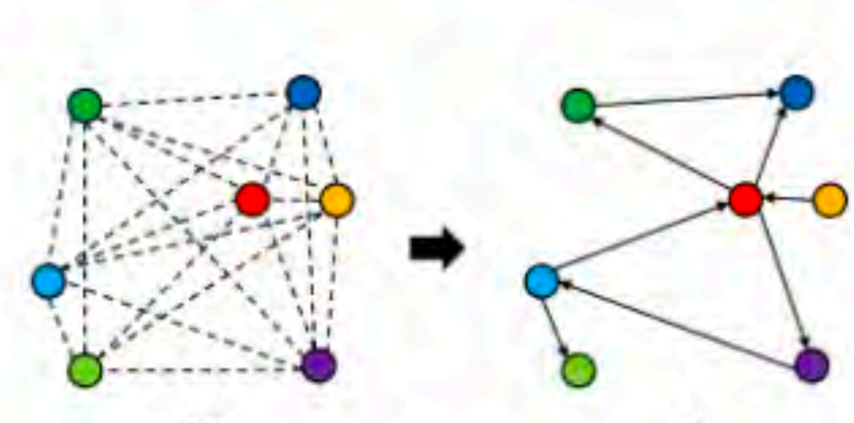
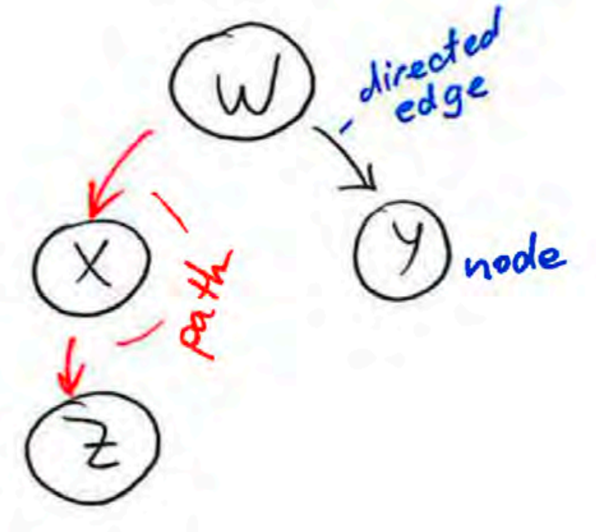
Unsupervised Representations towards Counterfactual Predictions

