Exploring Attention Mechanisms in Transformers for Data-Efficient Model-Based Reinforcement Learning

Daniel De Dios Allegue[1]   supervised by Dr. Frans Oliehoek[1], Dr. Jinke He[1]

[1]EEMCS, Delft University of Technology, The Netherlands

Abstract

Transformer-based world models in model-based RL enable effective planning but struggle in environments driven by short-term memory dependencies. We introduce two inductive biases; local attention and Gaussian adaptive attention, to steer the model toward the most relevant recent observations. Matched to an environment's inherent memory dependencies, both mechanisms outperform causal attention, achieving higher data efficiency with fewer interactions. These results demonstrate that influence-based priors can produce more effective, data-efficient attention for world-model learning.

Background

Model-Based Reinforcement Learning (MBRL): Agents learn a predictive model of the environment to plan future actions, improving data efficiency over model-free methods.

UniZero: Replace RNNs with causal self-attention to capture long-range dependencies, enabling more effective planning in complex tasks [1]. However, in environments dominated by rapid, local dynamics (e.g. Pong, Boxing), the causal self-attention mechanisms in UniZero disperse focus across irrelevant tokens, i.e. *attention dilution*, causing **performance degradation** with respect to previous models.

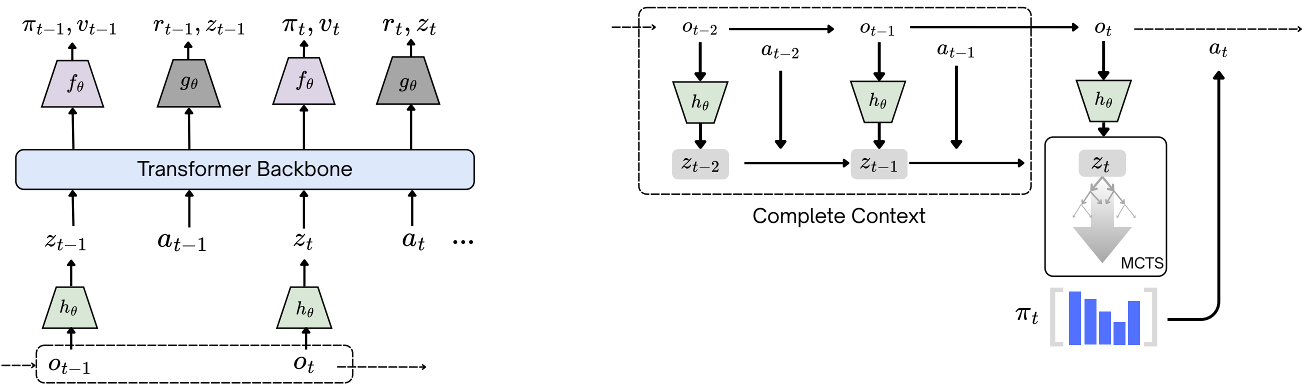


Figure 1. UniZero Architecture: The encoder maps raw observations to latent states; a Transformer attends over past (masked) observations and actions at each timestep t; the dynamics head g_θ predicts next latent state and reward, and the decision head f_θ outputs policy and value estimates. At inference, the root of the Monte Carlo Tree z_t benefits from a richer context by explicitly using a sequence of raw observations $o_{t-H:t}$

Influence-Based Priors In a partial observability setting (modelled with a POMDP), the agent maintains a belief state $b_t = P(s_t \mid H_t)$ conditioned on the full history $H_t = (o_1, a_1, \ldots, o_t)$. In reality, only a small subset of past observations and actions is needed to predict future outcomes as accurately as the entire history [2]. By identifying and focusing on these truly influential past states z_t or actions a_t, we can build a compact summary of H_t, reducing model complexity and dramatically improving data efficiency.

