A Symmetric and Object-centric World Model for Stochastic Environments

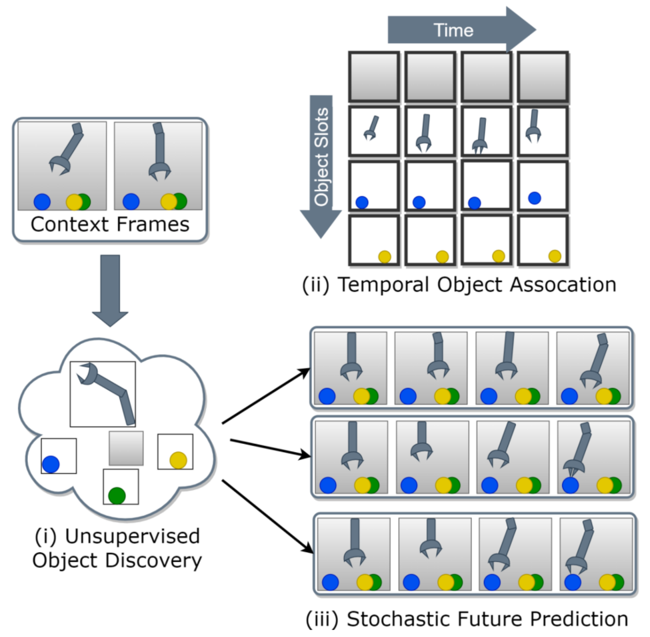
Patrick Emami, Pan He, Anand Rangarajan, Sanjay Ranka

UNIVERSITY of FLORIDA

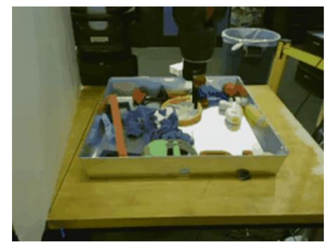
https://github.com/pemami4911/symmetric-and-object-centric-world-models

What's the goal?

Train an object-centric world model on the task shown on the right for **real-world**, **stochastic** environments.
The learned latent spatiotemporal object-centric representations (ii) can be re-used, e.g., for visual model-based RL.

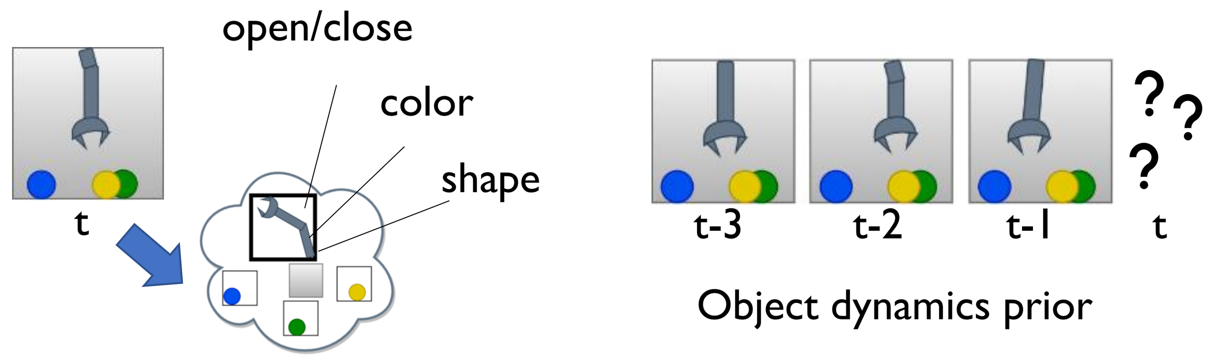


Why is real world + stochastic hard?



- Must handle complex object morphologies. With perceptual grouping (aka segmentation)?
- The predictive model of the world must account for many possible futures (e.g., due to dynamics uncertainty)
- We want to capture inductive biases like natural symmetries; e.g., learning a single dynamics model shared by all objects

Why should we think critically about latent SSM design for the world model?



These two distributions serve distinct functions for the object-centric world model and their variances fit different aspects of environment stochasticity!

What do we do?

