Appendix

Domain Descriptions

IPPC 2011/2014 Domains The IPPC 2011/2014 domains benchmarks consist of 10 instances per domain, of generally increasing difficulty. We provide brief descriptions of each domain below:

- **AcademicAdvising** This is a goal-oriented problem in which the objective is to obtain an academic degree. Specific courses must be taken within a time budget, while taking into account their prerequisites in the correct order. This domain heavily emphasizes sequential logical reasoning and backwards induction.
- **CooperativeRecon** The objective is to control a rover with three different tools for detecting the presence of life on a planet's surface. Sensors are noisy, they can be damaged and repaired by visiting a base, and they can contaminate the object with repeated measurement. The difficulty of this problem is the presence of complex logical preconditions, dead ends, and the need for careful sequential planning.
- **CrossingTraffic** In a grid, a robot must get to a goal cell, while avoiding obstacles arriving randomly and moving in one direction. Both collision and a goal state are modelled as absorbing states in the Markov chain. These also serve as dead ends in the domain, presenting a challenge for many replanning methods.
- **Elevators** The goal is to control a set of elevators to pick up passengers arriving randomly, by moving the elevators between floors and opening or closing the door. Another example where sequential reasoning is required, although we point out that the version as it was used in the competition contains bugs, making myopic policies perform considerably well.
- **GameOfLife** Encodes the Conway's cellular automata "game of life" on a grid. One gets a reward for generating patterns that keep the most cells alive. This domain is highly stochastic.
- **Navigation** A robot must get to a goal. However, every cell traversed causes the agent to disappear with a fixed probability, which decreases for longer paths taken to the goal. In addition to dead ends, this domain poses challenges for determinization methods, which typically prefer the shorter routes with the highest probability of failure over the full trajectory.
- **SkillTeaching** The objective is to teach a series of skills to a student through the use of hints and multiple choice questions. Similar to AcademicAdvising, some skills are prerequisites of others, emphasizing sequential logical reasoning.
- **SysAdmin** The objective is to reboot non-operational computers in a network, which fail with probability that depends on how many connected computers in the network are currently operational. Like GameOfLife, this domain is also highly stochastic.
- **Tamarisk** The aim of this domain is to either eradicate or replace with native species an invasive species that spreads out across space over time. This domain has highly stochastic transitions.
- **Traffic** Modelled using a cell transition model (CTM) of traffic flows, the overall aim of this domain is to advance the traffic signals at an intersection to optimally control traffic through the intersection.
- **TriangleTireworld** The goal is to arrive at a goal from a starting point by traversing a road network. There is a chance of getting a flat tire at each location, which could potentially be replaced with a (single) spare tire equipped by the car. It was intended to be difficult for determinization and replanning approaches, since the highest probability path to the goal is longer than other possible paths.
- **Wildfire** Similar in some respect to SysAdmin, the goal is to control the spread of fire in a grid by either putting out the fire or cutting the fuel. Each cell can combust with probability dependent on the number of neighbours on fire, making it a highly stochastic problem.

IPPC 2023 Domains The mixed discrete-continuous domains from IPPC 2023 contained 5 instances per domain, and are described below:

- **HVAC** The goal is to control the heating system (continuous actions) in a building with multiple interconnected rooms to maintain a specific temperature (continuous state) in each room. Occupancy is a Boolean stochastic variable, which should be taken into account in order to save on heating costs.
- **MarsRover** The goal is to navigate a set of rovers in a continuous space in order to harvest as many high-value minerals scattered in the space as possible, with harvesting controlled by Boolean actions. This problem is challenging due to its sparse reward nature, making it difficult for replanning methods.
- **MountainCar** The goal is to push a cart up a hill, by taking advantage of the surrounding terrain. The problem is designed such that a direct route to the top of the hill is not possible, requiring building up momentum in the valley below. This domain only provides a reward at the top of the hill, making it highly sparse reward.
- **PowerGen** The idea is to control a power distribution network, consisting of different types of power generation units with different cost characteristics, to meet demand for power in a region that is based on temperature. The challenge this domain poses is the unit commitment, in which some power units are costly to put into or out of operation (but perhaps cheap to run), while dealing with highly stochastic temperature transitions.
- **RaceCar** The goal is to move a vehicle from a starting location in a continuous space to a desired target location, while navigating around hard nonlinear boundaries. The problem is complicated by the highly sparse reward nature, since reward is only received close to the goal.
- **Reservoir** The goal is control the continuous flow of water in a series of interconnected reservoirs. This problem is difficult due to its stochastic transitions and high state/action dimension.

