

SIA: Symbolic Interpretability for Anticipatory Deep Reinforcement Learning in Network Control

MohammadErfan Jabbari^{*†}, Abhishek Dutttagupta^{*†}, Claudio Fiandrino^{*}, Leonardo Bonati[§],
Salvatore D'Oro[§], Michele Polese[§], Marco Fiore^{*}, and Tommaso Melodia[§]

^{*}IMDEA Networks Institute, Spain, Email: {name.surname}@networks.imdea.org

[§]Northeastern University, Boston, USA, Email: {l.bonati, s.doro, m.polese, t.melodia}@northeastern.edu

[†]Universidad Carlos III de Madrid, Spain

Abstract—Deep Reinforcement Learning (DRL) promises adaptive control for future mobile networks but conventional agents remain reactive: they act on past and current measurements and cannot leverage short-term forecasts of exogenous Key Performance Indicators (KPIs) such as bandwidth. Augmenting agents with predictions can overcome this temporal myopia, yet uptake in networking is scarce because forecast-aware agents act as closed-boxes; operators cannot tell whether predictions guide decisions or justify the added complexity. We propose SIA, the first interpreter that exposes in real time how forecast-augmented DRL agents operate. SIA fuses Symbolic AI abstractions with per-KPI Knowledge Graphs to produce explanations, and includes a new Influence Score (IS) metric. SIA achieves sub-millisecond speed, over $200\times$ faster than existing EXplainable Artificial Intelligence (XAI) methods. We evaluate SIA on three diverse networking use cases, uncovering hidden issues, including temporal misalignment in forecast integration and reward-design biases that trigger counter-productive policies. These insights enable targeted fixes: a redesigned agent achieves a 9% higher average bitrate in video streaming, and SIA's online Action-Refinement module improves RAN-slicing reward by 25% without retraining. By making anticipatory DRL transparent and tunable, SIA lowers the barrier to proactive control in next-generation mobile networks.

I. INTRODUCTION

Next-generation mobile networks promise significant performance improvements but must cope with growing traffic demands and rapidly changing network conditions [1], [2]. Deep Reinforcement Learning (DRL) is a promising approach for adaptive network control, with proven successes in resource allocation [3], [4], scheduling [5], [6], and parameter optimization [7], [8]. DRL agents learn, through trial and error, policies that maximize cumulative rewards by weighting both the immediate and long-term outcomes of their actions [9].

A fundamental limitation, however, stems from how DRL agents interact with their environment. Mobile networks feature two distinct classes of Key Performance Indicators (KPIs) with different relationships to agent actions [10], [11]. *Controllable KPIs* respond directly to an agent's decisions. For example, scheduling a user directly affects their transmitted data, and activating base station antennas impacts power consumption. Agents naturally learn to predict how their actions influence these controllable KPIs using temporal difference learning [12], [13]. In contrast, *exogenous KPIs* change regardless of the agent's actions [14]. Factors like channel conditions shifting with user mobility or available bandwidth fluctuating due to external network congestion [15] fall into this category. This leaves agents with a *temporal myopia* when dealing with

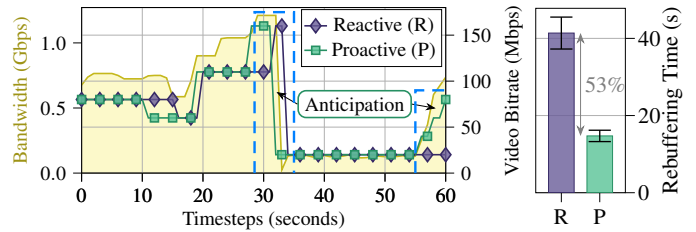


Fig. 1. Agents using future network bandwidth estimates achieve higher QoE by acting proactively.

exogenous KPIs [16]; they are blind to upcoming changes and can only react after they happen [17].

This reactive behavior leads to poor performance. In Adaptive Bitrate Streaming (ABR), for instance, an agent learns how bitrate choices affect buffer levels (a controllable KPI) but cannot anticipate sudden bandwidth drops (an exogenous KPI). Such drops cause rebuffering [18], which severely degrades user Quality of Experience (QoE) [19]–[21]. Recent approaches tackle this limitation by learning internal predictive models [22] or directly equipping agents with forecasts of future exogenous KPIs [23], [24]. As Figure 1 shows, a forecast-augmented agent can anticipate a bandwidth drop and proactively reduce its bitrate, reducing rebuffering time by 53%.

However, equipping agents with forecasts creates a new challenge. While performance improves, the lack of insight into agent behavior is a major barrier to real-world deployment [25], making it difficult to justify the added computational cost. This opacity raises critical questions: *Do agents truly use forecasts to inform their decisions, or do they rely on current observations mostly? When forecasts are used, how far ahead does an agent look, and does this horizon change with network dynamics? Finally, how do forecasts alter an agent's strategy, for example, by causing it to sacrifice short-term gains for long-term benefits?* Understanding these trade-offs is essential to align agent behavior with network operator goals [26].

These questions fall within the domain of EXplainable Artificial Intelligence (XAI), a sub-field of Artificial Intelligence (AI) that develops techniques to interpret model decisions. However, existing methods in this field are designed specifically for reactive agents and cannot separate the influence of current observations from the impact of future predictions [27], [28]. This critical diagnostic gap hinders the trust, debugging, and optimization of these forecast-augmented systems.

To close this gap, we present SIA (Symbolic Interpretability for Anticipatory DRL), a framework designed to reveal how

DRL agents exploit forecasts. Its primary goal is to transform opaque neural network decisions into human-interpretable explanations using Symbolic Artificial Intelligence (Symbolic AI) and per-KPI knowledge graphs. Additionally, SIA includes an optional Action Refinement module that uses forecasted values to improve an existing agent’s performance, eliminating the need for costly retraining. Our core contributions are:

- We design and implement SIA, a novel interpretation framework for anticipatory DRL agents. By using symbolic abstractions and scalable, per-KPI Knowledge Graphs (KGs), SIA uniquely isolates the influence of current observations from future predictions in real time.
- We introduce the Influence Score (IS), a computationally efficient local explanation metric derived from our symbolic framework. The IS is the first metric to quantify and disentangle the influence of a KPI’s current state from its predicted future trend.
- We demonstrate SIA’s practical impact across three diverse networking tasks: ABR streaming [15], massive Multiple-Input Multiple-Output (MIMO) scheduling [5], and Radio Access Network (RAN) slicing [29]. The insights from SIA guide an ABR agent redesign that improves average bitrate by 9%, while its Action Refinement module boosts the reward for a RAN-slicing agent by 25% without retraining.

Upon acceptance, we will make our artifacts available for reproducibility and to stimulate the research in the area further.

II. BACKGROUND AND RELATED WORK

A. Background on Anticipatory DRL

DRL agents learn optimal policies through dynamic interaction with an environment. At each timestep t , an agent in state $s_t \in \mathcal{S}$ selects action $a_t \in \mathcal{A}$ using its policy $\pi(a_t | s_t)$ and receives a reward r_t . The agent’s goal is to maximize the cumulative reward $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t]$, where γ is a discount factor [9]. The value function, $V^\pi(s) = \mathbb{E}[r_t + \gamma V^\pi(s_{t+1}) | s_t = s]$, provides long-term reasoning, helping agents anticipate how their actions affect *controllable* KPIs. The controllable part of the next state, s_{t+1}^c , is a function of the agent’s decision, $s_{t+1}^c = f(s_t, a_t)$, giving it direct influence over these outcomes.

This anticipation fails for *exogenous* KPIs, where the transition to the exogenous part of the next state, s_{t+1}^e , is independent of the agent’s actions ($s_{t+1}^e \perp a_t$). In these cases, DRL agents can only react to environmental changes. In video streaming, for example, an agent can learn to manage buffer levels (a controllable KPI) through bitrate selection, but it cannot preempt disruptions from a sudden drop in network bandwidth (an exogenous KPI).

