Better Decisions through the Right Causal World Model

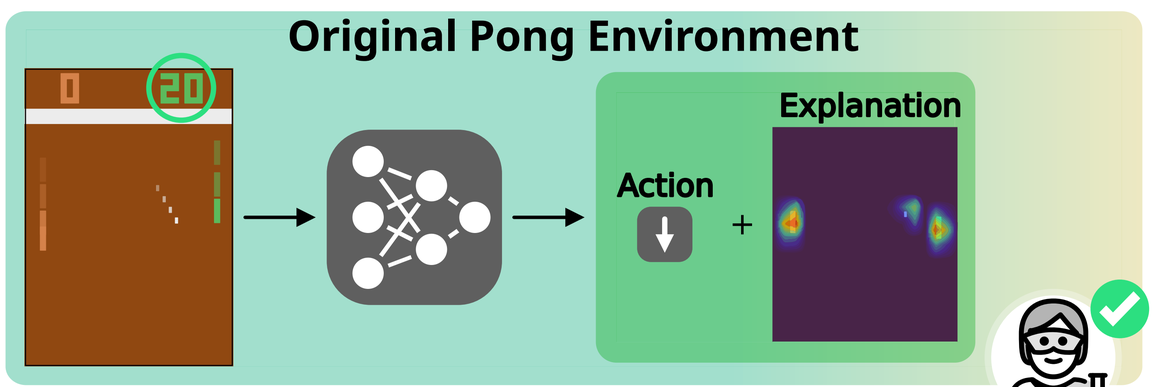
Elisabeth Dillies [1*], Quentin Delfosse [2,3*], Jannis Blüml [2,4],
Raban Emunds [2], Florian Peter Busch [2], Kristian Kersting [2,4,5,6]

RLDM 2025

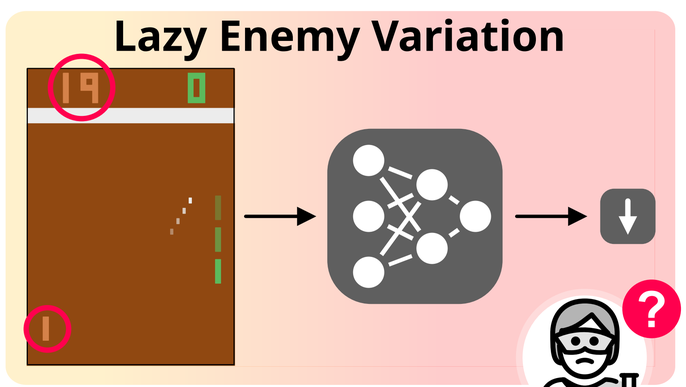
RL agents learn hidden shortcuts. Use interpretable causal world models.

Motivation

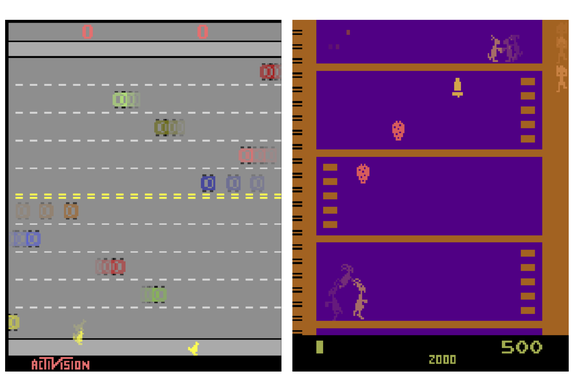
- Traditional opaque **deep RL approaches are prone to shortcut learning**, hidering human-like generalization [1,2].
- We develop a novel algorithm that leverages object-centric RL environments [3] to automatically **extract causal world models**.
- RL agents with causal world models **take the right actions for the right reasons**, rather than based on spurious correlations.



