Boris Sedlak, Víctor Casamayor Pujol,
Praveen Kumar Donta, Schahram Dustdar

Intelligent Service Adaptations through Active Inference Agents

Problem Definition

-- Internet of Things (IoT) devices supply smart environments with sensory observations, e.g., video frames from traffic junctions

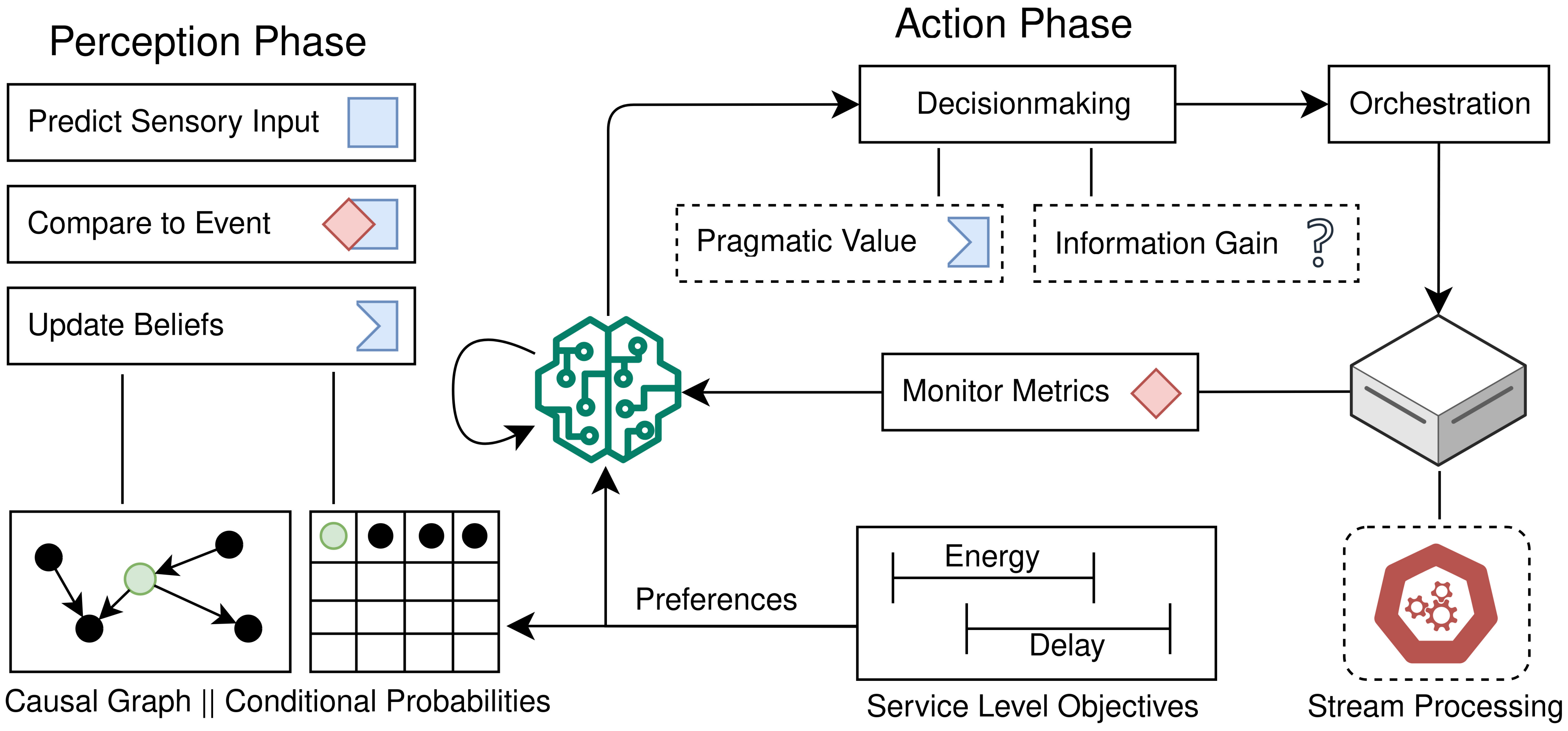
-- Data processing (e.g., visual analysis) occurs at nearby devices

-- Processing characterized by internal requirements (e.g., latency or quality) that must be continuously evaluated and ensured

-- However, logic to ensure requirements mostly confined to distant Cloud centers; no causal understanding how to configure process

Fig. 1: Processing video frames with Yolov8 to analyze traffic; computation occurs at nearby edge devices with low latency