To overcome this limitation, anticipatory DRL equips agents with forecasts of exogenous KPIs [30], [31]. A forecaster g predicts future states $\hat{s}_{t+i}^e = g(s_{t-L:t}^e)$ for $i \in [1, h]$, where L is the lookback window and h is the prediction horizon. The agent then uses an augmented state $\bar{s}_t = [s_t, \hat{s}_{t+1}^e, \dots, \hat{s}_{t+h}^e]$ to learn a proactive policy $\bar{\pi} : \bar{\mathcal{S}} \rightarrow \mathcal{A}$. For instance, LUMOS [24], a decision-tree-based throughput predictor, showed that augmenting a model-predictive control (MPC) agent with its forecasts resulted in a 19.2% QoE improvement

over baseline ABR algorithm. Similarly, congestion forecasting has been shown to enable 14% cellular capacity gains in load balancing [23]. Other techniques include learning predictive latent-space models [22] or redefining the state to include future predictions [32], [33].

B. Background on Symbolic AI

Symbolic AI supports interpretable reasoning by encoding knowledge in explicit, human-readable structures [34]. For anticipatory DRL, it can abstract raw numerical KPIs and actions into logical constructs that capture temporal relationships, such as state transitions and forecasted trends. This symbolic abstraction provides a clearer view of decision processes than analyzing the raw numerical KPIs the policy network operates on. First-Order Logic (FOL), a formal language for symbolic reasoning [35], uses components like:

- **Predicates:** Relations (e.g., *allocate* or *schedule*)
- **Constants:** Domain entities (e.g., $\{Low, High\}$)
- **Variables:** Dynamic properties (e.g., *bandwidth*)
- **Quantifiers:** Scoping operators (\forall, \exists)
- **Connectives:** Logical operators (\wedge, \vee, \neg)

This formalism allows policies to be interpreted intuitively. Consider the rule: “If latency exceeds 100 ms *and* packet loss exceeds 2%, switch to the low-latency routing path.” The corresponding FOL representation is:

$$\forall t [(highLatency(t) \wedge highPacketLoss(t)) \Rightarrow switchPath(lowLatency)]$$

- $highLatency(t)$: Latency $> 100ms$ at t
- $highPacketLoss(t)$: Loss rate $> 2\%$ at t
- $switchPath$: Action to change routing configuration

Raw measurements map directly to symbolic predicates, enabling both causal analysis and human-understandable rules [36], [37]. This transparency is essential for analyzing anticipatory DRL agents in networks.

C. Related Work

XAI reveals how AI models operate, helping build trust and enable debugging [25]. However, most existing XAI approaches focus on reactive agents and therefore fail to capture the temporal complexity inherent in forecast-augmented DRLs.

Post-hoc interpreters like Local Interpretable Model-agnostic Explanations (LIME) [27] and SHapely Additive exPlanations (SHAP) [28] attribute importance to each of an agent’s k input features. Other approaches examine attention mechanisms [38] or analyze variable importance through gradient-based methods [39]. However, these methods share common limitations: their utility in real-time network control is limited [40]. Model-agnostic methods are computationally prohibitive, with methods like KernelSHAP having an exponential complexity in k , restricting them to offline analysis. While more efficient, model-specific variants exist, all versions share a more fundamental flaw: they treat all inputs as a static set, unable to distinguish the influence of current observations from future predictions. Even DRL-specific tools fall short. METIS [41] uses decision trees that ignore temporal constraints and requires slow retraining

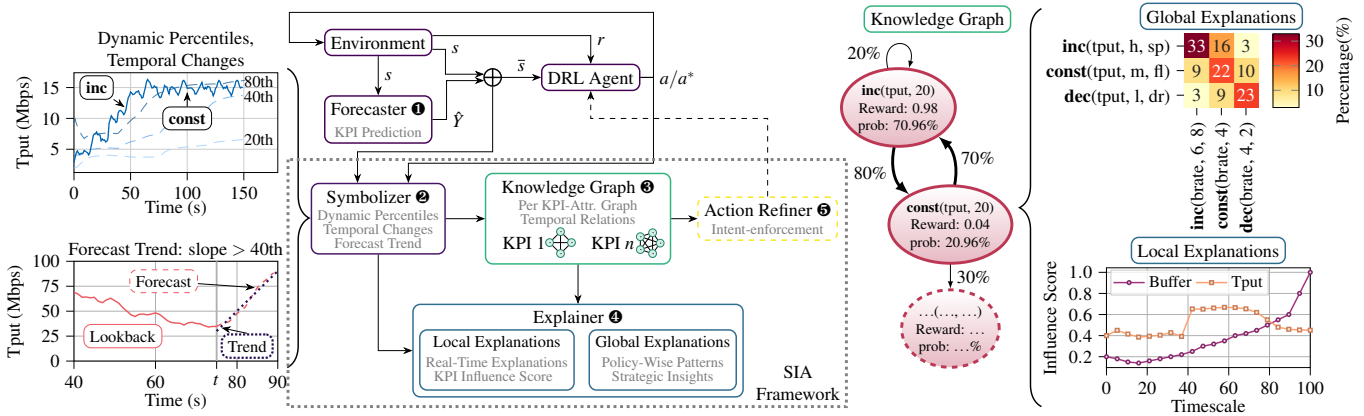


Fig. 2. Architecture of SIA, showing the core modules and information flow from raw KPIs to the explanations

for updates. EXPLORA [42] builds massive state-action graphs that consume excessive memory and cannot handle new actions.

Other symbolic or programmatic approaches also fall short, as methods for rule extraction (e.g., NSRL [43]), policy synthesis (e.g., PIRL [44]), or interpretation for reactive agents (e.g., SymbXRL [45]) all lack the temporal awareness to disentangle the influence of current observations from future predictions. Finally, Large Language Model (LLM)-based methods [46], [47], despite having natural-language outputs, cannot meet the near-real-time demands (on the order of tens of milliseconds) of network-control loops and risk producing hallucinated explanations.

Two gaps therefore persist for agents that use forecasting. *First*, existing tools lack the real-time performance needed for practical use in network-control operations. *Second*, they are fundamentally unable to distinguish the influence of current observations from future predictions, a critical requirement for interpreting anticipatory agents. SIA bridges these gaps by delivering efficient, temporally-grounded explanations for both proactive and reactive DRL agents. It uniquely reveals how forecasts reshape reactive decisions into proactive strategies through a flexible symbolic representation.

III. THE SIA FRAMEWORK

Figure 2 illustrates SIA’s modular architecture, which is designed to operate alongside an anticipatory DRL agent. Raw KPIs pass through five modules: the Forecaster (1) generates predictions; the Symbolizer (2) converts them into symbolic representations; the Knowledge Graph (KG) (3) builds structured state-action relationships; the Explainer (4) transforms these into human-readable insights; and finally, the optional Action Refiner (5) uses these insights to improve agent decisions in real time.

A. SIA Core Modules

1) *Forecaster (1)*: SIA uses the existing forecasting component from a standard anticipatory DRL system (see §II-A). SIA is agnostic to the forecast source; it uses the predictions (\hat{Y}_t) as provided, without modification. This module operates independently of other SIA components, and can accommodate any forecasting approach (e.g., statistical, machine learning, or

Algorithm 1: Symbolizer State Transformation Logic

Data: Current value v_t , previous value v_{t-1} , percentile sketch P , forecast series F , sensitivity θ

Result: Symbolic state s_{sym}

- 1 *Change detection*;
- 2 **if** $|v_t - v_{t-1}|/|v_{t-1}| > \theta$ **then**
- 3 $predicate \leftarrow \begin{cases} \text{inc} & \text{if } v_t > v_{t-1}; \\ \text{dec} & \text{otherwise} \end{cases}$;
- 4 **else**
- 5 $predicate \leftarrow \text{const}$;
- 6 **end**
- 7 *Dynamic categorization*;
- 8 $perc \leftarrow \text{PERCENTILERANK}(v_t, P)$;
- 9 $category \leftarrow \text{MAPTOBUCKET}(perc)$;
- 10 *Trend incorporation*;
- 11 **if** $F \neq \emptyset$ **then**
- 12 $slope \leftarrow \text{LINREG}(F).coef$;
- 13 $s_perc \leftarrow \text{PERCENTILERANK}(slope, SlopeHist)$;
- 14 $trend \leftarrow \text{MAPTOTREND}(s_perc)$;
- 15 **return** $\langle predicate, category, trend \rangle$;
- 16 **else**
- 17 **return** $\langle predicate, category \rangle$;
- 18 **end**

hybrid). As we demonstrate in §V, this modular design allows established forecasters to work effectively.

