Mind Dreamer: Untethering Imagination via Active Counterfactual Reasoning on Latent Manifolds

Anonymous Authors[1]

Abstract

Model-Based Reinforcement Learning (MBRL) leverages latent imagination for sample efficiency, yet remains constrained by **Historical Tethering**: imagination is typically initialized from observed states. This creates a learning asymmetry, where the world model's manifold discovery outpaces the policy's sparse-reward optimization. We propose **Mind Dreamer (MD)**, a framework that operationalizes **Active Counterfactual Reasoning (ACR)** to transcend Markovian continuity. MD reformulates discovery as the minimization of a global Relay Manifold Expected Free Energy (R-EFE); by invoking a latent-space *do*-operator, MD utilizes an adversarial generator to synthesize non-continuous **latent jumps** to epistemic blind spots that are physically plausible yet cognitively challenging. To resolve the credit assignment paradox across these spatial ruptures, we derive the **Relay Value Function (RVF)** and **Relay Uncertainty Function (RUF)**. These potentials treat synthesized anchors as latent bridges, propagating pragmatic and epistemic value through a principled Bellman-style formulation. Notably, we prove that the Uncertainty propagation across discontinuities necessitates a quadratic discount γ^2, establishing a formal epistemic horizon. Theoretically, MD acts as an optimal importance sampler that expands the manifold's spectral gap, reducing the hitting time to critical bottleneck states. Empirically, MD achieves a **1.67× average speedup** over DreamerV3 on DeepMind Control Suite, reaching **8.8×** in sparse-reward tasks.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning . Do not distribute.

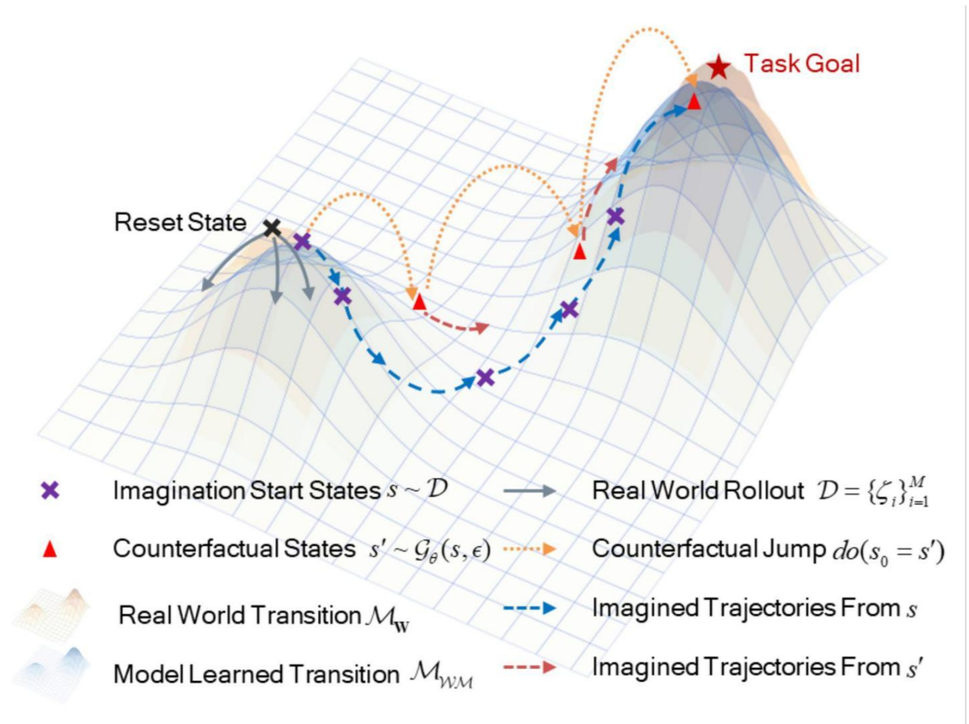

Figure 1. **Untethering Imagination via Active Counterfactual Reasoning**. Unlike standard MBRL (blue) which is **tethered** to historical observations ($s \sim \mathcal{D}$), Mind Dreamer enables proactive manifold intervention beyond historical support. An adversarial generator $\mathcal{G}$ performs non-continuous **latent jumps** via *do*-interventions (orange) to synthesize counterfactual anchors (red). These anchors bridge epistemic blind spots, enabling goal-directed imagination on vast manifold.

1. Introduction

Model-Based Reinforcement Learning (MBRL) has emerged as a cornerstone of sample-efficient AI, predicated on the power of *latent imagination*—the ability to simulate future outcomes within a learned world model (Ha & Schmidhuber, 2018; Sutton, 1991; Hafner et al., 2025). This paradigm is conceptually anchored by the **Manifold Hypothesis**: while sensory observations are high-dimensional, they reside on a low-dimensional manifold $\mathcal{M}$ governed by the environment's underlying physics (Bengio et al., 2013). Under this view, MBRL is a dual-process of geometric discovery: learning the tangent space of $\mathcal{M}$ (dynamics) and regressing a value functional over its topology.

However, a fundamental bottleneck remains: **Historical Tethering**. While modern world models rapidly recover the global structure of $\mathcal{M}$ via dense self-supervised objectives(LeCun, 2022; Bardes et al., 2021), imagination remains a "prisoner of history"—typically initialized only

from observed states in a replay buffer(Ecoffet et al., 2019; Dabney et al., 2021). This creates a critical **Learning Asymmetry**: the agent's internal atlas of the environment's physics (the manifold) expands far more rapidly than its policy can optimize for sparse rewards(Baker et al., 2022; Sekar et al., 2020). Consequently, imagination is confined to previously traversed geodesics, forcing the agent into a costly random walk when encountering "epistemic islands", even if its world model already possesses the structural knowledge to bridge them.

Prior attempts to mitigate exploration bottlenecks have largely relied on intrinsic motivation (Sekar et al., 2020; Houthooft et al., 2016) or goal-relabeling (Andrychowicz et al., 2017). While these methods incentivize the discovery of novel regions, they remain trajectory-bound. In this work, we operationalize this asymmetry via **Active Counterfactual Reasoning (ACR)**. We re-envision the world model as a Generative Oracle capable of targeted manifold intervention rather than a passive simulator By invoking a latent-space *do*-operator (Pearl, 2009), ACR transcends the constraints of Markovian continuity, enabling the agent to perform non-continuous "latent jumps" to synthesized anchors at the frontiers of its knowledge.

We introduce Mind Dreamer (MD), a framework that transforms MBRL from a passive replay mechanism into an information-theoretic stress test. MD utilizes an adversarial state generator $\mathcal{G}$ that lifts the principle of Expected Free Energy (EFE) (Friston, 2009; Tschantz et al., 2020) from a local policy-selection metric to a global discovery functional. This allows the agent to identify "Latent High-Leverage Anchors"—regions where pragmatic goal-alignment and epistemic uncertainty are optimally balanced(Friston et al., 2017b). However, evaluating a non-local jump presents a *credit assignment paradox*: temporal consistency is broken. To resolve this, we derive two recursive functionals: the **Relay Value Function (RVF)** and the **Relay Uncertainty Function (RUF)**. Unlike standard value functions that treat states as destinations, these Relay Potentials treat synthesized anchors as **Latent Bridges**, providing a Bellman-style formulation to evaluate the multi-step utility of a jump across spatial ruptures.

Our theoretical analysis demonstrates that Mind Dreamer acts as an **Optimal Importance Sampler** for manifold refinement. We prove that minimizing the R-EFE objective is mathematically equivalent to minimizing the variance of the world model's gradients, effectively performing a "Manifold Repair" curriculum. Crucially, we show that integrating informational shocks requires a **quadratic discount** γ^2, establishing a formal *Epistemic Horizon* that prevents the generator from chasing distal hallucinations. From a topological perspective, we prove that MD increases the manifold's *conductance*, effectively expanding the manifold's spectral gap, reducing the hitting time to critical states.

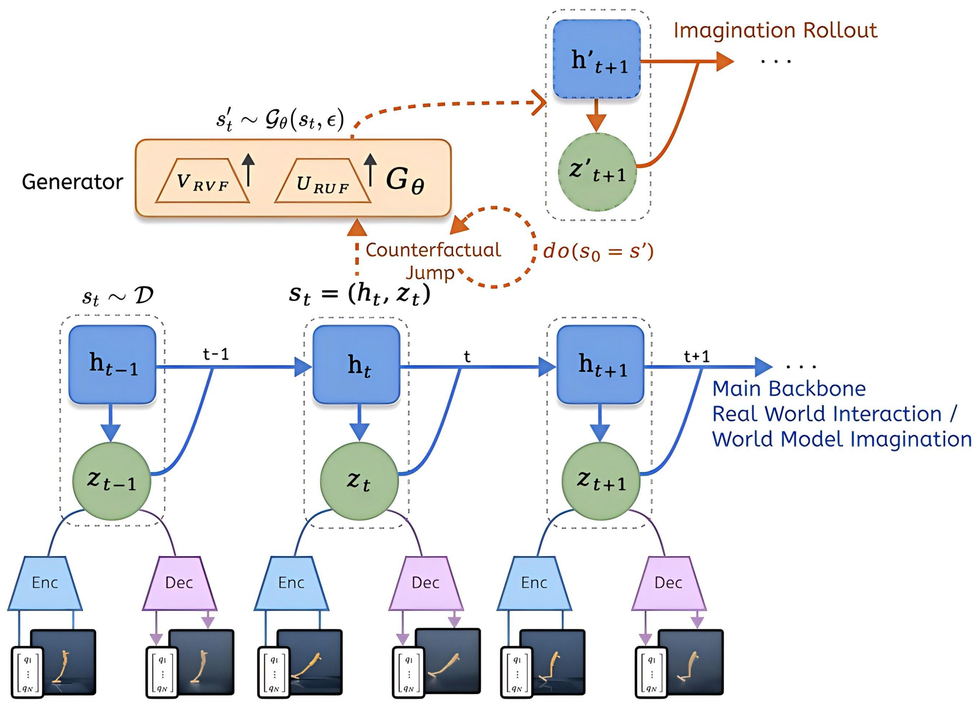

Figure 2. **The Mind Dreamer Paradigm.** Unlike standard MBRL tethered to historical observations, MD enables proactive discovery via: (1) **Adversarial Synthesis**, where a generator $\mathcal{G}_\theta$ uses *do*-interventions to create counterfactual anchors s' at manifold frontiers. (2) **Relay Guidance**, which leverages the Relay Value (V_{RVF}) and Uncertainty (V_{RUF}) functions to assign potential across spatial ruptures. This untethers imagination from past trajectories, targeting physically plausible yet epistemically rich regions for global manifold repair.

Our contributions are summarized as follows:

A New Paradigm: Active Counterfactual Reasoning (ACR). We identify the "Historical Tethering" limitation and introduce the latent *do*-operator to enable goal-directed, non-continuous imagination.

Bridging Discontinuity: Relay Potential Fields. We derive RVF and RUF as recursive functionals as credit assignment to bridge spatial ruptures, introducing the γ^2 discount for stable uncertainty propagation across "latent jumps".

Theoretical Optimality and Speedup. We prove that MD performs optimal importance sampling on manifold residuals, yielding a convergence speedup proportional to the expansion of the manifold's spectral gap.

Empirical Validation. On the DeepMind Control Suite, Mind Dreamer achieves a **1.67× average speedup** in sample efficiency over DreamerV3, with acceleration reaching **up to 8.8×** in challenging bottleneck tasks.

2. Related Works

The evolution of Model-Based RL has transitioned from simple Dyna-style planning (Sutton, 1991) to complex latent imagination (Ha & Schmidhuber, 2018; Hafner et al., 2023). However, the reliance on historical trajectory support

remains a bottleneck. Mind Dreamer sits at the intersection of Active Inference, Counterfactual Reasoning, and Manifold Topology.

2.1. From Passive Curiosity to Active Manifold Repair

Traditional exploration, such as prediction error (Pathak et al., 2017) or information gain (Houthooft et al., 2016) often relies on intrinsic bonuses to steer the agent's random walk. While recent active inference works (Tschantz et al., 2020; Mazzaglia et al., 2021; Millidge, 2020; Friston et al., 2017b; Çatal et al., 2020) treat control as inference, unified exploration and exploitation under Free Energy Principle (Friston, 2009), they remain *trajectory-bound*: Expected Free Energy (EFE) is typically evaluated along continuous Markovian paths. Plan2Explore (Sekar et al., 2020) optimizes for global curiosity but is still constrained by initializing rollouts from the replay buffer. MD transcends this by lifting EFE into a global curriculum for **Manifold Repair**, using an adversarial generator to synthesize counterfactual anchors in regions where the world model's geometry is most "brittle".

2.2. Beyond Trajectory Exploitation: Latent *do*-Intervention

Standard techniques like Hindsight Experience Replay (HER) (Andrychowicz et al., 2017) and Hierarchical RL (Nachum et al., 2018; Li et al., 2020) excel at trajectory exploitation—maximizing the utility of *collected* data through relabeling or abstraction. However, they remain "prisoners of the buffer", unable to reason beyond the convex hull of past experiences. Our approach shifts the paradigm toward **Manifold Intervention**. By invoking a latent-space *do*-operator (Pearl, 2009), MD decouples environmental physics from historical policy bias. Unlike Go-Explore (Ecoffet et al., 2021), which requires explicit environment resets to known states, MD performs mental teleportation to synthesized frontiers, performing intervention-based exploration (Parr & Friston, 2018) in unvisited latent biomes.

2.3. Credit Assignment Across Spatial Ruptures

Breaking Markovian continuity presents a significant credit assignment paradox. Standard temporal difference methods like TD(λ) (Sutton, 1988) or Eligibility Traces (Sutton & Barto, 2018; Seijen & Sutton, 2014) fail when trajectories are non-continuous. While Successor Representations (SR) (Dayan, 1993; Machado et al., 2018) decouple dynamics from rewards, they typically require continuous discovery to build the occupancy map. MD extends this logic via **Relay Potential Fields**. Unlike Goal-Conditioned RL (Schaul et al., 2015) where targets are terminal sinks, our RVF and RUF reformulate synthesized states as **Latent Bridges**. This allows the agent to propagate pragmatic and epistemic value across spatial ruptures identifying high-leverage via-points that connect disjoint regions of the manifold.

3. Methodology: Mind Dreamer

The core philosophy of MD is that efficient learning requires the agent to transcend its historical trajectories and proactively explore unmastered regions of the latent manifold $\mathcal{M}$ through structured "latent jumps."

3.1. Problem Formulation: MBRL on Latent Manifolds

We reformulate the standard POMDP through the lens of the Manifold Hypothesis(Cayton et al., 2005), by assuming that observations $o \in \mathcal{X}$ are realizations of an underlying low-dimensional latent manifold $\mathcal{M} \subset \mathbb{R}^d$, where $d \ll \dim(\mathcal{X})$, and an encoder $e_\psi : \mathcal{X} \to \mathcal{M}$ maps observations to latent states $s \in \mathcal{M}$. This perspective allows us to decompose MBRL into two distinct learning processes:

Learning the Tangent Bundle (World Model): The world model $p_\psi(s_{t+1}|s_t, a_t)$ approximates the transition geometry. By minimizing reconstruction and transition errors, the model learns the local vector field of the manifold.

Learning the Scalar Field (Value Function): The optimal value function $V^*_\phi(s)$ represents a scalar field over $\mathcal{M}$ that quantifies task-optimal potential.

3.2. Active Counterfactual Reasoning

To transcend the "Historical Tethering" of standard MBRL, we propose ACR. This mechanism shifts the world model's role from a passive simulator of past experiences to an active Generative Oracle for directed manifold discovery. The core of ACR is the operationalization of the Learning Asymmetry: the world model captures the transition geometry of the latent manifold $\mathcal{M}$ via dense self-supervised signals far more rapidly than the policy resolves sparse rewards. We exploit this by introducing the latent-space *do*-operator (Pearl, 2009).

Definition 3.1. Counterfactual Jump. An intervention $do(s_0 = s')$ allows the agent to initialize imagination at any anchor $s' \in \mathcal{M}$, provided s' satisfies the manifold consistency $\mathcal{L}_{mf}$ (see Sec. 3.5). This decouples discovery from historical buffer $\mathcal{D}$ while remaining anchored to the Markovian flow $s_t \sim p_\psi(\cdot|s_{t-1}, a_{t-1})$ from learned physics of the tangent bundle.

Unlike standard rollouts initialized from the replay buffer $s_0 \sim \mathcal{D}$, ACR allows the agent to perform an intervention $do(s_0 = s')$, where $s' \in \mathcal{M}$ is a counterfactual anchor synthesized by an adversarial generator $\mathcal{G}_\theta(s, \epsilon)$. This allows the imagination to "teleport" across spatial ruptures, investigating regions that are physically consistent with the environment's laws but have not been mastered by the current

policy. The generator $\mathcal{G}$ acts as an adversarial stress-tester that identifies regions where the latent manifold's tangent bundle $T\mathcal{M}$ is under-sampled or where the scalar field V_ϕ exhibits high curvature. $\mathcal{G}_\theta(s, \epsilon)$ operates exclusively during the imagination phase, ensuring that no extra interaction costs are imposed on the environment.

3.3. From Local EFE to Global Relay Discovery

To guide the state generator $\mathcal{G}$, we leverage the framework of Active Inference, replacing simple reward maximization with the minimization of Expected Free Energy (EFE) (Friston et al., 2017b). While standard MBRL is reactive, EFE provides a proactive objective that unifies goal-directed behavior (pragmatic value) with curiosity-driven exploration (epistemic value). In App. A.2, we demonstrate that MaxEnt RL is a special case of EFE (Millidge et al., 2020).

