Design**

The study adopts a controlled, quantitative experimental design comprising two experimental groups:

- Baseline PPO agent (visual observations only)
- Geometry-aware PPO agent (visual observations + pose information)

Both agents are evaluated under identical conditions across all experimental phases.

### 3.1.4 Variables

- **Independent Variable:** Policy architecture (baseline vs geometry-aware).
- **Dependent Variables:** Navigation success rate, episodic return, and an SPL-proxy derived from distance penalties.
- **Control Variables:** Simulator environment, scene layouts, hyperparameters, random seed, reward function, training budget, and evaluation protocol.

### 3.1.5 Study Phases

The experimental protocol is structured into three phases as shown in Figure 2:

- **Phase I: Single-Scene Training (FP1):** Evaluates learning stability and sample efficiency.
- **Phase II: Multi-Scene Generalisation (FP1-FP5):** Assesses transfer across structurally distinct but related environments.
- **Phase III: OOD Evaluation (FP6-FP8):** Tests robustness on unseen environments.

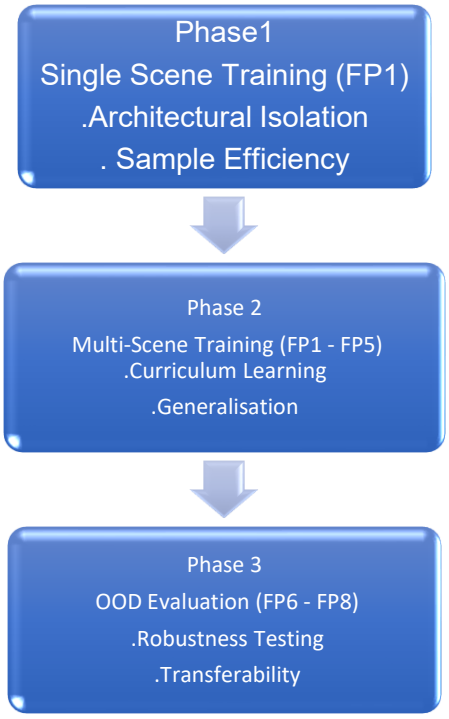


**Figure 2. Three-Phase Research Design Pipeline**

### 3.1.6 Proximal Policy Optimisation

Proximal Policy Optimisation (PPO) is employed as the core learning algorithm due to its stability under on-policy optimisation and its widespread adoption in embodied navigation research. PPO constrains policy updates via a clipped surrogate objective that limits destructive gradient steps.

The PPO clipped objective is defined as (Schulman *et al.,* 2017):

$$\mathcal{L}_{\text{CLIP}}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \qquad (3.1)$$

with

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} \qquad (3.2)$$

and $\hat{A}_t$ denotes the Generalised Advantage Estimate (GAE) (Schulman *et al.,* 2015).

The full PPO loss optimised during training is:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{CLIP}}(\theta) - c_v\mathcal{L}_{\text{VF}}(\theta) + c_e H(\pi_\theta) \qquad (3.3)$$

where $\mathcal{L}_{\text{VF}}$is the value function loss and $H(\pi_\theta)$ is an entropy regularisation term. The full PPO objective combines the clipped policy loss, value-function regression, and entropy regularisation (Schulman *et al.,* 2017).

### 3.1.7 Reward Model and Action Space

The reward function balances task completion and navigation efficiency:

- +1.0 terminal reward for reaching the goal
- −0.01 per-step penalty
- −0.02 penalty for failed actions

The action space comprises standard discrete navigation primitives: MoveAhead, RotateLeft, and RotateRight.

### 3.1.8 Observations and State Representations

Both agents receive RGB observations resized to 84×84 pixels. The geometry-aware agent additionally receives a low-dimensional geometric vector encoding distance-to-goal, initial distance, and normalised progress toward the goal. No depth, semantic, map-based, or recurrent memory information is provided.

## Geometry-aware PPO Architecture

### 3.1.9 State and Policy Formalisation

Navigation is formulated as a POMDP, where state representations may be augmented with auxiliary geometric variables. The baseline and geometry-aware agents differ only in state representation:

$$s_t^{base} = o_t, \qquad s_t^{geom} = (o_t, g_t) \qquad (3.4)$$

where $o_t$ denotes the visual observation and $g_t$ denotes the low-dimensional geometric vector (Kaelbling, Littman and Cassandra, 1998; Liu, Suganuma and Okatani, 2024).

Correspondingly, the policy parameterisations are (Liu, Suganuma and Okatani, 2024):

$$\pi_\theta^{base}(a_t \mid o_t),\ \pi_\theta^{geom}(a_t \mid o_t, g_t) \qquad (3.5)$$

### 3.1.10 Feature Fusion via Geometric Conditioning

Feature-level conditioning follows a FiLM-style modulation mechanism (Perez *et al.,* 2018), adapted for geometry-aware navigation (Liu, Suganuma and Okatani, 2024). The geometry-aware policy incorporates geometric information through feature-level conditioning:

$$z_t = \gamma(g_t) \odot f(o_t) + \beta(g_t) \qquad (3.6)$$

where $f(o_t)$ is the visual feature embedding and $\gamma(\cdot)$, $\beta(\cdot)$ are learned modulation functions. The fused representation $z_t$ is shared by the policy and value networks.

### AI2-THOR Environment Design

AI2-THOR is selected for its photorealistic indoor environments, deterministic physics, and diversity of layouts (Kolve *et al.,* 2017). These properties enable controlled comparison across agents.

### 3.1.11 Scene Grouping

Scenes are partitioned as depicted in Figure 3. At the first stage both agents were subjected to uniform training on Floor Plan 1, which has the simplest layer out. The

next phase of training incorporated FP1 to FP5 with increasing scene complexity. The last stage was out-of-distribution on FP6 to FP8.

$$S_{train} = \{FP1, FP2, FP3, FP4, FP5\}, S_{OOD} = \{FP6, FP7, FP8\} \quad (3.7)$$

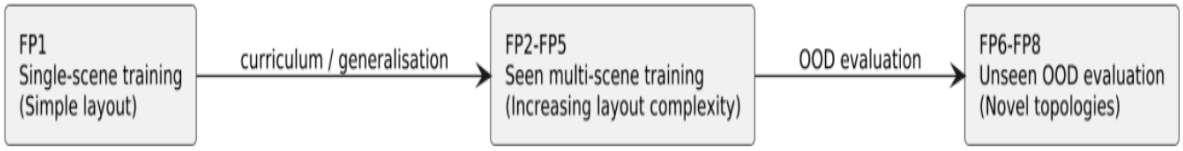


**Figure 3. illustrates the Floorplan grouping strategy.**

**Dataset and Training Regimes**

The study uses the predefined AI2-THOR scene dataset. No manual annotation or data collection is performed. Scene layouts and goal locations are fixed and reliable, directly supporting controlled variation in spatial structure. Limitations such as static lighting and absence of dynamic obstacles are acknowledged.

### 3.1.12 Curriculum Sampling

Scene sampling follows either a uniform or weighted curriculum:

$$p(FP_i) = \begin{cases} \frac{1}{5}, & \text{uniform curriculum} \\ w_i, & \text{weighted curriculum, } \sum_i w_i = 1 \end{cases} \quad (3.8)$$

Weighted curricula are adjusted adaptively based on observed failure modes.

### Hyperparameter Configuration

All experiments use identical PPO hyperparameters for both agents (Table 3.1), ensuring fair comparison. The number of parallel environments varies across experimental phases to balance computational constraints and training stability, while core optimisation parameters remain unchanged.

**Table 3.1: Hyperparameter Configuration**

| **Parameter** | **Value** |
|---|---|
| Learning rate | 3e−4 (linear decay) |
| n_steps | 1024 |
| Batch size | 2048 |
| Discount factor (γ) | 0.99 |
| GAE λ | 0.95 |
| Clip range | 0.2 |
| Entropy coefficient | 0.01 |
| Value loss coefficient | 0.5 |
| Max gradient norm | 0.5 |
| Parallel environments | 8 |

## Evaluation Protocol

### 3.1.13 Metrics

- Success Rate (primary)
- Episodic Return
- SPL-proxy based on distance penalties

### 3.1.14 Episode Counts

- FP1: 200 episodes
- FP1–FP5: 50 episodes per scene
- FP6–FP8: 50 episodes per scene

### 3.1.15 Determinism

All final evaluations use deterministic policies as depicted in the evaluation pipeline below (Figure 4).

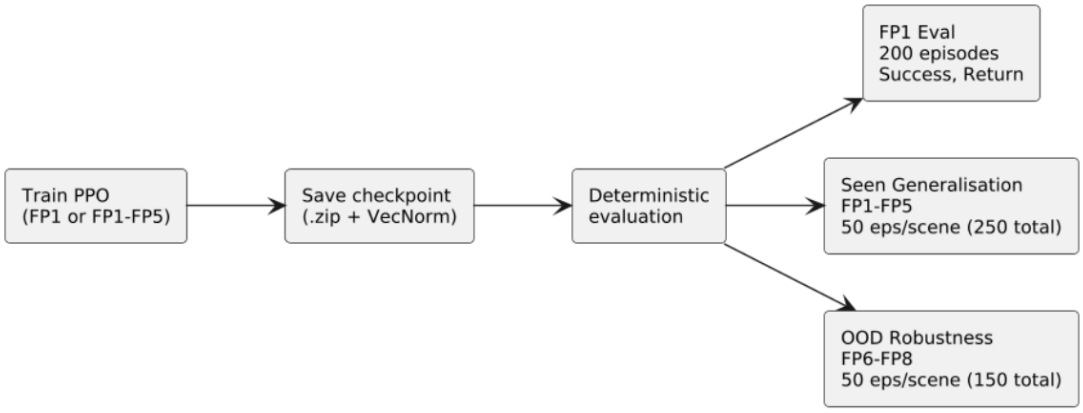


**Figure 4. illustrates the evaluation pipeline.**

## Reproducibility and Experimental Controls

Reproducibility was ensured through deterministic execution, fixed random seeds, synchronous vectorised environments, and strict software version locking. All

experiments were conducted on identical hardware and software configurations to eliminate variability arising from computational factors. Table 3.2 documents the complete system configuration used consistently across all training and evaluation runs.