2) *Symbolizer (2)*: The Symbolizer converts raw KPIs into symbolic representations using FOL, handling both current observations and forecasts. It consists of the three-stage logic detailed in Algorithm 1, namely:

i) Change detection. It identifies the direction of immediate changes between data points, as increasing (**inc**), decreasing (**dec**), or stable (**const**) by applying a tunable sensitivity threshold, θ (default 5%), to filter out insignificant noise.

ii) Dynamic categorization. KPI values are mapped to a configurable number of percentile-based categories. For example, a five-category bucket scheme would span 20 percentiles each (e.g., **VeryLow** [0-20th], **Low** [20-40th], etc.). This process adapts dynamically as network conditions evolve, with streaming quantile estimators (P^2) maintaining the required percentiles with $O(1)$ [48], [49].

iii) Trend incorporation. For forecasts with horizon $h > 1$, the regression slope of the forecasted time-series is compared against a historical distribution of slopes and categorized

into a configurable set of trends, such as **Dropping** [<40th percentile], **Fluctuating** [40-60th], or **Spiking** [>60th].

Listing 1. Example of state symbolization

```
# Throughput KPI symbolization
current_tput = 15.7 # Mbps, previous_tput = 12.3 # Mbps
# Change detection:
change = (15.7 - 12.3) / 12.3 = 0.276 = 27.6%
threshold = 5%
27.6% > 5% => significant change detected
15.7 > 12.3 => increasing
predicate = inc
# Dynamic categorization:
percentiles = [7.2, 10.5, 13.8, 18.2] # P20,40,60,80
15.7 Mbps in percentile rank => 72nd percentile
Buckets: VeryLow[0-20], Low[20-40], Medium[40-60],
         High[60-80], VeryHigh[80-100]
72nd percentile => High category
# Trend incorporation:
forecast = [14.9, 13.2, 11.8, 9.7] # next 4 time steps
linear regression slope = -1.68 Mbps/step
slope percentile rank = 23rd percentile
Buckets: Dropping[0-40], Fluctuating[40-60],
         Spiking[60-100]
23rd percentile => Dropping trend
# Final symbolic state:
=> inc(tput, High, Dropping)
# Interpretation: "Throughput increased to High level but
# is forecasted to be Dropping"
```

The symbolization process also extends to the agent’s actions. The logic is configured based on the nature of the action space, allowing SIA to adapt to different agents. For discrete, ordered actions like bitrate selection, it can represent the change relative to the previous action (e.g., `dec(bitrate, 1200.0, 750.0)`). For categorical actions, such as selecting a scheduling policy, it can map to specific predicates (e.g., `toPolicy(WF)`). This unified symbolic representation of both states and actions is a fundamental input for all subsequent modules in the SIA framework.

Adapting SIA to a new agent involves tuning three elements: i) the *change-detection threshold*, θ , to a KPI’s volatility (e.g., a low θ for stable metrics like packet-loss versus a higher one for bursty signals like Channel State Information (CSI)); we found a consistent $\theta = 5\%$ effective for our use cases; ii) the *category scheme* to set the contextual resolution, where a given number of category buckets and their percentile boundaries are defined; and iii) the *action representation*, which mirrors the agent’s action space and can reuse standard KPI logic or be customized for richer insights. This streamlined configuration allowed the same Symbolizer to operate across the ABR, MIMO, and RAN-slicing agents in our evaluations (see § IV).

3) **Knowledge Graph** (⊕): The Knowledge Graph (KG) module addresses the state-explosion problem common in existing XAI methods (see §II-C). While prior work often encodes the joint state space in a monolithic structure with exponential complexity, SIA leverages KPI independence by maintaining a separate directed, attributed graph for each KPI. Nodes hold symbolic states (e.g., `inc(tput, High, Dropping)`), edges represent agent actions, and attributes capture empirical action probabilities, reward estimates, and transition counts.

This design yields four key benefits: (i) *Bounded complexity*, as each graph contains at most 45 nodes (3 predicates \times 5 categories \times 3 trends); (ii) *Efficient graph updates*, with per-timestep time complexity $O(k)$. The per-KPI factorization

Algorithm 2: Forecast Aware Action Refinement

Data: Current state s_t ; agent action a_t ; forecasts \hat{Y} ; knowledge graphs \mathcal{KG} ; threshold τ

Result: Refined action a_{refined}

```
1  $s_{\text{current}} \leftarrow \text{Symbolizer}(s_t)$ ;
2  $s_{\text{future}} \leftarrow \text{Symbolizer}(\hat{Y})$ ;
3  $a_{\text{refined}} \leftarrow a_t$ ; // default: keep the agent action
4 for each KPI  $k$  with a forecast do
5   if  $\text{edge}(s_{\text{current}}^k \rightarrow s_{\text{future}}^k) \in \mathcal{KG}_k$  then
6      $a_{\text{best}}^k \leftarrow \arg \max_a \bar{R}(a \mid s_{\text{current}}^k \rightarrow s_{\text{future}}^k)$ ;
7      $r_{\text{best}} \leftarrow \bar{R}(a_{\text{best}}^k \mid s_{\text{current}}^k \rightarrow s_{\text{future}}^k)$ ;
8      $r_{\text{agent}} \leftarrow \bar{R}(a_t \mid s_{\text{current}}^k)$  if available, else 0;
9     if  $r_{\text{best}} > r_{\text{agent}} + \tau$  then
10       $a_{\text{refined}} \leftarrow a_{\text{best}}^k$ ; // override with the
11      better action
12      break;
13   end
14 end
15 return  $a_{\text{refined}}$ ;
```

enables independent, parallelizable operations (see § V-C); (iii) *Causal querying*, for agent-strategy queries, such as determining the most likely actions when throughput is **High** now but the forecasted trend is **Dropping**; and (iv) *Efficient memory usage*, with space complexity $O(k, |S_{\text{sym}}|, |\mathcal{A}|)$, an exponential reduction from $O(|S|^k |\mathcal{A}|)$ in monolithic approaches, enabling real-time operation at scale.

4) **Explainer** (⊕): The Explainer module converts the symbolic representations and KG data into human-readable insights. It uses (1) the per-KPI KGs; (2) the agent’s current action; and (3) historical action distributions; to produce two complementary explanation types.

Local explanations. These quantify each KPI’s influence on an individual decision using the IS (see §III-B2). In contrast to perturbation-based analyses like SHAP or LIME, which must generate and analyze many variations of each data point, our method efficiently reuses pre-computed KG statistics. It runs in $O(k)$ time, returning an explanation in approximately 0.65 ms per decision (see §V-C).

Global explanations. These reveal policy-level patterns. First, mutual information analysis between KPIs and actions pinpoints which metrics most strongly steer decisions. Second, action-focused policy graphs visualize state-action sequences, exposing biases and strategies that traditional feature-importance tools often miss.

5) **Action Refiner** (⊕): The Action Refiner module augments a reactive agent with forecast awareness without retraining. As Algorithm 2 shows, the module first symbolizes the current and forecasted states (lines 1–2). It then queries the per-KPI KG for each forecasted KPI to find the historically best action, a_{best}^k , for the corresponding state transition (lines 4–6). If this proactive action’s expected reward exceeds the agent’s original choice by at least τ , the module overrides the decision (lines 8–10). The computational complexity of this process is $O(k_f \times |\mathcal{A}|)$, where k_f is the number of forecasted KPIs and $|\mathcal{A}|$ is the action-space size. We report the empirical latency measurements and a breakdown of per-component costs for this process in §V-C.

B. SIA's Explanations

SIA's unified pipeline generates both local and global explanations from its symbolic representations, a key distinction from other XAI tools. Interpreters like LIME provide only local scores, while others like METIS and EXPLORA offer only global summaries. Although SHAP can produce both, it does so by aggregating numerous computationally expensive local scores. SIA, in contrast, derives both explanations efficiently from its KG structure, without altering the agent.

