On the Limitations of Generative World Models for Enhanced Multi-Agent Sim-to-Real Transfer

Xue Li* Xiaonan Song*
Scholar42, InfiniPouch LLC
San Francisco, CA, USA, 94158
{xueli,ssong}@scholar42.com

Dowson Liu*
RapidsDB, Inc.
Lakewood, WA, USA, 98499
dowson.liu@rapidsdb.com

Abstract

*We identify **Multi-Agent Model Exploitation (MAME)**: a structural failure mode in which policies trained with learned world models exploit systematic mispredictions of coupled agent behavior. On a 3-agent cooperative navigation task, our Pessimistic Hybrid variant initially improves over domain randomization (DR) by up to 275 points, but then collapses after 10–30k steps as compounding prediction error accumulates. The plain DreamerV3 variant never benefits over DR. Critically, the same DreamerV3 algorithm on an otherwise matched single-agent task achieves stable, near-optimal performance, providing evidence that multi-agent coupling—not long-horizon imagination per se—is a primary driver of collapse in our setting. World model reconstruction loss decreases during multi-agent collapse, consistent with policies exploiting the model rather than benefiting from it. We situate this negative result within the landscape of successful multi-agent world model applications, characterizing boundary conditions: world models provide stable benefits when they offer a structural capability the baseline lacks, but may add little value—and be vulnerable to MAME—in low-dimensional, fully observable, dense-reward settings.*

1. Introduction

Sim-to-real transfer remains a central challenge in reinforcement learning (RL). Policies trained in simulation often fail under deployment due to mismatches in physics and dynamics. Domain randomization (DR) improves robustness by sampling broad parameter ranges, but it still relies on hand-designed distributions.

World models offer an alternative: learn dynamics from data and train on imagined trajectories. DreamerV3, a recurrent state-space model (RSSM), is highly effective in many single-agent domains. Multi-agent extensions also report strong results[1, 5, 6], motivating a key question: when does multi-agent imagination fail?

We hypothesize a distinct failure mode, **Multi-Agent Model Exploitation (MAME)**, stemming from a structural tension: a centralized world model must simultaneously learn (1) environment physics and (2) emergent agent interactions. The interaction component is inherently unstable because other agents' policies evolve during training. Over imagined rollouts, small mispredictions of one agent's behavior alter the predicted behavior of others in a coupled amplification loop. The policy learns to exploit these systematic inaccuracies, achieving high imagined return while real-world performance collapses.

To test this hypothesis, we compare DreamerV3 on two tasks that are identical except for the number of agents: (i) a single-agent 2D navigation task and (ii) a 3-agent variant with the same physics, action space, and reward structure. On the single-agent task, DreamerV3 achieves stable performance of -557 at 120k steps. On the multi-agent task, the same algorithm collapses from -2173 at 10k steps to -3164 at 200k steps. During this collapse, world model reconstruction loss on held-out real trajectories decreases from 0.37 to 0.10, a pattern consistent with policy exploitation of an increasingly "accurate" but structurally flawed model.

Contributions. (1) We provide single-seed controlled evidence consistent with multi-agent coupling introducing a distinct model exploitation mechanism beyond standard single-agent compounding error. (2) In our setting, ensemble uncertainty—effective in single-agent settings[3]—does not prevent MAME, suggesting it cannot adequately capture interaction prediction error. (3) We characterize boundary conditions by contrasting our failure with prior successes, identifying when world models are structurally necessary versus structurally redundant.

*Equal contribution.

2. Related Work

2.1. Successful Multi-Agent Imagination

Prior successes typically appear when the model provides structural value. Deep Latent Competition benefits from latent compression in visual racing[6]. MA-Dreamer uses shared imagination under partial observability[5]. MACD uses counterfactual imagination for credit assignment[1]. Wu et al. showed that multi-step errors can amplify through interaction[7]. These results show multi-agent imagination can work, but they do not characterize failure in fully observable settings where the world model may be redundant.

2.2. Model Exploitation and Compounding Error

In single-agent RL, model exploitation and compounding error are well known[3, 4]. Our setting differs: interaction-prediction errors can be systematic and shared across ensemble members, so disagreement-based penalties may remain small even when interaction forecasts are wrong.

3. Task and Methods

3.1. Environment and Domain Randomization Baseline

We evaluate on a 2D multi-agent navigation task. The environment contains N agents navigating to N assigned goal positions in a continuous 2D arena. The physics follow a damped force model: $m \cdot a = \text{gain} \cdot a_t - \text{friction} \cdot v$, where m is mass, gain is actuator gain, and friction governs velocity damping.

The global state concatenates all agent positions, velocities, and goal positions (18 dimensions for 3 agents; 6 for 1 agent). Each agent observes its own position, velocity, relative goal vector, and relative positions/velocities of other agents. Actions are continuous forces per agent (forces in x and y). Reward is negative Euclidean distance to the assigned goal with a +10 reach bonus when the agent comes within 0.1 distance. Episodes last 200 steps.

