EVENT-AWARE TENSOR-NETWORK INFERENCE FOR PATH RISK IN WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating whether an event occurs along a future trajectory requires information that accurate terminal-state prediction need not preserve. Matrix product states (MPS) can compress this computation for structured stochastic dynamics, but standard singular-value truncation minimizes local error without regard to its effect on the queried probability. We propose event-aware tensor-network inference: an approximate backward survival function weights forward truncation residuals by their effect on the target event probability, and a cost-aware rule reallocates bond dimension across space and time, with backward propagation, allocation, and recomputation charged to the inference budget. The resulting indicators track the true effect of each compression better than residual magnitude, especially under nonlocal coupling (median Spearman 0.68 vs. 0.39), and ablations confirm that both event sensitivity and the cost term contribute. On structured benchmarks with exact references, the method matches fixed-rank accuracy at equal or lower recorded cost, and in a visual path-event proof of concept it improves Brier score under moderate occlusion. Benefits depend on correlation structure and the evaluated scale is limited; exact propagation remains preferable on small systems.

1 INTRODUCTION

Visual world models learn predictive representations from images to support simulation and decision-making (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020). Object-centric representations and relational dynamics make this prediction problem more structured by separating entities and modeling their interactions (Battaglia et al., 2016; Kipf et al., 2018; Locatello et al., 2020). Building on these ideas, models such as STOVE, G-SWM, and SlotFormer predict the evolution of scenes through object-level representations (Kossen et al., 2020; Lin et al., 2020; Wu et al., 2023). For many applications, however, the quantity of interest is a particular future event: whether objects collide, a failure propagates, or an unsafe configuration is reached within a prediction window. Physical reasoning benchmarks emphasize this need to reason about events and their consequences (Yi et al., 2020; Bakhtin et al., 2019). Estimating such probabilities requires more than predicting an accurate terminal state. Two trajectories can end in similar configurations while differing in whether an unsafe event occurred earlier. Uncertainty in the observed state further induces multiple correlated future trajectories, whose contributions must be aggregated without counting repeated failures more than once. Even with a fixed dynamics model, computing a path-event probability can therefore be expensive. Monte Carlo offers a general solution, but the relative precision of ordinary sampling deteriorates as events become rare. Specialized estimators, including adaptive multilevel splitting, provide important alternatives (Cérou & Guyader, 2007; Bréhier et al., 2016).

Tensor networks offer another approach when the predictive distribution has compressible structure. Matrix product states (MPS), closely related to tensor trains, represent high-dimensional tensors through coupled low-dimensional cores (Oseledets, 2011; Schollwöck, 2011). Their bond dimensions control the capacity to represent dependencies across variable partitions. Although developed extensively in quantum many-body computation (Vidal, 2003), these tools also support classical stochastic propagation, trajectory statistics, and probabilistic inference (Johnson et al., 2010; Causer et al., 2022; Crotti & Braunstein, 2023). The relevant opportunity is classical computational compression; it does not require quantum hardware or complex probability amplitudes.

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.

A central question is how to allocate this limited capacity to a specified query. Standard canonical SVD truncation minimizes a local Euclidean approximation error at a prescribed rank. The magnitude of its discarded singular values, however, does not by itself determine the resulting error in a path-event probability. A small residual can align with a direction that strongly affects the event, while a larger residual may have little influence on that query. Goal-oriented numerical analysis addresses the analogous issue by weighting local residuals with adjoint sensitivities (Becker & Rannacher, 2001; Giles & Süli, 2002). This principle motivates allocating tensor capacity according to the downstream effect of compression, together with the cost of retaining additional rank.

We develop event-aware tensor-network inference for world models with finite object states and structured stochastic transitions (Figure 1). A visual belief initializes an unnormalized MPS that tracks the probability mass surviving the target event. An approximate backward survival function weights forward truncation residuals by their effect on the queried probability, and a cost-aware rule uses these indicators to revise bond dimensions across space and time, with backward propagation, allocation, and replay all charged to the inference cost. The formulation assumes a structured finite-state interface rather than a continuous latent world model. Isolated-truncation analysis and ablations show that event sensitivity ranks compression decisions by their true impact on the queried probability, most clearly under nonlocal coupling. On structured tasks with exact references, the method matches fixed-rank accuracy at equal or lower recorded runtime; in visual path-event prediction it improves over spectrum-adaptive MPS under moderate occlusion. The benefit is conditional: exact propagation remains preferable on small systems, and nonlocal interactions and retained visual uncertainty both raise cost. Our contributions are: 1. We formulate event-sensitive truncation indicators for tensorized survival propagation by combining forward residuals with approximate backward survival functions. 2.We develop a cost-aware procedure for allocating bond dimensions across space and time, accounting for the recomputation required to realize revised rank schedules. 3.We evaluate the resulting accuracy–cost trade-offs on structured dynamics and visual path-event prediction, and analyze component contributions, truncation sensitivity, and structural limitations.