Methodology



Perception Phase

Predict Sensory Input

Compare to Event

Update Beliefs

Causal Graph || Conditional Probabilities

Action Phase

Decisionmaking

Orchestration

Pragmatic Value

Information Gain

Monitor Metrics

Preferences

Energy

Delay

Service Level Objectives

Stream Processing

Fig. 2: Continuous observation and adaptation of processing services according to Service Level Objectives (SLOs)

-- Constrain processing through Service Level Objectives (SLOs)

-- AIF agents perceive their environment and enact on it

-- Perception phase predicts expected SLO fulfillment and adjusts the generative model

-- Action phase reconfigures local processing environment to minimize FE and fulfill SLOs

Markov Blankets

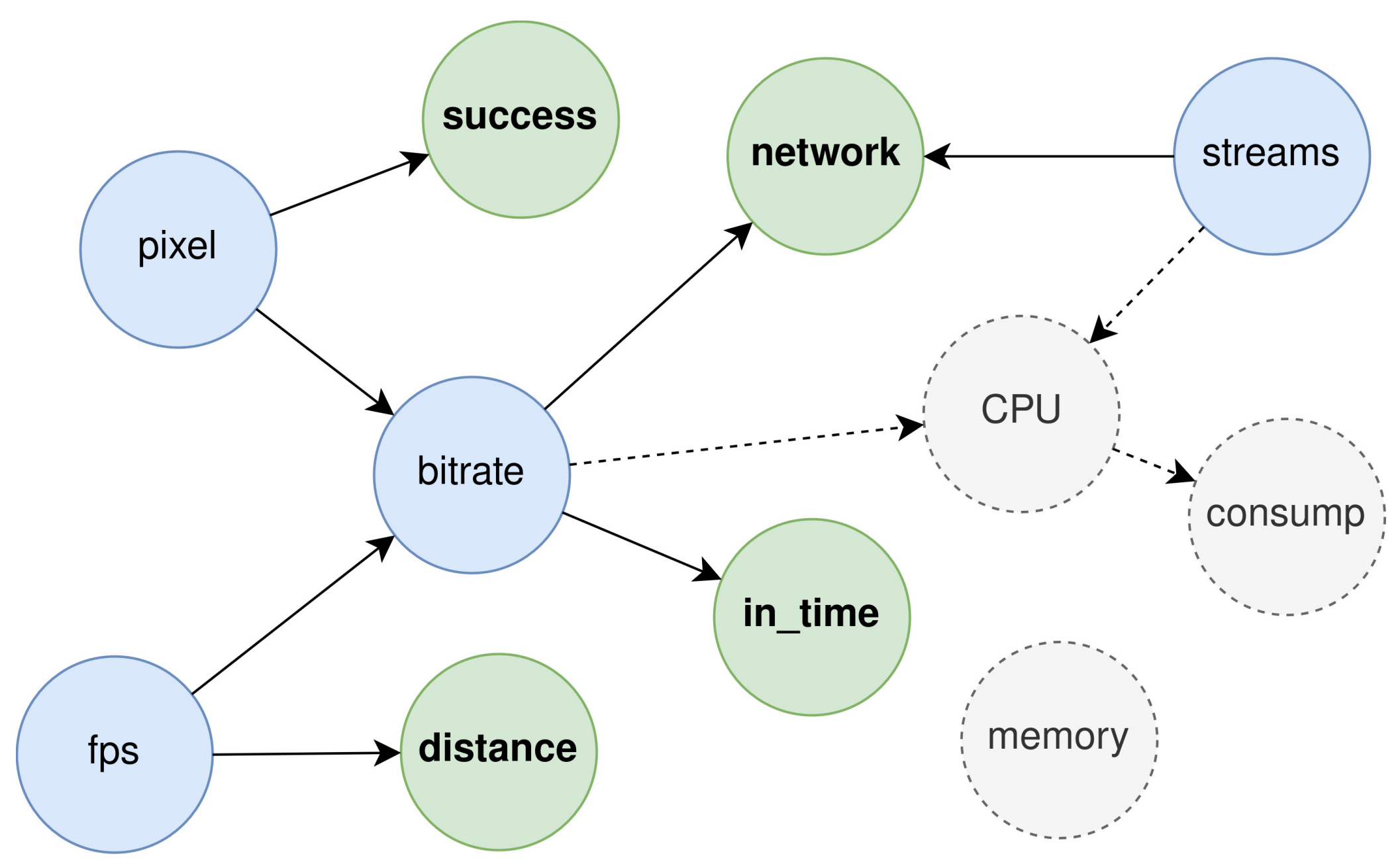


Fig. 3: Structural causal model trained by the AIF agent to interpret the system and infer how to ensure SLO fulfillment

