GenParticles: Probabilistic Particle-Based Modeling for Object-Centric Motion

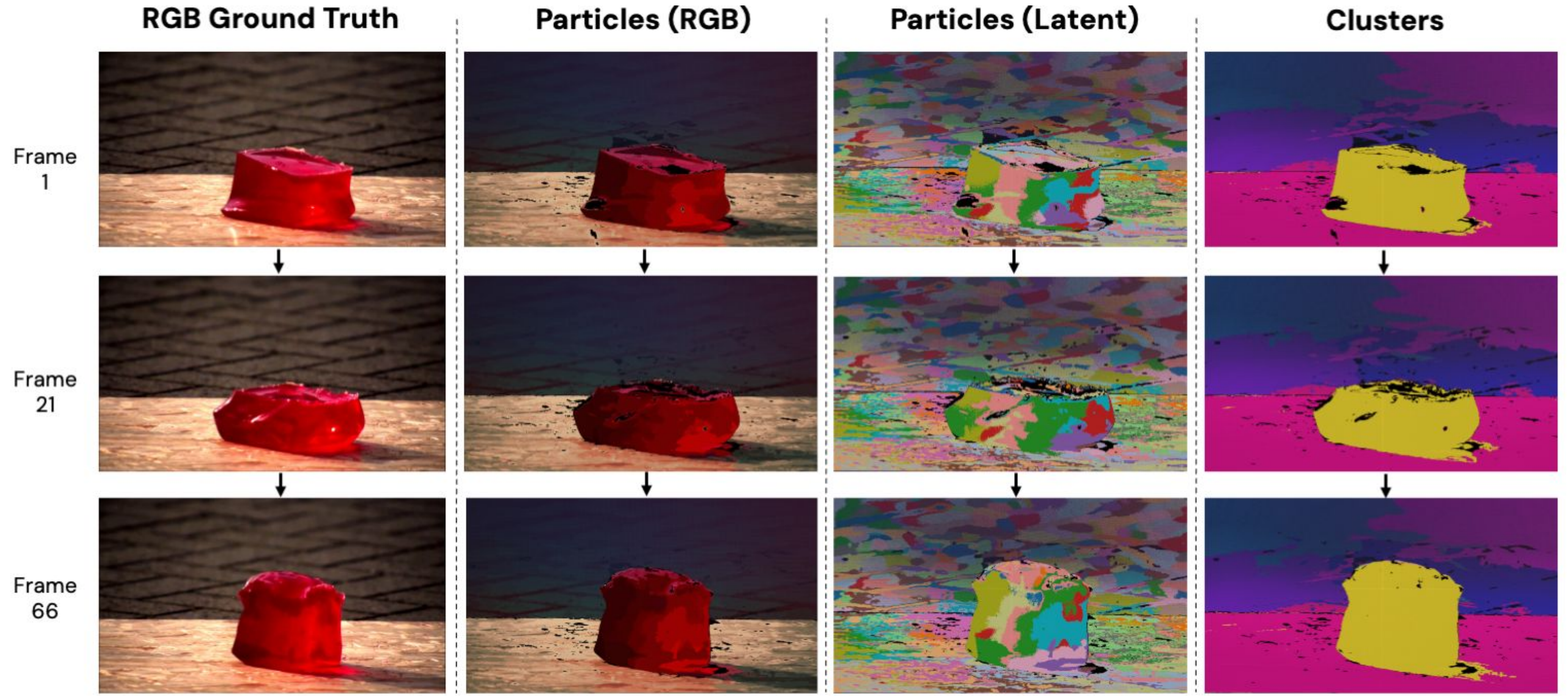
Arijit Dasgupta*, Eric Li*, Mathieu Huot, William T. Freeman, Vikash Mansinghka, Joshua B. Tenenbaum

Structured World Models for Robotic Manipulation

RSS 2025 | Los Angeles, USA | June 21, 2025

Introduction

- Object **motion** and **structure** inference from visual input is key for robotic manipulation in dynamic, disturbed or cluttered scenes.
- Humans use motion cues to infer object identity and deformation, inspiring models for flexible, persistent representations of objects.