- **UAV** A set of unmanned aerial vehicles (UAVs) is to be flown in a 3-dimensional space towards a set of target locations. The challenge of this domain is that some UAVs are not controllable, instead moving according to random walks. This stresses the credit assignment capability of planners, which must learn to identify the noisy but practically meaningless state variables.

Detailed Empirical Results per Instance

In this section, we provide the complete average unnormalized returns per instance and domain for all relevant baselines, which was used to compute all normalized and aggregated results presented in the Main paper.

Figures 4, 5 and 6 illustrate the results on IPPC 2011/2014 domains using a 1, 3 and 5 second timeout per decision, respectively. Figures 7, 8 and 9 illustrate the corresponding results on IPPC 2023 domains.

Additional Hyper-Parameters

In addition to the hyper-parameters tuned during the experiment, JaxPlan used the Adam optimizer with default parameters, and a batch size of 32 for all domains. In addition, to ensure the inverse of the sigmoid exists, and to avoid potential saturation or overflow in the calculation of the soft Boolean actions $\tilde{a}_i$, we clip actions $\tilde{a}_i$ to the range $[\delta, 1-\delta]$, with $\delta = 0.001$ prior to computing θ_i'. For GurobiPlan, we set $\epsilon = 1e - 5$ for all constraints added to the MINLP. The batch size, δ and ε were selected prior to the experiment. For other baselines, we used the default values of any hyper-parameters as used in IPPC 2011/2014/2023, but we note that DiSProD also performs its own hyper-parameter tuning procedure.

