Sparse Hierarchical Imagination for Efficient Transformer World Models

1. Motivation

Transformer-based world models (TWMs) have recently shown strong performance in sample-efficient reinforcement learning, particularly in domains with high-dimensional visual observations. Notable examples include IRIS [Micheli et al., 2023], STORM [Zhang et al., 2023], DART [Agarwal et al., 2024], TransDreamer [Chen et al., 2022], and Robine et al. [2023], which employ autoregressive Transformer architectures to model sequences of discrete latent tokens, typically obtained through vector-quantized autoencoders (VQ-VAE). These models achieve good control performance and demonstrate the viability of Transformer-based architectures for modeling complex visual dynamics.

However, Transformer world models face persistent limitations related to the computational cost and stability of long-horizon imagination. Autoregressive rollouts are computationally expensive, as attention cost grows quadratically with sequence length and the model is required to generate full latent reconstructions at each step. Moreover, rollouts often suffer from compounding error over time, particularly when modeling fine-grained visual details that are not critical for downstream control. Sparse Imagination [Chun et al., 2025] addresses part of this challenge by dropping tokens during rollout to reduce computation and improve policy learning. Yet, the approach applies flat token-level dropout, without considering the inherent structure of the scene or the relative importance of different tokens over time. In many environments, the visual state is naturally hierarchical: global context and goal-related information should be preserved across the rollout, dynamic objects must be modeled accurately to maintain control-relevant dynamics, while background details can often be sparsified without affecting policy performance. We propose Sparse Hierarchical Imagination (SHI), a framework that explicitly organizes tokens into a semantic hierarchy and applies level-specific, temporally adaptive sparsification during imagination. The goal is to improve rollout efficiency and stability by focusing model capacity on the most relevant components of the state for control.

2. Related Work

Transformer-based world models such as IRIS [Micheli et al., 2023], DART [Agarwal et al., 2024], TransDreamer [Chen et al., 2022], and STORM [Zhang et al., 2023] have demonstrated that modeling sequences of discrete latent tokens can lead to improved sample efficiency in reinforcement learning. Sparse Imagination [Chun et al., 2025] introduced token dropout for rollout acceleration, but the method applies flat dropout without considering token hierarchy or scene semantics.

Latent imagination for policy learning has also been explored in the Dreamer family of models [Hafner et al., 2020, 2021, 2023], which operate in compact latent spaces. However, Dreamer models

do not address structured sparsification within Transformer-based architectures, where the latent space is larger and structured at the token level.

Our work is closely related to SPARTAN [Lei et al., 2024], which learns sparse causal graphs over object tokens in physical domains. SPARTAN focuses on modeling interactions but does not address rollout efficiency through sparsification. Hierarchical latent models such as THICK [Gumbsch et al., 2024] introduce adaptive temporal abstractions in discrete latent dynamics, learning to sparsely update higher-level context variables. In contrast, our method focuses on sparsification of token sequences in Transformer-based world models, integrating causal and uncertainty-based signals to guide token selection during imagination. While both approaches aim to improve rollout efficiency, THICK leverages temporal abstraction whereas our method targets structured token sparsification conditioned on rollout context.

Object-centric representations, as explored by Slot Attention [Locatello et al., 2020] and SlotFormer [Wu et al., 2023], motivate treating object-level tokens distinctly during imagination. Our method draws on this insight by explicitly applying level-specific sparsification policies.

Finally, Sparse Transformers [Child et al., 2019] and Routing Transformers [Roy et al., 2020] propose adaptive attention mechanisms for improving Transformer efficiency. While related in spirit, these techniques operate at the attention level and are not directly applicable to the rollout sparsification problem addressed in this work.

3. Method

We consider an agent interacting with an environment that produces visual observations $\mathbf{x}_t$ and actions a_t. Our goal is to learn an efficient world model that can perform long-horizon imagination in latent space with structured sparsification. We follow the standard Transformer world model pipeline, where each observation $\mathbf{x}_t$ is first encoded into a sequence of discrete latent tokens via a hierarchical VQ-VAE or token clustering procedure. Specifically, the encoder maps $\mathbf{x}_t$ into a set of latent tokens $\mathbf{z}_t = [z_t^1, z_t^2, \ldots, z_t^K]$. We assume that the token set can be partitioned into L levels of semantic granularity, such that $\mathbf{z}_t = \bigcup_{l=1}^{L} \mathbf{z}_t^{(l)}$, with $\mathbf{z}_t^{(l)} = [z_t^{(l,1)}, \ldots, z_t^{(l,K_l)}]$. For example, Level 1 may contain global scene descriptors, Level 2 may contain object tokens, and Level 3 may contain fine-grained visual details.

The goal of Sparse Hierarchical Imagination (SHI) is to apply structured, level-specific sparsification during Transformer rollouts, focusing modeling capacity on control-relevant aspects of the state while improving computational efficiency and rollout stability. In contrast to Sparse Imagination [Chun et al., 2025], which applies flat token-level dropout, SHI leverages a hierarchical token organization and applies level-specific, temporally adaptive masking. Moreover, unlike Dreamer-based models [Hafner et al., 2020, 2021, 2023], which perform latent imagination in compact representations, SHI operates directly on Transformer-based models with large, structured token spaces. Finally, although SPARTAN [Lei et al., 2024] learns causal graphs over object tokens, it does not employ these graphs for controlling rollout sparsification. Our method integrates these components in a unified pipeline.