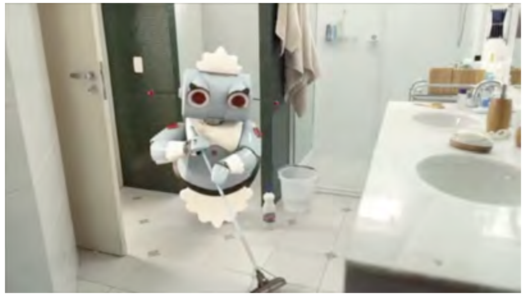




Animesh Garg | UNIVERSITY OF TORONTO VECTOR INSTITUTE

Compositional Representations

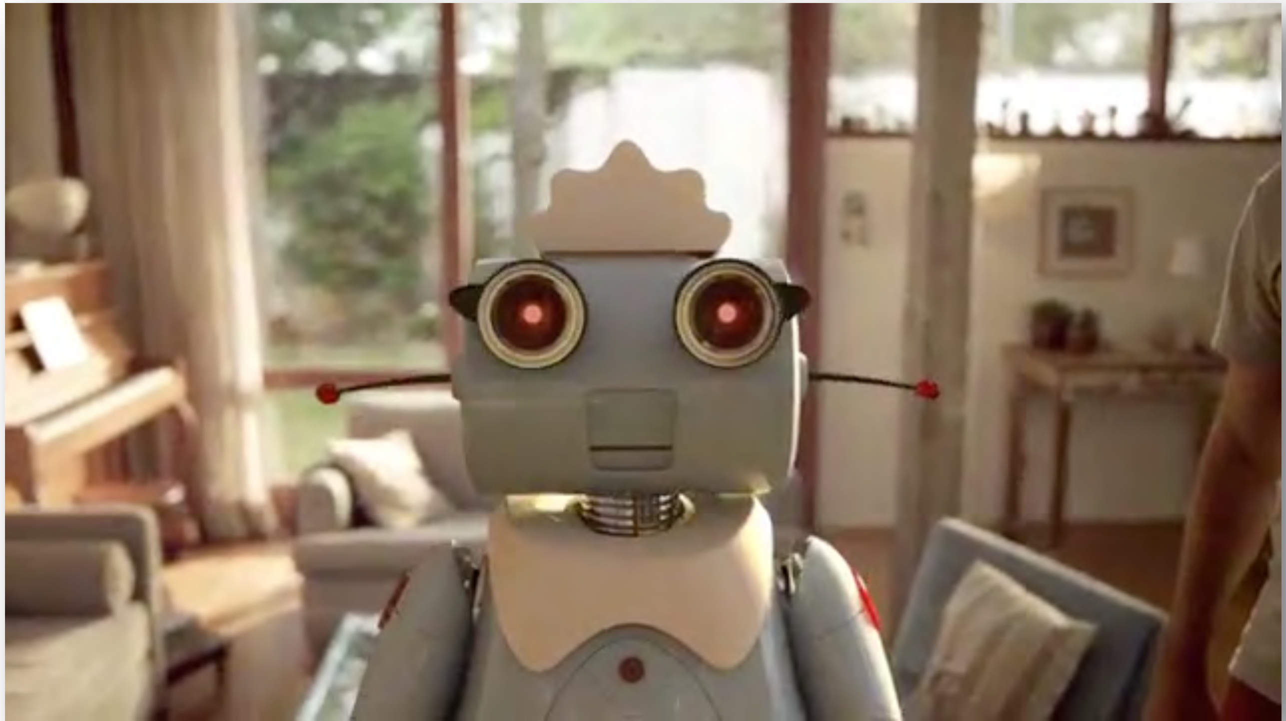
Vacuuming



Sweeping/Mopping

Cooking

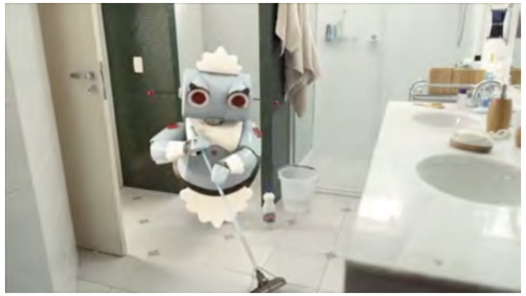
Laundry



Compositional Representations

• • •

Vacuuming



Sweeping/Mopping

Cooking

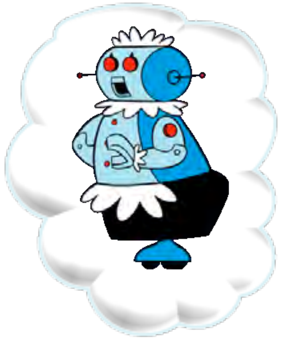
Laundry

• • •

Diversity:
New Scenes,
Tools,…

Complexity:
Long-term
Settings

Compositional Representations



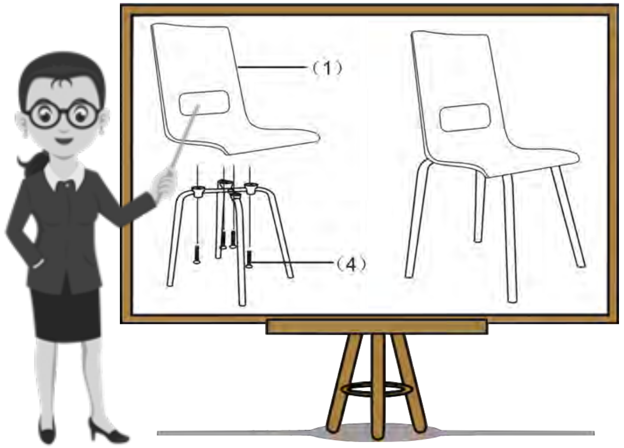
Dartmouth AI Me

UNIM
1st Indus

1956 '61 1968 2013 2018 2020

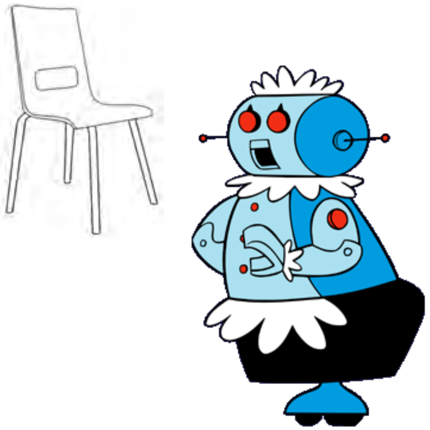
Compositional Representations

Supervision



Input

Task Imitation

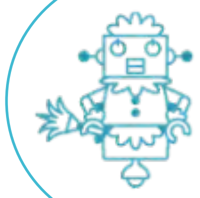


Task Performance in
Data Scarce Set-up

Generalization

New Task Variations