Definition 3.2. The Local Objective. For a synthesized anchor s', the local EFE $G(s')$ quantifies its immediate utility (inference details in Appendix A.1):

$$G(\pi, \tau) = -\beta \underbrace{\mathcal{I}(s_\tau; o_\tau|\pi)}_{\text{Epistemic Value } \mathcal{E}(s')} - \eta \underbrace{\mathbb{E}_{q(o_\tau)}[\ln p(o_\tau)]}_{\text{Pragmatic Value} \mathcal{P}(s')} \quad (1)$$

where $\mathcal{P}(s')$ aligns with the agent's goal-prior $p(o|C)$ (Levine, 2018) and $\mathcal{E}(s')$ measures the Mutual Information $\mathcal{I}$, identifying regions where the world model's tangent space is ill-defined (Houthooft et al., 2016).

However, Point-wise EFE fails for non-local jumps because it ignores the **reachability** and **distal potential** beyond the jump. We therefore lift $G(s')$ into the Relay EFE (R-EFE), denoted as $\Psi(s, s')$.

$$\Psi(s, s') = \mathbb{E}_q \left[\sum_{k=1}^{H} \gamma^k G(s_k) \middle| s_0 = s, s' \in \tau \right] \quad (2)$$

R-EFE represents a path-integral of free energy, evaluating not just the immediate utility of an anchor s', but the cumulative potential of all trajectories branching from s'.

3.4. Bridging the Gap: Relay Potential Fields

Evaluating the R-EFE objective across spatial ruptures presents a significant credit assignment challenge: as standard value functions depend on local temporal consistency ($s_t \to s_{t+1}$), fail to propagate gradients ($s_t \to s'_{t+1}$). We decompose the Eq. 2 into two recursive **Relay Potential Fields**. These fields act as tractable proxies, allowing the generator $\mathcal{G}$ to perform functional gradient ascent on the manifold's utility landscape.

Pragmatic Proxy (RVF): Unlike goal-conditioned RL typically treats s' as a terminal destination (Andrychowicz et al., 2017; Schaul et al., 2015), Relay Value Function V_{RVF} treats the anchor s' as a **Latent Bridge** rather than a terminal destination. It integrates the cost of reaching s' with the value function $V_\phi(s')$ available thereafter (Formal definition in Appendix C.2):

$$V_{RVF}(s, s') \triangleq \max_\pi \mathbb{E}_{\substack{\pi \\ s_k = s'}} \left[\sum_{t=0}^{k-1} \gamma^t r_t + \gamma^k V_\phi(s') \right] \quad (3)$$

This ensures $\mathcal{G}$ identifies "Pragmatic Gateways"—states that are not necessarily high-reward themselves but are strategically positioned to unlock high-density reward regions.

Epistemic Proxy (RUF): The Relay Uncertainty Function V_{RUF} quantifies the **informational density** of a jump. Crucially, it aggregates expected information gain $\mathcal{I}$ using a **quadratic discount** γ^2 (Derivation in Appendix B.3):

$$V_{RUF}(s, s') \triangleq \mathbb{E}_{\substack{\pi \\ s_k = s'}} \left[\sum_{t=0}^{k-1} \gamma^{2t} \mathcal{I}_{t+1} + \gamma^{2k} U_{\phi_u}(s') \right] \quad (4)$$

Where $U_{\phi_u}(s')$ is the Bellman Uncertainty Function (O'Donoghue et al., 2018a). As detailed in 4.4, the γ^2 term is not a hyperparameter but a requisite for stable variance propagation, ensuring the generator targets regions of genuine epistemic volatility while remaining robust to distal hallucinations where model variance compounds uncontrollably.

3.5. The Adversarial Curriculum: Structural Constraints and Co-evolution

By redefining EFE minimization through RVF and RUF maximization, we transform the active inference problem into a standard RL optimization task for the generator $\mathcal{G}_\theta(s, \epsilon)$:

$$\max_\theta \mathbb{E}_{\substack{s \sim \mathcal{D} \\ s' \sim \mathcal{G}_\theta(s,\epsilon)}} \left[\eta V_{RVF}(s, s') + \beta V_{RUF}(s, s') - \lambda \mathcal{L}_{mf}(s') \right] \quad (5)$$

This creates a "Hacker-Security" dynamic: $\mathcal{G}$ acts as a "hacker" proactively identifying the weakest links (high RUF) or shortest paths (high RVF) in the agent's current atlas. To prevent $\mathcal{G}$ from generating latent hallucinations, we constrain it to the manifold learned by the world model $p_\psi(s_{t+1}|s_t, a_t)$ through structural self-consistency. The manifold constraint $\mathcal{L}_{mf}$ includes:

Dynamics Coherence: Transition entropy $\mathcal{H}(p_\psi(\cdot|s', a))$ is penalized to prevent $\mathcal{G}$ from exploiting ill-defined "cracks" in the world model's dynamics. This effectively defines a "Pessimistic Trust Region" where the model's predictions remain valid (Kidambi et al., 2020).

Cycle-Consistency: Enforces $D_{KL}[\text{Enc}(\text{Dec}(s'))\|s']$ to anchor jumps within the verifiable reconstruction manifold (Zhu et al., 2017), ensuring that synthesized states remain

Algorithm 1 Mind Dreamer: Active Counterfactual Reasoning

Initialize: World Model $\mathcal{WM}$ (Encoder e_ψ, RSSM p_ψ), Value V_ϕ, Uncertainty U_{ϕ_u}, Policy π_ω, Generator $\mathcal{G}_\theta$, Relay Function V_{RVF} and V_{RUF}, Buffer $\mathcal{D}$
for each training step **do**
  1. **World Model Learning**: Update $\{e, p\}$ using $\mathcal{D}$ (Standard RSSM loss)
  2. Mind Dreamer (Latent A* Search)
  Sample $s_0 \sim \mathcal{D}$
  $s' \leftarrow \mathcal{G}_\theta(s_0, \epsilon)$ where $\epsilon \sim \mathcal{N}(0, I)$ {Recursive Latent Composition}
  $\Psi \leftarrow -\eta V_{RVF}(s', s_0) - \beta V_{RUF}(s', s_0)$ {Pragmatic + Epistemic}
  $\mathcal{L}_{mf} \leftarrow \mathcal{H}(p_\psi(\cdot|s', a)) + D_{KL}[\text{Enc}(\text{Dec}(s'))\|s']$
  Update $\theta \leftarrow \nabla_\theta (\Psi + \lambda \mathcal{L}_{mf})$ {Minimize Ψ under manifold constraints}
  3. **Policy Optimization**:
    Start imagination rollouts from generated $\{z_{s'}\}$
    Update π, V_ϕ and U_{ϕ_u} via TD-learning
    Update V_{RVF} and V_{RUF} via k-horizon HER with non-recursive bootstrap target (Eq. 3 and Eq. 4)
  4. Environment Interaction
  Execute π_ω in real environment, collect data to $\mathcal{D}$
end for

representable by the World Model. This constraint ensures that $\mathcal{L}_{mf}$ acts as a trust-region boundary for the adversarial generator (Schulman et al., 2015).

Generator-Policy Co-evolution. As $\mathcal{G}$ shifts the imagination distribution, the policy π must remain stable. We employ quasi-static Target Potential Networks for V_{RVF} and V_{RUF} to ensure non-recursive bootstrap stability (Mnih et al., 2015), preventing the divergence typically seen in off-policy estimation on non-linear manifolds. Furthermore, the World Model p_ψ is updated at a lower frequency relative to the generator $\mathcal{G}$, providing a more stationary transition landscape for state synthesis (Heusel et al., 2017).

4. Theoretical Analysis: From Local Actions to Global Manifold Interventions

We formally ground Mind Dreamer (MD) by demonstrating that RVF and RUF are not merely heuristics but represent the decomposition of Expected Free Energy (EFE) into recursive functionals, and that "jumping" to high-EFE states is a kind of optimal sampling on the global manifold that accelerates manifold refinement and optimal value function convergence.

4.1. Operator Foundations: Mapping EFE to Relay Potentials

We first establish that RVF and RUF are not merely heuristics but represent the decomposition of Expected Free Energy (EFE) into recursive functionals.

Definition 4.1. Relay Potential Operators. Let s be an anchor and s' a counterfactual jump. Given $\tau_{s'} = \inf\{t \geq 0 : s_t = s'\}$ as the first hitting time of s', we define the Pragmatic Relay Operator $\mathcal{T}_V$ and Epistemic Relay Operator $\mathcal{T}_U$ as:

$$(\mathcal{T}_V V)(s, s') = \mathbb{E}_\pi[\sum_{t=0}^{\tau_{s'}-1} \gamma^t r_t + \gamma^{\tau_{s'}} V_\phi(s')] \tag{6}$$

$$(\mathcal{T}_U U)(s, s') = \mathbb{E}_\pi[\sum_{t=0}^{\tau_{s'}-1} \gamma^{2t} \mathcal{I}_{t+1} + \gamma^{2\tau_{s'}} U_{\phi_u}(s')] \tag{7}$$

Theorem 4.2. ***Path-Integral Equivalence***. *The fixed points V^*_{RVF} and V^*_{RUF} uniquely recover the k-step path-integrated constituents of the Expected Free Energy (EFE)(Kappen, 2005).*

Proof Sketch: Under an instrumental prior $p(o|C) \propto \exp(V^*(s)/\lambda)$, maximizing V_{RVF} minimizes cumulative pragmatic risk (see Appendix C.1), while under Gaussian posterior approximations (Millidge et al., 2021) V_{RUF} providing a recursive measure of epistemic leverage across manifold ruptures, maximizing V_{RUF} is equivalent to maximizing the expected reduction in model surprisal (Friston et al., 2015; Houthooft et al., 2016), where informational shocks $\mathcal{I}$ propagate with a quadratic discount γ^2.

Lemma 4.3. ***Contraction and Uniqueness***. *On a connected latent manifold $\mathcal{M}$, for any $s' \neq s$, the operator $\mathcal{T}_V$ and $\mathcal{T}_U$ is a γ and γ^2 contraction mapping under the L_∞ norm.*

Proof Sketch: Since $\tau_{s'} \geq 1$ for any non-trivial jump, the discount factor applied to the bootstrap term $V_\phi(s')$ satisfies $\gamma^{\tau_{s'}} \leq \gamma < 1$. This ensures that $\mathcal{T}_V$ obeys the Banach Fixed-Point Theorem, admitting a unique fixed point V^*_{RVF} that represents the maximum potential of all policy flows passing through the gateway s'. The same applies to $\mathcal{T}_U$, guaranteeing a stable global landscape for the state generator. Detailed proof in Appendix C.4

Remark 4.4. **The Epistemic Horizon**: The γ^2 discount in RUF (Eq. 4) is not an empirical tuning parameter but a necessity of variance propagation. While pragmatic rewards propagate via γ, epistemic informational shocks decay quadratically (O'Donoghue et al., 2018b) (Derivation in Appendix B.3). This quadratic decay ensures that the generator ignores "distal hallucinations" where compounding model variance renders information unverifiable.

4.2. Optimal Imagination: The Dual Accelerators

We formalize the Generator $\mathcal{G}$ as an Active Importance Sampler that minimizes the variance of stochastic gradients for both the world model and the value function.

Theorem 4.5. ***Optimal Proposal and Variance Reduction***. *Let $\mathcal{J}(\theta, \phi)$ be the joint manifold error. The variance of the gradient estimators $\hat{\mathbf{g}}_\theta$ and $\hat{\mathbf{g}}_\phi$ is minimized if $\mathcal{G}$ samples from a distribution q^* such that $q^*(s) \propto \rho(s)\|\nabla G(s)\|_2$. In the MD framework, the R-EFE potential Ψ acts as a path-integrated proxy for this gradient magnitude (Alain et al., 2016). (Proof in Appendix B.8).*

Proof Sketch: Following (Kahn & Marshall, 1953), importance sampling variance is minimized when the proposal density is proportional to the integrand's magnitude. By maximizing Ψ, $\mathcal{G}$ aligns $q_\mathcal{G}$ with q^*, concentrating imagination on "high-leverage" regions where gradients are steepest.

Accelerator I: RUF for Manifold Repair: While standard MBRL passively updates the world model via observed transitions, MD actively repairs the latent manifold by identifying structural gaps.

Proposition 4.6. ***RUF as a Fisher Information Catalyst***. *Under a Gaussian world model p_ψ, the RUF potential V_{RUF} is a first-order surrogate for the trace of the Fisher Information Matrix (FIM): $\mathbb{E}[\mathcal{I}] \approx \frac{1}{2}Tr(\mathcal{F}(\theta)\Sigma_\theta)$. (Derivation in Appendix B.5).*

Proof Sketch: By the Bernstein-von Mises Theorem, epistemic gain $\mathcal{I}$ aligns with the parameter-space curvature. Maximizing RUF forces $\mathcal{G}$ to synthesize anchors s' at high-curvature regions where observations provide maximal parameter refinement, effectively performing active manifold repair (Amari, 1998).

Accelerator II: RVF for Value Convergence: Simultaneously, MD accelerates the value field optimization by transforming value propagation from local diffusion into global targeted updates.

Theorem 4.7. ***Pragmatic Gradient Maximization***. *Let $\Delta V(s) = \sup_{s'} V_{RVF}(s, s') - V_\phi(s)$ be the Relay Advantage. The convergence $V_\phi \to V^*$ is maximized when $\mathcal{G}$ identifies anchors s' that maximize the Variational Bellman Residual $\Delta V(s)$ (See Appendix C.1).*

Proof Sketch: $\Delta V(s)$ represents the latent gap between the current belief V_ϕ and the optimal k-step potential. Maximizing this gap is equivalent to Prioritized Sweeping (Moore & Atkeson, 1993) on a global scale. By "teleporting" to Residual Peaks—states where the value mismatch is greatest—MD collapses value error across the manifold at an accelerated rate, bypassing the limitations of trajectory-bound sampling.

Remark 4.8. **The Symbiotic Convergence**. The generator $\mathcal{G}$ acts as a dual-objective optimizer: by maximizing RUF, it performs Manifold Repair (reducing $\mathcal{R}_\mathcal{M}$), and by maximizing RVF, it performs Policy Alignment (reducing ΔV). This ensures that the imagination process is not just 'dreaming' of novel states, but specifically targeting states that collapse the joint error of the world model and the value field.

4.3. Topological Acceleration: Breaking the Conductance Bottleneck

To quantify the efficiency gain of MD, we compare its convergence rate against trajectory-constrained agents (e.g., DreamerV3). Trajectory-based exploration is fundamentally bounded by the conductance Φ of the latent manifold $\mathcal{M}$. In environments with "epistemic islands" or bottleneck states, $\Phi \to 0$, leading to an exponential hitting time $\tau \approx \Phi^{-2}$ for distal discovery (Jerrum & Sinclair, 1989).

Theorem 4.9. ***Conductance Expansion and Speedup***. *The $do(s')$ intervention in MD induces a synthetic expansion of the manifold's spectral gap. The acceleration ratio η relative to trajectory-sampling scales with the χ^2-divergence between the optimal EFE proposal q^* and the transition-constrained distribution q_{traj}:*

$$\eta = \frac{\tau_{traj}}{\tau_{MD}} \approx 1 + \chi^2(q^* \| q_{traj}) \propto \Phi^{-2}$$

Specifically, MD reduces the hitting time to the discovery frontier from $\mathcal{O}(poly(\Phi^{-1}))$ to $\mathcal{O}(\log |\mathcal{M}|)$ by establishing non-local latent bridges. (Full derivation in Appendix D.4 and D.5).

Remark 4.10. **The Epistemic Island Paradox**. Standard MBRL suffers from an "exponential wall" when $q_{traj} \to 0$ in sparse-reward regimes (Osband et al., 2016). By "teleporting" the imagination directly to high-EFE regions, MD collapses the χ^2 divergence, transforming exploration from a random walk into targeted manifold intervention (See Case Study in Appendix D.4).

4.4. Manifold Anchoring: Stability and Safety Jumps

A primary challenge in adversarial MBRL is preventing the generator from exploiting "manifold cracks"—high-frequency artifacts where the world model is inaccurate. We establish the safety margin for such non-local jumps.

Theorem 4.11. ***Hallucination Error Bound***. *Let $\delta = \|s' - Proj_\mathcal{M}(s')\|$ represent the manifold deviation of a jump, regularized by $\mathcal{L}_{mf}$. Under the assumption of L-Lipschitz continuity of the value field, the value estimation error ϵ_V is bounded by:*

$$\epsilon_V = \|V_{RVF}(s') - V^*(s')\| \leq \frac{L \cdot \delta}{1 - \gamma^n}$$

where the quadratic discount γ^2 in RUF (Eq. 4) ensures that distal epistemic errors decay faster than pragmatic rewards,

shielding the policy from recursive hallucination (Proof in Appendix D.9).

Proof Sketch: The estimation error δ propagates through the Bellman recursion. By minimizing $\mathcal{L}_{cycle}$ and $\mathcal{L}_{dyn}$, MD functions as an implicit spectral filter on the latent Jacobian $\nabla_s p_\psi$, effectively anchoring adversarial jumps to the physically verifiable support of $\mathcal{M}$. The use of γ^2 for uncertainty propagation ensures a finite Epistemic Horizon, preventing the generator from chasing compounding model variance (Asadi et al., 2018).

Corollary 4.12. ***Gradient Stability***. *Under the Manifold Anchoring bound, the variance of the policy gradient $\nabla_\omega J(\pi_\omega)$ remains $L_{\nabla V} \cdot \delta$ stable, precluding the catastrophic collapses typical of unconstrained adversarial dreaming (Details in Appendix D.11).*

5. Experiments

Our experiments aim to answer four key questions: (1) Does MindDreamer (MD) improve sample efficiency and final performance by untethering imagination? (2) Can MD identify and fill "epistemic blind spots" on the latent manifold? (3) Does our manifold safeguarding effectively bound the risk of latent hallucinations?