**Table 3.2: Software and Hardware Configuration Used for All Experiments**

| Component | Specification |
|---|---|
| GPU | NVIDIA RTX 5070 (Ada/Blackwell-era, 8 GB VRAM) |
| CPU | Intel® Core™ i7-14650HX |
| CPU Cores / Threads | 12 cores / 24 threads |
| System Memory (RAM) | 16 GB |
| Operating System | Ubuntu 22.04 LTS (WSL2, Microsoft Hypervisor) |
| CUDA Version | CUDA 12.8 |
| Python Version | Python 3.12 (virtual environment) |
| PyTorch Version | PyTorch 2.3 (GPU-enabled) |
| Stable-Baselines3 | SB3 2.2+ |
| Gymnasium | 0.29.1 |
| AI2-THOR Version | 5.1.1 |
| Rendering Backend | Unity (headless via Xvfb) |
| Execution Mode | Deterministic, synchronous vectorised environments |

### Methodological Limitations

The methodology focuses on a single navigation task, uses an SPL-proxy rather than official SPL, and does not employ domain randomisation. These limitations are addressed in Chapter 6.

### Summary

This chapter established a rigorous, reproducible methodological framework for evaluating geometry-aware PPO in embodied navigation. By formalising state representations, objectives, curricula, and evaluation protocols, it provides a sound foundation for the implementation and empirical analysis presented in subsequent chapters.

# CHAPTER 4

# IMPLEMENTATION

## Introduction to the Implementation Strategy

This chapter presents the concrete implementation of the experimental framework defined in Chapter 3. While the preceding chapter established the research design, optimisation objectives, evaluation protocol, and theoretical formulations, the present chapter focuses exclusively on how the experiments were executed in practice. Emphasis is placed on software realisation, environment integration, architectural instantiation, training pipelines, and execution controls.

The implementation was structured as a reproducible pipeline comprising:(i) AI2-THOR environment integration through a custom Gym-compatible wrapper, (ii) instantiation of a baseline PPO agent, (iii) instantiation of a geometry-aware PPO agent consistent with the architectural formulation in Chapter 3, (iv) execution of uniform and weighted multi-scene curricula, and (v) deterministic evaluation on both in-distribution and out-of-distribution scenes.

All components were implemented using fixed software versions, deterministic execution settings, and consistent scripts to ensure that any empirical differences observed in Chapter 5 arise solely from architectural differences rather than procedural variation.

## Software Environment and Execution Context

All experiments were executed on a single workstation using the fixed hardware and software configuration documented in Table 3.2 (Chapter 3). To avoid redundancy,

hardware specifications, framework versions, and CUDA configuration are not restated here.

From an implementation perspective, several execution constraints were enforced:

- Training was conducted on a single GPU without distributed or multi-GPU parallelism.
- Vectorised environments were executed synchronously to preserve deterministic rollout order.
- Unity rendering was performed headlessly using Xvfb to support WSL2 execution.
- all experiments were run inside a Python virtual environment with strict dependency isolation.

These constraints ensured consistency across baseline and geometry-aware agents and eliminated non-deterministic sources of variation.

**AI2-THOR Navigation Environment Implementation**

**4.1.1 Custom Gym-Compactible Environment Wrapper**

AI2-THOR does not natively expose a Gym interface compatible with Stable-Baselines3. To bridge this gap, a custom Gym-compatible wrapper (thor_nav_env.py) was implemented. This wrapper provides a unified interface that allows both PPO variants to interact with the simulator through identical APIs.

The wrapper encapsulates the following responsibilities:

- Deterministic scene initialisation and reset logic.
- Agent spawning and goal placement using fixed protocols.
- RGB frame capture and resizing to 84×84 resolution.
- Computation of goal distance and progress signals.
- Reward computation consistent with the formulation in Chapter 3.

- Execution of discrete navigation actions and detection of failed actions.
- Termination and truncation handling.

For the geometry-aware agent, an additional wrapper (thor_nav_geom_env.py) augments the base environment by exposing low-dimensional geometric signals derived from agent-goal relations. Importantly, the underlying environment dynamics and reward structure remain unchanged.

To maintain clarity and conciseness in this chapter, the full implementation of the environment wrapper is provided in Appendix B, with Figure 5 illustrating the interaction between the base AI2-THOR environment, the geometry wrapper, and the PPO agent.

```
class ThorPointNav84(gym.Env):
    def reset(self, *, seed=None, options=None):
        self.ctrl.reset(self.scene); self.ctrl.step(action="Initialize")
        self._goal = self._random_reachable(); start = self._random_reachable()
        while _euclid(start, self._goal) < 1.5: start = self._random_reachable()
        self._teleport_agent(start); ev = self.ctrl.last_event
```

**Figure 5. AI2-THOR Gym Environment Wrapper**

### 4.1.2 Action and Observation Pipelines

#### Action Space

The discrete action space mirrors standard embodied navigation benchmarks and includes:

- MoveAhead (0.35 m),
- RotateLeft (90°),
- RotateRight (90°).

**Observation Space**

- Visual observations: Both agents receive RGB frames resized to 84×84 pixels.
- Geometric observations: The geometry-aware agent additionally receives a three-dimensional vector comprising current distance-to-goal, initial distance-to-goal, and normalised progress.

This design ensures that architectural differences are isolated to the policy network rather than the environment interface.

**Baseline PPO Implementation**

**4.1.3 Network Instantiation**

The baseline agent was instantiated using the unmodified CnnPolicy provided by Stable-Baselines3. The architecture consists of:

- a convolutional encoder operating on RGB inputs.
- a shared latent representation of fixed dimensionality.
- separate multilayer perceptron heads for policy and value estimation.

No architectural extensions, auxiliary losses, or memory modules were introduced. The baseline implementation serves as a strict control condition against which the geometry-aware extension is evaluated. A simplified diagram of the architecture is given in Figure 6 below.

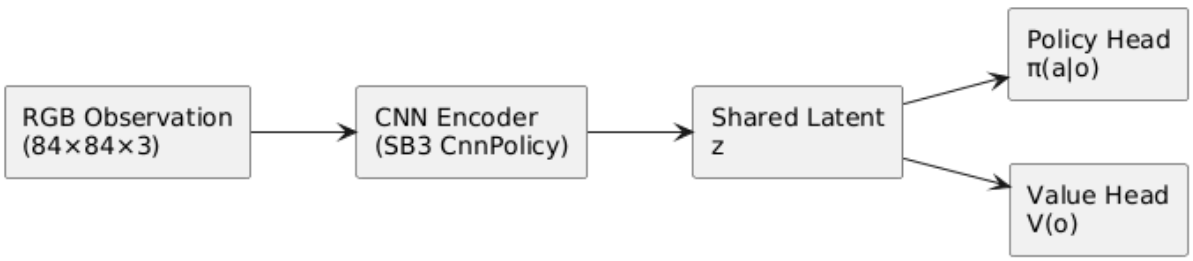

**Figure 6. Baseline PPO Network Architecture**

### 4.1.4 Training Pipeline (FP1)

Baseline training on FloorPlan1 (FP1) was executed incrementally to assess optimisation behaviour and learning stability. Training followed the PPO optimisation objective defined in Chapter 3 and used identical hyperparameters across runs. A snippet of baseline PPO training script is given in Figure 7, while full training script is given in appendix B.

The training pipeline followed a fixed sequence:

1. environment instantiation via the custom wrapper.
2. PPO agent initialisation.
3. synchronous rollout collection.
4. periodic checkpoint saving.
5. deterministic evaluation using frozen policies.

Training diagnostics including entropy, value loss, policy loss, KL divergence, and explained variance were logged continuously to monitor optimisation dynamics.

```
env = DummyVecEnv([make_env()])
model = PPO("CnnPolicy", env, n_steps=1024, batch_size=1024, n_epochs=4,
            gamma=0.99, gae_lambda=0.95, clip_range=0.2, ent_coef=0.01,
            vf_coef=0.5, learning_rate=3e-4, device="cuda", seed=42)
model.learn(total_timesteps=300_000, progress_bar=True)
```

**Figure 7. Baseline PPO Training Script**

**Geometry-Aware PPO Implementation**

### 4.1.5 Architectural Extensions

The geometry-aware PPO agent extends the baseline architecture by conditioning policy learning on explicit geometric signals derived from agent-goal relations. The extension introduces three implementation components:

- a low-dimensional geometric input vector (3-D).
- feature-level fusion via concatenation with visual embeddings.
- a shared latent representation processed jointly by the policy and value heads.

No feature-wise linear modulation (FiLM) or multiplicative gating was implemented in the final system. All other components including the visual encoder, optimiser, policy head, and value head remain unchanged relative to the baseline.

This design reflects a deliberate trade-off between architectural simplicity and geometric inductive bias. Figure 8 depicts a simplify version of the design.

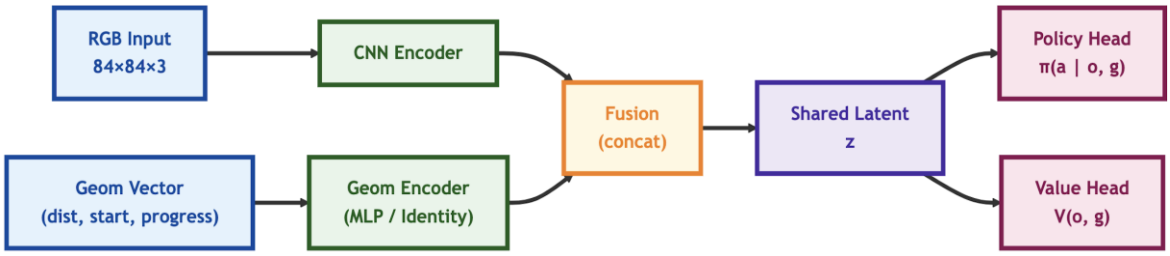


**Figure 8. Geometry-Aware PPO Architecture**

### 4.1.6 Geometry-Aware PPO Training Pipeline

Training scripts (Figure 9) for the geometry-aware agent mirror those used for the baseline PPO, differing only in the policy class and observation specification. This

ensures that any performance differences are attributable to architectural inductive bias rather than training procedure. Full training scripts are provided in Appendix B,

```
vec_env = VecTransposeImage(SubprocVecEnv(env_fns, start_method="forkserver"))
model = PPO("MultiInputPolicy", vec_env, n_steps=1024, batch_size=4096, n_epochs=4,
            gamma=0.99, gae_lambda=0.95, clip_range=0.2, ent_coef=0.01,
            vf_coef=0.5, learning_rate=3e-4, device="cuda")
model.learn(total_timesteps=300_000, progress_bar=True)
```

**Figure 9. Geometry-Aware PPO Policy Definition**

**Multi-Scene Curriculum Implementation**

### 4.1.7 Uniform Curriculum Execution

An initial uniform curriculum was implemented across FP1–FP5, assigning equal sampling probability (20%) to each scene. Scene selection was handled at environment reset using a deterministic sampler. A snippet of the script is given in Figure 10. We attached the full script in appendix B.