Notation. Let \mathcal{K} be the set of all KPIs. For a given KPI $k \in \mathcal{K}$, its symbolic state is $s_k \in \mathcal{S}_k$. The agent selects an action $a_t \in \mathcal{A}$ from the set of all possible actions at time t .

1) *Global Explanations:* Global explanations reveal the agent's overall policy through two complementary techniques.

Mutual-information analysis quantifies the statistical dependence between each symbolic KPI and the action distribution:

$$MI(k; a) = \sum_{s_k \in \mathcal{S}_k} \sum_{a \in \mathcal{A}} p(s_k, a) \log \frac{p(s_k, a)}{p(s_k) p(a)}, \quad k \in \mathcal{K}. \quad (1)$$

Here, $p(s_k, a)$ is the joint probability, while $p(s_k)$ and $p(a)$ are the marginals. Ideal for symbolic data, this metric captures nonlinear relationships without distributional assumptions or discretization and has low $O(|\mathcal{S}_k| |\mathcal{A}|)$ complexity.

Action-focused policy graphs visualize decision-making logic by representing actions as nodes and state transitions as directed edges. As shown in §V-A1, these graphs expose hidden strategies and biases by annotating transitions with metrics like frequency and average reward.

2) *Local Explanations: The Influence Score:* For real-time decision analysis, we introduce the Influence Score (IS) to quantify each KPI's contribution to a specific action. This score is derived from the per-KPI KG in four steps:

I) **Extract Conditional Probabilities:** For each KPI k , compute its conditional action distribution, $P_k(a|s_k) = \frac{\text{count}(a, s_k)}{\text{count}(s_k)}$, where $\text{count}(a, s_k)$ is the historical count derived from node and edge annotations in the KPI's KG.

II) **Establish a Baseline:** Compute a baseline action distribution, $P_0(a) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} P_k(a|s_k)$, which serves as the average action probability across all KPIs.

III) **Calculate Information Contribution:** Measure the influence of a KPI using its Kullback-Leibler (KL)-divergence from the baseline ($D_{\text{KL}}(P_k || P_0)$), which quantifies how much the KPI's state reduces uncertainty about the agent's action.

IV) **Apply Alignment Weighting:** Weight the information contribution by an alignment function, $\delta(a_t, a_k^*)$, which filters the influence based on how well the agent's chosen action a_t aligns with the KPI's most likely action, $a_k^* = \arg \max_a P_k(a|s_k)$.

Conceptually, a large divergence means that the KPI provides a distinctive signal relative to average behavior. The complete formulation combines these steps:

$$IS_k = D_{\text{KL}}(P_k || P_0) \times \delta(a_t, a_k^*). \quad (2)$$

The alignment function δ adapts the IS to action types. For continuous actions (e.g., power allocation, bitrate selection), we use $\delta_{\text{decay}}(a_t, a_k^*) = \exp(-d(a_t, a_k^*)^2 / 2\sigma^2)$, where $d(\cdot, \cdot)$ measures action distance and σ controls sensitivity. This approach achieves $O(k)$ complexity for bounded action spaces, enabling

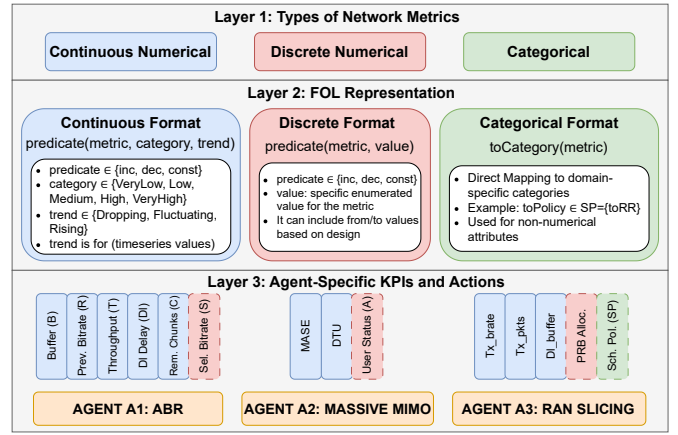


Fig. 3. Symbolic FOL representations for the KPIs of our evaluation agents. real-time explanations that capture temporal dependencies, a capability absent in static methods like SHAP or LIME.

IV. USE CASES AND EXPERIMENTAL SETUP

To evaluate SIA, we test it across three distinct use cases spanning different network layers: ABR, Massive MIMO scheduling, and RAN slicing. This diversity demonstrates SIA's flexibility. In this section, we detail each use case, including the DRL agent's design, state, reward function, and the forecasting model chosen for the task. While the agents are heterogeneous, SIA unifies them by converting their raw KPIs into a consistent symbolic representation based on FOL.

Our evaluation is guided by three research questions (RQs):

- RQ_1 : How does SIA explain a forecast-aware agent's policy compared to a reactive one, and how do these explanations improve agent performance?
- RQ_2 : How does SIA's IS compare against established interpreters like SHAP and LIME?
- RQ_3 : How effective is SIA's Action Refinement module at boosting performance without retraining the agent?

Figure 3 illustrates the general FOL representations used across all agents. Continuous KPIs are mapped to `predicate(metric, category, trend)`, discrete values to `predicate(metric, value)`, and categorical variables to `toCategory(metric)`.

A. A1: ABR for Video Streaming

For our first use case, we adapt the Pensieve DRL agent [15] to an ABR scenario that aims to maximize video quality while minimizing stalling. The agent's state includes buffer levels, the last chosen bitrate, download delay and throughput (Thput) histories (e.g., symbolically as `inc(tput, High, Dropping)`), and the number of chunks remaining. The agent's action is to select the next bitrate from a discrete set, while its reward function balances the competing goals of high bitrate, low rebuffering, and smooth quality transitions:

$$R_t = \underbrace{\sum_{n=1}^N q(R_n)}_{\text{Bitrate Utility}} - \underbrace{\mu \sum_{n=1}^N T_n}_{\text{Rebuffering}} - \underbrace{\sum_{n=1}^{N-1} |q(R_{n+1}) - q(R_n)|}_{\text{Quality Variation(Smoothness)}}. \quad (3)$$

To give the agent foresight, we equip it with forecasts of exogenous KPIs, selecting the optimal model for each task. For the A1-P agent, which performs *univariate* forecasting of bandwidth (Bwd), we employ PatchTST [50], a State-of-the-Art (SOTA) Transformer model recognized for its superior performance in long-horizon univariate forecasting. In contrast, the A1-P+SIA agent is redesigned using insights from SIA (see Section V-A3) and requires a simultaneous, *multivariate* forecasting of Bwd, Thput, and delay. For this, we use a lightweight Multi-layer Perceptron (MLP) paired with Reversible Instance Normalization (RevIN) [51]. This choice is data-driven: our evaluation shows the MLP-RevIN model achieves 20% Mean Absolute Percentage Error (MAPE), significantly outperforming alternatives from literature like Lumos (80% MAPE) [24] and Xatu (30% MAPE) [52]. All agents are tested on datasets from [53], [54]; Table I summarizes their configurations.

B. A2: User Scheduling in Massive MIMO

In this use case, we use the Massive MIMO scheduler from [5]. The agent’s action is to decide for each of the seven users whether to schedule them or not. To interpret the policy’s effect on interference, SIA abstracts these collective actions into a higher-level symbolic representation, `Alloc(User Group, Percentage)`, which captures the percentage of resources given to each user group. The agent’s state includes per-user KPIs like Maximum Available Spectral Efficiency (MASE), Data Transmitted of User (DTU), and their assigned User Group Label (G). Agent’s reward signal balances system throughput and Jain-Fairness-Index (JFI) [55], as

$$R_t = \underbrace{\beta \gamma_t^{\text{total}}}_{\text{System Data Transmission}} + \underbrace{(1 - \beta) \text{JFI}_t}_{\text{Data Fairness}}, \quad \beta = 0.5. \quad (4)$$

To predict the exogenous MASE KPI for the forecast-augmented agent (A2-P), we again use PatchTST [50], as this is another univariate forecasting task. By using these forecasts, the anticipatory agent learns an effective policy 24 episodes faster than the reactive baseline (A2-R).

