# Adaptive Context Length Optimization with Low-Frequency Truncation for Multi-Agent Reinforcement Learning

**Wenchang Duan**[1], **Yaoliang Yu**[2], **Jiwan He**[1], **Yi Shi**[1*]
[1]Shanghai Jiao Tong University [2]University of Waterloo
{duanwenchang, kih020429, yishi}@sjtu.edu.cn
yaoliang.yu@uwaterloo.ca

## Abstract

Recently, deep multi-agent reinforcement learning (MARL) has demonstrated promising performance for solving challenging tasks, such as long-term dependencies and non-Markovian environments. Its success is partly attributed to conditioning policies on large fixed context length. However, such large fixed context lengths may lead to limited exploration efficiency and redundant information. In this paper, we propose a novel MARL framework to obtain adaptive and effective contextual information. Specifically, we design a central agent that dynamically optimizes context length via temporal gradient analysis, enhancing exploration to facilitate convergence to global optima in MARL. Furthermore, to enhance the adaptive optimization capability of the context length, we present an efficient input representation for the central agent, which effectively filters redundant information. By leveraging a Fourier-based low-frequency truncation method, we extract global temporal trends across decentralized agents, providing an effective and efficient representation of the MARL environment. Extensive experiments demonstrate that the proposed method achieves state-of-the-art (SOTA) performance on long-term dependency tasks, including PettingZoo, MiniGrid, Google Research Football (GRF), and StarCraft Multi-Agent Challenge v2 (SMACv2).

## 1 Introduction

Multi-agent reinforcement learning (MARL) has drawn increasing interest in recent years, which provides a promise for facing many complex real-world challenging problems such as transportation management [1], robot control [2], and finance [3]. However, due to the long-term dependencies and non-rigorous Markovianity of complex tasks, contextual information is introduced to assist policy making [4, 5, 6]. Accordingly, this places a substantial demand on how to leverage contextual information and to what extent [7, 8].

Existing methods are mostly applied to single-agent reinforcement learning (RL), where contextual information performs reasonably well in simple tasks [9, 10, 11]. In comparison, multi-agent reinforcement learning (MARL) involves significantly more complex tasks [12, 13], where relying solely on short context lengths or individual observations often results in suboptimal performance. To address this, one natural approach is to extend the context length. However, the expansion of the context length leads to two significant challenges: the first is an increase in necessary computation, and the second is the difficulty of high dimensionality of the input representation and generalization [14].

*Corresponding author: Yi Shi (yishi@sjtu.edu.cn).
†Project code is available at: https://github.com/duanwenchang/ACL-LFT.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

To address the challenge of increasing computation, references [9, 15] optimized the needed context length and the utilization efficiency; references [16, 17] adopted parallel computation; and reference [18] enhanced the performance of modern hardware. However, the above methods involve a long-time pre-training process, and eventually only obtain static context length. These static context lengths are difficult to adapt to changing environments, which potentially leads to suboptimal solutions or inefficient use of computational resources.

The challenge of input representation and generalization remains unresolved in the field of multi-agent reinforcement learning (MARL) [14]. While recent improvements in processing long sequences that came with attention models significantly alleviate requirements for generalization, an effective representation is still crucial for obtaining optimal contextual information [19, 20, 21]. Regarding the fields where the contextual information is also significant, natural language processing (NLP) typically involves leveraging large language models (LLMs) to autonomously learn and generate prompts, which does not align well with the principles of MARL [22, 23]. Therefore, a tailored input representation for MARL is required.

According to the aforementioned analysis, we propose an adaptive context length optimization with low-frequency truncation (ACL-LFT) for MARL. Specifically, a senior central agent is introduced to adaptively optimize context length, and a tailored attention-based reward is designed to align with the central agent. Via real-time interacting with environment, the central agent determines the optimal context length to address the challenge of increasing computation. Besides, we apply the Fourier transform to map the data from the time domain to the frequency domain, facilitating more effective redundancy filtering compared to direct processing in the time domain. Via truncating the low-frequency band, we obtain an effective input representation for the central agent, which captures the global temporal trends from the decentralized agents. With the above designs, our method effectively solves the dual challenges of increasing context length, achieving efficient leverage of the contextual information.

We benchmark the proposed method across various environments including Sample Spread in PettingZoo [24]; MiniGrid Soccer Game in OpenAI Gym [25]; Academy 3 vs 1 with Keeper, and Academy Counterattack-Hard in Google Research Football (GRF) [26]; *3s5z_vs_3s6z*, *5m_vs_6m*, and *corridor* in StarCraft Multi-Agent Challenge v2 (SMACv2) environments [27]. Combined with several types of experiments, including state-of-the-art (SOTA) sequence processing algorithms and different fixed-length methods, we show that the proposed method significantly enhances the performance of the baseline algorithm in changing environments. The main contributions of this paper are summarized as follows:

- To the best of our knowledge, ACL-LFT is the first framework to systematically address the dual challenges of increasing context length in MARL. Equipped with the central agent, our framework achieves adaptive and efficient leverage of contextual information to enhance the decision-making of decentralized agents. Additionally, we present a theorem to theoretically demonstrate the superior performance of adaptive context length over static ones in dynamic environments.
- We propose a novel Fourier-based low-frequency truncation to obtain the global temporal trends from context, effectively addressing the challenge of representing the MARL environment and providing an efficient input for the central agent.
- We empirically demonstrate that the proposed method outperforms SOTA sequence processing algorithms across various long-term dependency environments. We also provide experimental results to demonstrate the superior performance of the proposed method over different fixed lengths in dynamic environments.

## 2 Preliminaries

### 2.1 Decentralized Partially Observable Markov Decision Process with Historical Information

The Decentralized Partially Observable Markov Decision Process with historical information is defined as a tuple $\mathcal{M} = (N, S, A, P, R, \gamma)$, where $N$ is the set of $n$ agents, $S$ denotes the global state space, and $A$ represents the joint action space. At time $t$, the environment evolves according to the transition function $P(s'_t|s_t, a_t)$, which specifies the probability of reaching the next state $s'_t$ given the current global state $s_t = \{s_t^1, s_t^2, \cdots, s_t^n\}, s \in S$ and the joint action $a_t = \{a_t^1, a_t^2, \cdots, a_t^n\}$. The

environment then produces a global reward $r_t = R(s_t, a_t)$. In this decentralized setting, each agent follows a local policy $\pi$ that seeks to maximize the expected cumulative discounted reward, given by $J(\pi) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \,\Big|\, \pi\Big]$, where $\gamma \in [0, 1)$ is a discount factor that balances the importance of immediate versus future rewards. The classical Markov property assumes that state transitions depend only on the current state and action.

However, in decentralized partially observable environments, agents cannot directly access the full global state $s_t$; instead, they rely on local observations that are incomplete and noisy. In many scenarios, the assumption that decision-making can be based solely on the current observation is insufficient. To address this, the framework incorporates **historical information** to approximate the underlying dynamics and capture long-term dependencies. Formally, the extended model can be written as $\widetilde{\mathcal{M}} = (N, \widetilde{S}, \widetilde{A}, \widetilde{P}, \widetilde{R}, \gamma)$, where $\widetilde{S}$ includes not only the current state $S$ but also a contextual information space $S^{-1}$ representing observation histories, and $\widetilde{A} = A$. At time $t$, the transition is expressed as

$$P(\widetilde{s_t'}|\widetilde{s}_t, \widetilde{a}_t) = P\big(\widetilde{s}' = s_t' \cup s_t'^{,-1} \,\big|\, \widetilde{s}_t = s_t \cup s_t^{-1}, \widetilde{a}_t = a_t\big),$$

and the reward is given by

$$R(\widetilde{s}_t, \widetilde{a}_t) = R(\widetilde{s}_t = s_t \cup s_t^{-1}, \widetilde{a}_t = a_t).$$

Unlike the standard Markov Decision Process, this extended Decentralized Partially Observable framework enables modeling of complex dynamic environments with long-term temporal dependencies, where leveraging historical information is essential for effective coordination among agents.

### 2.2 The Fourier Transform and Littlewood–Paley Theory

The Fourier transform provides a fundamental tool for analyzing functions in the frequency domain by decomposing signals into their constituent frequency components. Formally, for a function $f \in L^1(\mathbb{R}^d)$, the Fourier transform is defined as:

$$\mathcal{F}f(\xi) = \hat{f}(\xi) = \int_{\mathbb{R}^d} e^{-i(x|\xi)} f(x)\, dx, \tag{1}$$

where $(x|\xi)$ denotes the inner product in $\mathbb{R}^d$. As a continuous linear map from $L^1(\mathbb{R}^d)$ into $L^\infty(\mathbb{R}^d)$, it satisfies $|\hat{f}(\xi)| \leq \|f\|_{L^1}$, ensuring boundedness in the transformed domain. Besides, for any function $\varphi \in L^1$ and an automorphism $L$ on $\mathbb{R}^d$, the transformation obeys:

$$\mathcal{F}(\varphi \circ L) = \frac{1}{|\det L|} \hat{\varphi} \circ L^{-1}. \tag{2}$$

By mapping state representations from the time domain to the frequency domain, the Fourier transform captures underlying structural patterns, where low-frequency components effectively encode global trends while filtering high-frequency noise [28][29].

Littlewood–Paley theory provides a decomposition that functions or distributions are easier to deal with if split into countable sums of smooth functions whose Fourier transforms are compactly supported in a ball or an annulus [30]. In the complex non-Markovianity environments, such decomposition renders a localization procedure in frequency space, which the derivatives act almost as homotheties on distributions. This property establishes fundamental bounds on the behavior of derivatives in different $L^p$ spaces, leading to the following Bernstein inequalities. Let $\mathcal{C}$ be an annulus and $B$ a ball. There exists a constant $C$ such that for any nonnegative integer $k$, any pair $(p, q) \in [1, \infty]^2$ with $q \geq p \geq 1$, and any function $u \in L^p$, the following holds:

$$Supp \quad \hat{u} \subset \lambda B \Rightarrow \|D^k u\|_{L^q} \overset{\text{def}}{=} \sup_{|\alpha|=k} \|\partial^\alpha u\|_{L^q} \leq C^{k+1} \lambda^{k+d(\frac{1}{p}-\frac{1}{q})} \|u\|_{L^p}, \tag{3}$$

$$Supp \quad \hat{u} \subset \lambda \mathcal{C} \Rightarrow C^{-k-1} \lambda^k \|u\|_{L^p} \leq \|D^k u\|_{L^p} \leq C^{k+1} \lambda^k \|u\|_{L^p}. \tag{4}$$

The above inequalities highlight a key property: if a function's Fourier spectrum is restricted within frequency $\delta$, its $\alpha$-th order derivative amplifies high-frequency components by a factor of $\delta^{|\alpha|}$. This property enables an efficient truncation of low-frequency information, which serves as an effective representation of contextual information while preserving stability in the decision-making process.

