NashDreamer

Model-Based Reinforcement Learning for Zero-Sum Imperfect-Information Games

Tomáš Holeček **Viliam Lisý**

Artificial Intelligence Center, FEE, Czech Technical University in Prague

NeurIPS 2026

Two-Player Zero-Sum Imperfect-Information Environments

Two players **against each other**: one player's gain is the other's loss
Neither of them **sees everything**

Recent progress: reinforcement learning, learning only from interaction with a simulator, can be **game-theoretically sound** in these games

R-NaD, MMD, ...

But: RL often needs **abundant simulator access** $\Rightarrow$ problematic when the simulator is **expensive**

Model-Based RL: Learn the Simulator Too



- Learn a model of the environment, then plan in it (MuZero) or train in imagination (Dreamer)
- Strong results in **single-agent** domains
- Two-player zero-sum imperfect-information games: underexplored

Can Dreamer-style imagination be made **game-theoretically sound**?

Goal: same policy optimizer, **fewer real environment interactions**

Why Not an Independent World Model per Player?

e.g. Perturbed Rock-Paper-Scissors

	R	P	S
R	0	-2	$+1$
P	$+2$	0	-2
S	-1	$+2$	0

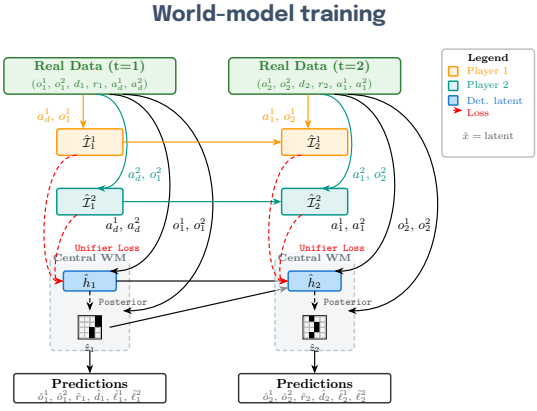
against the Nash opponent:

$$\hat{r}(\mathrm{R}) = \hat{r}(\mathrm{P}) = \hat{r}(\mathrm{S}) = 0$$

- Straightforward: an **independent** MBRL instance for each player
- Problematic: the **opponent's actions** now also dictate what happens
- Against a Nash opponent, all actions in the support of our own Nash have identical payoffs $\Rightarrow$ we find our Nash only by **getting lucky**

Without **centralization**, we might not be able to learn the **actual rules** of the game

NashDreamer: Centralized Model, Individual Views

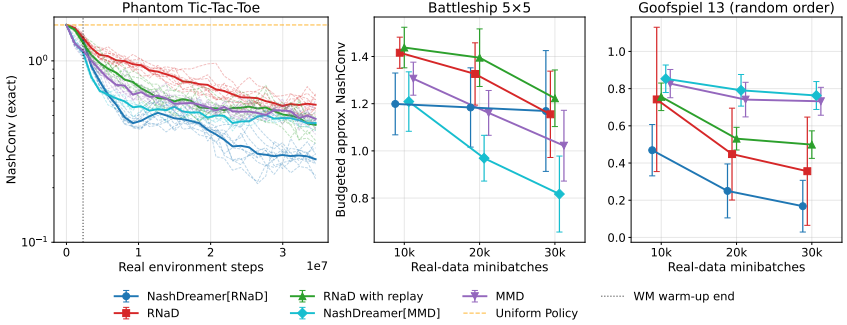


- **Centralized world model**: sees the actions and observations of **both** players ⇒ can learn the rules independent of the opponent's strategy
- **Infoset models**: each player's **individual view**, built only from their own observations and actions
- Actors act on their own view, trained in imagination with any game-theoretic policy gradient: NashDreamer[RNaD], NashDreamer[MMD]

The world model architecture otherwise largely follows **DreamerV3**, with smaller technical changes (see the paper)

Results: Strong Policies from Fewer Real Interactions

Proof-of-concept validation on benchmark games. Shown here: the larger games.



exact NashConv

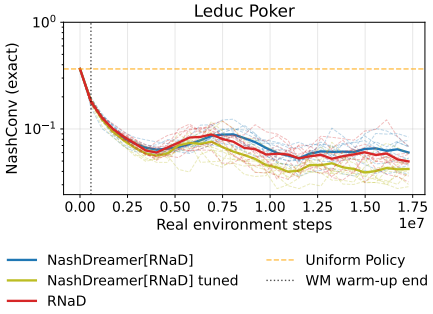
NashConv approximated by a PPO best response against the fixed policy

NashDreamer learns **strong policies** from **fewer real interactions**, mostly in deterministic games

DreamerV3 Is Vulnerable to Posterior Collapse

Finding. In the DreamerV3 architecture, the "perfect" model can be *globally* suboptimal: a model with **collapsed posterior** attains a lower loss

- Many chance outcomes with small effect on rewards and observations ⇒ ignoring chance is cheaper
- In games, chance is often **strategically decisive**
- **Leduc Poker**: the model cannot disambiguate all deals. With default coefficients, NashDreamer[RNaD] ends above R-NaD
- Tuning the DreamerV3 KL-balancing coefficients can still help in practice to recover performance



Open challenge: posterior collapse of KL balancing in stochastic games

Thank you for your attention!







Code `github.com/aicenter/nash_dreamer`

Questions? `holecto6@fel.cvut.cz`

Supported by the Czech Science Foundation (GA25-18353S). Computational resources provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.