C. A3: RAN Slicing and Policy Scheduling

Our third use case, from [29], involves joint RAN slicing and scheduling for three traffic slices, running as an xApp on the O-RAN near-RT RIC and emulated in Colosseum [56]. State KPIs include: `tx_bitrate` (e.g., `inc(tx_brate, High)`), `tx_packets`, and `DWL_buffer_size`, while actions control Physical Resource Block (PRB) allocation and scheduling policies for each slice. We chose this complex agent to test SIA’s action refinement ability, as retraining is impractical due to the large action space. To forecast the exogenous KPIs (`tx_bitrate` and `DWL_buffer_size`), we use a multivariate MLP architecture paired with RevIN.

The Symbolizer’s parameters were chosen to balance detail and stability. The change-detection threshold, θ , is configurable for each KPI to handle different numerical scales; we found $\theta = 5\%$ effective for our use cases. The number of percentile buckets is also configurable and chosen for desired granularity: five for primary KPIs and three for forecasted trends.

TABLE I
AGENT CONFIGURATIONS. h DENOTES THE FORECAST HORIZON

Agent	Config	Forecasted KPIs	Forecast Model	h
<i>A1: ABR Streaming [15]</i>				
A1-R	Reactive	None	N/A	0
A1-P	Proactive	Bwd	PatchTST	4
A1-P+SIA	SIA-guided design	Bwd, Thput, Delay	MLP-RevIN	4
<i>A2: MIMO Scheduling [5]</i>				
A2-R	Reactive	None	N/A	0
A2-P	Proactive	MASE	PatchTST	4
A2-R+SIA	Action-Refined	MASE	PatchTST	4
<i>A3: RAN Slicing [29]</i>				
A3-R	Reactive	None	N/A	0
A3-R+SIA	Action-Refined	tx_brate, dl_buffer	MLP-RevIN	4

V. EVALUATION RESULTS

This section evaluates SIA using the agent configurations from §IV. We use agents A1 and A2 to address RQ_1 and RQ_2 , and agents A2 and A3 for RQ_3 . We compare SIA’s explanations to SOTA interpreters and demonstrate how it improves agent performance without retraining.

A. Analysis of Global Explanations

To address RQ_1 , we demonstrate how SIA’s global explanations reveal anticipatory policies, uncover design flaws, and guide performance improvements using policy graphs and Mutual Information (MI) analysis.

1) *Agent Policy Analysis*: SIA’s policy graphs visualize an agent’s strategy by mapping symbolic actions to nodes and transitions to edges. Figures 4 and 5 compare the policies of the reactive (A1-R) and proactive (A1-P) agents in the ABR task [54]. The graphs reveal key behavioral differences:

- A1-P maintains a higher steady-state quality, spending over 58% of its time at 1200 kbps, while A1-R defaults to a lower 750 kbps baseline for 43% of its time.
- A1-R uses incremental transitions (750 \rightarrow 1850 \rightarrow 2850 kbps) before retreating, whereas A1-P makes decisive, forecast-guided shifts. This stability is evident in its stronger self-loops (e.g., an 80.65% probability of staying at 2850 kbps versus A1-R’s 62.5%).
- A1-P leverages network forecasts to sustain the maximum bitrate for extended periods, spending nearly four times longer at peak quality than A1-R (11.01% vs. 2.84%).
- A1-P employs different recovery paths, such as dropping to 300 kbps vs 750 kbps for A1-R to rebuild its buffer; this forward-looking strategy is missing in the reactive A1-R.

These proactive strategies yield measurable gains: Table II shows A1-P achieves a 1.7% higher reward and a 1.8% higher bitrate. Unlike traditional interpreters like Metis [41], which produce large decision trees (3200–5000 nodes) with long retraining cycles (~ 30 min), SIA’s bounded policy graphs offer an interpretable, real-time visualization of the agent’s temporal action policies that are hidden by previous methods.

2) *Revealing Design Biases*: SIA’s policy graphs can also expose subtle flaws in an agent’s reward design. Figure 6 shows the policy for the reactive MIMO scheduling agent (A2-R), which, by design, should aim to schedule users from Group 0 (group of users with the best channel quality).

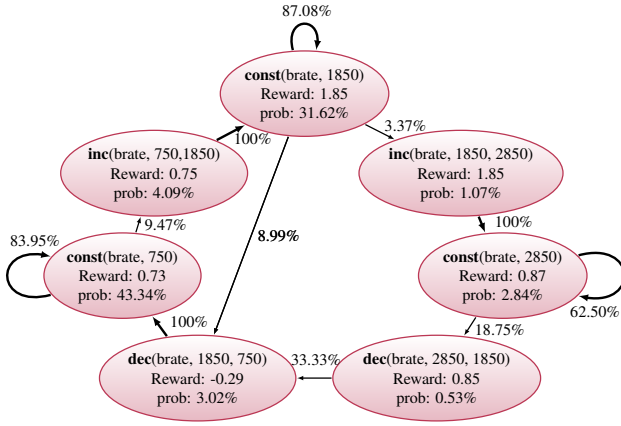


Fig. 4. Policy graph of the reactive agent (A1-R).

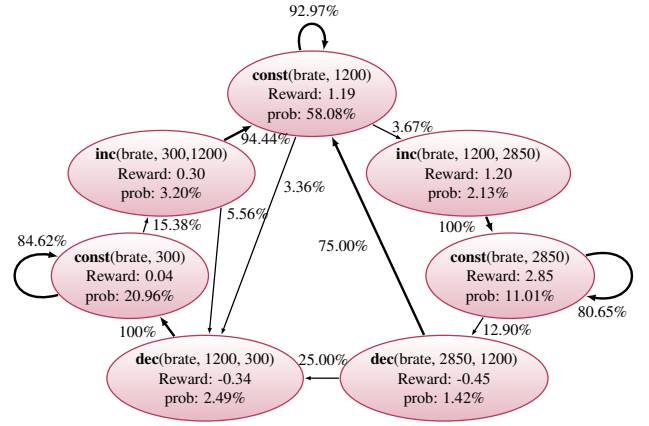


Fig. 5. Policy graph of the proactive agent (A1-P).

TABLE II
PERFORMANCE OF PROACTIVE AGENT (A1-P) AND SIA GUIDED PROACTIVE (A1-P+SIA) COMPARED TO THE REACTIVE BASELINE (A1-R).

Agent	Reward Increase	Bitrate Increase	P-value
A1-P	+1.7%	+1.8%	< 0.01
A1-P+SIA	+3.4%	+9.0%	< 0.01

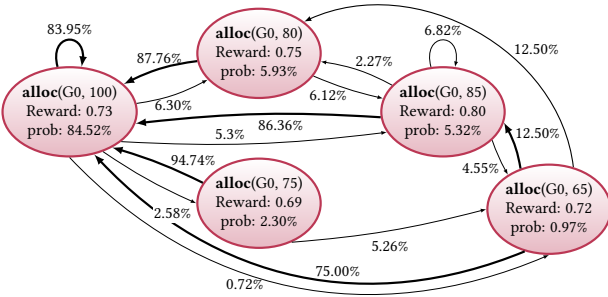


Fig. 6. Policy graph of the reactive agent (A2-R) for Group 0 users.

The graph reveals a counterintuitive behavior: the agent’s most probable action, allocating 100% of resources to Group 0, yields a lower reward (0.73) than a partial allocation (0.80 for an 85% allocation). This result is unexpected, as scheduling other groups should introduce interference. The root cause lies in the reward function (Eq. (4)): the JFI is based on cumulative DTU and quickly saturates near 1.0. As these values grow large, the index becomes insensitive to momentary allocation differences, causing the agent to prioritize short-term throughput at the cost of system-wide interference.

This apparent contradiction (why the agent favors an action with lower reward) is explained by opportunity. Multiple user groups are only present in 17% of timesteps. The policy flaw is thus revealed during these critical steps, where the agent diverts resources from Group 0 in 72% of cases, causing interference. Traditional feature-importance methods would miss this policy-level flaw, which SIA exposes by showing how a faulty reward design can lead to a counter-productive policy.

3) *Quantifying Forecast Impact Through MI*: We use MI analysis to reveal a critical design flaw in the proactive agent (A1-P) arising from naively integrating forecasts with inconsistent temporal structures. Figure 7 shows the MI values for our three ABR agent variants.