## 3 Methodology

In this section, we propose a novel MARL framework for obtaining adaptive and effective contextual information, which systematically tackles the dual challenges of increasing context length in MARL. The overall framework is shown in Fig. 1, which is comprised of three main components: (1) the Fourier-based low-frequency truncation module, (2) a central agent of adaptive select contextual information, and (3) the structure of learning with spatio-temporal decoupling.

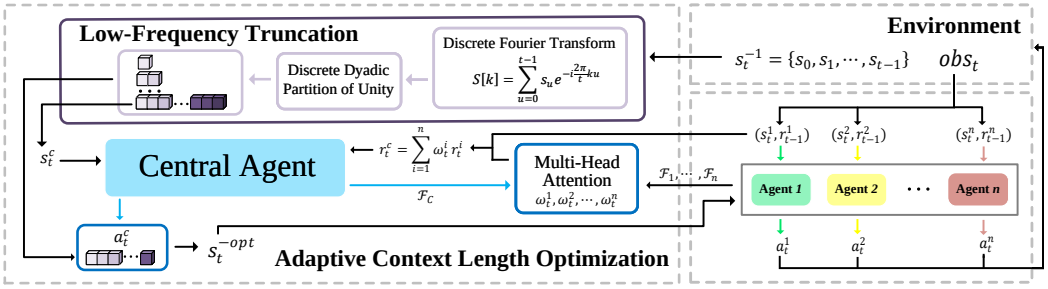



Figure 1: Schematics of our ACL-LFT. At each time $t$, the historical state $s_t^{-1}$ is first processed via the Fourier-based low-frequency truncation module. The central agent leverages the truncated information $s_t^c$ as input and then adaptively optimizes the context length. Subsequently, the decentralized agents then integrate the optimized contextual information $s_t^{-opt}$ with the current state to achieve decision-making.

### 3.1 Fourier-based Low-Frequency Truncation

To achieve an efficient representation of the MARL environment and enhance its effectiveness as input for the central agent, we introduce a low-frequency truncation method. By filtering out high-frequency fluctuations while preserving low-frequency parts, this method captures global temporal trends across decentralized agents and provides a more stable basis for downstream decision-making. Specifically, given the discrete nature of historical state data $s_t^{-1} = \{s_j\}_{j=0}^{t-1}$, we first leverage the Discrete Fourier Transform (DFT) to convert time domain data to frequency domain data:

$$S[k] = \sum_{u=0}^{t-1} s_u e^{\frac{-i2\pi ku}{t}}, \quad k = 0, 1, \ldots, t-1, \tag{5}$$

where $S[k]$ represents the frequency-domain coefficient corresponding to frequency index $k$, while each coefficient encodes a particular oscillatory component of the historical sequence. For real-valued signals, the DFT exhibits conjugate symmetry: $S[k] = S[t-k]^*$, reflecting periodicity in the frequency domain. This transformation effectively disentangles different frequency components of the input sequence, allowing for a more interpretable and structured representation of historical states.

Building upon the Littlewood–Paley theory, we then introduce the Dyadic Partition of Unity method and extend it to the discrete space to truncate the low-frequency information. The Dyadic Partition of Unity method for measurable functions is provided in Appendix A.1. We extend this method to adapt discrete frequency domain historical states. Specifically, we aim for the sum of the window functions to approximate unity across the entire frequency domain:

$$X[k] + \sum_{j=0}^{J-1-m} \Phi_j[k] \approx 1, \quad \forall k = 0, 1, \ldots, N-1, \tag{6}$$

where $X[k]$ is a low-pass window function that retains only the low-frequency components. $\Phi_j[k]$ represents band-pass window functions that separate different frequency bands in a dyadic manner. $J$ is the maximum decomposition level, and for simplicity, we assume that $t = 2^J$. $m$ is a tunable parameter that determines the truncation frequency, ensuring that $2^m < t/2$.

Within this method, the low-frequency region is defined for $k \leq 2^m$ or $k \geq N - 2^m$, where we set $X[k] = 1$ and ensure that $\Phi_j[k] = 0$ for all $j$, thereby preserving the low-frequency information $s_c$.

In the band-pass regions, corresponding to frequency indices satisfying $2^{j+m} \leq k < 2^{j+m+1}$ (or their symmetric counterparts), a single window function $\Phi_j[k]$ is activated with a value of 1, while all other $\Phi_{j\prime}[k]$ remain zero for $j\prime \neq j$, ensuring a well-defined partitioning of frequency bands. At transition points, such as $k = 2^{j+m}$, minor gaps may arise due to non-overlapping support. But as the signal length $t$ increases, the gaps become negligible with an error proportionally decreasing as $O(1/t)$, thereby maintaining a stable approximation of equation 6.

The details of Dyadic Partition of Unity in Discrete Form and its rigorous proof are provided in Appendix A.2. By leveraging low-frequency truncation, the above method effectively captures the global temporal trends across decentralized agents, reducing the redundancy of contextual information and serving as an efficient input representation for the subsequent central agent.

### 3.2 The Central Agent of Adaptive Contextual Information Selection

The central agent in our framework serves as a global information processor, adaptively determining the optimal contextual information length for decentralized agents. It is designed to process and analyze only historical information, without directly handling the current state. Specifically, its decision-making process is structured around three key components: state representation, action space, and reward formulation.

Firstly, the state of the central agent is derived through the Fourier-based low-frequency truncation module, which is elaborated in section 3.1. For the discrete historical states $s_t^{-1} = \{s_j\}_{j=0}^{t-1}$ at time $t$, this module extracts the truncated representation $s_t^c$, which effectively represents the global temporal trends across decentralized agents.

Then, the action space of the central agent $A_c$ is defined as the selection of different low-frequency truncation levels, given by:

$$a_t^c \in A_c = \{m_1, m_2, \ldots, m_M\}, \tag{7}$$

where $M$ represents the dimension of $A_c$, and each action $m_i$ corresponds to a different range of preserved low-frequency bands, with each band representing temporal trends with varying degrees of long-term dependency. Given the selected truncation level $m_i$, the corresponding optimal contextual information $s_t^{-opt}$ is obtained by truncation domain.

Finally, to guide its adaptation process, the reward of the central agent is tailored via the multi-head attention mechanism, which weights the influence of decentralized agents. Specifically, the value function estimates and the policy distributions of decentralized agents serve as keys, while the value function estimate and the policy distributions of the central agent serve as the query. We denote the concatenated representation of the value estimate and the policy distribution of agent $i$ as $\mathcal{F}_i$, and that of the central agent as $\mathcal{F}_c$. For each attention head $g$, $\mathcal{F}_i$ and $\mathcal{F}_c$ are projected into the query and key spaces via transformation matrices $W_Q^g$ and $W_K^g$:

$$\mathcal{Q}_c^g = W_Q^g \mathcal{F}_c, \quad \mathcal{K}_i^g = W_K^g \mathcal{F}_i. \tag{8}$$

The attention weight assigned to each decentralized agent is then computed as:

$$\omega_i^g = \frac{\exp\left(\frac{\mathcal{Q}_c^g \cdot (\mathcal{K}_i^g)^T}{\sqrt{d_k}}\right)}{\sum_i \exp\left(\frac{\mathcal{Q}_c^g \cdot (\mathcal{K}_i^g)^T}{\sqrt{d_k}}\right)}, \tag{9}$$

where $d_k$ represents the dimensionality of the key matrix $\mathcal{K}$, ensuring numerical stability. The final attention weight for each agent at time $t$ is obtained by averaging across all heads $\omega_t^i = \frac{1}{head} \sum_{g=1}^{head} \omega_i^g$. At time $t$, for the weights $\{\omega_t^i\}_{i=1}^n$, the reward for the central agent is then derived as a weighted aggregation of the rewards of decentralized agents $r_t^i$:

$$r_t^c = \sum_{i=1}^{n} \omega_t^i r_t^i. \tag{10}$$

where $\sum_{i=1}^n \omega_t^i = 1$.

Combined with these three components, the parameters of the central agent are updated using gradient-based optimization with advantage estimation. The value function $V(s_t^c)$ is trained to approximate

the expected return through the temporal difference error:

$$\delta_t^c = r_t^c + \gamma V(s_{t+1}^c) - V(s_t^c), \tag{11}$$

where $s_{t+1}^c$ denotes the next state. The policy parameters $\theta$ are then adjusted by maximizing the advantage-weighted objective:

$$\theta \leftarrow \theta + \zeta \nabla_\theta \log \pi(a_t^c | s_t^c) \delta_t^c, \tag{12}$$

while simultaneously minimizing the value function error $\|\delta_t^c\|^2$ through gradient descent.

Building upon the above design, the central agent achieves adaptive optimization of the context length, ensuring that decentralized agents receive the optimal contextual information $s_t^{-opt}$.

Furthermore, to theoretically establish a long-term advantage lower bound of the proposed method over fixed-length methods, we present Theorem 1.

**Theorem 1 (Long-Term Advantage Lower Bound of Adaptive Length)** *: At time $t$, let $L_{adap}$ be the adaptive context length, $L_{fix}$ be the fixed context length, and the mutual information loss of $L$ be denoted as $\mathcal{L}_t(L)$. The expected cumulative reward difference between adaptive and fixed context length satisfies the following regret bound:*

$$\begin{aligned} \sum_{t=1}^{T} \left(\mathcal{L}_t(L_{fix}) - \mathcal{L}_t(L_{adap})\right) &\geq \Omega(T) - O(T^\alpha) \\ &= \Omega(T) \quad (\textit{when } T \textit{ is sufficiently large}) \end{aligned} \tag{13}$$

*where $0 \leq \alpha < 1$, with $\alpha$ being a non-deterministic parameter whose formal definition is provided in Appendix B.*

This theorem demonstrates the long-term advantage of adaptive length policies with the increasingly unstable environment. The result suggests that adaptively adjusting the context length enables more effective information retention over time, leading to significantly lower regret accumulation. The details and proof are provided in Appendix B.

