ISSN 2093-0542

KOREAN SOCIETY FOR INTERNET INFORMATION

The 17th International Conference on Internet

(ICONI 2025)

Dec. 14-17, 2025 Okinawa Convention Center, Okinawa, Japan

http://www.iconi.org

Proceedings of ICONI 2025

| Organized by |

Korean Society for Internet Information (KSII)

| Sponsored by |

OCVB

Contents

KSII The 17th International Conference on Internet (ICONI) 2025.

Probing the Previous State Dependency in the Decision Transformer for Reducing the Number of RL Validation

Taehyun Han[1], Janghwan Kim[2], Sangho Lee[3], Chanyul Yang[4] and R. Young Chul Kim[5,*]
[125] Software Engineering Laboratory, Graduate School, Hongik University
Sejong, South Korea
[3] Rastech
[4] DAON INC
[e-mail: taehyun3172@g.hongik.ac.kr[1], {lentoconstante[2], bob[5]}@hongik.ac.kr,
project@rastech.co.kr[3], yangcy2000@da-on.com[4]]
*Corresponding author: R. Young Chul Kim

Abstract

Decision Transformer (DT) successfully reframed the offline reinforcement learning problem as a sequence modeling task. However, DTs are designed to take fixed-length input sequences, and it remains unclear whether this design represents the optimal history length for decision-making, primarily due to the model's black-box nature, which makes validating their underlying policies computationally expensive. Our probing study provides a lightweight diagnostic tool for these strategies, while prior work has explored adaptive history lengths for performance; however, no study has systematically investigated the effect of progressively shortening history on action consistency across different model qualities. In this paper, we propose an analytical approach that systematically varies the input sequence length to examine how the decisions of DT change. Specifically, we compare the original baseline result, which uses the full, fixed history length ($K = 20$), with results from progressively shorter history lengths, denoted as k (ranging from 1 to 19). By applying this method to three different models—*Expert, Medium, and Medium-Replay*—and quantifying consistency using L2 Norm and Cosine Similarity, we provide new insights into each model's dependency on past information and its underlying decision-making behavior.

Keywords: Decision Transformer, Offline Reinforcement Learning, The Previous State Dependency (History Dependency)

1. Introduction

Offline reinforcement learning (RL) focuses on learning policies from pre-collected datasets. Recently, Transformer-based approaches have emerged in this domain, with the Decision Transformer (DT) achieving notable success by reframing RL as a sequence modeling task [1,2]. However, DT relies on a fixed input history length ($K = 20$), a design choice whose optimality remains unexamined due to the black-box nature of deep models [3]. Prior work has explored DT's internal mechanisms through attention pattern analysis or studies on token importance [3]. Recent variants, such as Long-Short DT [4] and Elastic DT [5], include ablation studies on history length—but primarily to

This research was supported by Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports and Tourism (MCST) in 2025 (Project Name: Artificial Intelligence-based User Interactive Storytelling 3D Scene Authoring Technology Development, Project Number: RS-2023-00227917307820870502O1) and National Research Foundation (NRF), Korea, under project BK21 Four.

validate their own architectural improvements, rather than to systematically probe how action consistency in the original DT evolves as history varies from k to K.

This paper fills this gap with a dedicated probing study. Understanding the previous state dependency is critical for enabling efficient, real-time inference in DT-based agents [4]. Furthermore, it provides a lightweight method to diagnose a model's strategy, offering a potential pathway to reducing the number of costly RL validation cycles. We introduce a systematic methodology that controls the observed input sequence length and quantifies changes in decision-making behavior via L2 norm and cosine similarity of predicted actions. Applied to three DT variants (*Medium*, *Expert*, *Medium-Replay*) on the *Hopper-Expert-v2 dataset* [1], our analysis reveals stark differences in history dependency (previous state dependency)—suggesting that training data quality fundamentally shapes decision patterns.

