

# Interpreting Anticipatory Deep Reinforcement Learning for Proactive Mobile Network Control

MohammadErfan Jabbari\*, Abhishek Duttagupta\*<sup>†</sup>, Claudio Fiandrino\*, Leonardo Bonati<sup>§</sup>,  
Salvatore D'Oro<sup>§</sup>, Michele Polese<sup>§</sup>, Marco Fiore\*, and Tommaso Melodia<sup>§</sup>

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

<sup>§</sup>Northeastern University, Boston, USA, Email: {l.bonati, s.doro, m.polese, t.melodia}@northeastern.edu

<sup>†</sup>Universidad Carlos III de Madrid, Spain

**Abstract**—Deep Reinforcement Learning (DRL) is widely used for adaptive control in mobile networks, yet most agents remain reactive. This limitation is particularly problematic for exogenous Key Performance Indicators (KPIs), whose dynamics cannot be directly controlled by agent action and evolve independently. Anticipatory DRL addresses this issue by augmenting the state with short-horizon KPIs forecasts, but it remains unclear whether such information truly influences decisions. We use SIA, a symbolic interpretability tool, to explain whether and how anticipatory information is actually exploited by the policy, enabling principled redesign of forecast inputs and performance improvements. Using policy graphs and Mutual Information (MI) over symbolic temporal features, SIA distinguishes proactive and reactive behaviors. Using a standard Pensieve ABR agent augmented with throughput forecasts, experiments on real-world 5G traces show a 3% average reward improvement, with anticipatory policies spending more time at high bitrates while reducing unnecessary oscillations.

## I. INTRODUCTION AND MOTIVATION

Anticipatory DRL equips standard agents with short-term future knowledge to enable proactive rather than reactive control. This capability will be especially important for next-generation mobile networks, where abrupt fluctuations driven by mobility, dynamic scheduling, and volatile radio conditions render reactive control insufficient for maintaining stable Quality of Service (QoS) [1]. As shown in Fig. 1, anticipatory agents adjust bitrate ahead of degradation, reducing rebuffering and improving Quality of Experience (QoE).

Typically, DRL in networking use cases features two types of KPIs: *controllable* and *exogenous*. Controllable KPIs respond directly to agent actions and can be learned through temporal-difference learning. In contrast, exogenous KPIs evolve largely independently of actions, preventing the agent from learning a control policy over them. Hence, anticipatory DRL design is not just a matter of state augmentation and raises key questions:

- 1) Does the agent learn how to exploit anticipatory information, or are performance gains due to incidental feature addition?
- 2) Does exploiting anticipatory information translate into proactive decision-making?

To answer these questions, we use **SIA** (Symbolic Interpretability for Anticipatory DRL) [2], a recently proposed explanatory tool tailored to network DRL systems. SIA operates in the symbolic domain, using First-Order Logic (FOL)-based definitions of exogenous and controllable KPIs to reveal the agent's decision-making process, and leverages policy

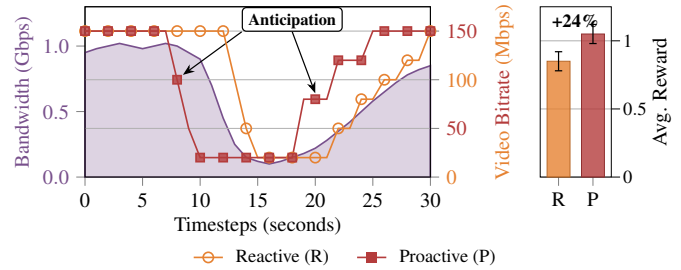


Fig. 1. Proactive Agents with forecasted bandwidth achieve higher QoE [2]. graphs and MI-based temporal analysis to quantify the impact of the forecast. Specifically, we demonstrate the effectiveness of the approach with a practical Adaptive Bitrate Streaming (ABR) streaming agent [3] operating on real-world 5G traces [4]. Our results demonstrate that the interpretability granted by SIA can reveal whether and how anticipatory information is used effectively in the target DRL agent.