The analysis of static KPIs shows consistent influence across all agents (Fig. 7a), but a critical misalignment is exposed when examining the temporal KPIs. The issue is that the agent receives correlated inputs with inconsistent time horizons. For agent A1-P, both bandwidth and throughput are fed as vectors of 8 timesteps. However, the bandwidth input is structured around its forecast, containing 3 past, 1 current, and 4 future values. In contrast, the throughput input retains its original reactive structure from agent A1-R, containing only 7 past values and 1 current value, with no forecast. Figure 7(b-d) shows the consequence: the agent learns from bandwidth that the strongest predictive signal comes from the values near its horizon ($t = 1$ and $t = 2$). It then incorrectly applies this heuristic to the throughput data, learning to focus on timesteps just before the horizon. Since the throughput horizon is at $t = 0$, this causes the MI peak to incorrectly shift to $t = -2$.

This suggests *the agent learns relative temporal patterns rather than understanding time inherently*. Guided by this insight, we designed A1-P+SIA with a *uniform* temporal structure (7 past, 1 current, and 4 future values) across all correlated KPIs. This SIA-guided redesign corrected the misalignment and, as shown in Table II, produced substantial gains, achieving a 9% higher bitrate and a 3.4% higher overall reward than the baseline. This uncovers a key principle for anticipatory DRL: *correlated temporal inputs must be presented with consistent formatting to prevent policy misinterpretation*.

B. Analyses of Local Explanations

To answer RQ_2 , this section demonstrates SIA’s ability to generate real-time, temporally-aware local explanations. We show how the IS enables operational monitoring and reveals strategic adaptations that are invisible to interpreters like SHAP and LIME.

1) *Real-Time Monitoring with Influence Scores*: Figure 8 demonstrates SIA’s real-time monitoring of the reactive MIMO agent, A2-R. The plots track the Delta Data Transmitted for User (DDTU) KPI, its corresponding IS, and the agent’s resource allocation for Group 0, revealing a clear causal chain. For instance, after an allocation drop at timestep 9 that causes the DDTU to fall, its IS spikes to nearly 0.8 upon entering a **VeryLow** category at timestep 11. This drop flags the DDTU

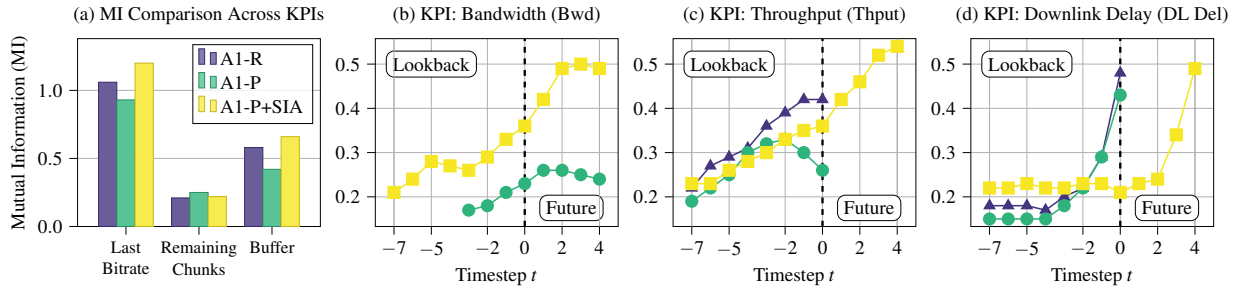


Fig. 7. MI analysis between input KPIs and agent actions for the reactive (A1-R), proactive (A1-P), and SIA-guided (A1-P+SIA) agents.

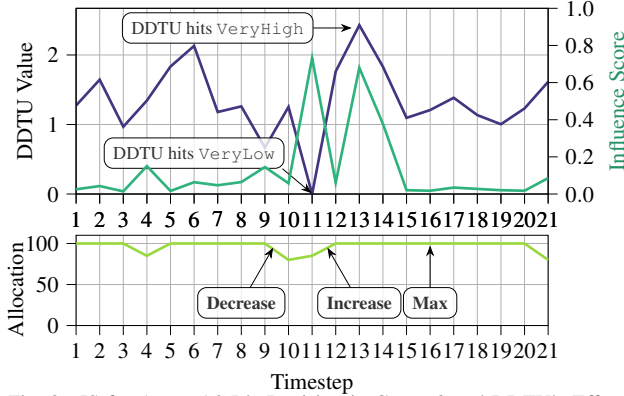


Fig. 8. IS for Agent A2-R's Decision in Group 0 and DDTU's Effect

as the dominant factor, prompting an immediate increase in resource allocation for Group 0. This causal analysis, explaining both when and why an agent reacts, is generated in just 0.65 ms mean latency. In contrast, obtaining a single explanation from SHAP or LIME on the same task takes 141 ms and 159 ms, respectively (see §V-C), enabling true real-time observability.

2) *Revealing Strategic Adaptation Over Time*: Beyond explaining single decisions, the IS can reveal how an agent's strategy dynamically evolves over an entire episode. Figure 9 tracks the IS (y-axis) over the normalized episode duration (x-axis) for three key static KPIs influencing the reactive agent's (A1-R) streaming strategy on the 5G dataset [53].

The analysis reveals a sophisticated, multi-faceted policy. In the initial phase (0–80%), the agent prioritizes stable playback, evidenced by the dominance of the Last Bitrate KPI ($IS \approx 0.3$), reflecting the smoothness term in the reward function (3). Concurrently, the Buffer influence peaks at startup and again near the 60% mark, showing an intermittent focus on refilling buffer to prevent stalls, reflecting the rebuffering term (3).

A dramatic shift occurs in the final quality maximization phase (80–100%), where the influence of Remaining Chunks triples to become the dominant KPI ($IS \approx 0.45$). With few chunks left, the rebuffering penalty risk diminishes, allowing the agent to aggressively pursue a higher bitrate. This adaptation suggests how network operators can align resources with the agent's policy: provide stable bandwidth initially and then offer bursty, high-throughput resources toward the episode's end.

3) *Distinguishing Temporal from Static Features*: Figure 10 highlights SIA's unique temporal awareness compared to SHAP and LIME. While all methods produce similar importance distributions for static features (left panel), a clear distinction emerges for temporal ones (right panel). Here, SIA's

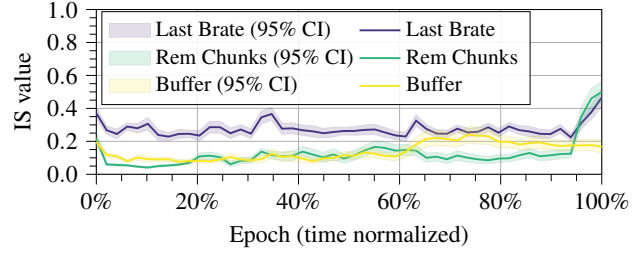


Fig. 9. IS for agent A1-R, showing KPIs importance

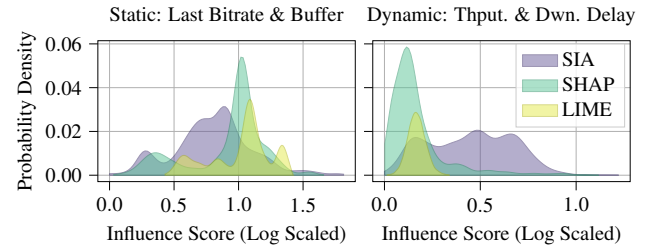


Fig. 10. Comparing IS of different interpreters

IS distribution is distinctly bimodal as a direct result of its symbolic encoding and the mechanics of its formula (2).

This bimodality arises because the symbolic state s_k for a temporal KPI is a composite of category and trend components. The IS, driven by the D_{KL} term, yields a high score in two different scenarios. The first peak represents decisions driven by the KPI's current *category* (e.g., throughput is *VeryHigh*), while the second peak reflects decisions driven by its future *trend* (e.g., throughput is *Dropping*), which can contradict the current category.

For instance, SIA can differentiate whether a bitrate reduction stems from a *Low* current throughput (a category-driven decision) or from a *Dropping* trend despite a *High* current throughput (a trend-driven decision). In contrast, SHAP and LIME average the importance across all temporal data points, blending these distinct contexts into a single, less informative peak. This ability to disentangle the influence of current state versus future predictions is a critical capability for explaining anticipatory agents, directly addressing RQ_2 .

