GENERAL AGENTS MUST LEARN ONLINE

Anonymous authors
Paper under double-blind review

ABSTRACT

It has been widely anticipated that scaling offline training could enable artificial general intelligence (AGI) in agents with *frozen policies*. However, such agents rely on offline knowledge, restricting their capability in big worlds that exceed their capacity. We formalize this limitation by showing frozen agents require *a model size* that grows with environment complexity and the accuracy demanded across rewards. Moreover, frozen agents require substantial amounts of *training data* to acquire this capability offline. Finally, we establish an achievable tradeoff between frozen information and *online experience*, showing that general agents are attainable with limited persistent capacity. These results motivate architectures that integrate offline training with online experience as an alternative route toward AGI.

1 INTRODUCTION

Humans draw on accumulated knowledge to meet changing demands and continually update that knowledge through experience, a process supported by neural *plasticity* (Kolb & Whishaw, 1998). Artificial general intelligence (AGI) seeks to reproduce this generality, aiming for human-level capabilities across cognitive domains rather than excellence on a pre-specified set of goals (Hendrycks et al., 2025). In real-world decision-making, this ambition requires an agent to apply acquired knowledge across *diverse and previously unspecified goals* without goal-specific redesign.

A prominent route toward this ambition is to acquire broadly useful knowledge through *large-scale pre-training before deployment*. This approach is evident across domains, e.g., contemporary language models develop reasoning, coding, and tool use capabilities through broad pre-training and post-training, including reinforcement learning (OpenAI, 2025; Anthropic, 2025; GLM-5 Team, 2026), while visual world models combine visual pre-training with training on recorded robot interactions to support planning and control (Ye et al., 2026; Mur-Labadia et al., 2026; NVIDIA, 2026a). However, if learning stops at deployment, the knowledge retained from training must support decision-making across all downstream goals.

The *big world hypothesis* turns this requirement into a scaling problem because the environment can contain far more structure potentially relevant to decision making than an agent can retain (Javed & Sutton, 2024). An agent pursuing a specific goal can ignore environmental distinctions that do not affect its decisions. A general agent, however, must also support goals for which those same distinctions may be essential. The breadth of goals therefore determines how much of the environment an agent needs to understand. This raises the central question of *how much environmental information generality requires and whether that information must be retained before deployment or can instead be acquired through online experience*.

Recent theory makes this information requirement concrete by showing sufficiently broad capabilities constrain what an agent must encode about its environment. Richens & Everitt (2024) show that robustness across local interventions implies an approximate causal world model, while Richens et al. (2025) show that bounded performance across a sufficiently rich family of compositional goals implies a predictive world model. These work characterize what general agents must know about their environments, but these characterizations alone do not establish whether that knowledge can be acquired and retained with limited resources. This gap is critical in big worlds, where the information needed across goals may exceed an agent's capacity.

Learning during deployment offers an alternative in which agents acquire environmental knowledge through interaction and use it to revise their continuation policies (Grigsby et al., 2024a; Duan et al., 2016). Examples include robotic controllers that update policy parameters through physical

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

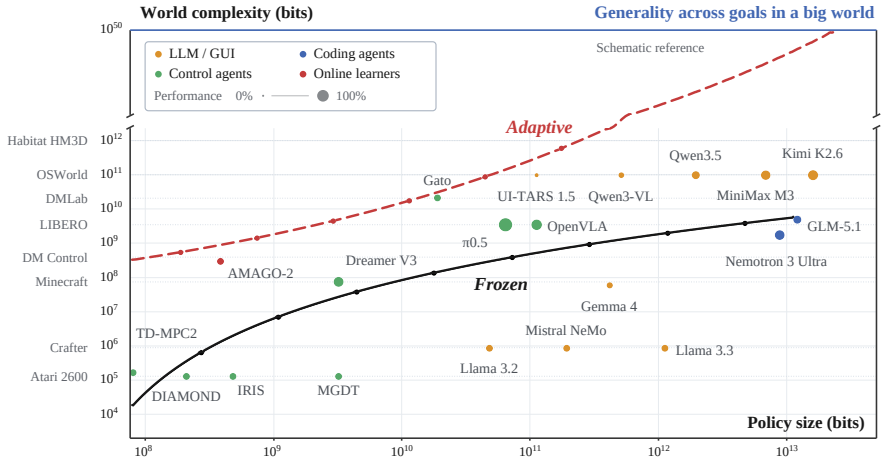


Figure 1: An illustrative visualization of why **general agents need to learn online** under the big world hypothesis. Here, policy size is measured by the number of bits required to store the policy parameters, while the world complexity is proxied by the number of bits required to store a simulator. The markers denote representative agents from several categories, using the success rates reported in the corresponding simulator benchmarks. Appendix F lists sources and clarifies measurements.

interaction and human corrections (Luo et al., 2025), and language agents that revise explicit memories using execution feedback to guide subsequent decisions (Zhang et al., 2026). We use *online learning* to describe this adaptation from deployment experience, without restricting it to a particular learning algorithm or architecture. By learning what is needed for the specified goal during interaction, agents can reduce the information that must be acquired and retained in advance. Thus, **frozen** and **adaptive** agents represent two deployment regimes, distinguished by whether their continuation policies remain fixed before interaction begins or are allowed to evolve with experience. This distinction becomes consequential as world complexity grows.

Figure 1 highlights a possible limit to pursuing generality through model scaling alone. Pretrained models such as GPT-6 Astra (OpenAI, 2026), Claude Opus 5.5 (Anthropic, 2026), and Gemini 3.8 Flash (Google DeepMind, 2026b) demonstrate broad capabilities, making further scaling of offline training an appealing route toward general agents. Yet, as world complexity grows, achieving human-level performance across all future tasks may require frozen agents to retain more information before deployment than a *physically achievable* model can store. The adaptive trend illustrates a different possibility, in which interaction supplies the knowledge needed for the task, reducing what must be retained in advance.

This raises the question of whether learning during deployment changes how the resources such as the policy size and number of offline training samples needed on general human-level tasks in the real world. To theoretically compare these two policy paradigms, we consider *reward-conditioned reinforcement learning*, in which offline training **precedes** the specification of the downstream reward. Agents must therefore support goals that were unspecified during training. After deployment, the reward is revealed, and a general agent needs to perform well regardless of which reward is specified. We quantify world complexity through the *metric entropy* of the transition family, which measures the information needed to specify the dynamics at a prescribed resolution. This formulation yields three main results in the following:

- *The model size of a frozen general agent must scale with world complexity.* We first show that to perform uniformly well across all rewards, a frozen agent must retain environment-dependent information that grows with the complexity of the underlying dynamics, regardless of how much offline data or computation is available (see Theorem 1). For example, a frozen agent in a simplified household setting needs at least 10^{49} parameters[1], close to the number of atoms that Earth contains (Mazur, 2015) (see **Example 2** in Section 4.3).
- *Training a general agent offline is data-hungry.* In addition, even with unlimited retained capacity, training a frozen general agent requires an amount of offline data that grows with world complexity and inverse-quadratically with the target regret (see Theorem 2).

[1]At 16 bits per parameter, with all retained environmental information stored in model weights.

- *Learning during deployment.* In contrast, we show that for any finite capacity, sufficiently many online interactions can yield uniformly good performance across all rewards, establishing an achievable tradeoff between information retained after offline training and information acquired online (see Theorem 3 and Section 4.4).

Section 4.3 and 4.4 discuss the broader significance of these results. Section 4.3 places the lower bounds for frozen general agents on a concrete scale: even a highly restricted household setting yields an astronomical gap between the information required for generality and the capacity of current agentic models. Section 4.4 then sharpens the implications of Theorem 3: a single interaction history can support the entire reward family. Moreover, in our reward-conditioned RL setting, scalable implementations may exploit reward-specific memory and thereby avoid online-memory requirements of the same order as the persistent capacity required by frozen agents.

2 PRELIMINARIES

In this section, we will formalize the setting and the problem.

Environments. We model the environments as Markov Decision Processes (MDPs) (Sutton & Barto, 2018), which are characterized by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \mathcal{R}, \gamma)$. Specifically, $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, respectively; $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel; $\mathcal{R}$ is the family of all **bounded transition rewards** $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$; and $\gamma \in (0, 1)$ is the discount factor. We denote a sequence of state-action pairs as a *trajectory* $\tau = (s_0, a_0, s_1, a_1, \dots)$, and a finite prefix of τ as a *history* $h_t = (s_0, a_0, \dots, s_t)$. We then make the following assumption to simplify our analysis:

Assumption 1. *We assume MDPs with at least* 5 *states and* 2 *actions, providing generative access that allows the agent to repeatedly interact from any state* $s \in \mathcal{S}$.

Here, we assume $|\mathcal{S}| \geq 5$ to simplify our results, and $|\mathcal{A}| \geq 2$ to allow non-trivial policies. Generative access is a standard assumption in theoretical analysis of sample complexity in MDPs (Zurek & Chen, 2025; Gheshlaghi Azar et al., 2013).

Agents. In recent literature (Richens et al., 2025), general agents are often formulated as goal-conditioned agents (Liu et al., 2022). Similarly, we study *reward-conditioned reinforcement learning*, in which agents are modeled as reward-conditioned agents π_r. Here, we distinguish two types of general agents: we first introduce *frozen general agents.* Intuitively, a frozen agent is trained before deployment and does not update its policy once an episode begins. Since each specified reward induces a discounted MDP, a stationary Markovian policy is sufficient for optimal control (Ingebrand et al., 2024). We then model a frozen general agent as a **state-response** policy:

Definition 1 (frozen agent)**.** *An agent is frozen if it does not update its policy in response to interactions after the episode begins. We restrict the frozen agents to stationary Markovian policies*

$$\pi_r : \mathcal{S} \times r \to \Delta(\mathcal{A})$$

in zero-shot MDP settings considered in this paper.

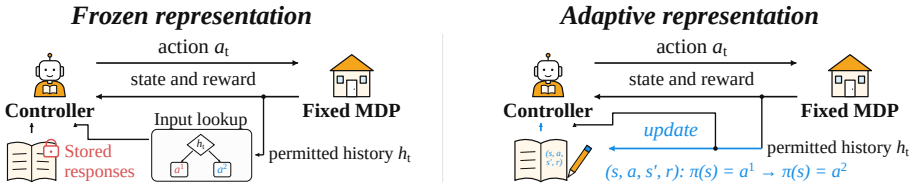


Figure 2: **Frozen and adaptive agents.** The frozen agent stores responses to all possible inputs in advance, including permitted histories. The adaptive agent updates its policy through interaction.

In contrast, we consider *adaptive general agents*, which behave as online learners by updating their policies from experience throughout the episode. Although the policy used at each time step may be Markovian, the adaptation itself depends on previous interactions. Hence, the induced policy over the episode is history-dependent. Formally, an adaptive policy is $\pi_r : h_t \times r \to \Delta(\mathcal{A})$. For simplicity, we assume policies are deterministic throughout this work.

Remark 1. Importantly, history dependence alone does not determine whether an agent is still learning. We say an agent is learning online only when its continuation policy changes in response to experience acquired through interactions after deployment (Abel et al., 2023a;b).

Value functions. We use s_t and a_t to denote the state and action at time t, respectively. The value function is then defined by

$$\mathcal{V}^{\pi_r}(s) \coloneqq \mathbb{E}_{\pi_r}\Big(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \mid s_0 = s\Big).$$

Naturally, we define the optimal value function as $\mathcal{V}^*(s) \coloneqq \max_{\pi_r} \mathcal{V}^{\pi_r}(s)$ for all $s \in \mathcal{S}$. The optimal policy is then $\pi^* \coloneqq \arg\max_{\pi_r} \mathcal{V}^{\pi_r}(s)$ for all $s \in \mathcal{S}$. However, in practice, general agents are only expected to achieve bounded regret relative to the optimal value function.

Definition 2 (bounded regret general agent)**.** *Given a desired regret δ, we define a bounded regret general agent as an agent π_r satisfying*

$$\mathcal{V}^{\pi_r}(s) \geq \mathcal{V}^*(s) - \delta \tag{1}$$

for every transition reward r and every initial state $s \in \mathcal{S}$.

This requirement reflects a **standard notion of general agents** as systems that *generalize across a broad family of rewards*, rather than being specialized to particular rewards (Richens et al., 2025).

Motivation. Silver & Sutton (2025) suggest that general agents will advance through *online interactions* with the world. However, **is online learning unavoidable for a general agent?** We connect this question to world complexity through *world model theorems* (Richens & Everitt, 2024; Richens et al., 2025): if the same reward-conditioned policy π_r achieves bounded regret δ across $\mathcal{R}$ under two transition kernels P and P', then the regret guarantee allows us to upper bound the discrepancy between P and P'. This in turn defines a fidelity threshold over world models: transition kernels within this threshold can be treated as equivalent, whereas those beyond it must be distinguished. Therefore, in this paper, we aim to provide a formal answer to the following fundamental question:

Can frozen agents retain environment representations of sufficient fidelity to achieve bounded regret?

Moreover, if not, *does online learning offer a scalable route for general agents?*

3 BIG WORLD HYPOTHESIS

In this section, we aim to characterize big worlds and measure their complexity.

How to simulate big worlds? Recent works embed a finite state automaton within a computationally universal environment, such as a universal Turing machine, thereby making the agent intrinsically smaller than the world (Lewandowski et al., 2025), or consider arbitrarily complex goal sets (Lu et al., 2026) to formalize big worlds. However, although environments may contain arbitrarily complex transition dynamics, the reward family $\mathcal{R}$ may only require agents to retain accurate knowledge of finite parts of environments. Hence, we distinguish between the information an agent must preserve to achieve bounded regret and the intrinsic complexity of the world itself.

Randomness of environments. In order to quantify the complexity of environments, we consider the full tabular transition kernel family $\mathcal{F} \coloneqq (\Delta_{|\mathcal{S}|-1})^{\mathcal{S}\times\mathcal{A}}$, where $\Delta_{|\mathcal{S}|-1}$ denotes the probability simplex of dimension $|\mathcal{S}| - 1$ (Lidayan et al., 2025). At the outset, a transition kernel P is sampled from an arbitrary full-support distribution over $\mathcal{F}$, and we denote the realized kernel by P_θ, where θ indexes the realized environment. Once the environment is realized, the same P_θ governs the entire procedure: the agent first undergoes offline training in this environment, and is then deployed in the same environment after a reward $r \in \mathcal{R}$ is specified.

However, how should we measure the environmental information that general agents must retain? Existing world model theorems show that robust agents can recover the dynamics accurately (Richens & Everitt, 2024; Richens et al., 2025). We therefore consider ϵ-separated families of transition kernels under the uniform total variation distance:

Definition 3 (ϵ-separated world model family)**.** *A world model family is an ϵ-**separated world model family** $\mathcal{F}_\epsilon \subseteq \mathcal{F}$ if the following inequality holds:*

$$\|P_\theta - P_{\theta'}\|_{\infty,\mathrm{TV}} \coloneqq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|P_\theta(\cdot \mid s,a) - P_{\theta'}(\cdot \mid s,a)\|_{\mathrm{TV}} > \epsilon \tag{2}$$

for any two distinct parameterized transition kernels $P_\theta, P_{\theta'} \in \mathcal{F}_\epsilon$, where $\epsilon \in (0,1)$.

Definition 3 treats world models that differ by less than or equal to ϵ as equivalent at the required fidelity. Since $\mathcal{F}$ is a compact finite-dimensional metric space, the ϵ-separated world model family $\mathcal{F}_\epsilon$ is also finite. In particular, if bounded regret forces agents to distinguish every pair of ϵ-separated transition kernels, then the maximum cardinality of $\mathcal{F}_\epsilon$ lower bounds the number of distinguishable persistent representations agents must support. We provide an example below to further illustrate this.

Example 1. Suppose a model-based agent needs to distinguish four representative transition kernels to satisfy a given regret δ. Then the agent must retain enough information to identify which of these four environments it inhabits, requiring at least $\log_2 4 = 2$ bits of persistent information, even though its actual implementation may require many more parameters.

Finally, we measure such complexity by entropy. For a given approximation precision ϵ, different partitions of $\mathcal{F}$ may yield different numbers of representative world models, motivating two complementary geometric constructions. We first introduce the **ϵ-packing entropy:** we select an ϵ-separated world model family of maximum cardinality (Colomer et al., 2025). Accordingly, we define the following lower bound on the ϵ-packing entropy under total variation distance:

Definition 4 (ϵ-packing entropy lower bound)**.** *For any distance $\epsilon \in (0, \frac{1}{4})$ and for the full tabular transition kernel family $\mathcal{F}$, we write the lower bound on ϵ-packing entropy as*

$$\underline{\mathbb{H}}_\epsilon(\mathcal{F}) \coloneqq |\mathcal{S}||\mathcal{A}|(|\mathcal{S}| - 1) \log_2 \frac{1}{4\epsilon}, \tag{3}$$

where $|\mathcal{S}||\mathcal{A}|(|\mathcal{S}| - 1)$ captures the degrees of freedom for transition kernels.

This quantity measures the number of bits needed to distinguish a maximally large set of ϵ-separated world models. Moreover, we measure how much information suffices to represent the world at precision ϵ. This motivates a complementary **ϵ-covering entropy:** we select a world model family of minimum cardinality such that every transition kernel lies within distance ϵ of at least one representative (Colomer et al., 2025). The corresponding upper bound on the ϵ-covering entropy is defined as follows:

Definition 5 (ϵ-covering entropy upper bound)**.** *For any distance $\epsilon \in (0, 1)$ and for the full tabular transition kernel family $\mathcal{F}$, we write the upper bound on ϵ-covering entropy as*

$$\overline{\mathbb{H}}_\epsilon(\mathcal{F}) \coloneqq |\mathcal{S}||\mathcal{A}|(|\mathcal{S}| - 1) \log_2 \left(1 + \frac{2}{\epsilon}\right), \tag{4}$$

where $|\mathcal{S}||\mathcal{A}|(|\mathcal{S}| - 1)$ captures the degrees of freedom for transition kernels.

Definition 5 characterizes the information sufficient to represent any world model to precision ϵ using a shared family of representatives, providing a **representational upper bound** for our analysis.

Remark 2. While we use entropy to characterize agent complexity, under quantization with a fixed precision, the number of bits required to represent a parameterized agent scales linearly with its parameter count. Therefore, entropy provides a tractable measure of complexity.