5.1. Experimental Setup and Baselines

We emphasize that Mind Dreamer (MD) is a modular framework compatible with various MBRL architectures. In this study, we implement MD upon the state-of-the-art **DreamerV3** (Hafner et al., 2023) to demonstrate its potential, use identical RSSM backbones and interaction budgets to ensure fairness. We evaluate against: (1) **DreamerV3** as the primary baseline for trajectory-tethered imagination; (2) **DreamerV2** (Hafner et al., 2020b) to isolate the gains from backbone advancements versus our ACR mechanism; and (3) **Plan2Explore** (Sekar et al., 2020) for curiosity-driven exploration comparison.

5.2. Visualizing ACR: The Three-Ring Manifold Case Study

To demystify the "Latent Jump" mechanism, we design a synthetic **Three-Ring Manifold** environment where the underlying geometry $\mathcal{M}$ is known but topologically challenging. The agent resides on three coplanar coaxial rings, with the top view of the field as observation. Each ring acts as a Local Attractor (simulating a local optimum), to transition to the higher reward ring, the agent must move against the gravitational pull toward the manifold boundary—a feat nearly impossible for random walk exploration.

As shown in Fig. 3, standard DreamerV3 trajectories remain trapped within the first ring, as its imagination is initialized

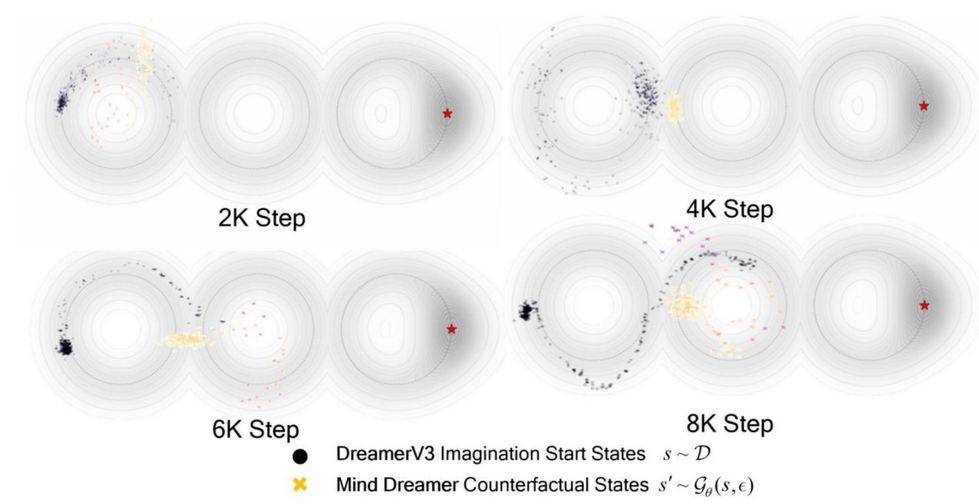

Figure 3. **Visualization of Sampling Dynamics on the Synthetic Three-Ring Manifold.** We compare the distribution of imagined states between DreamerV3 and Mind Dreamer across training snapshots. DreamerV3 exhibits Historical Tethering, its sampling distribution is strictly coupled with historical occupancy, hard to escape the ring attractor. Mind Dreamer demonstrates Active Manifold Refinement, the sampling distribution rapidly aggregates at the topological junctions between rings. This illustrates MD's ability to proactively bridge manifold bottlenecks, enabling the agent to transcend the limits of the replay buffer and linearize the discovery of global task potential.

from the replay buffer which is dominated by high-density attractor data. In contrast, MD's generator $\mathcal{G}$ proactively identifies Latent Gateways at the rings' boundaries where V_{RUF} (uncertainty) is high. By invoking $do(s')$, MD teleports imagination to the boundary ruptures, allowing the policy to discover the transition to the next ring without requiring thousands of random physical steps, thus achieves nearly a **4.2**$\times$ acceleration for first hitting time over DreamerV3. This confirms that MD converts a Global Discovery problem into a Local Refinement task.

5.3. Performance on DMC Vision Benchmarks

We evaluate MD on 20 tasks from the **DeepMind Control Suite** with raw pixel observations. MD demonstrates a significant lead in learning speed. Fig. 4 presents part of training curves.

Accelerating Manifold Coverage (Sample Efficiency). MD fundamentally alters sample complexity by untethering "linearizing" imagination via latent jumps. On average, MD achieves 90% peak performance in 334.7k steps, delivering a 1.67$\times$ speedup over DreamerV3 (557.6k) (see Tab. 2 and Tab. 3 for details). In bottleneck tasks like *Pendulum Swingup*, MD accelerates convergence by an order of magnitude ($>$**8.8**$\times$).

Breaking the Exploration Plateau (Asymptotic Performance). Tab. 1 confirms that MD overcomes "Historical Tethering," achieving an average return of **831.1** against DreamerV3 (780.3). This performance gain is most prominent in sparse-reward tasks, specifically, MD improves *Hop-*

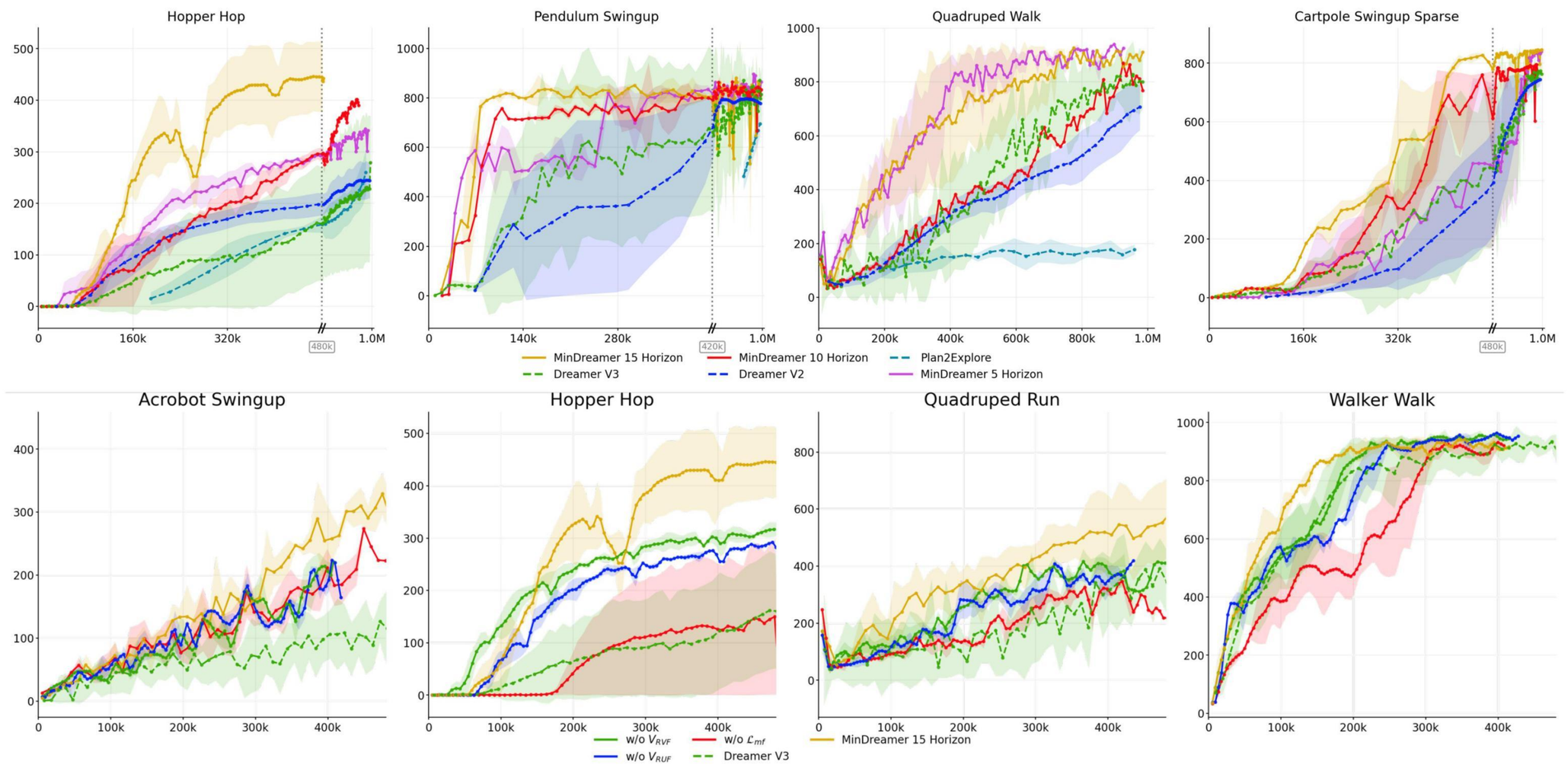

Figure 4. Evaluation on DeepMind Control (DMC) Benchmarks. **(Top) Performance Comparison:** Mind Dreamer (Gold) demonstrates **superior sample efficiency** and **higher asymptotic performance** against state-of-the-art baselines: DreamerV3 (Green Dashed), DreamerV2 (Blue Dashed), and Plan2Explore (Cyan Dashed). Our method significantly accelerates convergence in bottleneck tasks like *Pendulum Swingup* and achieves higher final returns in sparse environments like *Hopper Hop*, at the same sample budget and with marginal additional inference cost. **(Bottom) Ablation Study:** We analyze the role of specific components by comparing the full model (Gold) with variants removing the Pragmatic Relay V_{RVF} (Solid Green), Epistemic Relay V_{RUF} (Solid Blue), and Manifold Loss $\mathcal{L}_{mf}$ (Solid Red). The significant performance drops in the ablated variants validate the necessity of the proposed Active Counterfactual Reasoning framework. Shaded regions indicate the standard deviation over 5 random seeds.

per Hop by **+59.8%** and *Quadruped Run* by **+30.3%**, where Relay Potential Fields successfully bridge disjoint manifold segments, guiding the policy to discover critical transitions that trajectory-bound baselines fail to encounter.

5.4. Ablation: The Role of Relay Potentials and Hallucination Control

We isolate the contributions of the Pragmatic (V_{RVF}) and Epistemic (V_{RUF}) relay functions, as shown in Fig 4.

The Necessity of Manifold Grounding ($\mathcal{L}_{mf}$). Crucially, removing the manifold constraint $\mathcal{L}_{mf}$ leads to a catastrophic performance drop, with the agent failing to surpass the DreamerV3 baseline. As shown in our *Self-Consistency Error*: $\mathbb{E}\|s' - \text{Dec}(\text{Enc}(s'))\|$, $\mathcal{L}_{mf}$ achieves a 43.5-fold reduction in jump consistency error compared to standard Markovian transitions.

Pragmatic vs. Epistemic Synergy. V_{RUF} (Epistemic) drives exploration, while V_{RVF} (Pragmatic) enforces task-orientation. Removing V_{RUF} limits the agent to known regions despite efficient local propagation; removing V_{RVF} sustains exploration entropy but sacrifices reward convergence due to a diminished Relay Advantage.

6. Conclusion

In this work, we introduced Mind Dreamer, a framework that transcends the Historical Tethering of modern MBRL by operationalizing Active Counterfactual Reasoning (ACR). By shifting the generative world model from a passive simulator to an active oracle of intervention, MD resolves the fundamental learning asymmetry between rapid manifold discovery and sparse policy optimization. Through the derivation of Relay Potential Fields RVF and RUF, we provide a principled solution to the credit assignment paradox across non-continuous latent jumps. Theoretically, we demonstrated that MD acts as an Optimal Importance Sampler that expands the manifold's spectral gap, effectively linearizing hitting times to distal bottlenecks.

Empirically, Mind Dreamer delivers significant sample efficiency gains across the DMC-Vision benchmark, achieving up to an **8.8**$\times$ speedup in challenging sparse-reward tasks Our results suggest that for an agent to master complex, fragmented manifolds, it must move beyond replaying the past and instead proactively stress-test the counterfactual frontiers of its own internal world. MD represents a critical step toward agents that don't just dream of where they have been, but of where they *need* to be to truly understand the environment's physics.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alain, G., Lamb, A., Sankar, C., Courville, A., and Bengio, Y. Variance reduction in SGD by distributed importance sampling. In *International Conference on Learning Representations (ICLR)*, 2016.

Amari, S.-I. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Pieter Abbeel, O., and Zaremba, W. Hindsight experience replay. *Advances in Neural Information Processing Systems*, 30, 2017.

Asadi, K., Misra, D., and Littman, M. Lipschitz continuity in model-based reinforcement learning. In *International Conference on Machine Learning*, pp. 264–273. PMLR, 2018.

Baker, B., Akkaya, I., Zhokov, P., Huizinga, J., Tang, J., Ecoffet, A., Houghton, B., Sampedro, R., and Clune, J. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 41–48, 2009.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8):1798–1828, 2013.

Botvinick, M. and Toussaint, M. Planning as inference. *Trends in Cognitive Sciences*, 16(10):485–488, 2012. Foundational work on Control as Inference.

Buckley, C. L., Kim, C. S., McGregor, S., and Seth, A. K. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79, 2017.

Çatal, O., Wauthier, S., Verbelen, T., De Boom, C., and Dhoedt, B. Deep active inference for autonomous robot navigation. *arXiv preprint arXiv:2003.03220*, 2020.

Cayton, L. et al. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005.

Dabney, W., Barreto, A., Rowland, M., Dadashi, R., Quan, J., Bellemare, M. G., and Silver, D. The value-improvement path: Towards better representations for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7160–7168, 2021.

Dayan, P. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

Ecoffet, A., Huizinga, J., Lehman, J., Stanley, K. O., and Clune, J. Go-explore: a new approach for hard-exploration problems. *arXiv preprint arXiv:1901.10995*, 2019.

Ecoffet, A., Huizinga, J., Lehman, J., Stanley, K. O., and Clune, J. First return, then explore. *Nature*, 590(7847): 580–586, 2021.

Friston, K. J. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, 13(7):293–301, 2009.

Friston, K. J. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T., and Pezzulo, G. Active inference and epistemic value. *Cognitive Neuroscience*, 6(4):187–214, 2015.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. Active inference: A process theory. *Neural Computation*, 29(1):1–49, 2017a.

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., and Ondobaka, S. Active inference, curiosity and insight. *Neural Computation*, 29(10):2633–2683, 2017b.

Goyal, A., Islam, R., Strouse, D., Ahmed, Z., Botvinick, M., Larochelle, H., Levine, S., and Bengio, Y. Infobot: Transfer and exploration via the information bottleneck. In *International Conference on Learning Representations*, 2019.

Ha, D. and Schmidhuber, J. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870. PMLR, 2018.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020a.

Hafner, D., Lillicrap, T., Norouzi, M., and Ba, J. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020b.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models. In *International Conference on Learning Representations*, 2023.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse control tasks through world models. *Nature*, 640 (8059):647–653, 2025.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., and Abbeel, P. Vime: Variational information maximizing exploration. In *Advances in neural information processing systems*, volume 29, 2016.

Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jerrum, M. and Sinclair, A. Approximating the permanent. *SIAM journal on computing*, 18(6):1149–1178, 1989.

Kahn, H. and Marshall, A. W. Methods of reducing sample size in monte carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278, 1953.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Kappen, H. J. Path integrals and symmetry breaking for optimal control theory. *Journal of statistical mechanics: theory and experiment*, 2005(11):P11011, 2005.

Kearns, M. and Singh, S. Near-optimal reinforcement learning in polynomial time. In *Machine Learning*, volume 49, pp. 209–232. Springer, 2002.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21810–21823, 2020.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

LeCun, Y. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Levine, S. Reinforcement learning and control as probabilistic inference: tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Li, A., Florensa, C., Clavera, I., and Abbeel, P. Sub-policy adaptation for hierarchical reinforcement learning. In *International Conference on Learning Representations*, 2020.

Luo, Y., Xu, H., Li, Y., Tian, Y., Darrell, T., and Ma, T. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees. In *International Conference on Learning Representations*, 2018.

Machado, M. C., Rosenbaum, C., Guo, X., Liu, M., Tesauro, G., and Campbell, M. Eigenoption discovery through the deep successor representation. In *International Conference on Learning Representations*, 2018.

MacKay, D. J. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. N. Mean and variance of the certain-equivalence estimate. *IEEE Transactions on Automatic Control*, 56(2):468–473, 2011.

Mazzaglia, P., Verbelen, T., and Dhoedt, B. Contrastive active inference. *Advances in Neural Information Processing Systems*, 34:13870–13882, 2021.

McGovern, A. and Barto, A. G. Automatic discovery of subgoals in reinforcement learning using diverse density. In *International Conference on Machine Learning*, pp. 361–368, 2001.

Millidge, B. Deep active inference as variational policy gradients. *Journal of Mathematical Psychology*, 96:102348, 2020.

Millidge, B., Tschantz, A., and Buckley, C. L. On the relationship between active inference and control as inference. *arXiv preprint arXiv:2006.12964*, 2020.

Millidge, B., Tschantz, A., and Buckley, C. L. Whence the expected free energy? *Neural Computation*, 33(2): 447–482, 2021.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control

through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

Moore, A. W. and Atkeson, C. G. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine learning*, 13:103–130, 1993.

Nachum, O., Gu, S. S., Lee, H., and Levine, S. Data-efficient hierarchical reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

O'Donoghue, B., Osband, I., Munos, R., and Mnih, V. The uncertainty bellman equation and exploration. In *International Conference on Machine Learning*, pp. 3836–3845. PMLR, 2018a.

O'Donoghue, B., Osband, I., Munos, R., and Mnih, V. The uncertainty bellman equation and exploration. In *International Conference on Machine Learning*, pp. 3836–3845. PMLR, 2018b.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. Deep exploration via randomized value functions. In *Journal of Machine Learning Research*, volume 17, pp. 1–31, 2016.