The remainder of this paper is organized as follows. Section 2 reviews the related work on the Decision Transformer and Transformer interpretability—Section 3 details our proposed probing methodology, including the attention masking technique and evaluation metrics. Section 4 presents experimental results from applying our method to the three different models. Section 5 provides an in-depth discussion of these results, interpreting the models' distinct decision-making strategies. Finally, Section 6 concludes the paper by summarizing our findings, acknowledging limitations, and highlighting the practical implications for reducing RL validation.

2. Related Work

2.1 Decision Transformer

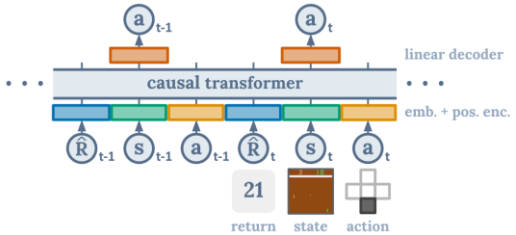


Fig. 1. DT architecture (Source: [1])

DT takes Return-to-Go (RTG), state, and action tokens as input and predicts the following action in an auto-regressive manner, similar to GPT. When a user sets a target reward as input, DT predicts the optimal action from the sequence data to achieve this target [1]. Various follow-up studies are currently underway, such as improving DT's performance or adapting its architecture for online RL environments [4, 5, 7].

2.2 Transformer Interpretability

Deep learning neural networks, such as Transformers, have a 'black-box' characteristic, making it difficult to understand the basis for their outputs [2,3] clearly. Various studies have been conducted to analyze Transformers in different ways, aiming to understand their characteristics. Representative techniques include methods that measure importance by masking parts of the input or methods that directly analyze the model's internal attention weights [2]. Furthermore, specific to DT, some studies have examined which inputs—such as RTG or state tokens—have a more significant impact on the model's decision-making [3].

3. Methodology

This chapter outlines the proposed probing to analyze the history dependency (or previous state dependency) of DT.

3.1 Target Models and Data

This study analyzes pre-trained DT models based on the D4RL benchmark's *Hopper-v2* environment. Following the original DT paper [1], we use models trained with a history length of $K = 20$. To compare how history dependency (previous state dependency) varies with the quality of training data, we use three models available on HuggingFace [6]: Expert, Medium, and Medium-Replay.

3.2 Probing: Attention Masking

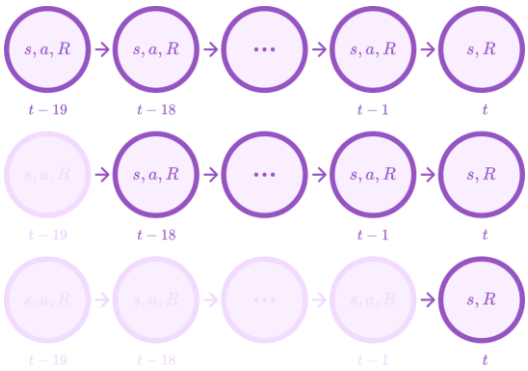


Fig. 2. Proposed Probing Methodology

Calculate Baseline Action (a_P) : First, we establish a baseline using the action predicted with the full history length of $K = 20$. As shown in the top row of **Fig. 2**(baseline), all information from $t - 19$ to t (note: s, a, R for past steps, s, R for the current step) is fed into the model. The resulting action predicted at the final timestep t is defined as the baseline action $a_P = a_{K=20}$.

Calculate Probed Action ($a_Q(k)$): Next, we vary the 'available recent history length' k from 1 to 19. As shown in the middle and bottom rows of **Fig. 2** (illustrating $k = 19$ and $k = 1$, respectively), for each k, we calculate the number of past steps to mask ($m = K - k$)and set the attention mask values for the oldest m steps to 0 (visualized as faded nodes). The action predicted using this modified mask is defined as the probed action $a_Q(k)$.