### 3.3 Structure of Learning with Spatio-Temporal Decoupling

In this section, we discuss how to leverage the spatio-temporal decoupling to train the proposed learning framework. Specifically, the training process is divided into two components: the central agent, which is responsible for selecting the optimal contextual information $s_t^{-opt}$; and the decentralized agents, which leverage this information and their current state $s_t$ to optimize their policies. In this framework, the central agent is trained independently to optimize the temporal information component, while the decentralized agents undergo joint training to refine their policies with filtered temporal information and their spatial information. The policy-making and training process of the ACL-LFT algorithm is illustrated in the pseudocode provided in Appendix C.3.

The global optimization objective of the framework is given by the expected sum of discounted rewards over time for decentralized agents:

$$J_i(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_i(\widetilde{s}_t, \widetilde{a}_t) \mid \pi\right] \tag{14}$$

Besides, the central agent's objective is to ensure goal alignment with decentralized agents through gradient correlation. Specifically, the central objective is defined as

$$\nabla_\pi J_c(\pi) = \sum_{t=0}^{\infty} \gamma^t \sum_{i=1}^{n} \omega_t^i \nabla_\pi \mathbb{E}[R_i \mid \pi]. \tag{15}$$

$$\nabla_\pi \left(\sum_{j=1}^{n} J_j(\pi)\right) = \sum_{t=0}^{\infty} \gamma^t \sum_{j=1}^{n} \nabla_\pi \mathbb{E}[R_j \mid \pi]. \tag{16}$$

Goal alignment holds when the gradients of the central and decentralized objectives are positively correlated, that is,

$$\left\langle \nabla_\pi J_c(\pi), \nabla_\pi \sum_{j=1}^{n} J_j(\pi) \right\rangle > 0 \quad \text{when } \omega_t^i > 0.$$

This formulation ensures that the central agent's optimization direction remains consistent with the aggregated learning objectives of decentralized agents, thereby enhancing the stability and cooperative efficiency of the overall framework.

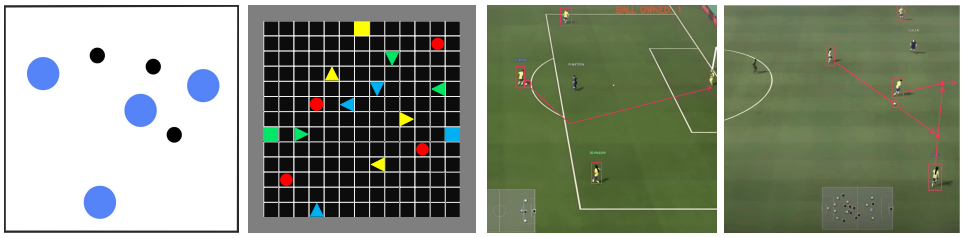
By structuring training in this manner, our framework mitigates the challenge of the excessively large parameter search space that typically arises from the joint optimization of both contextual and current information, a factor known to hinder convergence. As a result, our framework not only accelerates the learning process but also ensures that agents can efficiently leverage temporal trends, thereby improving decision-making in complex multi-agent environments.

## 4 Experiments and Analysis

This section evaluates the proposed ACL-LFT framework across various MARL environments. Section 4.1 outlines the experimental setup and baselines. Section 4.2 and Section 4.3 compare ACL-LFT with sequence processing and fixed-length methods. Section 4.4 and Section 4.5 present ablation and case analyses. Finally, Section 4.6 examines the decentralized setting without cross-agent information sharing.

### 4.1 Experiment Setup

**Environments** We consider various tasks, including Sample Spread in PettingZoo [24], MiniGrid Soccer Game in OpenAI Gym [25], Academy 3 vs 1 with Keeper, and Academy Counterattack-Hard in Google Research Football (GRF) [26]. The overview of environments is shown in Fig. 2, while the details of environments and their reward design are provided in Appendix C.1. All experiments are implemented based on the Multi-Agent Proximal Policy Optimization (MAPPO) algorithm [31]. Furthermore, to verify the effectiveness of our proposed method under both complex and large-scale scenarios, as well as to analyze the impact of removing the MAPPO backbone, we conduct additional experiments based on MAPPO, QMIX [32], and QPLEX [33] in the StarCraft Multi-Agent Challenge v2 (SMACv2) environments [27], including *3s5z_vs_3s6z*, *5m_vs_6m*, and *corridor*. The detailed experimental results are presented in Appendix C.4.



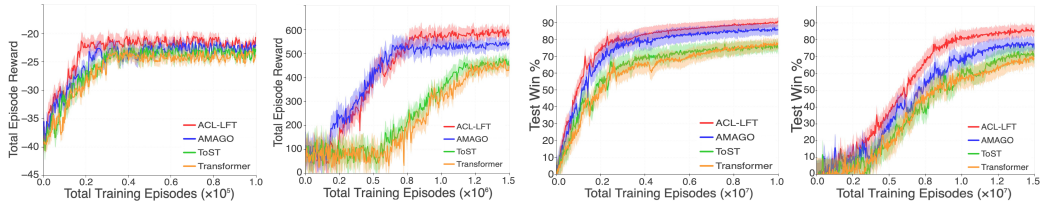
(a) Sample Spread (b) Minigrid Soccer Game (c) 3 vs 1 with Keeper (d) Counterattack-Hard

Figure 2: Sample Spread (a) is a search game where agents learn to cover all the landmarks while avoiding collisions. Minigrid Soccer Game (b) is a 15×15 environment where agents (triangles) earn rewards by kicking the ball (circle) into same-colored goalmouths (squares). Academy 3 vs 1 with Keeper (c) is a scenario where three offensive agents attempt to score against one defender and a goalkeeper. Academy Counterattack-Hard (d) is a scenario where four agents must execute a rapid counterattack while avoiding defenders.

Given the varying maximum episode lengths across different environments, we adapt the context length accordingly for each case. Specifically, the maximum episode steps for Sample Spread, MiniGrid Soccer Game, Academy 3 vs 1 with Keeper, and Academy Counterattack-Hard are 25, 512, 400, and 400, respectively. Therefore, the corresponding context lengths are set to 4, 64, 64, and 64 steps. These values define the maximum selectable context lengths for the proposed method, consistent with the central agent's input dimension. Further details are provided in Appendix C.2.

**Baselines** We first benchmark the sequence processing algorithms, including Transformer [34], Token Statistics Transformer (ToST) [35], and AMAGO [36]. The introduction of these methods is provided in Appendix C.2. Then, we benchmark the proposed method against different fixed context lengths.

### 4.2 Performance Comparison with Sequence Processing Methods




(a) Sample Spread (b) Minigrid Soccer Game (c) 3 vs 1 with Keeper (d) Counterattack-Hard

Figure 3: Performance Comparison with Sequence Processing Methods in Four Environments

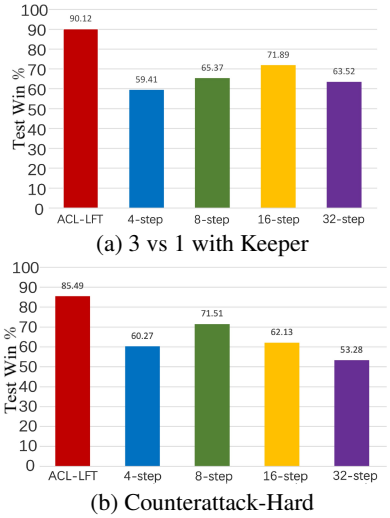
In this section, we benchmark the proposed method against Transformer, ToST, and AMAGO. As shown in Fig. 3, the performances are depicted via data from every 100 episodes and averaged over 5 seeds. It is seen that the proposed method outperforms in all scenarios. Specifically, Transformer and ToST need more time to explore sophisticated policies and demonstrate large oscillations in the exploration process. AMAGO demonstrates strong performance during the policy exploration phase, owing to its effective handling of long sequences in parallel. However, due to the fact that its contextual information is of fixed length, which tends to have a lot of noise, it performs poorly compared to the proposed method after convergence.

In the Sample Spread, Academy 3 vs 1 with Keeper and Academy Counterattack-Hard, the proposed method demonstrates the fastest exploration efficiency and consistently achieves the highest post-convergence performance. Notably, as the complexity of the scenarios increases—from Sample Spread to Academy 3 vs 1 with Keeper and Academy Counterattack-Hard—the performance gap between the proposed method and the baseline methods becomes more evident. In the Minigrid Soccer Game, the proposed method and AMAGO exhibit comparable exploration efficiency; however, AMAGO converges prematurely and fails to achieve strong final results. In contrast, the proposed method utilizes low-frequency truncation of historical information, significantly mitigating the impact of redundant data. By adaptively selecting the optimal context length, the proposed method achieves superior performance across all environments.

### 4.3 Performance Comparison with Fixed-Length

In section 3.2, we presented the Theorem 1, which demonstrates the long-term advantage of adaptive length policies over fixed-length. In this section, we benchmark the proposed method against different fixed context lengths (8, 16, 32, and 64 steps) in Academy 3 vs 1 with Keeper and Academy Counterattack-Hard.

We computed the average performance of the proposed method and four fixed-length methods during two relatively stable periods after convergence. Specifically, the results were averaged from the 0.9 millionth to the 1 millionth episode in the Academy 3 vs 1 with Keeper, and from the 1.4 millionth to the 1.5 millionth episode in the Academy Counterattack-Hard. As shown in Fig. 4, the proposed method significantly outperforms all fixed-length methods, further demonstrating the effectiveness and efficiency of adaptive context length optimization. Notably, among the fixed-length methods, the 16-step and 8-step show the best performance in Academy 3 vs 1 with Keeper and Academy Counterattack-Hard, respectively. These findings suggest that longer context lengths do not necessarily lead to better performance, as excessive historical




(a) 3 vs 1 with Keeper

(b) Counterattack-Hard

Figure 4: Performance Comparison with Different Fixed Lengths on GRF

information may introduce significant noise. This observation further underscores the critical role of our low-frequency truncation in enhancing overall performance.

### 4.4 Ablation Experiments

To understand the contribution of each component in the proposed ACL-LFT framework, we carry out ablation studies to test the contribution of adaptive context length (ACL) and low-frequency truncation (LFT). Specifically, we utilize the best-performing fixed-length configurations in the Academy 3 vs 1 with Keeper (32-step) and Academy Counterattack-Hard (16-step) environments. These configurations are applied to evaluate ACL-LFT-NO-ACL and ACL-LFT-Raw, which test the performance of the methods without ACL and without both ACL and LFT, respectively. Furthermore, the ACL-LFT-NO-LFT is input with 32 and 32 steps, which align with the maximum step that can be selected by the ACL-LFT.