Expresses the relations of processing metrics, including parameters (blue) that can be adjusted by the AIF agent

Target variables are constrained with SLOs (green) so that AIF agents can infer how to adjust dependent params

Variables that are not included in the MB of the SLOs (grey) are disregarded; speeding up training and inference

Pragmatic Value

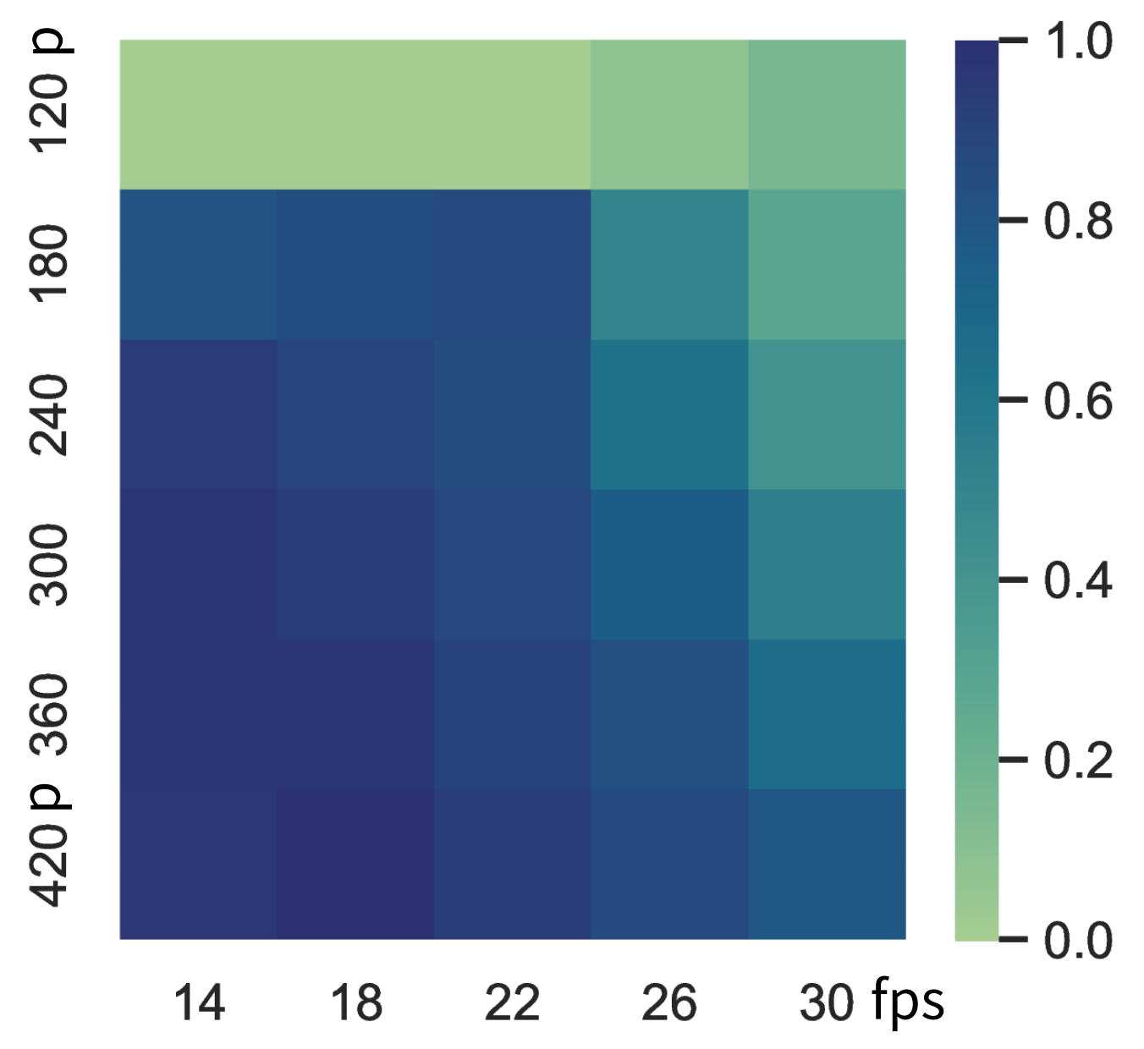


Fig. 4: PV heatmap for two configuration parameters (i.e., fps & pixel) that the AIF agent can actively adjust