This phase served as a controlled baseline for assessing the impact of curriculum structure on both agents.

```
def reset(self, **kwargs):
    if self._env is not None: self._env.close()
    scene = self._rng.choice(self._scenes)  # uniform curriculum
    self._env = ThorPointNavGeom84(scene=scene, width=84, height=84, max_steps=200)
    return self._env.reset(**kwargs)
```

**Figure 10. Uniform Curriculum Sampler**

### 4.1.8 Weighted Curriculum refinement

Weighted curricula were implemented to refine performance on scenes exhibiting poor success rates. Scene weights were adjusted iteratively based on observed failures following the weighting logic in Figure 11. Fixed ordering or schedule was not enforced.

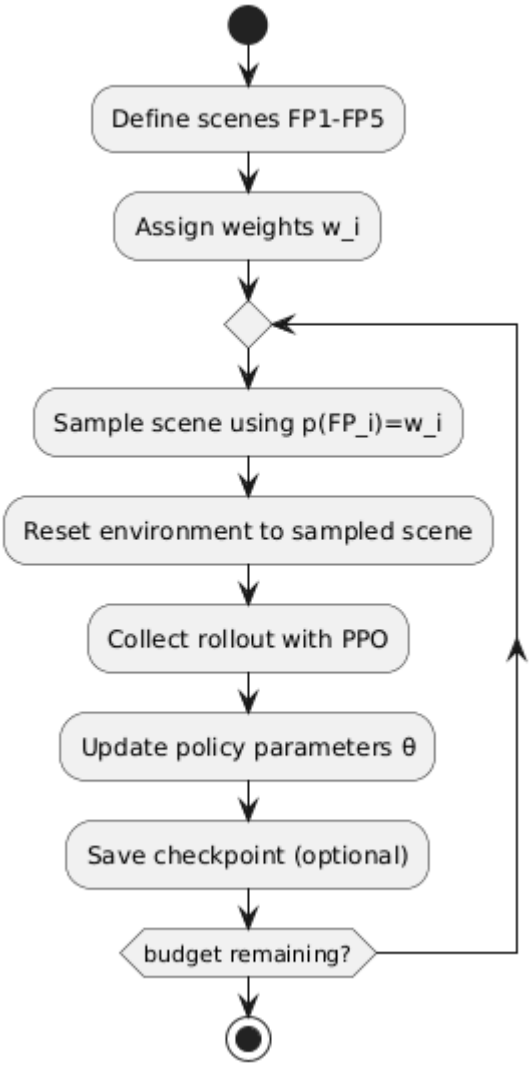


**Figure 11. Curriculum and Scene-Weighting Logic**

### Out-of-Distribution Evaluation Pipeline

Out-of-distribution evaluation was performed on FP6–FP8 using frozen checkpoints. No training updates were applied during evaluation. The same deterministic evaluation script was used across all agents and scenes, and a snippet of the script is contained in Figure 12 while the full code can be found in appendix B.

```
for scene in ["FloorPlan6","FloorPlan7","FloorPlan8"]:
    env = EnvCls(scene=scene, width=84, height=84, max_steps=200, seed=1234)
    succ = sum(run_episode(model, env) for _ in range(50))
    print(f"{scene}: success {succ}/50 ({succ/50:.1%})")
```

**Figure 12. OOD Evaluation Script**

### Reproducibility and Execution Controls

All implementation scripts enforce the experimental controls defined in Chapter 3, including fixed random seeds, deterministic PyTorch settings, synchronous vectorised environments, and consistent Unity rendering parameters. Checkpoint naming and directory structure were standardised across all experiments.

### Chapter Summary

This chapter detailed the practical implementation of the experimental framework defined in Chapter 3. By adhering strictly to fixed architectures, controlled execution pipelines, and reproducible scripts, the implementation ensures that all empirical differences analysed in Chapter 5 arise from architectural design rather than procedural variation. The resulting checkpoints, logs, and evaluation artefacts provide the empirical foundation for the results and analysis presented next.

# CHAPTER 5

# RESULTS AND ANALYSIS

**Overview**

This chapter reports and analyses the empirical results from a controlled comparison between a baseline Proximal Policy Optimisation (PPO) agent and a geometry-aware PPO agent evaluated within the AI2-THOR navigation environment. Results are organised across three experimental stages: (i) single-scene learning on FloorPlan1 (FP1), (ii) multi-scene generalisation across FP1-FP5, and (iii) out-of-distribution (OOD) evaluation on unseen scenes FP6-FP8.

All reported results are obtained using deterministic evaluation, ensuring that observed performance differences reflect learned policy behaviour rather than stochastic action sampling. Success rate is treated as the primary metric, supported by episodic return (mean ± standard deviation) and training stability diagnostics where relevant.

Tables 5.1-5.7 and Figures 13-18 present the quantitative evidence used to address Research Questions RQ1-RQ5.

**Single-Scene Learning Performance (FP1)**

### 5.1.1 Resource Requirements

**Table 5.1. FP1 Single-Scene Performance at Key Training Checkpoints**

| Model | Checkpoint Stage | Training Steps | FP1 Success (%) | Mean Return | Return SD |
|---|---|---|---|---|---|
| **Baseline PPO** | Early warm-up | 10,000 | 0.0 | - | - |
| **Baseline PPO** | Continued training | 210,000 | 0.0 | - | - |
| **Baseline PPO** | Continued training | 310,000 | 5.0 | - | - |
| **Baseline PPO** | Final FP1 model | 610,000 | 62.5 | 1.419 | 2.955 |
| **Geometry-Aware PPO** | Initial FP1 model | 300,000 | 18.5 | - | - |
| **Geometry-Aware PPO** | Continued FP1 model | 600,000 | 82.5 | 2.913 | 1.654 |

Table 5.1 reports FP1 success rates for both agents at successive training checkpoints. The baseline PPO exhibits prolonged stagnation, achieving 0% success up to approximately 210k environment steps and only marginal improvement (5% success) at 310k steps. Substantial performance gains occur only after extended optimisation, with the baseline converging to (125/200 successful episodes) 62.5%

with a mean return of 1.419 ± 2.955 success following approximately 610k training steps. This slow progression reflects the difficulty of resolving sparse rewards and partial observability when relying solely on visual input.

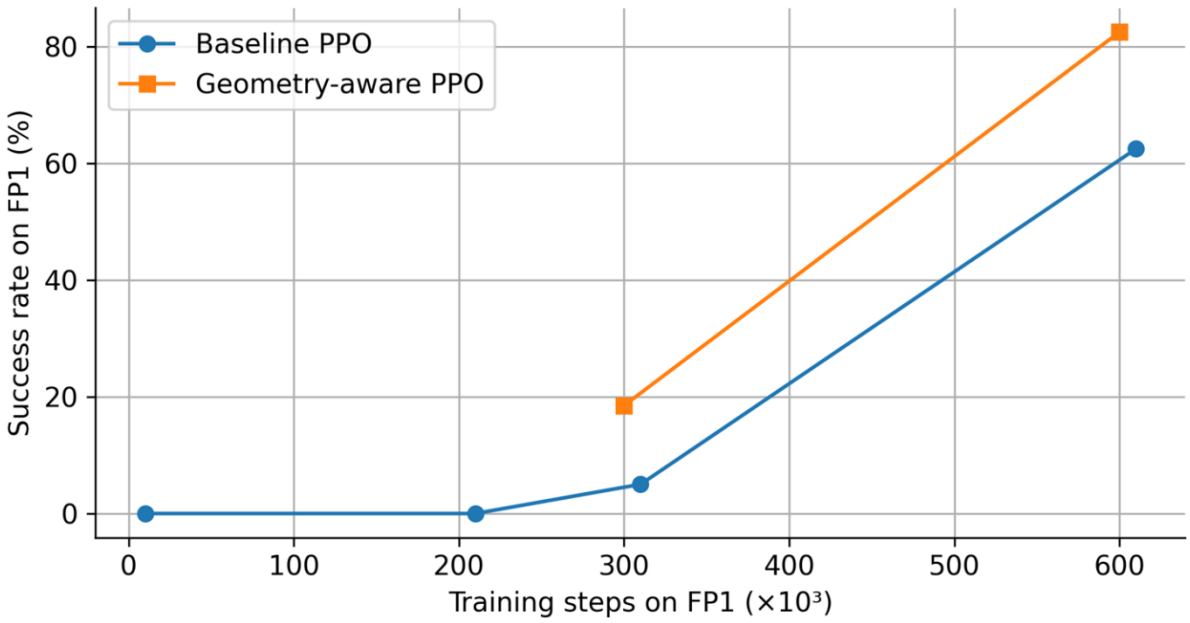


**Figure 13. FP1 Learning Curves: Baseline PPO vs Geometry-Aware PPO**

In contrast to the baseline, the geometry-aware PPO demonstrates significantly faster improvement. At 300k steps, the agent achieves 18.5% success, and by 600k steps reaches 82.5% success (165/200 successful episodes) with a higher and more concentrated mean return of 2.913 ± 1.654. Hence, it outperforms the baseline while requiring fewer total interactions. Episodic return trends mirror this pattern, with earlier increases and reduced variance indicating improved optimisation stability. The difference in FP1 success is substantive: the normal-approximate 95% confidence interval (CI) for the baseline success rate (0.625 ± ≈0.07) does not overlap that of the geometry-aware agent (0.825 ± ≈0.06). FP1 returns display a similar pattern; the

standard error of the mean is roughly twice as large for the baseline as for the geometry-aware agent. This reflects more volatile trajectories.

Figure 13 (learning curves) illustrates the contrasting learning dynamics. The baseline curve exhibits a long plateau, which is then followed by a relatively late but noisy rise in success. The geometry-aware curve climbs earlier and stabilises at a higher plateau. This is consistent with logs from the trainings. Near 300k step, the explained variance of the baseline fluctuates in a low range (≈0.06-0.27). On the other hand, the geometry-aware agent reaches and maintains substantial higher value (≈0.72-0.75) at comparable steps. At the same time, the value loss for the geometry-aware agent is an order of magnitude lower (≈0.08-0.18) than for the baseline (≈0.38-1.10), which indicates a better-calibrated value function.