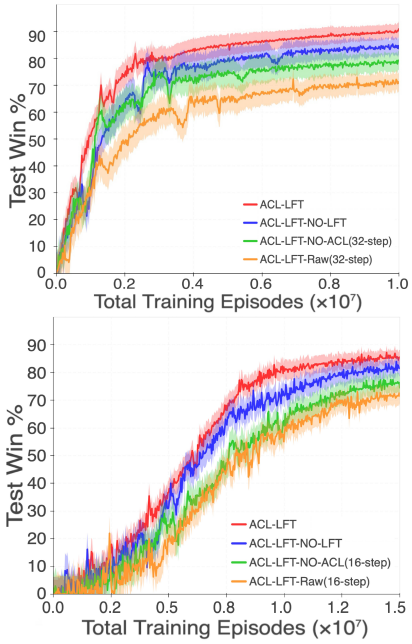



Figure 5: Comparison of ablation studies: (a) 3 vs 1 with Keeper; (b) Counterattack-Hard.

As shown in Fig. 5, the most impact on performance is the ablation of ACL, which further demonstrates the significant effect of adaptive context length optimizing. Additionally, the results reveal that ACL-LFT-NO-ACL outperforms ACL-LFT-Raw, indicating that low-frequency truncation (LFT) contributes notably to the overall performance of the proposed framework. In conclusion, the results highlight the critical importance of both ACL and LFT in enhancing the efficiency and effectiveness of the overall framework. Moreover, they demonstrate these two components complement each other, contributing synergistically to performance improvement, especially in environments with varying complexities and temporal dynamics.

### 4.5 Case Study

To intuitively demonstrate the performance benefits brought by the adaptive context length mechanism, we conduct a case study on the MiniGrid Soccer Game. This environment features a tailored reward function, which is highly sensitive to the agent's ability to leverage historical information for effective path planning and cooperative strategies. Specifically, we set the maximum step of environment to 64. Accordingly, all baseline methods use a fixed context length of 16, which aligns with the maximum selectable length for the proposed method.

Table 1: Step-Reward Comparison on MiniGrid Soccer Game

| Step | ACL-LFT | Transformer | ToST | AMAGO |
|---|---|---|---|---|
| 0 | 0.00 (0) | 0.00 | 0.00 | 0.00 |
| 5 | 2.71 (8) | 2.96 | 0.00 | 0.00 |
| 10 | 1.52 (16) | 1.73 | 2.17 | 2.63 |
| 15 | 1.77 (8) | 1.94 | 1.59 | 2.08 |
| 20 | 2.41 (4) | 1.65 | 1.26 | 1.54 |
| 25 | 3.19 (4) | 2.36 | 1.85 | 1.93 |
| 30 | 3.85 (2) | 2.07 | 2.26 | 2.45 |
| 35 | 3.30 (8) | 2.85 | 2.51 | 2.16 |
| 40 | 4.06 (2) | 3.57 | 2.90 | 2.48 |
| **41** | **14.31 (1)** | 3.45 | 2.97 | 2.65 |
| 45 | / | 4.12 | 3.36 | 3.03 |
| **47** | / | **13.98** | 3.52 | 3.41 |
| 50 | / | / | 3.79 | 3.66 |
| 55 | / | / | 4.09 | 3.85 |
| **56** | / | / | **14.25** | 4.02 |
| **59** | / | / | / | **13.62** |

As shown in Table 1, we record the reward value every 5 steps and highlight the time step at which the first goal is achieved in bold. The numbers in parentheses indicate the context length

adaptively selected by ACL-LFT at each time step. The results show that ACL-LFT quickly adjusts its context length after obtaining positive rewards (e.g., step = 15, 20), enabling timely re-planning to avoid inefficient exploration, while other methods remain in the aimless exploration phase. Notably, ACL-LFT dynamically selects shorter yet effective context lengths (e.g., length 2 at step 40), significantly improving path efficiency and ultimately achieving a goal by step 41. In contrast, methods with fixed-length contexts adapt more slowly, require longer to identify viable paths. This indicates that ACL-LFT enhances exploration efficiency and mitigates the impact of redundant information via adaptive context length optimization, thereby achieving superior performance.

### 4.6 Ablation on the Absence of Cross-Agent Historical Information

To further examine the influence of centralized sequence processing and verify that our method does not rely on cross-agent information sharing, we conduct an additional ablation study in which each agent can only access its own local historical observations and actions, without any global or inter-agent historical information. In this variant, the central agent is disabled from aggregating histories across agents; instead, it independently processes the local sequences for each agent, ensuring that no centralized communication channel exists during decision-making.

Table 2: Performance comparison without cross-agent historical information.

| Task | AMAGO | Mamba | **ACL-LFT** |
|---|---|---|---|
| 3s5z vs 3s6z | $76.1 \pm 2.9$ | $72.6 \pm 3.2$ | $\mathbf{78.9 \pm 2.8}$ |
| 5m_vs_6m | $48.1 \pm 4.0$ | $46.2 \pm 4.5$ | $\mathbf{52.7 \pm 4.2}$ |
| corridor | $74.3 \pm 4.8$ | $69.0 \pm 5.9$ | $\mathbf{77.9 \pm 5.3}$ |

As shown in Table 2, ACL-LFT consistently outperforms existing temporal modeling methods such as AMAGO and Mamba even when no cross-agent historical information is available. This confirms that the performance improvement of ACL-LFT originates from its proposed low-frequency temporal representation and adaptive contextual-length optimization rather than from any implicit inter-agent information sharing. Moreover, this ablation validates that ACL-LFT preserves the partially observable nature of the SMACv2 environment and remains effective under purely decentralized historical settings, demonstrating the soundness and generality of our framework.

## 5 Conclusion

To systematically address the dual challenges of increasing context length in MARL, we propose an adaptive context length optimization with low-frequency truncation (ACL-LFT) for MARL. The proposed method adaptively optimizes context length via a central agent. Equipped with a Fourier-based low-frequency truncation, we address the challenge of representing the MARL environment and provide an efficient input for the central agent. The experimental results demonstrate the proposed method significantly enhances the performance of the baseline algorithm in changing environments. In addition, we demonstrate both theoretically and experimentally the long-term advantage of adaptive context length over fixed-length.

## References


[1] Yu Han, Meng Wang, and Ludovic Leclercq. Leveraging reinforcement learning for dynamic traffic control: A survey and challenges for field implementation. *Communications in Transportation Research*, 3:100104, 2023.

[2] Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martín-Martín, and Peter Stone. Deep reinforcement learning for robotics: A survey of real-world successes. *Annual Review of Control, Robotics, and Autonomous Systems*, 8, 2024.

[3] Noella Nazareth and Yeruva Venkata Ramana Reddy. Financial applications of machine learning: A literature review. *Expert Systems with Applications*, 219:119640, 2023.

[4] Vinzenz Thoma, Barna Pásztor, Andreas Krause, Giorgia Ramponi, and Yifan Hu. Contextual bilevel reinforcement learning for incentive alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[5] Guy Tennenholtz, Nadav Merlis, Lior Shani, Martin Mladenov, and Craig Boutilier. Reinforcement learning with history dependent dynamic contexts. In *International Conference on Machine Learning*, pages 34011–34053. PMLR, 2023.

[6] Andrew Wang, Andrew C Li, Toryn Q Klassen, Rodrigo Toro Icarte, and Sheila A McIlraith. Learning belief representations for partially observable deep rl. In *International Conference on Machine Learning*, pages 35970–35988. PMLR, 2023.

[7] Annie Wong, Thomas Bäck, Anna V Kononova, and Aske Plaat. Deep multiagent reinforcement learning: Challenges and directions. *Artificial Intelligence Review*, 56(6):5023–5056, 2023.

[8] Ashish Kumar Shakya, Gopinatha Pillai, and Sohom Chakrabarty. Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications*, 231:120495, 2023.

[9] Botao Hao, Tor Lattimore, and Chao Qin. Contextual information-directed sampling. In *International Conference on Machine Learning*, pages 8446–8464. PMLR, 2022.

[10] Chris Lu, Yannick Schroecker, Albert Gu, Emilio Parisotto, Jakob Foerster, Satinder Singh, and Feryal Behbahani. Structured state space models for in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36:47016–47031, 2023.

[11] Xudong Gong, Feng Dawei, Kele Xu, Bo Ding, and Huaimin Wang. Goal-conditioned on-policy reinforcement learning. *Advances in Neural Information Processing Systems*, 37:45975–46001, 2025.

[12] Xinran Li and Jun Zhang. Context-aware communication for multi-agent reinforcement learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 1156–1164, 2024.

[13] Muning Wen, Jakub Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang. Multi-agent reinforcement learning is a sequence modeling problem. *Advances in Neural Information Processing Systems*, 35:16509–16521, 2022.

[14] Matthew Riemer, Khimya Khetarpal, Janarthanan Rajendran, and Sarath Chandar. Balancing context length and mixing times for reinforcement learning at scale. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[15] Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4643–4663, 2024.

[16] ZiRui Wang, DENG Yue, Junfeng Long, and Yin Zhang. Parallelizing model-based reinforcement learning over the sequence length. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[17] Hao Liu and Pieter Abbeel. Blockwise parallel transformers for large context models. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Norman P Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th annual international symposium on computer architecture*, pages 1–12, 2017.

[19] Jonathan Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[20] Taku Yamagata, Ahmed Khalil, and Raul Santos-Rodriguez. Q-learning decision transformer: Leveraging dynamic programming for conditional sequence modelling in offline rl. In *International Conference on Machine Learning*, pages 38989–39007. PMLR, 2023.

[21] Shengchao Hu, Li Shen, Ya Zhang, Yixin Chen, and Dacheng Tao. On transforming reinforcement learning with transformers: The development trajectory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[22] Eghbal Hosseini and Evelina Fedorenko. Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023.

[24] Jordan Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:15032–15043, 2021.

[25] Arnaud Fickinger. Multi-agent gridworld environment for openai gym. `https://github.com/ArnaudFickinger/gym-multigrid`, 2020.

[26] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, et al. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4501–4510, 2020.

[27] Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Foerster, and Shimon Whiteson. Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36:37567–37593, 2023.

[28] Jihyeong Jeon, Jiwon Park, Chanhee Park, and U Kang. Frequant: A reinforcement-learning based adaptive portfolio optimization with multi-frequency decomposition. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1211–1221, 2024.