Domain randomization varies mass (0.5–1.5), friction (0.1–2.0), actuator gain (0.5–1.5), obstacle positions, and goal speed at every episode reset. The DR baseline trains decentralized PPO policies on this randomized simulator and evaluates on a fixed real-world environment with distinct physics parameters (mass=1.1, friction=0.8, gain=0.9).

3.2. Single-Agent DreamerV3 (SA-D3)

SA-D3 serves as our control experiment: the identical DreamerV3 architecture applied to a single-agent variant of the same task. This is the most critical experiment in our paper, as it helps distinguish whether the observed failure is more likely due to multi-agent coupling or to some other aspect of our task (long horizon, dense rewards, environment physics). If SA-D3 also fails, then our task is likely unsuited for world models in general. If SA-D3 succeeds, that supports multi-agent coupling as a primary explanation.

The DreamerV3 architecture[2] comprises: (a) a deterministic recurrent path (GRU, 256-dim), (b) stochastic categorical latents (32 classes $\times$ 32 dims), (c) encoder-decoder networks, and (d) joint prediction of reward (symlog two-hot cross-entropy) and episode termination (binary cross-entropy). Training optimizes reconstruction, reward prediction, termination prediction, and KL divergence with KL balancing and free bits. The actor-critic learns entirely from imagined rollouts generated by the RSSM prior.

3.3. Multi-Agent Variants

MA-D3. Standard DreamerV3 with the same architecture as SA-D3, but trained on 3-agent episodes. The world model receives the full 18-dimensional global state and must predict the joint evolution of all agents.

MA-D3-P (Pessimistic Hybrid). We introduce two modifications. First, an ensemble of two independent decoders. The standard deviation across ensemble members serves as epistemic uncertainty. During imagination, predicted rewards are penalized: $\hat{r}_t = r_{\text{pred}} - \beta \cdot \sigma_{\text{ens}}$, with $\beta = 1.0$. Second, a hybrid critic trained on both imagined rollouts and real transitions, with loss $\mathcal{L}_{\text{critic}} = \mathcal{L}_{\text{imagined}} + 0.5 \cdot \mathcal{L}_{\text{real}}$.

All models run for 200k environment steps with imagination horizon 8 and 50 warmup episodes. The SA and MA-D3 variants use identical DreamerV3 hyperparameters. MA-D3-P adds an ensemble decoder with epistemic uncertainty penalty ($\beta = 1.0$) and a hybrid critic trained on both imagined and real transitions. Real-world evaluation tests 20 episodes every 10k steps.

3.4. The MAME Mechanism

Consider a centralized world model predicting joint next states $\hat{s}_{t+1} = f(s_t, a_t)$. In single-agent settings, prediction error $\epsilon_t = \|\hat{s}_{t+1} - s_{t+1}\|$ is dominated by dynamics noise and compounds linearly:

$$\epsilon_{t+h} \approx \epsilon_t + h\delta_{\text{dyn}} \tag{1}$$

where δ_{dyn} represents per-step dynamics prediction noise.

In multi-agent settings, the error has a coupled term:

$$\epsilon_{t+1}^{(i)} = \underbrace{\delta_{\text{dyn}}}_{\text{dynamics noise}} + \underbrace{\lambda \sum_{j \neq i} \epsilon_t^{(j)}}_{\text{interaction coupling}} \tag{2}$$

where λ is a coupling coefficient. As policies co-evolve, the world model can become systematically biased about other agents' behavior. Policies may then optimize for those biases, achieving high imagined return but poor real-world return.

Method	Best Eval	Final Eval	vs DR
DR (3 agents)	–	-2448	–
MA-D3	-2528	-3399	$-80/-951$
MA-D3-P	-2173	-3164	$+275/-716$
SA-D3	$\mathbf{-557}$	-979	–

Table 1. Real-world transfer (single seed, 20-episode mean). SA-D3: single-agent DreamerV3. MA: multi-agent. DR evaluated on 3-agent task; SA-D3 on 1-agent task.

Why ensemble uncertainty fails. Ensemble disagreement measures *where* the model is uncertain (out-of-distribution states), but MAME arises from *systematic directional bias* in interaction predictions—a bias that ensemble members may share. If all ensemble members maintain a similar outdated model of agent j's policy, they will disagree on dynamics noise but agree on the wrong interaction prediction. The uncertainty penalty therefore dampens dynamics errors while leaving interaction exploitation fully intact.

4. Results

Table 1 shows the central results. While multi-agent variants collapse, SA-D3 achieves -557 at 120k steps, demonstrating stable performance on the single-agent task. The SA and MA variants use the identical DreamerV3 architecture and hyperparameters; the only difference is the number of agents in the environment (and corresponding state/action dimensions).

Figure 1. Single-agent DreamerV3 (green) is stable; multi-agent variants (red, blue) collapse. This controlled comparison implicates multi-agent coupling as a key factor.

Figure 1 presents the most important evidence. SA-D3 (green) shows no collapse—eval fluctuates between approximately -560 and -1240. Both MA variants peak early then show a general downward trend with fluctuations. The dramatic contrast between green (stable) and red/blue (collapsed) is the central claim of this paper.