1. We propose a latent SSM in which the variance for the discovery and dynamics distributions are learned separately
2. We also introduce the best-of-many rollouts (BMR) training objective for fitting the dynamics variance
3. Demonstrate the world model's effectiveness on real-world robotic manipulation videos with noisy actions

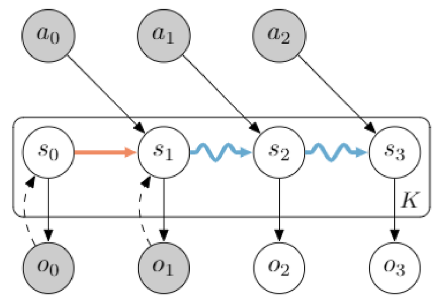
State space model (SSM) & objectivez

Conditional latent variable model with K object latents at each time step + separate discovery and dynamics priors:

$$p(o_{T\le t\le H}, \mathbf{s}_{\le T+H} \mid o_{<T}, a_{\le T+H-1})$$
$$= p_O(\mathbf{s}_0 \mid o_0) \prod_{t=1}^{T-1} p_O(\mathbf{s}_t \mid o_t, \mathbf{s}_{t-1}, a_{t-1}) \prod_{t=T}^{T+H} p(o_t \mid \mathbf{s}_t) p_D(\mathbf{s}_t \mid \mathbf{s}_{t-1}, a_{t-1}).$$

Gaussian discovery prior

- Uses past information to predict K means with a relational net
- The means are used to initialize a discovery posterior mean during iterative inference which helps *associate objects over time*
- Variance is learned as model parameter **fixed** across time steps $t < T$ and slots



Gaussian latent (unimodal) dynamics

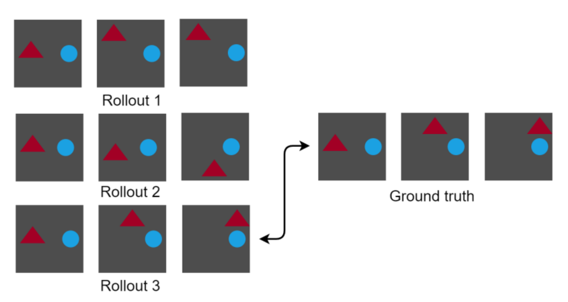
- Shares its K means with discovery prior
- K variances are predicted at each time step $t \le T \le H$ by the relational net

At time steps $0 < t < T$ and at the i^{th} step of iterative inference we have:

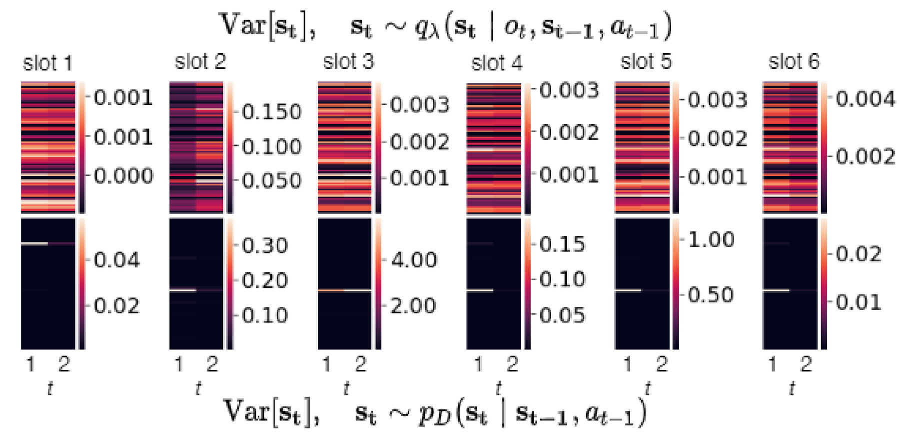
$$\mathcal{L}^{(i)}_{t,\text{o.d.}} = -\mathbb{E}_{\mathbf{s}^{(i)}_t \sim \mathcal{N}(\boldsymbol{\lambda}^{(i)}_t)}[\log p(o_t \mid \mathbf{s}^{(i)}_t)] + D_{KL}(\mathcal{N}(\boldsymbol{\lambda}^{(i)}_t) \parallel \underbrace{p_O(\mathbf{s}_t \mid \mathbf{s}_{t-1}, a_{t-1})}_{\text{Gaussian discovery prior}})$$

$$\mathcal{L}_{\text{BMR}} = \sum_{t=0}^{T-1}\Big(\sum_{i=1}^{I} \frac{i}{I}\mathcal{L}^{(i)}_{t,\text{o.d.}}\Big) - \max_j \Big\{\sum_{t=T}^{T+H} \mathbb{E}[\log p(o_t \mid \mathbf{s}^{(j)}_t)]\Big\}_{j=1}^{J}$$