We adopt a two-stage training procedure. We first pretrain a Transformer world model autoregressively on full token sequences. The model learns to model $p(\mathbf{z}_{t+1}|\mathbf{z}_{\leq t}, a_{\leq t})$ without any sparsification. In parallel, we train a SPARTAN-style causal graph over token sequences, which learns directed edges between tokens and captures their causal dependencies. The causal graph provides a dynamic notion of which tokens influence future dynamics and informs the masking decisions during imagination. The Transformer and SPARTAN components are trained jointly or in alternating phases.

During rollout, the imagination pipeline proceeds as follows. Given an initial observation $\mathbf{x}_0$, we encode it into $\mathbf{z}_0$ and initialize the Transformer's hidden state. At each rollout step t, we apply level-specific, temporally adaptive masking to the current token set. Masked token subsets are defined as $\tilde{\mathbf{z}}_t^{(l)} = \mathbf{z}_t^{(l)} \odot m_t^{(l)}$, where $m_t^{(l)} \in \{0,1\}^{K_l}$ is a binary mask controlling which tokens at level l are retained at time step t. The autoregressive rollout distribution is then $p(\tilde{\mathbf{z}}_{t+1} | \tilde{\mathbf{z}}_{\leq t}, a_{\leq t})$.

The masking functions $m_t^{(l)}$ are parameterized by a small controller network. At each time step t and level l, the controller outputs $m_t^{(l)} = \sigma(f^{(l)}(\mathbf{h}_t, u_t, \mathbf{c}_t))$, where $\mathbf{h}_t$ summarizes the imagination context, for example through a memory token as in DART [Agarwal et al., 2024], u_t is a time-step embedding, and $\mathbf{c}_t$ contains auxiliary signals. The auxiliary signals $\mathbf{c}_t$ include token relevance scores derived from the SPARTAN causal graph, token-level uncertainty estimates computed from a stochastic Transformer architecture following STORM [Zhang et al., 2023], and attention scores to a memory token. Token relevance and uncertainty help prioritize which tokens should be retained for accurate imagination, while attention to a memory token encourages maintaining tokens that are salient in the current rollout context.

In addition, inspired by IRIS [Micheli et al., 2023], we apply image reconstruction loss selectively at a subset of rollout steps, specifically at checkpoints $t \in \{t_0, t_5, t_{10}\}$. This encourages semantic consistency during imagination while avoiding unnecessary modeling of visual details at intermediate steps.

4. Hypotheses and Evaluation

We hypothesize that applying Sparse Hierarchical Imagination (SHI) will improve both computational efficiency and rollout stability compared to existing approaches. In particular, we expect that SHI will reduce the number of floating point operations required per rollout step relative to flat token dropout as implemented in Sparse Imagination [Chun et al., 2025], when evaluated at equal rollout depth. Furthermore, by preserving semantically relevant tokens and sparsifying those less critical to control, we expect that SHI will reduce both token-level and image-level reconstruction error at longer rollout horizons.

We also hypothesize that improvements in rollout stability and efficiency will yield better sample efficiency during policy learning. This will be measured by the final median human-normalized score on Atari 100k and Crafter benchmarks. To test these hypotheses, we will conduct experiments comparing SHI against DART, IRIS, STORM, and Sparse Imagination baselines. We will report rollout compute cost, rollout error, and final RL performance. Additional ablation studies will investigate the impact of hierarchical masking, causal graph guidance, and uncertainty-based masking on the performance of SHI.

5. Conclusion

We have described a method for structured sparsification of Transformer world model rollouts through Sparse Hierarchical Imagination. By organizing latent tokens into semantic levels and learning adaptive, temporally aware masking functions, this approach aims to improve rollout efficiency and stability without sacrificing control performance. We will evaluate this method on established benchmarks and conduct ablation studies to better understand its contributions.

References

Pranav Agarwal, Sheldon Andrews, and Samira Ebrahimi Kahou. Learning to play atari in a world of tokens, 2024. URL https://arxiv.org/abs/2406.01361.

Chang Chen, Yi-Fu Wu, Jaesik Yoon, and Sungjin Ahn. Transdreamer: Reinforcement learning with transformer world models. *arXiv preprint arXiv:2202.09481*, 2022. URL https://arxiv.org/abs/2202.09481.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019. URL https://arxiv.org/abs/1904.10509.

Junha Chun, Youngjoon Jeong, and Taesup Kim. Sparse imagination for efficient visual world model planning. *arXiv preprint arXiv:2506.01392*, 2025. URL https://arxiv.org/abs/2506.01392.

Christian Gumbsch, Noor Sajid, Georg Martius, and Martin V. Butz. Learning hierarchical world models with adaptive temporal abstractions from discrete latent dynamics. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Z8CO8chIA7.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=1ikK0kHjvjG.

Danijar Hafner, Aravind Srinivas, Timothy Lillicrap, and Mohammad Norouzi. Mastering diverse domains through world models, 2023. URL https://arxiv.org/abs/2301.04104. arXiv preprint arXiv:2301.04104.

Anson Lei, Bernhard Schölkopf, and Ingmar Posner. Spartan: A sparse transformer learning local causation, 2024. URL https://arxiv.org/abs/2411.06890.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention, 2020. URL https://arxiv.org/abs/2006.15055.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models, 2023. URL https://arxiv.org/abs/2209.00588.

Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world models are happy with 100k interactions, 2023. URL https://arxiv.org/abs/2303.07109.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers, 2020. URL https://arxiv.org/abs/2003.05997.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models, 2023. URL https://arxiv.org/abs/2210.05861.

Weipu Zhang, Gang Wang, Jian Sun, Yetian Yuan, and Gao Huang. Storm: Efficient stochastic transformer based world models for reinforcement learning, 2023. URL https://arxiv.org/abs/2310.09615.