C. Performance and Scalability Analysis

SIA is designed for real-time operation. We report worst-case performance, measured on our most complex agent (A1-P), which has the highest number of input KPIs. The core pipeline for a local explanation, including symbolizing inputs, updating the KG, and calculating the IS, completes in a mean time of just 0.65 ms. Table III details the sub-millisecond latency of each component. This makes the full process over $200\times$ faster

TABLE III
LATENCY PER SIA COMPONENT AND TIMESTEP OF THE A1-P AGENT.

Component	Mean Latency	Std. Dev.
Symbolizer	0.099 ms	0.009 ms
Knowledge Graph Update	0.265 ms	0.055 ms
Influence Score	0.280 ms	0.030 ms
Action Refinement	0.330 ms	0.040 ms
Global Explanation	0.610 ms	0.080 ms

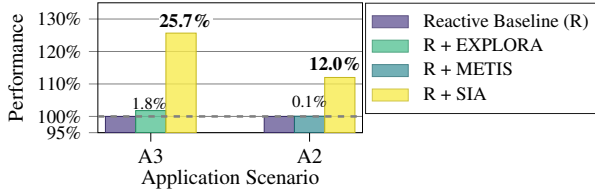


Fig. 11. Performance gains via SIA’s action refinement.

than traditional interpreters like SHAP (141 ms) and LIME (159 ms) on the same task.

This efficiency stems from SIA’s architecture. Because its complexity scales linearly ($O(k)$) with the number of KPIs, unlike the exponential complexity of competing methods, its performance is predictable. These worst-case latencies ensure SIA operates comfortably within the one-second control loop of the O-RAN near-RT RIC constraints.

D. Improving Agent Performance via Action Refinement

To address RQ_3 , we evaluate SIA’s Action Refinement module, which gives reactive agents forecast-awareness without retraining. This module yields considerable gains, as shown in Figure 11. Applying it to the RAN slicing agent (A3-R+SIA) achieves a 25.7% cumulative reward boost, substantially outperforming EXPLORA’s 1.8% gain. Similarly, the refined MIMO agent (A2-R+SIA) provides a 12.0% reward increase, far exceeding the 0.1% from Metis.

These gains are possible because SIA’s bounded symbolic state space and efficient KG queries enable rapidly identifying the historically optimal action for a forecasted state transition. The refiner overrides the agent’s decision if this action’s expected reward exceeds the original by a configurable threshold, τ (set to 3% in our experiments), as detailed in Algorithm 2. This override check completes in just 0.33 ± 0.04 ms (see §V-C), meeting the O-RAN near-RT RIC’s timing budget and offering a practical path to proactive control in production networks where retraining is often prohibitive.

These gains are robust to moderate forecast inaccuracies. Because the refiner operates on symbolic categories, performance is robust to forecast errors unless they are large enough to shift a KPI’s value across a percentile boundary, a resilience mechanism inherent to SIA’s design.

VI. DISCUSSION AND LIMITATIONS

Discussion. SIA offers a practical methodology for interpreting anticipatory DRL. Its utility is shown by its generalizability across diverse use cases and its ability to generate real-time explanations using our novel IS metric. Its use of per-KPI KGs avoids the state-explosion problem of monolithic approaches, ensuring its overhead and memory footprint remain scalable for production. We show SIA’s insights are actionable: they expose

actual design flaws, such as the MIMO agent’s reward bias, and guide an ABR agent redesign that increases bitrate by 9%. Moreover, its Action Refinement module boosts RAN-slicing agent’s reward by 25% without retraining.

Limitations. Through an ablation analysis of SIA’s operating boundaries, we identified three key limitations and deployment considerations: (i) both SIA’s explanations and the Action Refiner’s suggestions (see §V-D) are affected by forecast accuracy. However, SIA’s symbolic foundation provides resilience, as performance is largely unaffected unless forecast errors are large enough to push a KPI across a category boundary. While this implies higher sensitivity for stable metrics with narrow percentile bands, such KPIs are typically forecasted with high accuracy, which mitigates the risk. (ii) SIA exhibits a cold-start phase because the per-KPI KGs are initially sparse. In practice, this can be mitigated by pre-populating the KGs from offline traces. (iii) The framework is sensitive to the Symbolizer’s configuration. We found the number of categories is a critical hyperparameter: using fewer than three produces overly generic insights, while more than seven prolongs the cold-start period. An odd number of categories (e.g., five) is preferable, as it provides a distinct middle category (i.e., *Medium*). These findings led to our default of five and three categories for a KPI’s value and trend, respectively (see §III-A2). Together, these aspects define the practical conditions for SIA’s successful deployment.

VII. CONCLUSIONS

This paper presented a new paradigm for interpreting anticipatory DRL agents by introducing SIA, a symbolic framework that brings transparency and trust to their operation in mobile networking. By leveraging scalable per-KPI knowledge graphs and a novel IS metric, SIA delivers real-time, actionable insights that are beyond the reach of existing methods. Our evaluations demonstrated that these insights are not merely diagnostic; they enable both targeted agent redesigns and automated performance enhancements, boosting key network metrics by up to 25%. Ultimately, by making proactive control understandable and tunable, SIA lowers a critical barrier to its adoption in next-generation networks.

ACKNOWLEDGMENTS

This work is partially supported by BRAIN project PID2021-128250NB-I00 funded by MCIN/AEI/10.13039/501100011033/ and the European Union ERDF “A way of making Europe”; by Agile-6G Project PID2024-163089NB-I00 funded by MICIU/AEI/10.13039/501100011033; C. Fiandrino is a Ramón y Cajal awardee (RYC2022-036375-I), funded by MCIU/AEI/10.13039/501100011033 and the ESF+. This work is also supported by U.S. NSF under grants CNS-2112471 and CNS-2434081, and by OUSD(R&E) through ARL CA W911NF-24-2-0065. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

