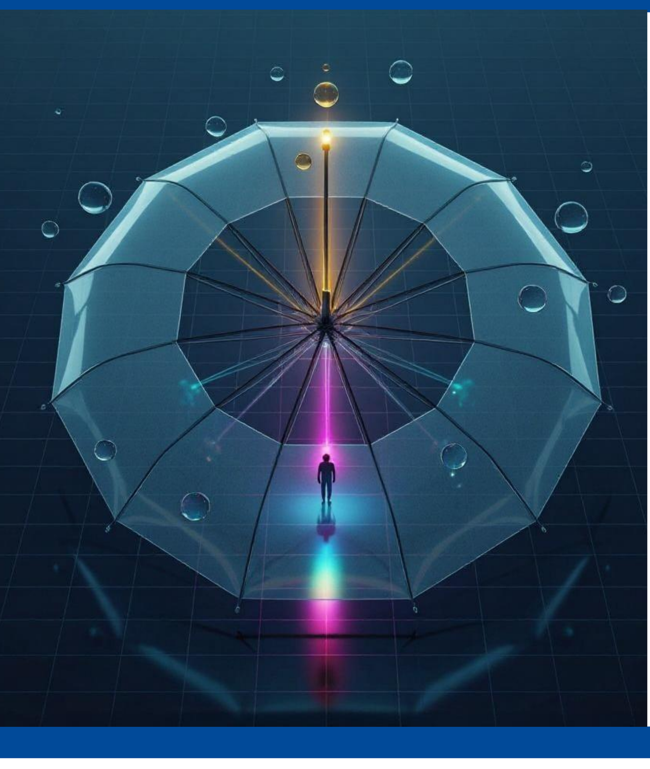


MetaIRN
Meta-Gradient-based Intrinsic Rewards via Attention Network

Philipp Link, Julius Heidmann, Luan Liebig-Schultz
{p.link, julius.heidmann, luan.liebig-schultz}@stud.uni-hannover.de

Leibniz Universität Hannover

LUH|AI

1 Motivation

- **Sparse Reward Environments** make Learning for Agents difficult → **slow/unstable**
- How to **facilitate Training**? → Use **Intrinsic Rewards** to **enrich Feedback** through **Temporal Credit Assignment (TCA)**
- Which **Actions** lead to **Future Success**?
- **TCA**: Associate Actions with Temp. distant Rewards and assign Credit (int. Rewards)

2 Related Work & Setting

SAIR[1]/ **LIRPG**[2] [Jiang et al. 2021], [Zheng et al. 2018]

- **Meta-Gradient** to train an **Attention Network** producing **Intrinsic Rewards** to facilitate PPO learning

Decoupling Memory from **TCA**[4] [Ni et al. 2023]

- Exactly **distinguish** between **TCA** and **Memory ability** of an Agent
  - → Memory Problem[4]:

    Definition 1.A (Reward memory length $m^{\mathcal{M}}_{\text{reward}}$). *For a POMDP $\mathcal{M}$, $m^{\mathcal{M}}_{\text{reward}}$ is the smallest $n \in \mathbb{N}$ such that the expected reward conditioned on recent n observations is the same as the one conditioned on full history,* i.e., $\mathbb{E}[r_t \mid h_{1:t}, a_t] = \mathbb{E}[r_t \mid h_{t-n+1:t}, a_t], \forall t, h_{1:t}, a_t$.
  - → Credit Assignment Problem[4]:

    $$c(h_{1:t};\pi) := \min_{1 \le n \le T-t+1} \{n \mid \exists a_t^* \in A_t^*, \text{s.t. } G_n^\pi(h_{1:t}, a_t^*) > G_n^\pi(h_{1:t}, a_t'), \forall a_t' \notin A_t^*\}$$

3 Approach

- Can a **Combination** of **Meta-Gradient Optimization** and **Attention-based intrinsic Rewards** enhance the **Agent's Ability** to **assign Credit** in **TCA-only scenarios**?
- Based on **SAIR**[1]/ **LIRPG**[2] **Meta-Gradient**:
  - **Inner Loop:** PPO Update (Ex.&In. Rew.)
  - **Outer Loop:** Attention Net Update (Ex. Rew.)
- **This Agent** is trained in **Umbrella Length**[3] **(TCA-Only)** Env, which is **isolated** to **focus** exclusively on **TCA problems**
- **Comparison** to **Vanilla PPO Performance**

Meta-Gradient Update[1]**:**

PPO Network:

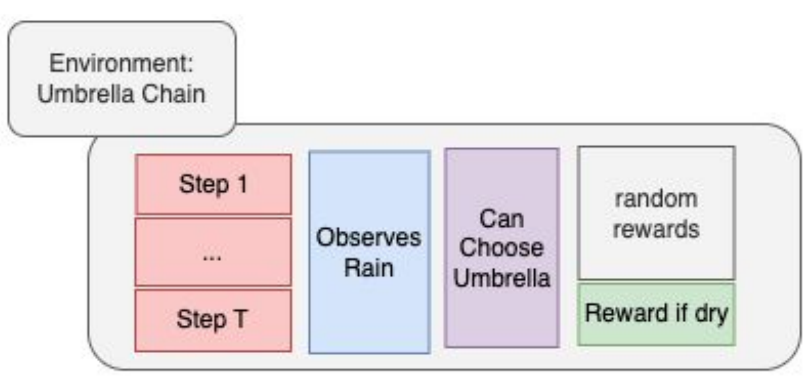
$$\theta' \leftarrow \theta + \alpha \nabla_\theta J^{ex+in}(\theta|\mathcal{D})$$

Intrinsic Network:

$$\eta' \leftarrow \eta + \alpha \nabla_\eta J^{ex}(\theta'|\mathcal{D}) \nabla_\eta \theta'$$

4 Umbrella Length[3] Env

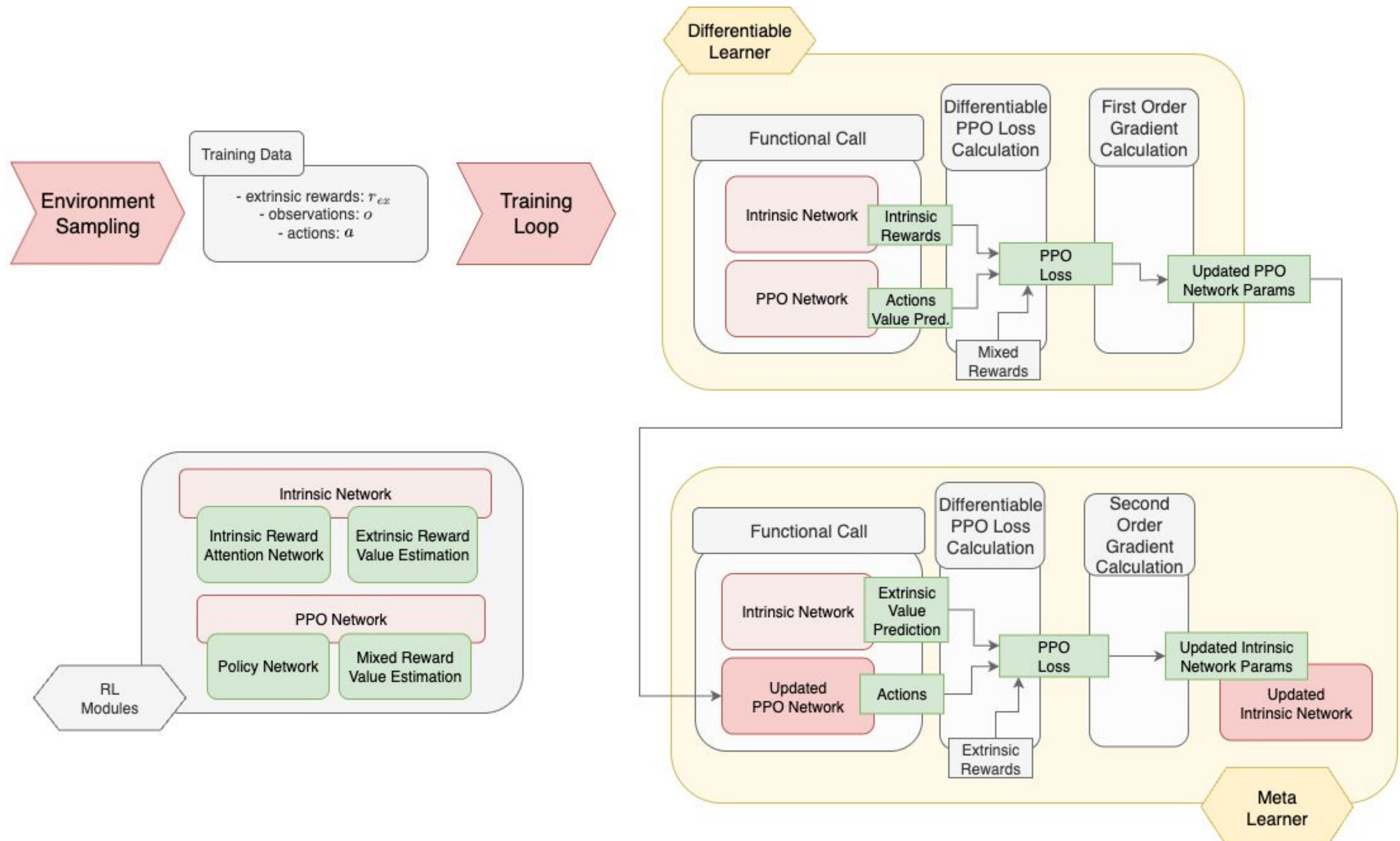
- **Fully observable MDP**
- **Intermediate Reward** $R_i \in \{-0.1, 0.1\}$
- **Final Reward** $R_f \in \{-10, 10\}$ depends only on First Action