in Novel Environments

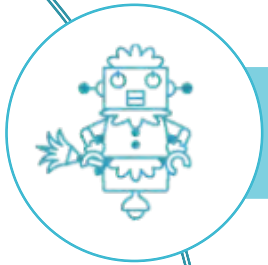
Compositional Representations


Learning Disentanglement


Learning Keypoints


Learning Causal Graphs

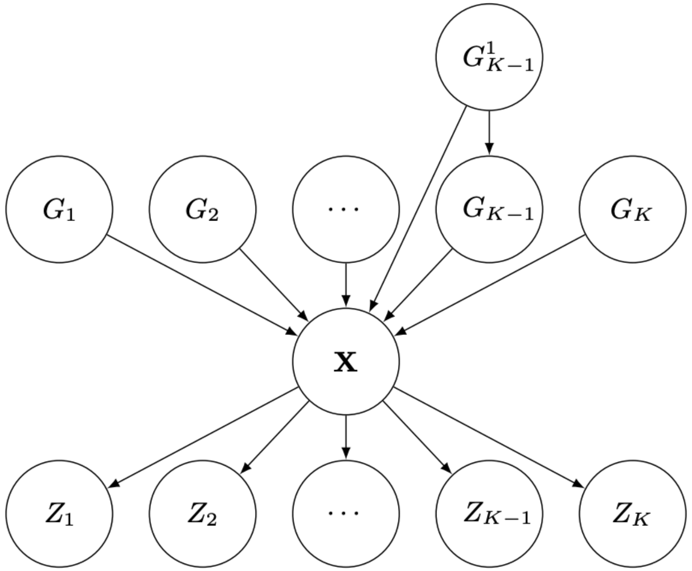
Compositional Representations



Learning Disentanglement

Learning Keypoints

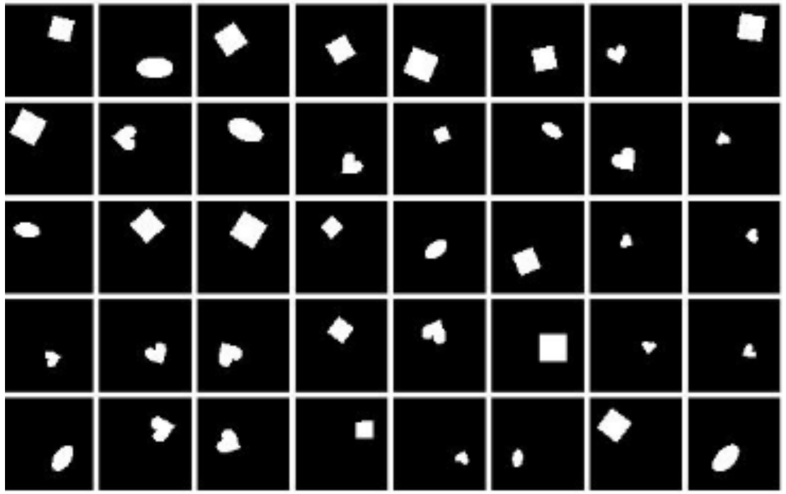
Learning Causal Graphs



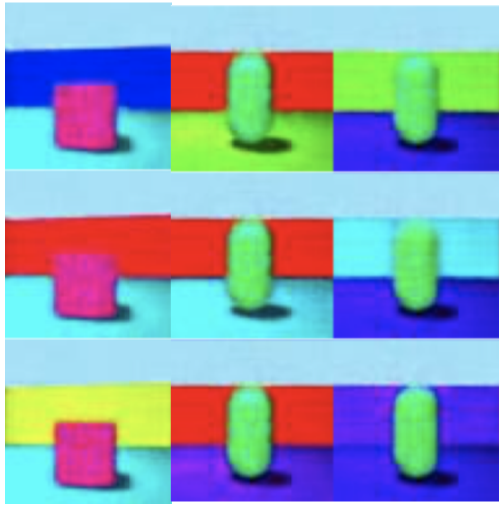
Generative Models: Disentanglement

Objectives of Disentanglement

- Compositional Representations
- Controllable Sample Generation

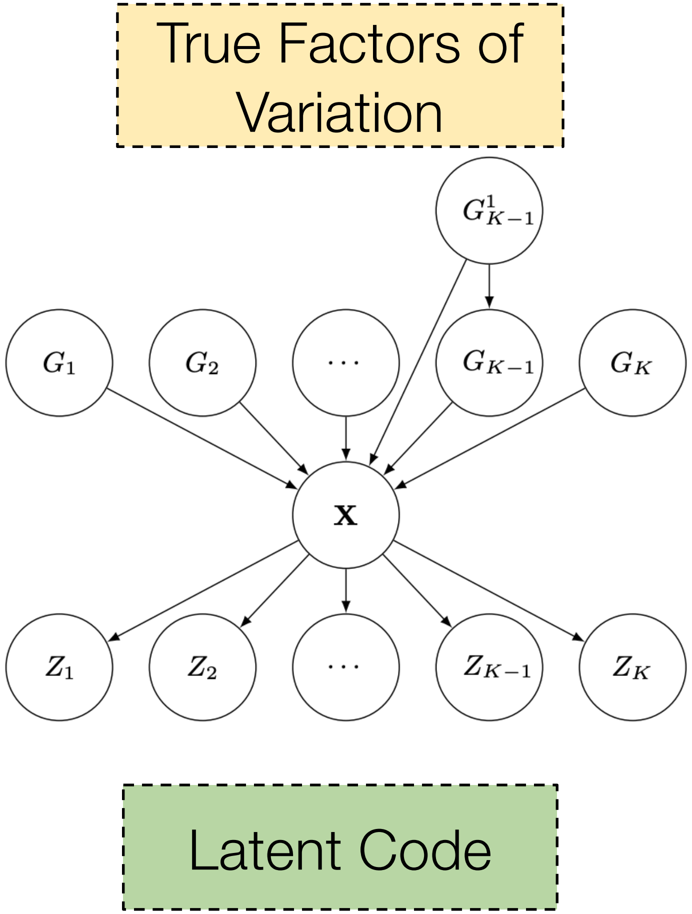


dSprites



3DShapes

Existing datasets in unsupervised disentanglement learning



Disentanglement: Challenges

✗ High-Resolution Output

✗ Non-identifiability in Unsupervised setting

✗ Metrics focus on learning disentangled representations

Locatello et al. ICML 2019.