4 Theoretical characterizations

In this section, we will characterize how frozen and adaptive general agents differ under the big world hypothesis by comparing the complexity of these agents and their environments.

4.1 Frozen general agents

How large are frozen general agents? Recent works have shown that trained agents can retain environment-dependent knowledge in internal representations (Hafner et al., 2024; Assran et al., 2025). Hence, we abstract the persistent information retained after offline training as a code $c \in \mathcal{C}$, which together with the specified reward determines its policy π_r. Since $|\mathcal{C}|$ distinguishable codes can carry at most $\log_2 |\mathcal{C}|$ bits of persistent information, we use this persistent capacity to measure the complexity of frozen general agents $\pi_r : \mathcal{S} \times r \to \Delta(\mathcal{A})$ in this paper.

With both world and agent complexity in place, we now return to the central question posed earlier by relating the complexity of general agents to that of their environments. The following theorem shows that bounded regret across the reward family requires this capacity to at least match environment complexity at the corresponding precision.

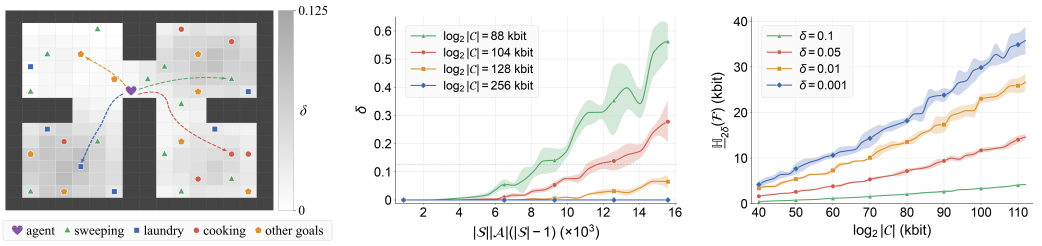
Theorem 1 (frozen agents are as complex as worlds)**.** *Suppose a frozen agent π_r satisfies bounded regret for a given regret $\delta \in (0, 0.125)$ over the reward family for any $P \in \mathcal{F}$. Then, we have*

$$\log_2 |\mathcal{C}| \geq \underline{\mathbb{H}}_{2\delta}(\mathcal{F})$$

where $\underline{\mathbb{H}}_{2\delta}(\mathcal{F})$ is the ϵ-packing entropy lower bound in Definition 4 with $\epsilon = 2\delta$.

The proof of Theorem 1 is detailed in Appendix B. This theorem reveals an intrinsic information cost of generality, and consequently, a frozen general agent must retain enough information about the dynamics to satisfy the desired regret over the entire reward family $\mathcal{R}$. Since this constraint holds regardless of the offline training procedure, *even abundant data or computational resources cannot compensate for insufficient persistent capacity*. Therefore, **frozen general agents are incompatible under the big world hypothesis**.

Furthermore, we provide a numerical analysis of Theorem 1 in Figure 3, with detailed settings specified in Appendix E. Figure 3b shows how capacity limitations can remain hidden in small worlds, where agents with substantially different capacities could achieve similarly low regret. However, their performance diverges as the world complexity grows. To maintain general capabilities, an agent's **capacity must keep pace with environment complexity**. Moreover, Figure 3c shows that the complexity of environments in which a frozen agent can sustain this performance grows approximately linearly with its capacity.



(a) (b) (c)

Figure 3: **Numerical analysis of Theorem 1.** a) visualizes the household general agent in **Example 2**, with shading regions illustrating regret. b) shows regret δ increasing with the number of degrees of freedom of the transition kernels for different agent capacities. c) shows how environment complexity scales with agent capacity for different desired regret δ. In both panels, shaded bands indicate 95% confidence intervals for the mean over 5 independently generated environments.

Moreover, how much data do frozen general agents need? Beyond retaining sufficient information, frozen general agents must acquire knowledge through offline training. We let $\boldsymbol{\tau} := (\tau_T^1, \tau_T^2, \ldots, \tau_T^n)$ denote the offline training dataset, where n is the number of trajectories used to train the agent and each trajectory contains T observed transitions. Since the queried state-action pairs (s, a) are known to the sampling procedure, each transition contributes only its resulting state. We use $\texttt{Data}$ to denote the bits required to represent those transitions. The following theorem evaluates the bounded regret guarantee in expectation over the random offline training data.

Theorem 2 (frozen agents need large amounts of data)**.** *Fix a desired regret $\delta \in (0, \frac{1}{8})$. Even with unrestricted capacity, any frozen agent π_r whose expected worst-case regret over all rewards and initial states is at most δ for any transition kernel $P \in \mathcal{F}$ must satisfy*

$$\texttt{Data} \geq \frac{7\underline{\mathbb{H}}_{\frac{1}{4(|\mathcal{S}|-1)}}(\mathcal{F})}{10368\delta^2}.$$

Here, $\underline{\mathbb{H}}_{\frac{1}{4(|\mathcal{S}|-1)}}(\mathcal{F})$ is the ϵ-packing entropy lower bound in Definition 4 with $\epsilon = \frac{1}{4(|\mathcal{S}|-1)}$.

The proof of Theorem 2 is shown in Appendix C. Intuitively, Theorem 2 shows an acquisition cost of generality. Here, the resolution $\epsilon = \frac{1}{4(|\mathcal{S}|-1)}$ matches the intrinsic dimensional scale of the tabular transition kernel family. Though we place no restriction on persistent capacity, satisfying bounded regret still requires substantial information to be acquired. In particular, the required data grows linearly with world complexity, while the cost of improving accuracy scales as $\Omega(\frac{1}{\delta^2})$. Hence, sufficient capacity alone does not make generality available offline (Silver & Sutton, 2025).

Together, Theorem 1 and 2 expose two impossibilities for frozen general agents under the big world hypothesis: the information that scales with world complexity must first be acquired offline and then retained after training. In Section 4.3, we further illustrate these impossibility results with real-world examples. This motivates a distinct possibility—acquiring the information after deployment *through interactions with the world*.

4.2 Adaptive general agents

Under the big world hypothesis, online information is necessary for generality. We consider adaptive general agents $\pi_r : h_t \times r \to \Delta(\mathcal{A})$. In this paradigm, generality is not only about how much knowledge an agent possesses, but also about its ability to acquire knowledge online in big worlds when the retained information is not enough. Specifically, we let K denote the number of online interactions and h_K the corresponding history.

Intuition. With limited persistent capacity, adaptive general agents may not retain all the environment-dependent information needed to distinguish all the world models. Instead, online learning provides an additional source of information: different environments induce different distributions over histories, and hence the *realized trajectory can progressively reduce the uncertainty relevant to future behavior* (in information theory, the **mutual information** between environments and histories will accumulate). This suggests a direct tradeoff between persistent information and online experience, with the continuation policy evaluated after K interactions:

Theorem 3 (general agents can trade capacity for online experience)**.** *For any regret $\delta \in (0,1)$ and for any finite capacity, there exists an adaptive agent π_r that satisfies bounded regret simultaneously for all $r \in \mathcal{R}$ with probability at least $\alpha \in (0,1)$ for any $P \in \mathcal{F}$ after K interactions, provided that:*

$$\log_2 |\mathcal{C}| + \frac{(1-\gamma)^4\delta^2}{512 \ln 2} K \geq 2\overline{\mathbb{H}}_{\frac{(1-\gamma)^2\delta}{16}}(\mathcal{F}) + |\mathcal{S}||\mathcal{A}| \log_2 \left(\frac{4|\mathcal{S}||\mathcal{A}|}{1-\alpha}\right)$$

where $\overline{\mathbb{H}}_{\frac{(1-\gamma)^2\delta}{16}}(\mathcal{F})$ is the ϵ-covering entropy upper bound in Definition 5 with $\epsilon = \frac{(1-\gamma)^2\delta}{16}$.

We defer the proof of Theorem 3 to Appendix D. Theoretically, Theorem 3 establishes an achievable tradeoff between persistent capacity and online learning. In particular, the two terms on the left hand side represent two distinct sources of the required information: retaining it after offline training or acquiring it through online interactions. Theorem 3 is an *existence result* established through constructions inspired by random coding: it guarantees such adaptive agents exist, but does not prescribe how this tradeoff should be realized algorithmically.

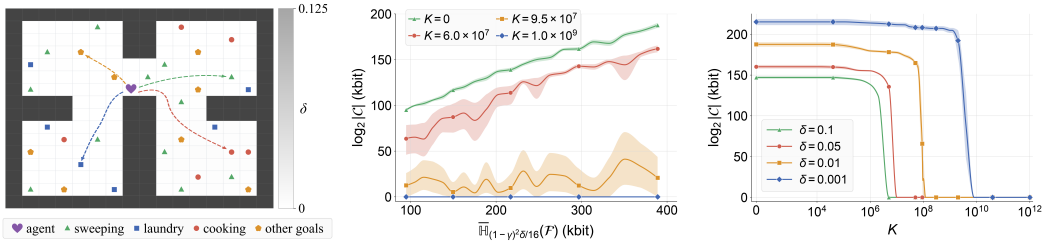
Previous work has studied hypothesis class complexity–sample complexity, memory–sample complexity, and information acquisition–regret tradeoffs in online learning and reinforcement learning (Gheshlaghi Azar et al., 2013; Steinhardt et al., 2016; Arumugam & Van Roy, 2021). Theorem 3 differs in the quantities it relates: it jointly characterizes **persistent capacity**, the **number of online interactions**, and the **resulting regret** for a single agent that must remain *general* across the reward family. Moreover, this tradeoff is governed directly by **world complexity**, capturing how online experience can substitute for persistent knowledge as worlds become increasingly complex.

We also provide a numerical analysis in Figure 4, and we detail the setup in Appendix E. Specifically, Figure 4b illustrates that *learning during deployment* changes how generality scales with world complexity: part of the growing information burden can be shifted from persistent capacity to online experience. Figure 4c further shows that the extent of this substitution depends strongly on the desired accuracy, with smaller regret making online compensation increasingly costly.

Significantly, in our reward-conditioned RL settings, Theorem 3 provides an even stronger guarantee than adaptation to a specified reward. We further discuss this distinction and the practical implications of Theorem 3 in Section 4.4.

4.3 Interpretation for impossibility results

How large are real-world systems? Javed & Sutton (2024) argue that, in many real-world decision-making problems, the environment is vastly larger than the agent, highlighting both the large volume of sensory data and interactions with other complex agents. Specifically, based on the estimated human sensory data rate, a person receives approximately 8.64×10^{13} bits of information per day



(a) (b) (c)

Figure 4: **Numerical analysis of Theorem 3.** a) visualizes the household general agent in **Example 2**, with shading regions illustrating regret. b) shows how the persistent capacity scales with world complexity at different interaction horizons K and a fixed regret $\delta = 0.01$. c) shows how the required persistent capacity decreases with interaction horizon K for different regret δ. In both figures, shaded bands indicate 95% confidence intervals for the mean over 20 independently generated environments.

(Zheng & Meister, 2024). Moreover, the environment includes other agents whose policies contribute to the complexity of its dynamics. Therefore, accurately modeling these dynamics requires predicting their responses across various situations and rewards. We illustrate this with a simplified example below drawn from a simplified *household environment*, showing that frozen general agents require astronomical capacity to achieve uniformly low regret across the reward family.

Example 2. Consider a household environment containing only 50 movable household items, each of which can occupy one of 10 admissible coarse positions. This highly simplified abstraction can admit 10^{50} distinct household configurations. A general household agent that maintains low regret across all household goals must therefore remain near-optimal across this entire configuration space since deployment may start from any of these states. For a frozen agent, this capability must be supported entirely by information retained before deployment. Even **without** invoking the full tabular transition family, this restricted environment with only 2 actions and 2 outcomes for each action already requires approximately 2×10^{50} bits of persistent information since the agent needs to distinguish $2^{|\mathcal{S}||\mathcal{A}|}$ worlds following the interpretation in **Example 1.**

Frozen general agents become prohibitive in big worlds. Javed & Sutton (2024) explicitly point out that representing value functions for *Go* across all states has no simple solution, and our simplified household example requires far more capacity than the parameter bits available in current open-weight agentic models. If all retained environment-dependent information resides in these parameters, even granting that every bit is useful, the required **capacity, measured in bits already slightly exceeds the estimated number of atoms Earth contains**. Moreover, a real-world general agent *must contend with many more interacting agents, operate in environments with vastly richer configurations, and face a much broader range of goals*, making the relevant decision-making complexity potentially far greater. Hence, frozen agents at current capacity scales cannot maintain bounded regret over general rewards in big worlds, motivating agents whose capabilities scale with online experience.

4.4 Implications of Theorem 3

Online information provides an alternative route to generality. Notably, our impossibility results only require the agent to satisfy this regret bound for **a small subset of rewards** constructed in Appendix B. In contrast, Theorem 3 gives a stronger guarantee than standard goal-conditioned reinforcement learning: whereas GCRL is evaluated with respect to the goal currently specified to the agent (Liu et al., 2022), a single interaction history suffices to support bounded regret for **every reward** $r \in \mathcal{R}$. This resembles reward-free reinforcement learning (Zhang et al., 2021), where exploration must support arbitrary downstream rewards, and its minimax interaction complexity is $\Omega(\frac{|\mathcal{S}|^2|\mathcal{A}|}{\epsilon^2})$. However, once the reward is specified, tabular reinforcement learning (Gheshlaghi Azar et al., 2013) has a minimax sample complexity scaling as $\Omega(\frac{|\mathcal{S}||\mathcal{A}|}{\epsilon^2})$. Therefore, the persistent capacity $\log_2 |\mathcal{C}|$ and interaction requirements K in Theorem 3 are *consistent* with known lower bound scalings in the world complexity and regret δ up to the logarithmic factor.

What makes adaptive agents scalable? A direct implementation of Theorem 3 could estimate the transition kernel from the history, but this would require the online memory of the same order as the persistent capacity in Theorem 1. Hence, the online memory of such direct model-based approaches

still scales with the world complexity. However, in our reward-conditioned reinforcement learning setting, classical tabular $\mathcal{Q}$-learning keeps only $\mathcal{O}(|\mathcal{S}||\mathcal{A}|)$ action-value entries (Li et al., 2021) for a specified reward. Moreover, online experience can also be acquired efficiently. Recent empirical work shows that autonomous exploration can acquire environment-specific knowledge through targeted and parallel interactions (Zhu et al., 2026). Together, these mechanisms suggest another *scalable* route to online learning in reward-conditioned reinforcement learning with reward-specific memory.

This also exposes the main algorithmic challenge left by Theorem 3: **how can we design end-to-end agents that integrate offline knowledge with online experience?** Recent systems introduce post-training adaptation, but still freeze the resulting memory for downstream goals (Zhu et al., 2026). Although in-context reinforcement learning demonstrates adaptation through interaction histories with frozen network weights (Grigsby et al., 2024a), scaling such adaptation to sustained online learning in big worlds remains an open challenge. The broader vision in Theorem 3 is to develop adaptive agents that continually update their decision-relevant memory and strategies during downstream episodes, sustaining these abilities as experience accumulates. By **end-to-end**, we mean that offline training should learn not only what information to retain, but also how to acquire, compress, and use new experience for online decision-making. Therefore, a natural future direction is to realize the capacity–experience tradeoff in Theorem 3 through adaptive agents whose capabilities could scale efficiently with online experience.

5 Related work

Big world hypothesis. The big world hypothesis posits that many decision-making environments are fundamentally more complex than the agents operating within them. Formally, agents cannot fully perceive the world, nor represent exact values or optimal policies for every possible situation (Javed & Sutton, 2024). Recent works develop big world simulators to study environments whose available information exceeds what a bounded agent can retain (Kumar et al., 2024). Moreover, computationally embedded agents formalize an intrinsic mismatch between finite agents and effectively unbounded environments (Lewandowski et al., 2025). Recent empirical studies have begun to instantiate this perspective in continual learning benchmarks in which agents are expected to keep adapting rather than converge to a fixed policy (Tang et al., 2026; Mohamed et al., 2026). In contrast, we go beyond these formulations by *deriving quantitative capacity limitations and establishing a tradeoff between offline and online learning* for general agents in big worlds.

Online agents. Recent in-context reinforcement learning methods train online learners that adapt from past interactions (Grigsby et al., 2024a). Other works study cross-domain adaptation with given model weights (Polubarov et al., 2025), or meta-train agents in randomized environments to examine how experience and task diversity affect generalization (Wang et al., 2025). While these works demonstrate particular mechanisms for adaptation, we show that, in big worlds, *online adaptation provides a scalable route for finite-capacity agents to maintain general capabilities* as world complexity grows. Moreover, under Definition 1 and Remark 1, some current ICRL methods (Chen & Paternain, 2025; Polubarov et al., 2025) that use a pre-committed mapping remain frozen if online experience does not update their continuation policies, making them structurally distinct from the scalable online learners whose existence we establish in Theorem 3.

In Appendix A, we discuss more related work in details.

6 Conclusions

In this work, we study whether generality can be sustained by frozen knowledge under the big world hypothesis. We show that frozen general agents face both capacity and offline data costs that scale with world complexity, while adaptive agents trade persistent capacity for online experience.

Takeaway. *Learning is not inherently a one-shot process*: knowledge is continually revised as experience accumulates. However, once an agent is frozen, all of its future behavior must be supported by a finite snapshot of what was learned in the past. For a frozen agent facing an open world whose complexity continues to grow after training ends, learning is effectively reduced to a one-shot transfer from the finite past to the infinite future. From this perspective, *online learning could be a fundamental way for generality to remain scalable.*

AI USE STATEMENT

In this work, we only used generative AI tools for checking correctness of proofs and polishing writing. We have not used generative AI tools for finding related work, research ideation, drafting content, generating synthetic datasets or proving mathematical claims. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We believe this work raises no ethical concerns.

REPRODUCIBILITY STATEMENT

We clearly state the problem setting and assumptions in the main text, and provide complete mathematical proofs in the appendices. The details of all the experiments are described in Appendix E.

ACKNOWLEDGMENTS

We thank all the anonymous reviewers and area chair(s) for their valuable feedback.

REFERENCES

David Abel, Andre Barreto, Benjamin Van Roy, Doina Precup, Hado van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=ZZS9WEWYbD.

David Abel, André Barreto, Hado van Hasselt, Benjamin Van Roy, Doina Precup, and Satinder Singh. On the convergence of bounded agents, 2023b. URL https://arxiv.org/abs/2307.11044.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in Atari. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.12399.