2 Related Work

2.1 Object-Centric Visual World Models

Object-centric learning represents scenes through entities that can be tracked and composed. MONet and IODINE learn decompositions of images into object-level latent variables (Burgess et al., 2019; Greff et al., 2019), while Slot Attention provides an iterative mechanism for extracting object-centric representations (Locatello et al., 2020). Video models extend this approach to temporally consistent representations: SAVi uses motion-based supervision, SAVi++ incorporates depth prediction, and STEVE combines slots with a transformer decoder for complex video reconstruction (Kipf et al., 2022; Elsayed et al., 2022; Singh et al., 2022). These methods motivate structured visual interfaces, but learned slots alone do not specify normalized probabilities over the finite states required by our inference procedure. A complementary line learns how objects interact and evolve. Visual Interaction Networks infer physical dynamics from images (Watters et al., 2017); STOVE jointly models object states and dynamics (Kossen et al., 2020); and contrastive structured world models learn object-factorized latent transitions (Kipf et al., 2020). G-SWM studies object-centric generative prediction, while OP3 integrates entity representations with visual model-based control (Lin et al., 2020; Veerapaneni et al., 2020). SlotFormer models future object representations autoregressively and uses them for downstream reasoning and planning (Wu et al., 2023). Our focus is the subsequent inference problem: given a fixed belief and structured dynamics, how should computation be allocated to estimate a path-event probability? We therefore hold the predictive model fixed when comparing risk estimators, separating allocation effects from changes in representation learning.

2.2 Tensor Networks for Stochastic Inference

MPS and tensor trains provide compact representations of high-dimensional arrays through adaptive internal ranks (Oseledets, 2011; Schollwöck, 2011). Their use for classical probability distributions is well established. Johnson et al. (2010) adapt MPS evolution techniques to classical stochastic systems, and Barthel et al. (2018) compress trajectory-dependent messages for stochastic dynamics on networks. Tensor-train representations also support multivariate probability approximation and sampling (Dolgov et al., 2020). These works supply the representation and propagation foun-

dations of our method. Tensor networks have also been used for event-related quantities and rare trajectories. Causer et al. (2021) construct sampling dynamics using MPS approximations associated with dynamical large deviations, and Causer et al. (2022) extend tensor-network calculations to finite-time dynamical statistics. Matrix Product Belief Propagation addresses reweighted stochastic dynamics on graphs (Crotti & Braunstein, 2023). Related approaches approximate committor functions in tensor format (Chen et al., 2023) or construct importance distributions using tensor-train transformations (Cui et al., 2024). Thus, neither tensorized stochastic inference nor the use of event-related backward information is new in isolation. Our method combines approximate backward survival with local compression residuals to guide a cost-constrained rank schedule for finite-horizon path-event queries. Rank adaptation also arises in numerical methods for low-rank evolution and linear systems. The time-dependent variational principle projects dynamics onto an MPS manifold (Haegeman et al., 2011), while alternating minimal energy methods use residual information to enrich tensor-train approximations (Dolgov & Savostyanov, 2014). Our allocation objective is the accuracy of a specified event probability under limited computation. We retain ordinary singular-value ordering within each compression and adapt the rank assigned to different locations and times.

2.3 Rare-Event Estimation and Goal-Oriented Computation

Rare-event methods improve sampling efficiency by concentrating computation on consequential trajectories. Subset simulation decomposes a small failure probability into conditional probabilities of more accessible events (Au & Beck, 2001). Adaptive multilevel splitting progressively selects trajectories using an importance function (Cérou & Guyader, 2007), with unbiasedness results for generalized constructions under their stated conditions (Bréhier et al., 2016). Importance sampling and cross-entropy methods instead modify sampling distributions (Owen & Zhou, 2000; Rubinstein, 1999). These approaches are relevant comparators for tensor propagation; their performance depends on the event, proposal or reaction coordinate, and the associated adaptation cost. Our experiments characterize the evaluated configurations rather than establish a general ordering between sampling and tensor methods. The methodological basis of our indicator is goal-oriented error estimation. Dual-weighted residual methods assess numerical error through its effect on a target functional (Becker & Rannacher, 2001), and adjoint methods support both output-error analysis and correction (Giles & Süli, 2002). We apply this principle to realized tensor compression residuals and use the resulting indicators for computational allocation. Because the backward function is itself compressed, the indicators remain approximate and do not certify the final risk error.

Finally, numerical inference accuracy and predictive quality require different evaluations. Reference probabilities under a fixed model assess the former, whereas proper scoring rules evaluate probabilistic predictions against realized outcomes (Gneiting & Raftery, 2007). We use this distinction to organize the experiments: structured tasks isolate inference error, while Brier score and NLL assess visual event prediction under shared perception and dynamics models.

3 Methodology

We estimate path-event probabilities under a frozen structured world model. Visual observations initialize a belief over finite object states, which we propagate as a matrix product state (MPS). Our allocation strategy combines forward truncation residuals with backward event sensitivity to distribute bond dimensions under a computational budget.

3.1 Visual Beliefs and Path-Event Queries

Let $o_{\leq 0} = o_{-L+1:0}$ denote the observed context and $x = (x^1, \ldots, x^J) \in \mathcal{X} = \prod_{j=1}^{J} \mathcal{X}_j$ the state of a fixed set of objects, with $|\mathcal{X}_j| = d_j$. We use a mixture of product distributions as the perception–inference interface:

$$b_0(x \mid o_{\leq 0}) = \sum_{c=1}^{M} \pi_c(o_{\leq 0}) \prod_{j=1}^{J} b_{c,j}(x^j \mid o_{\leq 0}), \tag{1}$$

where π and each $b_{c,j}$ are normalized categorical distributions. The shared mixture index represents dependencies between objects and admits an MPS representation with bond dimension at most M. The inference procedure also accepts other initial beliefs supplied in MPS form.