[29] Nate Gillman, Daksh Aggarwal, Michael Freeman, Saurabh Singh, and Chen Sun. Fourier head: Helping large language models learn complex probability distributions. *arXiv preprint arXiv:2410.22269*, 2024.

[30] Hajer Bahouri. *Fourier analysis and nonlinear partial differential equations*. Springer, 2011.

[31] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

[32] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

[33] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. Qplex: Duplex dueling multi-agent q-learning. *arXiv preprint arXiv:2008.01062*, 2020.

[34] A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. Attention is all you need. In *NIPS*, 2017.

[35] Ziyang Wu, Tianjiao Ding, Yifu Lu, Druv Pai, Jingyuan Zhang, Weida Wang, Yaodong Yu, Yi Ma, and Benjamin D Haeffele. Token statistics transformer: Linear-time attention via variational rate reduction. *arXiv preprint arXiv:2412.17810*, 2024.

[36] Jake Grigsby, Linxi Fan, and Yuke Zhu. Amago: Scalable in-context reinforcement learning for adaptive agents. In *The Twelfth International Conference on Learning Representations*, 2023.

[37] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[38] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[39] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[40] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

**Appendix Overview:** In this Appendix we provide important details that could not be included in the main text due to space constraints. First in Appendix A, we provide detailed proofs for the low-frequency truncation in discrete space. Next in Appendix B, we provide the proof of the approximately long-term advantage lower bound of the contextual information over static counterparts in dynamic environments. Finally, in Appendix C we provide additional details about our experiments discussed in the main text.

# A Proof of Low-Frequency Truncation in Discrete Space

## A.1 Dyadic Partition of Unity

Let $C$ be the annulus defined as $C = \{\xi \in \mathbb{R}^d \mid \frac{3}{4} \leq |\xi| \leq \frac{8}{3}\}$. There exist measurable functions $\chi$ and $\varphi$, taking values in the interval $[0, 1]$, belonging respectively to $\mathcal{D}(B(0, \frac{4}{3}))$ and $\mathcal{D}(C)$, such that:

$$\forall \xi \in \mathbb{R}^d, \quad \chi(\xi) + \sum_{j \geq 0} \varphi(2^{-j}\xi) = 1, \tag{17}$$

and for all:

$$\forall \xi \in \mathbb{R}^d \setminus \{0\}, \sum_{j \in \mathbb{Z}} \varphi(2^{-j}\xi) = 1. \tag{18}$$

These functions satisfy the disjoint support conditions:

$$|j - j\prime| \geq 2 \Rightarrow \operatorname{Supp} \varphi(2^{-j}\cdot) \cap \operatorname{Supp} \varphi(2^{-j'}\cdot) = \emptyset, \tag{19}$$

$$j \geq 1 \Rightarrow \operatorname{Supp} \chi \cap \operatorname{Supp} \varphi(2^{-j}\cdot) = \emptyset. \tag{20}$$

Defining the translated annulus $C$ as $Cdef = B(0, \frac{2}{3}) + C$, we note that $C$ remains an annulus and satisfies:

$$|j - j\prime| \geq 5 \Rightarrow 2^{j'}C \cap 2^{j}C = \emptyset. \tag{21}$$

Furthermore, the functions $\chi$ and $\varphi$ satisfy the bounds:

$$\forall \xi \in \mathbb{R}^d, \quad \frac{1}{2} \leq \chi^2(\xi) + \sum_{j \geq 0} \varphi^2(2^{-j}\xi) \leq 1, \tag{22}$$

$$\forall \xi \in \mathbb{R}^d \setminus \{0\}, \quad \sum_{j \in \mathbb{Z}} \varphi^2(2^{-j}\xi) \leq 1. \tag{23}$$

The aforementioned method establishes a smooth dyadic decomposition of frequency space using radial functions $\chi$ and $\varphi$ , ensuring a partition of unity while maintaining disjoint support conditions. The construction guarantees that the frequency space is effectively covered while avoiding excessive overlap, making it well-suited for applications in harmonic analysis and function space theory. The inequalities further confirm that the decomposition remains stable, with bounded sums ensuring proper reconstruction properties.

## A.2 Proof of Dyadic Partition of Unity in Discrete Form

To extend this formulation to the discrete setting, we consider a similar approach where the continuous frequency domain is replaced by a discrete grid. In this section, we first supplement the details of some important definitions. Then supply the essential proofs of our method.

To facilitate a dyadic-like decomposition in the discrete frequency domain, we introduce a set of window functions (the low-pass window function and the band-pass window functions) that partition the frequency spectrum into complementary regions. These functions ensure that different frequency components of a signal are captured separately, allowing for a structured analysis of its spectral content. Specifically, the low-pass window function $X[k]$ is defined as:

$$X[k] = \begin{cases} 1, & k \leq 2^m \text{ or } k \geq t - 2^m, \\ 0, & \text{otherwise}, \end{cases} \tag{24}$$

where $m$ is an integer satisfying $0 < m < J$ , which determines the cutoff for low-frequency retention. This function effectively captures the low-frequency trends of a signal while discarding

high-frequency components.
In contrast, the band-pass window functions $\Phi_j[k]$ isolate specific frequency bands and are defined as:

$$\Phi_j[k] = \begin{cases} 1, & 2^{j+m} \leq k < 2^{j+m+1} \text{ or } t - 2^{j+m+1} < k \leq t - 2^{j+m}, \\ 0, & \text{otherwise}, \end{cases} \tag{25}$$

for $j = 0, 1, \ldots, J - 1 - m$, each function spans a high-frequency dyadic band. These band-pass functions systematically cover the frequency spectrum, ensuring that different frequency components are separately analyzed while maintaining a structured partitioning. Using these window functions, the signal $s[u]$ can be decomposed into distinct components. The low-frequency component, which encapsulates the long-term trend of the signal is obtained as discrete inverse Fourier transform. And the band-pass components, which capture fluctuations at specific dyadic frequency scales are given by:

$$\Delta_j s[u] = \text{IDFT}(\Phi_j[k] \cdot S[k])[u], \quad j = 0, 1, \ldots, J - 1 - m. \tag{26}$$

By summing these components, the original signal can be approximately reconstructed as:

$$s[u] \approx \Delta_{-1} s[u] + \sum_{j=0}^{J-1-m} \Delta_j s[u]. \tag{27}$$

This decomposition demonstrates that the essential features of the signal are effectively captured across multiple frequency scales, allowing for a detailed analysis of its spectral characteristics.

A key property of our method is that the window functions form an approximate partition of unity in the frequency domain. This ensures the decomposition provides a stable and comprehensive representation of the signal. Specifically, the functions satisfy the relation:

$$X[k] + \sum_{j=0}^{J-1-m} \Phi_j[k] \approx 1, \tag{28}$$

for most frequency indices $k$. This property can be verified by examining different frequency regions.

1. **Low-Frequency Regions:** When the frequency index falls within the low-frequency range, the low-pass function fully retains the frequency components, while all band-pass functions are inactive:

   $$X[k] = 1, \quad \Phi_j[k] = 0, \quad k \leq 2^m \cup k \geq t - 2^m, \quad \forall j. \tag{29}$$

   Consequently, the summation property holds:

   $$X[k] + \sum_{j=0}^{J-1-m} \Phi_j[k] = 1 + \sum 0 = 1. \tag{30}$$

   This ensures that the low-frequency components are preserved without interference from high-frequency bands.

2. **Band-Pass Regions:** In the band-pass regions, the low-pass function does not contribute, while exactly one band-pass function is active:

   $$X[k] = 0, \quad \exists!\, j \text{ s.t. } \Phi_j[k] = 1, \quad 2^{j+m} \leq k < 2^{j+m+1}. \tag{31}$$

   This guarantees that the sum remains unity:

   $$X[k] + \sum_{j=0}^{J-1-m} \Phi_j[k] = 0 + 1 = 1. \tag{32}$$

   Thus, each frequency component is assigned uniquely to one of the band-pass filters, ensuring no overlap or redundancy.

3. **Boundary Points:** At the boundary points ($k = 2^{j+m}, 2^{j+m+1}, t - 2^{j+m}, t - 2^{j+m+1}$) where transitions occur between different frequency bands, both the low-pass and band-pass window functions may be inactive, leading to:

   $$X[k] = 0, \quad \Phi_j[k] = 0, \quad \forall j. \tag{33}$$

Consequently, at these discrete boundary points, the summation deviates from unity:

$$X[k] + \sum_{j=0}^{J-1-m} \Phi_j[k] = 0. \tag{34}$$

To further analyze the convergence properties of this decomposition, we define the error function:

$$E[k] = 1 - \left( X[k] + \sum_{j=0}^{J-1-m} \Phi_j[k] \right). \tag{35}$$

The set of indices where $E[k] \neq 0$ is given by:

$$\mathcal{E} = \{k \in \mathbb{Z} \mid E[k] \neq 0\}. \tag{36}$$

Since $E[k]$ is nonzero only at a finite number of boundary points, the measure of its support satisfies:

$$|\mathcal{E}| \ll t. \tag{37}$$

Thus, as $t \to \infty$, the fraction of affected indices vanishes, leading to an almost everywhere convergence:

$$\lim_{t\to\infty} \frac{|\mathcal{E}|}{t} = 0. \tag{38}$$

The above proofs ensure that the partitioning scheme provides a stable and asymptotically exact decomposition in the frequency domain.

Another fundamental aspect of our method is the disjointness and independence of the window functions, which guarantees that the extracted components remain distinct and do not interfere with each other. This separation is maintained through the following two properties:

1. **Between different band-pass windows:** For any $j \neq j\prime$, the support intervals of the band-pass window functions are disjoint:

$$[2^{j+m}, 2^{j+m+1}) \cap [2^{j\prime+m}, 2^{j\prime+m+1}) = \emptyset. \tag{39}$$

As a result, the corresponding window functions satisfy:

$$\Phi_j[k] \cdot \Phi_{j\prime}[k] = 0, \quad \forall k. \tag{40}$$

2. **Between the low-pass and band-pass windows:** The low-pass function $X[k]$ is supported in the low-frequency regions $k \leq 2^m \quad \text{or} \quad k \geq t - 2^m$. On the other hand, each band-pass function $\Phi_j[k]$ is supported in the range $2^{j+m} \leq k < 2^{j+m+1}$. Since $2^{j+m} > 2^m$, it follows that the support of $X[k]$ and $\Phi_j[k]$ are mutually exclusive, ensuring:

$$X[k] \cdot \Phi_j[k] = 0, \quad \forall j, k.$$

Based on the above method, the annulus $C$ is substituted with a corresponding set in the discrete Fourier domain, and the dyadic scaling operations are adapted to respect the discrete nature of the transform. The goal remains to construct a stable decomposition that preserves the essential properties of the continuous case while accommodating the constraints imposed by discrete sampling.