Anthropic. System card: Claude Opus 4 & Claude Sonnet 4. Technical report, Anthropic, May 2025. URL https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf.

Anthropic. Introducing Claude Opus 5.5, September 2026. URL https://www.anthropic.com/claude-opus-5-5. Released September 22, 2026. Accessed September 26, 2026.

Dilip Arumugam and Benjamin Van Roy. Deciding what to learn: A rate-distortion approach. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 373–382. PMLR, 18–24 Jul 2021.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Charles Beattie, Joel Z. Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, Julian Schrittwieser, Keith Anderson, Sarah York, Max Cant, Adam Cain, Adrian Bolton, Stephen Gaffney, Helen King, Demis Hassabis, Shane Legg, and Stig Petersen. DeepMind Lab, 2016. URL https://arxiv.org/abs/1612.03801.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279, 2013. doi: 10.1613/jair.3912. URL https://jair.org/index.php/jair/article/view/10819.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 2025. URL https://proceedings.mlr.press/v305/black25a.html.

ByteDance Seed. UI-TARS-1.5 model, 2025. URL https://huggingface.co/ByteDance-Seed/UI-TARS-1.5-7B. Official model card. Accessed September 26, 2026.

Weiqin Chen and Santiago Paternain. Random policy enables in-context reinforcement learning within trust horizons. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=mAiMKnr9r5. Featured Certification.

Pau Colomer, Christian Deppe, Holger Boche, and Andreas Winter. Deterministic identification over channels with finite output: A dimensional perspective on superlinear rates. *IEEE Transactions on Information Theory*, 71(5):3373–3396, May 2025. ISSN 1557-9654. doi: 10.1109/tit.2025.3531301. URL http://dx.doi.org/10.1109/TIT.2025.3531301.

Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl^2: Fast reinforcement learning via slow reinforcement learning, 2016. URL https://arxiv.org/abs/1611.02779.

Matteo Gallici, Mattie Fellows, Benjamin Ellis, Bartomeu Pou, Ivan Masmitja, Jakob Nicolaus Foerster, and Mario Martin. Simplifying deep temporal difference learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7IzeL0kflu.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Mach. Learn.*, 91(3): 325–349, June 2013. ISSN 0885-6125. doi: 10.1007/s10994-013-5368-1. URL https://doi.org/10.1007/s10994-013-5368-1.

GLM-5 Team. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Google DeepMind. Gemma 4, 2026a. URL https://deepmind.google/models/gemma/gemma-4/. Official model family page. Accessed September 26, 2026.

Google DeepMind. Gemini 3.8 Flash: Model card, September 2026b. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/. Published September 2, 2026. Accessed September 26, 2026.

Jake Grigsby, Linxi Fan, and Yuke Zhu. AMAGO: Scalable in-context reinforcement learning for adaptive agents. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=M6XWoEdmwf.

Jake Grigsby, Justin Sasek, Samyak Parajuli, Daniel Adebi, Amy Zhang, and Yuke Zhu. AMAGO-2: Breaking the multi-task barrier in meta-reinforcement learning with transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=OSHaRf4TVU.

Danijar Hafner. Benchmarking the spectrum of agent capabilities. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2109.06780.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024. URL https://arxiv.org/abs/2301.04104.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FzpfPa6unv.

Dan Hendrycks, Dawn Song, Christian Szegedy, Honglak Lee, Yarin Gal, Erik Brynjolfsson, Sharon Li, Andy Zou, Lionel Levine, Bo Han, Jie Fu, Ziwei Liu, Jinwoo Shin, Kimin Lee, Mantas Mazeika, Long Phan, George Ingebretsen, Adam Khoja, Cihang Xie, Olawale Salaudeen, Matthias Hein, Kevin Zhao, Alexander Pan, David Duvenaud, Bo Li, Steve Omohundro, Gabriel Alfour, Max Tegmark, Kevin McGrew, Gary Marcus, Jaan Tallinn, Eric Schmidt, and Yoshua Bengio. A definition of agi, 2025. URL https://arxiv.org/abs/2510.18212.

Tyler Ingebrand, Amy Zhang, and ufuk topcu. Zero-shot reinforcement learning via function encoders. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=tHBLwSYnLf.

Khurram Javed and Richard S. Sutton. The big world hypothesis and its ramifications for artificial intelligence. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*, 2024. URL https://openreview.net/forum?id=Sv7DazuCn8.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model, 2024. URL https://arxiv.org/abs/2406.09246.

Bryan Kolb and Ian Q. Whishaw. Brain plasticity and behavior. *Annual Review of Psychology*, 49: 43–64, 1998. doi: 10.1146/annurev.psych.49.1.43.

Saurabh Kumar, Hong Jun Jeon, Alex Lewandowski, and Benjamin Van Roy. The need for a big world simulator: A scientific challenge for continual learning. In *Finding the Frame: An RLC Workshop for Examining Conceptual Frameworks*, 2024. URL https://openreview.net/forum?id=10XMwt1nMJ.

Kuang-Huei Lee, Ofir Nachum, Mengjiao Yang, Lisa Lee, Daniel Freeman, Winnie Xu, Sergio Guadarrama, Ian Fischer, Eric Jang, Henryk Michalewski, and Igor Mordatch. Multi-game decision transformers. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2205.15241.

Alistair Letcher, Mattie Fellows, Alexander D. Goldie, Jonathan Richens, Jakob N. Foerster, and Oliver Richardson. Inverting the bellman equation: From q-values to world models, 2026. URL https://arxiv.org/abs/2606.21173.

Alex Lewandowski, Aditya A. Ramesh, Edan Meyer, Dale Schuurmans, and Marlos C. Machado. The world is bigger! a computationally-embedded perspective on the big world hypothesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=gJclyLFSdU.

Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Tightening the dependence on horizon in the sample complexity of q-learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6296–6306. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/li21b.html.

Aly Lidayan, Michael D Dennis, and Stuart Russell. BAMDP shaping: a unified theoretical framework for intrinsic motivation and reward shaping. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tijmpS9Vy2.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44776–44791, 2023. doi: 10.52202/075280-1939. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8c3c666820ea055a77726d66fc7d447f-Abstract-Datasets_and_Benchmarks.html.

Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-conditioned reinforcement learning: Problems and solutions. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 5502–5511. International Joint Conferences on Artificial Intelligence Organization, 2022. URL https://doi.org/10.24963/ijcai.2022/770.

Yikai Lu, Yifei Wu, Xinyu Lu, and Tongxin Li. World models in pieces: Structural certification for general agents. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=BfeHBPYkZ0.

Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. *Science Robotics*, 10(105):eads5033, 2025. doi: 10.1126/scirobotics.ads5033. URL https://doi.org/10.1126/scirobotics.ads5033.

Eric Mazur. *Principles & Practice of Physics*. Pearson Education, global edition, 2015. ISBN 9781292076447. URL https://api.pageplace.de/preview/DT0400.9781292076492_A25428739/preview-9781292076492_A25428739.pdf. Figure 1.9, p. 25.

Meta. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models, September 2024a. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Published September 25, 2024. Accessed September 26, 2026.

Meta. Llama-3.3-70B-Instruct, 2024b. URL https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Official model card. Accessed September 26, 2026.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vhFu1Acb0xb.

MiniMax. MiniMax-M3, 2026. URL https://huggingface.co/MiniMaxAI/MiniMax-M3. Official model card. Accessed September 26, 2026.

Mistral AI. Mistral NeMo, July 2024. URL https://mistral.ai/news/mistral-nemo/. Published July 18, 2024. Accessed September 26, 2026.

Mohamed A. Mohamed, Kateryna Nekhomiazh, Vedant Vyas, Marcos M. Jose, Andrew Patterson, and Marlos C. Machado. The cell must go on: Agar.io for continual reinforcement learning, 2026. URL https://arxiv.org/abs/2505.18347.

Mojang Studios. Minecraft: Java and Bedrock edition for PC, 2026. URL https://www.minecraft.net/en-us/store/minecraft-java-bedrock-edition-pc. Official product page. Accessed September 26, 2026.

Moonshot AI. Kimi-K2.6, 2026. URL https://huggingface.co/moonshotai/Kimi-K2.6. Official model card. Accessed September 26, 2026.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning, 2026. URL https://arxiv.org/abs/2603.14482.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI, 2026a. URL https://arxiv.org/abs/2606.02800.

NVIDIA. NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16, 2026b. URL https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16. Official model card. Accessed September 26, 2026.

OpenAI. GPT-5 system card. Technical report, OpenAI, August 2025. URL https://cdn.openai.com/gpt-5-system-card.pdf.

OpenAI. GPT-6 Astra, 2026. URL https://developers.openai.com/api/docs/models/gpt-6-astra. API model documentation. Accessed September 26, 2026.

Andrei Polubarov, Lyubaykin Nikita, Alexander Derevyagin, Ilya Zisman, Denis Tarasov, Alexander Nikulin, and Vladislav Kurenkov. Vintix: Action model via in-context reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=gi9MOXNfw2.

Qwen Team. Qwen3.5-397B-A17B, 2026. URL https://huggingface.co/Qwen/Qwen3.5-397B-A17B. Official model card for one member of the Qwen3.5 family. Accessed September 26, 2026.

Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai, 2021. URL https://arxiv.org/abs/2109.08238.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=1ikK0kHjvj.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pOoKI3ouv1.

Jonathan Richens, Tom Everitt, and David Abel. General agents need world models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=dlIoumNiXt.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9339–9347, 2019. URL https://arxiv.org/abs/1904.01201.

David Silver and Richard Sutton. Welcome to the era of experience, 2025. URL https://api.semanticscholar.org/CorpusID:277919528.

Jacob Steinhardt, Gregory Valiant, and Stefan Wager. Memory, communication, and statistical queries. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1490–1516, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/steinhardt16.html.

Richard S Sutton and Andrew G Barto. Reinforcement learning: an introduction, 2nd edn. adaptive computation and machine learning, 2018.

Steven Tang, Xinze Xiong, Anna Hakhverdyan, Andrew Patterson, Jacob Adkins, Jiamin He, Esraa Elelimy, Parham Mohammad Panahi, Martha White, and Adam White. Forager: a lightweight testbed for continual learning with partial observability in rl, 2026. URL https://arxiv.org/abs/2605.01131.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind Control Suite, 2018. URL https://arxiv.org/abs/1801.00690.

Fan Wang, Pengtao Shao, Yiming Zhang, Bo Yu, Shaoshan Liu, Ning Ding, Yang Cao, Yu Kang, and Haifeng Wang. Towards large-scale in-context reinforcement learning by meta-training in randomized worlds. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=b6ASJBXtgP.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, volume 37, pp. 52040–52094, 2024. doi: 10.52202/079017-1650. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies, 2026. URL https://arxiv.org/abs/2602.15922.

Z.ai. GLM-5.1, 2026. URL https://huggingface.co/zai-org/GLM-5.1. Official model card. Accessed September 26, 2026.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10008343.

Zihan Zhang, Simon Du, and Xiangyang Ji. Near optimal reward-free reinforcement learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12402–12412. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhang21e.html.

Jieyu Zheng and Markus Meister. The unbearable slowness of being: Why do we live at 10 bits/s?, 2024. URL https://arxiv.org/abs/2408.10234.

Sibo Zhu, Shicheng Fan, Xinyue Wang, Wenyi Wu, Kun Zhou, and Biwei Huang. Rsiagent: Autonomous exploration for recursive self-improvement in new environments, 2026. URL https://arxiv.org/abs/2609.15364.

Matthew Zurek and Yudong Chen. The plugin approach for average-reward and discounted MDPs: Optimal sample complexity analysis. In *36th International Conference on Algorithmic Learning Theory*, 2025. URL https://openreview.net/forum?id=WXGDyFsooW.

A EXTENDED RELATED WORK

World model theorems. Existing world model theorems establish identifiability results linking an agent's capabilities to the recovery of underlying dynamics: robustness to local interventions can recover approximate causal models (Richens & Everitt, 2024); sufficiently rich goals can reveal predictive dynamics (Richens et al., 2025); compositional goals can certify local transition probabilities (Lu et al., 2026); and rich action-value information can identify transition dynamics (Letcher et al., 2026). These works characterize what environmental knowledge capable agents must encode. However, our focus is not on such causality results, but on characterizing *the resource cost of acquiring and retaining such knowledge, and showing how online learning changes this scaling*.

Continual learning. Continual learning studies agents that continually adapt as their environments or goal distributions evolve over time (Abel et al., 2023a). Previous work on bounded agents revisits notions of convergence under persistent adaptation (Abel et al., 2023b), and recent benchmarks evaluate continual learning under partial observability (Tang et al., 2026). While these works characterize adaptation to evolving environments and goals, we focus on *when online adaptation becomes necessary for generality even under stationary dynamics*.

B PROOF OF THEOREM 1

The proof of Theorem 1 is divided into **2** steps. **Step 1:** we first generalize the world model theorem in MDP settings, where previous world model theorems are shown in controlled Markov Process (cMP) settings (Richens et al., 2025; Lu et al., 2026), $\mathcal{Q}$ learning settings (Letcher et al., 2026), and causality settings (Richens & Everitt, 2024). **Step 2:** finally, we lower bound agent size by world model complexity using packing arguments.

Step 1: we show the world model theorem in MDP settings. Consider an arbitrary helper constant $z \in [0, 1]$. We begin by constructing the bounded transition reward r as follows:

$$r_{s,a,s',z}(x, y, x') = z + \mathbf{1}\{x = s, y = a\}(\mathbf{1}\{x' = s'\} - z). \tag{5}$$

Naturally, $r_{s,a,s',z}(x, y, x') = \mathbf{1}\{x' = s'\}$ once (s, a) is the desired pair, and $r_{s,a,s',z}(x, y, x') = z$ otherwise. It is easy to verify that the construction $r_{s,a,s',z}(x, y, x') \in [0, 1]$.

Since we focus on finite MDPs, we have

$$P(s' \mid s, a) = \sum_{s_{t+1} \in \mathcal{S}} \mathbf{1}\{s_{t+1} = s'\} P(s_{t+1} \mid s_t = s, a_t = a). \tag{6}$$

We now fix an arbitrary **Markovian** deterministic policy π_r conditioned on this construction. Applying *Eq.(5)* to the value function gives:

$$\begin{aligned}
\mathcal{V}^{\pi_r}(s) &= \mathbb{E}_{\pi_r}\Big(\sum_{t=0}^{\infty} \gamma^t r_{s,a,s',z}(s_t, a_t, s_{t+1}) \mid s_0 = s \Big) \\
&= \mathbb{E}_{\pi_r}\left(\sum_{t=0}^{\infty} \gamma^t \Big(z + \mathbf{1}\{s_t = s, a_t = a\}(\mathbf{1}\{s_{t+1} = s'\} - z) \Big) \mid s_0 = s \right) \\
&= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\pi_r}\Big(z + \mathbf{1}\{s_t = s, a_t = a\}(\mathbf{1}\{s_{t+1} = s'\} - z) \mid s_0 = s \Big) \\
&= \sum_{t=0}^{\infty} \gamma^t z + \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\pi_r}\Big(\mathbf{1}\{s_t = s, a_t = a\}(\mathbf{1}\{s_{t+1} = s'\} - z) \mid s_0 = s \Big),
\end{aligned} \tag{7}$$

where the last equality is valid since two separated terms are both finite. Then, applying *Eq.(6)* to the second term, we obtain:

$$\begin{aligned}
&\mathbb{E}_{\pi_r}\Big(\mathbf{1}\{s_t = s, a_t = a\}(\mathbf{1}\{s_{t+1} = s'\} - z) \Big) \\
&\quad = \mathbb{E}_{\pi_r}\Big(\mathbb{E}_{\pi_r}\big(\mathbf{1}\{s_t = s, a_t = a\}(\mathbf{1}\{s_{t+1} = s'\} - z) \mid s_t, a_t \big) \Big) \\
&\quad = \mathbb{E}_{\pi_r}\Big(\mathbf{1}\{s_t = s, a_t = a\} \sum_{s_{t+1} \in \mathcal{S}} (\mathbf{1}\{s_{t+1} = s'\} - z) P(s_{t+1} \mid s_t, a_t) \Big) \\
&\quad = \mathbb{E}_{\pi_r}\Big(\mathbf{1}\{s_t = s, a_t = a\} \big(\sum_{s_{t+1} \in \mathcal{S}} \mathbf{1}\{s_{t+1} = s'\} P(s_{t+1} \mid s_t, a_t) - z \sum_{s_{t+1} \in \mathcal{S}} P(s_{t+1} \mid s_t, a_t) \big) \Big) \\
&\quad = \mathbb{E}_{\pi_r}\Big(\mathbf{1}\{s_t = s, a_t = a\} \big(P(s' \mid s, a) - z \big) \Big).
\end{aligned} \tag{8}$$

Combining *Eq.(7)* and *Eq.(8)*, we have

$$\mathcal{V}^{\pi_r}(s) = \frac{z}{1 - \gamma} + \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\pi_r}\Big(\mathbf{1}\{s_t = s, a_t = a\} \big(P(s' \mid s, a) - z \big) \Big). \tag{9}$$

Based on this decomposition, we only need to use the following lemmas to constrain the policy π_r using bounded regret shown in *Eq.(1)*, then we can show world model theorems.

Lemma B.1 (Proof in Appendix B.1)**.** *We fix an arbitrary deterministic policy $\pi_r : \mathcal{S} \times r \to \mathcal{A}$. Suppose π_r satisfies Eq.(1) for $r_{s,a,s',z}$, then the reward conditioned policy π_r must take action a at state s if*

$$P(s' \mid s, a) - z > \delta. \tag{10}$$

Moreover, we can prove a complementary constraint on policy π_r:

Lemma B.2 (Proof in Appendix B.2)**.** *We fix an arbitrary deterministic policy $\pi_r : \mathcal{S} \times r \to \mathcal{A}$. Suppose π_r satisfies Eq.(1) for $r_{s,a,s',z}$, then the reward conditioned policy π_r must **not** take action a at state s if*

$$P(s' \mid s, a) - z < -\delta. \tag{11}$$

Finally, we can prove world model theorems in MDP settings using lemmas stated above.