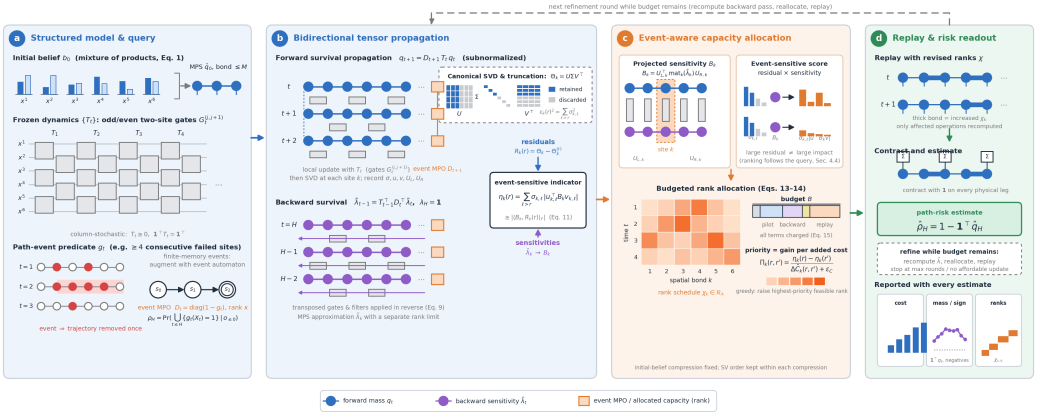

Figure 1: **Event-aware tensor-network inference.** (a) A structured world model: an initial belief b_0 in MPS form, frozen column-stochastic dynamics T_t built from odd/even layers of two-site gates, and a path-event predicate g_t realized as a diagonal MPO D_t. (b) Forward survival propagation $q_{t+1} = D_{t+1} T_t q_t$ with canonical SVD truncation records local residuals $R_k(r)$; an approximate backward survival function $\tilde{\lambda}$ supplies the sensitivity of ρ_H to each compression, and the two combine into the indicator $\eta_k(r)$. (c) A cost-aware rule allocates bond dimension across sites and time steps under budget B, ranking candidate increases by gain per added cost. (d) The affected propagation is replayed with the revised ranks and $\hat{\rho}_H = 1 - \mathbf{1}^\top \hat{q}_H$ is read out; setup, pilot, backward, allocation and replay costs are all charged. Heatmap and diagnostics are illustrative.

The frozen dynamics use column-stochastic operators T_t, with $[T_t]_{x',x} = p_\theta(X_{t+1} = x' \mid X_t = x)$. Thus $T_t \geq 0$ and $\mathbf{1}^\top T_t = \mathbf{1}^\top$. Given failure predicates $g_t : \mathcal{X} \to \{0, 1\}$, the target probability is

$$\rho_H = \Pr\left(\bigcup_{t=1}^{H} \{g_t(X_t) = 1\} \,\middle|\, o_{\leq 0}\right) = 1 - \mathbb{E}\left[\prod_{t=1}^{H} (1 - g_t(X_t)) \,\middle|\, o_{\leq 0}\right]. \tag{2}$$

Predicates are evaluated after complete transition steps; the initial state is excluded. Finite-memory events can be handled by augmenting the state with an event automaton. Our numerical target is the risk under the specified belief and dynamics, distinct from model error relative to the environment.

3.2 Tensorized Survival Propagation

Let $q_t(x)$ denote the joint probability of reaching state x without triggering the event by time t. With $D_t = \mathrm{diag}(1 - g_t)$,

$$q_0 = b_0, \qquad q_{t+1} = D_{t+1} T_t q_t, \qquad \rho_H = 1 - \mathbf{1}^\top q_H. \tag{3}$$

The mass of q_t is the survival probability. Filtering removes failed trajectories once, so q_t must remain subnormalized throughout propagation.

For chain-structured dynamics, a complete transition consists of odd- and even-edge layers:

$$T_t = T_t^{\mathrm{even}} T_t^{\mathrm{odd}}, \qquad T_t^a = \prod_{j \in \mathcal{B}_a} G_t^{(j,j+1)}, \quad a \in \{\mathrm{odd}, \mathrm{even}\}, \tag{4}$$

where each $G_t^{(j,j+1)}$ is a two-object stochastic gate and gates within a layer act on disjoint pairs. Event filtering follows both layers. When the safety predicate admits a finite-state scan,

$$1 - g_t(x) = \ell_t^\top W_t^{[1]}(x^1) \cdots W_t^{[J]}(x^J) r_t, \tag{5}$$

it defines a diagonal matrix product operator (MPO). Its bond dimension depends on the event; for consecutive failures, the auxiliary state tracks the current run length. We represent surviving mass as

$$\widehat{q}_t(x) = A_t^{[1]}(x^1) \cdots A_t^{[J]}(x^J), \qquad A_t^{[j]}(x^j) \in \mathbb{R}^{\chi_{t,j-1} \times \chi_{t,j}}, \tag{6}$$

with boundary dimensions equal to one. Gate and MPO applications are followed by canonical SVD compression. At compression site k, indexing time and location, write the updated unfolding as $Q_k = U_{L,k}\Theta_k U_{R,k}^\top$, with orthonormal environment matrices. For candidate rank r,

$$\begin{aligned} \Theta_k &= \sum_{\ell=1}^{s_k} \sigma_{k,\ell} u_{k,\ell} v_{k,\ell}^\top, \qquad \Theta_k^{(r)} = \sum_{\ell=1}^{r} \sigma_{k,\ell} u_{k,\ell} v_{k,\ell}^\top, \\ R_k(r) &= \Theta_k - \Theta_k^{(r)}, \qquad \epsilon_k(r)^2 = \|R_k(r)\|_F^2 = \sum_{\ell>r} \sigma_{k,\ell}^2. \end{aligned} \tag{7}$$

Canonicalization makes $\epsilon_k(r)$ equal to the Euclidean norm of the full-tensor perturbation. This residual measures discarded magnitude, but not its effect on the event probability.