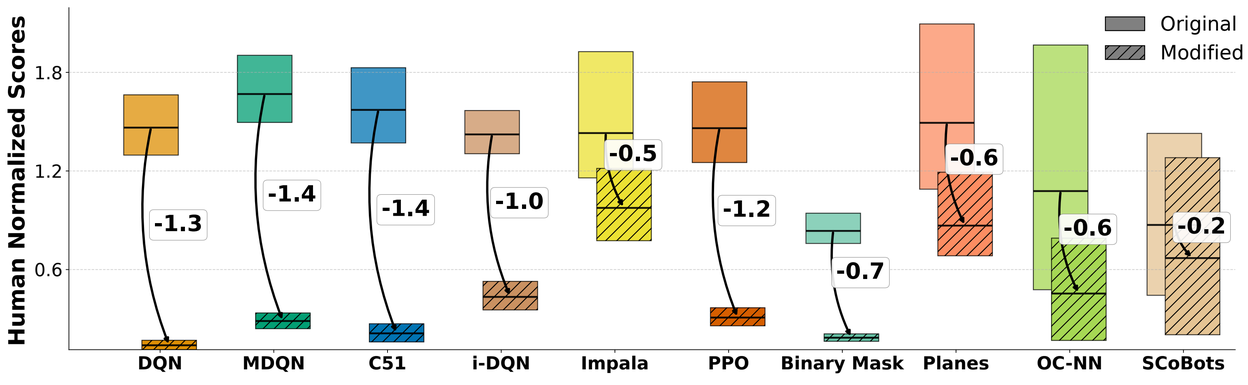
Evaluating on training environment leads to perfect score ✅, consistent actions ✅, and intuitive explanation maps ✅.



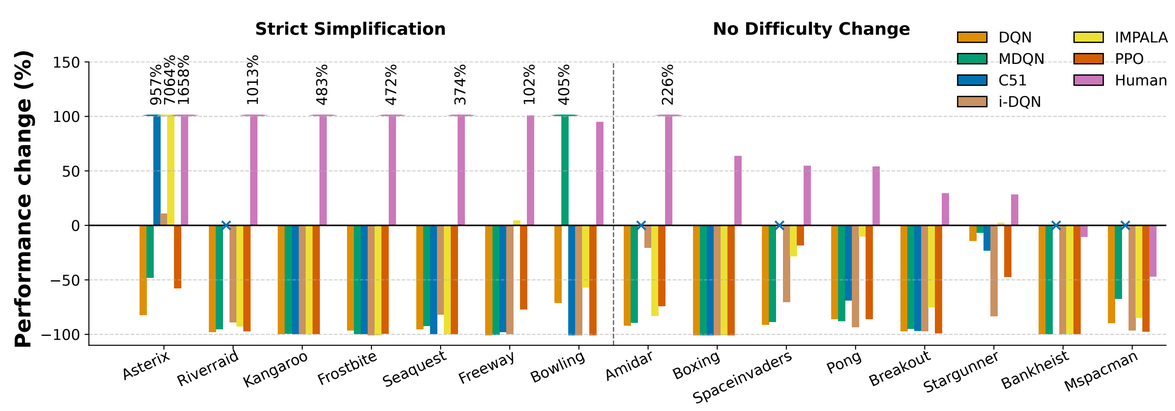
However, changing the enemy's behavior destroys the agent's policy.



