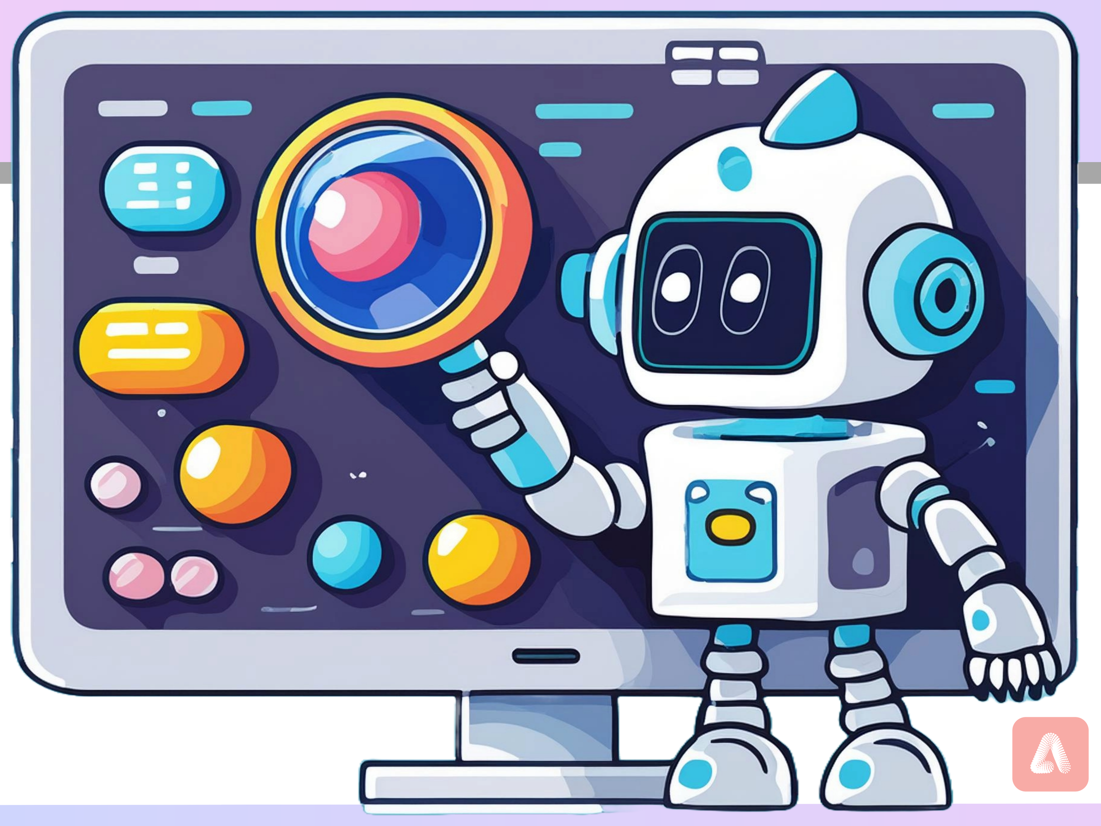
OCAtari: Object-Centric Atari 2600 Reinforcement Learning Environments

Quentin Delfosse[*,1,2] Jannis Blüml[*,1,3] Bjarne Gregory[1]
Sebastian Sztwiertnia[1,4] Kristian Kersting[1,3,5,6]

Reinforcement Learning Conference

We need neurosymbolic agents for interpretable decision making. Use resource-efficient and precise object-centric Atari environments.