Lemma B.3 (Proof in Appendix B.3, world model theorems)**.** *Suppose a reward conditioned policy π_r satisfies bounded regret for all transition rewards r under two transition kernels P and P', then we have the following inequality holds:*

$$\|P - P'\|_{\infty,\mathrm{TV}} \leq 2\delta. \tag{12}$$

Importantly, world model theorems shown above do not require general agents to handle arbitrarily complex goals (Richens et al., 2025; Lu et al., 2026), but only require general agents to satisfy bounded regret over all possible rewards in the sense that we do **not exclude myopic agents** ($\gamma \to 0$).

Step 2: we lower bound the capacity of the persistent code. For any transition kernel $P_\theta \in \mathcal{F}$, let $c_{P_\theta} \in \mathcal{C}$ denote persistent information retained by Markovian policy π_r after offline training under P_θ. We further use θ' to denote the transition kernel that must be distinguished by π_r. Applying Lemma B.3, suppose $c_{P_\theta} = c_{P_{\theta'}}$, then

$$\|P_\theta - P_{\theta'}\|_{\infty,\mathrm{TV}} \leq 2\delta. \tag{13}$$

We then choose a 2δ-separated world model family $\mathcal{F}_{2\delta} \subseteq \mathcal{F}$ of maximum cardinality. For two distinct $P_\theta, P_{\theta'} \in \mathcal{F}_{2\delta}$, we have $c_{P_\theta} \neq c_{P_{\theta'}}$. Therefore,

$$|\mathcal{F}_{2\delta}| = |c_{P_\theta} : P_\theta \in \mathcal{F}_{2\delta}| \leq |\mathcal{C}|. \tag{14}$$

For any $P_{\theta'} \in \mathcal{F}$, we define $\mathcal{B}_\epsilon(P_{\theta'}) \coloneqq \{P_\theta \in \mathcal{F} : \|P_\theta - P_{\theta'}\|_{\infty,\mathrm{TV}} \leq \epsilon\}$. Every $P_\theta \in \mathcal{F}$ lies in one of the balls with centers in $\mathcal{F}_{2\delta}$, otherwise $\mathcal{F}_{2\delta} \cup \{P_\theta\}$ would be 2δ-separated with cardinality $|\mathcal{F}_{2\delta}| + 1$. Consequently, we have $\mathcal{F} = \bigcup_{P_{\theta'} \in \mathcal{F}_{2\delta}} \mathcal{B}_{2\delta}(P_{\theta'})$.

We measure volume in transition probability coordinates. Enumerate the states as $s^1, \ldots, s^{|\mathcal{S}|}$ and define the coordinate simplex as:

$$\Delta_{|\mathcal{S}|-1} \coloneqq \{x \in \mathbb{R}^{|\mathcal{S}|-1} : x_i \geq 0,\ \sum_{i=1}^{|\mathcal{S}|-1} x_i \leq 1\}. \tag{15}$$

For every (s, a), the vector $(P_\theta(s^i \mid s, a))_{i=1}^{|\mathcal{S}|-1}$ lies in this simplex, and $P_\theta(s^{|\mathcal{S}|} \mid s, a) = 1 - \sum_{i=1}^{|\mathcal{S}|-1} P_\theta(s^i \mid s, a)$. Let vol denote Lebesgue measure in these $|\mathcal{S}| - 1$ coordinates for a row, and in $|\mathcal{S}||\mathcal{A}|(|\mathcal{S}| - 1)$ coordinates for a kernel. All products over (s, a) range over $\mathcal{S} \times \mathcal{A}$. Since $[0, 1/(|\mathcal{S}| - 1)]^{|\mathcal{S}|-1} \subseteq \Delta_{|\mathcal{S}|-1} \subseteq [0, 1]^{|\mathcal{S}|-1}$,

$$0 \le \mathrm{vol}(\Delta_{|\mathcal{S}|-1}) \le 1,$$
$$\mathrm{vol}(\mathcal{F}) = \prod_{(s,a)} \mathrm{vol}(\Delta_{|\mathcal{S}|-1}) = \big[\mathrm{vol}(\Delta_{|\mathcal{S}|-1})\big]^{|\mathcal{S}||\mathcal{A}|} > 0. \tag{16}$$

Fix $P_{\theta'} \in \mathcal{F}$ and $P_\theta \in \mathcal{B}_\epsilon(P_{\theta'})$. For every (s,a), expanding total variation gives

$$\begin{aligned}
&\sum_{i=1}^{|\mathcal{S}|-1} |P_\theta(s^i \mid s,a) - P_{\theta'}(s^i \mid s,a)| \\
&\le \sum_{i=1}^{|\mathcal{S}|-1} |P_\theta(s^i \mid s,a) - P_{\theta'}(s^i \mid s,a)| + |1 - \sum_{i=1}^{|\mathcal{S}|-1} P_\theta(s^i \mid s,a) - 1 + \sum_{i=1}^{|\mathcal{S}|-1} P_{\theta'}(s^i \mid s,a)| \\
&= \sum_{i=1}^{|\mathcal{S}|} |P_\theta(s^i \mid s,a) - P_{\theta'}(s^i \mid s,a)| \le 2\epsilon.
\end{aligned} \tag{17}$$

By the definition of $\Delta_{|\mathcal{S}|-1}$ in *Eq.(15)*, the reflected region is therefore contained in the scaled simplex $2\epsilon\Delta_{|\mathcal{S}|-1}$. Scaling each of its $|\mathcal{S}|-1$ coordinates by 2ϵ gives

$$\mathrm{vol}(2\epsilon\Delta_{|\mathcal{S}|-1}) = (2\epsilon)^{|\mathcal{S}|-1}\,\mathrm{vol}(\Delta_{|\mathcal{S}|-1}). \tag{18}$$

Since there are at most $2^{|\mathcal{S}|-1}$ coordinate patterns, *Eq.(18)* yields

$$\begin{aligned}
&\mathrm{vol}\left(\left\{\big(P_\theta(s^i \mid s,a) - P_{\theta'}(s^i \mid s,a)\big)_{i=1}^{|\mathcal{S}|-1} : P_\theta \in \mathcal{B}_\epsilon(P_{\theta'})\right\}\right) \\
&\le 2^{|\mathcal{S}|-1}\,\mathrm{vol}(2\epsilon\Delta_{|\mathcal{S}|-1}) \\
&= 2^{|\mathcal{S}|-1}(2\epsilon)^{|\mathcal{S}|-1}\,\mathrm{vol}(\Delta_{|\mathcal{S}|-1}) \\
&= (4\epsilon)^{|\mathcal{S}|-1}\,\mathrm{vol}(\Delta_{|\mathcal{S}|-1}).
\end{aligned} \tag{19}$$

Applying *Eq.(19)* to every (s,a) gives

$$\begin{aligned}
\mathrm{vol}(\mathcal{B}_\epsilon(P_{\theta'})) &= \mathrm{vol}\left(\left\{\big(P_\theta(s^i \mid s,a) - P_{\theta'}(s^i \mid s,a)\big)_{(s,a),\,1\le i\le|\mathcal{S}|-1} : P_\theta \in \mathcal{B}_\epsilon(P_{\theta'})\right\}\right) \\
&\le \prod_{(s,a)} (4\epsilon)^{|\mathcal{S}|-1}\,\mathrm{vol}(\Delta_{|\mathcal{S}|-1}) \\
&= (4\epsilon)^{|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)}\big[\mathrm{vol}(\Delta_{|\mathcal{S}|-1})\big]^{|\mathcal{S}||\mathcal{A}|} \\
&= (4\epsilon)^{|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)}\,\mathrm{vol}(\mathcal{F}).
\end{aligned} \tag{20}$$

Applying the definition of $\mathcal{B}_\epsilon(P_{\theta'})$ to *Eq.(20)* with $\epsilon = 2\delta$, we obtain

$$\begin{aligned}
1 &= \frac{\mathrm{vol}(\mathcal{F})}{\mathrm{vol}(\mathcal{F})} \\
&= \frac{\mathrm{vol}\left(\bigcup_{P_{\theta'}\in\mathcal{F}_{2\delta}} \mathcal{B}_{2\delta}(P_{\theta'})\right)}{\mathrm{vol}(\mathcal{F})} \\
&\le \sum_{P_{\theta'}\in\mathcal{F}_{2\delta}} \frac{\mathrm{vol}(\mathcal{B}_{2\delta}(P_{\theta'}))}{\mathrm{vol}(\mathcal{F})}
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{P_{\theta'} \in \mathcal{F}_{2\delta}} \frac{(4 \cdot 2\delta)^{|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)} \operatorname{vol}(\mathcal{F})}{\operatorname{vol}(\mathcal{F})} \\
&= |\mathcal{F}_{2\delta}| (8\delta)^{|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)} \\
&\leq |\mathcal{C}| (8\delta)^{|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)}, \qquad (21)
\end{aligned}
$$

where the last inequality uses *Eq.(14)*. Rearranging and taking base two logarithms gives

$$
\begin{aligned}
\log_2 |\mathcal{C}| &\geq \log_2 \left((8\delta)^{-|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)} \right) \\
&= -|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1) \log_2 (8\delta) \\
&= |\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1) \log_2 \frac{1}{4(2\delta)} \\
&= \underline{\mathbb{H}}_{2\delta}(\mathcal{F}), \qquad (22)
\end{aligned}
$$

which completes the proof.

B.1 Proof of lemma B.1

We prove this by contradiction: we assume $\pi_r(y = a \mid x = s) = 0$. Then, we have

$$
\mathbf{1}\{s_t = s, a_t = a\} = 0 \qquad (23)
$$

for every trajectory τ and every $t \geq 0$ since π_r is stationary and deterministic.

Naturally, combining *Eq.(9)* and *Eq.(23)*, we have $\mathcal{V}^{\pi_r}(s) = \frac{z}{1-\gamma}$.

We now construct a Markovian deterministic policy $\tilde{\pi}_r$. When at state $x = s$, we require $\tilde{\pi}_r(y = a \mid x = s) = 1$, and we do not make any constraint on it when $x \neq s$. We have

$$
\begin{aligned}
\mathcal{V}^{\tilde{\pi}_r}(s) &= \frac{z}{1-\gamma} + \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\tilde{\pi}_r} \Big(\mathbf{1}\{s_t = s, a_t = a\} \big(P(s' \mid s, a) - z \big) \Big) \\
&= \frac{z}{1-\gamma} + \mathbb{E}_{\tilde{\pi}_r} \Big(\mathbf{1}\{s_0 = s, a_0 = a\} \big(P(s' \mid s, a) - z \big) \Big) \\
&\quad + \sum_{t=1}^{\infty} \gamma^t \mathbb{E}_{\tilde{\pi}_r} \Big(\mathbf{1}\{s_t = s, a_t = a\} \big(P(s' \mid s, a) - z \big) \Big) \\
&= \frac{z}{1-\gamma} + \big(P(s' \mid s, a) - z \big) \\
&\quad + \sum_{t=1}^{\infty} \gamma^t \mathbb{E}_{\tilde{\pi}_r} \Big(\mathbf{1}\{s_t = s, a_t = a\} \big(P(s' \mid s, a) - z \big) \Big) \\
&\geq \frac{z}{1-\gamma} + \big(P(s' \mid s, a) - z \big). \qquad (24)
\end{aligned}
$$

Applying *Eq.(24)*, we obtain

$$
\begin{aligned}
\mathcal{V}^*(s) - \mathcal{V}^{\pi_r}(s) &\geq \mathcal{V}^{\tilde{\pi}_r}(s) - \mathcal{V}^{\pi_r}(s) \\
&\geq \frac{z}{1-\gamma} + \big(P(s' \mid s, a) - z \big) - \frac{z}{1-\gamma} \\
&= P(s' \mid s, a) - z \\
&> \delta, \qquad (25)
\end{aligned}
$$

which contradicts to *Eq.(1)*. Hence, we complete the proof.

B.2 PROOF OF LEMMA B.2

The proof is similar to the last lemma: we assume $\pi_r(y = a \mid x = s) = 1$. We then have

$$\mathbf{1}\{s_0 = s, a_0 = a\} = 1. \tag{26}$$

since π_r is stationary and deterministic.

We can now upper bound the value function of π_r using *Eq.(26)*:

$$\begin{aligned}
\mathcal{V}^{\pi_r}(s) &= \frac{z}{1-\gamma} + \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\pi_r}\Big(\mathbf{1}\{s_t = s, a_t = a\}\big(P(s' \mid s, a) - z\big)\Big) \\
&= \frac{z}{1-\gamma} + \mathbb{E}_{\pi_r}\Big(\mathbf{1}\{s_0 = s, a_0 = a\}\big(P(s' \mid s, a) - z\big)\Big) \\
&+ \sum_{t=1}^{\infty} \gamma^t \mathbb{E}_{\pi_r}\Big(\mathbf{1}\{s_t = s, a_t = a\}\big(P(s' \mid s, a) - z\big)\Big) \\
&= \frac{z}{1-\gamma} + \big(P(s' \mid s, a) - z\big) \\
&+ \sum_{t=1}^{\infty} \gamma^t \mathbb{E}_{\pi_r}\Big(\mathbf{1}\{s_t = s, a_t = a\}\big(P(s' \mid s, a) - z\big)\Big) \\
&\leq \frac{z}{1-\gamma} + \big(P(s' \mid s, a) - z\big)
\end{aligned} \tag{27}$$

since $P(s' \mid s, a) - z \leq 0$.

Then, we construct a Markovian deterministic policy $\tilde{\pi}_r$. When at state $x = s$, we require $\tilde{\pi}_r(y = a \mid x = s) = 0$, and we do not make any constraint on it when $x \neq s$. Obviously, we then have $\mathcal{V}^{\tilde{\pi}_r}(s) = \frac{z}{1-\gamma}$.

Applying *Eq.(27)*, we obtain:

$$\begin{aligned}
\mathcal{V}^*(s) - \mathcal{V}^{\pi_r}(s) &\geq \mathcal{V}^{\tilde{\pi}_r}(s) - \mathcal{V}^{\pi_r}(s) \\
&\geq \frac{z}{1-\gamma} - \Big(\big(P(s' \mid s, a) - z\big) + \frac{z}{1-\gamma}\Big) \\
&= -P(s' \mid s, a) + z \\
&> \delta,
\end{aligned} \tag{28}$$

contradicting to *Eq.(1)*, which proves the claim.

B.3 PROOF OF LEMMA B.3

We use the above lemmas to do binary search within $[0, 1]$ to prove world model theorems.

First, initializing $l_0 = 0$ and $u_0 = 1$, we define the midpoint as

$$z_k = \frac{l_k + u_k}{2}$$

and construct r_{s,a,s',z_k}.

At iteration $k = 0, 1, \ldots, n-1$, we query the policy π_r and do the following updates. If $\pi_r(s) = a$, then update $l_{k+1} = z_k$; and we update $u_{k+1} = z_k$ otherwise.

We can then prove by **induction** that, for every $k \geq 0$, the following inequality holds:

$$l_k - \delta \leq P(s' \mid s, a) \leq u_k + \delta. \tag{29}$$

Consider base case: when $k = 0$, the claim obviously holds since $P(s' \mid s, a) \in [0, 1]$.

Consider inductive steps: Suppose *Eq.(29)* holds at iteration time k, we consider the iteration at time $k + 1$.

If $\pi_r(s) = a$, then the contrapositive of Lemma B.2 gives

$$l_{k+1} \leq P(s' \mid s, a) + \delta \tag{30}$$

since $l_{k+1} = z_k$. The upper bound remains valid since $u_{k+1} = u_k$.

If $\pi_r(s) \neq a$, then the contrapositive of Lemma B.1 gives

$$u_{k+1} \geq P(s' \mid s, a) - \delta \tag{31}$$

since $u_{k+1} = z_k$. The lower bound remains valid since $l_{k+1} = l_k$.

Combining *Eq.(30)* and *Eq.(31)*, we can conclude *Eq.(29)*.

Moreover, since every update halves the interval $[0, 1]$, we have

$$u_n - l_n = 2^{-n}. \tag{32}$$

We then define the estimate transition kernel as:

$$\hat{P}_n(s' \mid s, a) = \frac{l_n + u_n}{2}.$$

Applying the lower bound in *Eq.(29)*, we obtain:

$$\begin{aligned} \hat{P}_n(s' \mid s, a) - P(s' \mid s, a) &\leq \frac{l_n + u_n}{2} - (l_n - \delta) \\ &= \delta + \frac{u_n - l_n}{2} \\ &= \delta + 2^{-n-1}. \end{aligned}$$

Similarly, applying the upper bound in *Eq.(29)*, we have:

$$\begin{aligned} P(s' \mid s, a) - \hat{P}_n(s' \mid s, a) &\leq (u_n + \delta) - \frac{l_n + u_n}{2} \\ &= \delta + \frac{u_n - l_n}{2} \\ &= \delta + 2^{-n-1}. \end{aligned}$$

Putting two pieces together, we obtain

$$|\hat{P}_n(s' \mid s, a) - P(s' \mid s, a)| \leq \delta + 2^{-n-1}. \tag{33}$$

Naturally, we take $n \to \infty$, then binary search algorithms can at most determine estimate transition kernels $\hat{P}(s' \mid s, a)$ with an upper bounded error δ.

We now prove *Eq.(12)* toward a **contradiction**: suppose that, there exists two transition kernels P and P' that have a TV distance larger than 2δ, and the identical policy π_r behaves exactly the same in those two transition kernels.

Consequently, there exists a target state space $\tilde{\mathcal{S}} \subseteq \mathcal{S}$ and a state-action pair $(\tilde{s}, \tilde{a})$ such that $P(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a}) - P'(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a}) > 2\delta$. We construct $r_{\tilde{s},\tilde{a},\tilde{\mathcal{S}},z} = z + \mathbf{1}\{x = \tilde{s}, y = \tilde{a}\}(\mathbf{1}\{x' \in \tilde{\mathcal{S}}\} - z)$, and we assume without loss of generality that:

$$P(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a}) > P'(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a})$$

and set $z = \frac{P(\tilde{s}' \in \tilde{\mathcal{S}} | \tilde{s}, \tilde{a}) + P'(\tilde{s}' \in \tilde{\mathcal{S}} | \tilde{s}, \tilde{a})}{2}$. It is obvious that Lemma B.1 and B.2 also hold for this reward. Then, we have

$$\begin{aligned} P(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a}) - z &= P(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a}) - \frac{P(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a}) + P'(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a})}{2} \\ &= \frac{P(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a}) - P'(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a})}{2} \\ &> \delta. \end{aligned} \tag{34}$$

However, we also have

$$\begin{aligned} P'(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a}) - z &= P'(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a}) - \frac{P(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a}) + P'(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a})}{2} \\ &= -\frac{P(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a}) - P'(\tilde{s}' \in \tilde{\mathcal{S}} \mid \tilde{s}, \tilde{a})}{2} \\ &< -\delta. \end{aligned} \tag{35}$$

We substitute Lemma B.1 to *Eq.(34)* under P, which gives $\pi_r(\tilde{a} \mid \tilde{s}) = 1$.