HackAtari altered environments: Freeway and Kangaroo

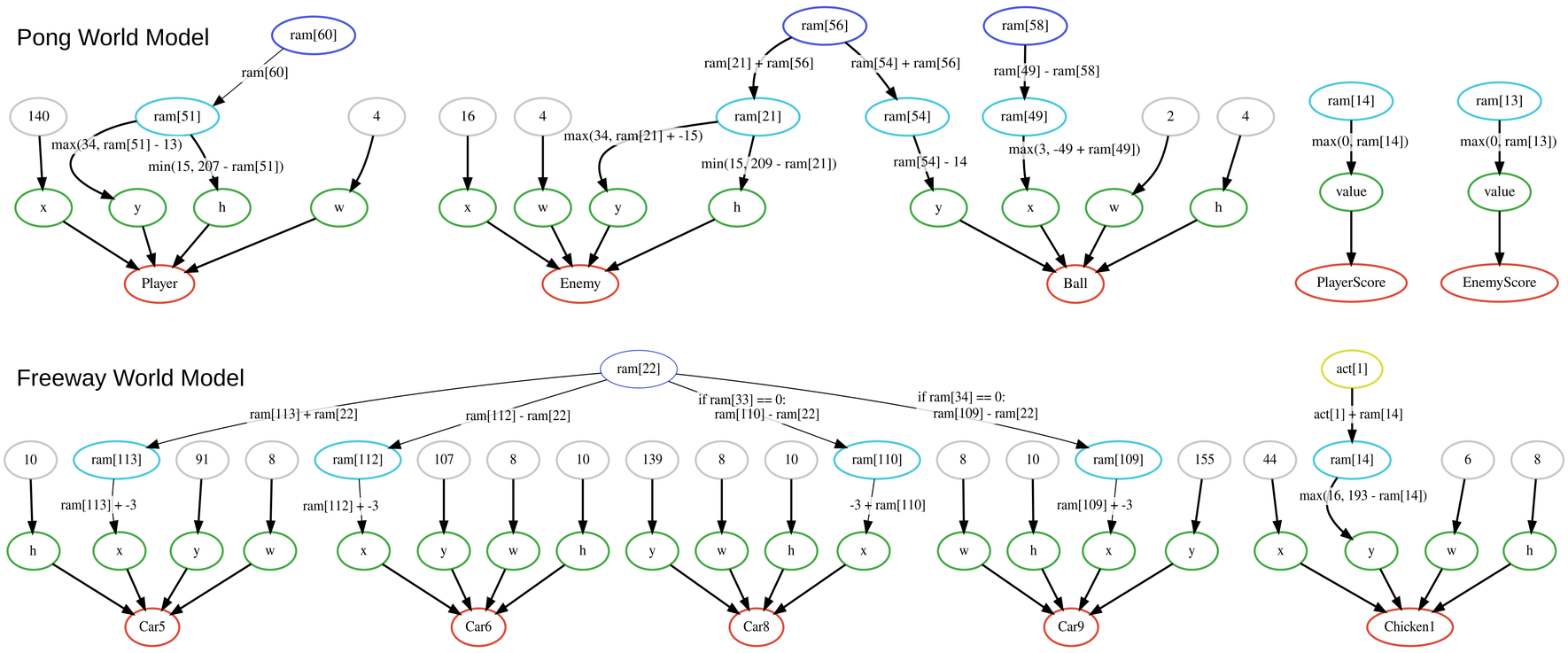


Average performances drop of RL algorithms on tasks simplifications.



Detailed (per-game) performance changes of RL agents and humans.

Causal Object-centric Model Extraction Tool



World models extracted by COMET for Pong and Freeway games.

- **Extract the objects and their properties** from the observations
- **Identify the environment's internal states** corresponding to the extracted objects' properties
- **Model object-centric transitions** uncovering causal relationships governing the objects' dynamics
- **Add semantic inference** using LLMs to annotate causal variables and enhance interpretability
- This process allows us to extract the true causal relations of the games and to **reprogram them in JAX.**

[1] Di Langosco et al. "Goal misgeneralization in deep reinforcement learning" (2022)
[2] Delfosse et al. "Deep Reinforcement Learning Agents are not even close to Human Intelligence" (2025).
[3] Delfosse et al. "OCAtari: Object-Centric Atari 2600 Reinforcement Learning Environments" (2024)


Elisabeth Dillies


Quentin Delfosse


Jannis Blüml

OCAtari HackAtari JAXAtari

[1] Sorbonne Université
[2] AIML Lab TU Darmstadt
[3] ATHENE
[4] hessian.AI
[5] TUDa Centre for Cognitive Science
[6] German Research Center for AI