Disentanglement: Challenges

✗ High-Resolution Output

StyleGAN based backbone (~1%)

New high-resolution synthetic datasets: Falcor3D and Isaac3D

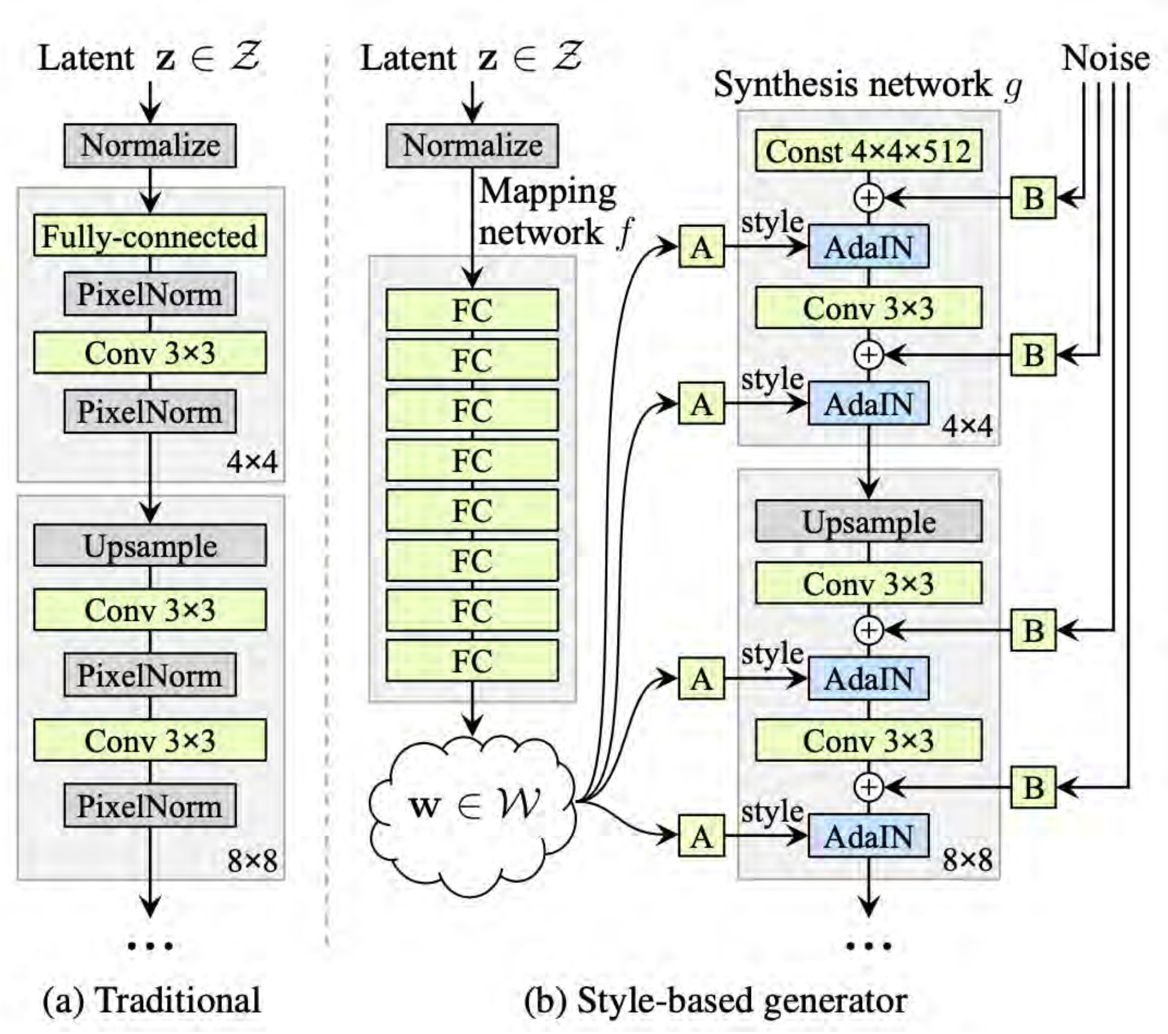
✗ Non-identifiability in Unsupervised setting

Limited Supervision (~1%)