5 Future Works

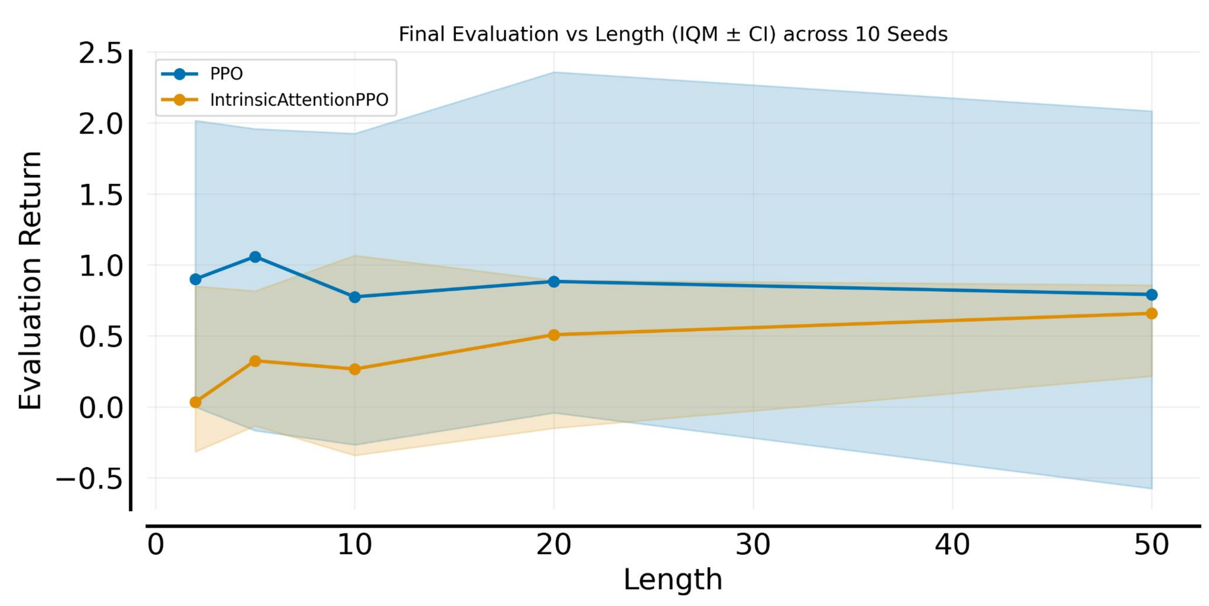
- **TMaze Env** (Memory-Only Test)
- **Meta-Gradient Implementation**
- **Research Ext. Value Pred. Behaviour**