## B Proof of Long-Term Advantage Lower Bound of Adaptive Length

**Theorem 1** (Advantage Lower Bound of Adaptive Length): At time $t$, let $L_{\text{adap}}$ be the adaptive context length, $L_{\text{fix}}$ be the fixed context length, and the mutual information loss of $L$ be denoted as $\mathcal{L}_t(L_t)$. The expected cumulative reward difference between adaptive and fixed context length satisfies the following bound:

$$\begin{aligned} \sum_{t=1}^{T} \left(\mathcal{L}_t(L_{\text{fix}}) - \mathcal{L}_t(L_{\text{adap}})\right) &\geq \Omega(T) - O(T^{\alpha}) \\ &= \Omega(T) \quad (\text{when } T \text{ is sufficiently large}) \end{aligned} \tag{41}$$

where $0 \leq \alpha < 1$, with $\alpha$ being a non-deterministic parameter that varies with the environment.

**Proof:** To prove this theorem, we first model the non-stationary environment, including the observation and its latent variable. Specifically, at each time $t$, the agent receives noisy observations $o_t = g(\xi_t) + \epsilon_t$, where $\{\xi_t\}_{t=1}^T$ represents the latent variable, and the error term $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The latent variable sequence $\{\xi_t\}_{t=1}^T$ governs environmental dynamics and follows a diffusion process:

$$d\xi_t = \mu(\xi_t)dt + \eta(\xi_t)dW_t, \tag{42}$$

where $W_t$ represents a standard Brownian motion, $\mu(\cdot)$ and $\eta(\cdot)$ denote the drift and diffusion coefficients, respectively.

For any window length $L$, the mutual information $I(a_t; \xi_t|L)$ quantifies the information content of the action $a_t$ about the latent variable $\xi_t$:

$$I(a_t; \xi_t|L) = H(\xi_t) - H(\xi_t|a_t, L), \tag{43}$$

where $H(\xi_t)$ represents the entropy of $\xi_t$ (prior entropy), and $H(\xi_t|a_t, L)$ represents the conditional entropy of $\xi_t$ given $a_t$ and window length $L$. With the theoretical optimal window length $L_t^*$, the mutual information loss of $L$ can be obtained:

$$\mathcal{L}_t(L) = I(a_t; \xi_t|L_t^*) - I(a_t; \xi_t|L) \tag{44}$$

Since the action $a_t$ is generated from the observation sequence $o_{t-L:t} = \{o_{t-L}, \ldots, o_t\}$ via the stochastic policy $\pi(a_t|o_{t-L:t})$, the data processing inequality implies:

$$I(a_t; \xi_t|L) \leq I(o_{t-L:t}; \xi_t|L). \tag{45}$$

With the equation 43, we obtain:

$$I(o_{t-L:t}; \xi_t|L) = H(\xi_t) - H(\xi_t|o_{t-L:t}). \tag{46}$$

Thus, the upper bound on the mutual information loss is given by:

$$\mathcal{L}_t(L) \leq I(o_{t-L:t}; \xi_t|L_t^*) - I(o_{t-L:t}; \xi_t|L) = H(\xi_t|o_{t-L:t}) - H(\xi_t|o_{t-L_t^*:t}). \tag{47}$$

Based on the Fourier-based truncation, the policies can approximately fully utilize the observed information; thereby, we obtain:

$$\mathcal{L}_t(L) \approx H(\xi_t|o_{t-L:t}) - H(\xi_t|o_{t-L_t^*:t}). \tag{48}$$

where $\sigma_{t|L}^2$ represents the estimation accuracy of $\xi_t$ given $o_{t-L:t}$

Next, considering the general situation that the latent variable $\xi_t$ can be regarded as a Gaussian form for the posterior distribution $p(\xi_t|o_{t-L:t})$. This setting is reasonable under several mild and widely satisfied conditions: (i) the dynamics and observation models can be locally approximated as linear or weakly nonlinear in the neighborhood of the true latent state $L_t^*$; (ii) the observation noise is Gaussian, consistent with the structure of the Kalman filter; and (iii) the functions $\mu(\xi_t)$ and $\eta(\xi_t)$, often parameterized by neural networks, are smooth and differentiable, allowing for local Taylor expansions or moment-based approximations such as sigma-point propagation. Therefore, the posterior distribution can be reasonably approximated as:

$$p(\xi_t|o_{t-L:t}) = \mathcal{N}(\hat{\xi}_t, \sigma_{t|L}^2), \tag{49}$$

where $\hat{\xi}_t$ denotes the posterior mean and $\sigma_{t|L}^2$ the posterior variance. Consequently, the conditional entropy can be computed as:

$$H(\xi_t|o_{t-L:t}) = \frac{1}{2}\log(2\pi e\sigma_{t|L}^2). \tag{50}$$

Notably, this Gaussian approximation is particularly justified in policy optimization algorithms such as PPO [37], where updates are constrained to remain close to the current policy, effectively preserving local linearity in the latent-to-observation mapping.

Due to the fact that $L_t^*$ is the optimal window length at time step $t$, we expect that the posterior variance $\sigma_{t|L}^2$ behaves in a manner where it decreases as the window length approaches the optimal value $L_t^*$. This behavior is common in many estimation problems, where the system's performance

improves as it gets closer to the optimal configuration. In particular, reference [38] suggests that the most useful information for decision-making is concentrated around certain "bottleneck" points, and small deviations from the optimal choice result in progressively diminishing returns in terms of information processing. Given this intuition, the posterior variance, $\sigma^2_{t|L}$, can be regarded as behaving convexly in the neighborhood of $L_t^*$, because this ensures that small changes around the optimal window length lead to an increase in variance, which represents a deterioration in the quality of the estimation. Specifically, there exists a positive constant $k > 0$ such that:

$$\sigma^2_{t|L} \geq \sigma^2_{\min} + \frac{1}{2}k(L - L_t^*)^2, \tag{51}$$

where $\sigma^2_{\min} = \sigma^2_{t|L_t^*}$ denotes the posterior variance achieved at the optimal window length.

Combining equations 48 and 50, the mutual information loss can be lower bounded in terms of the posterior variance ratio:

$$\mathcal{L}_t(L) \approx \frac{1}{2}\log\left(\frac{\sigma^2_{t|L}}{\sigma^2_{\min}}\right). \tag{52}$$

Substituting the lower bound in equation 51 into equation 52, we obtain:

$$\mathcal{L}_t(L) \geq \frac{1}{2}\log\left(1 + \frac{k(L - L_t^*)^2}{2\sigma^2_{\min}}\right). \tag{53}$$

Applying the inequality $\log(1 + x) \geq \frac{x}{1+x}$ for $x > -1$, we further derive:

$$\mathcal{L}_t(L) \geq \frac{1}{2} \cdot \frac{\frac{k(L-L_t^*)^2}{2\sigma^2_{\min}}}{1 + \frac{k(L-L_t^*)^2}{2\sigma^2_{\min}}} = \frac{k(L - L_t^*)^2}{4\sigma^2_{\min} + 2k(L - L_t^*)^2}. \tag{54}$$

For a fixed window length $L_{\text{fix}}$, the total information loss over $T$ time steps can then be bounded from below:

$$\sum_{t=1}^{T} \mathcal{L}_t(L_{\text{fix}}) \geq \sum_{t=1}^{T} \left(\frac{k}{4\sigma^2_{\min}}(L_{\text{fix}} - L_t^*)^2 \cdot \frac{1}{1 + \frac{k}{2\sigma^2_{\min}}(L_{\text{fix}} - L_t^*)^2}\right). \tag{55}$$

When $|L_{\text{fix}} - L_t^*|$ exceeds a threshold $\zeta$, the denominator in equation 54 is bounded, and the loss is lower bounded by a positive constant $c = \frac{k\zeta^2}{4\sigma^2_{\min} + 2k\zeta^2}$. On the other hand, when $|L_{\text{fix}} - L_t^*| < \delta$, the quadratic term dominates and the loss scales as $\frac{k}{4\sigma^2_{\min}}(L_{\text{fix}} - L_t^*)^2$. In both cases, we have that the total loss satisfies:

$$\sum_{t=1}^{T} \mathcal{L}_t(L_{\text{fix}}) = \Omega(T), \tag{56}$$

demonstrating that any fixed window length incurs linear cumulative loss over time unless it tracks the optimal $L_t^*$.

Considering the adaptive strategy for selecting the context length $L_{\text{adap}}$ at each time $t$. We adopt the most universal polynomial rate to describe the adaptive policy converging to the optimal window length:

$$|L_{\text{adap},t} - L_t^*| \leq \epsilon_t = O(t^{-\beta}), \quad \beta > 0. \tag{57}$$

which are commonly used in general convex problems and in stochastic algorithms. In contrast, exponential convergence, etc., often requires extremely strong assumptions such as strong convexity and global smoothness.

Next, based on the equation 52, for small deviations $|L - L_t^*| < \delta$, applying the first-order approximation $\log(1 + x) \approx x$, we obtain:

$$\mathcal{L}_t(L) \approx \frac{k}{4\sigma^2_{\min}}(L - L_t^*)^2. \tag{58}$$

This shows that in a local region $|L - L_t^*| < \delta$, the mutual information loss behaves quadratically with respect to $(L - L_t^*)^2$. This local quadratic behavior is compatible with the global Lipschitz condition via the boundedness of gradients and compactness of the parameter space.

Moreover, in reinforcement learning algorithms such as TRPO [39] or PPO [37], the policy $\pi(a_t \mid o_{t-L:t})$ is updated under a trust-region constraint (e.g., a KL divergence threshold), which enforces local smoothness of the policy:

$$\|\theta_{t+1} - \theta_t\| \leq \delta. \tag{59}$$

This constraint implies a Lipschitz condition on the policy distribution in terms of Total Variation (TV) distance:

$$\|\pi(\cdot \mid o_{t-L:t}) - \pi(\cdot \mid o_{t-L^*:t})\|_{\text{TV}} \leq C_\pi |L - L^*|, \tag{60}$$

where $C_\pi$ is a constant dependent on the network structure.