SCAN ME

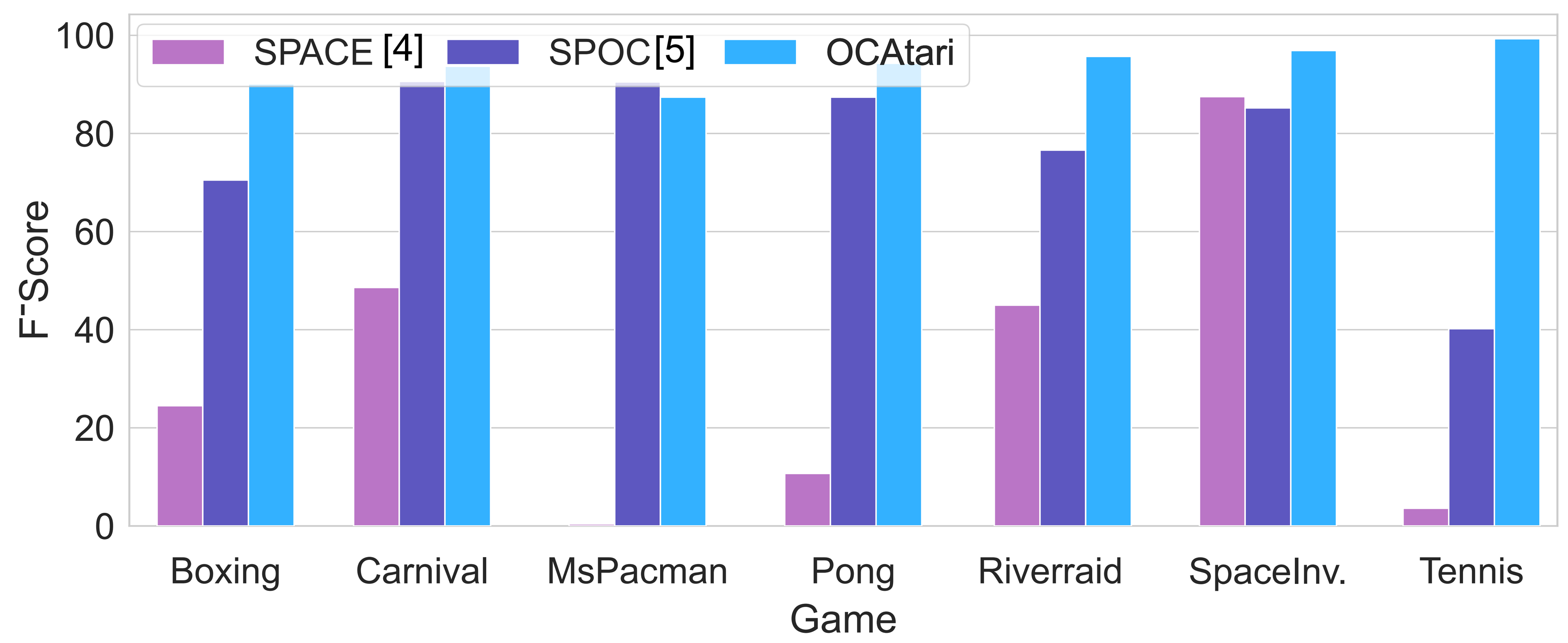


Goal: Object-Centric RL Agents

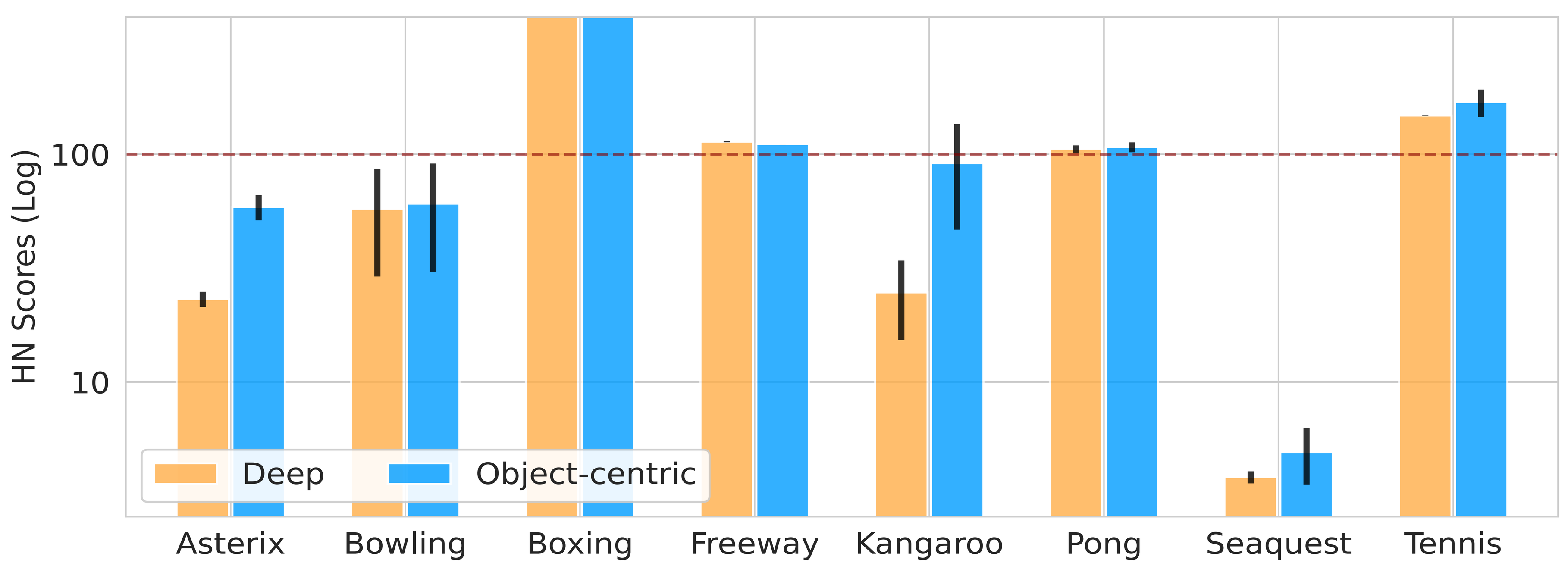
- Traditional deep (pixel-based) RL approaches are opaque and do not result in corrigible agents, prone to shortcut learning [1,2,3].
- Effective decision-making in RL relies on understanding and interacting with distinct objects within an environment.