- [1] H. Tataria, M. Shafi *et al.*, “6G wireless systems: Vision, requirements, challenges, insights, and opportunities,” *Proc. of the IEEE*, vol. 109, no. 7, pp. 1166–1199, 2021.
- [2] Z. Zhang, Y. Xiao *et al.*, “6G wireless networks: Vision, requirements, architecture, and key technologies,” *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, 2019.
- [3] H. Mao, M. Alizadeh *et al.*, “Resource management with deep reinforcement learning,” in *Proc. of ACM HotNets*, 2016, pp. 50–56.
- [4] Y. Cai, P. Cheng *et al.*, “Deep reinforcement learning for online resource allocation in network slicing,” *IEEE Trans. Mobile Comput.*, vol. 23, no. 6, pp. 7099–7116, 2023.
- [5] Q. An, S. Segarra *et al.*, “A deep reinforcement learning-based resource scheduler for massive MIMO networks,” *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 1, pp. 242–257, 2023.
- [6] H. Ye, G. Y. Li *et al.*, “Power of deep learning for channel estimation and signal detection in OFDM systems,” *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 114–117, 2017.
- [7] C. Ge, Z. Ge *et al.*, “Chroma: Learning and using network contexts to reinforce performance improving configurations,” in *Proc. of ACM MobiCom*, 2023.
- [8] F. Vannella, A. Proutiere *et al.*, “Learning optimal antenna tilt control policies: A contextual linear bandits approach,” *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 12666–12679, 2024.
- [9] B. Andrew and S. Richard S, “Reinforcement learning: an introduction,” 2018.
- [10] T. G. Dietterich, “Hierarchical reinforcement learning with the maxq value function decomposition,” *J. Artif. Intell. Res.*, vol. 13, pp. 227–303, 2000.
- [11] A. Y. Ng, D. Harada *et al.*, “Policy invariance under reward transformations: Theory and application to reward shaping,” in *Proc. of ICML*, 1999, pp. 278–287.
- [12] R. S. Sutton, D. Precup *et al.*, “Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning,” *Artif. Intell.*, vol. 112, no. 1-2, pp. 181–211, 1999.
- [13] R. S. Sutton, “Learning to predict by the methods of temporal differences,” *Mach. Learn.*, vol. 3, no. 1, pp. 9–44, 1988.
- [14] L. Chen, J. Lingys *et al.*, “Auto: Scaling deep reinforcement learning for datacenter-scale automatic traffic optimization,” in *Proc. of ACM SIGCOMM*, 2018, pp. 191–205.
- [15] H. Mao, R. Netravali *et al.*, “Neural adaptive video streaming with pensieve,” in *Proc. of ACM SIGCOMM*, 2017, pp. 197–210.
- [16] D. A. Worthy, A. R. Otto *et al.*, “Working-memory load and temporal myopia in dynamic decision making,” *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 38, no. 6, p. 1640, 2012.
- [17] D. Arumugam and B. Van Roy, “Deciding what to learn: A rate-distortion approach,” in *Proc. of ICML*, 2021, pp. 373–382.
- [18] K. Chen, B. Wang *et al.*, “DeeProphet: Improving http adaptive streaming for low latency live video by meticulous bandwidth prediction,” in *Proc. of ACM Web Conf.*, 2023, pp. 2991–3001.
- [19] R. A. K. Fezeu, C. Fiandrino *et al.*, “Unveiling the 5G mid-band landscape: From network deployment to performance and application QoE,” in *Proc. of ACM SIGCOMM*, 2024, pp. 358–372.
- [20] C. G. Bampis and A. C. Bovik, “Learning to predict streaming video QoE: Distortions, rebuffering and memory,” *arXiv:1703.00633*, 2017.
- [21] S. S. Krishnan and R. K. Sitaraman, “Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs,” in *Proc. of ACM IMC*, 2012, pp. 211–224.
- [22] S. Fujimoto, W.-D. Chang *et al.*, “For SALE: State-action representation learning for deep reinforcement learning,” in *Proc. of NIPS*, A. Oh, T. Naumann *et al.*, Eds., vol. 36, 2023, pp. 61 573–61 624.
- [23] S. Chinchali, P. Hu *et al.*, “Cellular network traffic scheduling with deep reinforcement learning,” in *Proc. of AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.
- [24] G. Lv, Q. Wu *et al.*, “Accurate throughput prediction for improving QoE in mobile adaptive streaming,” *IEEE Trans. Mobile Comput.*, 2023.
- [25] D. Gunning and D. Aha, “DARPA’s explainable artificial intelligence (XAI) program,” *AI magazine*, vol. 40, no. 2, pp. 44–58, 2019.
- [26] A. Clemm, L. Ciavaglia *et al.*, “Intent-based networking-concepts and definitions,” *IETF RFC*, 2022, rFC 9315, Oct.
- [27] M. T. Ribeiro, S. Singh *et al.*, ““Why should I trust you?” explaining the predictions of any classifier,” in *Proc. of ACM SIGKDD*, 2016, pp. 1135–1144.
- [28] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. of NIPS*, 2017, pp. 4768–4777.
- [29] M. Polese, L. Bonati *et al.*, “CoIO-RAN: Developing machine learning-based xapps for open RAN closed-loop control on programmable experimental platforms,” *IEEE Trans. Mobile Comput.*, vol. 22, no. 10, pp. 5787–5800, 2022.
- [30] M. Morari, C. E. Garcia *et al.*, “Model predictive control: Theory and practice,” *IFAC Proc. Vol.*, vol. 21, no. 4, pp. 1–12, 1988.
- [31] T. M. Moerland, J. Broekens *et al.*, “Model-based reinforcement learning: A survey,” *Found. Trends Mach. Learn.*, vol. 16, no. 1, pp. 1–118, 2023.
- [32] S. P. Singh, M. L. Littman *et al.*, “Learning predictive state representations,” in *Proc. of ICML*, 2003, pp. 712–719.
- [33] M. Littman and R. S. Sutton, “Predictive representations of state,” in *Proc. of NIPS*, vol. 14. The MIT Press, 2001.
- [34] A. d. Garcez and L. C. Lamb, “Neurosymbolic AI: The 3rd wave,” *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 12387–12406, 2023.
- [35] S. Russell and P. Norvig, “Artificial intelligence: a modern approach. 3rd,” *Upper Saddle River, EUA: Prentice-Hall*, 2010.
- [36] L. De Raedt, S. Dumancic *et al.*, “From statistical relational to neuro-symbolic artificial intelligence,” in *Proc. of IJCAI*, vol. 5, 2020, pp. 4943–4950.
- [37] A. Graves, G. Wayne *et al.*, “Hybrid computing using a neural network with dynamic external memory,” *Nature*, vol. 538, no. 7626, pp. 471–476, 2016.
- [38] A. Bibal, R. Cardon *et al.*, “Is attention explanation? an introduction to the debate,” in *Proc. of ACL*, 2022, pp. 3889–3900. [Online]. Available: <https://aclanthology.org/2022.acl-long.269/>
- [39] J. Ish-Horowicz, D. Udwin *et al.*, “Interpreting deep neural networks through variable importance,” *arXiv preprint arXiv:1901.09839*, 2019.
- [40] S. Milani, N. Topin *et al.*, “Explainable reinforcement learning: A survey and comparative review,” *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1–36, 2024.
- [41] Z. Meng, M. Wang *et al.*, “Interpreting deep learning-based networking systems,” in *Proc. of ACM SIGCOMM*, 2020, pp. 154–171.
- [42] C. Fiandrino, L. Bonati *et al.*, “Explora: AI/ML explainability for the open RAN,” *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 1, no. CoNEXT3, pp. 1–26, 2023.
- [43] Z. Ma, Y. Zhuang *et al.*, “Learning symbolic rules for interpretable deep reinforcement learning,” *arXiv preprint arXiv:2103.08228*, 2021.
- [44] A. Verma, V. Murali *et al.*, “Programmatically interpretable reinforcement learning,” in *Proc. of ICML*, 2018, pp. 5045–5054.
- [45] A. Duttagupta, M. Jabbari *et al.*, “SymbXRL: Symbolic explainable deep reinforcement learning for mobile networks,” in *Proc. of IEEE INFOCOM*, 2025, pp. 1–10.
- [46] M. Ameer, B. Brik *et al.*, “Leveraging LLMs to explain DRL decisions for transparent 6G network slicing,” in *Proc. of IEEE NetSoft*, 2024, pp. 204–212.
- [47] F. Petroni, T. Rocktäschel *et al.*, “Language models as knowledge bases?” in *Proc. of EMNLP-IJCNLP*, 2019, pp. 2463–2473. [Online]. Available: <https://aclanthology.org/D19-1250/>
- [48] R. Jain and I. Chlamtac, “The P2 algorithm for dynamic calculation of quantiles and histograms without storing observations,” *Commun. ACM*, vol. 28, no. 10, pp. 1076–1085, 1985.
- [49] M. Greenwald and S. Khanna, “Space-efficient online computation of quantile summaries,” *ACM SIGMOD Rec.*, vol. 30, no. 2, pp. 58–66, 2001.
- [50] Y. Nie, N. H. Nguyen *et al.*, “A time series is worth 64 words: Long-term forecasting with transformers,” in *Proc. of ICLR*, 2023.
- [51] T. Kim, J. Kim *et al.*, “Reversible instance normalization for accurate time-series forecasting against distribution shift,” in *Proc. of ICLR*, 2021.
- [52] Y. S. Nam, J. Gao *et al.*, “Xatu: Richer neural network based prediction for video streaming,” *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 5, no. 3, pp. 1–26, 2021.
- [53] A. Narayanan, E. Ramadan *et al.*, “Lumos5G dataset,” 2021. [Online]. Available: <https://dx.doi.org/10.1145/3419394.3423629>
- [54] H. Riiser, P. Vigmostad *et al.*, “Commute path bandwidth traces from 3G networks: Analysis and applications,” in *Proc. of ACM MMSys*, 2013, pp. 114–118.
- [55] R. K. Jain, D.-M. W. Chiu *et al.*, “A quantitative measure of fairness and discrimination,” *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, vol. 21, no. 1, 1984.
- [56] M. Polese, L. Bonati *et al.*, “Colosseum: The open RAN digital twin,” *IEEE Open J. Commun. Soc.*, 2024.