Generative Particle Model

Algorithm 1 Generative Particle Model

Input:
K, L, N ▷ Number of clusters, particles, and observed data points
Priors: $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $(\boldsymbol{\mu}^{\mathcal{H}}, \sigma^2_{\mu^{\mathcal{H}}}, \boldsymbol{\Psi}^{\mathcal{H}}, \nu^{\mathcal{H}})$, $(\boldsymbol{\Psi}^{\mathcal{B}}, \nu^{\mathcal{B}})$, σ^2_V, $(\boldsymbol{\Psi}^{\mathcal{V}}, \nu^{\mathcal{V}})$
Sample cluster weights: $\boldsymbol{\pi}^{\mathcal{H}} \sim \mathrm{Dir}(\boldsymbol{\alpha})$
Sample particle weights: $\boldsymbol{\pi}^{\mathcal{B}} \sim \mathrm{Dir}(\boldsymbol{\beta})$
for $k = 1$ to K **do**
  Sample cluster covariance: $\boldsymbol{\Sigma}^{\mathcal{H}}_k \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}^{\mathcal{H}}, \nu^{\mathcal{H}})$
  Sample cluster mean: $\boldsymbol{\mu}^{\mathcal{H}}_k \sim \mathcal{N}(\boldsymbol{\mu}^{\mathcal{H}}, \sigma^2_{\mu^{\mathcal{H}}}\mathbf{I})$
  Sample cluster translation: $\mathbf{t}_k \sim \mathrm{DiscreteNormal}(\mathbf{0}, s^2\mathbf{I})$
  Sample cluster rotation: $\mathbf{R}_k \sim \mathrm{DiscreteVMF}(\kappa^{\mathrm{vmf}}, \theta_{\max})$
end for
for $\ell = 1$ to L **do**
  Sample cluster assignment: $z^{\mathcal{H}}_\ell \sim \mathrm{Cat}(\boldsymbol{\pi}^{\mathcal{H}})$
  Let $k = z^{\mathcal{H}}_\ell$
  Sample particle covariance: $\boldsymbol{\Sigma}^{\mathcal{B}}_\ell \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}^{\mathcal{B}}, \nu^{\mathcal{B}})$
  Sample particle mean: $\boldsymbol{\mu}^{\mathcal{B}}_\ell \sim \mathcal{N}(\boldsymbol{\mu}^{\mathcal{H}}_k, \boldsymbol{\Sigma}^{\mathcal{H}}_k)$
  Compute cluster-induced velocity: $\bar{\mathbf{v}}_\ell = \mathbf{t}_k + (\mathbf{R}_k - \mathbf{I})(\boldsymbol{\mu}^{\mathcal{B}}_\ell - \boldsymbol{\mu}^{\mathcal{H}}_k)$
  Sample particle velocity mean: $\mathbf{v}_\ell \sim \mathcal{N}(\bar{\mathbf{v}}_\ell, \sigma^2_V\mathbf{I})$
  Sample particle velocity covariance: $\boldsymbol{\Sigma}^{\mathcal{V}}_\ell \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}^{\mathcal{V}}, \nu^{\mathcal{V}})$
end for
for $n = 1$ to N **do**
  Sample particle assignment: $z^{\mathcal{B}}_n \sim \mathrm{Cat}(\boldsymbol{\pi}^{\mathcal{B}})$
  Let $\ell = z^{\mathcal{B}}_n$
  Sample data point position: $\mathbf{x}_n \sim \mathcal{N}(\boldsymbol{\mu}^{\mathcal{B}}_\ell, \boldsymbol{\Sigma}^{\mathcal{B}}_\ell)$
  Sample data point velocity: $\mathbf{v}_n \sim \mathcal{N}(\mathbf{v}_\ell, \boldsymbol{\Sigma}^{\mathcal{V}}_\ell)$
end for

Two-layer hierarchical Gaussian mixture model of **particles** and **data points** with instantaneous motions and rigidity constraints