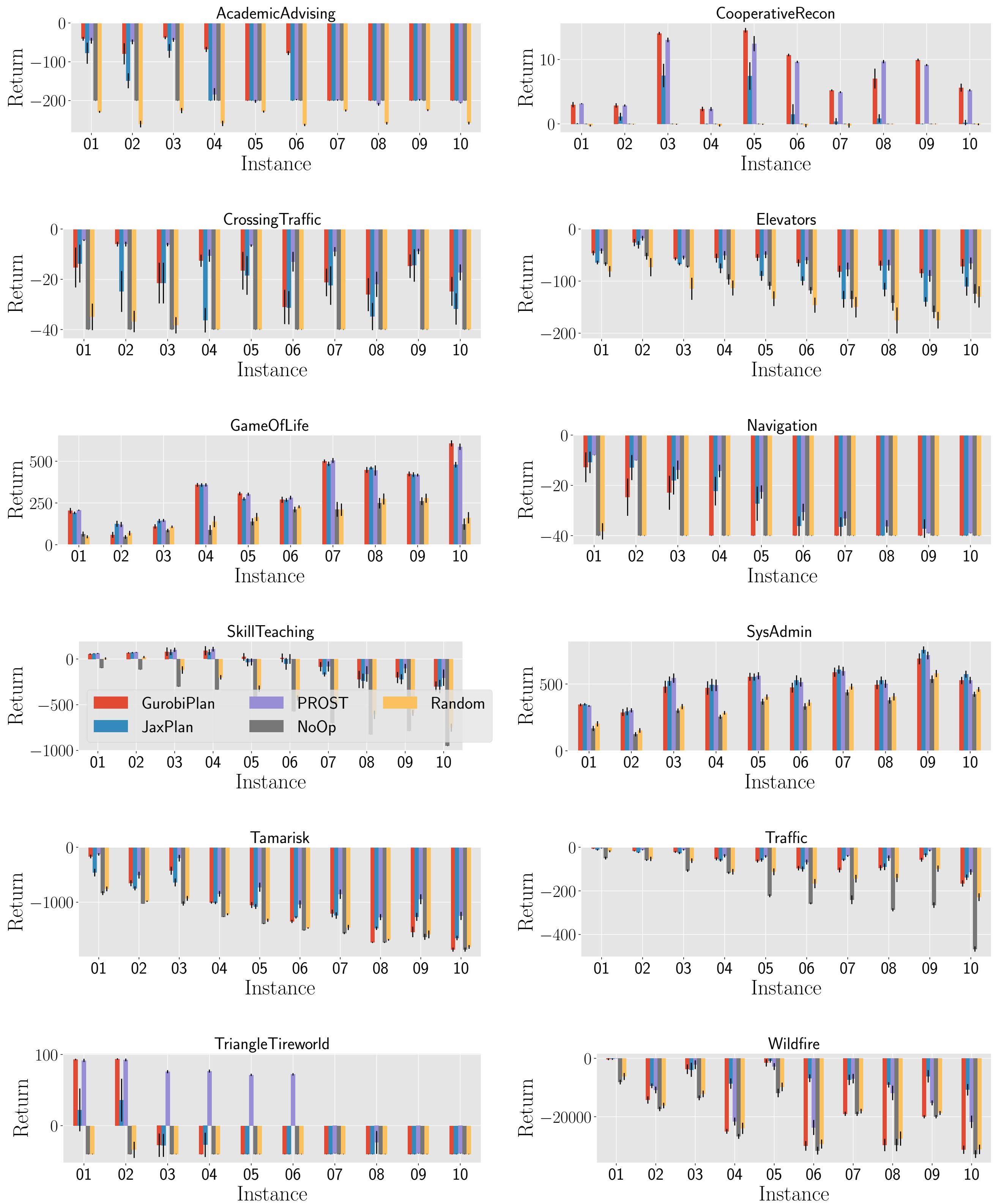

Figure 4: Unnormalized average return per domain and per instance on the IPPC 2011/2014 domains, using a 1-second time budget per decision.