3.3 Event-Sensitive Truncation Indicators

Expand propagation into elementary linear operations $L_1, \ldots, L_K$, including gates and event filters. The exact and compressed computations satisfy

$$z_k = L_k z_{k-1}, \qquad \widehat{z}_k = L_k \widehat{z}_{k-1} + e_k, \qquad z_0 = b_0, \quad z_K = q_H. \tag{8}$$

Define the backward survival function by $\lambda_K = \mathbf{1}$ and $\lambda_{k-1} = L_k^\top \lambda_k$. Telescoping the propagation error gives

$$\widehat{\rho}_H - \rho_H = -\lambda_0^\top e_0 - \sum_{k=1}^{K} \lambda_k^\top e_k, \qquad e_0 = \widehat{z}_0 - z_0. \tag{9}$$

This identity holds for the realized residuals even when ranks are selected adaptively. It identifies backward survival as the sensitivity of the final query to each compression.

We approximate λ_k by an MPS $\widetilde{\lambda}_k$, using reverse application of the transposed operators with a separate rank limit. Contracting it with the forward canonical environments gives

$$B_k = U_{L,k}^\top \operatorname{mat}_k(\widetilde{\lambda}_k) U_{R,k}. \tag{10}$$

For rank-r SVD compression, $\operatorname{mat}_k(e_k) = -U_{L,k} R_k(r) U_{R,k}^\top$. The approximate contribution to risk error and its allocation indicator are therefore

$$\begin{aligned} -\widetilde{\lambda}_k^\top e_k &= \langle B_k, R_k(r) \rangle_F, \\ \eta_k(r) &= \sum_{\ell>r} \sigma_{k,\ell} \left| u_{k,\ell}^\top B_k v_{k,\ell} \right| \;\geq\; \left| \langle B_k, R_k(r) \rangle_F \right|. \end{aligned} \tag{11}$$

Taking absolute values componentwise avoids cancellation within the discarded tail. The indicator selects where to increase rank; compression still retains the leading singular components in their standard order. Backward approximation introduces uncertainty:

$$\left| (\lambda_k - \widetilde{\lambda}_k)^\top e_k \right| \leq \|\lambda_k - \widetilde{\lambda}_k\|_2 \, \|e_k\|_2. \tag{12}$$

Accordingly, η_k is an allocation surrogate, not a certified error bound. It is nonincreasing in r for a fixed local environment, but need not predict monotonic improvement after replay.

3.4 Cost-Aware Spatiotemporal Bond Allocation

Let $\boldsymbol{\chi} = (\chi_1, \ldots, \chi_K)$ specify retained ranks at compression sites. Using the current forward and backward passes, we construct the budgeted surrogate

$$\min_{\chi_k \in \mathcal{R}_k} \sum_{k=1}^{K} \eta_k(\chi_k) \quad \text{subject to} \quad C_{\text{spent}} + \widehat{C}_{\text{remaining}}(\boldsymbol{\chi}) \leq B, \tag{13}$$

where $\mathcal{R}_k$ contains feasible candidate ranks and B is the inference budget. Initial-belief compression is held fixed. For a candidate increase $r \to r'$, we prioritize

$$\Pi_k(r, r') = \frac{\eta_k(r) - \eta_k(r')}{\widehat{\Delta C}_k(r, r') + \varepsilon_C}, \qquad \varepsilon_C > 0. \tag{14}$$

The nonnegative incremental cost estimate accounts for affected contractions and recomputation. We greedily select the highest positive feasible priority and update the proposed schedule and cost estimate. This is a local heuristic: changing one rank can alter downstream spectra and costs.

Revised ranks are realized by replaying from an unchanged checkpoint before the earliest affected operation, or from the initial belief. Increasing already truncated core dimensions alone cannot recover discarded information. All query-specific work contributes to measured cost:

$$C_{\text{total}} = C_{\text{setup}} + C_{\text{pilot}} + C_{\text{backward}} + C_{\text{allocation}} + C_{\text{replay}}. \tag{15}$$

These terms include required transfers, checkpoint operations, and repeated passes. Reusable calibration is separated from query latency and included when reporting first-use cost.

For uniform local dimension d and maximum forward rank χ, an MPS requires $O(Jd\chi^2)$ storage, and a sweep over the local gates across H steps costs $O(HJd^3\chi^3)$ using dense two-site SVDs. These estimates describe forward gate propagation, not the complete adaptive procedure. An event MPO of rank κ can enlarge intermediate bonds to $\kappa\chi$, while backward propagation and checkpointing incur additional computation and storage. We therefore evaluate efficiency using the measured total cost in Eq. equation 15, including construction of the sensitivity indicators and replay.

4 Experiments

We evaluate inference accuracy, computational cost, and contribution of event-sensitive allocation. Structured dynamics provide exact reference risks; visual experiments evaluate predictions against observed event outcomes. We then analyze individual components and computational limitations.

4.1 Experimental Setup

Structured tasks. We consider binary stochastic chains $x_t \in \{0,1\}^J$, initialized at the all-zero state. Each complete step applies odd- and even-edge layers of two-site stochastic gates. The target event is the occurrence of at least four consecutive failed sites within H steps:

$$p^\star = \Pr\Big(\exists\, t \in \{1,\dots,H\},\ j \in \{1,\dots,J-3\} : x_t^j = \cdots = x_t^{j+3} = 1\Big). \tag{16}$$

Events are checked after complete steps, excluding the initial state, and each trajectory is counted only once. The main local task uses $J = 20$ and $H = 50$. The topology comparison uses $J = 16$ and $H = 50$, adding one gate per step on edge $(7, 8)$, $(6, 10)$, or $(4, 12)$. Gate parameters, frequency, and update order are held fixed. Each topology has its own Exact DP reference.