**Table 5.2. FP1 training diagnostics averaged over the 3 final updates**

| **Metric** | **Baseline PPO** | **Geometry-Aware PPO** |
|---|---|---|
| **Approx. KL Divergence** | 0.0031 | 0.0078 |
| **Entropy Loss** | -0.361 | -0.191 |
| **Explained Variance (value)** | 0.145 | 0.737 |
| **Value Loss** | 0.915 | 0.173 |
| **Mean Episode Length** | 69.2 | 43.5 |
| **Mean Episode return** | 2.83 | 3.26 |
| **Training Steps at Checkpoint** | ≈ 3.0 × 105 | ≈ 3.0 × 105 |

Approximate KL and entropy trends are different for the two agents, see Table 5.2 above. For baseline FP1 run, KL remains small (≈0.002-0.0035) while the magnitude of the entropy is a bit higher than supposed (≈−0.31 to −0.39). This suggests that the policy stays diffuse even late in training. Geometry-aware runs on the other hand, shows larger KL spike early on (up to ≈0.013) with a corresponding lower entropy (≈−0.18 to −0.21). This observed behaviour is consistent with more decisive updates and earlier policy sharpening.

### 5.1.2 Interpretation

These results directly address RQ1, demonstrating that geometric conditioning improves both convergence speed and asymptotic performance in single-scene navigation. Under identical optimisation settings, the geometry-aware PPO:

- Reaches higher final FP1 success (82.5% vs 62.5%)
- Achieves this with a smaller training budget (600k vs 610k steps) for the single scene run, and lower overall budget when multi-scene runs are considered
- Exhibits more stable training signals, with higher explained variance and lower value loss.

These gains are not marginal. The non-overlapping Cis for success, the tighter return distribution, and the diagnostic trajectories all point in the same direction. And the direction is that conditioning on pose information reduces partial observability, stabilises learning, and improves sample efficiency. The improvement is achieved without altering the optimisation algorithm, reward structure, or training protocol, isolating the effect to the policy representation itself. The baseline on the other hand,

did not failed. It does reach a respectable FP1 performance, making the comparison meaningful and not trivial.

However, these advantages remain scene-local at this stage. FP1 alone does not test whether the geometry-aware agent is simply overfitting to a single layout more efficiently than the baseline. That question is addressed by the generalisation and OOD experiments.

**Single-Scene transfer without Curriculum**

**Table 5.3. Zero-shot transfer performance from FP1 to FP5**

| **Model** | **FP1 (%)** | **FP2 (%)** | **FP3 (%)** | **FP4 (%)** | **FP5 (%)** |
|---|---|---|---|---|---|
| **Baseline PPO (FP1-trained)** | 58.0 | 8.0 | 14.0 | 8.0 | 0.0 |
| **Geometry-Aware PPO (FP1-trained)** | 82.0 | 0.0 | 14.0 | 12.0 | 0.0 |

To test the zero-shot transfer agents that are trained only on FP1 are tested on FP2-FP5, without further training. As expected, both agents show a large performance drop outside of training environment. The results (Table 5.3) suggest that geometric priors alone are not enough for generalisation across unseen environments. The need for exposure to more than one scene during training has been identified even when policies are augmented with explicit spatial information.

**Multi-Scene Generalisation**

### 5.1.3 Balanced-Budget Comparison

**Table 5.4. Balanced-budget multi-scene generalisation across FP1 – FP5**

| Agent | Scene | Episodes | Success | Success Rate | Mean Return | Return SD |
|---|---|---|---|---|---|---|
| **Baseline PPO** | FP1 | 50 | 6 | 58.0% | 0.968 | 2.910 |
| **Baseline PPO** | FP2 | 50 | 15 | 36.0% | 0.770 | 2.555 |
| **Baseline PPO** | FP3 | 50 | 35 | 32% | -1.889 | 4.053 |
| **Baseline PPO** | FP4 | 50 | 10 | 66.0% | 1.052 | 2.384 |
| **Baseline PPO** | FP5 | 50 | 33 | 6% | -0.780 | 2.184 |
| **Baseline PPO** | **Overall** | **250** | **98** | **39.2%** | - | - |
| **Geometry-Aware** | FP1 | 50 | 23 | 46.0% | 1.160 | 2.527 |
| **Geometry-Aware** | FP2 | 50 | 19 | 38.0% | 0.345 | 2.713 |
| **Geometry-Aware** | FP3 | 50 | 31 | 62.0% | 1.459 | 2.354 |
| **Geometry-Aware** | FP4 | 50 | 19 | 38% | 0.743 | 2.196 |
| **Geometry-Aware** | FP5 | 50 | 12 | 24% | -0.047 | 2.015 |
| **Geometry-Aware** | **Overall** | **250** | **104** | **41.6%** | - | - |

Multi-scene training gives a better picture of the interaction of geometry and curriculum exposure. Under similar training conditions, the PPO with geometry awareness gets more even performance on FP1-FP5, especially on the scenes with complex layout (See Table 5.4).

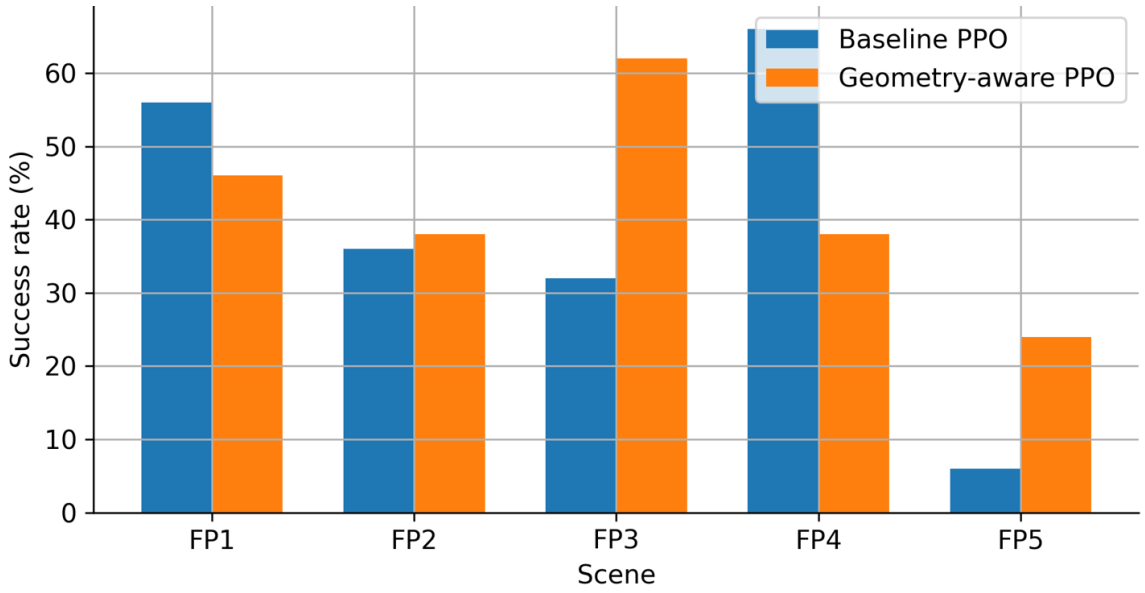


**Figure 14. multi-Scene FP1-FP5 Performances (Balanced Budgets)**

The balanced budget bar chart and the heatmap in Figure 14 and Figure 15 respectively make the pattern visually explicit. Geometry compresses the spread of success across scenes. We observed sacrifices of some peak performance (like on FP5) by the geometry-aware agent. It, however, gains more uniform performance across FP1-FP4. On the side of baseline PPO, there is exhibition of strong peaks and deep troughs, which is atypical symptom of curriculum-induced over-specialisation.

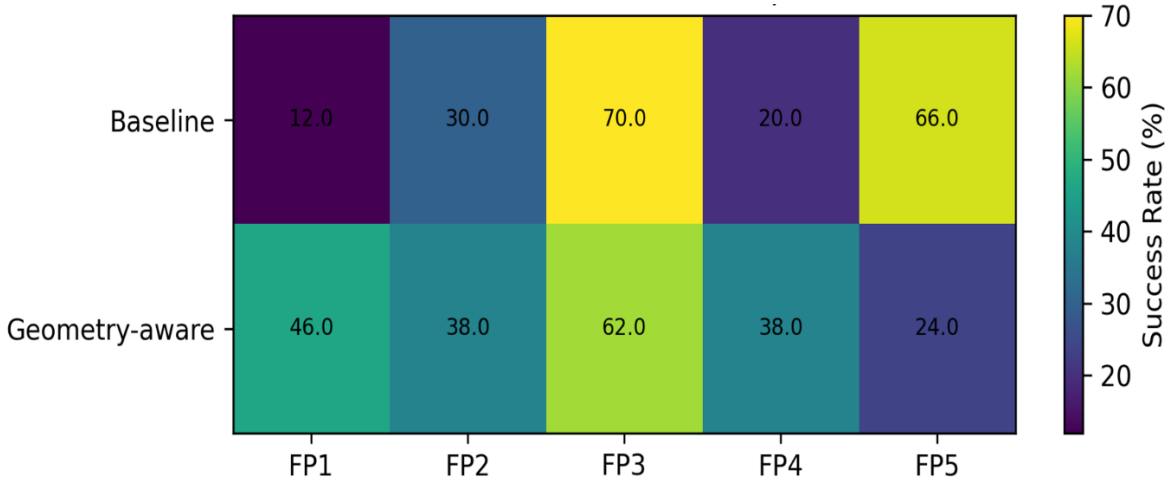


**Figure 15. Balanced-Budget Scene performance Heatmap**

Figure 16 plots mean returns per scene for the balanced-budget comparison. Both agents produce similar mean returns, even though the success profiles differ. The geometry-aware agent avoids the strongly negative returns observed for the baseline, specifically at FP3 under the 400k-step balance regime, with mean −1.889 ± 4.053. This suggests that geometry has less wasteful trajectories than the baseline.