However, we can also apply Lemma B.2 over P' to *Eq.(35)*, which gives $\pi_r(\tilde{a} \mid \tilde{s}) = 0$.

Hence we prove the claim since π_r is deterministic.

C PROOF OF THEOREM 2

The proof constructs an adversarial world model family and is divided into two parts. **Part 1**: we lower bound the number of transition observations required by bounded regret δ by constructing adversarial transition kernels. **Part 2**: we convert this observation lower bound into the data size `Data` using the fixed width coding.

Part 1: lower bound transition observations required by bounded regret δ. We enumerate $2\lfloor \frac{|\mathcal{S}|}{2} \rfloor$ states as $s^1, s^2, \ldots, s^{2\lfloor |\mathcal{S}|/2 \rfloor}$. We construct a sign array $\boldsymbol{\sigma} = \left(\sigma_{(s,a),l}\right)_{(s,a)\in\mathcal{S}\times\mathcal{A}, 1\le l\le \lfloor |\mathcal{S}|/2\rfloor} \in \{-1,1\}^{|\mathcal{S}||\mathcal{A}|\lfloor |\mathcal{S}|/2\rfloor}$. We then define $P^{\boldsymbol{\sigma}}$ by:

$$P^{\boldsymbol{\sigma}}(s^{2l-1} \mid s,a) \coloneqq \frac{1+6\sigma_{(s,a),l}\delta}{2\lfloor |\mathcal{S}|/2\rfloor};$$
$$P^{\boldsymbol{\sigma}}(s^{2l} \mid s,a) \coloneqq \frac{1-6\sigma_{(s,a),l}\delta}{2\lfloor |\mathcal{S}|/2\rfloor}$$

for every $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $l \in \{1,\ldots,\lfloor |\mathcal{S}|/2\rfloor\}$. We then assign the transition kernel P of the remaining state $s^{|\mathcal{S}|}$ with probability zero. We can easily verify the validity of this construction.

We now construct the adversarial transition kernels. For any index $(s,a),l$ we constructed above, we consider the sign array $\tilde{\boldsymbol{\sigma}}$ obtained by flipping:

$$\tilde{\sigma}_{(\tilde{s},\tilde{a}),\tilde{l}} \coloneqq \begin{cases} -\sigma_{(s,a),l}, & ((\tilde{s},\tilde{a}),\tilde{l}) = ((s,a),l), \\ \sigma_{(s,a),l}, & ((\tilde{s},\tilde{a}),\tilde{l}) \neq ((s,a),l). \end{cases}$$

Notice that this $\boldsymbol{\sigma}$ is not the value of parameter θ, and one can construct $\theta_{(s,a),l} = P^{\boldsymbol{\sigma}}(s^l \mid s,a)$ to parameterize since the construction is valid.

The transition kernels $P^{\boldsymbol{\sigma}}$ and $P^{\tilde{\boldsymbol{\sigma}}}$ differ only in the probabilities of s^{2l-1} and s^{2l} in (s,a). Hence,

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}\left(P^{\boldsymbol{\sigma}}(\cdot \mid s,a) \| P^{\tilde{\boldsymbol{\sigma}}}(\cdot \mid s,a)\right) &= \frac{1+6\delta}{2\lfloor |\mathcal{S}|/2\rfloor} \ln\frac{1+6\delta}{1-6\delta} + \frac{1-6\delta}{2\lfloor |\mathcal{S}|/2\rfloor} \ln\frac{1-6\delta}{1+6\delta} \\
&= \frac{(1+6\delta)-(1-6\delta)}{2\lfloor |\mathcal{S}|/2\rfloor} \ln\frac{1+6\delta}{1-6\delta} \\
&= \frac{6\delta}{\lfloor |\mathcal{S}|/2\rfloor} \ln\frac{1+6\delta}{1-6\delta} \\
&= \frac{6\delta}{\lfloor |\mathcal{S}|/2\rfloor} \left(2\sum_{t=0}^{\infty} \frac{(6\delta)^{2t+1}}{2t+1}\right) \\
&\le \frac{12\delta}{\lfloor |\mathcal{S}|/2\rfloor} \sum_{t=0}^{\infty} (6\delta)^{2t+1} \\
&= \frac{12\delta}{\lfloor |\mathcal{S}|/2\rfloor} \frac{6\delta}{1-36\delta^2} \\
&\le \frac{12\delta}{\lfloor |\mathcal{S}|/2\rfloor} \frac{6\delta}{1-(3/4)^2} \\
&= \frac{1152\delta^2}{7\lfloor |\mathcal{S}|/2\rfloor}.
\end{aligned} \tag{36}$$

Let $\mathbb{P}^{\boldsymbol{\sigma}}$ denote the probabilistic distribution of trajectories τ with time step T under $P^{\boldsymbol{\sigma}}$. Formally, $\boldsymbol{\tau}_T \coloneqq (\tau_T^1, \ldots, \tau_T^n)$ is the set of all training trajectories τ. Therefore,

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}^{\boldsymbol{\sigma}}\|\mathbb{P}^{\tilde{\boldsymbol{\sigma}}}\right) &= \sum_{i=1}^{n}\sum_{t=0}^{T-1}\mathbb{E}_{\boldsymbol{\sigma}}\Big(\mathbb{D}_{\mathrm{KL}}\big(P^{\boldsymbol{\sigma}}(\cdot\mid s_{i,t},a_{i,t})\|P^{\tilde{\boldsymbol{\sigma}}}(\cdot\mid s_{i,t},a_{i,t})\big)\Big)\\ &= \sum_{i=1}^{n}\sum_{t=0}^{T-1}\mathbb{P}^{\boldsymbol{\sigma}}(s_{i,t}=s,a_{i,t}=a)\mathcal{D}_{\mathrm{KL}}\big(P^{\boldsymbol{\sigma}}(\cdot\mid s,a)\|P^{\tilde{\boldsymbol{\sigma}}}(\cdot\mid s,a)\big)\\ &\le \frac{1152\delta^2}{7\lfloor|\mathcal{S}|/2\rfloor}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\mathbb{P}^{\boldsymbol{\sigma}}(s_{i,t}=s,\,a_{i,t}=a).\end{aligned} \tag{37}$$

To simplify the notation, we write $\sum_{\boldsymbol{\sigma}}$ to sum over all possible values in our constructions. We now apply the above equation to Pinsker's inequality:

$$\begin{aligned}&\frac{1}{2^{|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}\sum_{\boldsymbol{\sigma}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{l=1}^{\lfloor|\mathcal{S}|/2\rfloor}\left\|\mathbb{P}^{\boldsymbol{\sigma}}-\mathbb{P}^{\tilde{\boldsymbol{\sigma}}}\right\|_{\mathrm{TV}}\\ &\le\left[\frac{1}{2^{|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}\sum_{\boldsymbol{\sigma}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{l=1}^{\lfloor|\mathcal{S}|/2\rfloor}\frac{1}{2}\mathcal{D}_{\mathrm{KL}}\left(\mathbb{P}^{\boldsymbol{\sigma}}\|\mathbb{P}^{\tilde{\boldsymbol{\sigma}}}\right)\right]^{\frac{1}{2}}\\ &\le\left[\frac{576\delta^2}{7\cdot 2^{|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}\sum_{\boldsymbol{\sigma}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{l=1}^{\lfloor|\mathcal{S}|/2\rfloor}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\frac{\mathbb{P}^{\boldsymbol{\sigma}}(s_{i,t}=s,\,a_{i,t}=a)}{\lfloor|\mathcal{S}|/2\rfloor}\right]^{\frac{1}{2}}\\ &=\left[\frac{576\delta^2}{7\cdot 2^{|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}\sum_{\boldsymbol{\sigma}}\sum_{i=1}^{n}\sum_{t=0}^{T-1}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mathbb{P}^{\boldsymbol{\sigma}}(s_{i,t}=s,\,a_{i,t}=a)\right]^{\frac{1}{2}}\\ &=\left[\frac{576\delta^2}{7\cdot 2^{|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}\sum_{\boldsymbol{\sigma}}\sum_{i=1}^{n}\sum_{t=0}^{T-1}1\right]^{\frac{1}{2}}\\ &=\left[\frac{576nT\delta^2}{7|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}\right]^{\frac{1}{2}}\\ &=24\delta\sqrt{\frac{nT}{7|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}}.\end{aligned} \tag{38}$$

For each realization of the offline training data, let π_r denote the frozen policy produced by the training procedure. Among the finite family of transition kernels constructed above, we define $\tilde{P}$ as a transition kernel under which this realized policy has the smallest worst case regret.

Since the true transition kernel $P^{\boldsymbol{\sigma}}$ is itself contained in this finite family, the worst case regret of the same realized policy under $\tilde{P}$ is no larger than that under $P^{\boldsymbol{\sigma}}$. Therefore, Lemma B.3 applied with this policy's realized worst case regret under $P^{\boldsymbol{\sigma}}$ as the tolerance, bounds the transition kernel distance by twice that regret. Taking expectations gives:

$$\mathbb{E}_{\boldsymbol{\sigma}}\|\tilde{P}-P^{\boldsymbol{\sigma}}\|_{\mathrm{TV},\infty}\le 2\delta. \tag{39}$$

For all $(s,a),l$ with $\sigma_{(s,a),l}=1$, we can obtain the lower bound of the distance:

$$\begin{aligned}&\frac{1}{2^{|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}|\mathcal{S}||\mathcal{A}|}\sum_{\boldsymbol{\sigma}}\mathbb{E}_{\boldsymbol{\sigma}}\Big(\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\|\tilde{P}(\cdot\mid s,a)-P^{\boldsymbol{\sigma}}(\cdot\mid s,a)\|_{\mathrm{TV}}\Big)\\ &=\frac{1}{2^{|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}|\mathcal{S}||\mathcal{A}|}\sum_{\boldsymbol{\sigma}}\mathbb{E}_{\boldsymbol{\sigma}}\Big(\frac{1}{2}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}|\tilde{P}(s'\mid s,a)-P^{\boldsymbol{\sigma}}(s'\mid s,a)|\Big)\end{aligned}$$

$$
\begin{aligned}
&\geq \frac{1}{2^{|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}|\mathcal{S}||\mathcal{A}|}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{l=1}^{\lfloor|\mathcal{S}|/2\rfloor}\sum_{\boldsymbol{\sigma}_{(s,a),l}=1}\left(\frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\Big(|\tilde{P}(s^{2l-1}\mid s,a)-P^{\boldsymbol{\sigma}}(s^{2l-1}\mid s,a)|\right.\\
&\left.+|\tilde{P}(s^{2l}\mid s,a)-P^{\boldsymbol{\sigma}}(s^{2l}\mid s,a)|\Big)+\frac{1}{2}\mathbb{E}_{\tilde{\boldsymbol{\sigma}}}\Big(|\tilde{P}(s^{2l-1}\mid s,a)-P^{\tilde{\boldsymbol{\sigma}}}(s^{2l-1}\mid s,a)|+|\tilde{P}(s^{2l}\mid s,a)-P^{\tilde{\boldsymbol{\sigma}}}(s^{2l}\mid s,a)|\Big)\right)\\
&\geq \frac{1}{2^{|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}|\mathcal{S}||\mathcal{A}|}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{l=1}^{\lfloor|\mathcal{S}|/2\rfloor}\sum_{\boldsymbol{\sigma}_{(s,a),l}=1}\int\frac{1}{2}\Big(|\tilde{P}(s^{2l-1}\mid s,a)-P^{\boldsymbol{\sigma}}(s^{2l-1}\mid s,a)|+|\tilde{P}(s^{2l}\mid s,a)\\
&-P^{\boldsymbol{\sigma}}(s^{2l}\mid s,a)|+|\tilde{P}(s^{2l-1}\mid s,a)-P^{\tilde{\boldsymbol{\sigma}}}(s^{2l-1}\mid s,a)|+|\tilde{P}(s^{2l}\mid s,a)-P^{\tilde{\boldsymbol{\sigma}}}(s^{2l}\mid s,a)|\Big)\min\left\{d\mathbb{P}^{\boldsymbol{\sigma}},d\mathbb{P}^{\tilde{\boldsymbol{\sigma}}}\right\}\\
&\geq \frac{1}{2^{|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}|\mathcal{S}||\mathcal{A}|}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{l=1}^{\lfloor|\mathcal{S}|/2\rfloor}\sum_{\boldsymbol{\sigma}_{(s,a),l}=1}\frac{1}{2}\Big(|P^{\boldsymbol{\sigma}}(s^{2l-1}\mid s,a)-P^{\tilde{\boldsymbol{\sigma}}}(s^{2l-1}\mid s,a)|\\
&+|P^{\boldsymbol{\sigma}}(s^{2l}\mid s,a)-P^{\tilde{\boldsymbol{\sigma}}}(s^{2l}\mid s,a)|\Big)\int\min\left\{d\mathbb{P}^{\boldsymbol{\sigma}},d\mathbb{P}^{\tilde{\boldsymbol{\sigma}}}\right\}\\
&=\frac{1}{2^{|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}|\mathcal{S}||\mathcal{A}|}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{l=1}^{\lfloor|\mathcal{S}|/2\rfloor}\sum_{\boldsymbol{\sigma}_{(s,a),l}=1}\frac{1}{2}\left(\frac{6\delta}{\lfloor|\mathcal{S}|/2\rfloor}+\frac{6\delta}{\lfloor|\mathcal{S}|/2\rfloor}\right)\int\min\left\{d\mathbb{P}^{\boldsymbol{\sigma}},d\mathbb{P}^{\tilde{\boldsymbol{\sigma}}}\right\}\\
&=\frac{6\delta}{2^{|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{l=1}^{\lfloor|\mathcal{S}|/2\rfloor}\sum_{\boldsymbol{\sigma}_{(s,a),l}=1}\left(1-\|\mathbb{P}^{\boldsymbol{\sigma}}-\mathbb{P}^{\tilde{\boldsymbol{\sigma}}}\|_{\mathrm{TV}}\right)\\
&=3\delta\left[1-\frac{1}{2^{|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor-1}|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{l=1}^{\lfloor|\mathcal{S}|/2\rfloor}\sum_{\boldsymbol{\sigma}_{(s,a),l}=1}\|\mathbb{P}^{\boldsymbol{\sigma}}-\mathbb{P}^{\tilde{\boldsymbol{\sigma}}}\|_{\mathrm{TV}}\right]\\
&=3\delta\left[1-\frac{1}{2^{|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}\sum_{\boldsymbol{\sigma}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{l=1}^{\lfloor|\mathcal{S}|/2\rfloor}\|\mathbb{P}^{\boldsymbol{\sigma}}-\mathbb{P}^{\tilde{\boldsymbol{\sigma}}}\|_{\mathrm{TV}}\right]\\
&\geq 3\delta\left(1-24\delta\sqrt{\frac{nT}{7|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}}\right),
\end{aligned}
\tag{40}
$$

where the last inequality applies *Eq.(38)*. From Lemma B.3, we have

$$
2\delta\geq 3\delta\left(1-24\delta\sqrt{\frac{nT}{7|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}}\right).
\tag{41}
$$

Rearranging *Eq.(41)* gives

$$
nT\geq\frac{7|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}{5184\delta^2}.
\tag{42}
$$

Part 2: convert the necessary transition observations into bits. By the definition of `Data`, we have

$$
\begin{aligned}
\texttt{Data}&=nT\lceil\log_2|\mathcal{S}|\rceil\\
&\geq\frac{7|\mathcal{S}||\mathcal{A}|\lfloor|\mathcal{S}|/2\rfloor}{5184\delta^2}\lceil\log_2|\mathcal{S}|\rceil\\
&\geq\frac{7|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)}{10368\delta^2}\lceil\log_2|\mathcal{S}|\rceil\\
&\geq\frac{7|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)}{10368\delta^2}\log_2(|\mathcal{S}|-1)
\end{aligned}
$$

$$= \frac{7\underline{\mathbb{H}}_{\frac{1}{4(|\mathcal{S}|-1)}}(\mathcal{F})}{10368\delta^2}, \tag{43}$$

where the second inequality uses

$$\left\lfloor \frac{|\mathcal{S}|}{2} \right\rfloor \geq \frac{|\mathcal{S}|-1}{2}.$$

This completes the proof.

D PROOF OF THEOREM 3

The proof draws the idea from random coding and we divide it into 4 steps. **Step 1:** we construct the reward grid and the finite reference family. **Step 2:** we define the persistent codes and the common policy rule. **Step 3:** we combine the per environment probability bound with the deterministic regret transfer. **Step 4:** we convert these bounds into the *capacity experience tradeoff*.

Notations. We use $\mathcal{H}_t$ to denote the random history and h_t to denote the realized history at time t. Since $\mathcal{S}$ and $\mathcal{A}$ are finite, $\mathcal{H}_t$ is finite. We use $\pi_{\theta,r}(h_K, c)$ to denote the history-dependent policy conditioned on P_θ and input h_K. Moreover, the first K transitions are generated by the generative access and we define

$$n_K := \left\lfloor \frac{K}{|\mathcal{S}||\mathcal{A}|} \right\rfloor \tag{44}$$

as the next state observations at every (s, a) in a fixed order.