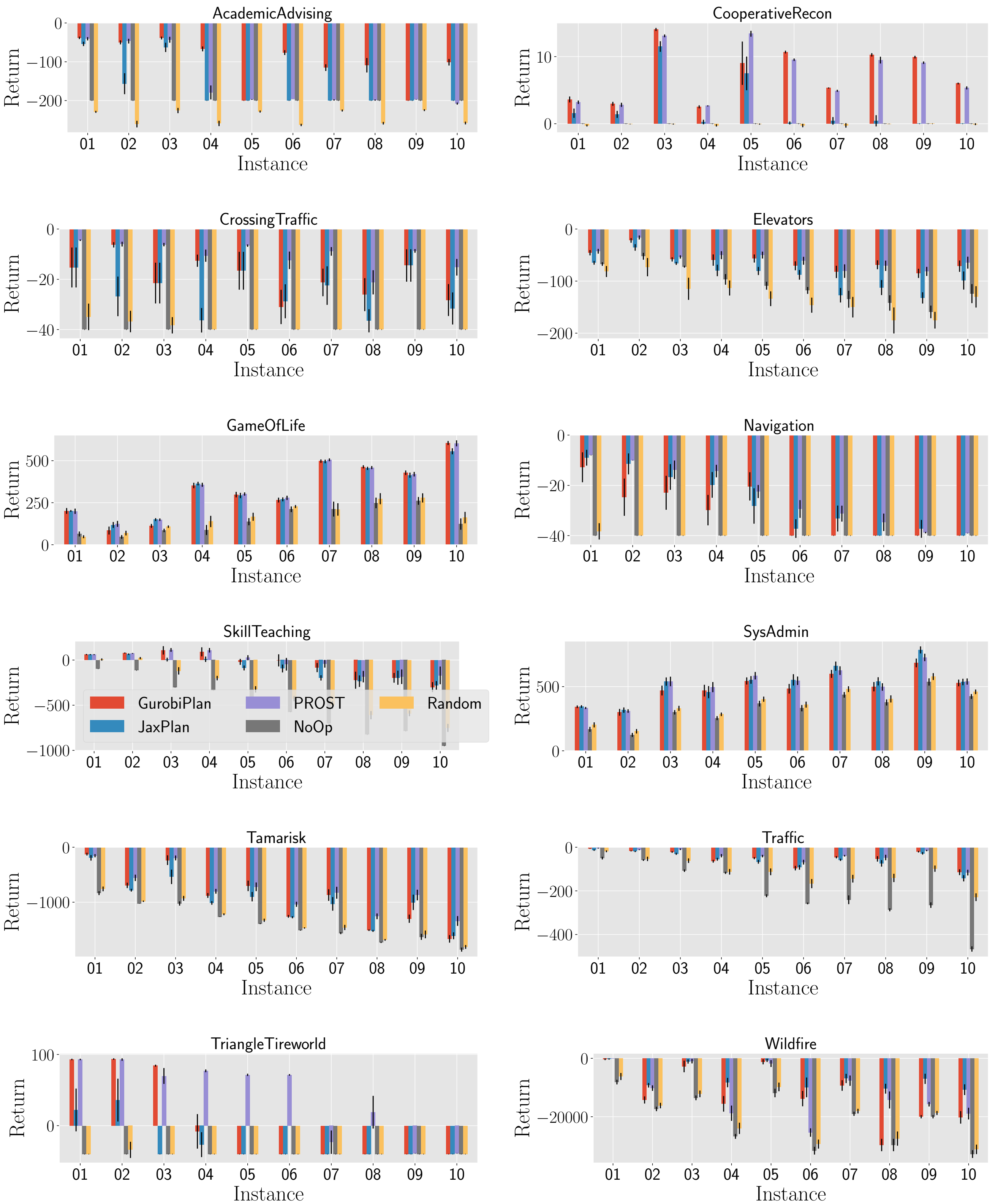

Figure 5: Unnormalized average return per domain and per instance on the IPPC 2011/2014 domains, using a 3-second time budget per decision.