We propose a variational objective that combines an object discovery loss with a sampling-based dynamics loss for future rollouts



Variances in the SSM



We show the variance for each latent attribute for $K = 6$ 64-dim slots at steps $t = 1, 2$ of a video.
The object discovery posterior variance (top) is ~uniform and has low magnitude across latent units and slots.
The dynamics model (bottom) learns to only predict high variance for latent attributes that may change over time.

Diverse generation

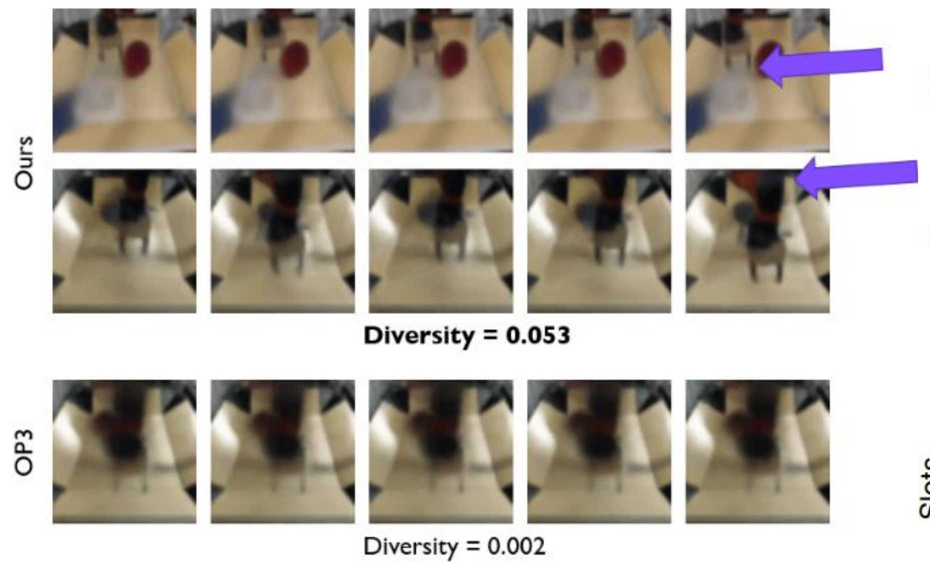
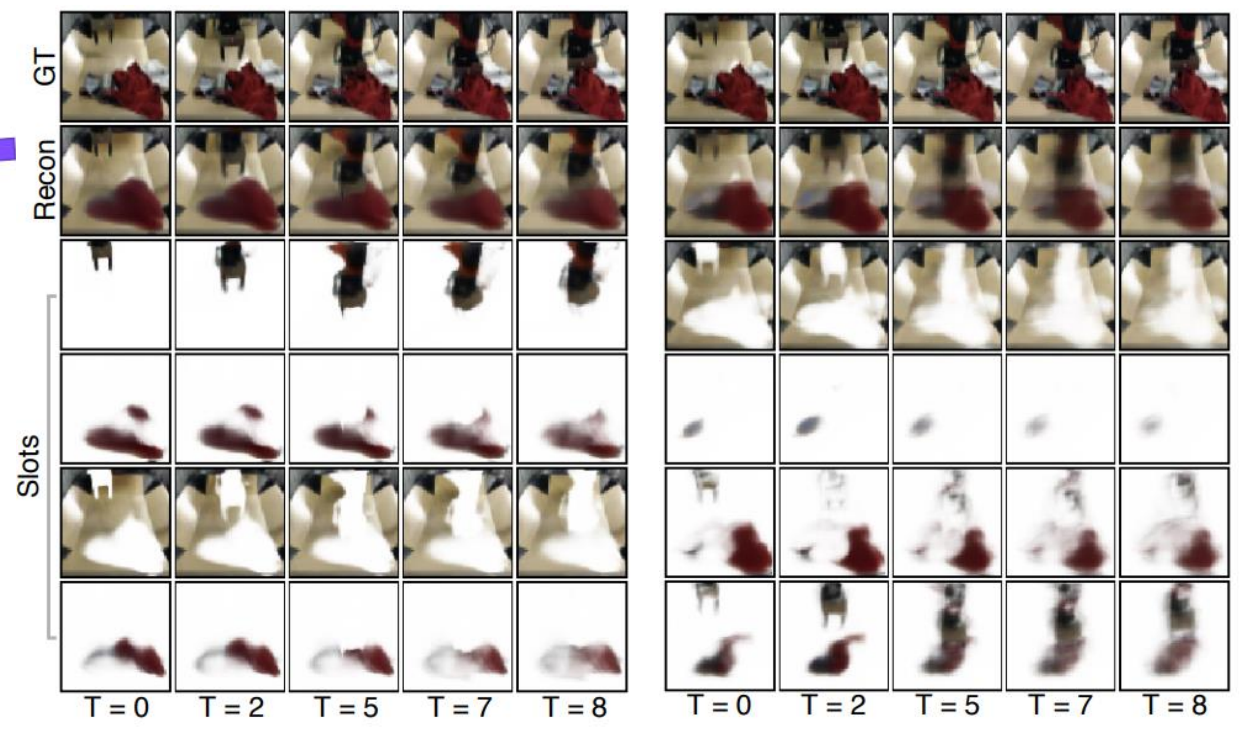