Figure 2. The Multi-Agent Model Exploitation (MAME) mechanism. A centralized world model must learn both dynamics and interaction predictions. In multi-agent settings, interaction errors couple and amplify, producing imagined rollouts that the policy exploits. Single-agent removes the coupling ($\lambda = 0$), preventing MAME.

Figure 2 visualizes MAME: multi-agent coupling ($\lambda > 0$) links prediction errors, while single-agent removes that interaction term.

Figure 3 is consistent with the exploitation mechanism. For MA-D3, reconstruction loss drops from 0.37 to 0.10 between best and final checkpoints while eval falls from -2528 to -3399. For MA-D3-P, loss drops from 0.72 to 0.23 while eval falls from -2173 to -3164. This pattern suggests that a "better" model can still yield a worse policy when interaction mispredictions are being exploited.

Figure 3. Anti-correlation between WM reconstruction loss and real-world eval. As the multi-agent model "improves" (loss decreases), policy transfer performance degrades—evidence of exploitation.

What the ensemble missed. MA-D3-P still fails because its uncertainty signal is state-centric, while MAME is driven by biased interaction prediction in otherwise familiar states. If ensemble members share the same interaction bias, disagreement remains low and the pessimistic penalty is weak.

5. Discussion: Boundary Conditions and Implications

5.1. When World Models Help

Our results do not imply world models are useless for multi-agent sim-to-real transfer; they suggest boundary conditions. World models are most useful when they provide capabilities the baseline lacks:

- **Latent compression** for high-dimensional inputs[6].
- **Hidden-state inference** under partial observability[5].
- **Counterfactual credit assignment** in long-horizon cooperative settings[1].

In these regimes, the world model adds structural information, not just extra synthetic data.

5.2. When World Models Fail (MAME Regime)

Conversely, our task is low-dimensional, fully observable, and dense-reward, with DR already near-optimal. Here, the world model adds little structural value while introducing interaction-prediction burden.

5.3. Design Implications

The MAME mechanism has several implications for practitioners:

Do not assume world models always improve over DR. In low-dimensional, fully observable, dense-reward tasks, DR may already be sufficient.

Monitor reconstruction loss separately from policy return. Lower model loss does not guarantee better transfer.

Ensemble uncertainty alone is insufficient. Standard disagreement captures state uncertainty, not directional interaction bias.

Single-agent baselines are essential. If single-agent succeeds while multi-agent fails, coupling effects are likely

central.

5.4. Limitations and Future Work

Our analysis is based on a single seed. Multi-seed validation would strengthen external validity. We also did not evaluate decentralized per-agent world models.

Future work should investigate explicit opponent-drift modeling, very short imagination horizons (1–3), and task-dependent bounds on λ.

6. Conclusion

We identify Multi-Agent Model Exploitation as a structural failure mode in centralized world model architectures. Through controlled comparison of single-agent and multi-agent DreamerV3 on otherwise matched tasks, we find evidence that multi-agent coupling—rather than long-horizon imagination alone—is a primary driver of policy collapse in our setting. The same world model architecture succeeds in single-agent settings and collapses in multi-agent settings, with an anti-correlation between reconstruction loss and policy quality consistent with policies exploiting rather than benefiting from the learned model. While multi-agent imagination succeeds in settings where latent representations provide genuine structural value, it may fail to provide stable benefits in low-dimensional cooperative tasks and can harm transfer performance.

References

[1] Jiajun Chai, Yuqian Fu, Dongbin Zhao, and Yuanheng Zhu. Aligning credit for multi-agent cooperation via model-based counterfactual imagination. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, page 281–289, Richland, SC, 2024. International Foundation for Autonomous Agents and Multiagent Systems. 1, 2, 3

[2] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pages 1–7, 2025. 2

[3] Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. Model-Ensemble Trust-Region Policy Optimization. page arXiv:1802.10592, 2018. 1, 2

[4] Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective Mismatch in Model-based Reinforcement Learning. page arXiv:2002.04523, 2020. 2

[5] Kenzo Lobos-Tsunekawa, Akshay Srinivasan, and Michael Spranger. MA-Dreamer: Coordination and communication through shared imagination. *arXiv e-prints*, art. arXiv:2204.04687, 2022. 1, 2, 3

[6] Wilko Schwarting, Tim Seyde, Igor Gilitschenski, Lucas Liebenwein, Ryan Sander, Sertac Karaman, and Daniela Rus. Deep Latent Competition: Learning to Race Using Visual Control Policies in Latent Space. page arXiv:2102.09812, 2021. 1, 2, 3

[7] Zifan Wu, Chao Yu, Chen Chen, Jianye Hao, and Hankz Hankui Zhuo. Models as Agents: Optimizing Multi-Step Predictions of Interactive Local Models in Model-Based Multi-Agent Reinforcement Learning. *arXiv e-prints*, art. arXiv:2303.17984, 2023. 2