Clusters: Globally coherent, rigidly moving particle groups

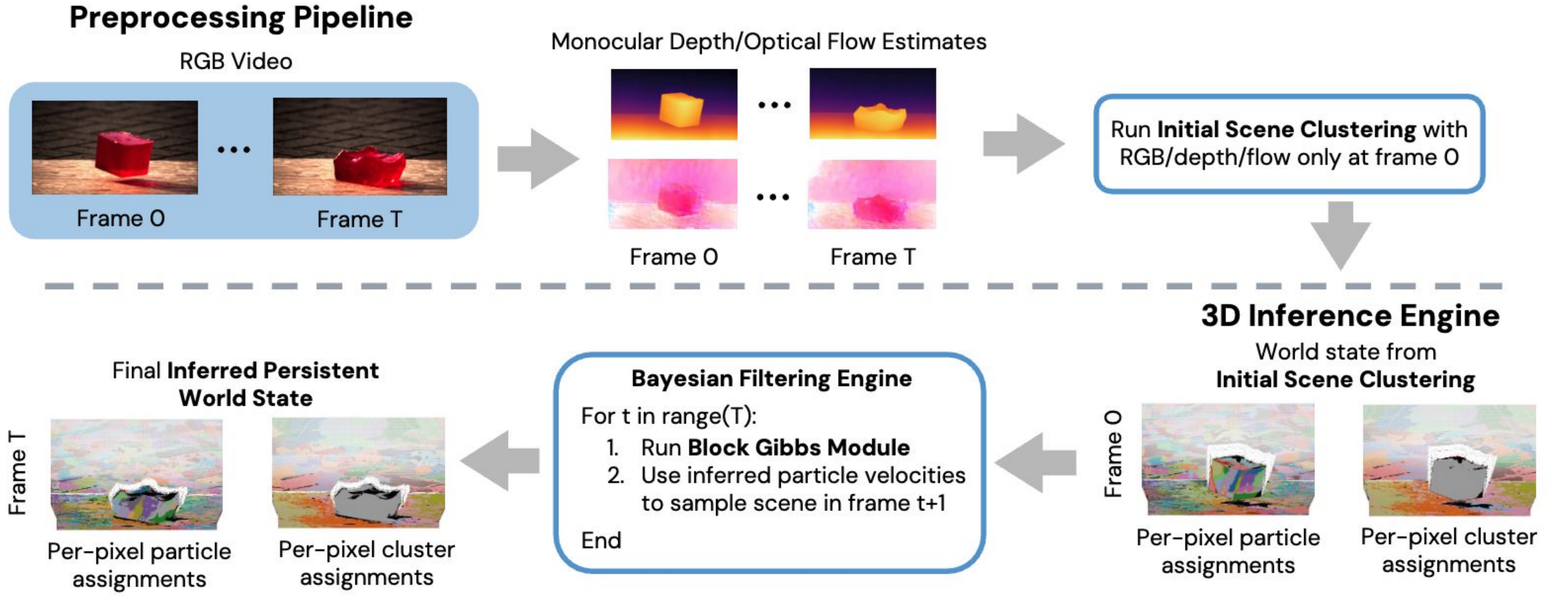
Particles: Localized components assigned to clusters, capturing spatial variability and approximate local translational motion

Data Points: point-level positions and instantaneous velocity

*Equal contribution

Approximate Inference

GenParticles performs sequential inference using a blocked Gibbs sampler that propagates <u>*inferred particles from the last time-step*</u> forward, assigns data points by spatial proximity, updates particle parameters, then infers cluster parameters in a strict order to ground higher-level structure while preserving object consistency with fixed assignments and covariances.



Blocked Gibbs Sampling

Latent Assignments

$$p(z^{\mathcal{B}}_n = \ell \mid \cdots) \propto \pi^{\mathcal{B}}_\ell \cdot \mathcal{N}(\mathbf{x}_n \mid \mu^{\mathcal{B}}_\ell, \Sigma^{\mathcal{B}}_\ell) \cdot \mathcal{N}(\mathbf{v}_n \mid \mathbf{v}_\ell, \Sigma^{\mathcal{V}}_\ell)$$

$$p(z^{\mathcal{H}}_\ell = k \mid \cdots) \propto \pi^{\mathcal{H}}_k \cdot \mathcal{N}(\mu^{\mathcal{B}}_\ell \mid \mu^{\mathcal{H}}_k, \Sigma^{\mathcal{H}}_k) \cdot \mathcal{N}(\mathbf{v}_\ell \mid \bar{\mathbf{v}}_{\ell,k}, \sigma^2_V I)$$

Mixture weights are updated via Dirichlet-Categorical Conjugacy

Particle Updates

Velocity and Spatial Covariances via *Normal Inverse-Wishart* Conjugacy $\quad \boldsymbol{\Sigma}^{\mathcal{B}}_\ell \quad \boldsymbol{\Sigma}^{\mathcal{V}}_\ell$

Velocities and Positions via *Normal-Normal* Conjugacy $\quad \boldsymbol{\mu}^{\mathcal{B}}_\ell \quad \mathbf{v}_\ell$

Cluster Updates

Rotations and Translations are sampled via full enumeration $\quad \mathbf{R}_k, \mathbf{t}_k$

All inference moves are functionally parallelized and matrix operations are vectorized via **GenJAX** *and* **JAX**