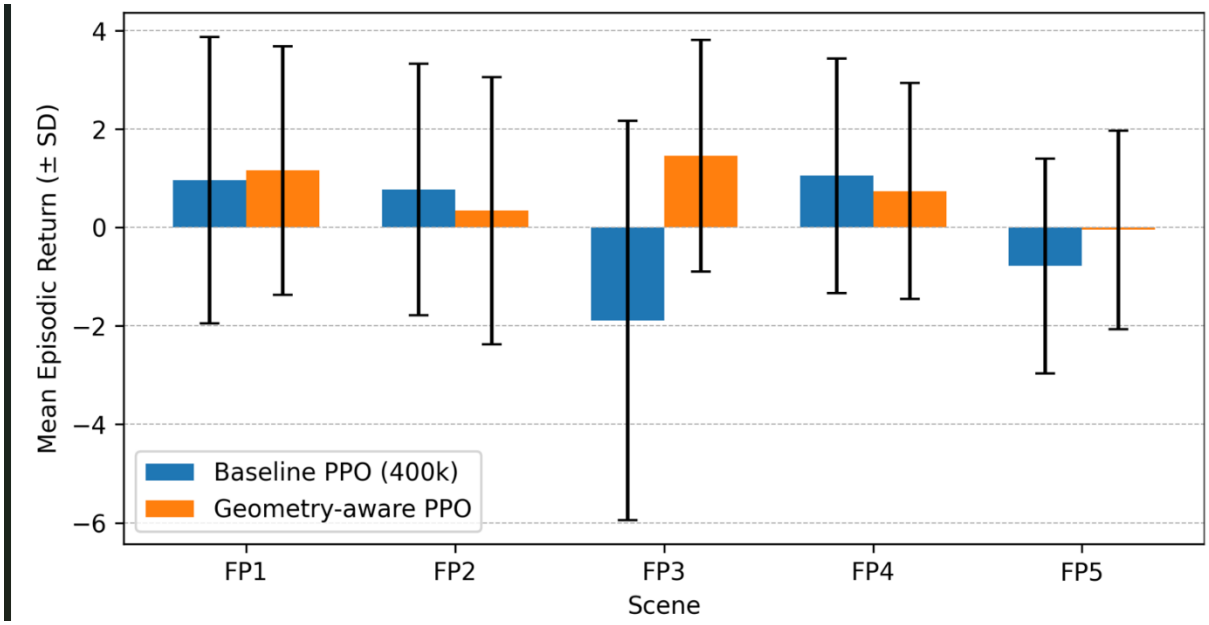


**Figure 16. Balanced-Budget mean return on FP1-FP5 in-distribution**

### 5.1.4 Baseline Stabilisation and Corrective Training

**Table 5.5. Baseline stabilisation multi-scene generalisation across FP1-FP5**

| Agent | Scene | Episodes | Success | Success Rate | Mean Return | Return SD |
|---|---|---|---|---|---|---|
| **Baseline PPO** | FP1 | 50 | 6 | 12.0% | -0.398 | 2.428 |
| **Baseline PPO** | FP2 | 50 | 15 | 30.0% | 0.417 | 2.244 |
| **Baseline PPO** | FP3 | 50 | 35 | 70% | 1.597 | 2.856 |
| **Baseline PPO** | FP4 | 50 | 10 | 20.0% | -1.990 | 2.937 |
| **Baseline PPO** | FP5 | 50 | 33 | 66% | 1.410 | 2.436 |
| **Baseline PPO** | **Overall** | **250** | **99** | **39.6%** | - | - |
| **Geometry-Aware** | FP1 | 50 | 23 | 46.0% | 1.160 | 2.527 |
| **Geometry-Aware** | FP2 | 50 | 19 | 38.0% | 0.345 | 2.713 |
| **Geometry-Aware** | FP3 | 50 | 31 | 62.0% | 1.459 | 2.354 |
| **Geometry-Aware** | FP4 | 50 | 19 | 38% | 0.743 | 2.196 |
| **Geometry-Aware** | FP5 | 50 | 12 | 24% | -0.047 | 2.015 |
| **Geometry-Aware** | **Overall** | **250** | **104** | **41.6%** | - | - |

To recover performance on collapsed FP5, the baseline PPO needs more stabilisation and corrective training. Even after this intervention, there are still uneven performance, in the sense that there is still degradation on some layouts. Scene-level variance is high with success ranging from 12% to 70%. Mean returns range from strong negative (FP4) to clearly positive (FP3 and FP5). The corresponding 95% Cis for mean returns are wide (due to SDs around 2 – 3 and n = 50 per scene). These indicate high behavioural variability across episodes. The overall success rate of the baseline PPO

improved from 39.2% to 39.6% as show in Table 5.5 above. The improvement, however, is still less than the overall success rate of the geometry PPO (41.6 %).

The geometry-aware agent achieved superior performance while using approximately 38% fewer training steps and without equivalent corrective effort. This does suggest that geometric conditioning increases robustness under curriculum-based training and makes the network less sensitive to scene ordering.

**Table 5.6. Multi-Scene (FP1 - FP5) PPO training diagnostics at the final checkpoints**

| Metric | Baseline PPO(Stabilised) | Geometry-Aware PPO |
|---|---|---|
| **Total Timesteps in Final Run Segment** | 212,992 | 100,352 |
| **Mean Episode Length** | 68.6 | 55.3 |
| **Mean Episode return** | 2.69 | 2.36 |
| **Approx. KL Divergence** | 0.0059 | 0.0108 |
| **Entropy Loss** | -0.232 | -0.261 |
| **Explained Variance (value)** | 0.405 | 0.625 |
| **Value Loss** | 0.621 | 0.303 |

Training diagnostics (Table 5.6) supports this interpretation. The final checkpoints of geometry-aware multi-scene run exhibits moderate KL (≈0.006-0.011), medium-low entropy (≈−0.26), and explained variance in the range of 0.52-0.63. This is suggestive of meaningful predictive signal. The stabilised baseline, on the other hand, shows lower explained variance (≈0.22–0.41) and higher entropy in at least one configuration (≈−0.61), which points to residual uncertainty in both value prediction and policy behaviour.

### 5.1.5 Critical Reading of Research Questions 2 and 3 (RQ2-RQ3)

For RQ2 (multi-scene generalisation), experimental evidence is mixed but informative:

- The geometry-aware agent does not dominate the baseline on every scene. FP5 remain challenging for both.
- Geometry, however, achieves competitive overall success (41.6% vs 39.6%) under a smaller training budget. Its performance across FP1-FP4 is more uniform.
- Training diagnostics point to a more stable value function and less extreme over-specialisation.

For RQ3 (curriculum and catastrophic forgetting), the performance difference of the two agents is more pronounced. The baseline needed explicit corrective stabilisation to recover from the catastrophic degradation on FP5 in earlier runs. Such repair was not required by the geometry-aware agent, although the FP5 performance remained modest. Geometry, therefore, does not eliminate forgetting but appears to cushion the adverse effects of curriculum. This supports more stable multi-scene learning at lower cost.

### Out-of-Distribution Evaluation

**Table 5.7 Out-of-distribution (OOD) Evaluation on Novel Scenes FP6 - FP8**

| Agent | Scene | Episodes | Success | Success Rate | Mean Return | Return SD |
|---|---|---|---|---|---|---|
| **Baseline PPO** | FP6 | 50 | 1 | 2.0% | -1.960 | 1.887 |
| **Baseline PPO** | FP7 | 50 | 3 | 6.0% | -1.305 | 3.518 |
| **Baseline PPO** | FP8 | 50 | 3 | 6.0% | -1.909 | 2.597 |
| **Baseline PPO** | **Overall** | **150** | **7** | **4.7%** | - | - |
| **Geometry-Aware** | FP6 | 50 | 3 | 6.0% | -1.567 | 1.603 |
| **Geometry-Aware** | FP7 | 50 | 2 | 4.0% | -1.763 | 1.454 |
| **Geometry-Aware** | FP8 | 50 | 7 | 14.0% | -1.392 | 1.839 |
| **Geometry-Aware** | **Overall** | **150** | **12** | **8.0%** | - | - |

Out-of-distribution evaluation is performed over 3 environments that were not used for training. Absolute success rates are low for both agents which reflect the challenge of zero-shot generalisation in embodied navigation. See Table 5.7 and Figure 17.

Using simple normal approximation, the 95% CI for geometry-aware OOD success is roughly 8.0% ± 4.3%, while that for the baseline is about 4.7% ± 3.4%. There is an overlap between the two intervals. This means that that the improvement is modest in statistical terms, but the relative gain (≈70% improvement in success rate) is still worthy considering the difficulty of the task. At scene level, both agents experienced consistent negative returns, a clear indication of long and mostly unsuccessful trajectories.

The variance patterns have some salient information. On the FP7, the SD in return (≈3.52) of the baseline is more than twice that of the geometry-aware agent (≈1.45). This implies more erratic behaviour and the reason for occasional very poor episodes. The lower variance of geometry suggests slightly more structured failure modes, even when success is rare.

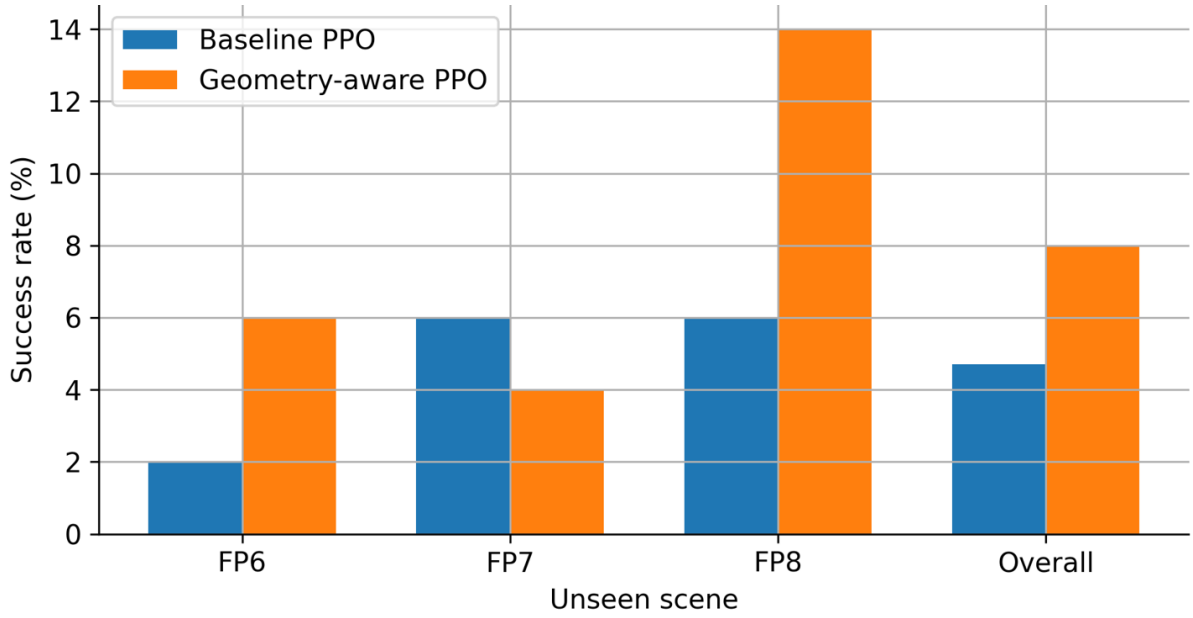


**Figure 17. Out-of-Distribution Evaluation**

### Training Efficiency and Budget Analysis

A comparison of the total training budgets reveals that the geometry-aware PPO achieves better FP1 performance, comparable FP1-FP5 performance, and higher OOD success with fewer interaction steps than the baseline PPO. Figure 18 gives a visual highlight of this trade-off. Geometry shifts the pareto frontier towards less data for more performance.