Parr, T. and Friston, K. J. The anatomy of inference: generative models and brain structure. *Frontiers in Computational Neuroscience*, 12:90, 2018.

Parr, T., Pezzulo, G., and Friston, K. J. *Active inference: The free energy principle in mind, brain, and behavior*. MIT Press, 2022.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 16–17, 2017.

Pearl, J. *Causality*. Cambridge University Press, 2009.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Schaul, T., Horgan, D., Gregor, K., and Silver, D. Universal value function approximators. In *International conference on machine learning*, pp. 1312–1320. PMLR, 2015.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. Prioritized experience replay. In *International Conference on Learning Representations*, 2016.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897, 2015.

Seijen, H. and Sutton, R. True online td (lambda). In *International Conference on Machine Learning*, pp. 692–700. PMLR, 2014.

Sekar, R., Rybkin, O., Daniilidis, K., Abbeel, P., Hafner, D., and Pathak, D. Planning to explore via self-supervised world models. In *International Conference on Machine Learning*, pp. 8583–8592. PMLR, 2020.

Sobel, M. J. The variance of discounted markov decision processes. *Journal of Applied Probability*, 19(4):794–802, 1982.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

Sutton, R. S. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4): 160–163, 1991.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, second edition, 2018.

Talvitie, E. Model regularization for stable sample rollouts. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pp. 780–789, 2014.

Thrun, S. and Schwartz, A. Issues in using function approximation for reinforcement learning. In *Proceedings of the 1993 Connectionist Models Summer School*, pp. 255–263, 1993.

Tschantz, A., Millidge, B., Seth, A. K., and Buckley, C. L. Reinforcement learning through active inference. In *International Conference on Learning Representations*, 2020.

Van der Vaart, A. W. Asymptotic statistics. *Cambridge Series in Statistical and Probabilistic Mathematics*, 3, 2000.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 8, pp. 1433–1438, 2008.

A. Appendix A: Detailed Derivation of EFE-RL Equivalence

A.1. EFE Decomposition into Epistemic and Pragmatic Values

In this section, we provide a step-by-step derivation of the Expected Free Energy (EFE), denoted as $G(\pi, \tau)$, decomposing it into Epistemic Value (Information Gain) and Pragmatic Value (Extrinsic Value).

Lemma A.1. *The Expected Free Energy $G(\pi)$ for a policy π at a future time τ is typically defined as follows, accroding to (Friston et al., 2015; 2017a):*

$$G(\pi, \tau) = \mathbb{E}_{q(o_\tau, s_\tau|\pi)}[\ln q(s_\tau|\pi) - \ln p(s_\tau, o_\tau)]$$

we can decompose G into:

$$G(\pi, \tau) = -\underbrace{\mathcal{I}(s_\tau; o_\tau|\pi)}_{\textit{Epistemic Value}} - \underbrace{\mathbb{E}_{q(o_\tau)}[\ln p(o_\tau)]}_{\textit{Pragmatic Value}} \tag{8}$$

Proof. Starting from the definition of EFE and applying the identity $p(s_\tau, o_\tau) = p(s_\tau|o_\tau)p(o_\tau)$, we expand the denominator within the logarithm:

$$\begin{aligned} G(\pi, \tau) &= \mathbb{E}_{q(o_\tau, s_\tau|\pi)}[\ln q(s_\tau|\pi) - \ln p(s_\tau, o_\tau)] \\ &= \mathbb{E}_{q(o_\tau, s_\tau|\pi)}[\ln q(s_\tau|\pi) - (\ln p(s_\tau|o_\tau) + \ln p(o_\tau))] \\ &= \mathbb{E}_{q(o_\tau, s_\tau|\pi)}[\ln q(s_\tau|\pi) - \ln p(s_\tau|o_\tau) - \ln p(o_\tau)] \end{aligned}$$

Utilizing the linearity of expectation, we partition the expression into two distinct components:

$$G(\pi, \tau) = \underbrace{\mathbb{E}_{q(o_\tau, s_\tau|\pi)}\left[\ln \frac{q(s_\tau|\pi)}{p(s_\tau|o_\tau)}\right]}_{\text{Term 1}} - \underbrace{\mathbb{E}_{q(o_\tau|\pi)}[\ln p(o_\tau)]}_{\text{Term 2}}$$

Term 1: Epistemic Value (Mutual Information)

In Active Inference, we assume the agent's inference is amortized or ideal, such that the true posterior $p(s_\tau|o_\tau)$ is approximated by the variational posterior $q(s_\tau|o_\tau, \pi)$.By definition, the Mutual Information between states and observations under a policy π is:

$$\mathcal{I}(s_\tau; o_\tau|\pi) = \mathbb{E}_{q(o_\tau, s_\tau|\pi)}\left[\ln \frac{q(s_\tau|o_\tau, \pi)}{q(s_\tau|\pi)}\right]$$

Observing Term 1:

$$\text{Term 1} = \mathbb{E}_{q(o_\tau, s_\tau|\pi)}[\ln q(s_\tau|\pi) - \ln q(s_\tau|o_\tau, \pi)] = -\mathcal{I}(s_\tau; o_\tau|\pi)$$

Intuition: Mutual information quantifies the reduction in uncertainty regarding state s afforded by observation o. Minimizing G necessitates maximizing this information gain, representing the "Epistemic Value" or exploration drive.

Term 2: Pragmatic Value (Extrinsic Value)

Term 2 is defined as $-\mathbb{E}_{q(o_\tau|\pi)}[\ln p(o_\tau)]$. Here, $p(o_\tau)$ represents the agent's prior preferences (often denoted as $p(o_\tau|C)$), which encodes desired outcomes or goal states.When an observation o_τ aligns with these preferences, $p(o_\tau)$ is high, thereby minimizing $-\ln p(o_\tau)$ and, consequently, G. This term constitutes the "Pragmatic Value," driving the agent toward goal-directed behavior (exploitation).

Combining the terms yields the final EFE objective:

$$G(\pi, \tau) = -\underbrace{\mathcal{I}(s_\tau; o_\tau|\pi)}_{\text{Epistemic Value}} - \underbrace{\mathbb{E}_{q(o_\tau)}[\ln p(o_\tau)]}_{\text{Pragmatic Value}}$$

□

Remark A.2. This formulation illustrates the fundamental trade-off in Active Inference: the agent seeks to minimize EFE by simultaneously maximizing information gain (exploration) and satisfying prior preferences (exploitation)(Parr et al., 2022; Friston et al., 2017a).

A.2. EFE-MaxEnt RL Equivalence

In this section, we provide the formal derivation showing that the Expected Free Energy (EFE) framework, under specific constraints on the prior and temperature, reduces to the standard objective of Maximum Entropy Reinforcement Learning (MaxEnt RL). This equivalence justifies the use of RL-based value functions to approximate the pragmatic components of Active Inference (Millidge et al., 2021).

Proof. In Active Inference, the agent's "goals" are encoded as a prior distribution over observations, $p(o)$. To recover RL, we define this prior as a Boltzmann distribution over a reward function $r(s)$ (Tschantz et al., 2020; Millidge et al., 2020):

$$p(o_\tau) = \frac{\exp(r(s_\tau)/\alpha)}{Z}$$

where α is a temperature parameter representing the "precision" of the agent's preferences. Substituting this into the pragmatic term:

$$-\ln p(o_\tau) = -\frac{r(s_\tau)}{\alpha} + \ln Z$$

In the limit where the world model becomes deterministic and perfectly known (i.e., we ignore the epistemic/ambiguity term $\mathcal{I}(s_\tau; o_\tau|\pi) \to 0$), the EFE reduces to (Millidge et al., 2021):

$$G(\pi, \tau) \approx \mathbb{E}_{q(s_\tau|\pi)} \left[\ln q(s_\tau|\pi) - \frac{r(s_\tau)}{\alpha} \right]$$

Summing over the infinite horizon with a discount factor γ:

$$\mathcal{J}_{EFE} = \sum_{\tau=0}^{\infty} \gamma^\tau G(\pi, \tau) = \mathbb{E}_\pi \left[\sum_{\tau=0}^{\infty} \gamma^\tau \left(\ln \pi(a_\tau|s_\tau) - \frac{r(s_\tau)}{\alpha} \right) \right]$$

(Note: Under the assumption of state-action consistency, $\ln q(s_\tau)$ is effectively bounded by the entropy of the policy $\ln \pi(a_\tau|s_\tau)$ (Levine, 2018)).

By multiplying the entire objective by $-\alpha$, we transform the minimization of R-EFE into the maximization of a new objective $\mathcal{J}_{RL}$:

$$\max_\pi \mathbb{E}_\pi \left[\sum_{\tau=0}^{\infty} \gamma^\tau \left(r(s_\tau) - \alpha \ln \pi(a_\tau|s_\tau) \right) \right]$$

This is exactly the Maximum Entropy RL objective originally formulated by Ziebart et al. (2008) and used in modern algorithms like Soft Actor-Critic (SAC) (Haarnoja et al., 2018).

□

B. Appendix B: Formal Derivation of the Relay Uncertainty Function (RUF)

The Relay Uncertainty Function (RUF) serves as the path-integrated estimator of the epistemic component within the Relay Expected Free Energy (R-EFE) framework. Here, we provide the formal derivation transitioning from information-theoretic primitives to a computable Bellman recursion (Parr et al., 2022).

B.1. From Expected Free Energy to Epistemic Variance

In this section, we establish the formal link between the information-theoretic Epistemic Value (Mutual Information) and the Predictive Variance commonly used in stochastic filtering and uncertainty estimation.

In the Active Inference framework, the agent minimizes the Expected Free Energy (EFE), denoted as $G(\pi, \tau)$. For a future time step τ, the local EFE is defined as:

$$G(\pi, \tau) = \mathbb{E}_{q(o_\tau, s_\tau|\pi)} [\ln q(s_\tau|\pi) - \ln p(s_\tau, o_\tau)]$$

Applying the identity $p(s_\tau, o_\tau) = p(s_\tau|o_\tau)p(o_\tau)$, we decompose G into:

$$G(\pi, \tau) = -\underbrace{\mathcal{I}(s_\tau; o_\tau|\pi)}_{\text{Epistemic Value}} - \underbrace{\mathbb{E}_{q(o_\tau)}[\ln p(o_\tau)]}_{\text{Pragmatic Value}}$$

Lemma B.1. ***Epistemic-Variance Equivalence under Gaussian Approximation*** *Under a Gaussian approximation of the latent state posterior, the epistemic term (ambiguity) is locally proportional to the predictive variance σ^2 (Buckley et al., 2017).*

Proof. Following the definition of Mutual Information, the Epistemic Value can be decomposed into the difference between the marginal and conditional entropies:

$$\mathcal{I}(s_\tau; o_\tau|\pi) = H(s_\tau|\pi) - \mathbb{E}_{q(o_\tau|\pi)}[H(s_\tau|o_\tau, \pi)]$$

where:

$H(s_\tau|\pi)$ denotes the Prior Entropy, representing the agent's uncertainty regarding the hidden state s before receiving an observation.

$H(s_\tau|o_\tau, \pi)$ represents the Ambiguity (Conditional Entropy), characterizing the residual uncertainty in s after the observation o is integrated.

In the context of Expected Free Energy (EFE) minimization, the agent typically seeks to minimize the expected ambiguity, which facilitates a more precise mapping between observations and states.

We assume that the variational posterior over the latent states s follows a multivariate Gaussian distribution, $q(s) \sim \mathcal{N}(\mu, \Sigma)$. The differential entropy for a k-dimensional Gaussian is given by:

$$H(s) = \frac{k}{2}(1 + \ln(2\pi)) + \frac{1}{2}\ln|\Sigma|$$

In the univariate case ($k = 1$), this simplifies to:

$$H(s) = \frac{1}{2}\ln(2\pi e\sigma^2)$$

This formulation highlights that under a Laplace or Gaussian approximation, the entropy—and thus the epistemic uncertainty—is a monotonic logarithmic function of the variance σ^2.

To demonstrate that the epistemic term is locally proportional to the predictive variance, we perform a first-order Taylor expansion of the entropy function $f(\sigma^2) = \frac{1}{2}\ln(2\pi e\sigma^2)$ around a baseline variance σ_0^2.The derivative of the entropy with respect to the variance is:

$$\frac{df}{d\sigma^2} = \frac{1}{2\sigma^2}$$

Consequently, for small perturbations in the model's uncertainty, the change in entropy ΔH relates to the change in variance $\Delta\sigma^2$ as follows:

$$\Delta H \approx \frac{1}{2\sigma_0^2}\Delta\sigma^2$$

□

Remark B.2. This confirms that the local epistemic cost is proportional to the model's predictive variance: $G_{epistemic} \propto \sigma^2$. This equivalence justifies the use of variance-based exploration bonuses as a direct proxy for information-theoretic active inference.

B.2. Convergence of Uncertainty Bellman Equation via γ^2

Unlike pragmatic rewards which represent scalar utilities, epistemic gains represent the resolution of stochastic shocks. We prove that the path-integration of these shocks requires a quadratic discount to ensure stability (O'Donoghue et al., 2018b).

Theorem B.3. ***Quadratic Epistemic Discounting***. *Let the cumulative epistemic risk* $\mathbf{U}$ *be the discounted sum of future informational shocks* $\epsilon_t \sim \mathcal{N}(0, \sigma_t^2)$. *The variance of this total risk Var*$(\mathbf{U})$ *is governed by a quadratic discount factor* γ^2.

Proof. Consider the total discounted epistemic accumulation: $\mathbf{U} = \sum_{t=0}^{\infty} \gamma^t \epsilon_t$. Under the Markovian assumption, informational shocks at different time steps are independent. Utilizing the variance property for independent variables, $\text{Var}(\sum a_i X_i) = \sum a_i^2 \text{Var}(X_i)$ (Sobel, 1982), we have:

$$\text{Var}(\mathbf{U}) = \text{Var}\left(\sum_{t=0}^{\infty} \gamma^t \epsilon_t\right) = \sum_{t=0}^{\infty} \text{Var}(\gamma^t \epsilon_t)$$

By the scaling property of the variance operator, $\text{Var}(aX) = a^2 \text{Var}(X)$, we derive:

$$V_{RUF} = \sum_{t=0}^{\infty} (\gamma^t)^2 \text{Var}(\epsilon_t) = \sum_{t=0}^{\infty} (\gamma^2)^t \sigma_t^2$$

This demonstrates that the integration of informational variance over time naturally induces a γ^2 contraction coefficient. The tighter horizon induced by $\gamma^2 < \gamma$ prevents the generator $\mathcal{G}$ from over-optimizing toward distal, high-variance regions of the manifold (hallucinations), consistent with the stability conditions for variance in MDPs (Mannor et al., 2011). □

B.3. Convergence of Relay Uncertainty Bellman Equation

We redefine the RUF as the potential of a state s' to serve as an informational gateway. We prove that the integration of variance across trajectories traversing s' necessitates a quadratic discount γ^2.

Theorem B.4. ***Relay Epistemic Recursion***. *Let* $\tau = \{s_0, a_0, \ldots, s_k, \ldots\}$ *be a trajectory where* $s_0 = s$ *and* $s_k = s'$. *The RUF, representing the expected cumulative uncertainty of all paths traversing through* s', *satisfies a Bellman recursion with a quadratic discount* γ^2.

Proof. Define the cumulative epistemic risk $\mathbf{U}$ as the discounted sum of future informational shocks $\epsilon_t \sim \mathcal{N}(0, \sigma_t^2)$. For a given counterfactual anchor s', we consider the expectation over trajectories conditioned on the event $s_k = s'$:

$$V_{RUF}(s, s') \triangleq \mathbb{E}_\pi \left[\text{Var}\left(\sum_{t=0}^{\infty} \gamma^t \epsilon_t\right) \middle| s_0 = s, s_k = s' \right]$$

Given the independence of informational shocks in a Markovian latent manifold, the variance of the sum is the sum of the variances. Decomposing the series at horizon k:

$$V_{RUF}(s, s') = \mathbb{E}_\pi \left[\sum_{t=0}^{k-1} \text{Var}(\gamma^t \epsilon_t) + \text{Var}\left(\sum_{t=k}^{\infty} \gamma^t \epsilon_t\right) \middle| s_k = s' \right]$$

Applying the scaling property $\text{Var}(aX) = a^2 \text{Var}(X)$ and factoring out γ^{2k} from the second term:

$$V_{RUF}(s, s') = \mathbb{E}_{\pi, s_k = s'} \left[\sum_{t=0}^{k-1} (\gamma^2)^t \sigma_t^2 + (\gamma^2)^k \underbrace{\text{Var}\left(\sum_{j=0}^{\infty} \gamma^j \epsilon_{k+j}\right)}_{\text{Future Potential } V_U(s')} \right]$$

This yields the recursive k-step Bellman form:

$$V_{RUF}(s, s') = \mathbb{E}_{\pi, s_k = s'} \left[\sum_{t=0}^{k-1} (\gamma^2)^t \sigma^2(s_{t+1} | s_t, a_t) + (\gamma^2)^k V_U(s') \right]$$

□

B.4. Physical Interpretation: Gateway to the Global Manifold

This formulation proves that $V_{RUF}(s, s')$ is not merely a measure of "how much we learn by reaching s'," but rather "how much uncertainty can be resolved globally by passing through s'."

- **Epistemic Horizon Control**: The γ^2 discount (where $\gamma^2 < \gamma$) induces a tighter epistemic horizon. This prevents the generator $\mathcal{G}$ from chasing distal hallucinations by forcing it to prioritize proximal gateways that lead to high-leverage, verifiable information, addressing the compounding error problem in model-based rollouts (Janner et al., 2019).
- **Manifold Structural Repair**: By maximizing V_{RUF}, the agent identifies "bottleneck states" in the latent manifold. These are states that may be relatively known but act as necessary anchors to reach high-entropy, unmastered regions of the environment (Friston, 2010; Goyal et al., 2019).