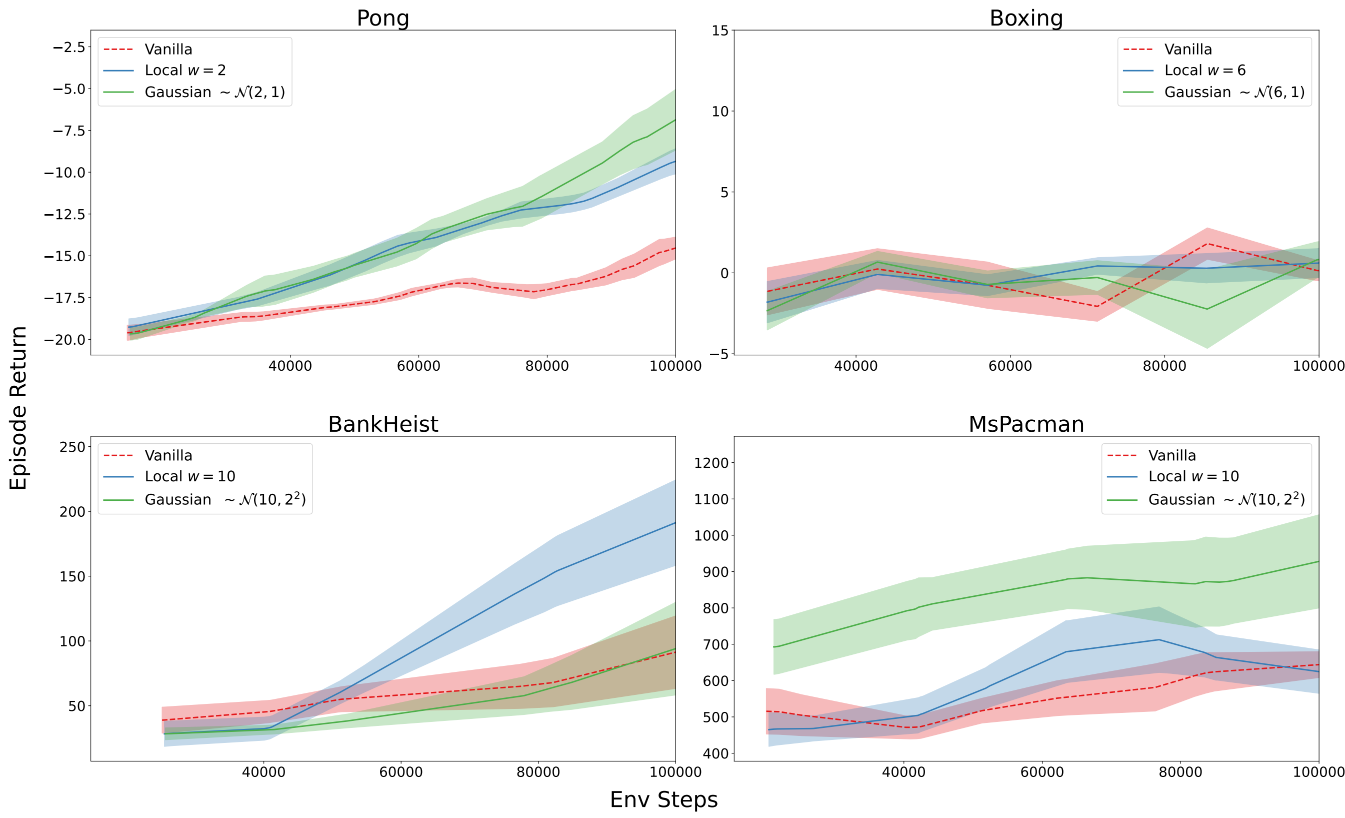
Results

Evaluation: We evaluate on Atari games—Pong and Boxing as short-memory tasks, BankHeist and MsPacman as memory-intensive tasks—under the Atari 100k setting (limited to 100k interactions) [3], comparing against the original UniZero transformer with causal self-attention.

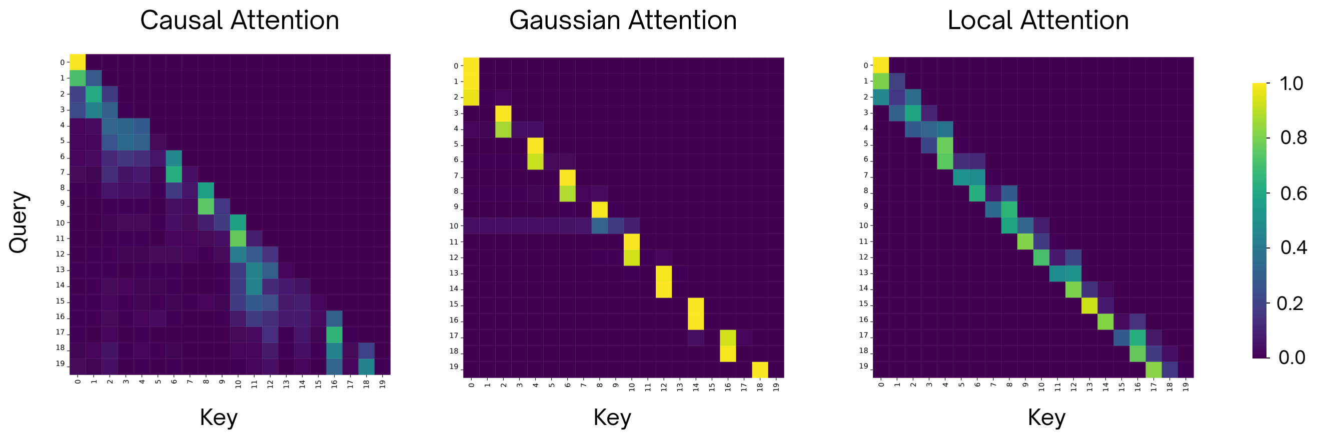
Table 1. **Atari Benchmark Score:** Mean episode returns (± standard error) across four Atari games.

