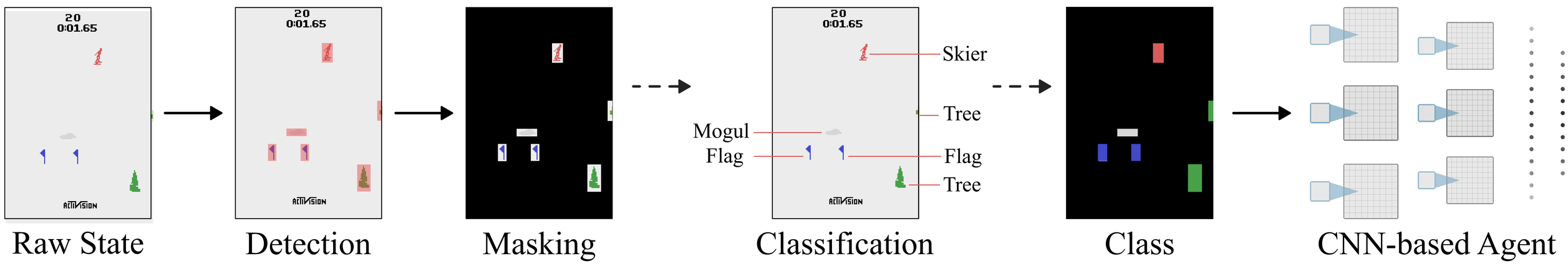
Balancing Abstraction and Spatial Relationships for Robust Reinforcement Learning

Jannis Blüml[1,2*], Cedric Derstroff[1,2*], Elisabeth Dillies[3], Quentin Delfosse[1,4], Kristian Kersting[1,2,5,6]

RLDM 2025

Focus on what matters. Simple, structured representations improve RL agents.

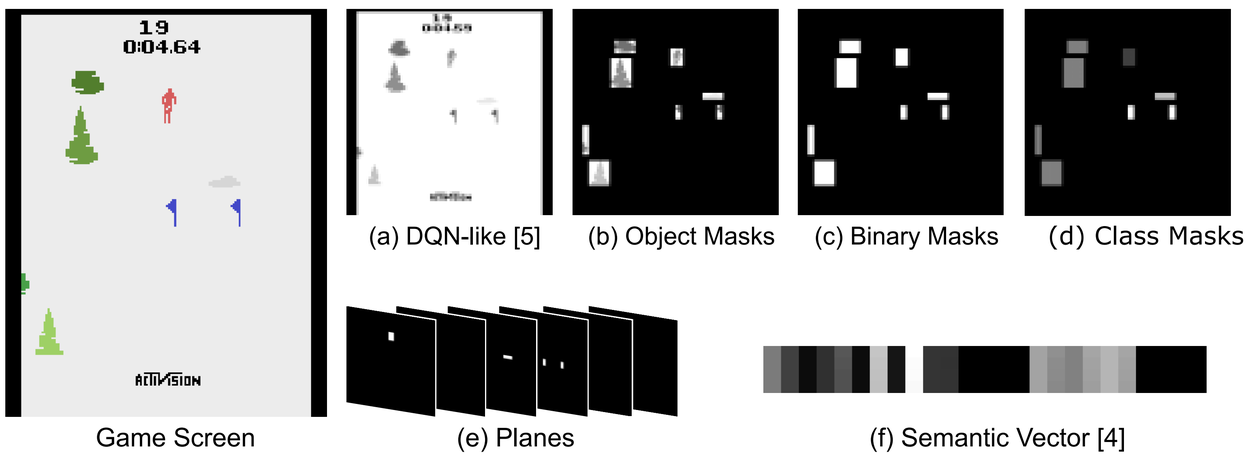
SCAN ME



Idea

- Traditional deep (pixel-based) RL approaches are opaque, do not result in corrigible agents and are prone to shortcut learning.[1,2]
- Inspired by human perception and advantages of object-centric RL, we propose **object-centric attention** to simplify inputs **by masking** irrelevant pixels.[3,4]
- We **retain only task-relevant objects and their spatial positions**, reducing input complexity without losing critical structure.