3.3 Evaluation Metrics

In this study, to quantify the difference between the baseline action a_p and the probed action $a_Q(k)$, we use the following two metrics. (The variable $past\_N$ in the code corresponds to k in this text.) L2 Norm (Euclidean Distance): Measures the magnitude of the difference between the two action vectors. Calculated as $D(k) = \left|\left|a_p - a_Q(k)\right|\right|_2$, a value closer to 0 indicates identical actions, while a larger value signifies a greater error. Cosine Similarity: Measures the directional agreement between the two action vectors. A value closer to 1 indicates that both actions point in the same direction, while values near 0 or -1 signify higher divergent directions.

4. Experiments and Results

This section presents the results of applying our Section 3 methodology to the three DT models.

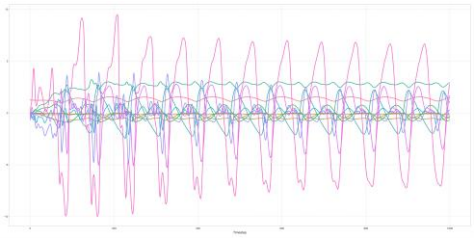


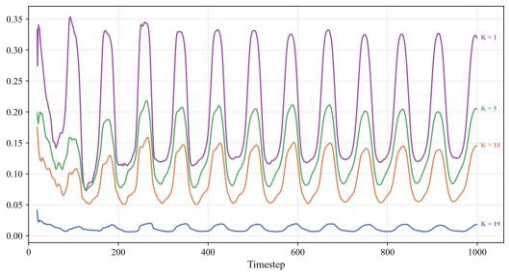
Fig. 3. State Parameter Trends of Average Expert Trajectories in Hopper-v2

This graph visualizes the changes in key state parameters (e.g., *torso height*, *joint angles*, *velocities*) over 1000 timesteps, averaged across all episodes from the Expert dataset. The periodic fluctuations observed in parameters correspond to the distinct phases of the Hopper agent's locomotion (e.g., jumping and landing cycles).

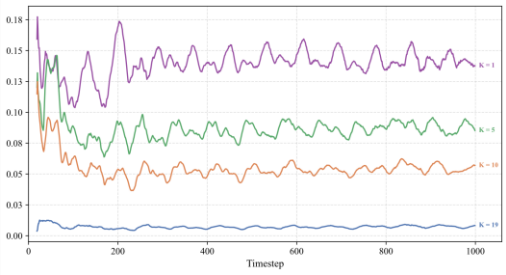
4.1 L2 Norm (Error) Analysis

The analysis results for the L2 Norm (Error) are shown in **Fig. 4**. These graphs illustrate the mean deviation from the baseline ($K = 20$), calculated by averaging the L2 Norm (Error) across all episodes for each corresponding timestep. For clear illustration, we plot the trends only for representative k^* values ($k^* = 1, 5, 10, 19$)

(a) Expert model



(b) Medium model



(c) Medium-Replay model

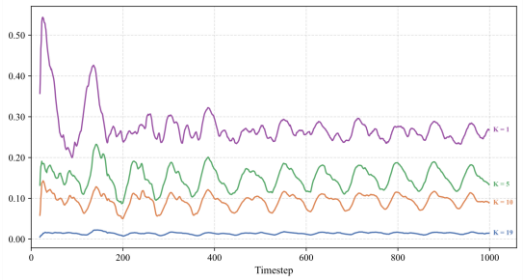


Fig. 4. L2 Norm (Error) Analysis by Model

Expert model (**Fig. 4(a)**): For $k = 1$ (purple lines), periodic spikes in the L2 Error are clearly observed, peaking at 0.35. This trend aligns with the periodic state changes (e.g., jumping, landing) observed in **Fig. 3**. As k increases to 5 (green line) and 10 (orange line), the error amplitude decreases significantly. At $k = 19$ (blue line), the error converges close to 0.