Step 1: reward approximation and a finite reference family. Fix $0 < \epsilon < \frac{1}{16}$ and define

$$m := \left\lceil \frac{|\mathcal{S}|}{2\epsilon^2} \right\rceil + |\mathcal{S}|, \qquad \eta := \frac{1}{2m}. \tag{45}$$

We then define the scalar grid and reward family as:

$$\mathcal{E}_\eta := \{0, \frac{1}{2m}, \dots, \frac{2m-1}{2m}, 1\}, \qquad \mathcal{R}_\eta := \{\tilde{r} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathcal{E}_\eta\}. \tag{46}$$

For every $r \in \mathcal{R}$, we select a nearest grid value at each coordinate. This give $\|r - \tilde{r}\|_\infty \le \frac{\eta}{2}$. For any fixed deterministic continuation policy $\pi_{\theta,r}(h_K, c)$ and any transition kernel P_θ, we have

$$\begin{aligned}
&|\mathcal{V}_r^{\pi_{\theta,r}(h_K,c)}(s) - \mathcal{V}_{\tilde{r}}^{\pi_{\theta,r}(h_K,c)}(s)| \\
&= |\mathbb{E}_{\pi_{\theta,r}(h_K,c)}\big(\sum_{t=0}^{\infty} \gamma^t \big(r(s_t, a_t, s_{t+1}) - \tilde{r}(s_t, a_t, s_{t+1})\big) \mid s_0 = s\big)| \\
&\le \mathbb{E}_{\pi_{\theta,r}(h_K,c)}\big(\sum_{t=0}^{\infty} \gamma^t |r(s_t, a_t, s_{t+1}) - \tilde{r}(s_t, a_t, s_{t+1})| \mid s_0 = s\big) \\
&\le \sum_{t=0}^{\infty} \gamma^t \frac{\eta}{2} \\
&= \frac{\eta}{2(1-\gamma)}.
\end{aligned} \tag{47}$$

Using the canonical optimal policy for r and evaluating the optimal policy under $\tilde{r}$ gives

$$\begin{aligned}
&\mathcal{V}_r^*(s) - \mathcal{V}_{\tilde{r}}^*(s) \\
&= \mathcal{V}_r^{\pi^*_{\theta,r}(h_K,c)}(s) - \mathcal{V}_{\tilde{r}}^*(s) \\
&\le \mathcal{V}_r^{\pi^*_{\theta,r}(h_K,c)}(s) - \mathcal{V}_{\tilde{r}}^{\pi^*_{\theta,r}(h_K,c)}(s) \\
&\le \frac{\eta}{2(1-\gamma)}.
\end{aligned} \tag{48}$$

Interchanging r and $\tilde{r}$ therefore yields

$$|\mathcal{V}_r^*(s) - \mathcal{V}_{\tilde{r}}^*(s)| \le \frac{\eta}{2(1-\gamma)}. \tag{49}$$

For the integer m we defined in *Eq.(45)*, we discretize each transition row by allowing every transition probability to take values in $\{0, 1/m, \ldots, 1\}$, subject to the probabilities over $s' \in \mathcal{S}$ summing to 1.

We denote $\mathcal{F}_m$ as the finite family of all tabular transition kernels whose rows satisfy this reference grid. Therefore, each state-action pair (s, a) has

$$N \coloneqq \binom{m + |\mathcal{S}| - 1}{|\mathcal{S}| - 1} \tag{50}$$

possibilities, and hence $|\mathcal{F}_m| = N^{|\mathcal{S}||\mathcal{A}|}$.

Lemma D.1 (Proof in Appendix D.1)**.** *For every $P_\theta \in \mathcal{F}$, there exists a reference transition kernel $P_{\hat{\theta}} \in \mathcal{F}_m$ satisfying*

$$P_{\hat{\theta}}(s' \mid s, a) \geq (1 - 2\epsilon^2) P_\theta(s' \mid s, a) \tag{51}$$

for every s, a, s'.

In *Eq.(51)*, both rows sum to one, so total variation equals the sum of their positive differences. Hence

$$\begin{aligned}
&\|P_\theta(\cdot \mid s, a) - P_{\hat{\theta}}(\cdot \mid s, a)\|_{\mathrm{TV}} \\
&= \sum_{i=1}^{|\mathcal{S}|} \mathbf{1}\{P_\theta(s^i \mid s, a) > P_{\hat{\theta}}(s^i \mid s, a)\}\big(P_\theta(s^i \mid s, a) - P_{\hat{\theta}}(s^i \mid s, a)\big) \\
&\leq \frac{|\mathcal{S}|}{m} \sum_{i=1}^{|\mathcal{S}|} \mathbf{1}\{P_\theta(s^i \mid s, a) > P_{\hat{\theta}}(s^i \mid s, a)\} P_\theta(s^i \mid s, a) \\
&\leq \frac{|\mathcal{S}|}{m} \\
&\leq 2\epsilon^2 \\
&\leq \epsilon.
\end{aligned} \tag{52}$$

Taking the maximum over (s, a), we obtain:

$$\|P_\theta - P_{\hat{\theta}}\|_{\infty,\mathrm{TV}} \leq \epsilon. \tag{53}$$

For any integer $u \geq v \geq 1$, we have

$$\binom{u}{v} \leq \frac{u^v}{v!} \leq \left(\frac{eu}{v}\right)^v.$$

The factorial inequality follows from $\ln(v!) = \sum_{i=1}^{v} \ln i \geq \int_1^v \ln x \, dx = v \ln v - v + 1 \geq v \ln v - v$.

By *Eq.(50)*, we have

$$\begin{aligned}
N &= \binom{m + |\mathcal{S}| - 1}{|\mathcal{S}| - 1} \\
&\leq \left(\frac{e(m + |\mathcal{S}| - 1)}{|\mathcal{S}| - 1}\right)^{|\mathcal{S}|-1} \\
&\leq \left(\frac{e|\mathcal{S}|}{|\mathcal{S}| - 1}\left(\frac{1}{2\epsilon^2} + 2\right)\right)^{|\mathcal{S}|-1} \\
&\leq \left(\frac{5e}{4\epsilon^2}\right)^{|\mathcal{S}|-1}
\end{aligned}$$

$$\leq \left(\frac{4}{\epsilon^2}\right)^{|\mathcal{S}|-1}$$
$$\leq \left(1+\frac{2}{\epsilon}\right)^{2(|\mathcal{S}|-1)}. \tag{54}$$

Here $m \leq \frac{|\mathcal{S}|}{2\epsilon^2} + |\mathcal{S}| + 1$, $\frac{|\mathcal{S}|}{|\mathcal{S}|-1} \leq \frac{5}{4}$, $\frac{1}{2\epsilon^2} + 2 \leq \frac{1}{\epsilon^2}$, and $\frac{5e}{4} < 4$. Consequently,

$$|\mathcal{S}||\mathcal{A}| \log_2 N \leq 2|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)\log_2\left(1+\frac{2}{\epsilon}\right)$$
$$= 2\overline{\mathbb{H}}_\epsilon(\mathcal{F}). \tag{55}$$

Step 2: persistent codes and the common policy family. We let

$$g \coloneqq \min\left\{N, \left\lfloor |\mathcal{C}|^{\frac{1}{|\mathcal{S}||\mathcal{A}|}} \right\rfloor\right\}. \tag{56}$$

For each (s,a), partition the N reference rows into g labeled groups of nearly equal size, each containing at most $\lceil N/g \rceil$ rows. The partitions may be sampled independently and uniformly at random, but are fixed before the true environment and online history are realized; the proof holds for every such balanced assignment.

For a reference kernel $P_{\tilde{\theta}}$, let $c^{s,a}_{P_{\tilde{\theta}}}$ denote its row group label and let $c_{P_{\tilde{\theta}}}$ collect all row labels. Since $g^{|\mathcal{S}||\mathcal{A}|} \leq |\mathcal{C}|$, these tuples can be identified with elements of $\mathcal{C}$. For an arbitrary true kernel, we define

$$c_{P_\theta} \coloneqq c_{P_{\hat{\theta}}}. \tag{57}$$

Thus, the same encoder is defined over the full tabular family.

For each reference kernel $P_{\theta'} \in \mathcal{P}_m$ and each $\tilde{r} \in \mathcal{R}_\eta$, we fix an arbitrary deterministic stationary optimal policy $\pi^*_{\theta',\tilde{r}}$. These choices are made once and kept fixed throughout the construction. We then define the following reference weight:

$$w_{\theta'}(h_K, c) \coloneqq N^{-|\mathcal{S}||\mathcal{A}|}\mathbf{1}\{c_{P_{\theta'}} = c\} \prod_{s\in\mathcal{S}}\prod_{a\in\mathcal{A}}\prod_{j=1}^{n_K} P_{\theta'}(s'_{s,a,j} \mid s,a). \tag{58}$$

These weights define the uniform reference distribution over the retained balanced observations and their codes. Normalizing them gives the corresponding reference posterior.

Given (h_K, c) with positive total reference weight, choose the policy family jointly over $\tilde{r} \in \mathcal{R}_\eta$ to maximize

$$\sum_{P_{\theta'}\in\mathcal{F}_m} w_{\theta'}(h_K, c) \prod_{\tilde{r}\in\mathcal{R}_\eta} \mathbf{1}\{\pi_{\theta,\tilde{r}}(h_K, c) = \pi^*_{\theta',\tilde{r}}\}. \tag{59}$$

The maximization is over tuples of deterministic stationary policies. At zero weight pairs, we use an arbitrary fixed policy family.

For an arbitrary reward r, let $\tilde{r}$ be its grid approximation and define

$$\pi_{\theta,r}(h_K, c) \coloneqq \pi_{\theta,\tilde{r}}(h_K, c). \tag{60}$$

The resulting adaptive rule depends only on the history, retained code, and reward. The subscript θ only labels the environment in which the policy is evaluated.

Lemma D.2 (Proof in Appendix D.2). *Fix a true kernel $P_\theta \in \mathcal{F}$ and a realizable history h_K with $c = c_{P_\theta}$. The policy family selected by Eq.(59) admits a reference witness $P_{\theta''} \in \mathcal{F}_m$ such that*

$$\prod_{j=1}^{n_K} P_{\theta''}(s'_{s,a,j} \mid s, a) \geq \prod_{j=1}^{n_K} P_{\hat{\theta}}(s'_{s,a,j} \mid s, a) \tag{61}$$

for every (s, a).

It is worth noting that, the witness is unique for the selected policy family and the maximizing reference weights themselves need not be distinct.

Step 3: reference error and regret transfer. We now separate the probability of selecting a distant reference witness from the regret incurred on a given history.

Lemma D.3 (Proof in Appendix D.3). *For any transition kernel $P_\theta \in \mathcal{F}$, we use $P_{\hat{\theta}}$ to denote its reference transition kernel. We then have the following inequality holds:*

$$\mathbb{P}_{P_\theta}(\|P_{\hat{\theta}} - P_{\theta''}\|_{\infty,\mathrm{TV}} > 3\epsilon) \leq |\mathcal{S}||\mathcal{A}|\big(\lceil \frac{N}{g} \rceil - 1\big) \exp\left(-\frac{n_K \epsilon^2}{2}\right). \tag{62}$$

We further use the below lemma to compare the value functions of our constructed policy.

Lemma D.4 (Proof in Appendix D.4). *We consider $P_\theta \in \mathcal{F}$ and a realizable pair (h_K, c) and let $P_{\theta''}$ be the reference witness. We upper bound the value loss by:*

$$\begin{aligned} &\max_{s\in\mathcal{S}} \sup_{r\in\mathcal{R}} \left\{ \mathcal{V}_r^*(s) - \mathcal{V}_r^{\pi_{\theta,r}(h_K, c_{P_\theta})}(s) \right\} \\ &\qquad \leq \frac{\eta}{1-\gamma} + \frac{2\|P_\theta - P_{\theta''}\|_{\infty,\mathrm{TV}}}{(1-\gamma)^2}. \end{aligned} \tag{63}$$

Specifically, if $\|P_{\hat{\theta}} - P_{\theta''}\|_{\infty,\mathrm{TV}} \leq 3\epsilon$, then the same left hand side is at most $\frac{\eta}{1-\gamma} + \frac{8\epsilon}{(1-\gamma)^2}$.

Combining Lemma D.3 and Lemma D.4 gives

$$\begin{aligned} &\mathbb{P}_{P_\theta}\left(\max_{s\in\mathcal{S}} \sup_{r\in\mathcal{R}} \left\{ \mathcal{V}_r^*(s) - \mathcal{V}_r^{\pi_{\theta,r}(\mathcal{H}_K, c_{P_\theta})}(s) \right\} > \frac{\eta}{1-\gamma} + \frac{8\epsilon}{(1-\gamma)^2} \right) \\ &\leq |\mathcal{S}||\mathcal{A}|\big(\lceil \frac{N}{g} \rceil - 1\big) \exp\left(-\frac{n_K \epsilon^2}{2}\right). \end{aligned} \tag{64}$$

For fixed K, the observation sequence has finite support. Thus the supremum of regret defines a measurable function of the observed sequence.

Step 4: we establish the capacity–experience tradeoff. We first choose

$$\epsilon = \frac{(1-\gamma)^2 \delta}{16},$$

and retain the values of m and η picked in *Eq.(45)*. Since $m \geq \frac{|\mathcal{S}|}{2\epsilon^2}$, we have

$$\begin{aligned} \eta = \frac{1}{2m} \leq \frac{\epsilon^2}{|\mathcal{S}|} \leq \frac{\epsilon}{|\mathcal{S}|}, \\ \frac{\eta}{1-\gamma} \leq \frac{(1-\gamma)\delta}{16|\mathcal{S}|} \leq \frac{\delta}{4}, \end{aligned}$$

$$\frac{8\epsilon}{(1-\gamma)^2} = \frac{\delta}{2},$$
$$\frac{\eta}{1-\gamma} + \frac{8\epsilon}{(1-\gamma)^2} \leq \frac{3\delta}{4} < \delta.$$

If $g = N$, each code group is a singleton, so no incorrect same code reference row exists. Hence $P_{\theta''} = P_{\hat{\theta}}$ and the failure probability is zero. Otherwise, $g = \lfloor |\mathcal{C}|^{\frac{1}{|\mathcal{S}||\mathcal{A}|}} \rfloor$. Applying *Eq.(50)* gives

$$\begin{aligned}\lceil \frac{N}{g} \rceil - 1 &\leq \frac{N}{g} \\ &\leq 2N|\mathcal{C}|^{-\frac{1}{|\mathcal{S}||\mathcal{A}|}}. \end{aligned} \tag{65}$$

The final bound holds when $g = N$ because its left hand side is zero. The sample allocation obeys

$$n_K \geq \frac{K}{|\mathcal{S}||\mathcal{A}|} - 1,$$

and therefore we have

$$-\frac{n_K\epsilon^2}{2} \leq -\frac{K\epsilon^2}{2|\mathcal{S}||\mathcal{A}|} + \frac{\epsilon^2}{2} \tag{66}$$

after rearranging.

Since $0 < \epsilon < \frac{1}{16}$, we also have $\exp(\frac{\epsilon^2}{2}) < \exp(\frac{1}{512}) < 2$. Consequently,

$$\begin{aligned}\exp\left(-\frac{n_K\epsilon^2}{2}\right) &\leq \exp(\frac{\epsilon^2}{2}) \exp\left(-\frac{K\epsilon^2}{2|\mathcal{S}||\mathcal{A}|}\right) \\ &\leq 2\exp\left(-\frac{K\epsilon^2}{2|\mathcal{S}||\mathcal{A}|}\right). \end{aligned} \tag{67}$$

Combining *Eq.(64)* with these estimates yields

$$\begin{aligned}&\mathbb{P}_{P_\theta}\left(\max_{s\in\mathcal{S}} \sup_{r\in\mathcal{R}} \left\{\mathcal{V}_r^*(s) - \mathcal{V}_r^{\pi_{\theta,r}(\mathcal{H}_K, c_{P_\theta})}(s)\right\} > \delta\right) \\ &\leq 4|\mathcal{S}||\mathcal{A}|N|\mathcal{C}|^{-\frac{1}{|\mathcal{S}||\mathcal{A}|}} \exp\left(-\frac{K(1-\gamma)^4\delta^2}{512|\mathcal{S}||\mathcal{A}|}\right) \end{aligned} \tag{68}$$

for every realized environment. It therefore suffices to impose

$$N|\mathcal{C}|^{-\frac{1}{|\mathcal{S}||\mathcal{A}|}} \exp\left(-\frac{K(1-\gamma)^4\delta^2}{512|\mathcal{S}||\mathcal{A}|}\right) \leq \frac{1-\alpha}{4|\mathcal{S}||\mathcal{A}|}. \tag{69}$$

Taking base two logarithms,

$$\begin{aligned}&\log_2 N - \frac{\log_2|\mathcal{C}|}{|\mathcal{S}||\mathcal{A}|} - \frac{K(1-\gamma)^4\delta^2}{512|\mathcal{S}||\mathcal{A}|\ln 2} \\ &\leq -\log_2\left(\frac{4|\mathcal{S}||\mathcal{A}|}{1-\alpha}\right). \end{aligned} \tag{70}$$

Multiplying by $|\mathcal{S}||\mathcal{A}|$ and using the sufficient condition in Theorem 3, we obtain

$$\begin{aligned}
&\log_2 |\mathcal{C}| + \frac{(1-\gamma)^4\delta^2}{512\ln 2}K \\
&\quad \geq 2\overline{\mathbb{H}}_{\frac{(1-\gamma)^2\delta}{16}}(\mathcal{F}) + |\mathcal{S}||\mathcal{A}|\log_2\left(\frac{4|\mathcal{S}||\mathcal{A}|}{1-\alpha}\right) \\
&\quad \geq |\mathcal{S}||\mathcal{A}|\log_2 N + |\mathcal{S}||\mathcal{A}|\log_2\left(\frac{4|\mathcal{S}||\mathcal{A}|}{1-\alpha}\right). \qquad (71)
\end{aligned}$$

Thus the probability of regret exceeding δ is at most $1-\alpha$ for every realized environment, with the same encoder and common continuation rule.

This concludes the proof.