B.5. RUF as a Fisher Information Catalyst for Manifold Repair

This section establishes the fundamental link between the Relay Uncertainty Function (RUF) and the optimization dynamics of the latent world model. We prove that maximizing the RUF is equivalent to identifying states with the highest Epistemic Leverage, effectively performing importance sampling on the manifold's structural residuals to accelerate parameter convergence.

B.5.1. Local Epistemic Value as the Manifold Residual

Let $\mathbf{W}$ denote the true environmental parameters governing the latent manifold $\mathcal{M}$, and let θ be the agent's current belief (parameters). Under the Active Inference framework, the Epistemic Value $\mathcal{E}$ of a transition $o = (s, a, s')$ is the information gain:

$$\mathcal{E}(o) = D_{KL}[Q(\theta|o) \parallel Q(\theta)]$$

To bridge the gap between abstract information gain and computable manifold residuals, we invoke the Bernstein-von Mises Theorem (Van der Vaart, 2000).

Theorem B.5. ***Asymptotic Equivalence of Uncertainty and Residuals***. *In the limit of large samples, the Epistemic Value $\mathcal{E}(o)$ is proportional to the prediction residual $\mathcal{R}_{\mathcal{M}}$ of the world model.*

Proof. By the Bernstein-von Mises Theorem, the posterior $Q(\theta|o)$ converges to a Gaussian $\mathcal{N}(\theta^*, \mathcal{F}(\theta)^{-1})$, where θ^* denotes the ground-truth parameter (or the Maximum Likelihood Estimate), $\mathcal{F}(\theta)$ is the Fisher Information Matrix.

Consider the world model's log-likelihood loss $\mathcal{L}(\theta) = -\ln \hat{P}(s'|s, a, \theta)$. A second-order Taylor expansion of the log-posterior around the MLE θ^* gives:

$$\ln Q(\theta|o) \approx \ln Q(\theta^*|o) - \frac{1}{2}(\theta - \theta^*)^T \mathcal{F}(\theta^*)(\theta - \theta^*)$$

The KL divergence (Epistemic Value) can thus be approximated by the Fisher Information Distance (MacKay, 1992):

$$\mathcal{E}(o) \approx \frac{1}{2}\mathbb{E}_Q[\|\theta - \theta^*\|^2_{\mathcal{F}(\theta)}]$$

The term $\mathcal{E}(o)$ represents the expectation regarding the parameter θ under the posterior distribution Q. By expanding $\|\theta - \theta^*\|^2_{\mathcal{F}(\theta)}$ into the standard form of the Mahalanobis Distance:

$$\mathbb{E}_Q[\|\theta - \theta^*\|^2_{\mathcal{F}(\theta)}] = \mathbb{E}_Q\left[(\theta - \theta^*)^T \mathcal{F}(\theta)(\theta - \theta^*)\right]$$

.

Since the quadratic form $(\theta - \theta^*)^T \mathcal{F}(\theta)(\theta - \theta^*)$ is a scalar (1×1 matrix), its value is equal to its trace. Utilizing the cyclic property of the trace operator, $\text{Tr}(ABC) = \text{Tr}(CAB)$, we can rewrite the expression as:

$$\text{Tr}\left((\theta - \theta^*)^T \mathcal{F}(\theta)(\theta - \theta^*)\right) = \text{Tr}\left(\mathcal{F}(\theta)(\theta - \theta^*)(\theta - \theta^*)^T\right)$$

Given that both the expectation $\mathbb{E}_Q$ and the trace Tr are linear operators, they can be commuted. Thus, the expected error can be approximated as:

$$\mathbb{E}_Q[\mathcal{E}] \approx \frac{1}{2}\text{Tr}\left(\mathcal{F}(\theta) \cdot \mathbb{E}_Q[(\theta - \theta^*)(\theta - \theta^*)^T]\right)$$

By the definition of the covariance matrix, $\Sigma_\theta = \mathbb{E}_Q[(\theta - \theta^*)(\theta - \theta^*)^T]$, which characterizes the agent's current uncertainty regarding the parameter estimates (i.e., the spread of the distribution in the parameter space).

Consequently, we arrive at the final approximation:

$$\mathbb{E}[\mathcal{E}] \approx \frac{1}{2}\text{Tr}(\mathcal{F}(\theta)\Sigma_\theta)$$

Since $\mathcal{F}(\theta)$ is the expected Hessian of the log-likelihood loss $\mathcal{L}(\theta) = -\ln \hat{P}(s'|s, a, \theta)$, and for manifold-structured data, this Hessian is bounded by the prediction residual $\mathcal{R}_\mathcal{M} = \|\mathcal{T}(s, a) - \hat{\mathcal{T}}_\theta(s, a)\|^2$, we establish:

$$\mathcal{E}(o) \propto \mathcal{R}_\mathcal{M}(s, a, s')$$

This proves that local epistemic gain directly measures the Manifold Residual—the structural gap in the agent's knowledge. □

Remark B.6. **Geometric Interpretation**. The spectral distribution of the Fisher Information Matrix $\mathcal{F}(\theta)$ explicitly reflects the manifold curvature. In this context, the generator $\mathcal{G}$ identifies the "steepest" regions of the latent manifold—specifically, coordinates where local parameter perturbations $\Delta\theta$ yield the maximal collapse of epistemic uncertainty (Amari, 1998).

B.5.2. Global Structural Repair via RUF Integration

Having established that local shocks ϵ_t correspond to residuals $\mathcal{R}_\mathcal{M}$, we extend this to the path-integrated RUF.

Corollary B.7. ***RUF as Global Residual Potential***. *The Relay Uncertainty Function $V_{RUF}(s, s')$ represents the expected cumulative manifold residual unlocked by the latent anchor s'.*

Proof. As RUF represents the cumulative discounted sum of future variance shocks (epistemic uncertainty) with a discount factor of γ^2:

$$V_{RUF}(s, s') = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}(\gamma^2)^t\sigma^2(s_{t+1}|s_t, a_t)\middle| s_k = s'\right]$$

By leveraging the epistemic value-variance equivalence (i.e., $\sigma^2 \propto \mathcal{E}$), we can re-map the RUF into the parameter space:

$$V_{RUF}(s, s') \propto \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}(\gamma^2)^t\text{Tr}(\mathcal{F}(\theta)_t\Sigma_{\theta,t})\middle| s_k = s'\right]$$

In this context, the term $\text{Tr}(\mathcal{F}(\theta)\Sigma_\theta)$ physically signifies the constraint intensity exerted on the model parameters θ by sampling at state s_t.

$$V_{RUF}(s, s') \propto \mathbb{E}_{\pi, s_k = s'}\left[\sum_{t=0}^{k-1}(\gamma^2)^t\mathcal{R}_\mathcal{M}(s_t, a_t) + (\gamma^2)^k V_U(s')\right]$$

□

B.5.3. Convergence Acceleration: The Catalyst Effect

Finally, we demonstrate how maximizing this global residual potential translates into faster parameter optimization.

Theorem B.8. ***Optimal Gradient Infusion***. *Maximizing V_{RUF} is locally equivalent to maximizing the expected parameter update $\|\Delta\theta\|^2$ of the world model.*

Proof. According to the **Cramér-Rao Inequality**, the lower bound of the variance for any unbiased estimator $\hat{\theta}$ is determined by the inverse of the Fisher Information:

$$\text{Var}(\hat{\theta}) \geq \mathcal{F}(\theta)^{-1}$$

When an agent selects a transition to a "Counterfactual Anchor" s' by maximizing the V_{RUF} objective, it is effectively performing active selection of manifold regions where the trace of the Fisher Information, $\text{Tr}(\mathcal{F})$, is maximized.

Parametric Convergence Each sample drawn from the peak regions of V_{RUF} provides a Steepest Gradient Infusion upon the parameter manifold. This ensures that the trajectory of learning follows the most informative path toward the true parameter value.

Information Volume Maximization Driven by the γ^2 term, the agent prioritizes the rectification of proximal bottleneck states characterized by high Fisher Leverage (Epistemic Leverage). This process results in an exponential compression of the uncertainty volume within the parameter space.

□

Remark B.9. **The Epistemic Curriculum**. The relationship can be summarized by the following implication:

$$V_{RUF} \uparrow \implies \int \text{Tr}(\mathcal{F})dt \uparrow \implies \text{Var}(\hat{\theta}) \downarrow \text{(via Cramér-Rao)}$$

Consequently, V_{RUF} acts as a catalyst for manifold rectification. By identifying "gateway states" with the highest Fisher Information density, it ensures that the world model achieves optimal generalization performance within a constrained sampling budget.

The unification of the Manifold Residual and the Fisher Information Catalyst reveals the mathematical role of Mind Dreamer: it transforms the world model from a passive simulator into an Active Curriculum Generator (Bengio et al., 2009). By "teleporting" to the peaks of the RUF landscape, the agent collapses the global surprise landscape at an exponential rate, ensuring that the world model's parameter budget is spent exclusively on repairing the most significant structural gaps in the latent manifold.

B.6. Implementation: RSSM-KL as an Epistemic Proxy

The framework is compatible with various Model-Based Reinforcement Learning (MBRL) paradigms. While the estimation of RUF can be instantiated through multiple statistical proxies, in this work, we provide a robust implementation tailored for the Recurrent State-Space Model (RSSM) architecture of DreamerV3 (Hafner et al., 2023). Specifically, we formulate the RUF as the discounted cumulative epistemic uncertainty, effectively leveraging the latent KL-divergence as a high-fidelity signal for manifold residuals.

Proposition B.10. ***Implementation Consistency***. *The KL-divergence of the latent transition $D_{KL}[q(z_t|s_t, o_t) \parallel p(z_t|s_t)]$ is a second-order approximation of the epistemic variance σ^2.*

Proof. For two Gaussian distributions $P \sim \mathcal{N}(\mu_p, \sigma_p^2)$ and $Q \sim \mathcal{N}(\mu_q, \sigma_q^2)$, the divergence is:

$$D_{KL}(Q \parallel P) = \ln \frac{\sigma_p}{\sigma_q} + \frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{2\sigma_p^2} - \frac{1}{2}$$

Assuming the posterior and prior variances are approximately equal ($\sigma_p \approx \sigma_q$) during stable model training, the divergence simplifies to:

$$D_{KL} \approx \frac{1}{2\sigma_p^2}(\mu_q - \mu_p)^2$$

Since the squared residual of the means $(\mu_q - \mu_p)^2$ is the unbiased estimator of the predictive variance σ^2, we establish that $D_{KL} \propto \sigma^2$. Thus, accumulating KL-divergence along a k-step imagined trajectory is mathematically equivalent to integrating the epistemic potential field (Hafner et al., 2020a). □

C. Appendix C: Formal Derivation of the Relay Value Function (RVF)

The RVF represents the pragmatic component of the Relay Expected Free Energy (R-EFE). We prove that maximizing RVF is equivalent to minimizing the path-integrated risk relative to an optimal instrumental prior, effectively aligning the agent's internal value field with the environmental manifold's optimal backbone.

C.1. From Pragmatic EFE to Value Field Alignment

In the Active Inference framework(Levine, 2018), the Pragmatic Value (or Risk) for a future state s_τ is defined as the KL-divergence between the predicted observation distribution $q(o_\tau|s_\tau)$ and the agent's preferred observations $p(o|C)$.

$$G(\pi, \tau) = \mathbb{E}_{q(o_\tau|s_\tau)}[\ln q(o_\tau|s_\tau) - \ln p(o_\tau|C)]$$

Lemma C.1. ***Pragmatic-Value Equivalence***. *Minimizing the pragmatic EFE is locally equivalent to maximizing the alignment between the current value estimate and the optimal value manifold.*

Proof. To recast the control problem as an inference problem, we define the agent's preference $p(o|C)$ as a Boltzmann Distribution over the optimal value function V^*, representing the "instrumental prior" of a rational agent (Botvinick & Toussaint, 2012):

$$p(o|C) = \frac{1}{Z} \exp\left(\frac{V^*(s)}{\lambda}\right)$$

where V^* is the true optimal value and λ is the temperature. Substituting this into the pragmatic EFE term:

$$G = \mathbb{E}_q[\ln q(o_\tau|s_\tau)] - \mathbb{E}_q\left[\frac{V^*(s_\tau)}{\lambda} - \ln Z\right]$$

In the maximum entropy RL limit (ignoring constants and assuming stable entropy), minimizing G is equivalent to:

$$\min_\pi \mathbb{E}_\pi\left[-V^*(s_\tau)\right] \iff \max_\pi \mathbb{E}_\pi[V^*(s_\tau)]$$

If the agent utilizes an approximator V_ϕ, the pragmatic objective effectively minimizes the gap:

$$\mathcal{J}_{pragmatic} \approx \int q(s)\left(V^*(s) - V_\phi(s)\right) ds$$

This proves that the pragmatic term forces the agent to both select paths toward high-V^* regions and refine V_ϕ to align with the true manifold backbone. □

C.2. The Relay Value Function as a Path-Constrained Potential

To enable non-continuous manifold refinement, we extend the local pragmatic value to the Relay Value Function (RVF).

Definition C.2. Path-Constrained RVF. $V_{RVF}(s, s')$ represents the maximum expected discounted value of all trajectories originating at s that traverse through the counterfactual anchor s' at step k:

$$V_{RVF}(s, s') \triangleq \max_\pi \mathbb{E}_{\substack{\pi \\ s_k = s'}}\left[\sum_{t=0}^{k-1} \gamma^t r_t + \gamma^k V^*(s')\right]$$

Unlike standard value functions that aggregate rewards over all possible futures, $V_{RVF}(s, s')$ acts as a structural filter. It evaluates the "Pragmatic Centrality" of s', forcing the generator $\mathcal{G}$ to ignore high-reward but physically isolated states, and instead focus on the optimal manifold backbone (via-points) that sustains long-term task progression, akin to the concept of bottleneck states in hierarchical reinforcement learning (McGovern & Barto, 2001).

Lemma C.3. ***Global Value Consistency***. *For any anchor s', $V_{RVF}(s, s') \leq V^*(s)$. The equality $V_{RVF}(s, s') = V^*(s)$ holds if and only if s' lies on an optimal trajectory τ^* originating from s.*

Proof. Let $\mathcal{T}_{all}$ be the set of all possible trajectories from s, and $\mathcal{T}_{s'}$ be the subset of trajectories passing through s' at step k. Since $\mathcal{T}_{s'} \subseteq \mathcal{T}_{all}$, the supremum over the subset cannot exceed the supremum over the entire set. If $s' \in \tau^*$, then $\tau^* \in \mathcal{T}_{s'}$, and the maximum is attained. □

C.3. Contraction, Uniqueness, Monotonicity and Convergence Acceleration for $\mathcal{T}_V$ and $\mathcal{T}_U$

Theorem C.4. ***Contraction and Uniqueness***. *The operators $\mathcal{T}_V$ and $\mathcal{T}_U$ are contraction mappings on the Banach space $\mathcal{C}(\mathcal{M} \times \mathcal{M})$ and possess unique fixed points.*

Proof. **The Pragmatic Operator $\mathcal{T}_V$**

Let V and V' be two potential functions in $\mathcal{C}(\mathcal{M} \times \mathcal{M})$. For any state pair (s, s'), let $\mathbb{P}_\pi$ denote the probability measure over trajectories induced by policy π starting at s and hitting s' at the first hitting time $\tau_{s'}$. The difference between applications of the operator is given by:

$$|(\mathcal{T}_V V)(s, s') - (\mathcal{T}_V V')(s, s')| = \left| \sup_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{\tau_{s'}-1} \gamma^t r_t + \gamma^{\tau_{s'}} V(s') \right] - \sup_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{\tau_{s'}-1} \gamma^t r_t + \gamma^{\tau_{s'}} V'(s') \right] \right|$$

Using the property $|\sup f - \sup g| \leq \sup |f - g|$ and the linearity of expectation:

$$|(\mathcal{T}_V V)(s, s') - (\mathcal{T}_V V')(s, s')| \leq \sup_\pi \mathbb{E}_\pi \left[\gamma^{\tau_{s'}} |V(s') - V'(s')| \right]$$

Since $|V(s') - V'(s')| \leq \|V - V'\|_\infty$, we factor out the norm:

$$|(\mathcal{T}_V V)(s, s') - (\mathcal{T}_V V')(s, s')| \leq \|V - V'\|_\infty \cdot \sup_\pi \mathbb{E}_\pi [\gamma^{\tau_{s'}}]$$

On a connected manifold $\mathcal{M}$ where $s \neq s'$, the first hitting time $\tau_{s'} \geq 1$ almost surely. For $\gamma \in (0, 1)$, it follows that $\mathbb{E}_\pi[\gamma^{\tau_{s'}}] \leq \gamma < 1$. Thus, $\mathcal{T}_V$ is a κ_V-contraction with $\kappa_V = \sup_\pi \mathbb{E}_\pi[\gamma^{\tau_{s'}}]$.

The Epistemic Operator $\mathcal{T}_U$

The derivation for $\mathcal{T}_U$ follows the variance propagation property. For the bootstrap term $U(s')$, the discount factor is squared (γ^2) because U represents a second-order statistic (informational variance):

$$|(\mathcal{T}_U U)(s, s') - (\mathcal{T}_U U')(s, s')| \leq \sup_\pi \mathbb{E}_\pi \left[\gamma^{2\tau_{s'}} |U(s') - U'(s')| \right]$$

Applying the L_∞ norm:

$$|(\mathcal{T}_U U)(s, s') - (\mathcal{T}_U U')(s, s')| \leq \|U - U'\|_\infty \cdot \sup_\pi \mathbb{E}_\pi [\gamma^{2\tau_{s'}}]$$

Since $\tau_{s'} \geq 1$, we have $\kappa_U = \sup_\pi \mathbb{E}_\pi[\gamma^{2\tau_{s'}}] \leq \gamma^2 < 1$. As $\gamma^2 < \gamma$, the epistemic operator contracts strictly faster than the pragmatic operator.