- We need to efficiently train object-centric RL agents that can recognize objects and reason on their relations.

Results: 50+ Object-Centric Atari Envs

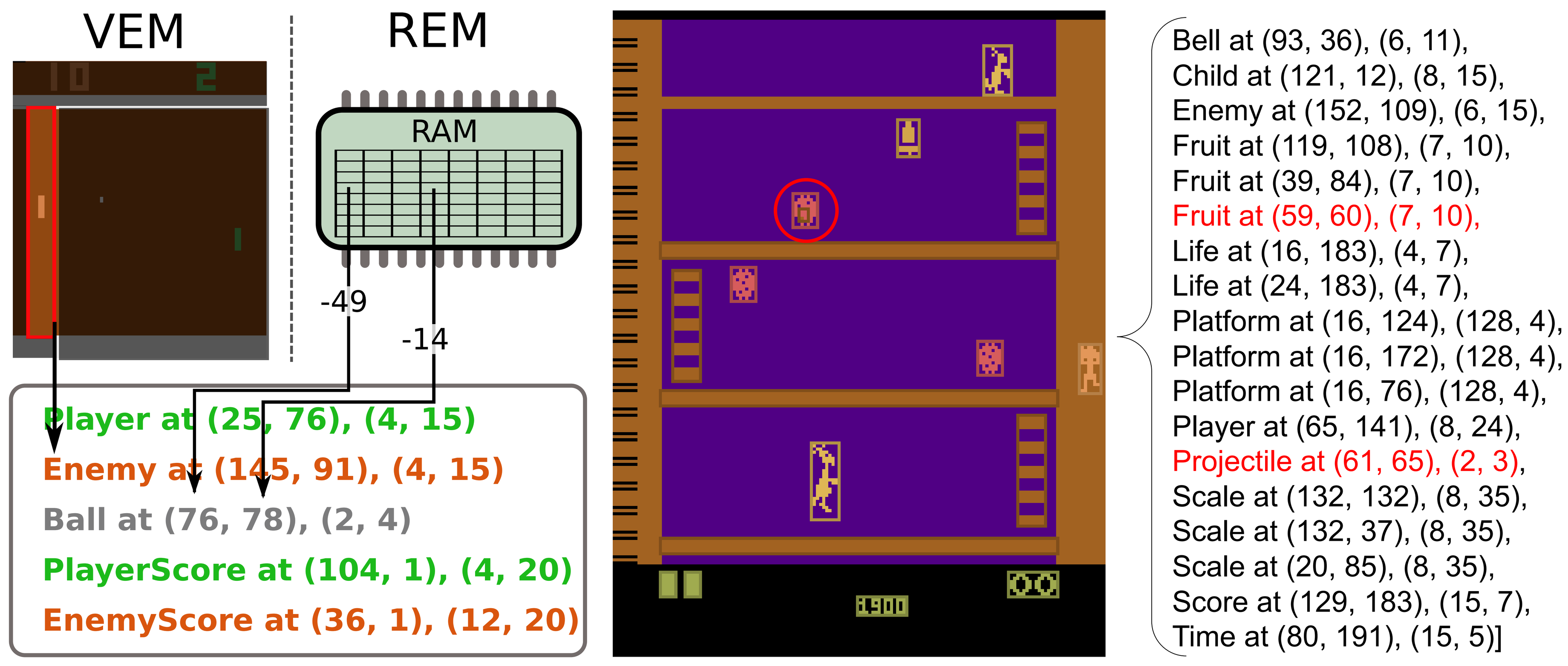
Accurate Object Detection on 50+ Atari Environments: OCAtari REM outperforms SOTA Atari object discovery methods. Results for all environments tested by SPACE [4] and SPACE+MOC [5].