D.1 Proof of Lemma D.1

Fix $P_\theta \in \mathcal{F}$ and enumerate states as $s^1, \ldots, s^{|\mathcal{S}|}$. At every (s,a),

$$\begin{aligned}
\sum_{i=1}^{|\mathcal{S}|}\left\lceil (m-|\mathcal{S}|)P_\theta(s^i \mid s,a)\right\rceil &\leq \sum_{i=1}^{|\mathcal{S}|}\Big((m-|\mathcal{S}|)P_\theta(s^i \mid s,a)+1\Big) \\
&= (m-|\mathcal{S}|)+|\mathcal{S}| \\
&= m. \qquad (72)
\end{aligned}$$

We choose integers $k_{s,a,i}$ initially equal to these ceilings and add any remaining integer mass to the last coordinate. We then obtain

$$k_{s,a,i} \geq \left\lceil (m-|\mathcal{S}|)P_\theta(s^i \mid s,a)\right\rceil, \qquad \sum_{i=1}^{|\mathcal{S}|} k_{s,a,i} = m. \qquad (73)$$

If the row already belongs to the reference grid, we instead set $k_{s,a,i} = mP_\theta(s^i \mid s,a)$. In this case, $mP_\theta(s^i \mid s,a)$ is an integer, and

$$\left\lceil (m-|\mathcal{S}|)P_\theta(s^i \mid s,a)\right\rceil \leq \left\lceil mP_\theta(s^i \mid s,a)\right\rceil = mP_\theta(s^i \mid s,a),$$

Hence, these integers also satisfy *Eq.(73)*. This fixes a deterministic rule for every (s,a) that is the identity on grid rows. We set $P_{\hat{\theta}}(s^i \mid s,a) \coloneqq \frac{k_{s,a,i}}{m}$, then for every coordinate,

$$\begin{aligned}
P_{\hat{\theta}}(s^i \mid s,a) &= \frac{k_{s,a,i}}{m} \\
&\geq \left(1-\frac{|\mathcal{S}|}{m}\right)P_\theta(s^i \mid s,a) \\
&\geq (1-2\epsilon^2)P_\theta(s^i \mid s,a), \qquad (74)
\end{aligned}$$

which completes the proof.

D.2 Proof of Lemma D.2

We first take two distinct transition kernels $P_{\tilde{\theta}}, P_{\theta'} \in \mathcal{F}_m$ and, without loss of generality, we choose a transition (s,a,s^i) such that

$$P_{\tilde{\theta}}(s^i \mid s,a) - P_{\theta'}(s^i \mid s,a) \geq \frac{1}{m} = 2\eta,$$

$$z \coloneqq \frac{P_{\tilde{\theta}}(s^i \mid s,a) + P_{\theta'}(s^i \mid s,a)}{2} \in \mathcal{E}_\eta.$$

Consequently,

$$P_{\tilde{\theta}}(s^i \mid s,a) - z = \frac{P_{\tilde{\theta}}(s^i \mid s,a) - P_{\theta'}(s^i \mid s,a)}{2} \geq \eta,$$
$$P_{\theta'}(s^i \mid s,a) - z = -\frac{P_{\tilde{\theta}}(s^i \mid s,a) - P_{\theta'}(s^i \mid s,a)}{2} \leq -\eta.$$

Use the reward from Appendix B:

$$r_{s,a,s^i,z}(x,y,x') \coloneqq z + \mathbf{1}\{x = s, y = a\}\big(\mathbf{1}\{x' = s^i\} - z\big). \tag{75}$$

We write $r \coloneqq r_{s,a,s^i,z}$. Then

$$r(x,y,x') \in \{0, z, 1\} \subseteq \mathcal{E}_\eta, \qquad r \in \mathcal{R}_\eta.$$

Since z is the midpoint of the two transition probabilities and their difference is at least 2η, we have each canonical policy is optimal and therefore satisfies bounded regret with tolerance $\frac{\eta}{2}$. Lemma B.1 and Lemma B.2 then imply

$$\pi^*_{\tilde{\theta},r}(s) = a, \qquad \pi^*_{\theta',r}(s) \neq a.$$

Because $r \in \mathcal{R}_\eta$, distinct reference kernels have distinct canonical policy families. Therefore, for every $P_{\tilde{\theta}}, P_{\theta'} \in \mathcal{F}_m$,

$$\prod_{\tilde{r} \in \mathcal{R}_\eta} \mathbf{1}\{\pi^*_{\tilde{\theta},\tilde{r}}(h_K, c) = \pi^*_{\theta',\tilde{r}}(h_K, c)\} = \mathbf{1}\{P_{\tilde{\theta}} = P_{\theta'}\}. \tag{76}$$

Fix h_K with $\mathbb{P}_{P_\theta}(\mathcal{H}_K = h_K) > 0$ and $c = c_{P_\theta} = c_{P_{\hat{\theta}}}$. Lemma D.1 and *Eq.(58)* give

$$\begin{aligned} w_{\hat{\theta}}(h_K, c) &= N^{-|\mathcal{S}||\mathcal{A}|} \prod_{x \in \mathcal{S}} \prod_{y \in \mathcal{A}} \prod_{j=1}^{n_K} P_{\hat{\theta}}(s'_{x,y,j} \mid x, y) \\ &\geq N^{-|\mathcal{S}||\mathcal{A}|}(1 - 2\epsilon^2)^{n_K|\mathcal{S}||\mathcal{A}|} \prod_{x \in \mathcal{S}} \prod_{y \in \mathcal{A}} \prod_{j=1}^{n_K} P_\theta(s'_{x,y,j} \mid x, y) \\ &> 0. \end{aligned} \tag{77}$$

Moreover, $\sum_{P_{\theta'} \in \mathcal{F}_m} w_{\theta'}(h_K, c) \geq w_{\hat{\theta}}(h_K, c) > 0$. For every candidate policy family, we have

$$\sum_{P_{\theta'} \in \mathcal{F}_m} \prod_{\tilde{r} \in \mathcal{R}_\eta} \mathbf{1}\{\pi'_{\theta,\tilde{r}}(h_K, c) = \pi^*_{\theta',\tilde{r}}(h_K, c)\} \leq 1.$$

Thus, for every $P_{\tilde{\theta}} \in \mathcal{F}_m$, applying *Eq.(59)* gives

$$\begin{aligned} w_{\tilde{\theta}}(h_K, c) &= \sum_{P_{\theta'} \in \mathcal{F}_m} w_{\theta'}(h_K, c) \mathbf{1}\{P_{\theta'} = P_{\tilde{\theta}}\} \\ &= \sum_{P_{\theta'} \in \mathcal{F}_m} w_{\theta'}(h_K, c) \prod_{\tilde{r} \in \mathcal{R}_\eta} \mathbf{1}\{\pi^*_{\tilde{\theta},\tilde{r}}(h_K, c) = \pi^*_{\theta',\tilde{r}}(h_K, c)\} \end{aligned}$$

$$\leq \sum_{P_{\theta'} \in \mathcal{F}_m} w_{\theta'}(h_K, c) \prod_{\tilde{r} \in \mathcal{R}_\eta} \mathbf{1}\{\pi_{\theta,\tilde{r}}(h_K, c) = \pi^*_{\theta',\tilde{r}}(h_K, c)\}$$
$$\leq \max_{P_{\theta'} \in \mathcal{F}_m} w_{\theta'}(h_K, c) \sum_{P_{\theta'} \in \mathcal{F}_m} \prod_{\tilde{r} \in \mathcal{R}_\eta} \mathbf{1}\{\pi_{\theta,\tilde{r}}(h_K, c) = \pi^*_{\theta',\tilde{r}}(h_K, c)\}$$
$$\leq \max_{P_{\theta'} \in \mathcal{F}_m} w_{\theta'}(h_K, c). \tag{78}$$

The selected family therefore matches a unique reference kernel $P_{\theta''}$, and

$$\sum_{P_{\theta'} \in \mathcal{F}_m} w_{\theta'}(h_K, c) \prod_{\tilde{r} \in \mathcal{R}_\eta} \mathbf{1}\{\pi_{\theta,\tilde{r}}(h_K, c) = \pi^*_{\theta',\tilde{r}}(h_K, c)\}$$
$$= \max_{P_{\theta'} \in \mathcal{F}_m} w_{\theta'}(h_K, c)$$
$$= w_{\theta''}(h_K, c)$$
$$> 0. \tag{79}$$

Subtracting *Eq.(58)*, we obtain:

$$0 < w_{\theta''}(h_K, c) = N^{-|\mathcal{S}||\mathcal{A}|} \mathbf{1}\{c_{P_{\theta''}} = c\} \prod_{x \in \mathcal{S}} \prod_{y \in \mathcal{A}} \prod_{j=1}^{n_K} P_{\theta''}(s'_{x,y,j} \mid x, y).$$

Consequently, $c_{P_{\theta''}} = c, \pi_{\theta,\tilde{r}}(h_K, c) = \pi^*_{\theta'',\tilde{r}}(h_K, c)$ for all $\tilde{r} \in \mathcal{R}_\eta$ and $\prod_{j=1}^{n_K} P_{\theta''}(s'_{s,a,j} \mid s, a) > 0$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. For a fixed (s, a), the row indexed maximum below range over its N distinct reference rows. For a candidate with $c^{s,a}_{P_{\theta'}} = c^{s,a}$, define

$$P_{\tilde{\theta}}(x' \mid x, y) := \begin{cases} P_{\theta'}(x' \mid s, a), & (x, y) = (s, a), \\ P_{\theta''}(x' \mid x, y), & (x, y) \neq (s, a). \end{cases}$$

The product construction gives

$$P_{\tilde{\theta}} \in \mathcal{F}_m, \qquad c_{P_{\tilde{\theta}}} = c, \qquad w_{\tilde{\theta}}(h_K, c) \leq w_{\theta''}(h_K, c).$$

Expanding the weights, we obtain

$$1 \geq \frac{w_{\tilde{\theta}}(h_K, c)}{w_{\theta''}(h_K, c)}$$
$$= \frac{N^{-|\mathcal{S}||\mathcal{A}|} \prod_{x \in \mathcal{S}} \prod_{y \in \mathcal{A}} \prod_{j=1}^{n_K} P_{\tilde{\theta}}(s'_{x,y,j} \mid x, y)}{N^{-|\mathcal{S}||\mathcal{A}|} \prod_{x \in \mathcal{S}} \prod_{y \in \mathcal{A}} \prod_{j=1}^{n_K} P_{\theta''}(s'_{x,y,j} \mid x, y)}$$
$$= \prod_{x \in \mathcal{S}} \prod_{y \in \mathcal{A}} \frac{\prod_{j=1}^{n_K} P_{\tilde{\theta}}(s'_{x,y,j} \mid x, y)}{\prod_{j=1}^{n_K} P_{\theta''}(s'_{x,y,j} \mid x, y)}$$
$$= \frac{\prod_{j=1}^{n_K} P_{\theta'}(s'_{s,a,j} \mid s, a)}{\prod_{j=1}^{n_K} P_{\theta''}(s'_{s,a,j} \mid s, a)}.$$

Hence, for every (s, a),

$$\prod_{j=1}^{n_K} P_{\theta''}(s'_{s,a,j} \mid s, a)$$

$$= \max_{P_{\theta'}(\cdot|s,a)} \left(\mathbf{1}\{c^{s,a}_{P_{\theta'}} = c^{s,a}\} \prod_{j=1}^{n_K} P_{\theta'}(s'_{s,a,j} \mid s,a) \right)$$
$$\geq \mathbf{1}\{c^{s,a}_{P_{\hat{\theta}}} = c^{s,a}\} \prod_{j=1}^{n_K} P_{\hat{\theta}}(s'_{s,a,j} \mid s,a)$$
$$= \prod_{j=1}^{n_K} P_{\hat{\theta}}(s'_{s,a,j} \mid s,a). \tag{80}$$

This proves the lemma.

D.3 Proof of Lemma D.3

We define the Hel-linger affinity for a fixed (s,a) :

$$\mathrm{Aff}_{s,a}(P_\theta, P_{\theta'}) \coloneqq \sum_{s' \in \mathcal{S}} \sqrt{P_\theta(s' \mid s,a) P_{\theta'}(s' \mid s,a)}. \tag{81}$$

This lies in $[0,1]$ since Cauchy–Schwarz bounds it by the square root of the product in $[0,1]$.

We can show that,

$$\sum_{s' \in \mathcal{S}} \min\{P_\theta(s' \mid s,a), P_{\theta'}(s' \mid s,a)\}$$
$$= \frac{1}{2} \sum_{s' \in \mathcal{S}} \left(P_\theta(s' \mid s,a) + P_{\theta'}(s' \mid s,a) - |P_\theta(s' \mid s,a) - P_{\theta'}(s' \mid s,a)|\right)$$
$$= 1 - \|P_\theta(\cdot \mid s,a) - P_{\theta'}(\cdot \mid s,a)\|_{\mathrm{TV}}, \tag{82}$$

and moreover,

$$\sum_{s' \in \mathcal{S}} \max\{P_\theta(s' \mid s,a), P_{\theta'}(s' \mid s,a)\}$$
$$= 2 - \sum_{s' \in \mathcal{S}} \min\{P_\theta(s' \mid s,a), P_{\theta'}(s' \mid s,a)\}$$
$$= 1 + \|P_\theta(\cdot \mid s,a) - P_{\theta'}(\cdot \mid s,a)\|_{\mathrm{TV}}. \tag{83}$$

For each outcome, the product of the minimum and maximum is the product of the original two probabilities. Applying Cauchy–Schwarz to their square roots gives

$$\mathrm{Aff}_{s,a}(P_\theta, P_{\theta'})^2 \leq \left(\sum_{s' \in \mathcal{S}} \min\{P_\theta(s' \mid s,a), P_{\theta'}(s' \mid s,a)\} \right) \left(\sum_{s' \in \mathcal{S}} \max\{P_\theta(s' \mid s,a), P_{\theta'}(s' \mid s,a)\} \right)$$
$$= 1 - \|P_\theta(\cdot \mid s,a) - P_{\theta'}(\cdot \mid s,a)\|^2_{\mathrm{TV}}. \tag{84}$$

Fix a candidate row $P_{\theta'}(\cdot \mid s,a)$ for which

$$\|P_{\hat{\theta}}(\cdot \mid s,a) - P_{\theta'}(\cdot \mid s,a)\|_{\mathrm{TV}} > 3\epsilon. \tag{85}$$

Lemma D.1 implies

$$
\begin{aligned}
&\|P_\theta(\cdot \mid s,a) - P_{\theta'}(\cdot \mid s,a)\|_{\mathrm{TV}} \\
&\geq \|P_{\hat{\theta}}(\cdot \mid s,a) - P_{\theta'}(\cdot \mid s,a)\|_{\mathrm{TV}} - \|P_\theta(\cdot \mid s,a) - P_{\hat{\theta}}(\cdot \mid s,a)\|_{\mathrm{TV}} \\
&> 3\epsilon - 2\epsilon^2 \\
&> 2\epsilon.
\end{aligned} \tag{86}
$$

By *Eq.(51)*, $P_{\hat{\theta}}$ is positive on the support of P_θ. Hence, for every s' with $P_\theta(s' \mid s,a) > 0$,

$$
\begin{aligned}
&P_\theta(s' \mid s,a) \left(\frac{P_{\theta'}(s' \mid s,a)}{P_{\hat{\theta}}(s' \mid s,a)}\right)^{\frac{1}{2}} \\
&\quad \leq P_\theta(s' \mid s,a) \left(\frac{P_{\theta'}(s' \mid s,a)}{(1-2\epsilon^2)P_\theta(s' \mid s,a)}\right)^{\frac{1}{2}} \\
&\quad = \frac{\sqrt{P_\theta(s' \mid s,a) P_{\theta'}(s' \mid s,a)}}{\sqrt{1-2\epsilon^2}}.
\end{aligned}
$$

The same inequality holds for a zero probability outcome under the stated convention. Summing over outcomes gives

$$
\begin{aligned}
&\sum_{s' \in \mathcal{S}} P_\theta(s' \mid s,a) \left(\frac{P_{\theta'}(s' \mid s,a)}{P_{\hat{\theta}}(s' \mid s,a)}\right)^{\frac{1}{2}} \\
&\leq \frac{1}{\sqrt{1-2\epsilon^2}} \sum_{s' \in \mathcal{S}} \sqrt{P_\theta(s' \mid s,a) P_{\theta'}(s' \mid s,a)} \\
&= \frac{\mathrm{Aff}_{s,a}(P_\theta, P_{\theta'})}{\sqrt{1-2\epsilon^2}} \\
&\leq \left(\frac{1-4\epsilon^2}{1-2\epsilon^2}\right)^{\frac{1}{2}} \\
&= \left(1 - \frac{2\epsilon^2}{1-2\epsilon^2}\right)^{\frac{1}{2}} \\
&\leq \sqrt{1-\epsilon^2} \\
&= \exp\left(\tfrac{1}{2}\ln(1-\epsilon^2)\right) \\
&\leq \exp(-\frac{\epsilon^2}{2}).
\end{aligned} \tag{87}
$$

For the candidate row in *Eq.(85)*, $\mathbf{1}\{x \geq 1\} \leq \sqrt{x}$ for $x \geq 0$ yields

$$
\begin{aligned}
&\mathbb{P}_{P_\theta}\left(\prod_{j=1}^{n_K} P_{\theta'}(s'_{s,a,j} \mid s,a) \geq \prod_{j=1}^{n_K} P_{\hat{\theta}}(s'_{s,a,j} \mid s,a)\right) \\
&\leq \mathbb{E}_{P_\theta}\left(\prod_{j=1}^{n_K} \left(\frac{P_{\theta'}(s'_{s,a,j} \mid s,a)}{P_{\hat{\theta}}(s'_{s,a,j} \mid s,a)}\right)^{\frac{1}{2}}\right) \\
&= \prod_{j=1}^{n_K} \mathbb{E}_{P_\theta}\left(\left(\frac{P_{\theta'}(s'_{s,a,j} \mid s,a)}{P_{\hat{\theta}}(s'_{s,a,j} \mid s,a)}\right)^{\frac{1}{2}}\right) \\
&= \prod_{j=1}^{n_K} \left(\sum_{s' \in \mathcal{S}} P_\theta(s' \mid s,a) \left(\frac{P_{\theta'}(s' \mid s,a)}{P_{\hat{\theta}}(s' \mid s,a)}\right)^{1/2}\right)
\end{aligned}
$$

$$\leq \prod_{j=1}^{n_K} \exp(-\frac{\epsilon^2}{2})$$
$$= \exp(-\frac{n_K \epsilon^2}{2}). \tag{88}$$

The retained reset samples are independent under the fixed true kernel, so the expectation factorizes. We use the convention that an empty product equals one when $n_K = 0$.