Comparators and metrics. We compare Ours with Exact DP, streaming Monte Carlo (MC), generalized adaptive multilevel splitting (AMS), and canonical fixed-rank MPS. MC uses batches of 16,384 trajectories. AMS uses 2,048 particles, a longest-run reaction coordinate, and a nominal kill fraction of 0.1. For deterministic estimators, error is $100|\widehat{p} - p^\star|/p^\star$. For MC and AMS, we report relative RMSE, $100\sqrt{R^{-1}\sum_{r=1}^{R}(\widehat{p}^{(r)} - p^\star)^2}/p^\star$, over $R = 10$ independent runs. All reported reference risks are nonzero. Ours A and B identify the two local operating points; they are not different model architectures.

Measurement scope. Structured baseline measurements use CPU float64 arithmetic with one BLAS thread, Python 3.9.6, and NumPy 2.0.2 on macOS 15.6 (arm64). Exact DP and fixed-MPS times are medians of three isolated-process runs. Tables 1 and 2 report measured operating points rather than identical time budgets. Runtime comparisons are descriptive; repeated timings do not provide independent observations of deterministic inference accuracy. Visual protocols and component controls are specified in their respective subsections.

4.2 Path-Risk Inference on Structured Dynamics

Local dynamics. Table 1 reports the local task, with reference risk $p^\star \approx 9.13009 \times 10^{-3}$. Ours A and B achieve errors comparable to fixed MPS at $\chi = 4$ and $\chi = 8$, respectively, while reducing

Table 1: Local path-risk inference ($J = 20$, $H = 50$). Error is relative absolute error for deterministic methods and relative RMSE over ten runs for MC and AMS, expressed in percent. A and B denote the two evaluated Ours operating points. Rows are not matched in runtime.

Method	Configuration	Error (%) ↓	Time (s) ↓
Exact DP	Exact	0	6.434
MC	10^4 trajectories	6.51	0.315
MC	10^5 trajectories	4.02	3.047
MC	10^6 trajectories	1.02	30.902
AMS	2,048 particles	16.25	22.188
Fixed MPS	$\chi = 4$	0.194	1.287
Fixed MPS	$\chi = 8$	5.30×10^{-5}	1.398
Ours	A	0.186	1.113
Ours	B	4.70×10^{-5}	1.159

Table 2: Topology comparison ($J = 16$, $H = 50$). Distances $d = 1, 4, 8$ correspond to additional gates on $(7, 8)$, $(6, 10)$, and $(4, 12)$. Errors are relative absolute errors in percent. The respective reference risks are approximately 0.01209023, 0.00759351, and 0.00756458.

	$d = 1$		$d = 4$		$d = 8$	
Method	Error (%)	Time (s)	Error (%)	Time (s)	Error (%)	Time (s)
Exact DP	0	0.139	0	0.133	0	0.147
MPS, $\chi = 4$	0.109	0.986	9.121	1.043	13.801	1.109
MPS, $\chi = 8$	6.30×10^{-5}	1.097	0.0223	1.173	0.0828	1.283
MPS, $\chi = 16$	2.84×10^{-8}	1.214	3.52×10^{-7}	1.463	3.69×10^{-4}	1.787
Ours	5.50×10^{-5}	0.992	0.0212	1.088	0.0799	1.143

recorded runtime by 13.5% and 17.1%. The principal improvement at these operating points is therefore computational efficiency at similar accuracy.

The preferred estimator depends on the required precision. MC with 10^4 trajectories is the fastest reported configuration and attains 6.51% relative RMSE. Both MPS configurations and Ours attain errors below 1%, whereas MC with 10^6 trajectories reaches 1.02% relative RMSE at substantially higher cost. The evaluated AMS configuration is less competitive here; this finding concerns the specified reaction coordinate and implementation, rather than splitting methods in general.

Sensitivity to interaction topology. Table 2 shows that fixed MPS at $\chi = 4$ becomes less accurate for the tested longer-range interactions. Increasing capacity to $\chi = 8$ brings relative error below 0.1% in all three conditions. Ours achieves slightly lower errors than this baseline, with recorded runtime reductions of 9.6%, 7.2%, and 10.9% at interaction distances 1, 4, and 8.

Larger fixed ranks remain useful when higher precision is required: $\chi = 16$ achieves lower errors than Ours in all three conditions, at higher runtime. Exact DP is fastest at this system size. Moreover, changing the additional edge changes the dynamics and reference risk, so the comparison establishes sensitivity to these topologies rather than an isolated effect of interaction distance. Together, the results support a structure- and precision-dependent accuracy–cost trade-off. The contribution of the allocation rule is examined separately in Section 4.4.

4.3 Visual Path-Event Prediction

Task and evaluation. Given an observed video prefix $o_{-L+1:0}$ and an event query q, the task is to estimate the probability that the event occurs at least once in the next H steps. The corresponding outcome label is

$$y = \mathbb{I}[\exists\, t \in \{1, \ldots, H\} : g_q(x_t) = 1]\,. \tag{17}$$

All estimators share the visual front end and dynamics model. We evaluate clear, moderately occluded, and strongly occluded observations using Brier score and binary negative log-likelihood

Table 3: Visual path-event prediction. MC, spectrum-adaptive MPS, and Ours use multi-state beliefs unless marked otherwise. The MAP rows hold the inference algorithm fixed to Ours and replace the initial belief with a point state. Lower values are better; bold identifies the lowest reported prediction score in each observation condition.