Object Persistence in Video

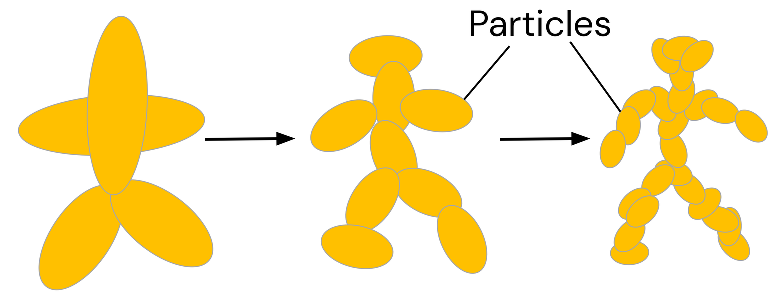
- We evaluate **GenParticles** on 33 single-object DAVIS videos by measuring object particle persistence, which is the average percentage of particles remaining inside the segmentation mask over time relative to their initialization, running on a single NVIDIA L4 24GB GPU.
- We compare against two *state-of-the-art* particle video baselines, CoTracker3 and SpaTracker, which run offline using full video access and structured 25×25 particle grids to represent objects, constrained by the same GPU memory limits for fair comparison.

DAVIS Video	GenParticles (Ours)	CoTracker3	SpaTracker	DAVIS Video	GenParticles (Ours)	CoTracker3	SpaTracker
boat	**100.00 ± 0.00**	90.69	92.14	bus	**93.33 ± 0.52**	82.16	82.79
car-turn	**100.00 ± 0.00**	97.15	99.89	dance-jump	**93.09 ± 1.80**	88.46	85.21
drift-chicane	**100.00 ± 0.00**	74.84	51.28	dog	92.78 ± 2.45	96.34	**97.09**
car-roundabout	**99.79 ± 0.25**	96.50	97.78	dance-twirl	**92.64 ± 1.12**	83.87	90.83
flamingo	**99.53 ± 0.42**	80.83	90.98	mallard-water	92.56 ± 1.00	**97.51**	94.28
breakdance-flare	**99.51 ± 0.25**	82.19	98.74	goat	**92.26 ± 2.82**	88.73	88.61
camel	**99.40 ± 0.53**	93.45	96.34	koala	**91.46 ± 1.03**	63.95	57.65
cows	**99.28 ± 0.89**	94.25	93.02	lucia	87.73 ± 2.18	93.77	**97.72**
rallye	98.96 ± 0.66	**100.00**	80.00	dog-agility	**84.78 ± 1.50**	78.65	74.20
rollerblade	**97.96 ± 1.92**	95.78	94.64	libby	**83.35 ± 3.19**	77.78	79.23
rhino	**97.50 ± 0.39**	88.19	87.87	parkour	78.57 ± 1.72	**84.96**	76.00
blackswan	96.89 ± 0.69	99.91	**100.00**	mallard-fly	69.23 ± 3.03	71.01	**85.27**
bear	**96.72 ± 1.29**	95.87	94.28	drift-turn	64.75 ± 2.76	**96.46**	91.47
elephant	**96.71 ± 1.11**	89.82	90.14	drift-straight	63.75 ± 12.47	**94.17**	74.00
breakdance	**96.31 ± 1.21**	69.06	94.83	***varanus-cage***	51.93 ± 9.78	**67.19**	64.93
hike	94.36 ± 2.56	92.67	**94.46**	***soccerball***	11.85 ± 0.22	**88.94**	87.22
car-shadow	94.11 ± 0.63	96.08	**97.28**	**Median Accuracy**	**94.11**	89.82	90.98

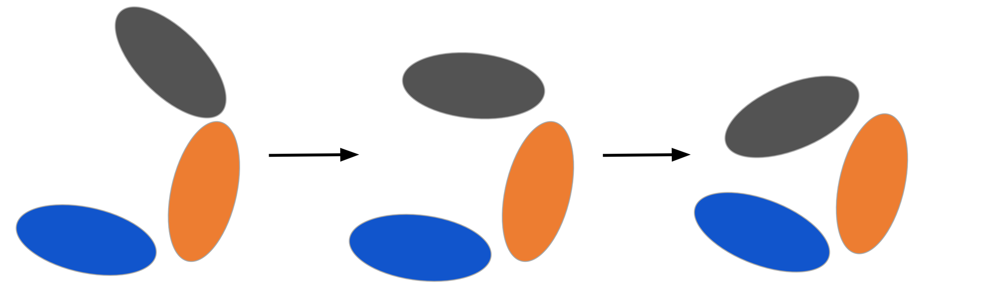
- GenParticles outperforms SpaTracker and CoTracker3 on **20 of 33** sequences, showing strong tracking but limited handling of occlusion and out-of-frame motion due to the absence of a dynamics prior.