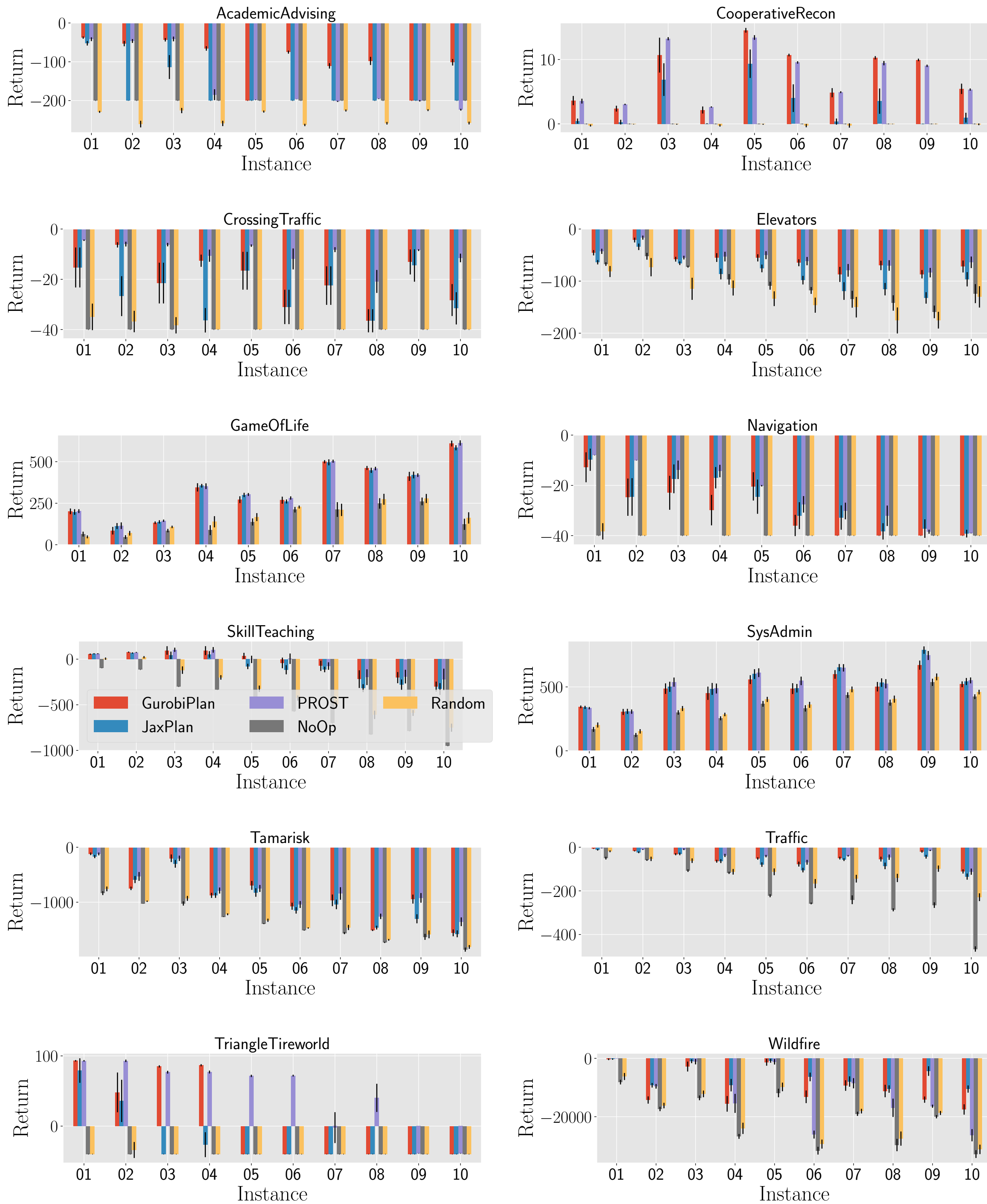

Figure 6: Unnormalized average return per domain and per instance on the IPPC 2011/2014 domains, using a 5-second time budget per decision.

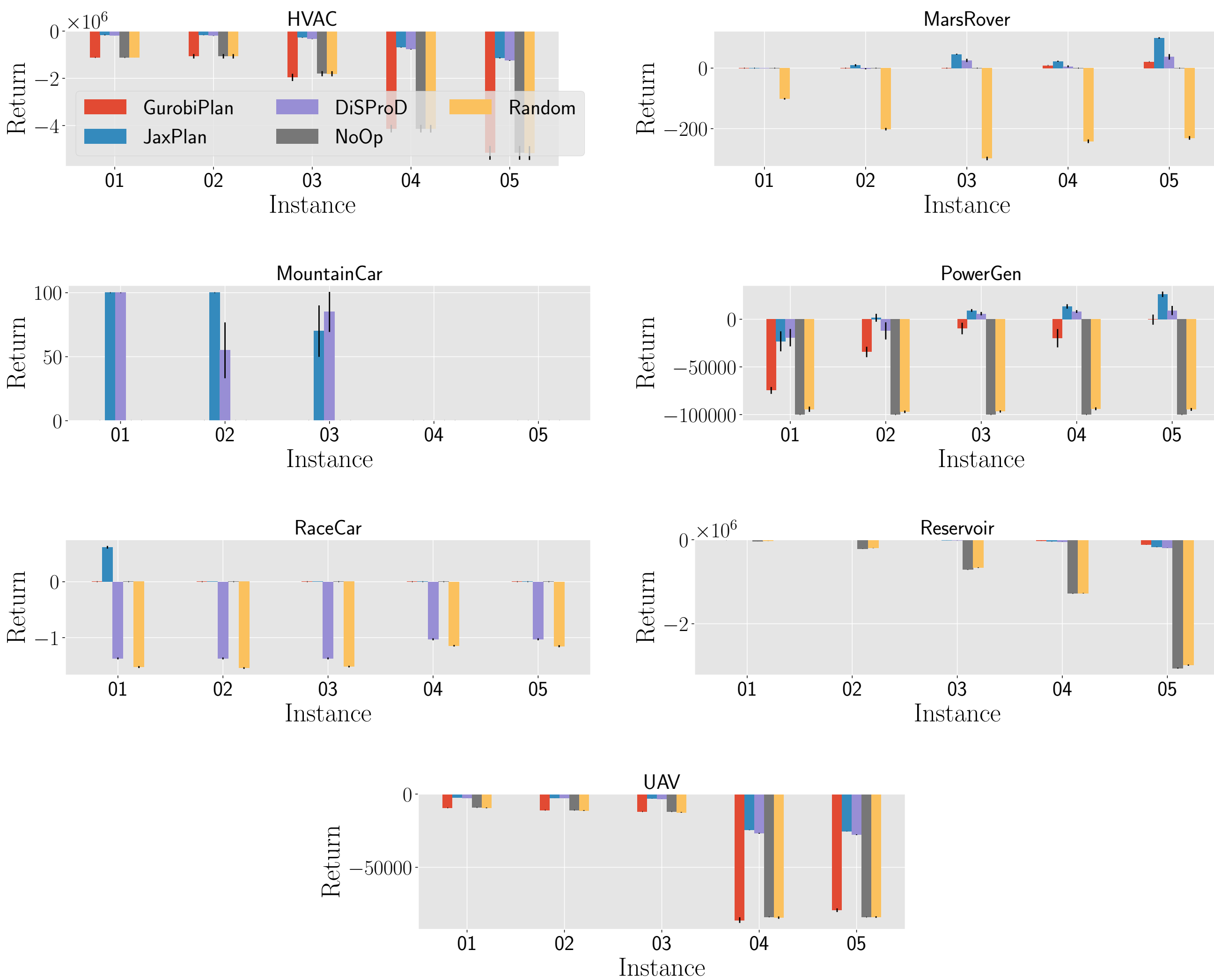

Figure 7: Unnormalized average return per domain and per instance on the IPPC 2023 domains, using a 1-second time budget per decision.