✗ Metrics focus on learning disentangled representations

New Metric to Trade-off between controllability and disentanglement

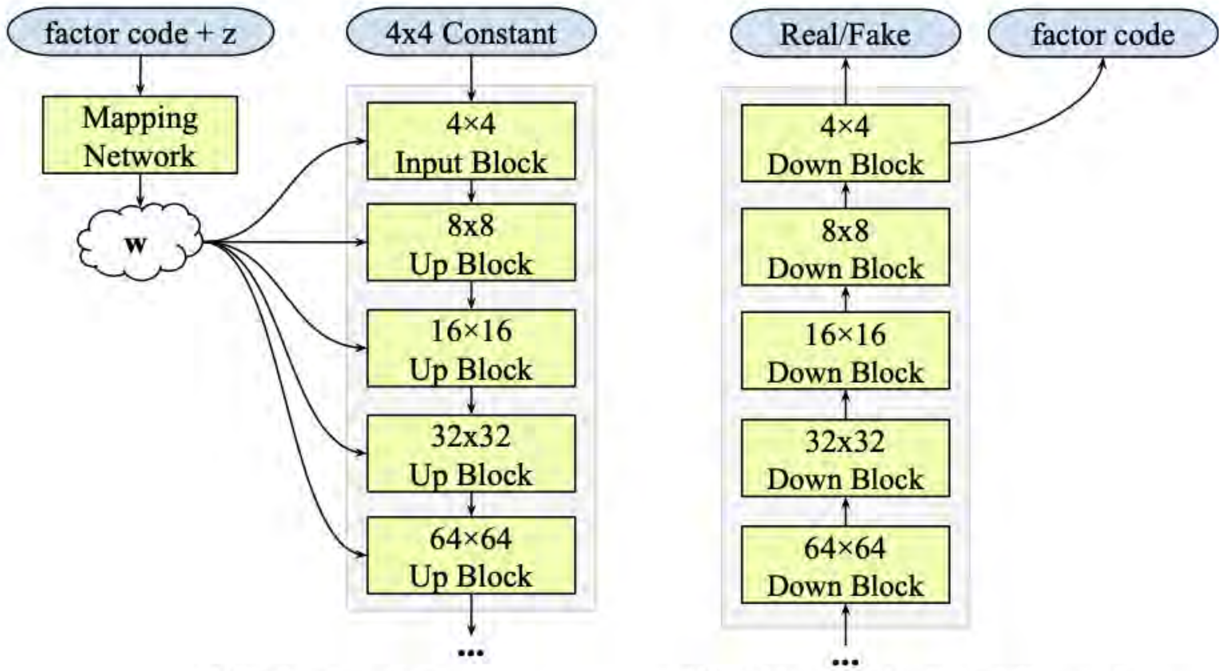
Disentanglement: StyleGAN



- Used a style-based generator to replace traditional generator
- Success at generating high-resolution realistic images

Kerras et al. CVPR 2019.

Disentanglement: Semi-StyleGAN



(i) Generator (ii) Discriminator / Encoder

Disentanglement in StyleGAN

Mapping Network in the generator conditions on the factor code and the encoder predicts its value

The semi-supervised loss is given by

$$\mathcal{L}^{(G)} = \mathcal{L}_{\text{GAN}} + \gamma_G \mathcal{L}_{\text{unsup}} + \alpha \mathcal{L}_{\text{sr}}$$

$$\mathcal{L}^{(D,E)} = -\mathcal{L}_{\text{GAN}} + \gamma_E \mathcal{L}_{\text{unsup}} + \beta \mathcal{L}_{\text{sup}} + \alpha \mathcal{L}_{\text{sr}}$$

with

$$\mathcal{L}_{\text{unsup}} = \sum_{c \sim \mathcal{C}, z \sim p_z} \|E(G(c,z)) - c\|_2$$ → unsupervised InfoGAN loss term