Observation	Method / initialization	Brier	NLL	Time (s)
Clear	MC	0.0432	0.169	1.20
	Spectrum-adaptive MPS	0.0418	0.164	1.19
	Ours	**0.0416**	**0.163**	1.18
	Ours, MAP state	0.0419	0.165	0.95
Moderate occlusion	MC	0.0684	0.246	1.20
	Spectrum-adaptive MPS	0.0648	0.234	1.19
	Ours	**0.0627**	**0.228**	1.20
	Ours, MAP state	0.0738	0.272	0.96
Strong occlusion	MC	0.0921	0.318	1.20
	Spectrum-adaptive MPS	**0.0886**	0.307	1.20
	Ours	0.0889	**0.306**	1.19
	Ours, MAP state	0.1035	0.365	0.96

Table 4: Component ablations. Errors are relative absolute risk errors in percent. All recorded runtimes fall within the 1.20 s budget. Comparisons are made within each task.

	Local		Nonlocal	
Variant	Error (%)	Time (s)	Error (%)	Time (s)
Full method	**0.180**	1.18	**0.080**	1.19
Without event sensitivity	0.188	1.17	0.105	1.18
Without cost term	0.205	1.19	0.096	1.19
Shuffled sensitivity	0.199	1.18	0.118	1.17

(NLL). These scores measure prediction quality against realized outcomes, rather than numerical error against an exact model probability.

Prediction accuracy and visual uncertainty. Table 3 compares MC, spectrum-adaptive MPS, and Ours with multi-state belief initialization at recorded inference times near 1.2 s. Under moderate occlusion, Ours reduces Brier score and NLL by 3.2% and 2.6% relative to spectrum-adaptive MPS. Differences under clear observations are small. Under strong occlusion, Ours has slightly lower NLL but slightly higher Brier score, indicating no consistent advantage across both metrics.

Retaining initial-state uncertainty has a larger effect under occlusion. Relative to MAP initialization, the multi-state belief reduces Brier score by 15.0% under moderate occlusion and 14.1% under strong occlusion; NLL decreases by approximately 16.2% in both cases. These gains require approximately 24–25% more runtime. The results support an accuracy–cost trade-off for belief propagation, distinct from the benefit of the allocation rule. The reported point estimates do not establish statistical significance.

4.4 Ablation and Mechanism Analysis

Component contributions. We compare the complete method with variants that remove event sensitivity, remove the cost term, or shuffle sensitivity information across allocation decisions. Table 4 reports local and nonlocal tasks under a 1.20 s inference budget.

The complete method has the lowest reported error in both tasks. Relative to removing event sensitivity, it reduces error by 4.3% locally and 23.8% nonlocally, indicating a larger contribution in the tested nonlocal setting. Removing the cost term also degrades accuracy at comparable runtime. Shuffling sensitivity information yields the largest nonlocal error, consistent with the importance of assigning query information to the appropriate truncation decisions.

Table 5: Correlation with isolated truncation impact. Entries are median Spearman correlations across 20 instances per condition; higher is better.

Condition	Forward residual	Event sensitivity
Weak coupling	0.86	0.87
Local coupling	0.57	0.76
Nonlocal coupling	0.39	0.68

Sensitivity to truncation impact. We assess whether the indicators rank truncations according to their effect on the queried probability. For a single truncation decision k, we measure its isolated impact as $\Delta_k = |p^{(k)} - p^\star|$, where $p^{(k)}$ applies only that truncation and otherwise uses exact propagation. Table 5 reports median within-instance Spearman correlations across 20 independent instances per condition. The indicators perform similarly under weak coupling. Event sensitivity provides stronger agreement under local and nonlocal coupling, where residual magnitude alone is less informative about the target probability. Together with the ablations, this supports using query information for allocation. The isolated-truncation analysis does not, however, bound accumulated error after multiple compressions or establish allocation optimality.

4.5 Computational Behavior and Limitations

Computational regime. Table 6 summarizes fixed-rank propagation costs and capacity requirements. On the local task ($J = 16$, $H = 25$), MPS with $\chi = 8$ achieves lower error than MC with 10^5 samples, using approximately $2.7\times$ less time and $3.3\times$ less peak memory. Exact DP is $4.4\times$ faster than MPS at this size, whereas the reported MPS configurations are faster at $J = 20$ (Table 1). On the tested nonlocal instance, increasing χ substantially reduces error, highlighting the capacity requirements that adaptive allocation must accommodate.

Table 6: Fixed-rank computational behavior. (a) Local task with $J = 16$ and $H = 25$; MC error is the mean absolute error across repeated runs. Peak RSS denotes process resident memory; "—" indicates an unreported value. (b) Absolute risk error on a separate nonlocal instance. These measurements exclude adaptive allocation.

(a) Local task: accuracy and cost

Method	Absolute error	Time (s)	RSS (MB)
Exact DP	Reference	0.113	—
MC (10^5)	1.01×10^{-4}	1.342	91.9
MPS ($\chi = 8$)	1.37×10^{-9}	0.500	27.5

(b) Nonlocal task: rank dependence

Bond dimension χ	4	8	16	32
Absolute error	1.04×10^{-3}	6.26×10^{-6}	2.79×10^{-8}	6.96×10^{-11}

These memory measurements concern fixed-rank propagation and exclude the additional costs of adaptive allocation. Moreover, SVD truncation does not guarantee nonnegativity, and computational benefits depend on compressible correlations and tractable event operators.