## II. APPLYING SIA ON ABR

### A. Use case Description and Forecast Augmentation

We adapt the Pensieve agent [3], a well-established DRL-based ABR agent that optimizes video quality while minimizing playout interruptions. The agent's state includes buffer level, previously bitrate, histories of throughput, and delay; actions select the next bitrate. The reward balances bitrate utility, rebuffering, and quality smoothness. To provide foresight, we augment the legacy Pensieve agent (*reactive*) with throughput forecasts, obtaining a new anticipatory DRL model (Pensieve-P, for *proactive*). The updated agent uses a lightweight MLP with RevIN [5] using a lookback window  $w = 10$  and prediction horizon  $h = 2$ , achieving 31% MAPE, a widely used metric for relative forecasting accuracy, outperforming ABR-specific methods such as Lumos [6] (38% MAPE).

### B. Quantifying Forecast Impact Through MI

SIA uses MI between symbolic KPI states and agent actions to identify which KPIs and *temporal slices* most influence the learned policy. For Pensieve-P, the exogenous throughput state spans seven past values  $(t-7) \dots (t-1)$ , the current  $(t)$ , and two forecasted values  $(t+1)$ ,  $(t+2)$ . The MI analysis reveals a clear temporal gradient, increasing from past to present and peaking on forecasted throughput:  $(t+1)$ ,  $(t+2)$  exhibit higher MI with actions than any past sample and slightly more than the current value. This confirms that the agent actively exploits anticipatory information when selecting bitrates.