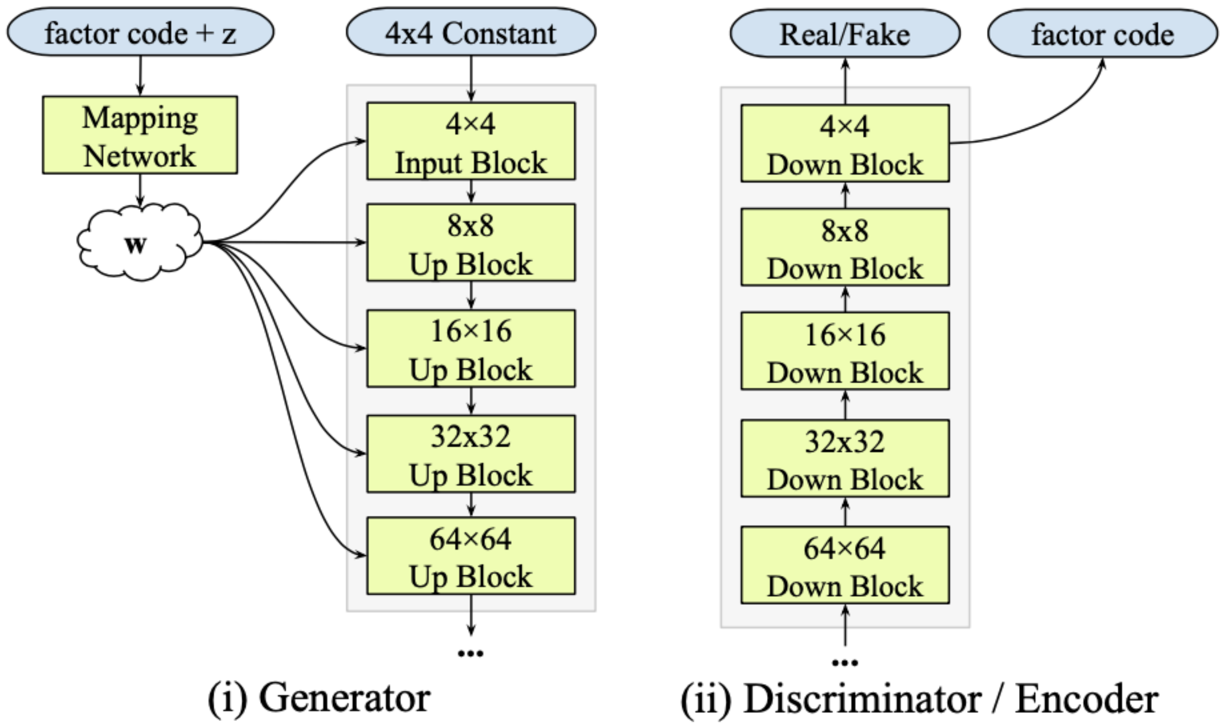
$$\mathcal{L}_{\text{sup}} = \sum_{(x,c) \sim \mathcal{J}} \|E(x) - c\|_2$$ → supervised label reconstruction term

$$\mathcal{L}_{\text{sr}} = \sum_{(x,c) \sim \mathcal{M}} \|E(x) - c\|_2$$ → smoothness regularization term

J set of all possible factor codes
X: set of labeled pairs of real image and factor code
M: mixed set of labeled and generated image-code pairs
E, G: encoder and generator neural networks

(ours) Li et al. ICML 2020.

Disentanglement: Semi-StyleGAN



Disentanglement in StyleGAN
Mapping Network in the generator conditions on the factor code and the encoder predicts its value

The semi-supervised loss is given by

$$\mathcal{L}^{(G)} = \mathcal{L}_{\text{GAN}} + \gamma_G \mathcal{L}_{\text{unsup}} + \alpha \mathcal{L}_{\text{sr}}$$

$$\mathcal{L}^{(D,E)} = -\mathcal{L}_{\text{GAN}} + \gamma_E \mathcal{L}_{\text{unsup}} + \beta \mathcal{L}_{\text{sup}} + \alpha \mathcal{L}_{\text{sr}}$$

with

$$\mathcal{L}_{\text{unsup}} = \sum_{c \sim \mathcal{C}, z \sim p_z} \|E(G(c,z)) - c\|_2$$ → unsupervised InfoGAN loss term

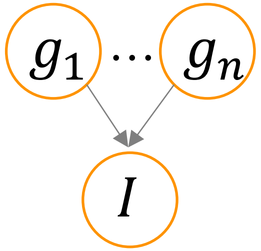
$$\mathcal{L}_{\text{sup}} = \sum_{(x,c) \sim \mathcal{J}} \|E(x) - c\|_2$$ → supervised label reconstruction term

$$\mathcal{L}_{\text{sr}} = \sum_{(x,c) \sim \mathcal{M}} \|E(x) - c\|_2$$ → smoothness regularization term

Labeled Data $|\mathcal{J}| \ll |\mathcal{X}|$ Unpaired Data
$\mathcal{M}$: Artificially Augmented Data

Li et al. ICML 2020.

Semi-StyleGAN: CelebA (256x256)



I_i Data Point
<person_id>, long-hair

I'_i **Counterfactual** Data Point
<person_id>, bangs

Fixed-value

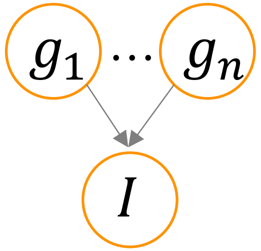
Bangs **Glasses** **Smiling**

Semi-StyleGAN on CelebA (0.5% of labeled data)

With very limited supervision, Semi-StyleGAN can achieve good disentanglement on real data

Li et al. ICML 2020.

Semi-StyleGAN: Isaac3D (512x512)



I_i Data Point
<object_id>, x-pos

I'_i **Counterfactual**
Data Point
<object_id>, x-pos_new

Fixed-value

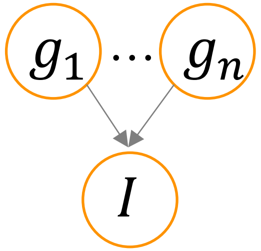
Robot X-Movement **Lighting Y-Dir** **Object Scale**

Semi-StyleGAN on Isaac3D (0.5% of labeled data)

Each factor in the interpolated images changes smoothly without affecting other factors

Li et al. ICML 2020.