Game	Causal (Baseline)	Local (Ours)	Gaussian (Ours)
Pong	-14.53 ± 0.66	-9.35 ± 0.76	-6.86 ± 1.82
Boxing	0.14 ± 0.63	0.62 ± 0.91	0.83 ± 1.10
MsPacman	643.93 ± 35.66	624.68 ± 59.89	928.12 ± 128.43
BankHeist	91.34 ± 28.03	191.30 ± 33.00	94.08 ± 35.80

Key Findings: Local attention yields strong gains in Pong and BankHeist. Gaussian adaptive attention delivers the largest improvements in Pong, Boxing, and MsPacman by flexibly capturing each game's temporal dependencies. Learning curves can be found below:



Attention Maps in Pong: Causal attention disperses focus, Local attention imposes a rigid diagonal window, and Gaussian attention learns a flexible, "soft" focus, explaining its robust performance across different task types.



Attention Mechanisms

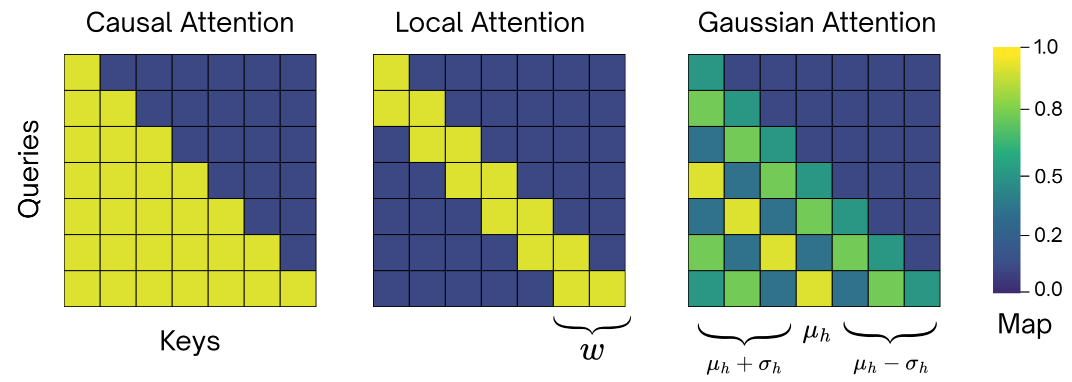
Causal Self-Attention : Each token i attends to all past tokens $j < i$ with equal capacity, using the standard scaled dot-product and a causal mask to prevent "peeking" into the future [4].

Local Attention: Applies a fixed window mask $W_{i,j} = 1$ if $|i - j| \leq w$ (else $-\infty$), introducing a *locality bias* that focuses each query on its w most recent tokens.

Gaussian Adaptive Attention: Learns a smooth, temporal bias by adding per-head mask logits

$$W_{ij}^{(h)} = \exp\left(-\frac{(\Delta_{ij} - \mu_h)^2}{2\sigma_h^2}\right) \tag{1}$$

to the standard dot-product scores before softmax [5]. Here, each head's μ_h and σ_h are trained end-to-end, centring attention on the most informative past tokens while retaining full flexibility.



Conclusions and Discussion

On Atari 100k, when these priors match game dynamics (e.g. Pong), **sample efficiency and final performance improve over causal attention.**

Local attention alone suffices for fixed longer-range dependencies (BankHeist), while Gaussian adaptivity becomes useful for environments with mixed or dynamic memory demands (Boxing, MsPacman).

Both mechanisms can approximate long-term dependencies often at lower cost than full attention, without needing the entire history.

These findings demonstrate that **simple inductive biases can make world models more data-efficient** and guide the design of targeted, resource-aware architectures.

[1] Y. Pu, Y. Niu, Z. Yang, J. Ren, H. Li, and Y. Liu, "Unizero: Generalized and efficient planning with scalable latent world models," *arXiv preprint arXiv:2406.10667*, 2024.

[2] F. Oliehoek, S. Witwicki, and L. Kaelbling, "Influence-based abstraction for multiagent systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, 2012, pp. 1422–1428.

[3] L. Kaiser *et al.*, "Model-based reinforcement learning for atari," *arXiv preprint arXiv:1903.00374*, 2019.

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

[5] G. Ioannides, A. Chadha, and A. Elkins, "Gaussian adaptive attention is all you need: Robust contextual representations across multiple modalities," *CoRR*, 2024.