Future Applications for Robotic Manipulation

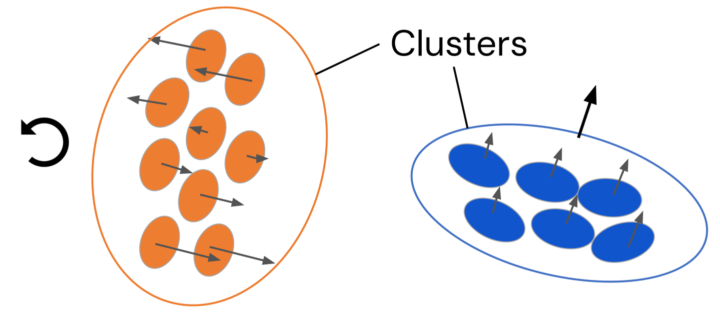
Adaptive Resolution



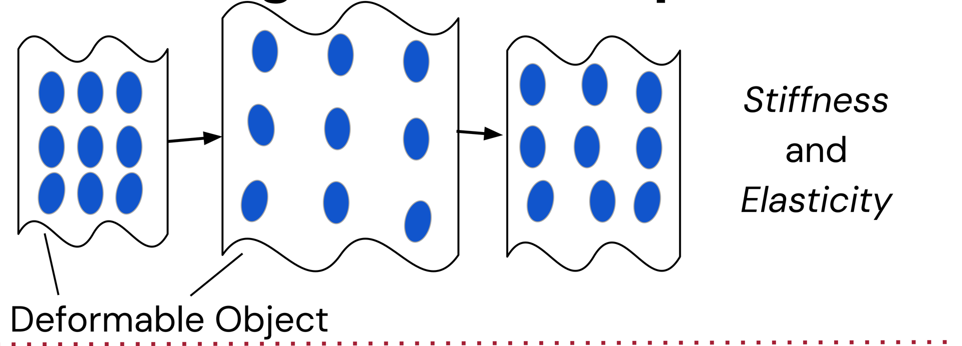
Persistent Object Representation



Interpretable Structure

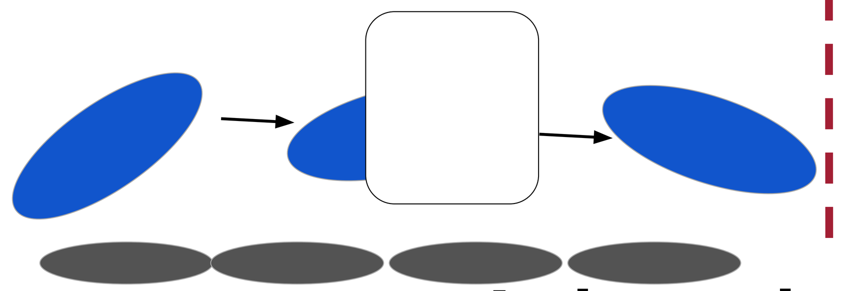


Inferring Internal Properties

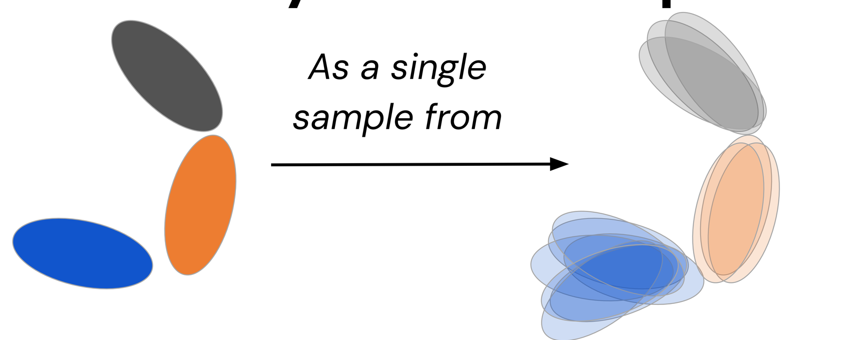


Physical Dynamics Priors

Motion under Occlusion



Uncertainty-aware Outputs



Acknowledgements

This work was supported in part by CoCoSys, one of seven centers in JUMP 2.0, Semiconductor Research Corporation (SRC) program sponsored by DARPA