Table 1: BAIR (Realism / Diversity / Accuracy)

Model	FVD (↓)	(Best - Worst)$_{100}$ SSIM (↑)	SSIM / PSNR (↑)
VRNN†	472.5 ± 15.2	0.089	0.72 / 19.72
OP3	642.3 ± 27.2	0.002	0.76 / 21.61
Ours	**564.8 ± 24.3**	**0.053**	**0.79 / 22.39**

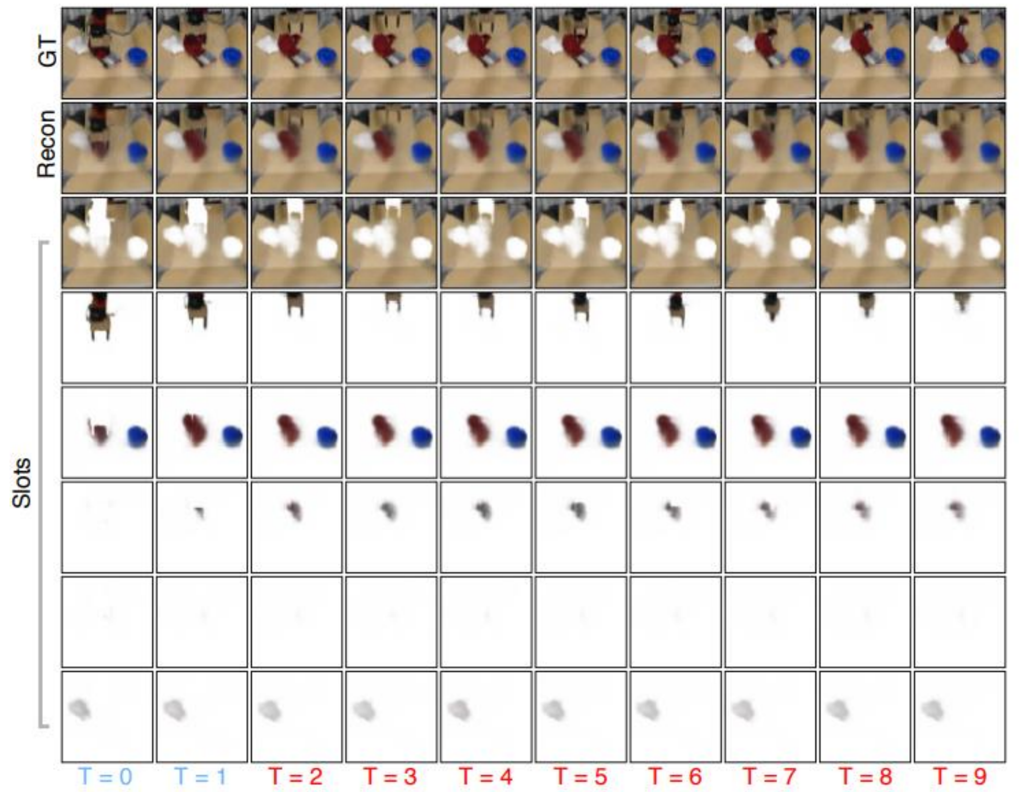
† No object discovery

Object-centric decompositions

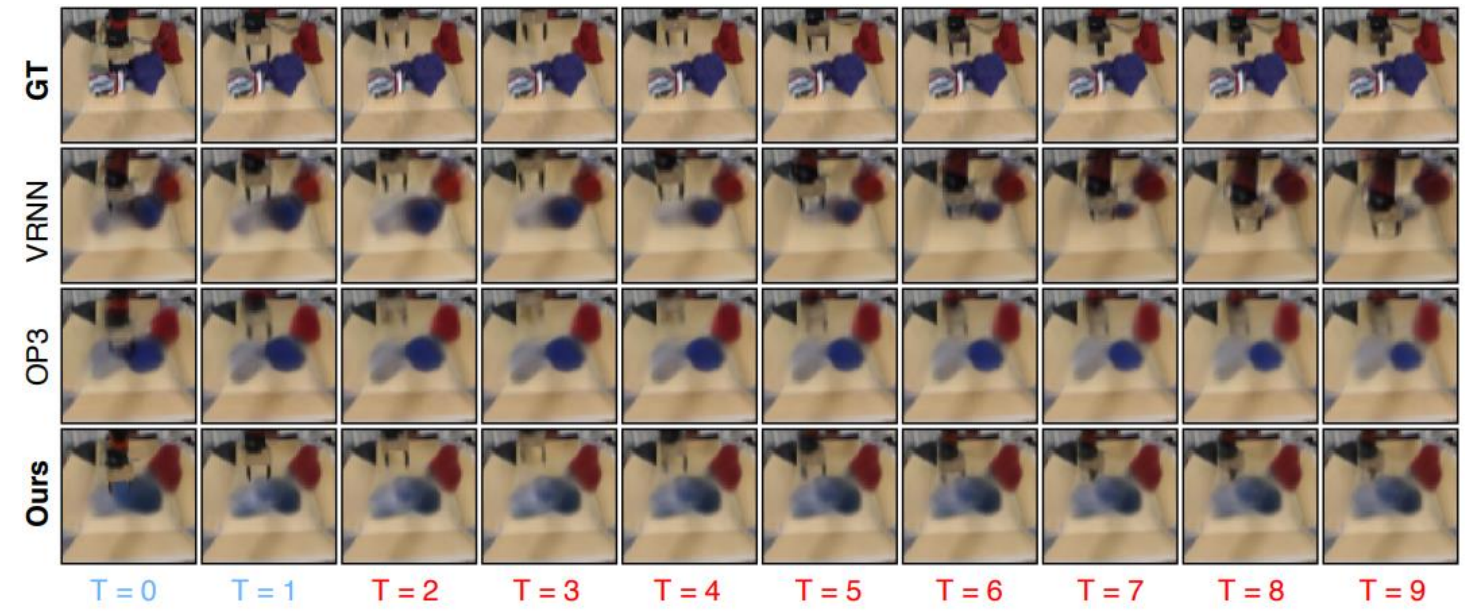


Object-centric decompositions (con't)

- **Background:** Segmented into the first slot by setting std. dev. to 0.09 and other slots' std. dev. to 0.11
- **Cloth:** Scenes contain multiple cloth items that are non-rigid, of highly variable size/shape, and with complex patterns. This leads the model to occasionally split them across slots or join two into one slot
- The multi-modal uncertainty over possible futures grows over time, causing blurriness to worsen at the end of rollouts



Rollout samples



Conclusions

- Next steps
  - Training with larger batch sizes and more steps for all models --- should lead to sharper rollouts
  - Add ablations
    - RSSM, J, replacing discovery prior with the dynamics model, not sharing the discovery & dynamics prior means, time-dependent discovery variance vs. time-independent (current), …
  - Multi-modal dynamics
  - BMR objective theoretical analysis
  - More environments and baselines
- Takeaways
  - We have introduced a perceptual-grouping based world model for real-world and stochastic environments
  - The proposed model combined with the BMR objective demonstrates an improvement in realism, accuracy and diversity of rollouts over OP3
  - Releasing a longer version for a journal soon!