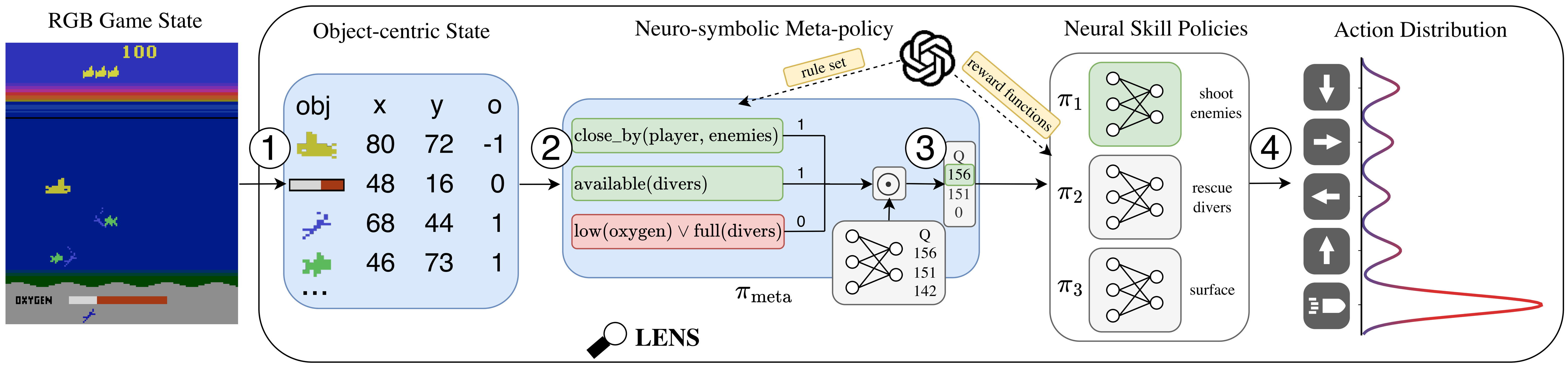
# Intepretable Reinforcement Learning via Meta-Policy Guidance

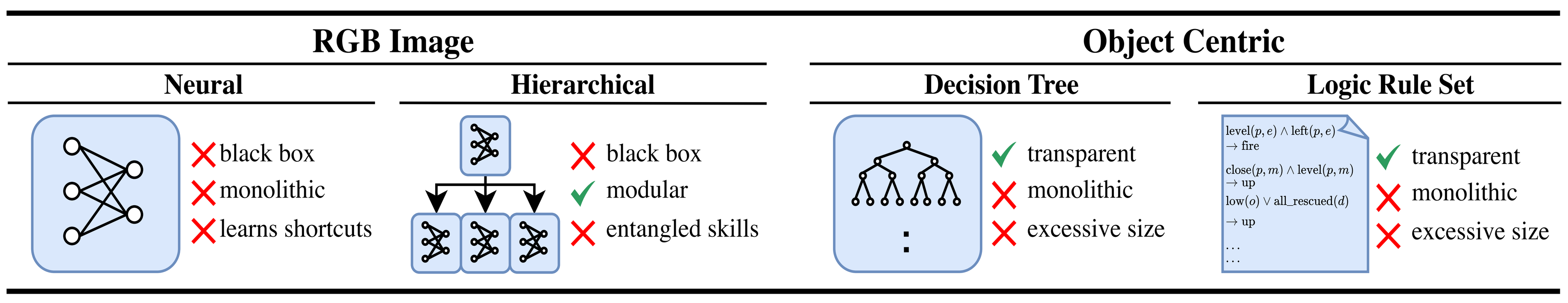
Raban Emunds[1], Jannis Blüml[1,2], Quentin Delfosse[1,3], Kristian Kersting[1,2,4,5]

EWRL18
September 17-19 2025, Tübingen

# Improve RL interpretability by combining symbolic meta-policies with neural skills.




## Motivation




- Traditional RL approaches often exhibit **misaligned behavior** that is **difficult to identify or correct** without interpretability.[1,2]
- But: Existing interpretable RL methods deployed on atomic actions quickly become **excessively complex.**
- Instead: Employ interpretable RL on **abstract skills** based on **object-centric** input and learn skills with **LLM-generated rewards**.