5 Conclusion

We presented event-aware tensor-network inference for path risk in structured stochastic dynamics. Approximate backward information guides bond allocation toward truncations relevant to the queried event. The reported experiments suggest benefits in selected settings, while visual prediction results vary with occlusion. These findings support query-dependent compression, with its practical value constrained by correlation structure and the cost of adaptation. Extending the approach to larger systems beyond exact propagation, and to learned dynamics and risk-sensitive planning, remains the main direction for future work.

REFERENCES

Siu-Kui Au and James L. Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, 16(4):263–277, 2001. doi: 10.1016/S0266-8920(01)00019-4.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. PHYRE: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1908.05656.

Thomas Barthel, Caterina De Bacco, and Silvio Franz. Matrix product algorithm for stochastic dynamics on networks applied to nonequilibrium Glauber dynamics. *Physical Review E*, 97(1): 010104, 2018. doi: 10.1103/PhysRevE.97.010104.

Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://arxiv.org/abs/1612.00222.

Roland Becker and Rolf Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001. doi: 10.1017/S0962492901000010.

Charles-Edouard Bréhier, Maxime Gazeau, Ludovic Goudenège, Tony Lelièvre, and Mathias Rousset. Unbiasedness of some generalized adaptive multilevel splitting algorithms. *The Annals of Applied Probability*, 26(6):3559–3601, 2016. doi: 10.1214/16-AAP1185.

Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. MONet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019. URL https://arxiv.org/abs/1901.11390.

Luke Causer, Mari Carmen Bañuls, and Juan P. Garrahan. Optimal sampling of dynamical large deviations via matrix product states. *Physical Review E*, 103(6):062144, 2021. doi: 10.1103/PhysRevE.103.062144.

Luke Causer, Mari Carmen Bañuls, and Juan P. Garrahan. Finite time large deviations via matrix product states. *Physical Review Letters*, 128(9):090605, 2022. doi: 10.1103/PhysRevLett.128.090605.

Frédéric Cérou and Arnaud Guyader. Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications*, 25(2):417–443, 2007. doi: 10.1080/07362990601139628.

Yian Chen, Jeremy Hoskins, Yuehaw Khoo, and Michael Lindsey. Committor functions via tensor networks. *Journal of Computational Physics*, 472:111646, 2023. doi: 10.1016/j.jcp.2022.111646.

Stefano Crotti and Alfredo Braunstein. Matrix product belief propagation for reweighted stochastic dynamics over graphs. *Proceedings of the National Academy of Sciences*, 120(47):e2307935120, 2023. doi: 10.1073/pnas.2307935120.

Tiangang Cui, Sergey Dolgov, and Robert Scheichl. Deep importance sampling using tensor trains with application to a priori and a posteriori rare events. *SIAM Journal on Scientific Computing*, 46(1), 2024. doi: 10.1137/23M1546981.

Sergey Dolgov, Karim Anaya-Izquierdo, Colin Fox, and Robert Scheichl. Approximation and sampling of multivariate probability distributions in the tensor train decomposition. *Statistics and Computing*, 30(3):603–625, 2020. doi: 10.1007/s11222-019-09910-z.

Sergey V. Dolgov and Dmitry V. Savostyanov. Alternating minimal energy methods for linear systems in higher dimensions. *SIAM Journal on Scientific Computing*, 36(5):A2248–A2271, 2014. doi: 10.1137/140953289.

Gamaleldin F. Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael C. Mozer, and Thomas Kipf. SAVi++: Towards end-to-end object-centric learning from real-world videos. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2206.07764.

Michael B. Giles and Endre Süli. Adjoint methods for PDEs: A posteriori error analysis and post-processing by duality. *Acta Numerica*, 11:145–236, 2002. doi: 10.1017/S096249290200003X.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2424–2433. PMLR, 2019. URL https://proceedings.mlr.press/v97/greff19a.html.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Jutho Haegeman, J. Ignacio Cirac, Tobias J. Osborne, Iztok Pižorn, Henri Verschelde, and Frank Verstraete. Time-dependent variational principle for quantum lattices. *Physical Review Letters*, 107(7):070601, 2011. doi: 10.1103/PhysRevLett.107.070601.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.01603.

T. H. Johnson, S. R. Clark, and D. Jaksch. Dynamical simulations of classical stochastic systems using matrix product states. *Physical Review E*, 82(3):036702, 2010. doi: 10.1103/PhysRevE.82.036702.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2688–2697. PMLR, 2018. URL https://proceedings.mlr.press/v80/kipf18a.html.

Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.12247.

Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2111.12594.

Jannik Kossen, Karl Stelzner, Marcel Hussing, Claas Voelcker, and Kristian Kersting. Structured object-aware physics prediction for video modeling and planning. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1910.02425.

Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, and Sungjin Ahn. Improving generative imagination in object-centric world models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6140–6149. PMLR, 2020. URL https://proceedings.mlr.press/v119/lin20f.html.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2006.15055.

Ivan V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011. doi: 10.1137/090752286.

Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000. doi: 10.1080/01621459.2000.10473909.

Reuven Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 1(2):127–190, 1999. doi: 10.1023/A:1010091220143.

Ulrich Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Annals of Physics*, 326(1):96–192, 2011. doi: 10.1016/j.aop.2010.09.012.

Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2205.14065.

Rishi Veerapaneni, John D. Co-Reyes, Michael Chang, Michael Janner, Chelsea Finn, Jiajun Wu, Joshua Tenenbaum, and Sergey Levine. Entity abstraction in visual model-based reinforcement learning. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 1439–1456. PMLR, 2020. URL https://proceedings.mlr.press/v100/veerapaneni20a.html.