There is a practical relevance for this. All experiments were run on a single NVIDIA RTX 5070 (8 GB VRAM) with deterministic, synchronous vectorised environments. Reducing total environment steps from ~1.71M to ~1.05M is not just a matter of wall-clock time; it also reduces energy consumption. The further consequence is reduction in the environmental footprint of training, which is a non-trivial consideration for large-scale RL studies.

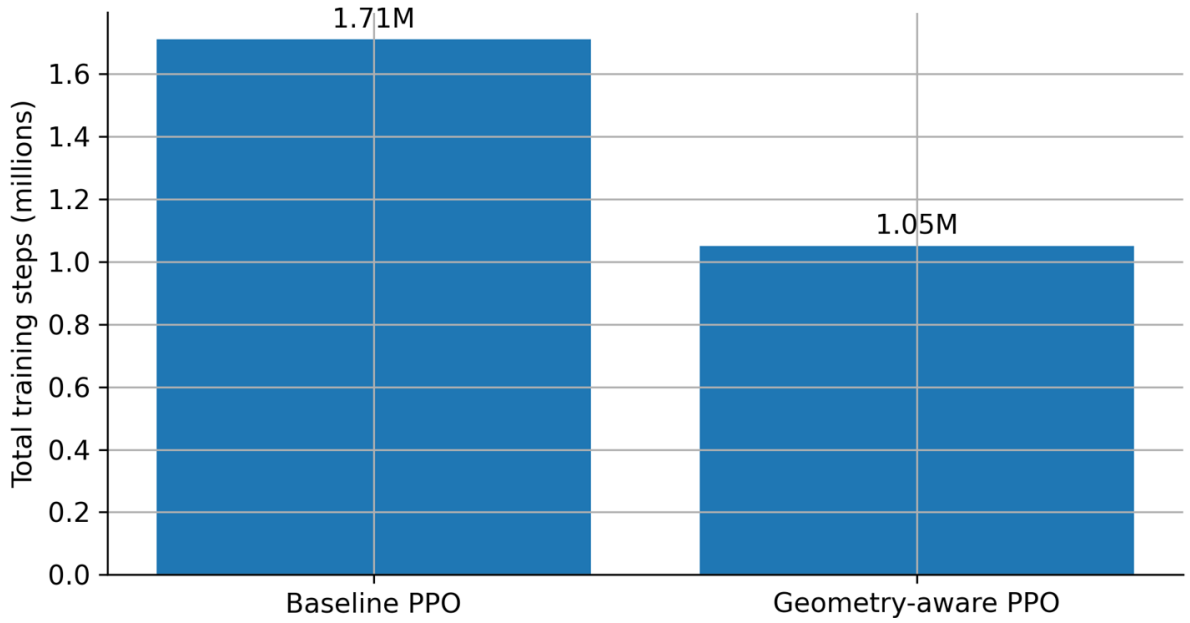


**Figure 18. Training Budget**

**Integrated Critical Discussion (RQ1-RQ5)**

Across FP1, FP1-FP5, and FP6-FP8, a consistent pattern emerges:

- Geometry-aware PPO improves learning efficiency and stability (RQ1).
- It delivers more uniform multi-scene performance at lower budget, with lower stabilisation training (RQ2, RQ3).
- It offers modest but real gains in OOD success and variance control (RQ4).
- It achieves these benefits with lower overall computational cost (RQ5).

However, the results also expose important limitations:

- Geometry alone does not prevent overfitting to specific scenes. FP5 remains a difficult case, with geometry-aware success limited to 24%.
- OOD success rates are still very low; both agents fail on most episodes in FP6-FP8.

The chapter therefore supports a balance view: geometry-aware PPO is a meaningful improvement over a strong baseline but does not provide full solution to the generalisation problem at hand. It however changes how the model fails and how quickly it learns.

**Summary of Key Findings**

The main empirical conclusions are:

- FP1 efficiency and stability: Geometry-aware PPO attains 82.5% on FP1 with fewer steps than the baseline needs to reach 62.5% and exhibits higher explained variance and lower value loss during training.
- Multi-scene generalisation: Under FP1-FP5 curricula, geometry provides more uniform in-distribution performance and reduces the need for corrective retraining.
- OOD robustness: The combination of geometry and curriculum moves the performance-budget trade-off in a favourable direction, reducing interaction cost while improving performance.

These findings form the empirical foundation for the broader conclusions and implications discussed in chapter 6.

# CHAPTER 6

# CONCLUSION AND FUTURE WORK

## Overall Conclusions

This dissertation investigated whether incorporating explicit geometric and pose-based inductive priors into Proximal Policy Optimisation (PPO) improves learning efficiency, generalisation, and robustness in embodied visual navigation. Using a controlled experimental framework in AI2-THOR, a standard vision-based PPO agent was compared against a geometry-aware PPO agent under identical reward functions, hyperparameters, curricula, and deterministic evaluation protocols.

The empirical results demonstrate that geometry-aware PPO delivers consistent, measurable advantages across all evaluated dimensions, though these advantages remain bounded and do not fully resolve the generalisation problem in embodied navigation.

- On FP1, geometry leads to faster convergence and higher final success (82.5% vs 62.5%), with more stable training diagnostics.
- Under multi-scene curricula across FP1-FP5, geometry yields more balanced performance and avoids the severe catastrophic forgetting that required stabilisation in the baseline regime.
- Under OOD evaluation on FP6-FP8, geometry achieves higher success and lower variance but still fails on most episodes.

- Across all phases, these gains are realised with lower total environment interaction, making the geometry-aware agent not only more capable but also more computationally efficient.

At the same time, the work confirms that geometry alone is not a silver bullet. Generalisation remains limited, particularly under strong structural shift. Some scenes (like FP5) remain challenging even for the geometry-aware agent. The study, therefore, positions pose-based inductive bias as a solid, empirically supported step towards more robust navigation, rather than as a complete solution.

**Synthesis of Research Questions**

Table 6.1 summarises how each research question was addressed.

**Table 6.1 Summary of How the Research Questions Were Addressed**

| Research Question | Methodological Approach | Key Empirical Evidence | Conclusion |
|---|---|---|---|
| RQ1: Does geometry-aware PPO improve learning efficiency and stability? | Controlled FP1 training under identical conditions | Geometry-aware PPO reaches 82.5% FP1 success vs 62.5% for baseline; higher explained variance and lower value loss at comparable steps | Geometry-aware PPO improves sample efficiency and optimisation stability |
| RQ2: Do geometric priors improve multi-scene generalisation? | FP1–FP5 uniform and weighted curricula | Geometry-aware agent achieved more consistent FP1–FP4 success and similar overall FP1-FP5 success (41.6% vs 39.6%) | Geometry improves generalisation when combined with curriculum learning but does not eliminate scene-specific weaknesses |

| | | | |
|---|---|---|---|
| | | under a smaller budget | |
| RQ3: Can curriculum learning mitigate catastrophic forgetting? | Progressive and weighted scene sampling | Baseline required corrective retraining to recover from collapse on FP5; geometry-aware PPO avoided catastrophic collapse but remained weak on FP% | Geometry makes curricula more robust but does not fully prevent forgetting |
| RQ4: Does geometry improve OOD robustness? | FP6–FP8 deterministic evaluation | Geometry-aware PPO achieved 8.0% OOD success vs 4.7% for baseline with lower variance on more scenes | Geometry provides modest but measurable robustness gains under structural novelty |
| RQ5: What are the efficiency trade-offs? | Training-budget comparison with fixed hardware and identical algorithms | Geometry-aware PPO attained better performance using ~1.05M steps vs ~1.71M for baseline multi-scene runs | Inductive bias reduces data requirements and computational cost |

### Achievement of Study Objectives

Table 6.2 summarises how the stated objectives were achieved through implementation and evaluation.

**Table 6.2 Summary of How the Study Objectives Were Achieved**

| Objective | Implementation Strategy | Outcome |
|---|---|---|
| Establish a strong PPO baseline | SB3 PPO (CnnPolicy) with deterministic evaluation on AI2-THOR and FP1-FP5 | Baseline reached 62.5% FP1 success and 39.6% FP1-FP5 success after stabilisation, providing a meaningful |
| Integrate geometric inductive priors | Augmentation of visual observations with low-dimensional geometric state via MultiInputPolicy and FiLM-like fusion | Geometry-aware PPO reached 82.5% FP1 success and exceeded baseline FP1-FP4 performance |
| Evaluate multi-scene generalisation | Uniform and weighted scene curricula across FP1-FP5, with matched evaluation protocols | Geometry-aware PPO achieved more stable cross-scene performance (41.6% overall success) under a smaller training budget |
| Assess OOD robustness | Deterministic zero-shot evaluation on FP6-FP8 for both agents | Geometry-aware PPO consistently outperformed baseline (8.0% vs 4.7% success), though absolute success remained low |
| Ensure methodological rigour | Fixed seeds, version locking, synchronous vectorised execution and shared evaluation scripts | Observed performance differences are attributable to architectural design rather than implementation artefacts |

### Implications for Embodied AI Design

The findings carry several implications for embodied AI research and system design:

1. **Role of inductive bias**: The results strengthen the argument that geometry-aware inductive biases are not optional extras but practical tools for improving training stability and efficiency. Geometry-aware PPO is not just a slightly better baseline; it shifts learning curves and diagnostics in a systematic manner and in a favourable direction.

2. **Generalisation beyond single scenes:** Multi-scene curricula alone does not guarantee generalisation. The behaviour of the baseline under weighted sampling illustrates how easy it is for PPO to overspecialise and forget. Geometry moderates this effect and produces more even performance across scenes at a lower cost.