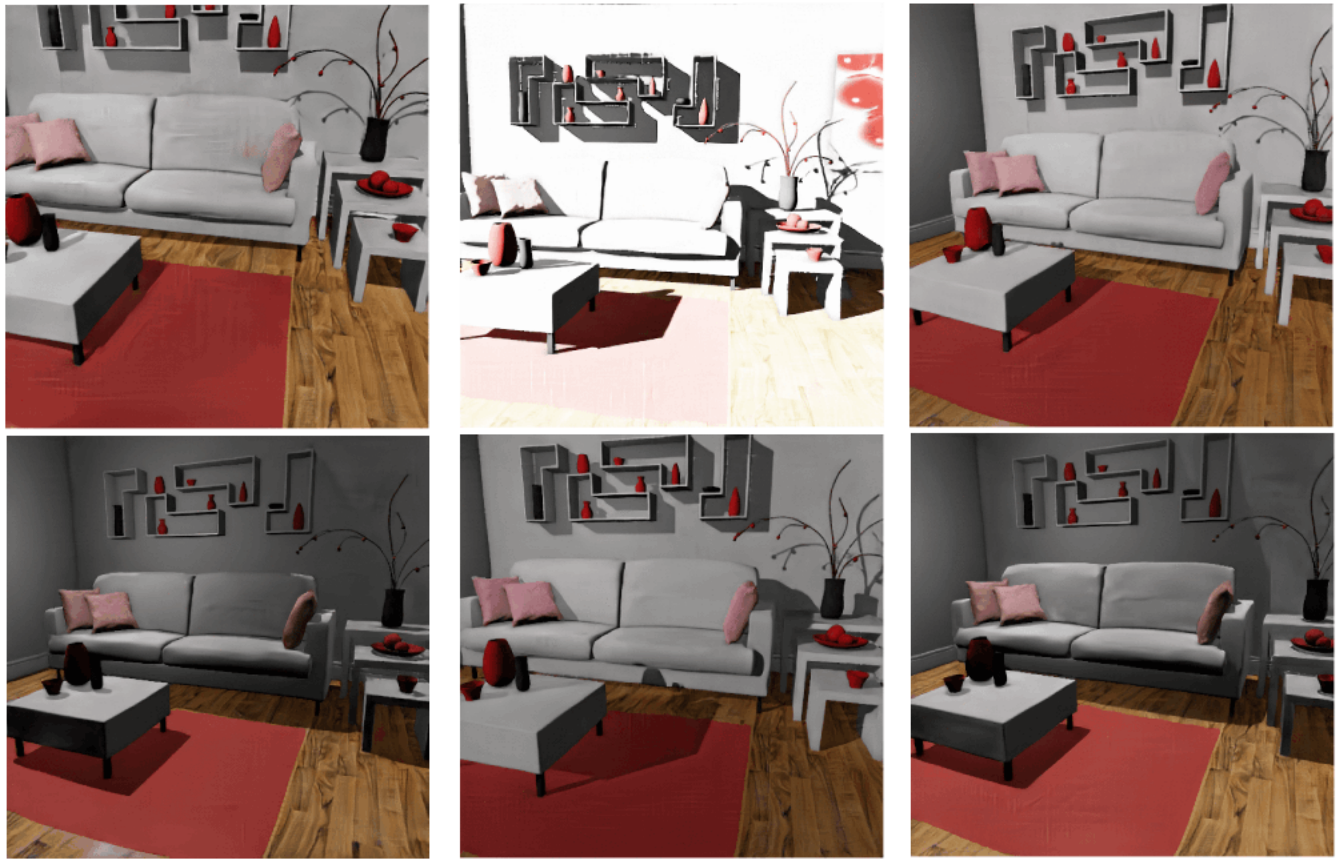
Semi-StyleGAN: Falcor3D (512x512)