Guifré Vidal. Efficient classical simulation of slightly entangled quantum computations. *Physical Review Letters*, 91(14):147902, 2003. doi: 10.1103/PhysRevLett.91.147902.

Nicholas Watters, Andrea Tacchetti, Theophane Weber, Razvan Pascanu, Peter Battaglia, and Daniel Zoran. Visual interaction networks: Learning a physics simulator from video. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.01433.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. SlotFormer: Unsupervised visual dynamics simulation with object-centric models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.05861.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1910.01442.

A Appendix

A Goal-Oriented Error Analysis

A.1 Exact Error Decomposition

Let $L_1, \ldots, L_K$ denote the linear operations in survival propagation, including transition gates and event filters. The exact and compressed computations satisfy

$$z_k = L_k z_{k-1}, \qquad \widehat{z}_k = L_k \widehat{z}_{k-1} + e_k, \tag{18}$$

where e_k is the realized compression residual. Define $\delta_k = \widehat{z}_k - z_k$ and $e_0 = \widehat{z}_0 - z_0$. The backward quantities satisfy

$$\lambda_K = \mathbf{1}, \qquad \lambda_{k-1} = L_k^\top \lambda_k. \tag{19}$$

Since $\lambda_k^\top \delta_k = \lambda_{k-1}^\top \delta_{k-1} + \lambda_k^\top e_k$, telescoping gives

$$\boxed{\widehat{\rho}_H - \rho_H = -\lambda_0^\top e_0 - \sum_{k=1}^{K} \lambda_k^\top e_k} \tag{20}$$

for $\rho_H = 1 - \mathbf{1}^\top z_K$. The initial term vanishes when the initial belief is represented exactly. The identity applies to the realized residuals even when ranks are selected adaptively.

A.2 Event-Sensitive Local Indicators

At a canonical compression site, the pre-truncation unfolding is $U_L \Theta U_R^\top$, where both environment matrices have orthonormal columns. For the singular value decomposition $\Theta = \sum_\ell \sigma_\ell u_\ell v_\ell^\top$, the residual after retaining rank r is

$$R_k(r) = \sum_{\ell > r} \sigma_\ell u_\ell v_\ell^\top, \qquad \mathrm{mat}(e_k) = -U_L R_k(r) U_R^\top. \tag{21}$$

Projecting the approximate backward state into these local bases gives

$$B_k = U_L^\top \,\mathrm{mat}(\widetilde{\lambda}_k) U_R. \tag{22}$$

The corresponding query-error contribution obeys

$$\begin{aligned} -\widetilde{\lambda}_k^\top e_k &= \langle B_k, R_k(r) \rangle_F, \\ |\langle B_k, R_k(r) \rangle_F| &\le \underbrace{\sum_{\ell > r} \sigma_\ell \left| u_\ell^\top B_k v_\ell \right|}_{\eta_k(r)}. \end{aligned} \tag{23}$$

Thus, $\eta_k(r)$ weights discarded singular directions by their relevance to the event query. Rank selection changes how many leading singular components are retained; it does not reorder them.

We compute B_k through left and right tensor environments rather than reconstructing the full state unfolding. The contractions depend on both forward and backward ranks, and their cost belongs to the adaptive computation.

A.3 Approximation Scope

The error introduced by the approximate backward state satisfies

$$\left| (\lambda_k - \widetilde{\lambda}_k)^\top e_k \right| \le \|\lambda_k - \widetilde{\lambda}_k\|_2 \|e_k\|_2. \tag{24}$$

Accordingly, the practical indicator is not a certified bound on total risk error without additional control of backward approximation. Changing an earlier rank also changes later states and residuals. Monotonicity of $\eta_k(r)$ for a fixed local decomposition therefore does not imply monotonic improvement of the final estimate.

B PATH-EVENT OPERATOR

B.1 SPATIAL AUTOMATON

The structured event is the occurrence of at least $m = 4$ consecutive failed sites. A binary state $x^j = 1$ denotes failure. To represent the complementary safety predicate, we use automaton states $a \in \{0, \dots, m-1\}$ that record the trailing failure-run length.

For $a, b \in \{0, \dots, m-1\}$, define

$$[A(0)]_{ab} = \mathbb{I}\{b = 0\}, \qquad [A(1)]_{ab} = \mathbb{I}\{b = a + 1,\ a < m - 1\}. \tag{25}$$

A transition completing a run of length m has zero weight. Let e_0 denote the basis vector for automaton state zero. The safety indicator is

$$s(x) = e_0^\top A(x^1) \cdots A(x^J)\mathbf{1}. \tag{26}$$

It equals one precisely when the configuration contains no run of length m.

The diagonal operator $D = \mathrm{diag}(s)$ admits an MPO representation with local cores $W^{x,x'} = \mathbb{I}\{x = x'\}A(x)$ and bond dimension at most m. The boundary vectors are absorbed into the first and last cores. This compact construction is specific to the event structure; arbitrary predicates need not admit a similarly small MPO.

B.2 TEMPORAL EVENT SEMANTICS

An event is evaluated after each complete time step, following both odd- and even-edge updates. The query covers $t = 1, \dots, H$ and excludes the initial state and intermediate substeps. Surviving probability mass evolves according to

$$q_0 = b_0, \qquad q_{t+1} = DT_t q_t, \qquad \rho_H = 1 - \mathbf{1}^\top q_H. \tag{27}$$