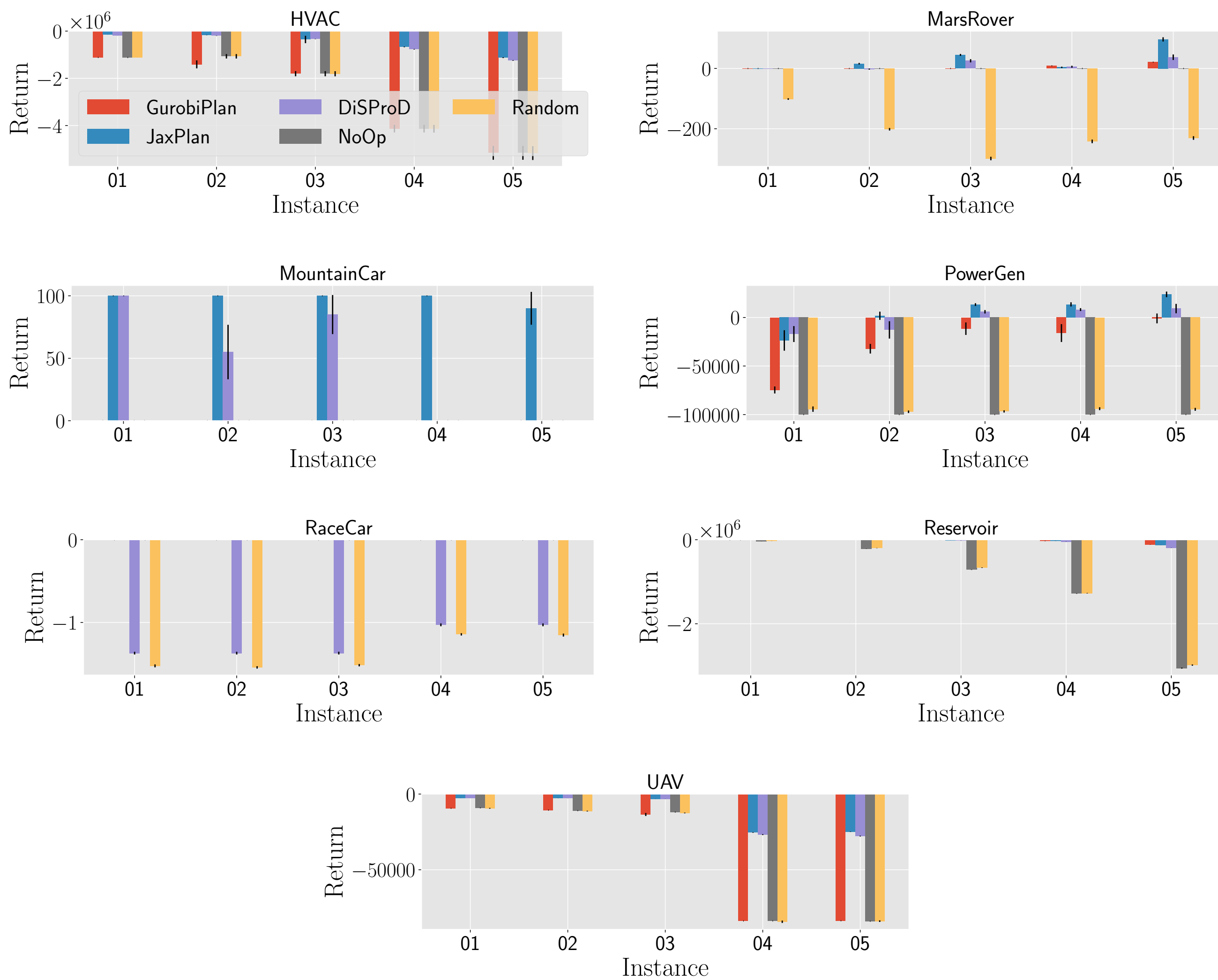

Figure 8: Unnormalized average return per domain and per instance on the IPPC 2023 domains, using a 3-second time budget per decision.