3. **Reproducibility and methodological value**: The controlled comparison, deterministic evaluation, and explicit reporting of training diagnostics provide a template for future navigation studies. Even when absolute performance is modest, the methodology itself including transparent budgets, clear diagnostics, and matched baselines, adds value to the field by clarifying which gains are real and which arise from hidden experimental choices.

**Limitations**

The study acknowledges the following limitations:

- The study focuses on point-goal navigation within a limited set of AI2_THOR kitchen layouts.
- SPL was approximated using distance penalties and success rates rather than computed via full geodesic path lengths.
- No memory mechanisms, mapping modules, or domain randomisation were employed.
- The number of OOD scenes remains small, limiting statistical power under extreme distribution shift.

These constraints were deliberate to isolate architectural effects but restrict the generality of conclusions.

**Future Research Directions**

Building on the present work, several natural extensions of this work are outlined as follows:

1. **Geometry and memory integration**: Combining pose-aware policies with recurrent memory or learned world models may address remaining partial observability.
2. **Scaling environment diversity**: Evaluating across larger and more diverse environments would test robustness under broader structural variation.

3. **Spatial representation learning**: Integrating pose priors with learned spatial maps or topological abstractions may further reduce state aliasing.
4. **Robustness under noise and dynamics**: Testing geometry-aware agents under noisy sensing, dynamic objects, or real-world perturbations would improve ecological validity.
5. **Environmental and ethical accounting**: Future work could instrument training runs with energy and carbon-footprint estimates. This will allow the sample-efficiency gains of geometry to be framed not only as academic but also as environmental contributions.

**Personal Reflection**

From a personal standpoint, this project has been as much an exercise in scientific discipline as in RL engineering. Early in the work, it was tempting to chase higher success rates by tweaking hyperparameters, changing reward functions, or discarding runs that did not give the desired results. Committing instead to a controlled comparison such as fixed seeds, shared code paths, explicit logging, forced a more rigorous mindset.

The experience of watching the baseline collapse under multi-scene, and then stabilising it only with extra targeted runs, made catastrophic forgetting feel concrete rather than abstract. Equally, seeing the geometry-aware agent reach stronger FP1 performance with cleaner diagnostics gave a tangible sense of what inductive bias can achieve when it is aligned with the task.

At the same time, the OOD results were a useful tool. It is easy to be impressed by an 82.5% success rate on FP1; it is harder to be satisfied when OOD success stagnates

around 8%. This contrast has reinforced the importance of evaluating models beyond the settings in which they are tuned, and of treating modest results as valuable evidence rather than as failures.

Finally, the project has highlighted the cost of RL experimentation, in the areas of time, compute, and energy, as well as the responsibility that comes with the cost. The sample-efficiency gains of the geometry-aware agent are not only a technical result, but also a small step towards more sustainable RL practice. Carrying that awareness into future work, both in embodied AI and beyond, will be an important part of my development as a researcher.

**Final Remarks**

In summary, this dissertation shows that geometry-aware reinforcement learning can make PPO-based navigation agents faster, more stable, and marginally more robust, without increasing training budgets. The work does not solve completely the generalisation problem in embodied AI, but it does sharpen our understanding of where geometric priors help, where they fall short, and how they interact with curriculum learning and evaluation design. Those insights, and the reproducible methodology that produced them, are the main enduring contributions of this research.

**WORD COUNT: 12608**

# REFERENCES


Anderson, P., Chang, A., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R. and Savva, M. (2018) On evaluation of embodied navigation agents. arXiv preprint arXiv:1807.06757, [Online], pp. Available from: https://arxiv.org/abs/1807.06757 . [Accessed 24 October 2025].

Brehmer, J., Bose, J., De Haan, P. and Cohen, T.S. (2023) Edgi: Equivariant diffusion for planning with embodied agents. Advances in Neural Information Processing Systems, [Online] 36 , pp. 63818–63834 Available from https://proceedings.neurips.cc/paper_files/paper/2023/hash/c95c049637c5c549c2a08e8d6dcbca4b-Abstract-Conference.html . [Accessed 20 October 2025].

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A. and Zhang, Y. (2017) Matterport3d: Learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158, [Online], pp. Available from: https://arxiv.org/abs/1709.06158 . [Accessed 21 October 2025].

Singh Chaplot, D., Jiang, H., Gupta, S. and Gupta, A. (2020) Semantic Curiosity for Active Visual Learning. arXiv e-prints, pp.arXiv-2006, [Online] Available from:https://link.springer.com/chapter/10.1007/978-3-030-58539-6_19 . [Accessed 13 October 2025].

Chen, J., Li, G., Kumar, S., Ghanem, B. and Yu, F. (2023) How to not train your dragon: Training-free embodied object goal navigation with semantic frontiers. arXiv preprint arXiv:2305.16925, [Online], pp. Available from https://arxiv.org/abs/2305.16925 . [Accessed 22 October 2025].

Cobbe, K., Hesse, C., Hilton, J. and Schulman, J. (2020) November. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning* (pp. 2048-2056). PMLR, [Online] Available from: http://proceedings.mlr.press/v119/cobbe20a.html . [Accessed 27 October 2025].

Goyal, A. and Bengio, Y. (2022) Inductive biases for deep learning of higher-level cognition. Proceedings of the Royal Society A, [Online] 478 (2266), pp. 20210068 Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2021.0068 . [Accessed 19 October 2025].

Huang, P., Xu, M., Zhu, J., Shi, L., Fang, F. and Zhao, D. (2022) Curriculum reinforcement learning using optimal transport via gradual domain adaptation. Advances in neural information processing systems, [Online] 35 , pp. 10656–10670 Available from:https://proceedings.neurips.cc/paper_files/paper/2022/hash/4556f5398bd2c61bd7500e306b4e560a-Abstract-Conference.html . [Accessed 14 October 2025].

Igl, M., Ciosek, K., Li, Y., Tschiatschek, S., Zhang, C., Devlin, S. and Hofmann, K. (2019) Generalization in reinforcement learning with selective noise injection and information bottleneck. Advances in neural information processing systems, [Online]

32 , pp. Available from: https://proceedings.neurips.cc/paper_files/paper/2019/hash/e2ccf95a7f2e1878fcafc8376649b6e8-Abstract.html . [Accessed 21 October 2025].

Jin, S., Wang, X. and Meng, Q. (2024) Spatial memory-augmented visual navigation based on hierarchical deep reinforcement learning in unknown environments. Knowledge-Based Systems, [Online] 285, pp. 111358 Available from: https://www.sciencedirect.com/science/article/pii/S0950705123011061?casa_token=rik6to-aCZIAAAAA:jZ1Yvl09jat4TuY_MycsZeL8Eq9iOGWWhjfd2EgHJSgM5tbQWPt-m011z-kT34IV1nGUN26hrnc . [Accessed 19 October 2025].

Kadian, A., Truong, J., Gokaslan, A., Clegg, A., Wijmans, E., Lee, S., Savva, M., Chernova, S. and Batra, D. (2020) Sim2real predictivity: Does evaluation in simulation predict real-world performance? IEEE Robotics and Automation Letters, [Online] 5 (4), pp. 6670–6677 Available from: https://ieeexplore.ieee.org/abstract/document/9158349/?casa_token=rlx3KNHFKRkAAAAA:iL6BzZpu1_gZ9M5CedTmXZQIcGZYB7Ni5FTHrllESzUPz3C6djmtGmfgt4WP5vIJr1-0GESgaKpXRA . [Accessed 17 October 2025].

Kaelbling, L.P., Littman, M.L. and Cassandra, A.R. (1998) Planning and acting in partially observable stochastic domains. Artificial Intelligence, [Online] 101 (1-2), pp. 99–134 Available from: https://www.sciencedirect.com/science/article/pii/S000437029800023X . [Accessed 17 November 2025].

Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Deitke, M., Ehsani, K., Gordon, D. and Zhu, Y. (2017) Ai2-thor: An interactive 3d environment for visual ai. arXiv preprint arXiv:1712.05474, [Online], pp. Available from: https://arxiv.org/abs/1712.05474 . [Accessed 6 October 2025].

Liu, S., Suganuma, M. and Okatani, T. (2024) Symmetry-aware neural architecture for embodied visual navigation. International journal of computer vision, [Online] 132 (4), pp. 1091–1107 Available from: https://link.springer.com/article/10.1007/s11263-023-01909-4 . [Accessed 15 October 2025].

Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M.E. and Stone, P. (2020) Curriculum learning for reinforcement learning domains: A framework and survey. Journal of Machine Learning Research, [Online] 21 (181), pp. 1–50 Available from: http://www.jmlr.org/papers/v21/20-212.html . [Accessed 13 October 2025].

Parisi, G.I., Kemker, R., Part, J.L., Kanan, C. and Wermter, S. (2019) Continual lifelong learning with neural networks: A review. Neural Networks, [Online] 113 , pp. 54–71 Available from: https://www.sciencedirect.com/science/article/pii/S0893608019300231 . [Accessed 23 October 2025].

Perez, E., Strub, F., De Vries, H., Dumoulin, V. and Courville, A. (2018) Film: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1), [Online] Available from:

https://ojs.aaai.org/index.php/AAAI/article/view/11671 . [Accessed 19 November 2025].

Romac, C., Portelas, R., Hofmann, K. and Oudeyer, P.Y. (2021) Teachmyagent: a benchmark for automatic curriculum learning in deep rl. In International Conference on Machine Learning (pp. 9052-9063). PMLR, [Online] Available from: http://proceedings.mlr.press/v139/romac21a . [Accessed 4 November 2025].

Sangalli, M., Blusseau, S., Velasco-Forero, S. and Angulo, J. (2022) Differential invariants for SE (2)-equivariant networks. In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 2216-2220). IEEE, [Online] Available from: https://ieeexplore.ieee.org/abstract/document/9897301/?casa_token=osDSWpCvtekAAAAA:H2kgFhF3Nf4-rztZLZ4OQgblSmmbaGbt_HX4mqrUzKnCqetUoq2pgGIFr6GbdfayJLPVHi9x5ZVAPw . [Accessed 8 November 2025].

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J. and Parikh, D. (2019) Habitat: A platform for embodied ai research. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9339-9347), [Online] Available from: http://openaccess.thecvf.com/content_ICCV_2019/html/Savva_Habitat_A_Platform_for_Embodied_AI_Research_ICCV_2019_paper.html . [Accessed 11 October 2025].

