Dynamic Mechanistic Interpretability for Real-Time State Transitions in World Models: Tracking Thoughts in Motions

Ali Mahdipour Shirayeh
PMCRT, University Health Network & Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada

Mechanistic interpretability has made impressive strides in opening up the "black box" of neural networks, but existing methods like sparse autoencoders and activation patching share a fundamental limitation: they treat models as static snapshots. This creates a major blind spot when analyzing world models, such as JEPAs or model-based RL agents—whose internal representations must dynamically evolve as environments shift. In this paper, we introduce dynamic mechanistic interpretability (DMI), a novel framework built to track, isolate, and causally verify time-varying computational circuits in real time. At the core of our approach are 'temporal sparse autoencoders' (t-SAEs), which enforce smoothness constraints across consecutive time steps to extract persistent, interpretable primitive features. Building on this, our dynamic circuit discovery (DCD) algorithm applies sliding-window causal patching to continuously map how computational subgraphs reconfigure during online adaptation. We evaluate DMI on world models operating across non-stationary continuous control environments. Our findings demonstrate that DMI isolates state-dependent sub-networks using under 15% of total edge weights while outperforming static baselines in reconstruction faithfulness. Finally, we show that targeted real-time interventions on these dynamic circuits allow us to directly manipulate an agent's internal belief state. By moving interpretability from static snapshots to dynamic systems, DMI provides a missing blueprint for understanding and controlling adaptive AI.

Keywords: World models, Sparse autoencoder, Dynamic circuit discovery, Neural representations, Causal activation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.