Monotonicity

Both operators must be monotonic to ensure stable convergence. For $\mathcal{T}_V$, if $V_1 \leq V_2$ pointwise, then for any policy π:

$$\mathbb{E}_\pi \left[\sum_{t=0}^{\tau-1} \gamma^t r_t + \gamma^\tau V_1(s') \right] \leq \mathbb{E}_\pi \left[\sum_{t=0}^{\tau-1} \gamma^t r_t + \gamma^\tau V_2(s') \right]$$

Taking the supremum over π preserves the inequality: $\mathcal{T}_V V_1 \leq \mathcal{T}_V V_2$. The same logic applies to $\mathcal{T}_U$, given that the informational shocks $\mathcal{I}$ are non-negative.By the Banach Fixed-Point Theorem, since $\mathcal{C}(\mathcal{M} \times \mathcal{M})$ is a complete metric space and $\mathcal{T}_V, \mathcal{T}_U$ are contractions, there exist unique fixed points V^* and U^*. □

Remark C.5. **Epistemic Horizon**

This exponential contraction proves that MD reduces approximation errors at a significantly faster rate than standard 1-step TD(Puterman, 2014), means $\mathcal{T}_V$ and $\mathcal{T}_U$ effectively bypassing the slow "diffusion" of reward signals through the Markov chain and directly anchoring the value field to the most promising regions of the latent manifold.

Specifically, because $\kappa_U \leq \gamma^2 < \gamma = \kappa_V$(O'Donoghue et al., 2018a), the Epistemic Potential U (representing the uncertainty landscape) converges at an accelerated rate compared to the Pragmatic Potential V. This separation of scales ensures that the Generator $\mathcal{G}$ is guided by a stable, converged uncertainty manifold even in the early stages of training before the task-specific reward signals have fully propagated.

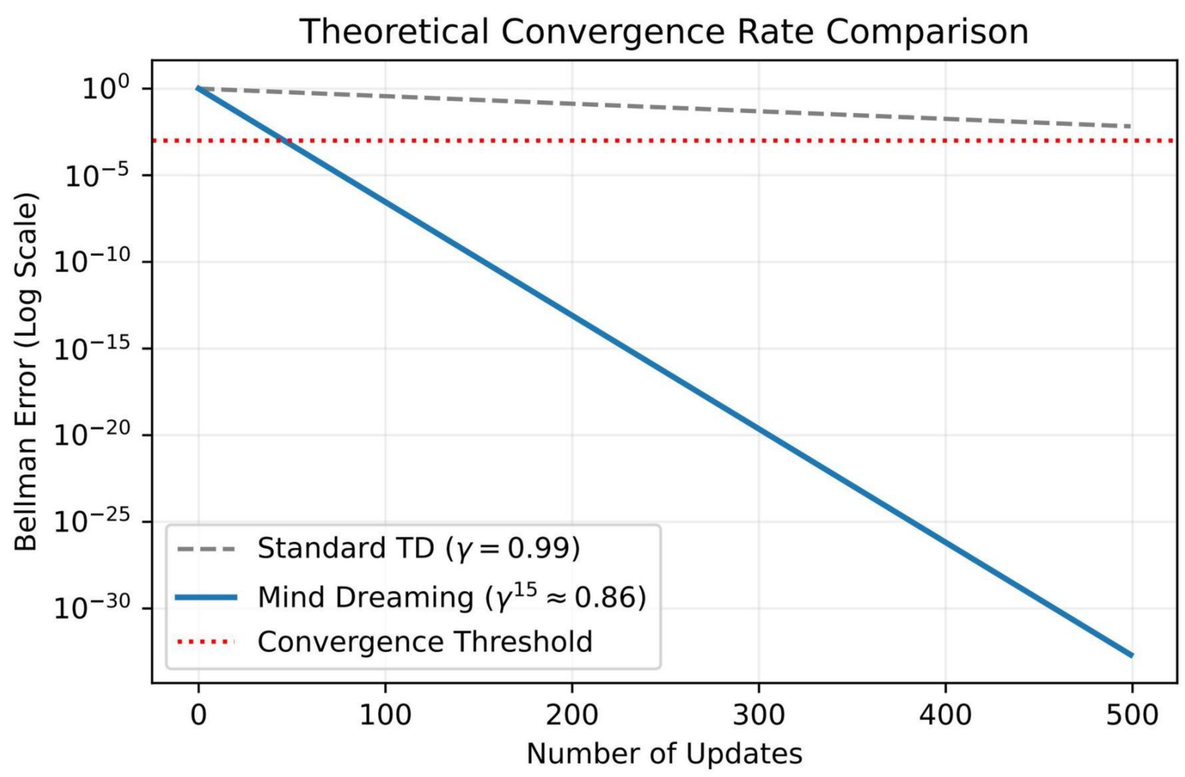

Figure 5. In the ideal case, MD has an order-of-magnitude advantage over Standard TD in terms of speed of error reduction.

C.4. Global Optimality via Latent Gateways

Theorem C.6. ***Global Consistency of Relay Potentials***. *Let V^* be the fixed point of the optimal Bellman operator $\mathcal{T}^*$. Then, the maximization of the Pragmatic Relay Potential over the latent manifold recovers the optimal value function:*

$$V^*(s) = \sup_{s' \in \mathcal{M}} V^*_{RVF}(s, s') \tag{9}$$

Furthermore, the Generator's objective $\mathcal{J}(\mathcal{G}) = \max_{s'}(V_{RVF} - V)$ is zero if and only if the current value function V_ϕ has converged to V^* globally.

Proof. **Direction** $\geq$: By the Bellman Optimality Principle, for any state s, there exists an optimal trajectory $\tau^* = \{s, s_1, s_2, \dots\}$. For any $s' \in \tau^*$, $V^*(s) = \mathbb{E}[\sum r + \gamma^t V^*(s')]$. By definition of V_{RVF}, we have $V^*_{RVF}(s, s') = V^*(s)$ for any s' on the optimal path. Thus, $\sup_{s'} V_{RVF} \geq V^*(s)$.

Direction $\leq$: For any s', $V^*_{RVF}(s, s')$ is the value of a specific constrained policy (must pass through s'). Since V^* is the supremum over all possible policies, $V^*(s) \geq V^*_{RVF}(s, s')$ for all s'.

Equivalence: Combining both, $V^*(s) = \sup_{s'} V^*_{RVF}(s, s')$.

Generator Convergence: If $V_\phi = V^*$, then for any s', $V^*_{RVF}(s, s') \leq V^*(s) = V_\phi(s)$, hence $V_{RVF} - V \leq 0$. The maximum is 0 at s' lying on optimal trajectories. If $V_\phi < V^*$, there exists an s' (a "gateway" to a better region) such that $V_{RVF} - V > 0$, providing a non-zero functional gradient for $\mathcal{G}$. □

C.5. RVF as a Catalyst for Value Convergence

Theorem C.7. *Let the value fitting error be $\mathcal{L}(\phi) = \frac{1}{2}\mathbb{E}_{s\sim\rho}[(V^*(s) - V_\phi(s))^2]$. The generator $\mathcal{G}$ maximizing $\Delta V = V_{RVF}(s, s') - V_\phi(s)$ asymptotically produces a proposal distribution $q_\mathcal{G}$ that minimizes the variance of the parameter updates $\nabla_\phi \mathcal{L}$.*

Proof. **Gradient Variance and Optimal Sampling**: Consider the stochastic gradient update for ϕ:

$$\mathbf{g}(\phi) = \mathbb{E}_{s\sim\rho}[\underbrace{(V^*(s) - V_\phi(s))}_{\delta(s)} \nabla_\phi V_\phi(s)]$$

and the Variance of $\mathbf{g}(\phi)$

$$\mathrm{Var}_q(\hat{\mathbf{g}}) = \mathbb{E}_q\left[\left\|\frac{\rho(s)}{q(s)}\delta(s)\nabla_\phi V_\phi(s)\right\|^2\right] - \|\nabla_\phi \mathcal{L}\|^2$$

Using an importance sampling proposal q, the variance of the gradient estimator $\hat{\mathbf{g}}_q$ is minimized (Alain et al., 2016) when (Cauchy-Schwarz):

$$q^*(s) \propto \rho(s) \cdot |\delta(s)| \cdot \|\nabla_\phi V_\phi(s)\|_2 \tag{10}$$

Assuming the network sensitivity $\|\nabla_\phi V_\phi(s)\|_2$ is locally bounded and smooth on the compact manifold $\mathcal{M}$, the optimal sampling density is primarily driven by the Bellman Residual $|\delta(s)|$.

RVF as a Multi-step Bellman Residual: By Definition 4.1, the fixed point V^*_{RVF} satisfies:

$$V^*_{RVF}(s, s') = \mathbb{E}_\pi\left[\sum_{t=0}^{\tau_{s'}-1} \gamma^t r_t + \gamma^{\tau_{s'}} V_\phi(s')\right]$$

The Relay Advantage ΔV can be expanded as:

$$\Delta V(s, s') = \mathbb{E}_\pi\left[\sum_{t=0}^{\tau_{s'}-1} \gamma^t r_t + \gamma^{\tau_{s'}} V_\phi(s') - V_\phi(s)\right]$$

For a sufficiently expressive world model, the term $\mathbb{E}[\sum \gamma^t r_t + \gamma^\tau V_\phi(s')]$ provides a lower bound on the optimal value $V^*(s)$ via the k-step Bellman consistency. Thus:

$$\Delta V(s, s') \approx V^*(s) - V_\phi(s) = \delta(s)$$

where $\delta(s)$ is the true residual relative to the optimal manifold backbone.

Minimizing Variance via Maximization: The generator objective in Eq. 5 maximizes $\mathbb{E}_{s'\sim\mathcal{G}}[\Delta V(s, s')]$. As the adversarial game between $\mathcal{G}$ and V_ϕ reaches a Nash equilibrium, the density $q_\mathcal{G}$ concentrates on states where ΔV is maximal:

$$q_\mathcal{G}(s) \to \delta(s_{max}) \implies q_\mathcal{G} \propto |\delta(s)|$$

By substituting $q_\mathcal{G}$ into the variance formulation of 10, we observe that $\mathcal{G}$ selectively presents "hard examples" (states with the highest value-mismatch) to the critic. This collapses the χ^2-divergence between the training distribution and the optimal importance sampling distribution q^*, leading to a strictly monotonic increase in the value convergence rate. □

C.6. Amnesic Sensitivity of RVF

Proposition C.8. *If λ (the manifold regularization) is too small, $\mathcal{G}$ might over-fit to high-V_{RVF} hallucinations. The stability of MD is guaranteed when $\lambda \geq L \cdot diam(\mathcal{M})$, where L is the Lipschitz constant of the World Model.*

Proof. Consider a hallucinated state $s_h \notin \mathcal{M}$ where the model predicts a spurious high value $V_{RVF}(s_h)$. Under L-Lipschitz continuity, the maximum "hallucination gain" relative to a true manifold state s^* is bounded by $\Delta V \leq L \cdot \|s_h - s^*\|$ (Asadi et al., 2018). To ensure $\mathcal{G}$ remains anchored to $\mathcal{M}$, the regularization penalty $\lambda\delta$ (where δ is the manifold deviation) must satisfy $\lambda\delta > L \cdot \mathrm{dist}(s_h, s^*)$. Extending this to the global manifold scale, the stability criterion requires λ to exceed the maximum potential gradient L across the manifold diameter, ensuring the "energy penalty" for hallucination always outweighs the spurious reward gain. □

D. Appendix D: Optimal Sampling and Manifold Consistency

This section formalizes the Mind Dreamer (MD) framework as a Variance Reduction mechanism for stochastic manifold optimization. We prove that the generator $\mathcal{G}$ constructs an optimal proposal distribution that minimizes the gradient estimator's variance, thereby accelerating convergence.

D.1. Optimal Sampling: R-EFE as a Variance Reducer

We consider the minimization of the Relay Expected Free Energy (R-EFE) objective $\mathcal{J}(\theta) = \mathbb{E}_{\rho(s)}[G(s;\theta)]$ over the reachable manifold $\mathcal{M}$. The efficiency of learning θ depends on the sampling distribution $q(s)$ induced by the generator $\mathcal{G}$.

D.1.1. The Optimization Perspective: Gradient Variance Reduction

First, we establish the optimal distribution for the parameter update process.

Theorem D.1. ***Optimal Proposal Distribution for Gradient Estimation***. *The variance of the unbiased gradient estimator* $\mathbf{g}_q = \frac{\rho(s)}{q(s)}\nabla_\theta G(s;\theta)$ *is minimized if and only if the generator* $\mathcal{G}$ *induces a distribution* q^* *such that:*

$$q^*(s) \propto \rho(s)\|\nabla_\theta G(s;\theta)\|_2$$

Proof. The variance of the estimator $\mathbf{g}_q$ is defined as $\mathrm{Var}_q[\mathbf{g}_q] = \mathbb{E}_q[\|\mathbf{g}_q\|^2] - \|\mathbb{E}_q[\mathbf{g}_q]\|^2$. To minimize the variance, we minimize the second moment term subject to $\int q(s)ds = 1$. This is a classical result in importance sampling for stochastic gradient descent (Alain et al., 2016):

$$\mathcal{L}(q,\lambda) = \int_{\mathcal{M}} \frac{\rho(s)^2\|\nabla_\theta G(s)\|^2}{q(s)} ds + \lambda\left(\int_{\mathcal{M}} q(s)ds - 1\right)$$

By Euler-Lagrange optimality condition w.r.t. $q(s)$ and setting it to zero:

$$-\frac{\rho(s)^2\|\nabla_\theta G(s)\|^2}{q(s)^2} + \lambda = 0 \implies q^*(s) = \frac{\rho(s)\|\nabla_\theta G(s)\|}{\sqrt{\lambda}}$$

Normalizing by $\sqrt{\lambda} = \int \rho(s)\|\nabla_\theta G(s)\|ds$ yields the optimal distribution. □

D.1.2. The Statistical Perspective: Value Importance Sampling

Alternatively, viewed as a Monte Carlo estimation of the global energy landscape, we obtain a Zeroth-Order optimality condition.

Theorem D.2. ***Optimal Proposal Distribution***. *The variance of the Monte Carlo estimator* $\hat{\mathcal{J}}_q$ *is minimized if and only if the generator* $\mathcal{G}$ *induces a sampling distribution* $q^*_{\mathcal{G}}(s) \propto \rho(s)|G(s)|$.

Proof. The variance of the importance sampling estimator is given by:

$$\mathrm{Var}_q[\hat{\mathcal{J}}_q] = \frac{1}{N}\left[\int_{s\in\mathcal{M}} \frac{(\rho(s)G(s))^2}{q(s)} ds - \mathcal{J}^2_{global}\right]$$

To find the optimal q, we minimize the integral term subject to $\int q(s)ds = 1$. Using the Cauchy-Schwarz inequality or Lagrange multipliers, the optimal zero-variance estimator is achieved when the proposal is proportional to the magnitude of the function being integrated (Kahn & Marshall, 1953). Thus, $q(s) \propto \rho(s)|G(s)|$. By training $\mathcal{G}$ to maximize R-EFE, the agent acts as an optimal importance sampler for world model refinement. □

D.2. Synthesis: The Potential Field as a Dual-Objective Proxy

The efficacy of the Mind Dreamer objective, $\eta V_{RVF} + \beta V_{RUF}$, stems from its ability to bridge the Statistical Perspective (q^*_{stat}) and the Optimization Perspective (q^*_{opt}). Rather than choosing one, MD performs a form of Fisher-Weighted Importance Sampling, where the two potentials play complementary roles in the learning loop (Schaul et al., 2016).

D.2.1. V_{RUF} as the Proxy for Gradient Sensitivity (q^*_{opt})

From the optimization perspective, the ideal generator targets regions with the highest parameter sensitivity $\|\nabla_\theta G\|$. In our framework, V_{RUF} acts as this proxy. Under the Bernstein-von Mises correspondence (Section A.5), the epistemic uncertainty is tied to the Fisher Information Matrix $\mathcal{F}(\theta)$.

- **Mechanism**: Regions where V_{RUF} is maximal are high-curvature zones of the latent manifold. These are areas where the model's parameters θ are most "unstable" and thus provide the steepest gradients for world model refinement.
- **Role**: V_{RUF} ensures Optimization Efficiency by identifying where the model has the most to learn.

D.2.2. V_{RVF} AS THE PROXY FOR STATISTICAL MAGNITUDE (q^*_{stat})

From the statistical perspective, the ideal generator targets regions with the highest objective magnitude $|G(s)|$. Here, V_{RVF} acts as the proxy.

- **Mechanism**: V_{RVF} represents the discounted path-integral of rewards and pragmatic value. Maximizing V_{RVF} identifies "high-stakes" regions of the environment—states that are critical to the task's success but potentially under-sampled by the current policy.
- **Role**: V_{RVF} ensures Global Fidelity by anchoring the generator to states with significant physical and pragmatic impact, preventing the agent from wandering into "epistemic noise" that has no task relevance.