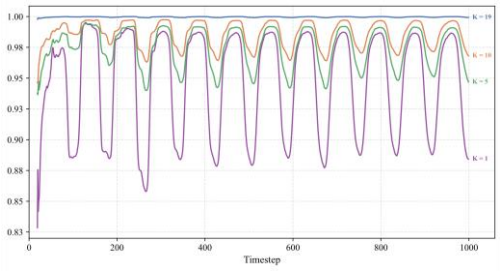
Medium Model (**Fig. 4(b)**): Even at $k = 1$, the L2 Error is markedly lower (around 0.15 to 0.18) and more stable than the Expert model, with very small spike amplitudes. The trend of error decreasing with larger k remains consistent.

Medium-Replay Model (**Fig. 4(c)**): This model exhibits the most significant error of all three models, spiking above 0.55 at $k = 1$ in the $t < 200$ region. Subsequently, it shows an irregular pattern with lower amplitudes than the Expert model.

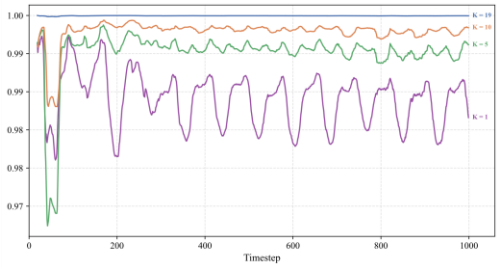
4.2 Cosine Similarity Analysis

While the L2 Norm in Section 4.1 measured the error magnitude, it cannot capture directional differences. Therefore, to analyze the directional agreement of the action vectors, Cosine Similarity was measured for the same k^* samples. This provides a complementary perspective on action consistency.

(a) Expert model



(b) Medium model



(c) Medium-Replay model

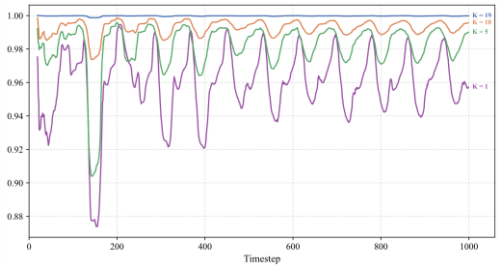


Fig. 5. Cosine Similarity Analysis by Model

The findings from the L2 Norm analysis are consistently mirrored in the Cosine Similarity results (**Fig. 5**).

Expert model (**Fig. 5(a)**): Exactly coinciding with the L2 Norm spikes, the similarity for $k = 1$ (purple line) shows periodic dips down to 0.83. This implies that when past information is insufficient, the model predicts an action in a completely different direction.

Medium Model (**Fig. 5(b)**): This model maintains a very high similarity (0.97 to 1.0) compared to the Expert model. Slight periodic dips are observable at $k = 1$, but the magnitude of the drop is minimal.

Medium-Replay Model (**Fig. 5(c)**): This model shows significant directional errors at similar points to the Expert model, with similarity dropping to 0.88 near $t = 150$ for $k = 1$.

5. Discussion

5.1 Previous State Dependency

Our probing study revealed a common phenomenon across all three models. The points where the L2 Norm error spikes and Cosine Similarity plummet (**Figs. 4, 5**) precisely align with the periodic state changes of the Hopper agent (**Fig. 3**). This suggests that these moments of dynamic state transitions—such as landing after a jump or preparing for a new one—represent 'critical decision points' where past trajectory information is most urgently required to determine the following action.

5.2 Comparison of Decision-Making Strategies by Model

The Expert model was the most sensitive to history length at these junctures (**Figs. 4(a), 5(a)**). When the history was extremely limited to $k = 1$, it predicted an action in a completely different direction from the baseline ($K = 20$), causing the L2 error to spike to 0.35 and Cosine Similarity to plummet to 0.83. This suggests that the Expert model, trained on consistent and high-quality data, is conditioned to rely heavily on a long history to assess these dynamic states accurately.

Conversely, the Medium model demonstrated remarkable stability, remaining almost entirely unaffected by history length at these same points (**Figs. 4(b), 5(b)**). Even at $k = 1$, the Cosine Similarity remained high (0.97-1.0) and the L2 error was minimal. This suggests that, due to training on lower-quality data, the model learned a robust policy that focused on the current state rather than meticulously following a specific trajectory. Its dependence on history is, therefore, the lowest of the three models.