Recall that, for some (s, a), we have

$$\|P_{\hat{\theta}}(\cdot \mid s, a) - P_{\theta''}(\cdot \mid s, a)\|_{\mathrm{TV}} > 3\epsilon.$$

By Lemma D.2, $P_{\theta''}(\cdot \mid s, a)$ has the same row code as $P_{\hat{\theta}}(\cdot \mid s, a)$. Hence the selected row is included among the candidate rows counted as:

$$\begin{aligned}
&\mathbf{1}\{\|P_{\hat{\theta}} - P_{\theta''}\|_{\infty,\mathrm{TV}} > 3\epsilon\} \\
&\leq \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{P_{\theta'}(\cdot|s,a)} \mathbf{1}\{c^{s,a}_{P_{\theta'}} = c^{s,a}_{P_{\hat{\theta}}}\} \mathbf{1}\{\|P_{\theta'}(\cdot \mid s, a) - P_{\hat{\theta}}(\cdot \mid s, a)\|_{\mathrm{TV}} > 3\epsilon\} \\
&\mathbf{1}\left\{\prod_{j=1}^{n_K} P_{\theta'}(s'_{s,a,j} \mid s, a) \geq \prod_{j=1}^{n_K} P_{\hat{\theta}}(s'_{s,a,j} \mid s, a)\right\}.
\end{aligned} \tag{89}$$

Taking expectations and applying *Eq.(88)* to candidates whose distance indicator is one gives

$$\begin{aligned}
&\mathbb{P}_{P_\theta}\big(\|P_{\hat{\theta}} - P_{\theta''}\|_{\infty,\mathrm{TV}} > 3\epsilon\big) \\
&\leq \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{P_{\theta'}(\cdot|s,a)} \mathbf{1}\{c^{s,a}_{P_{\theta'}} = c^{s,a}_{P_{\hat{\theta}}}\} \mathbf{1}\{\|P_{\theta'}(\cdot \mid s, a) - P_{\hat{\theta}}(\cdot \mid s, a)\|_{\mathrm{TV}} > 3\epsilon\} \\
&\mathbb{P}_{P_\theta}\left(\prod_{j=1}^{n_K} P_{\theta'}(s'_{s,a,j} \mid s, a) \geq \prod_{j=1}^{n_K} P_{\hat{\theta}}(s'_{s,a,j} \mid s, a)\right) \\
&\leq \exp(-\frac{n_K \epsilon^2}{2}) \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{P_{\theta'}(\cdot|s,a)} \mathbf{1}\{c^{s,a}_{P_{\theta'}} = c^{s,a}_{P_{\hat{\theta}}}\} \mathbf{1}\{\|P_{\theta'}(\cdot \mid s, a) - P_{\hat{\theta}}(\cdot \mid s, a)\|_{\mathrm{TV}} > 3\epsilon\} \\
&\leq \exp(-\frac{n_K \epsilon^2}{2}) \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} (\lceil \frac{N}{g} \rceil - 1) \\
&= |\mathcal{S}||\mathcal{A}|(\lceil \frac{N}{g} \rceil - 1) \exp(-\frac{n_K \epsilon^2}{2}),
\end{aligned} \tag{90}$$

where the last inequality uses the maximal group size and excludes the reference row itself, which proves our claim.

D.4 Proof of Lemma D.4

Fix $P_\theta \in \mathcal{F}$, a realizable (h_K, c), and the witness $P_{\theta''}$ in Lemma D.2. Let a_s be the action of $\pi_{\theta,r}(h_K, c)$ at state s. The difference of value functions under P_θ and $P_{\theta''}$ is:

$$\begin{aligned}
&\mathcal{V}^{\pi_{\theta,r}(h_K,c)}_{\theta,r}(s) - \mathcal{V}^{\pi_{\theta,r}(h_K,c)}_{\theta'',r}(s) \\
&= \sum_{s'\in\mathcal{S}} P_\theta(s' \mid s, a_s)\big(r(s, a_s, s') + \gamma \mathcal{V}^{\pi_{\theta,r}(h_K,c)}_{\theta,r}(s')\big)
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{s' \in \mathcal{S}} P_{\theta''}(s' \mid s, a_s)\big(r(s, a_s, s') + \gamma \mathcal{V}_{\theta'',r}^{\pi_{\theta,r}(h_K,c)}(s')\big) \\
&= \sum_{s' \in \mathcal{S}} \big(P_\theta(s' \mid s, a_s) - P_{\theta''}(s' \mid s, a_s)\big)\big(r(s, a_s, s') + \gamma \mathcal{V}_{\theta'',r}^{\pi_{\theta,r}(h_K,c)}(s')\big) \\
&+ \gamma \sum_{s' \in \mathcal{S}} P_\theta(s' \mid s, a_s)\big(\mathcal{V}_{\theta,r}^{\pi_{\theta,r}(h_K,c)}(s') - \mathcal{V}_{\theta'',r}^{\pi_{\theta,r}(h_K,c)}(s')\big). \qquad (91)
\end{aligned}
$$

Since $r \in [0, 1]$, the value lies in $[0, \frac{1}{1-\gamma}]$, and hence

$$
0 \le r(s, a_s, s') + \gamma \mathcal{V}_{\theta'',r}^{\pi_{\theta,r}(h_K,c)}(s') \le 1 + \frac{\gamma}{1-\gamma} = \frac{1}{1-\gamma}.
$$

The two transition rows sum to one. Their positive and negative signed differences therefore have the same total absolute mass, equal to the row total variation. In particular,

$$
\begin{aligned}
&- \frac{\|P_\theta(\cdot \mid s, a_s) - P_{\theta''}(\cdot \mid s, a_s)\|_{\mathrm{TV}}}{1-\gamma} \\
&= \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} \mathbf{1}\{P_\theta(s' \mid s, a_s) < P_{\theta''}(s' \mid s, a_s)\}\big(P_\theta(s' \mid s, a_s) - P_{\theta''}(s' \mid s, a_s)\big) \\
&\le \sum_{s' \in \mathcal{S}} \big(P_\theta(s' \mid s, a_s) - P_{\theta''}(s' \mid s, a_s)\big)\big(r(s, a_s, s') + \gamma \mathcal{V}_{\theta'',r}^{\pi_{\theta,r}(h_K,c)}(s')\big) \\
&\le \frac{1}{1-\gamma} \sum_{s' \in \mathcal{S}} \mathbf{1}\{P_\theta(s' \mid s, a_s) > P_{\theta''}(s' \mid s, a_s)\}\big(P_\theta(s' \mid s, a_s) - P_{\theta''}(s' \mid s, a_s)\big) \\
&= \frac{\|P_\theta(\cdot \mid s, a_s) - P_{\theta''}(\cdot \mid s, a_s)\|_{\mathrm{TV}}}{1-\gamma}.
\end{aligned}
$$

Thus the absolute value of the first sum in *Eq.(91)* is at most the row total variation divided by $1 - \gamma$. The second sum is at most γ times the supremum norm of the value difference. Taking the maximum over states gives

$$
\begin{aligned}
&\max_{s \in \mathcal{S}} |\mathcal{V}_{\theta,r}^{\pi_{\theta,r}(h_K,c)}(s) - \mathcal{V}_{\theta'',r}^{\pi_{\theta,r}(h_K,c)}(s)| \\
&\le \frac{\|P_\theta - P_{\theta''}\|_{\infty,\mathrm{TV}}}{1-\gamma} + \gamma \max_{s \in \mathcal{S}} |\mathcal{V}_{\theta,r}^{\pi_{\theta,r}(h_K,c)}(s) - \mathcal{V}_{\theta'',r}^{\pi_{\theta,r}(h_K,c)}(s)|.
\end{aligned}
$$

Rearranging gives:

$$
\max_{s \in \mathcal{S}} |\mathcal{V}_{\theta,r}^{\pi_{\theta,r}(h_K,c)}(s) - \mathcal{V}_{\theta'',r}^{\pi_{\theta,r}(h_K,c)}(s)| \le \frac{\|P_\theta - P_{\theta''}\|_{\infty,\mathrm{TV}}}{(1-\gamma)^2}. \qquad (92)
$$

The inequality holds for every deterministic stationary policy. Hence

$$
\begin{aligned}
\mathcal{V}_{\theta,r}^{*}(s) - \mathcal{V}_{\theta'',r}^{*}(s) &\le \mathcal{V}_{\theta,r}^{\pi_{\theta,r}^{*}(h_K,c)}(s) - \mathcal{V}_{\theta'',r}^{\pi_{\theta,r}^{*}(h_K,c)}(s) \\
&\le \frac{\|P_\theta - P_{\theta''}\|_{\infty,\mathrm{TV}}}{(1-\gamma)^2}. \qquad (93)
\end{aligned}
$$

Interchanging the transition kernels gives the same absolute bound for optimal values.

For any $r \in \mathcal{R}$ and its grid approximation $\tilde{r}$, Lemma D.2 makes the selected policy optimal under $P_{\theta''}$ for $\tilde{r}$. Therefore, we have:

$$
\begin{aligned}
&\mathcal{V}^{*}_{\theta,r}(s) - \mathcal{V}^{\pi_{\theta,r}(h_K,c)}_{\theta,r}(s) \\
&= \big(\mathcal{V}^{*}_{\theta,r}(s) - \mathcal{V}^{*}_{\theta,\tilde{r}}(s)\big) + \big(\mathcal{V}^{*}_{\theta,\tilde{r}}(s) - \mathcal{V}^{*}_{\theta'',\tilde{r}}(s)\big) + \big(\mathcal{V}^{*}_{\theta'',\tilde{r}}(s) - \mathcal{V}^{\pi_{\theta,\tilde{r}}(h_K,c)}_{\theta'',\tilde{r}}(s)\big) \\
&+ \big(\mathcal{V}^{\pi_{\theta,\tilde{r}}(h_K,c)}_{\theta'',\tilde{r}}(s) - \mathcal{V}^{\pi_{\theta,\tilde{r}}(h_K,c)}_{\theta,\tilde{r}}(s)\big) + \big(\mathcal{V}^{\pi_{\theta,\tilde{r}}(h_K,c)}_{\theta,\tilde{r}}(s) - \mathcal{V}^{\pi_{\theta,\tilde{r}}(h_K,c)}_{\theta,r}(s)\big) \\
&\leq \frac{\eta}{2(1-\gamma)} + \frac{\|P_\theta - P_{\theta''}\|_{\infty,\mathrm{TV}}}{(1-\gamma)^2} + 0 + \frac{\|P_\theta - P_{\theta''}\|_{\infty,\mathrm{TV}}}{(1-\gamma)^2} + \frac{\eta}{2(1-\gamma)} \\
&= \frac{\eta}{1-\gamma} + \frac{2\|P_\theta - P_{\theta''}\|_{\infty,\mathrm{TV}}}{(1-\gamma)^2}.
\end{aligned} \tag{94}
$$

Taking the supremum over rewards and the maximum over states proves *Eq.(63)*.

On the event $\|P_{\hat{\theta}} - P_{\theta''}\|_{\infty,\mathrm{TV}} \leq 3\epsilon$, we have

$$
\begin{aligned}
\|P_\theta - P_{\theta''}\|_{\infty,\mathrm{TV}} &\leq \|P_\theta - P_{\hat{\theta}}\|_{\infty,\mathrm{TV}} + \|P_{\hat{\theta}} - P_{\theta''}\|_{\infty,\mathrm{TV}} \\
&\leq 2\epsilon^2 + 3\epsilon \\
&\leq 4\epsilon.
\end{aligned} \tag{95}
$$

Substituting *Eq.(95)* into *Eq.(63)* gives

$$
\begin{aligned}
&\max_{s\in\mathcal{S}} \sup_{r\in\mathcal{R}} \Big\{\mathcal{V}^{*}_{r}(s) - \mathcal{V}^{\pi_{\theta,r}(h_K,c)}_{r}(s)\Big\} \\
&\qquad \leq \frac{\eta}{1-\gamma} + \frac{8\epsilon}{(1-\gamma)^2},
\end{aligned} \tag{96}
$$

which concludes the proof.

E DETAILS OF EXPERIMENTS

Here we introduce details of the experiment setup including the environments, agents, and rewards.

Environments. The environment is an MDP with varying numbers of states across experiments, 10 actions, and discount factor $\gamma = 0.99$, and satisfies Assumption 1. We independently sample each transition row $P(\cdot \mid s, a)$ uniformly from the probability simplex.

In the numerical analysis of Theorem 1, we sample 5 independent environments for every state space size $|\mathcal{S}|$, where we use $|\mathcal{S}| \in \{10, 15, 20, 24, 25, 26, \ldots, 40\}$ when we fix the persistent capacity, and $|\mathcal{S}| \in \{6, 7, \ldots, 40\}$ when we fix the desired regret δ. The sampled environments are reused across distinct persistent capacities and different regrets.

In the numerical analysis of Theorem 3, we again use the same sampling procedure and generate 20 independent environments for each $|\mathcal{S}| \in \{20, 21, \ldots, 40\}$. The experiment varying world complexity uses this full range. The experiment varying K fixes $|\mathcal{S}| = 40$ and reuses the corresponding 20 environments across all interaction budgets and regrets.

Agents. We implement a tabular reward-conditioned agent with a separate $\mathcal{Q}$ table for each of the 32 reward functions described below. Given a reward r, the agent selects the greedy deterministic policy π_r using the corresponding table. We train these tables using PQN (Gallici et al., 2025). We vary the persistent capacity $\log_2 |\mathcal{C}|$ by changing the precision used to store the tables.

In the numerical analysis of Theorem 1, the saved tables are frozen after training.

In the numerical analysis of Theorem 3, the adaptive agent updates them using online generative access described in Appendix D and stochastic TD updates. These online values are treated as online memory and are excluded from the persistent capacity.

Rewards. We construct 32 reward functions. For 16 of them, each transition reward $r(s, a, s')$ is sampled independently from $\mathrm{Unif}(-0.5, 0.5)$. The remaining 16 are local rewards: 8 are constructed using $r_{s,a,s',z}$ in Appendix B, and 8 using $r_{s,a,\tilde{\mathcal{S}},z}$ in Appendix B.3. Each reward is shifted by subtracting its minimum value so that its range lies in $[0, 1]$.

Moreover, since this paper focuses on theoretical explanations, we use such a finite reward set to approximate infinite choices of rewards. All the numerical analyses serve to illustrate and support our theoretical results.

F REFERENCE EXAMPLES IN FIGURE 1

Figure 1 uses representative models and environments to motivate the resource question studied in this paper. Table 1 and 2 provide the primary sources for these examples. The marker areas represent the task success rates reported for the corresponding model–simulator benchmarks. Since those benchmarks follow different evaluation protocols, the marker areas should be interpreted as indicative rather than directly comparable across systems.

Table 1: Model references for Figure 1. Below, M, B, and T denote 10^6, 10^9, and 10^{12} parameters, respectively. Reported counts refer to the cited configurations; checkpoint choice and storage precision remain unspecified for the plotted points.

Example	Primary reference	Reported configuration or scope
TD-MPC2	(Hansen et al., 2024)	1M, 5M, 19M, 48M, 317M; variant unspecified.
DIAMOND	(Alonso et al., 2024)	13M Atari agent; plotted setup unspecified.
IRIS	(Micheli et al., 2023)	Atari agent; plotted configuration unspecified.
MGDT	(Lee et al., 2022)	10M, 40M, 200M; variant unspecified.
AMAGO-2	(Grigsby et al., 2024b)	Architecture depends on the task domain.
Dreamer V3	(Hafner et al., 2024)	Model family; plotted variant unspecified.
Gato	(Reed et al., 2022)	79M, 364M, 1.18B; variant unspecified.
$\pi_{0.5}$	(Black et al., 2025)	Component sizes do not specify total stored policy size.
OpenVLA	(Kim et al., 2024)	7B model; evaluation depends on task and tuning.
UI-TARS 1.5	(ByteDance Seed, 2025)	Released 7B checkpoint differs from the research model.
Qwen3-VL	(Bai et al., 2025)	Dense and mixture of experts variants; checkpoint unspecified.
Qwen3.5	(Qwen Team, 2026)	Cited variant has 397B total, 17B active; plotted variant unspecified.
MiniMax M3	(MiniMax, 2026)	Approximately 428B total, 23B active.
Kimi K2.6	(Moonshot AI, 2026)	1T total, 32B active; vision encoder listed separately.
GLM-5.1	(Z.ai, 2026)	Model card identifies the release; counting convention unspecified.
Nemotron 3 Ultra	(NVIDIA, 2026b)	550B total, 55B active; cited checkpoint uses BF16.
Gemma 4	(Google DeepMind, 2026a)	Multiple model variants; checkpoint unspecified.
Llama 3.2	(Meta, 2024a)	Text models of 1B or 3B; vision models of 11B or 90B.
Mistral NeMo	(Mistral AI, 2024)	12B; base and instruction tuned variants.
Llama 3.3	(Meta, 2024b)	70B instruction tuned checkpoint.

Table 2: Environment references for Figure 1. Each storage coordinate requires a fixed implementation scope and a recorded byte count. The sources below identify the environments; the plotted storage scales remain illustrative.

Example	Primary reference	Required storage specification
Habitat HM3D	(Savva et al., 2019; Ramakrishnan et al., 2021)	Simulator release, HM3D asset split, and download format.
OSWorld	(Xie et al., 2024)	VM image, installed applications, task assets, and disk allocation convention.
DMLab	(Beattie et al., 2016)	DeepMind Lab build, level assets, and included dependencies.
LIBERO	(Liu et al., 2023)	Simulator, task suite, assets, and treatment of demonstrations.

Example	Primary reference	Required storage specification
DM Control	(Tassa et al., 2018)	Control Suite and MuJoCo versions, models, and dependencies.
Minecraft	(Mojang Studios, 2026)	Edition, release, mods, assets, and saved world scope.
Crafter	(Hafner, 2022)	Package release, assets, dependencies, and generated world scope.
Atari 2600	(Bellemare et al., 2013)	ALE reference; exact game ROM set, emulator, and package scope.

Measurement clarification. Quantifying policy size requires a specified checkpoint, parameter precision, and accounting for additional persistent memory. For N stored parameters at b bits each and M bits of additional retained information, the total storage budget is $Nb + M$ bits. For a mixture-of-experts model, this accounting uses the stored parameters, including inactive experts. Environment storage likewise requires a versioned package and asset manifest, with an explicit treatment of compression and dependencies, and with each byte counted as 8 bits. These implementation-level storage quantities are distinct from the transition family's metric entropy.

Quantifying success rate requires a specified model–environment pairing, including task set, reward definition, initial state distribution, interaction budget, and evaluation protocol. For each agent, we select a representative environment–benchmark pairing in these technical reports. The 10^{50} reference is derived from the lower bound in **Example 2**. While this lower bound cannot be further reduced by alternative representations without violating the same bounded-regret guarantee, it provides a conservative schematic reference for the scale of real-world complexity.