## Logically Enhanced Neural Skills

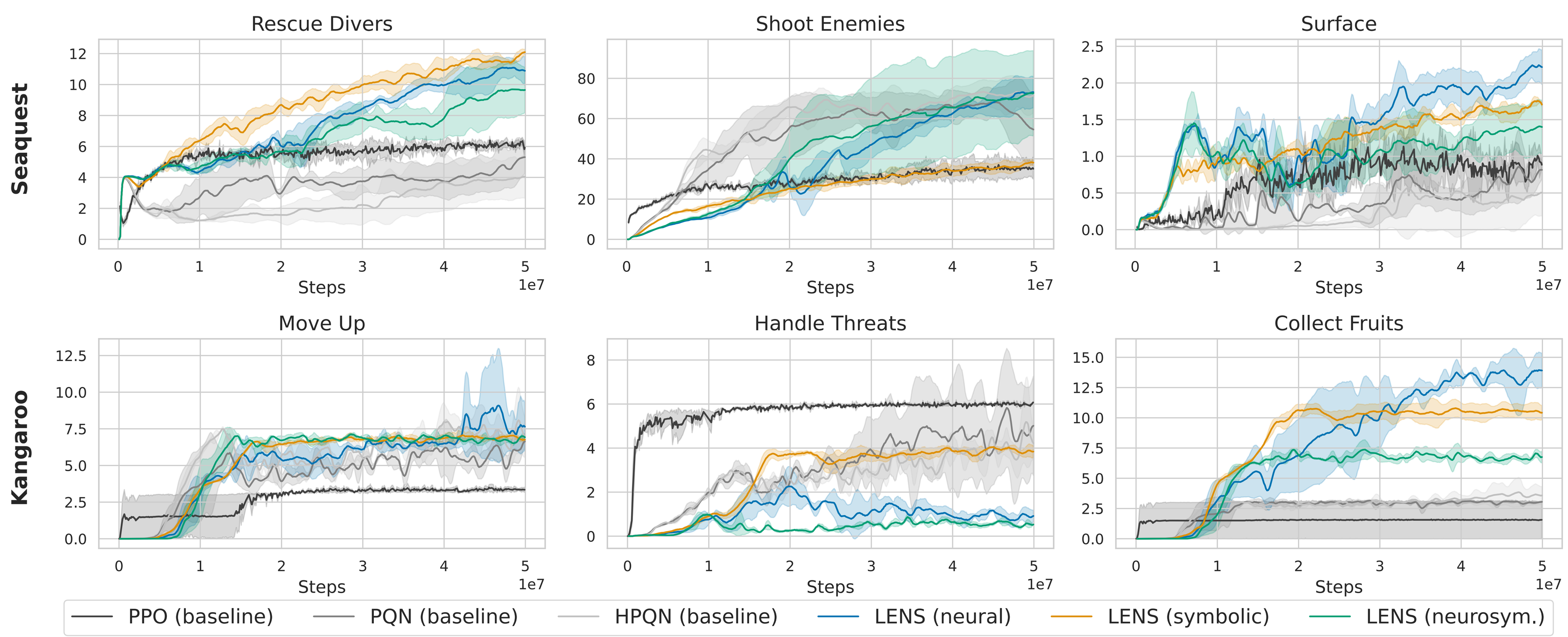
- **(1) Extract Objects + Attributes :** First, transform image state into object-centric (OC) state using existing methods (e.g. [3, 4, 5]).
- **(2) Filter Relevant Skills:** Remove skills not applicable in the current situation using LLM-generated filters (based on OC-input).
- **(3) Maximize Meta-Q-Values:** Select the neural skill to be executed by maximizing the Q-values of the meta policy with remaining skills.
- **(4) Execute Neural Skill:** Obtain the next action by maximizing the skill-specific Q-value function.
- Three meta-policy variations: **neural, symbolic, neuro-symbolic**

```
def meta_policy(st: state):
    if enemy_close(st.enemies,
    st.player):
        return fight_enemies()
    elif is_available(st.divers):
        return rescue_divers()
    elif is_low(st.oxygen):
        return surface()
    elif all_collected(st.divers):
        return surface()
    return rescue_divers()
```

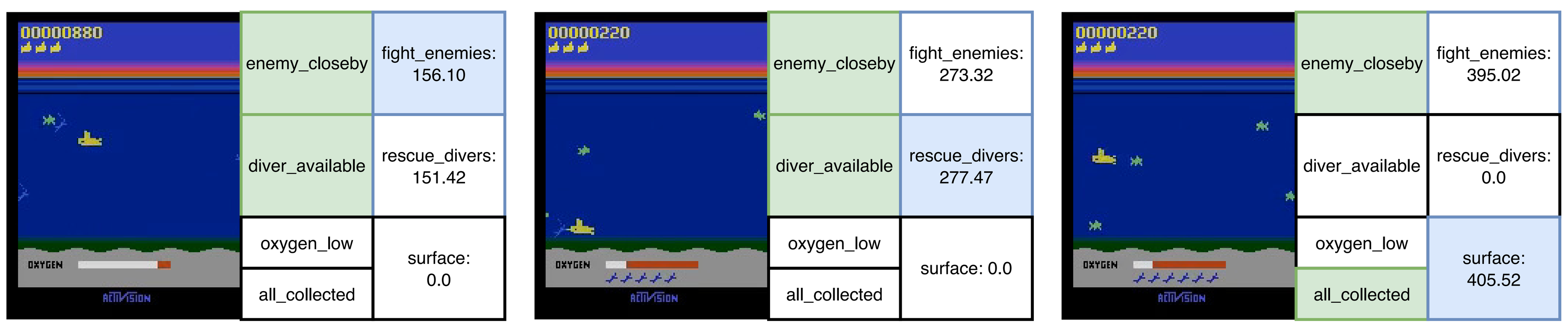
```
def meta_policy_rules(st: state):
    fight_enemies = False
    rescue_divers = True
    surface = False
    if enemy_close(st.enemies,
    st.player):
        fight_enemies = True
    if oxygen_low(st.oxygen):
        surface = True
    return [fight_enemies,
    rescue_divers, surface]
```