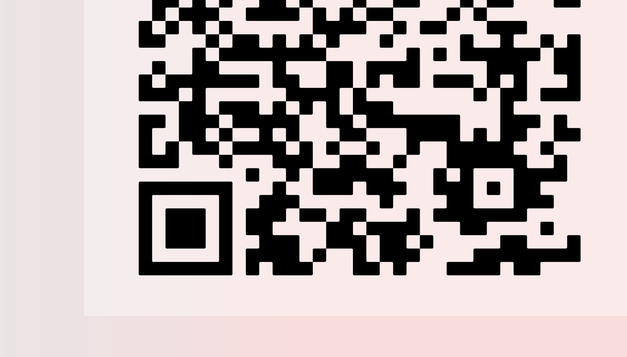
Object-Centric Attention via Masking



- **Filter what matters:** OCCAM masks out background noise, letting RL agents focus on relevant objects (just like human attention).
- **Preserve structure:** It keeps spatial relationships intact by retaining object positions in the input.
- **No complex extraction:** It works with simple object detection (e.g. bounding boxes), avoiding full symbolic processing or handcrafted features.
- **Plug-and-play:** OCCAM fits into standard CNN-based RL setups.

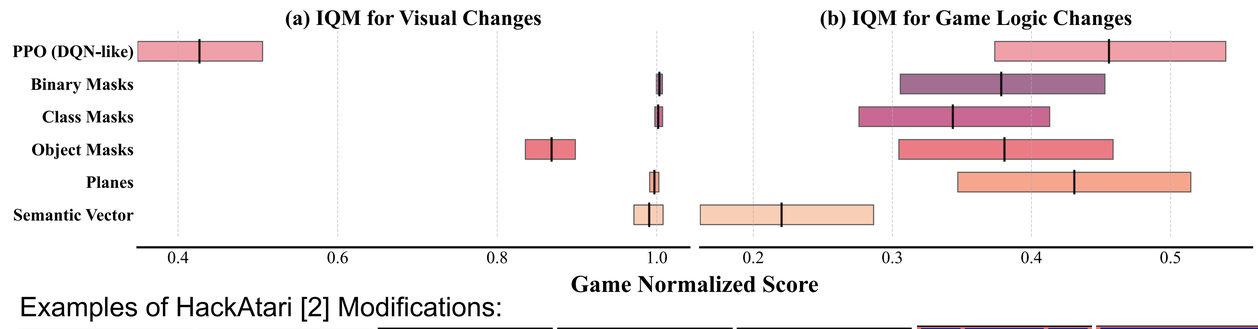
Results

1. Object-centric masking improves generalization and robustness, without hurting performance.

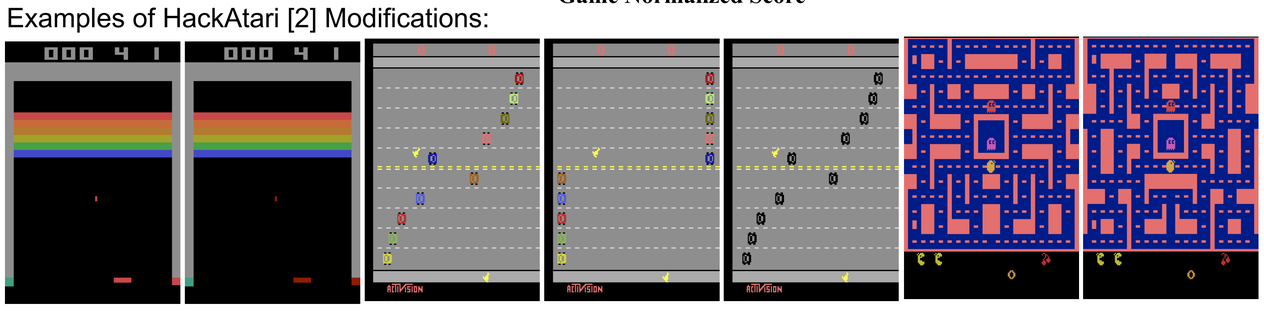


2. Removing spatial information decreases performance.
Preserving object relationships is important for effective decision-making.

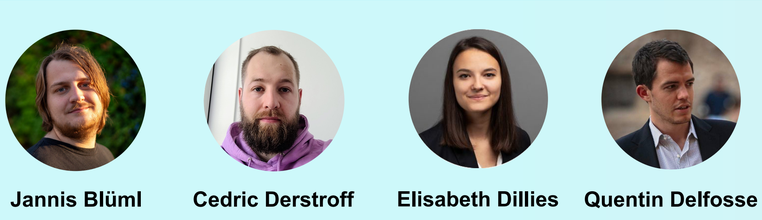
3. Abstraction alone is not enough. While improving robustness against visual distractions, agents remain vulnerable to changes in game mechanics.



Examples of HackAtari [2] Modifications:



[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] Davidson et al. "Investigating Simple Object Representations in Model-Free Deep Reinforcement Learning" (2020)
[4] Delfosse et al. "OCAtari: Object-Centric Atari 2600 Reinforcement Learning Environments" (2024)
[5] Mnih et al. "Playing Atari with Deep Reinforcement Learning" (2014)



Jannis Blüml
Cedric Derstroff
Elisabeth Dillies
Quentin Delfosse

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