Enhanced Decision-Making: OCAtari's object-centric agents match or surpass deep RL agents on multiple ALE games.



OCAtari: Efficient Object-Centric States



Object-Centric Extraction on the most popular benchmark: OCAtari covers **50+** Atari environments to train interpretable RL agents.

Efficient Processing: Develops resource-efficient ram-based extraction of the objects' attributes to ensure accurate object detection without compromising performance.

Vision Extraction Method (VEM): Uses computer vision techniques to identify objects using RGB values and positions from game frames (slow but accurate, used as a benchmark).

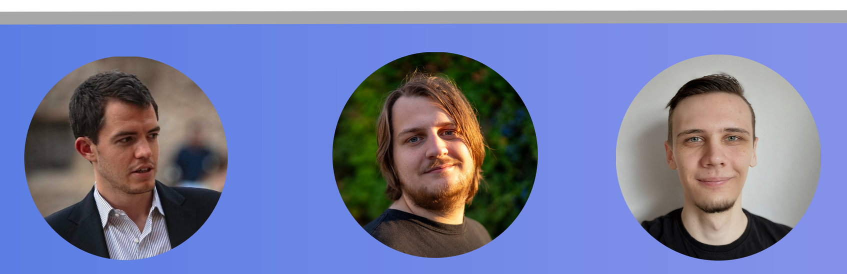
RAM Extraction Method (REM): Retrieves the objects' information directly from the RAM, efficiently providing neurosymbolic state representations.

Open Source Python implementation: pip installable public python package. SB3 and cleanRL integrations available.

Conclusion

We **need object extraction** for **interpretable** neurosymbolic **agents** to efficiently learn, **generalize** and **adapt** like humans. The **experimenter bias free** *Atari Learning Environment* is the **most used benchmark**, with **100 different games**. To **efficiently train** your **neurosymbolic RL agents**, use the **RAM extraction** of our ***Object-Centric Atari*** Environments.

[1] Di Langosco et al. "Goal misgeneralization in deep reinforcement learning." (2022)
[2] Delfosse et al. "Interpretable and explainable logical policies via neurally guided symbolic abstraction." (2024).
[3] Delfosse et al. "HackAtari: Learning Environments for Robust and Continual Reinforcement Learning." (2024).
[4] Lin et al. "SPACE: Unsupervised Object Scene Representation via Spatial Attention and Decomposition." (2020)
[5] Delfosse et al. "Boosting object representation learning via motion and object continuity." (2023).



Quentin Delfosse
Jannis Blüml
Bjarne Gregori

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