By Pinsker's inequality, when the deviation from the optimal context length is small, the TV distance can be related to KL divergence:

$$\|\pi(\cdot \mid L) - \pi(\cdot \mid L^*)\|_{\text{TV}} \approx \frac{1}{2}\sqrt{D_{\text{KL}}(\pi(\cdot \mid L)\|\pi(\cdot \mid L^*))}. \tag{61}$$

Combining the equation 60 and the equation 61:

$$D_{\text{KL}}(\pi(\cdot \mid L)\|\pi(\cdot \mid L^*)) \leq 2C_\pi^2 |L - L^*|^2. \tag{62}$$

Considering mutual information $I(a_t; \xi_t \mid L)$ is continuous with respect to the policy distribution [40], there exists a constant $C_I > 0$ such that:

$$|I(a_t; \xi_t \mid L) - I(a_t; \xi_t \mid L^*)| \leq C_I \cdot D_{\text{KL}}(\pi(\cdot \mid L)\|\pi(\cdot \mid L^*)). \tag{63}$$

Therefore, we obtain:

$$|I(a_t; \xi_t|L) - I(a_t; \xi_t|L^*)| \leq C_I C_\pi |L_{\text{adap}} - L^*| := K|L_{\text{adap}} - L^*|. \tag{64}$$

For any $L, L^*$, we consider two cases: **(i) Local region** ($|L - L^*| < \delta$): In this case, we directly apply the previously established equation 64 **(ii) Global region** ($|L - L^*| \geq \delta$): Since the mutual information is upper bounded by the entropy $H(\xi_t)$, we have:

$$|I(a_t; \xi_t|L) - I(a_t; \xi_t|L^*)| \leq H(\xi_t) \leq \frac{H(\xi_t)}{\delta} \cdot |L - L^*|. \tag{65}$$

Combining both cases, define:

$$K := \max\left\{C_I C_\pi, \frac{H(\xi_t)}{\delta}\right\}, \tag{66}$$

then for any $L, L^*$, the mutual information satisfies a global Lipschitz condition:

$$|I(a_t; \xi_t|L) - I(a_t; \xi_t|L^*)| \leq K|L - L^*|. \tag{67}$$

Further, we can obtain the following bound of adaptive context length on single-step mutual information loss:

$$\mathcal{L}_t(L_{adap}) = |I(a_t; \xi_t \mid L) - I(a_t; \xi_t \mid L^*)| \leq K|L_{\text{adap}} - L^*| = O(t^{-\beta}). \tag{68}$$

This establishes the global Lipschitz continuity of mutual information with respect to the context length $L$.

Summing over $t = 1$ to $T$, we get:

$$\sum_{t=1}^{T} \mathcal{L}_t(L_{adap}) \leq K \sum_{t=1}^{T} t^{-\beta} \tag{69}$$

Now we apply standard results from numerical analysis of p-series:

- If $\beta > 1$, then the series $\sum_{t=1}^{T} t^{-\beta}$ converges. Hence, the cumulative information loss is bounded: $\sum_{t=1}^{T} \mathcal{L}_t(L_{adap}) = O(1)$.
- If $\beta = 1$, then the series becomes harmonic and grows logarithmically: $\sum_{t=1}^{T} t^{-1} = O(\log T)$. Thus, $\sum_{t=1}^{T} \mathcal{L}_t(L_{adap}) = O(\log T)$.

- If $0 < \beta < 1$, then the series grows polynomially: $\sum_{t=1}^{T} t^{-\beta} = O(T^{1-\beta})$. Consequently, $\sum_{t=1}^{T} \mathcal{L}_t(L_{adap}) = O(T^{1-\beta})$.

In all cases, the cumulative information loss is sublinear in $T$, i.e., there exists $\alpha < 1$ such that:

$$\sum_{t=1}^{T} \mathcal{L}_t = O(T^{\alpha}). \tag{70}$$

Combine with the equation 56 and the equation 70, we obtain:

$$\begin{aligned} \sum_{t=1}^{T} (\mathcal{L}_t(L_{\text{fix}}) - \mathcal{L}_t(L_{adap})) &= \Omega(T) - O(T^{\alpha}) \\ &= \Omega(T) \quad (\text{when } T \text{ is sufficiently large}) \end{aligned} \tag{71}$$

# C Additional Details for Experiments

## C.1 Environments

**Sample Spread** The Sample Spread environment is a cooperative multi-agent task where agents must coordinate their movements to cover multiple static landmarks, aiming to minimize the overall distance between agents and landmarks while avoiding inter-agent collisions. In the original setting, the number of agents and landmarks is equal (3 each), which may lead to agents remaining stationary. To encourage exploration and promote more dynamic coordination behavior, we modify the setting to include 4 agents and 3 landmarks. The action shape is 5. The reward design is as follows:

- Each agent receives a local penalty of $-1.0$ for every collision with other agents, encouraging collision avoidance.
- The global reward is defined as $R = -\sum_{i=1}^{4} \min_j \|p_j - l_i\|$, where $p_j$ and $l_i$ denote the positions of agent $j$ and landmark $i$, respectively, encouraging agents to minimize the overall distance to landmarks.

The observation shape is 24:

Table 3: The Observation Features of Sample Spread

| **Feature** | **Dim** | **Description** |
|---|---|---|
| Self Velocity | 2 | Agent's own velocity vector |
| Self Position | 2 | Agent's own position in world coordinates |
| Landmark Relative Positions | 8 | Relative positions of 4 landmarks to the agent |
| Other Agents' Relative Positions | 6 | Relative positions of the other 3 agents to the agent |
| Communication Vectors | 6 | Communication features from the other 3 agents |
| **Total** | 24 | Final observation dimension per agent |

**Minigrid Soccer Game** The MiniGrid Soccer Game is a multi-agent, competitive and cooperative environment in which agents must coordinate to score goals using shared balls. Agents can pass the balls, intercept opponents, and strategically position themselves to influence the game outcome. In our specific configuration, we use 4 balls (red), 3 teams (blue, green, and yellow), with 3 agents per team, and each team is assigned a goal of corresponding color. The action-observation space is partially observable. The rewards are global-shared, which all agents receive the same reward whenever any agent scores or concedes a goal:

- Each agent receives a shared reward of $+1$ upon successfully picking up the ball from the ground. Each ball only provides this reward once.
- A shared reward of $+10$ is given when any agent scores a goal into its own team's designated goal area.
- A shared penalty of $-5$ is applied if the ball is accidentally scored into an opponent's goal.

- When holding the ball, agents receive a step-wise penalty of $-0.02 \times$ dist based on the distance between the ball and the agent's own goal.
- When holding the ball, agents receive a dense reward of $+0.2 \times$ progress based on the positive progress made toward their own goal.

Each agent observes a 3×3 local grid, and the shape of each cell is 6:

Table 4: The Observation Features of Minigrid Soccer Game

| Feature | Dim | Description |
|---|---|---|
| Object Type | 1 | Type index (wall/door/agent/key/...) |
| Color | 1 | Color index (green/blue/...) |
| State | 1 | Object state (0-2 for door open/closed/locked) |
| Carried Type | 1 | Type of carried object (0 if none) |
| Carried Color | 1 | Color of carried object (0 if none) |
| Direction/Marker | 1 | Agent direction (0-3) or current agent flag (0/1) |
| **Total** | **6** | |

**Academy 3 vs 1 with Keeper** Academy 3 vs 1 with Keeper environment is a multi-agent scenario where 3 offensive agents cooperate to score against a goalkeeper and a defender. The action space is discrete with 19 actions, covering basic football behaviors like passing, shooting, and movement. The reward structure is sparse: +100 for scoring a goal, –1 if the episode ends without scoring. Besides, we evaluate the model every 100 training episodes, each time over 30 test episodes to compute the average win rate. The observation shape is 26:

Table 5: The Observation Features of Academy 3 vs 1 with Keeper

| Feature | Dim | Description |
|---|---|---|
| Ego Player Position | 2 | $(x, y)$ position of the observing agent |
| Teammates Relative Positions | 4 | Relative positions of 2 teammates w.r.t. ego |
| Ego Player Direction | 2 | Velocity vector (direction) of the ego agent |
| Teammates Directions | 4 | Directions of 2 teammates |
| Opponents Relative Positions | 6 | Relative positions of 3 opponents w.r.t. ego |
| Opponents Directions | 6 | Directions of 3 opponents |
| Ball Relative Position | 2 | Ball position relative to ego |
| Ball Height | 1 | Ball $z$ coordinate (height) |
| Ball Direction | 3 | Ball velocity in $(x, y, z)$ |
| **Total** | **26** | |

**Academy Counterattack-Hard** Academy Counterattack-Hard environment is a multi-agent scenario where 4 offensive agents cooperate to score against a goalkeeper and a defender. The action space is discrete with 19 actions, covering basic football behaviors like passing, shooting, and movement. The reward structure is sparse: +100 for scoring a goal, –1 if the episode ends without scoring. Besides, we evaluate the model every 100 training episodes, each time over 30 test episodes to compute the average win rate. The observation shape is 34:

Table 6: The Observation Features of Academy Counterattack-Hard

| Feature | Dim | Description |
|---|---|---|
| Ego Agent Position | 2 | Agent's own $(x, y)$ coordinates |
| Teammates Relative Positions | 6 | Relative $(dx, dy)$ of 3 teammates |
| Ego Agent Direction | 2 | Movement vector $(v_x, v_y)$ |
| Teammates Directions | 6 | Movement vectors of 3 teammates $(v_x, v_y) \times 3$ |
| Opponents Relative Positions | 6 | Relative $(dx, dy)$ of 3 opponents |
| Opponents Directions | 6 | Movement vectors of 3 opponents $(v_x, v_y) \times 3$ |
| Ball Relative Position | 2 | Ball $(x, y)$ relative to ego agent |
| Ball Height | 1 | Ball $z$-coordinate (altitude) |
| Ball Direction | 3 | Ball velocity $(v_x, v_y, v_z)$ |
| **Total** | **34** | |

### C.2 Experiments with Sequence Processing Methods

All experiments were conducted using an NVIDIA A100 GPU, with the longest single training run taking approximately one month.