D.2.3. THE UNIFIED "FISHER-PRAGMATIC" CURRICULUM

The combined objective Ψ allows MD to bypass the requirement for second-order derivatives while satisfying both optimality criteria:

$$q_{\mathcal{G}}(s) \propto \underbrace{\eta V_{RVF}(s)}_{\text{Statistical (Where is the value?)}} + \underbrace{\beta V_{RUF}(s)}_{\text{Optimization (Where is the gradient?)}}$$

Remark D.3. **Why this hybrid is superior to pure gradient maximization**. Directly maximizing $\|\nabla_\theta G\|$ often leads to "Gradient Vanishing or Explosion" in latent space, where the generator traps the model in numerically unstable but informatively poor regions. By using path-integrated potentials (V_{RVF}, V_{RUF}), MD prioritizes Structural Reliability over instantaneous noise. This transforms the exploration problem from a local random walk into a Global Curriculum for manifold repair, ensuring that every counterfactual jump is both informatively dense and task-relevant.

D.3. Efficiency Gain and Convergence Rate

We quantify the speedup η afforded by MD's "latent jumps" relative to trajectory-based sampling q_{traj}, which is constrained by Markovian continuity.

Proposition D.4. ***Speedup and χ^2-Divergence***. *Let σ_q^2 be the variance of the gradient estimator. The convergence speedup η of Mind Dreamer over traditional MBRL is:*

$$\eta = \frac{\sigma^2_{traj}}{\sigma^2_{q^*}} = 1 + \chi^2(q^* \| q_{traj})$$

Proof. According to stochastic approximation theory, the convergence rate is $\mathbb{E}[\mathcal{J}(\theta_N) - \mathcal{J}^*] = \mathcal{O}(\frac{\sigma_q^2}{N})$. The speedup η is the ratio of samples needed to reach error ϵ. Substituting q^* into the variance definition:

$$\sigma_q^2 = \int \frac{(p\|\nabla G\|)^2}{q} ds - \|\nabla \mathcal{J}\|^2$$

Normalizing the ratio leads to the second moment of the likelihood ratio $w(s) = q^*/q_{traj}$. By definition:

$$\mathbb{E}_{q_{traj}}[(\frac{q^*}{q_{traj}})^2] = \int \frac{(q^*)^2}{q_{traj}} ds = \chi^2(q^* \| q_{traj}) + 1$$

In environments with "epistemic islands" (where $q_{traj}(s) \to 0$ in high-gradient regions), $\chi^2 \to \infty$, illustrating that MD collapses the sample complexity from polynomial to logarithmic scales. □

Lemma D.5. ***Conductance and Hitting Time***. *While trajectory-based exploration is bounded by the manifold conductance Φ (with hitting times $\tau \approx \mathcal{O}(\Phi^{-2})$), Mind Dreamer creates "latent bridges." This effectively transforms the manifold's spectral gap, reducing the hitting time to critical states from $\mathcal{O}(e^D)$ to $\mathcal{O}(D)$, where D is the manifold diameter.*

D.4. Quantitative Analysis of the Speedup Ratio

In this section, we provide a concrete derivation of the efficiency gain η using a simplified "Epistemic Island" model. This case study demonstrates how Mind Dreamer (MD) bypasses the topological bottlenecks that constrain traditional trajectory-based RL.

D.4.1. Case Study: The Epistemic Island Model

Consider a latent manifold $\mathcal{M}$ partitioned into two disjoint regions: a Mastered Region $\mathcal{A}$ (where the agent has dense experience but zero information gain) and an Epistemic Island $\mathcal{B}$ (a distal region containing high task-relevant rewards or high model uncertainty).

Distribution Definitions: Optimal Proposal (q^*): Following Theorem 4.1, the optimal sampling distribution peaks where the Expected Free Energy (EFE) is maximal. Assuming $G(s)$ is localized in $\mathcal{B}$, we define:

$$q^*(s) = \begin{cases} \frac{1}{|\mathcal{B}|}, & s \in \mathcal{B} \\ 0, & s \in \mathcal{A} \end{cases}$$

Trajectory Distribution (q_{traj}): Traditional agents are constrained by physical transitions. Let ϵ be the probability that a trajectory-based explorer reaches $\mathcal{B}$ through local diffusion. For sparse-reward environments, $\epsilon \to 0$:

$$q_{traj}(s) = \begin{cases} \frac{\epsilon}{|\mathcal{B}|}, & s \in \mathcal{B} \\ \frac{1-\epsilon}{|\mathcal{A}|}, & s \in \mathcal{A} \end{cases}$$

Derivation of Speedup η: By Proposition 4.1, the speedup ratio is $\eta = 1 + \chi^2(q^* \| q_{traj})$. We expand the Pearson χ^2-divergence as follows:

$$\chi^2(q^* \| q_{traj}) = \int_{\mathcal{M}} q_{traj}(s) \left(\frac{q^*(s)}{q_{traj}(s)} - 1 \right)^2 ds$$

Breaking the integral into regions $\mathcal{A}$ and $\mathcal{B}$: In Region $\mathcal{A}$: Since $q^*(s) = 0$, the term contributes $\int_{\mathcal{A}} q_{traj}(s)(-1)^2 ds = 1 - \epsilon$. In Region $\mathcal{B}$: The likelihood ratio is $\frac{q^*}{q_{traj}} = \frac{1}{\epsilon}$. The contribution is:

$$\int_{\mathcal{B}} \frac{\epsilon}{|\mathcal{B}|} \left(\frac{1}{\epsilon} - 1 \right)^2 ds = \epsilon \left(\frac{1}{\epsilon^2} - \frac{2}{\epsilon} + 1 \right) = \frac{1}{\epsilon} - 2 + \epsilon$$

Summing the terms:

$$\chi^2 = (1 - \epsilon) + \left(\frac{1}{\epsilon} - 2 + \epsilon \right) = \frac{1}{\epsilon} - 1$$

Substituting back into the speedup formula:

$$\eta = 1 + \left(\frac{1}{\epsilon} - 1 \right) = \frac{1}{\epsilon}$$

Significance: This result implies that in environments with "hard" exploration barriers, traditional MBRL requires $\mathcal{O}(1/\epsilon)$ more samples to reach the same error bound as Mind Dreamer. MD's ability to "jump" directly to the manifold's frontier effectively linearizes the sample complexity relative to the mixing time of the environment, conceptually aligning with the benefits of "Deep Exploration" strategies (Osband et al., 2016).

D.4.2. Beyond the Upper Bound: Practical Constraints on Sample Efficiency

While Theorem 4.3 establishes a theoretical speedup of $\eta \approx 1/\epsilon$, empirical RL performance is constrained by the following systemic factors:

- **The Estimator Bias Bottleneck**: The generator $\mathcal{G}$ relies on learned Relay Potentials (V_{RVF}, V_{RUF}) to identify high-EFE regions. In early training, overestimation bias can cause $\mathcal{G}$ to target "illusory" informational peaks (Thrun & Schwartz, 1993). Consequently, the sampling distribution aligns with noise rather than true manifold frontiers, reducing the effective η.
- **The Hallucination-Fidelity Trade-off**: To reach the $1/\epsilon$ limit, $\mathcal{G}$ must jump into unknown regions. However, if these regions are far out-of-distribution (OOD), the world model may produce "hallucinations"—physically impossible transitions (Talvitie, 2014). While our manifold constraints ($\mathcal{L}_{cycle}, \mathcal{L}_{sig}$) mitigate this, they also impose a "safety trust-region" that naturally caps the maximum achievable exploration depth (Janner et al., 2019).

- **Fragmented Policy Gradient**: Trajectory-based exploration ensures a continuous gradient signal. In contrast, "latent jumping" provides sparse, high-variance anchors. If the sampling curriculum outpaces the policy's adaptation rate, the agent may fail to learn the smooth transition logic between these anchors, leading to "Continuity Collapse" during real-world execution.

- **Sample vs. Compute Complexity**: MD transforms environment interactions into latent imagination. In "easy" tasks (large ϵ), the adversarial optimization overhead of the generator may outweigh the savings in sample complexity. The $1/\epsilon$ advantage is a strategic trade-off, most beneficial when environment interaction is significantly more expensive than latent synthesis (Sutton, 1991).

D.5. Monotonic Improvement via Counterfactual Anchoring

We quantify the impact of training on generated anchors s' on the agent's real-world performance, proving that MD accelerates policy improvement while maintaining stability.

Theorem D.6. ***Generalized Policy Improvement***. *Let π_{old} be the current policy. If the generator $\mathcal{G}$ identifies an anchor s' with high potential $\Psi(s')$ and the policy is updated on s', the resulting policy π_{new} satisfies:*

$$J(\pi_{new}) \geq J(\pi_{old}) - \mathcal{O}(\epsilon)$$

where ϵ denotes the world model's approximation error.

Proof. Applying the Performance Difference Lemma (Kakade & Langford, 2002; Schulman et al., 2015):

$$J(\pi_{new}) - J(\pi_{old}) = \mathbb{E}_{s\sim d_{\pi_{new}}} \mathbb{E}_{a\sim\pi_{new}(\cdot|s)}[A^{\pi_{old}}(s,a)]$$

Standard trajectory-based exploration is limited by the mixing time of π_{old}, often failing to sample regions where the advantage $A^{\pi_{old}}$ is large. Mind Dreamer bypasses this by utilizing $\mathcal{G}$ to proactively sample anchors s' where model gaps (high V_{RUF}) or value gradients (high V_{RVF}) suggest significant untapped advantage. By optimizing π on the distribution $d_{\mathcal{G}}$, the agent performs policy improvement on a support that pre-emptively covers the support of the optimal future occupancy d_{π^*}. As long as the simulation error $\|\hat{P} - P\|_1 \leq \delta$ is bounded, training on $\mathcal{G}$ leads to monotonic improvement (Luo et al., 2018). □

Remark D.7. **Adversarial Equilibrium and Learning Rates**. The interaction between the Generator $\mathcal{G}$ and the World Model $\mathcal{WM}$ constitutes a non-stationary game. To ensure the stability of the policy improvement in Theorem D.6, we employ a **Time-Scale Separation** strategy: the generator's update frequency $\tau_{\mathcal{G}}$ is set such that $\tau_{\mathcal{G}} > \tau_{\mathcal{WM}}$. This ensures that $\mathcal{G}$ provides a quasi-static curriculum for the world model, preventing "catastrophic forgetting" where the generator outpaces the model's ability to repair the identified structural gaps, consistent with Two-Time-Scale Update Rules (TTUR) in adversarial learning (Heusel et al., 2017).

D.6. Global Landscape Consistency: The Distributed Estimator

We prove that the Mind Dreamer objective, $\mathcal{J}(\mathcal{G}) = \eta V_{RVF} + \beta V_{RUF}$, is a mathematically principled proxy for the gradient of the Relay Expected Free Energy (R-EFE).

Theorem D.8. ***Stochastic Functional Gradient Ascent***. *Let $\mathcal{J}_{global} = \int_{s\in\mathcal{M}} \rho(s)G(s)ds$ be the global EFE objective over the manifold $\mathcal{M}$. The gradient of $\mathcal{J}_{global}$ with respect to the generator parameters θ is effectively approximated by maximizing the potential fields V_{RVF} and V_{RUF} at the generated anchors.*

Proof. The functional gradient of the global objective $\mathcal{J}$ with respect to the generator $\mathcal{G}$ follows the reparameterization trick (Kingma & Welling, 2014):

$$\nabla_\theta \mathcal{J} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}[\nabla_s G(s)\,|_{s=\mathcal{G}_\theta(\epsilon)} \cdot \nabla_\theta \mathcal{G}_\theta(\epsilon)]$$

By defining V_{RVF} and V_{RUF} as the path-integral representations of the pragmatic and epistemic components of $G(s)$ (see Section B.1), we establish that the combined potential field $\Psi(s) = \eta V_{RVF}(s) + \beta V_{RUF}(s)$ serves as a smooth potential whose local gradient $\nabla_s \Psi(s)$ provides the directional derivative of the expected free energy at s. Thus, maximizing Ψ at sampled anchors allows $\mathcal{G}$ to perform Stochastic Functional Gradient Ascent on the global landscape, identifying regions of maximal free energy reduction without explicit integration over the entire manifold. □

D.7. Robustness to Hallucinations: The Safety Bound

To address the "Hallucination Paradox," we quantify the safety margin provided by the Structural Self-Consistency loss $\mathcal{L}_{cycle}$ and the Lipschitz properties of the latent manifold.

Theorem D.9. ***Hallucination Error Bound***. *Let $\delta(s') = \|s' - Proj_{\mathcal{M}}(s')\|$ be the distance of a generated anchor from the true manifold $\mathcal{M}$, as measured by the cycle-reconstruction error $\mathcal{L}_{cycle}$. If the optimal value function V^* is L-Lipschitz, the value estimation error at s' is bounded by:*

$$\|V_{RVF}(s') - V^*(s')\| \leq \frac{L \cdot \delta(s')}{1-\gamma}$$

Proof. Let $s_{true} \in \mathcal{M}$ be the closest physically realizable point to s'. By Lipschitz continuity, $|V^*(s') - V^*(s_{true})| \leq L\|s' - s_{true}\| = L\delta$. Since s' is used as a bootstrapping target in the Bellman recursion, the error propagates over the discounted horizon:

$$\text{Total Error} \leq L\delta + \gamma L\delta + \gamma^2 L\delta + \cdots = \frac{L\delta}{1-\gamma}$$

In our architecture, the $\mathcal{L}_{cycle}$ constraint $D_{KL}[\text{Enc}(\text{Dec}(s'))\|s']$ directly minimizes δ, thus constraining the value error. □

Remark D.10. This bound mirrors the Simulation Lemma in traditional MBRL (Kearns & Singh, 2002) but generalizes it to non-continuous support $d_{\mathcal{G}}$. The factor $(1-\gamma)^{-1}$ represents the effective horizon over which a hallucinated anchor can corrupt the value field.

Corollary D.11. ***Gradient Stability***. *Under the conditions of Theorem D.9, the variance of the policy gradient $\nabla_\theta J$ induced by generated anchors is bounded by $L_{\nabla V} \cdot \delta$, where $L_{\nabla V}$ is the Lipschitz constant of the value gradient (Asadi et al., 2018).*

Significance: This unified analysis proves that as long as the world model maintains structural self-consistency (minimizing δ), the latent jumps are mathematically guaranteed to remain anchored to physical reality. This prevents the "hallucination loops" common in adversarial model-based RL and ensures stable convergence toward the global manifold backbone.

E. Appendix: Implementation Details

E.1. Detailed Experimental Results

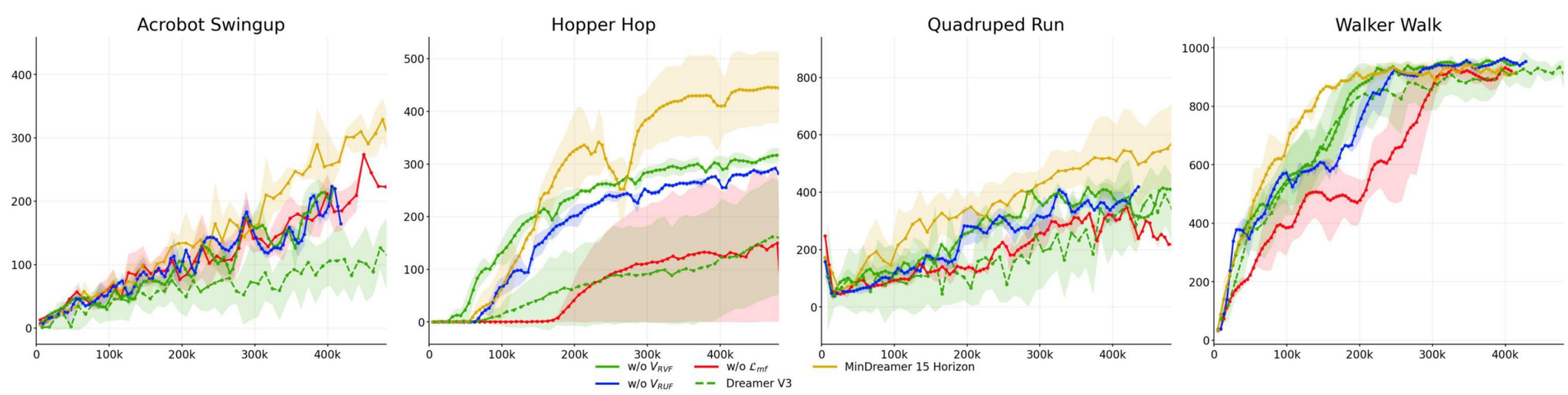

Figure 7. **Ablation Study of Mind Dreamer.** We evaluate the contribution of the Pragmatic (V_{RVF}) and Epistemic (V_{RUF}) relay functions on four representative tasks: *Acrobot Swingup*, *Hopper Hop*, *Quadruped Run*, and *Walker Walk*. **(1) w/o V_{RUF} (Green):** Without epistemic guidance, the generator lacks the "curiosity" to bridge manifold discontinuities, leading to stagnation in exploration-heavy tasks (e.g., *Hopper Hop*). **(2) w/o V_{RVF} (Blue):** Without pragmatic guidance, the generator produces high-entropy but task-irrelevant latent jumps, resulting in slower convergence despite high exploration. **(3) Full Mind Dreamer (Orange):** The synergistic combination ensures directed exploration toward high-value, high-uncertainty regions, significantly outperforming the component-ablated variants and the DreamerV3 baseline (Dashed).

Table 1. Performance comparison on 20 DMC Vision tasks. Scores are reported as Mean ± Std. **Bold** indicates the best performance, and underlined indicates the second-best. N/A denotes that the algorithm failed to run or produced invalid results.