The Medium-Replay model exhibited a more complex pattern (**Figs. 4(c), 5(c)**). After an initial period of instability ($t < 200$), it displays a clear periodicity like the Expert model. However, unlike the Medium model, it also reacts sensitively to history length. At $k = 1$, the drop in Cosine Similarity (to a low of 0.88) was far more significant than that of the Medium model (0.97 or higher). This suggests that while the Medium-Replay model also identifies these states as 'critical decision points' requiring past information (much like the Expert), the highly diverse and inconsistent trajectories in the 'Replay' dataset may have resulted in it learning a policy that is confused about which past experiences to reference.

In summary, we can interpret the models' behaviors at these 'critical decision points' as follows: the Expert model concludes, "The precise past trajectory is essential"; the Medium model concludes, "The current state is sufficient"; and the Medium-Replay model concludes, "The past is necessary, but it is unclear which experience to follow."

6. Conclusions

This paper proposes a systematic probing methodology using attention masking to analyze the effect of the DT's fixed history length (K = 20) on its decision-making. Our study dynamically manipulated the available history length k (from 1 to 19) for models trained with $K = 20$. It quantified the deviation from the baseline action ($K = 20$) using L2 Norm and Cosine Similarity metrics.

The experimental results confirmed that, regardless of training data quality, the models' history dependency (previous state dependency) commonly spiked at specific 'critical decision points' within the Hopper environment, such as 'jumping/landing'. However, the response to these points differed distinctly between models. The Expert model showed a high dependency on past trajectories at these junctures. In contrast, the Medium model exhibited a robust policy that was largely unaffected by history length. The Medium-Replay model recognized periodicity similarly to the Expert model but displayed a confounded dependency pattern, likely due to inconsistent training data. These findings suggest that the DT's decision-making mechanism, beyond mere performance, is fundamentally shaped by the quality and composition of its training data.

This study, however, has several limitations. First, our experiments were confined to the single Hopper-v2 dataset; different environments may yield different patterns of history dependency (or previous state dependency). Second, as a 'probing study', this work observes and quantifies

phenomena rather than fully elucidating the causal mechanisms within the transformer 'black-box'.

Despite these limitations, we hope this research contributes foundational data for future work on 'adaptive history selection' mechanisms, which could dynamically adjust k based on the agent's current state or uncertainty. Such advancements could lead to future research on reducing unnecessary computations from $O(K^2)$ to $O(k^2)$, potentially improving the viability of real-time deployment for DT-based agents. Furthermore, understanding the strategic importance of past states, as demonstrated vividly in this study, is also crucial for potentially and significantly reducing the number of costly RL validation cases.

References

[1] L. Chen, et al., "Decision transformer: Reinforcement learning via sequence modeling," *Advances in neural information processing systems*, vol.34, pp.15084-15097, 2021.

[2] A. Vaswani, et al., "Attention is all you need," *Advances in neural information processing systems*, vol.30, 2017.

[3] D. Thambi, P. Paruchuri, and P. Moodley, "Interpreting Decision Transformer: Insights from Continuous Control Tasks," in *Proc. of International Conference on Neural Information Processing*, Singapore: Springer Nature Singapore, 2024.

[4] J. Wang, et al., "Long-short decision transformer: Bridging global and local dependencies for generalized decision-making," in *Proc. of the Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025.

[5] Y. H. Wu, X. Wang, and M. Hamaya, "Elastic decision transformer," *Advances in neural information processing systems*, vol.36, pp.18532-18550, 2023.

[6] E. Beeching, Decision Transformer Gym Hopper Expert Model, Hugging Face model repository, 2021. [Online]. Available: https://huggingface.co/edbeeching/decision-transformer-gym-hopper-expert

[7] H. L. Hsu, et al., "D2t2: Decision transformer with temporal difference via steering guidance," in *Proc. of the Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024.