**Baseline Methods**

- Transformer [34]: A deep learning architecture that utilizes self-attention and positional encoding to model complex dependencies across sequences.
- Token Statistics Transformer (ToST) [35]: A recent Transformer variant, using a data-dependent low-rank projection based on the second moment statistics of input token features, and achieving linear computational complexity.
- AMAGO[36]: An in-context reinforcement learning algorithm that enables long-sequence Transformers to process entire trajectories in parallel, overcoming the memory capacity and long-term planning bottlenecks of traditional recurrent networks.

**Hyperparameter** We provide the hyperparameters used in each environments as follows:

Table 7: Hyperparameter Configuration

| **Parameter** | **Value** |
|---|---|
| Learning Rate | 0.001 |
| Discount Factor ($\gamma$) | 0.98 |
| GAE Coefficient ($\lambda$) | 0.95 |
| PPO Clip ($\epsilon$) | 0.2 |
| Training Epochs | 10 |
| Batch Size (Sample Spread) | 25 |
| Batch Size (Minigrid Soccer Game) | 128 |
| Batch Size (Academy 3 vs 1 with Keeper) | 50 |
| Batch Size (Academy Counterattack-Hard) | 50 |
| Entropy Coefficient ($\beta$) | 0.01 |
| MLP Hidden Layers (Sample Spread) | [256, 64, 16] |
| CNN Hidden Layers (Minigrid Soccer Game) | [16, 32, 64] |
| MLP Hidden Layers (Minigrid Soccer Game) | [256, 128, 64] |
| MLP Hidden Layers (Academy 3 vs 1 with Keeper) | [1024, 256, 64] |
| MLP Hidden Layers (Academy 3 vs 1 with Keeper) | [1024, 256, 64] |
| Activation Function | ReLU |
| Optimizer Type | Adam |

**Central Agent with Low-Frequency Truncation** Based on the Dyadic Partition of Unity in Discrete Form, we assign distinct low-frequency truncation lengths to each environment. Specifically, after applying the Discrete Fourier Transform (DFT), we retain the first 4, 128, 64, and 64 frequency components for the Sample Spread, MiniGrid Soccer Game, Academy 3 vs 1 with Keeper, and Academy Counterattack-Hard environments, respectively. The above frequency components are then inputted to the central agent, for which the input dimension is defined as $k_0$. The action space for the central agent in each environment is defined as follows:

Table 8: Central Agent Action Space

| **Environment (Step > Threshold)** | **Action Space** |
|---|---|
| Sample Spread | 0, 0, 1, 2, 4 |
| MiniGrid Soccer Game | 0, 1, 2, 4, 8, 16, 32, 64 |
| Academy 3 vs 1 with Keeper | 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 4, 8, 16, 32, 64 |
| Academy Counterattack-Hard | 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 4, 8, 16, 32, 64 |

The action space of the central agent in each environment is adaptively constructed based on the current time step using a dyadic partitioning rule. Specifically, for each environment, we define a threshold step value: Sample Spread ($t > 7$), MiniGrid Soccer Game ($t > 255$), Academy 3 vs 1

with Keeper and Academy Counterattack-Hard ($t > 127$). When the step $t$ exceeds the corresponding threshold, the action space is composed of a set of dyadic components $\{2^0, 2^1, \ldots, 2^k\}$, where $k = \min(\log_2(k_0), [\log_2 t] - 1)$, and the vector is left-padded with zeros to match the fixed dimension. The resulting component sets for each environment under this truncation rule are summarized in Table 8.

This work develops an adaptive context length optimization method with Fourier-based low-frequency truncation for multi-agent reinforcement learning (MARL). The proposed approach significantly improves the efficiency and effectiveness of MARL systems, enabling better decision-making in complex, dynamic environments.

The positive societal impacts of this research include advancing intelligent multi-agent systems in diverse domains such as transportation management, robotics, and resource allocation. These improvements can contribute to enhanced safety, reduced energy consumption, optimized traffic flows, and overall better management of complex systems benefiting society.

By promoting more efficient learning and adaptation in multi-agent environments, this work helps pave the way for scalable and practical AI applications that address real-world challenges with improved reliability and performance.

### C.3 ACL-LFT Algorithm

All methods in our experiments (including ours and the baselines) follow the same structural setting: a centralized module processes only the historical information and transmits it to distributed agents for decision-making. All methods share the same network architecture, ensuring a fair comparison across all baselines. The pseudocode of the ACL-LFT training process under this unified framework is provided in Algorithm 1.

**Algorithm 1:** ACL-LFT Policy-Making and Training Algorithm

**Input:** Distributed agents' policies $\{\theta_i\}_{i=1}^N$, shared value function $\phi$; central agent's policy $\theta_c$, value function $\phi_c$; horizon $T$; epochs $K_c$, $K_d$.
**Output:** Updated policies $\{\theta_i\}_{i=1}^N$, $\theta_c$ and value functions $\phi$, $\phi_c$.
Initialize per-agent episode buffers $\{B_i\}_{i=1}^N$; reset environment;
**for** *episode* $= 1$ *to max_epi* **do**
  **for** $t = 0$ *to* $T - 1$ **do**
    **for** $i = 1$ *to* $N$ **do**
      Extract historical state $s_t^{-1}$ and perform Fourier Transform;
      Obtain low-frequency section $s_t^c$;
      Obtain optimal contextual information $s_t^{-opt}$;
    **end**
    **for** $i = 1$ *to* $N$ **do**
      Obtain $a_t^i$ by $s_t^{-opt}$ and $s_t^i$; perform $a_t^i$;
      Obtain $s_{t+1}^i$ and $r_t^i$;
      Store transition $\tau_t^i$, $s_{t+1}^i$ in $B_i$;
    **end**
    Obtain $r_t^c$; store center transition $\tau_t^c$;
  **end**
  Compute advantages $A_t^c$ using GAE with $\phi_c$;
  **for** *epoch* $= 1$ *to* $K_c$ **do**
    Update policy $\theta_c$, value function $\phi_c$;
  **end**
  Construct centralized critic input $s_t^{global}$;
  Compute advantages $A_t^i$ using GAE with $\phi(s_t^{global})$;
  Collect $\tau_t^i$ into shared buffer $B$;
  **for** *epoch* $= 1$ *to* $K_d$ **do**
    Update shared policy $\theta_i$, value function $\phi$;
  **end**
**end**

### C.4 Additional Experiments

Table 9: Performance on SMACv2 with different MARL backbones and coordination mechanisms.

| SMACv2 Task | RL-Method | Transformer | ToST | AMAGO | Mamba | ACL-LFT |
|---|---|---|---|---|---|---|
| 3s5z vs 3s6z | MAPPO | $71.5 \pm 3.9$ | $72.2 \pm 3.4$ | $76.1 \pm 2.9$ | $72.6 \pm 3.2$ | $\mathbf{78.9 \pm 2.8}$ |
| | QMIX | $72.8 \pm 3.7$ | $73.7 \pm 3.6$ | $76.3 \pm 3.1$ | $75.1 \pm 3.7$ | $\mathbf{79.4 \pm 2.9}$ |
| | QPLEX | $73.6 \pm 3.3$ | $75.1 \pm 3.6$ | $77.5 \pm 2.8$ | $75.0 \pm 3.3$ | $\mathbf{80.1 \pm 3.0}$ |
| 5m_vs_6m | MAPPO | $44.5 \pm 4.9$ | $46.3 \pm 4.4$ | $48.1 \pm 4.0$ | $46.2 \pm 4.5$ | $\mathbf{52.7 \pm 4.2}$ |
| | QMIX | $44.3 \pm 5.5$ | $46.9 \pm 4.8$ | $48.3 \pm 4.3$ | $46.1 \pm 5.1$ | $\mathbf{52.4 \pm 4.6}$ |
| | QPLEX | $45.6 \pm 5.7$ | $47.3 \pm 4.6$ | $49.5 \pm 4.8$ | $47.8 \pm 4.9$ | $\mathbf{53.9 \pm 4.3}$ |
| corridor | MAPPO | $65.6 \pm 5.7$ | $68.1 \pm 6.6$ | $74.3 \pm 4.8$ | $69.0 \pm 5.9$ | $\mathbf{77.9 \pm 5.3}$ |
| | QMIX | $68.5 \pm 5.9$ | $70.3 \pm 6.8$ | $75.0 \pm 5.3$ | $71.2 \pm 6.2$ | $\mathbf{78.6 \pm 5.4}$ |
| | QPLEX | $70.6 \pm 4.7$ | $72.9 \pm 4.6$ | $76.3 \pm 5.8$ | $73.5 \pm 5.5$ | $\mathbf{79.2 \pm 4.9}$ |

As shown in Table 9, ACL-LFT consistently outperforms all evaluated baselines—including MAPPO, QMIX, and QPLEX—across all SMACv2 scenarios, with average improvements of +2.6% to +4.6% over the best-performing baseline (typically AMAGO).

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly summarize the paper's main contributions and accurately reflect its scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: he limitations of the work are briefly discussed in the appendix, including the assumptions made and directions for future improvement.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper includes all necessary assumptions and complete proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient methodological details and experimental settings to reproduce the main results.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide access to the core codebase and key experimental results in the supplemental material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides comprehensive details on training and testing settings, including data splits, hyperparameter choices, optimizer types, and selection criteria. Additional specifics are included in the appendix to ensure full reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars and/or confidence intervals for key experimental results. The sources of variability (e.g., random initialization, data splits) are clearly stated, and the methods for calculating error bars are described. Assumptions related to the statistical analysis are also discussed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: The paper specifies the hardware used and total computational resources required.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research presented in this paper fully adheres to the NeurIPS Code of Ethics. All experiments and data collection procedures comply with ethical standards, including respect for privacy, avoidance of harm, and fairness. No conflicts with applicable laws or regulations are present, and no special ethical concerns arise from the methods or applications discussed.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See Appendix for a discussion of potential broader societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper does not involve the release of high-risk models or datasets. It proposes a novel framework for multi-agent reinforcement learning that operates on simulated environments (e.g., PettingZoo, MiniGrid, GRF), which do not pose significant risks of misuse or require special safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper uses existing environments. All assets are properly cited in the paper with references to their original sources, and their licenses and terms of use have been fully respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release any new datasets, models, or other assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve any crowdsourcing or human subject research. All experiments are conducted in simulation environments without human interaction.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve any human subjects or crowdsourcing experiments. All experiments are conducted in simulation environments without human interaction.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The research does not use large language models (LLMs) as an important, original, or non-standard part of its methodology. Any LLM involvement, if any, was limited to general writing or editing assistance and does not impact the core scientific contributions.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.