4 Key Insights

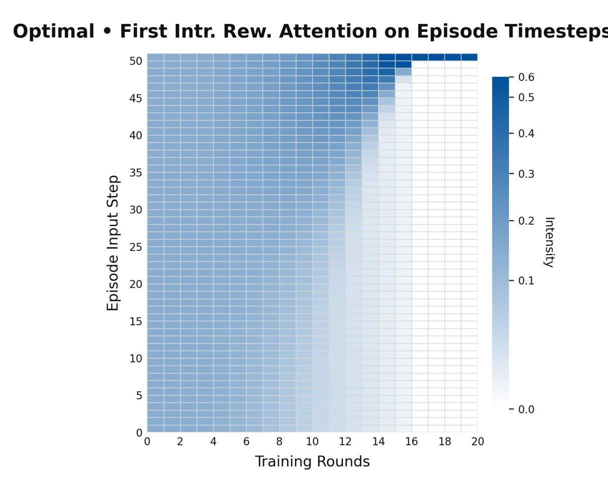


RLLib Implementation

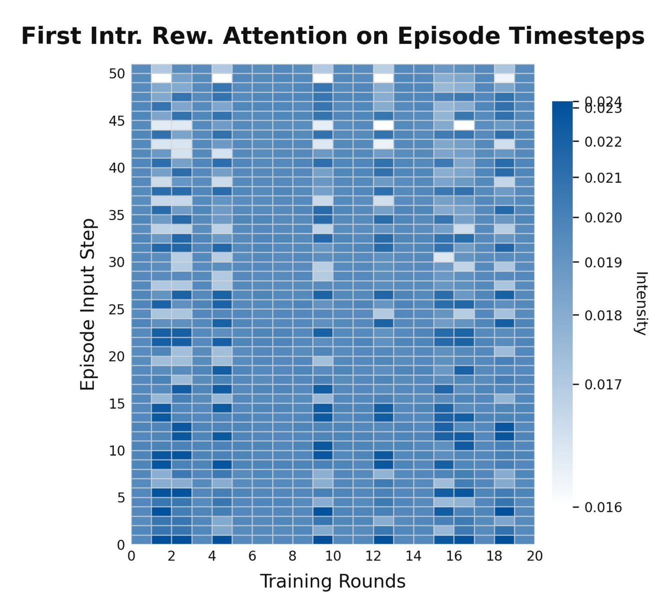
- **PPO** with **Intrinsic Attention & Vanilla**
- **Hyperparameter** based on **SMAC HPO** with **100 Configurations each**
- **10 Seeds Evaluations** for each **Environment Length (2, 5, 10, 20, 50)**

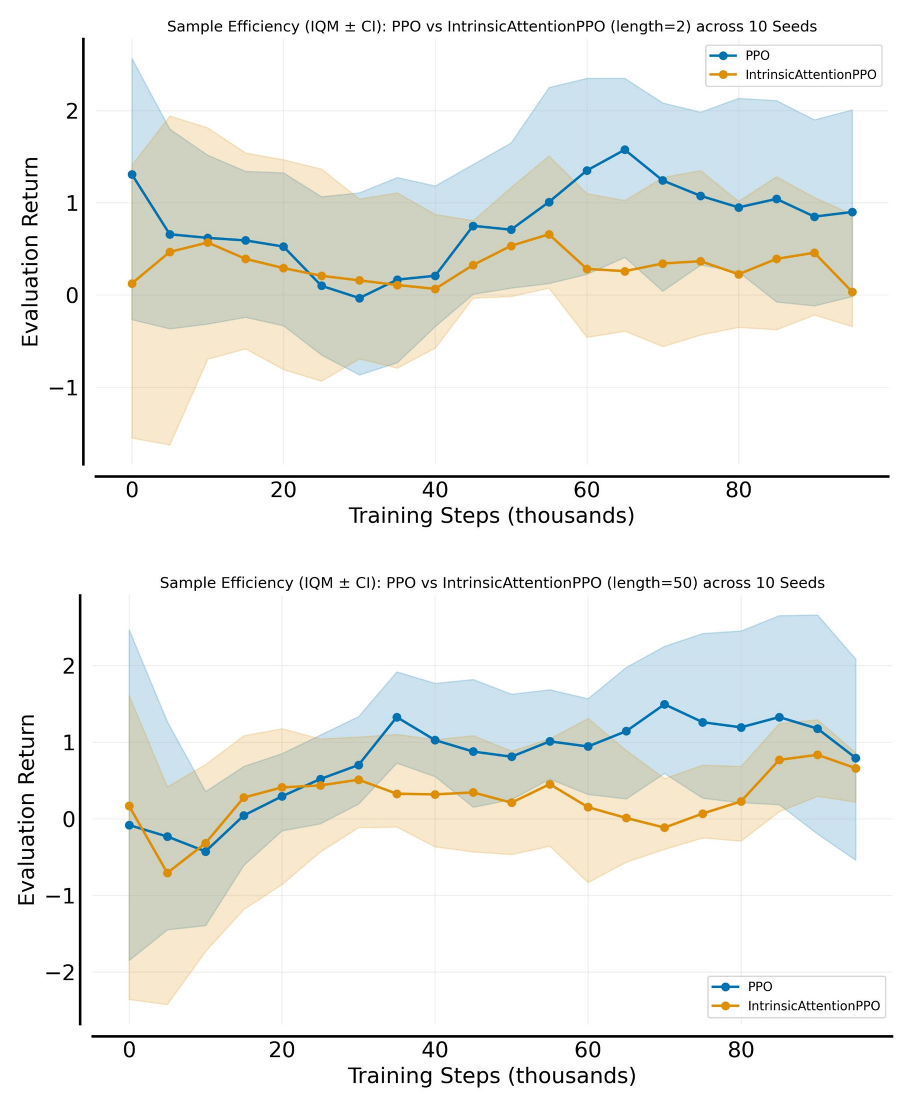


Performance over different environment lengths



Attention Matrix for Optimal Case and Trained Sample





Sample Efficiency for Selected Environment Lengths

Sources 1:https://ieeexplore.ieee.org/document/9727314 2:https://papers.nips.cc/paper_files/paper/2018/hash/51de85ddd068f0bc787691d356176df9-Abstract.html 3:https://openreview.net/forum?id=rygf-kSYwH
4:https://papers.nips.cc/paper_files/paper/2023/hash/9dc5accb1e4f4a9798eae145f2e4869b-Abstract-Conference.html