Used to rate the SLO fulfillment of processing configs; AIF agents aim to fulfill these preferences

Pragmatic value balanced with information gain to choose configurations that minimize EFE

Agent operates in n-dimensional space bounded by the number of variables and discrete states

Local Adjustments

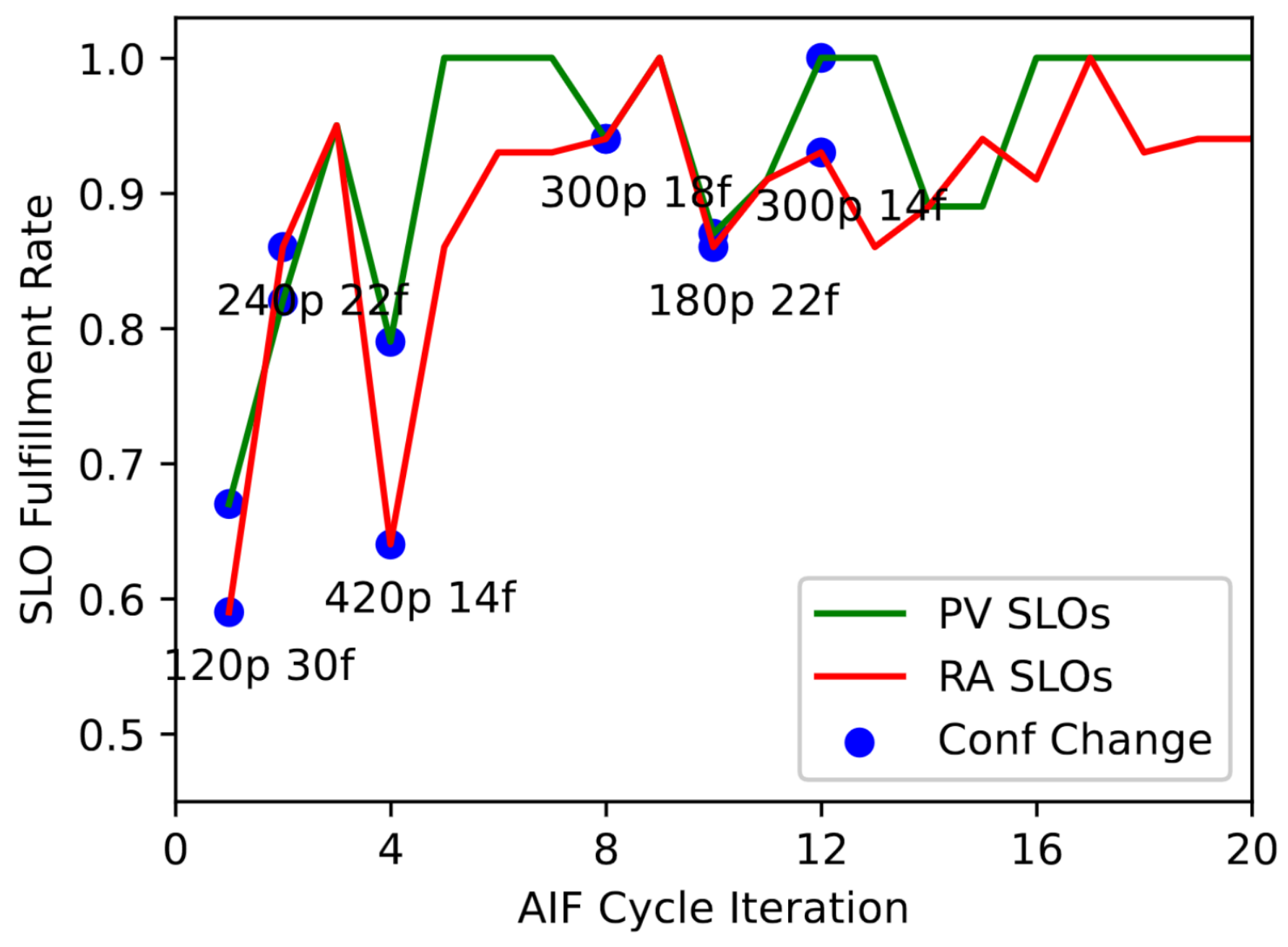


Fig. 5: AIF agents gradually find satisfying assignments

AIF agents gradually finds configurations that fulfill SLOs by exploring the solution space

Generative models can be exchanged between agents to speed up their SLO convergence

AIF agents are distributed over hierarchical processing tiers to supervise with finer granularity

Funded by
the European Union