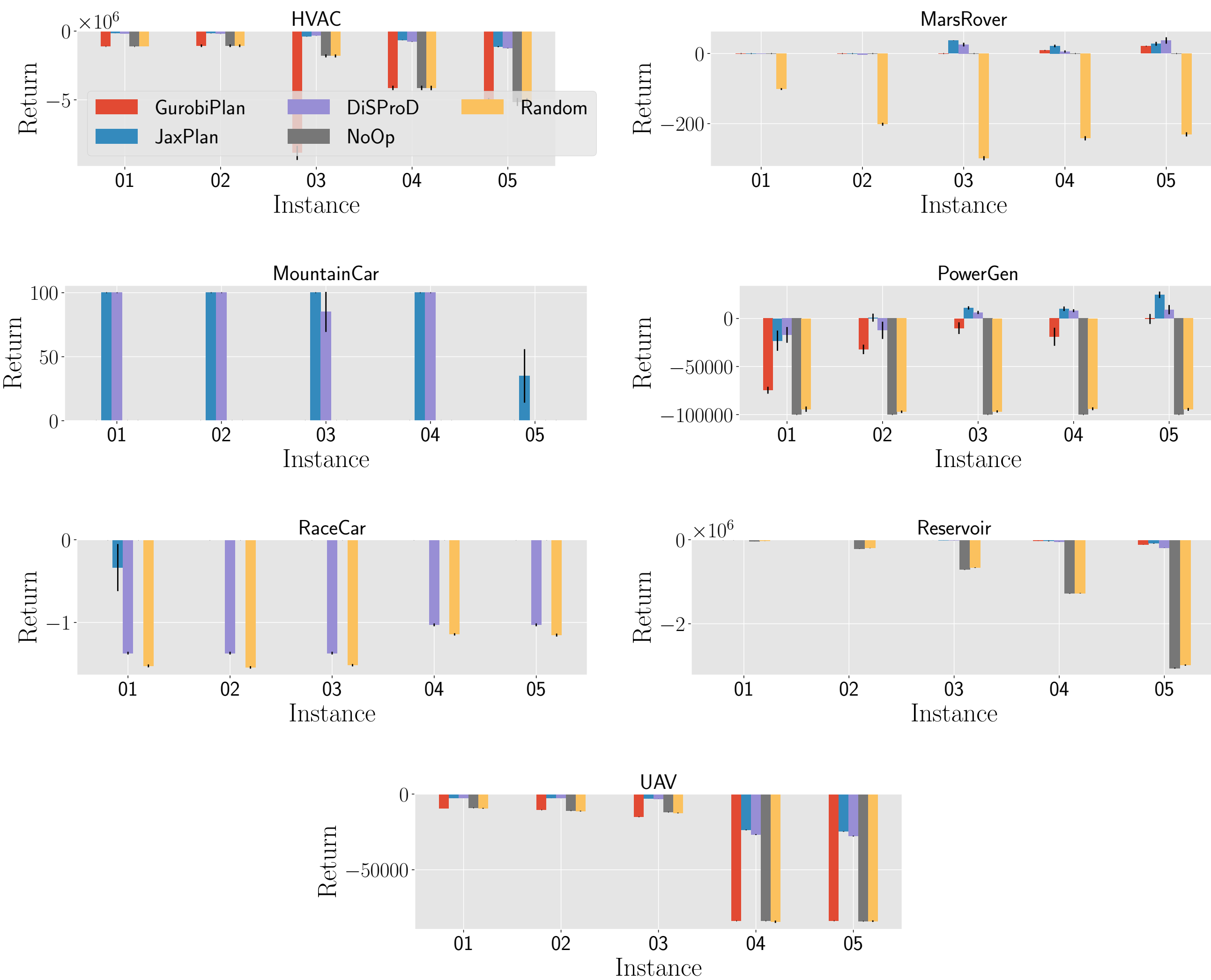

Figure 9: Unnormalized average return per domain and per instance on the IPPC 2023 domains, using a 5-second time budget per decision.