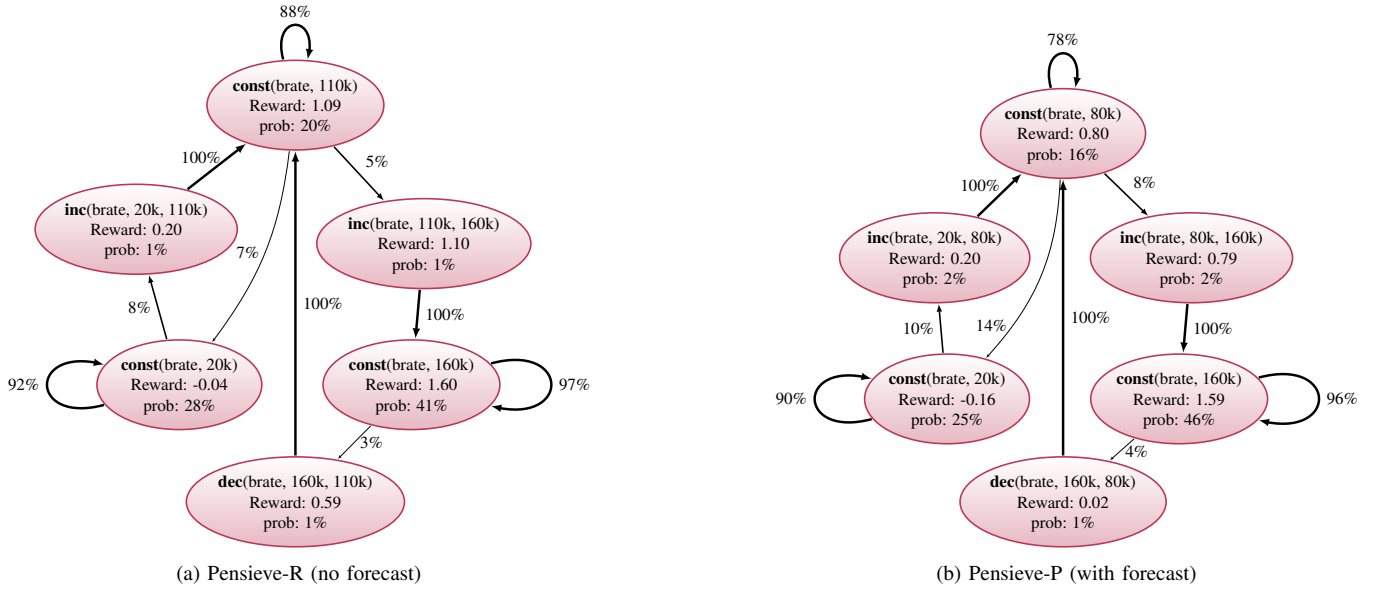


Fig. 2. SIA policy graphs showing how forecast augmentation restructures the agent’s operating points. Nodes represent symbolic actions with associated rewards and visit probabilities; edges show transition probabilities. Low probability nodes  $< 1\%$  were omitted for better visualization.

### C. Agent Policy Analysis

SIA analyzes policy-level behavior using policy graphs, which visualize action persistence and transitions under symbolic contexts. Nodes encode actions as predicate(`metric, from, to`) with directional predicates (`inc, dec, const`) with attributes like average reward and node probability, while edges capture transition dynamics. Fig. 2 reveals two key behavioral shifts when comparing Pensieve-R and Pensieve-P:

**Vanishing Intermediate Bitrate.** Pensieve frequently operates at `const(brate, 110k)` (20% of time), using it as an intermediate level between low and high quality. Pensieve-P virtually abandons this operating point, instead discovering `const(brate, 80k)` (16%) as its new middle ground. Without foresight, 110k is too aggressive under rapidly fluctuating throughput and imminent bandwidth drops, yet suboptimal when bandwidth remains stable. With forecasts, the agent can confidently commit to 160k when conditions permit, or retreat to the more conservative 80k when drops are anticipated.

**Forecast-Enabled Confidence.** Pensieve-P selects the highest bitrate `const(brate, 160k)` in 46% of decisions, compared to 41% for the reactive Pensieve baseline, while being more conservative at intermediate bitrates. This behavior yields a 3% average reward improvement, which is substantial in ABR settings. This apparent paradox (being both more aggressive at the top and more cautious in the middle) is resolved by anticipatory information: forecasts enable the agent to commit to high quality when safe and preemptively avoid risky states when degradation is predicted. This is clearly reflected in Fig. 1 during the recovery phase (timesteps 18–25): while the reactive agent waits for confirmed throughput increases to ramp-up, the proactive agent leverages the forecast to increase bitrates ahead of the bandwidth recovery curve.

**Takeaway Message.** By using foresight, the agent reduces

bitrates oscillations, yielding higher QoE.

### 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, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, “6G wireless systems: Vision, requirements, challenges, insights, and opportunities,” *Proc. of the IEEE*, vol. 109, no. 7, pp. 1166–1199, 2021.
- [2] M. Jabbari, A. Duttagupta, C. Fiandrino, L. Bonati, S. D’Oro, M. Polese, M. Fiore, and T. Melodia, “SIA: symbolic interpretability for anticipatory deep reinforcement learning in network control,” in *IEEE Conference on Computer Communications (INFOCOM 2026)*, May 2026, p. 9.94.
- [3] H. Mao, R. Netravali, and M. Alizadeh, “Neural adaptive video streaming with pensieve,” in *Proc. of ACM SIGCOMM*, 2017, pp. 197–210.
- [4] A. Narayanan, E. Ramadan, R. Mehta, X. Hu, Q. Liu, R. A. Fezeu, U. K. Dayalan, S. Verma, P. Ji, T. Li, F. Qian, and Z.-L. Zhang, “Lumos5G dataset,” 2021.
- [5] T. Kim, J. Kim, Y. Tae, C. Park, J.-H. Choi, and J. Choo, “Reversible instance normalization for accurate time-series forecasting against distribution shift,” in *Proc. of ICLR*, 2021.
- [6] G. Lv, Q. Wu, Q. Tan, W. Wang, Z. Li, and G. Xie, “Accurate throughput prediction for improving QoE in mobile adaptive streaming,” *IEEE Trans. Mobile Comput.*, 2023.