Schulman, J., Moritz, P., Levine, S., Jordan, M. and Abbeel, P. (2015) High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438, [Online], pp. Available from https://arxiv.org/abs/1506.02438 . [Accessed 12 October 2025].

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. (2017) Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, [Online], pp. Available from: https://arxiv.org/abs/1707.06347 . [Accessed 12 October 2025].

Simeonov, A., Du, Y., Lin, Y.C., Garcia, A.R., Kaelbling, L.P., Lozano-Pérez, T. and Agrawal, P. (2023) Se (3)-equivariant relational rearrangement with neural descriptor fields. In Conference on Robot Learning (pp. 835-846). PMLR, [Online] Available from: https://proceedings.mlr.press/v205/simeonov23a.html . [Accessed 10 October 2025].

Tatiya, G., Francis, J., Bondi, L., Navarro, I., Nyberg, E., Sinapov, J. and Oh, J. (2022) Knowledge-driven scene priors for semantic audio-visual embodied navigation. arXiv preprint arXiv:2212.11345, [Online], pp. Available from: https://arxiv.org/abs/2212.11345 . [Accessed 5 October 2025].

Wani, S., Patel, S., Jain, U., Chang, A. and Savva, M. (2020) Multion: Benchmarking semantic map memory using multi-object navigation. Advances in Neural Information Processing Systems, [Online] 33 , pp. 9700–9712 Available from: https://proceedings.neurips.cc/paper/2020/hash/6e01383fd96a17ae51cc3e15447e7533-Abstract.html . [Accessed 15 October 2025].

Wijmans, E., Kadian, A., Morcos, A., Lee, S., Essa, I., Parikh, D., Savva, M. and Batra, D. (2019) Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. arXiv preprint arXiv:1911.00357, [Online], pp. Available from: https://arxiv.org/abs/1911.00357 . [Accessed 25 October 2025].

Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J. and Savarese, S. (2018) Gibson env: Real-world perception for embodied agents. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 9068-9079), [Online] Available from: http://openaccess.thecvf.com/content_cvpr_2018/html/Xia_Gibson_Env_Real-World_CVPR_2018_paper.html . [Accessed 21 October 2025].

Xue, H., Hein, B., Bakr, M., Schildbach, G., Abel, B. and Rueckert, E. (2022) Using deep reinforcement learning with automatic curriculum learning for mapless navigation in intralogistics. Applied Sciences, [Online] 12 (6), pp. 3153 Available from: https://www.mdpi.com/2076-3417/12/6/3153 . [Accessed 28 October 2025].

Yang, F., Frivik, P., Hoeller, D., Wang, C., Cadena, C. and Hutter, M. (2025) Spatially-enhanced recurrent memory for long-range mapless navigation via end-to-end reinforcement learning. The International Journal of Robotics Research, [Online], pp. 02783649251401926 Available from: https://journals.sagepub.com/doi/abs/10.1177/02783649251401926 . [Accessed 2 January 2026].

Zheng, D., Huang, S., Zhao, L., Zhong, Y. and Wang, L., 2024. Towards learning a generalist model for embodied navigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13624-13634), [Online] Available from: http://openaccess.thecvf.com/content/CVPR2024/html/Zheng_Towards_Learning_a_Generalist_Model_for_Embodied_Navigation_CVPR_2024_paper.html . [Accessed 24 October 2025].

Zhu, F., Zhu, Y., Lee, V., Liang, X. and Chang, X. (2021) Deep learning for embodied vision navigation: A survey. arXiv preprint arXiv:2108.04097, [Online], pp. Available from: https://arxiv.org/abs/2108.04097 . [Accessed 12 October 2025].

# BIBLIOGRAPHY


Laskin, M., Lee, K., Stooke, A., Pinto, L., Abbeel, P. and Srinivas, A. (2020) Reinforcement learning with augmented data. Advances in neural information processing systems, 33, pp.19884-19895, [Online]. Available from: https://proceedings.neurips.cc/paper_files/paper/2020/hash/e615c82aba461681ade82da2da38004a-Abstract.html .[Accessed 11 October 2025].

Ma, G., Wang, Z., Yuan, Z., Wang, X., Yuan, B. and Tao, D. (2025) A comprehensive survey of data augmentation in visual reinforcement learning. International Journal of Computer Vision, 133(4), pp. 865–892, [Online]. Available from: https://link.springer.com/article/10.1007/s11263-025-02472-w . [Accessed 2 January 2026].

Mondal, A.K., Jiang, Y., Mukherjee, S., Sun, Q., Wang, L. and Li, J. (2022) Equivariant representations for data-efficient reinforcement learning. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, MD: PMLR, pp. 15842–15854, [Online]. Available from: https://proceedings.mlr.press/v162/mondal22a.html . [Accessed 27 October 2025].

Sun, J., Fang, H., Xu, C., Zhang, Y. and Liang, Y. (2024) A survey of object goal navigation. IEEE Transactions on Automation Science and Engineering, 21(2), pp. 667–682, [Online]. Available from: https://ieeexplore.ieee.org/abstract/document/10475904/?casa_token=V60AdVR3BIkAAAAA:PgpFZ3tv3SxbSyPY0D0JUSCc5ktthiTH3HpQmNfsJL1k_o6-OZiGZ0z6aatzYCS2OzOc8xUF7Aliyg . [Accessed 16 October 2025].

Wang, D., Walters, R. and Platt, R. (2022) SO(2)-equivariant reinforcement learning. arXiv preprint arXiv:2203.04439. [Online]. Available from: https://arxiv.org/abs/2203.04439 . [Accessed 19 October 2025].

Wong, L.H.K., Kang, X., Bai, K. and Zhang, J. (2025) A survey of robotic navigation and manipulation with physics simulators in the era of embodied AI. arXiv preprint arXiv:2505.01458. [Online]. Available from: https://arxiv.org/abs/2505.01458 . [Accessed 2 January 2026].

Wu, W., Gao, C., Chen, J., Lin, K.Q., Meng, Q., Zhang, Y., Qiu, Y., Zhou, H. and Shou, M.Z. (2025) Reinforcement Learning for Large Model: A Survey. arXiv preprint arXiv:2508.08189. [Online]. Available from: https://arxiv.org/abs/2508.08189 . [Accessed 3 January 2026].

Yarats, D., Fergus, R., Lazaric, A. and Pinto, L. (2021) Mastering visual continuous control: Improved data-augmented reinforcement learning. arXiv preprint arXiv:2107.09645. [Online]. Available from: https://arxiv.org/abs/2107.09645 . [Accessed 23 October 2026].

Zhao, L., Howell, O., Zhu, X., Park, J.Y., Zhang, Z., Walters, R. and Wong, L.L. (2024) Equivariant action sampling for reinforcement learning and planning. arXiv preprint arXiv:2412.12237. [Online]. Available from: https://arxiv.org/abs/2412.12237 . [Accessed 15 October 2026].

# APPENDICES

## APPENDIX A: GANTT CHART

| ID | Task | Start Date | End Date | Duration | 2025 | | | | | | | | | | | | 2026 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | J | F | M | A | M | J | J | A | S | O | N | D | J |
| 1 | Proposal preparation | 29/09/2025 | 06/10/2025 | 8 days | | | | | | | | | | | | | |
| 2 | Proposal writing and submission | 06/10/2025 | 13/10/2025 | 7 days | | | | | | | | | | | | | |
| 3 | Literature review | 13/10/2025 | 27/10/ 2025 | 14 days | | | | | | | | | | | | | |
| 4 | Methodology | 28/10/2025 | 10/11/ 2025 | 14 days | | | | | | | | | | | | | |
| 5 | Baseline PPO implementation | 10/11/ 2025 | 16/11/ 2025 | 7 days | | | | | | | | | | | | | |
| 6 | Geometry prior integration | 16/11/ 2025 | 29/11/ 2025 | 13 days | | | | | | | | | | | | | |
| 7 | Evaluation | 29/11/ 2025 | 03/12/ 2025 | 4 days | | | | | | | | | | | | | |
| 8 | Refinement | 03/12/ 2025 | 08/12/2025 | 5 days | | | | | | | | | | | | | |
| 9 | Results analysis | 08/12/ 2025 | 15/12/ 2025 | 7 days | | | | | | | | | | | | | |
| 10 | Conclusion and first draft | 15/12/ 2025 | 27/12/2025 | 12 days | | | | | | | | | | | | | |
| 11 | First draft submission and feedback | 27/12/ 2025 | 05/01/ 2026 | 9 days | | | | | | | | | | | | | |
| 12 | Corrections and final submission | 06 Jan 2026 | 12/01/2026 | 7 days | | | | | | | | | | | | | |

## APPENDIX B: IMPLEMENTATION AND REPRODUCIBILITY ARTEFACTS

To ensure transparency and reproducibility, all implementation artefacts for this study are archived in an external reproducibility bundle. Due to space constraints, only representative code excerpts are included in the main dissertation; full implementations, trained models, logs, and evaluation scripts are provided externally.

**Artefact bundle**:

**Google Drive link**: https://drive.google.com/drive/folders/1LUK3WzfCdSiZol-B84MPwOWiKJAygjYT?usp=drive_link

## APPENDIX B1: AI2-THOR GYM ENVIRONMENT WRAPPER

The navigation environment is implemented in **thor_nav_env.py**, with a geometry-aware wrapper **thor_nav_geom_env.py**. The wrapper augments observations with low-dimensional geometric signals while preserving identical dynamics and reward functions.
**Location:** 02_code/

## APPENDIX B2: TRAINING SCRIPTS

All PPO training procedures use fixed hyperparameters and controlled seeds. Scripts include baseline, geometry-aware, curriculum, and stabilisation runs.

**Location:** 02_code/

## APPENDIX B3: POLICY ARCHITECTURE

Geometry-aware policies incorporate pose-conditioned feature fusion prior to policy and value heads.
**Location:** 02_code/

## APPENDIX B4: CURRICULUM SAMPLING

Uniform and weighted curriculum strategies used for FP1–FP5 training are implemented in dedicated training scripts.
**Location:** 02_code/

## APPENDIX B5: EVALUATION AND OOD TESTING

Deterministic evaluation and OOD testing scripts generate all reported results.
**Location:** 02_code/

## APPENDIX B6: MODELS, LOGS, AND REPRODUCTION

Trained checkpoints, TensorBoard logs, and a concise reproduction guide are provided to enable independent verification.

**Locations:**03_models/, 04_logs/, 07_reproduce/