Examples of a **symbolic meta-policy** (left) and rule-set for the **neuro-symbolic meta-policy** (right) in Seaquest.

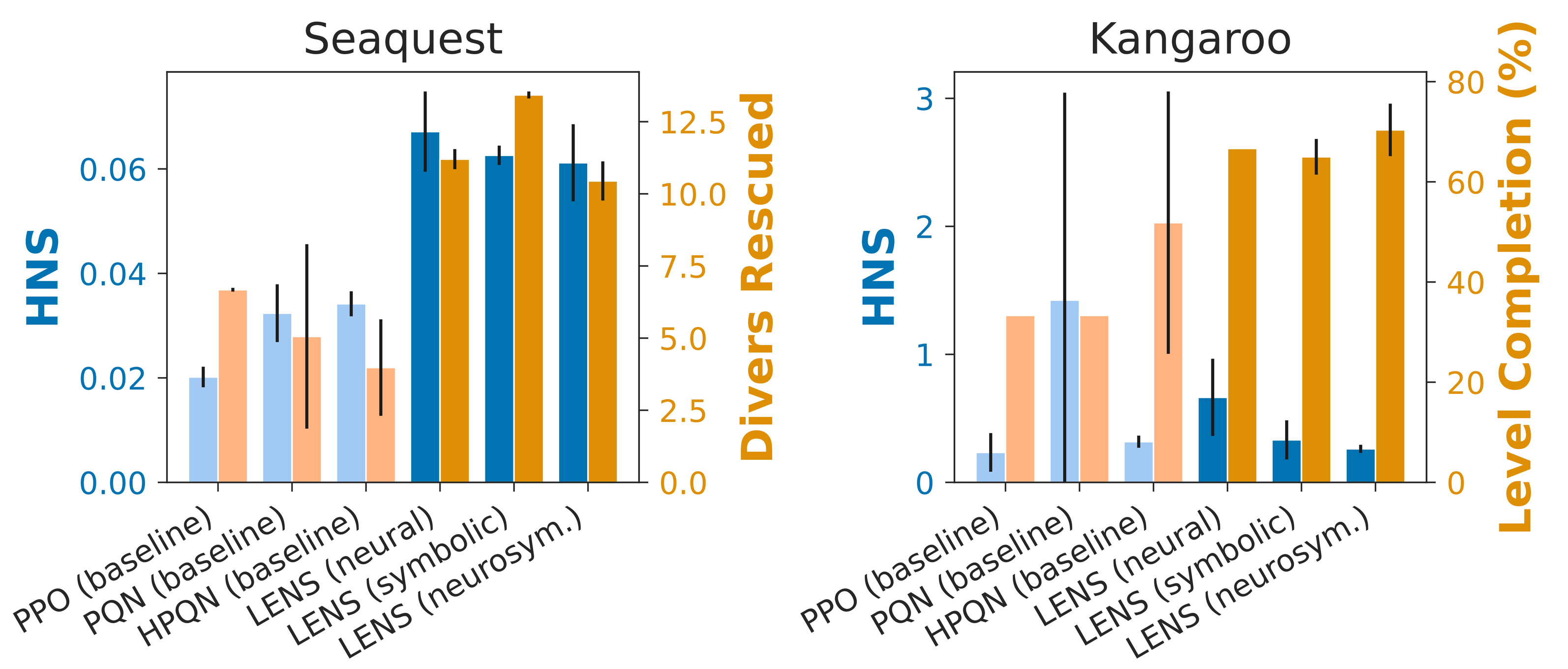
## Results




**1. LENS learns disentangled Skills jointly from off-policy data.**




**2. LENS produces interpretable yet flexible high-level plans.** Decisions in ambiguous situations are made intuitively.




**3. LENS is competitive and better aligned to actual environment goals.**

[1] Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." (2019)
[2] Delfosse et al. "Interpretable Concept Bottlenecks to Align Reinforcement Learning Agents" (2024)
[3] Li et al. "Object-sensitive Deep Reinforcement Learning" (2017)
[4] Locatello et al. "Object-Centric Learning with Slot Attention." (2020)
[5] Delfosse et al. "Boosting Object Representation Learning via Motion and Object Continuity" (2023)

Raban Emunds Jannis Blüml Quentin Delfosse

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