Task	**Dreamer V2**	**Dreamer V3**	**MinDreamer** 10 Horizon	**MinDreamer** 15 Horizon	**MinDreamer** 5 Horizon	**Plan2Explore**
Acrobot Swingup	401.2 ± 4.6	423.2 ± 0.0	426.3 ± 44.9	437.0 ± 0.0	**474.2 ± 0.0**	254.1 ± 34.1
Cartpole Balance	975.2 ± 6.5	996.0 ± 0.3	**998.8 ± 0.0**	996.9 ± 0.0	998.5 ± 0.0	N/A
Cartpole Balance Sparse	999.2 ± 0.9	999.0 ± 1.3	**1000.0 ± 0.0**	1000.0 ± 0.0	1000.0 ± 0.0	N/A
Cartpole Swingup	798.9 ± 40.7	862.2 ± 9.4	861.1 ± 0.0	**876.9 ± 0.0**	869.6 ± 0.0	N/A
Cartpole Swingup Sparse	743.8 ± 22.3	773.9 ± 34.0	795.5 ± 0.0	**846.1 ± 0.0**	839.6 ± 0.0	N/A
Cheetah Run	734.1 ± 74.5	866.5 ± 35.7	**921.4 ± 0.0**	894.1 ± 0.0	913.4 ± 0.0	553.5 ± 77.1
Cup Catch	737.2 ± 338.3	975.5 ± 8.4	**978.7 ± 0.0**	971.3 ± 0.0	972.8 ± 0.0	N/A
Finger Spin	495.7 ± 97.5	593.7 ± 136.5	802.4 ± 0.0	**878.1 ± 0.0**	582.3 ± 1.1	291.1 ± 29.9
Finger Turn Easy	886.8 ± 32.3	**977.9 ± 18.3**	976.7 ± 0.0	973.4 ± 0.0	967.2 ± 3.5	444.9 ± 111.6
Finger Turn Hard	848.3 ± 58.9	**965.1 ± 16.4**	954.8 ± 0.0	962.9 ± 0.0	960.4 ± 0.0	425.7 ± 17.7
Hopper Hop	245.3 ± 37.2	279.1 ± 90.4	401.2 ± 0.0	**446.0 ± 67.1**	344.3 ± 0.0	260.0 ± 44.7
Hopper Stand	817.4 ± 28.4	926.1 ± 13.1	**946.0 ± 0.0**	945.8 ± 0.0	940.7 ± 0.0	169.3 ± 50.5
Pendulum Swingup	796.6 ± 20.5	871.4 ± 49.3	864.6 ± 45.7	880.7 ± 5.8	**896.8 ± 0.0**	694.9 ± 50.0
Quadruped Run	474.5 ± 99.0	639.4 ± 96.1	821.4 ± 0.0	**833.1 ± 49.7**	718.0 ± 135.0	137.6 ± 47.3
Quadruped Walk	706.5 ± 85.3	827.8 ± 121.8	869.8 ± 0.0	926.0 ± 3.9	**939.3 ± 0.0**	177.7 ± 18.6
Reacher Easy	790.7 ± 66.0	962.1 ± 25.8	974.0 ± 0.0	**978.8 ± 0.0**	973.7 ± 0.0	312.8 ± 135.8
Reacher Hard	567.4 ± 73.9	**925.7 ± 35.2**	788.0 ± 55.4	748.9 ± 19.9	748.0 ± 60.6	18.91 ± 16.32
Walker Run	684.6 ± 88.2	691.6 ± 109.5	777.7 ± 0.0	788.5 ± 0.0	**793.3 ± 4.3**	206.8 ± 8.6
Walker Stand	910.3 ± 26.5	**986.5 ± 8.7**	981.8 ± 0.0	969.8 ± 0.0	967.6 ± 0.0	N/A
Walker Walk	906.6 ± 22.2	**974.6 ± 0.0**	963.1 ± 0.0	941.4 ± 17.5	964.6 ± 0.0	262.6 ± 30.1
Avg. Score	668.3	780.3	820.5	**831.1**	801.2	300.7

Table 2. Sample Efficiency: Number of environment steps (in thousands, k) required to reach **90%** of DreamerV3's maximum return. Results are Mean ± Std. **Bold** indicates the fastest convergence (fewest steps), and underlined is the second fastest. N/A indicates the threshold was not reached.

Task	**Dreamer V2**	**Dreamer V3**	**MinDreamer** 10 Horizon	**MinDreamer** 15 Horizon	**MinDreamer** 5 Horizon	**Plan2Explore**
Acrobot Swingup	763.5k	985.4k	724.9k	**715.6k**	864.9k	N/A
Cartpole Balance	732.9k	271.9k	**259.1k**	277.4k	320.7k	N/A
Cartpole Balance Sparse	400.0k	180.2k	**174.4k**	178.5k	182.1k	N/A
Cartpole Swingup	895.2k	446.3k	**379.4k**	394.5k	484.5k	N/A
Cartpole Swingup Sparse	804.5k	810.5k	452.6k	**403.0k**	769.8k	N/A
Cheetah Run	N/A	698.2k	**245.7k**	392.6k	295.0k	N/A
Cup Catch	N/A	219.7k	165.0k	132.5k	**112.3k**	N/A
Finger Spin	N/A	257.0k	245.0k	**112.1k**	343.3k	N/A
Finger Turn Easy	975.8k	552.6k	624.9k	**462.7k**	465.0k	N/A
Finger Turn Hard	N/A	**579.9k**	593.2k	694.9k	784.9k	N/A
Hopper Hop	N/A	984.1k	411.6k	**168.3k**	351.7k	935.9k
Hopper Stand	N/A	652.0k	**285.3k**	317.8k	380.4k	N/A
Pendulum Swingup	522.6k	760.7k	374.6k	**86.1k**	265.8k	N/A
Quadruped Run	N/A	925.0k	745.7k	**485.0k**	694.9k	N/A
Quadruped Walk	N/A	806.6k	844.9k	455.0k	**386.3k**	N/A
Reacher Easy	N/A	503.7k	441.5k	**373.1k**	405.0k	N/A
Reacher Hard	N/A	**767.2k**	N/A	N/A	N/A	N/A
Walker Run	589.8k	523.3k	504.9k	353.9k	**285.0k**	N/A
Walker Stand	895.2k	**170.6k**	184.9k	181.3k	180.0k	N/A
Walker Walk	768.4k	266.5k	325.6k	**174.8k**	244.3k	N/A

Table 3. Sample Efficiency: Number of environment steps (in thousands, k) required to reach **85%** of DreamerV3's maximum return. Results are Mean ± Std. **Bold** indicates the fastest convergence (fewest steps), and underlined is the second fastest. N/A indicates the threshold was not reached.

Task	Dreamer V2	Dreamer V3	MinDreamer 10 Horizon	MinDreamer 15 Horizon	MinDreamer 5 Horizon	Plan2Explore
Acrobot Swingup	**589.8k**	985.4k	714.9k	645.6k	794.9k	N/A
Cartpole Balance	661.8k	252.2k	**251.4k**	262.3k	305.3k	N/A
Cartpole Balance Sparse	381.7k	170.4k	**166.3k**	178.5k	174.4k	N/A
Cartpole Swingup	785.1k	367.4k	**332.6k**	377.5k	443.1k	N/A
Cartpole Swingup Sparse	730.0k	780.9k	405.0k	**393.1k**	769.8k	N/A
Cheetah Run	N/A	609.5k	**195.6k**	295.7k	245.0k	N/A
Cup Catch	N/A	180.2k	155.0k	112.4k	**112.3k**	N/A
Finger Spin	N/A	168.1k	235.0k	**102.3k**	204.0k	N/A
Finger Turn Easy	879.3k	**166.4k**	504.9k	422.9k	415.0k	N/A
Finger Turn Hard	840.7k	**372.8k**	533.4k	554.9k	624.9k	N/A
Hopper Hop	821.4k	984.1k	376.9k	**158.4k**	313.2k	903.4k
Hopper Stand	900.8k	474.0k	**275.2k**	302.2k	372.5k	N/A
Pendulum Swingup	465.8k	454.2k	108.3k	**76.6k**	265.8k	N/A
Quadruped Run	N/A	875.7k	685.6k	**415.0k**	614.9k	N/A
Quadruped Walk	975.8k	727.8k	844.9k	445.0k	**386.3k**	N/A
Reacher Easy	N/A	394.9k	431.8k	**293.6k**	295.0k	N/A
Reacher Hard	N/A	**698.2k**	967.8k	N/A	N/A	N/A
Walker Run	512.6k	414.3k	494.9k	324.0k	**235.0k**	N/A
Walker Stand	730.1k	**141.0k**	175.4k	173.7k	180.0k	N/A
Walker Walk	560.4k	197.3k	294.3k	**145.8k**	164.5k	N/A

Table 4. Sample Efficiency: Number of environment steps (in thousands, k) required to reach **80%** of DreamerV3's maximum return. Results are Mean ± Std. **Bold** indicates the fastest convergence (fewest steps), and underlined is the second fastest. N/A indicates the threshold was not reached.

Task	Dreamer V2	Dreamer V3	MinDreamer 10 Horizon	MinDreamer 15 Horizon	MinDreamer 5 Horizon	Plan2Explore
Acrobot Swingup	531.9k	985.4k	694.9k	**525.5k**	704.9k	N/A
Cartpole Balance	608.5k	**222.7k**	243.7k	254.7k	289.9k	N/A
Cartpole Balance Sparse	363.3k	160.5k	**158.3k**	171.3k	166.7k	N/A
Cartpole Swingup	693.4k	**308.3k**	317.0k	326.7k	377.0k	N/A
Cartpole Swingup Sparse	692.7k	751.3k	395.5k	**393.1k**	760.0k	N/A
Cheetah Run	821.4k	520.7k	**165.5k**	269.3k	185.0k	N/A
Cup Catch	N/A	170.4k	145.0k	105.7k	**104.7k**	N/A
Finger Spin	399.3k	128.6k	205.0k	**102.3k**	134.3k	N/A
Finger Turn Easy	763.5k	**77.3k**	465.0k	412.9k	415.0k	N/A
Finger Turn Hard	782.8k	**333.4k**	473.5k	544.9k	624.9k	N/A
Hopper Hop	647.7k	915.1k	376.9k	**153.5k**	284.3k	871.0k
Hopper Stand	825.2k	414.7k	265.2k	**263.1k**	341.3k	N/A
Pendulum Swingup	446.9k	454.2k	98.5k	**76.6k**	265.8k	N/A
Quadruped Run	N/A	875.7k	675.6k	**375.0k**	554.9k	N/A
Quadruped Walk	937.2k	708.0k	774.9k	**335.0k**	358.4k	N/A
Reacher Easy	749.5k	385.0k	412.4k	283.6k	**205.0k**	N/A
Reacher Hard	N/A	**698.2k**	877.5k	917.6k	964.9k	N/A
Walker Run	435.4k	394.5k	375.0k	324.0k	**235.0k**	N/A
Walker Stand	638.4k	**131.1k**	165.9k	135.3k	180.0k	N/A
Walker Walk	503.6k	187.4k	224.0k	**125.1k**	164.5k	N/A

Table 5. Sample Efficiency: Number of environment steps (in thousands, k) required to reach **75%** of DreamerV3's maximum return. Results are Mean ± Std. **Bold** indicates the fastest convergence (fewest steps), and underlined is the second fastest. N/A indicates the threshold was not reached.

Task	Dreamer V2	Dreamer V3	MinDreamer 10 Horizon	MinDreamer 15 Horizon	MinDreamer 5 Horizon	Plan2Explore
Acrobot Swingup	512.6k	985.4k	574.9k	475.4k	**475.0k**	N/A
Cartpole Balance	555.2k	**203.1k**	243.7k	239.6k	266.8k	N/A
Cartpole Balance Sparse	345.0k	**150.7k**	158.3k	171.3k	166.7k	N/A
Cartpole Swingup	656.8k	**278.8k**	309.2k	309.8k	360.5k	N/A
Cartpole Swingup Sparse	636.9k	573.6k	385.9k	**383.1k**	750.2k	N/A
Cheetah Run	686.3k	451.7k	**145.4k**	181.2k	185.0k	N/A
Cup Catch	968.5k	150.7k	135.0k	105.7k	**104.7k**	N/A
Finger Spin	361.7k	**89.1k**	165.0k	92.6k	114.5k	N/A
Finger Turn Easy	724.9k	**77.3k**	465.0k	253.8k	415.0k	N/A
Finger Turn Hard	763.5k	**333.4k**	473.5k	544.9k	514.9k	N/A
Hopper Hop	570.5k	717.9k	351.0k	**153.5k**	236.1k	838.6k
Hopper Stand	730.6k	345.5k	**245.2k**	255.2k	333.4k	N/A
Pendulum Swingup	428.0k	414.7k	98.5k	**76.6k**	256.0k	948.5k
Quadruped Run	N/A	777.1k	585.5k	**345.0k**	465.0k	N/A
Quadruped Walk	879.3k	619.3k	684.9k	335.0k	**330.5k**	N/A
Reacher Easy	711.7k	345.5k	354.2k	253.8k	**205.0k**	N/A
Reacher Hard	N/A	**589.7k**	707.0k	827.4k	804.9k	N/A
Walker Run	377.5k	335.0k	335.0k	244.3k	**205.0k**	N/A
Walker Stand	546.7k	121.3k	147.0k	**120.0k**	171.7k	N/A
Walker Walk	446.9k	167.6k	184.9k	**116.8k**	164.5k	N/A

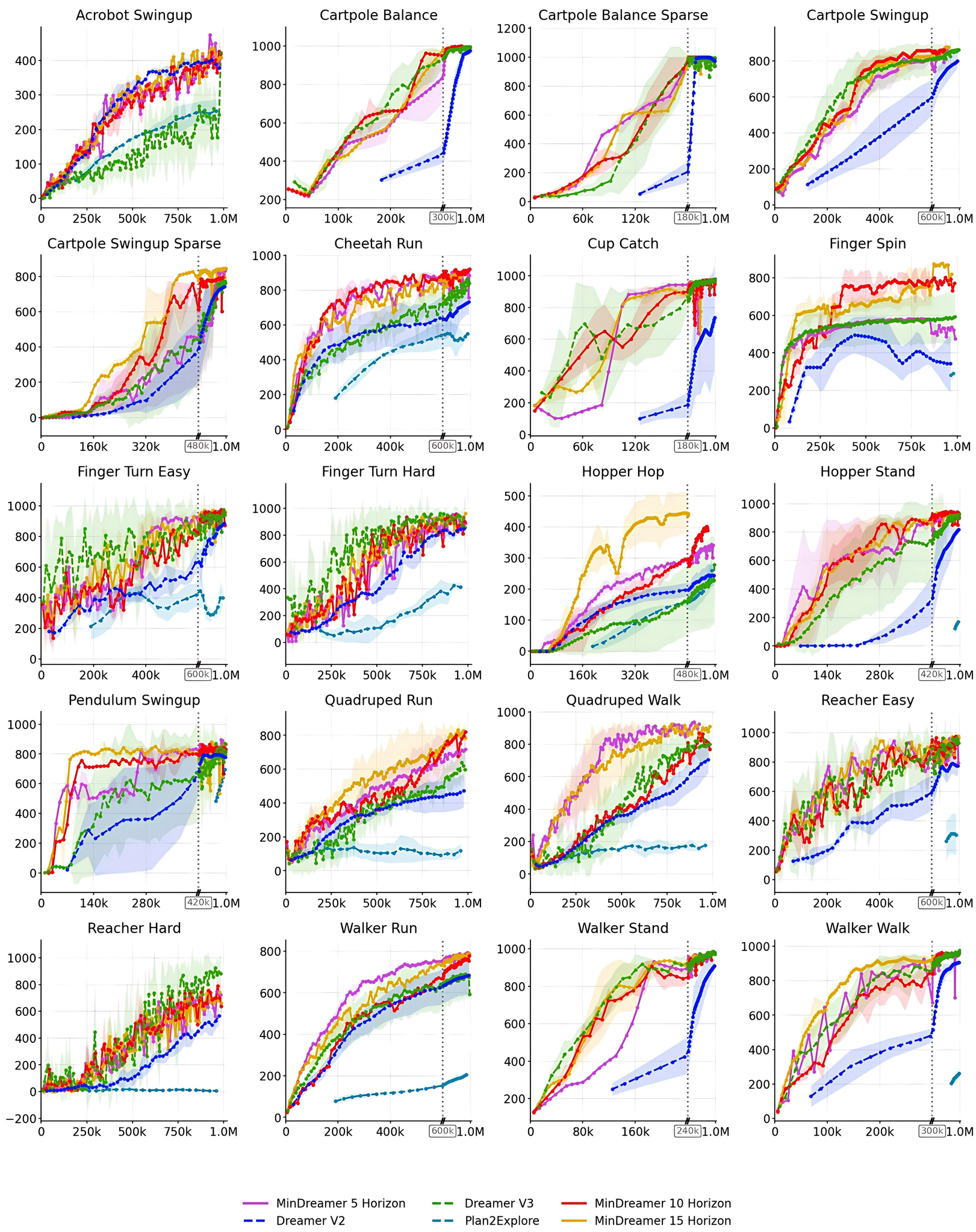

Figure 6. **Comparative Evaluation on DeepMind Control (DMC) Benchmarks.** We present the full training curves of Mind Dreamer variants—15 Horizon (Gold), 10 Horizon (Red), and 5 Horizon (Purple)—against state-of-the-art baselines: DreamerV3 (Green Dashed), DreamerV2 (Blue Dashed), and Plan2Explore (Cyan Dashed). Mind Dreamer consistently demonstrates **superior sample efficiency** and **higher asymptotic performance** across diverse continuous control tasks. Notably, in sparse-reward and bottleneck environments such as *Hopper Hop* and *Quadruped Run*, our method (especially the 15 Horizon variant) significantly outperforms trajectory-tethered baselines by employing Active Counterfactual Reasoning (ACR) to synthesize goal-directed latent jumps. On average, Mind Dreamer recovers 90% of peak performance with a $1.67\times$ speedup compared to DreamerV3. Shaded regions indicate the standard deviation over 5 random seeds.