I_i Data Point
<lighting>, camera_pos

I_i' **Counterfactual**
Data Point
< lighting >, camera_new

Fixed-value



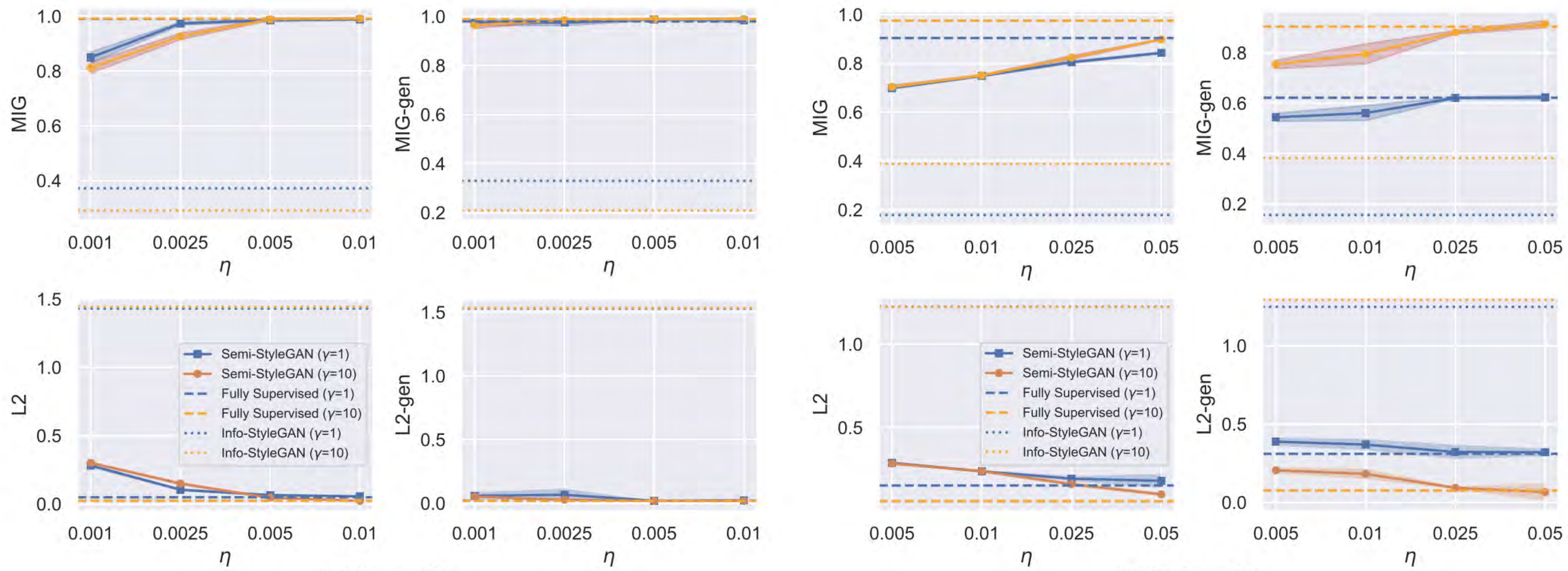
Lighting Intensity **Lighting X-Dir** **Camera X-Pos**

Semi-StyleGAN on Falcor3D (1% of labeled data)

Each factor in the interpolated images changes smoothly without affecting other factors

Li et al. ICML 2020.

Semi-StyleGAN: Role of Limited Supervision



(a) Isaac3D (b) Falcor3D

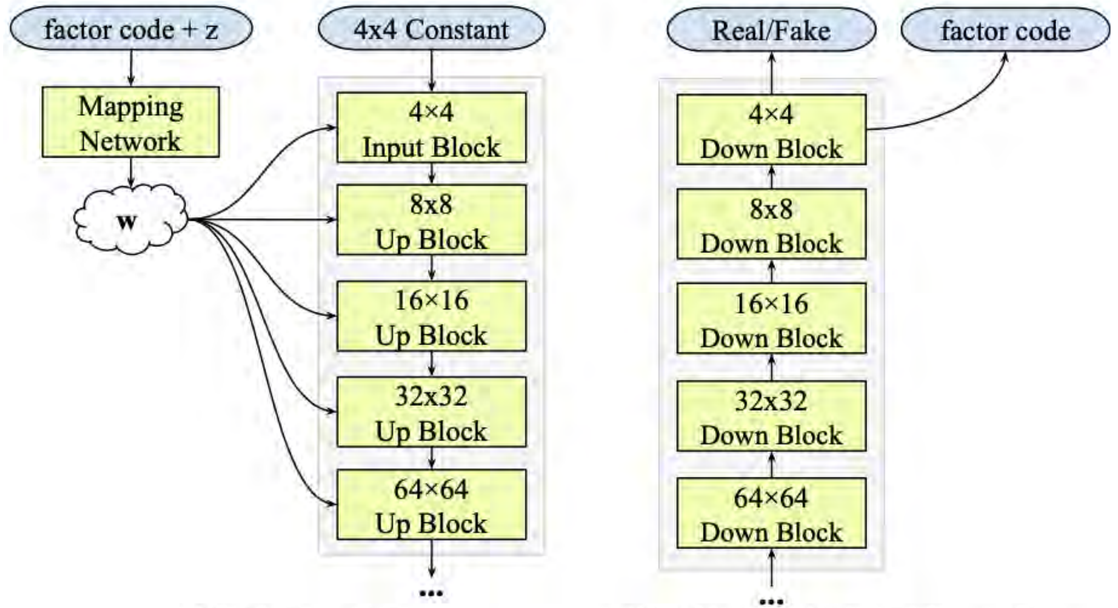
(L2 and L2-gen: lower is better, MIG and MIG-gen: higher is better)

Semi-StyleGAN with the default setting $\gamma_G = \beta = \gamma, \gamma_E = 0, \alpha = 1$

Only using 0.25~2.5% of labeled data at par with supervised disentanglement

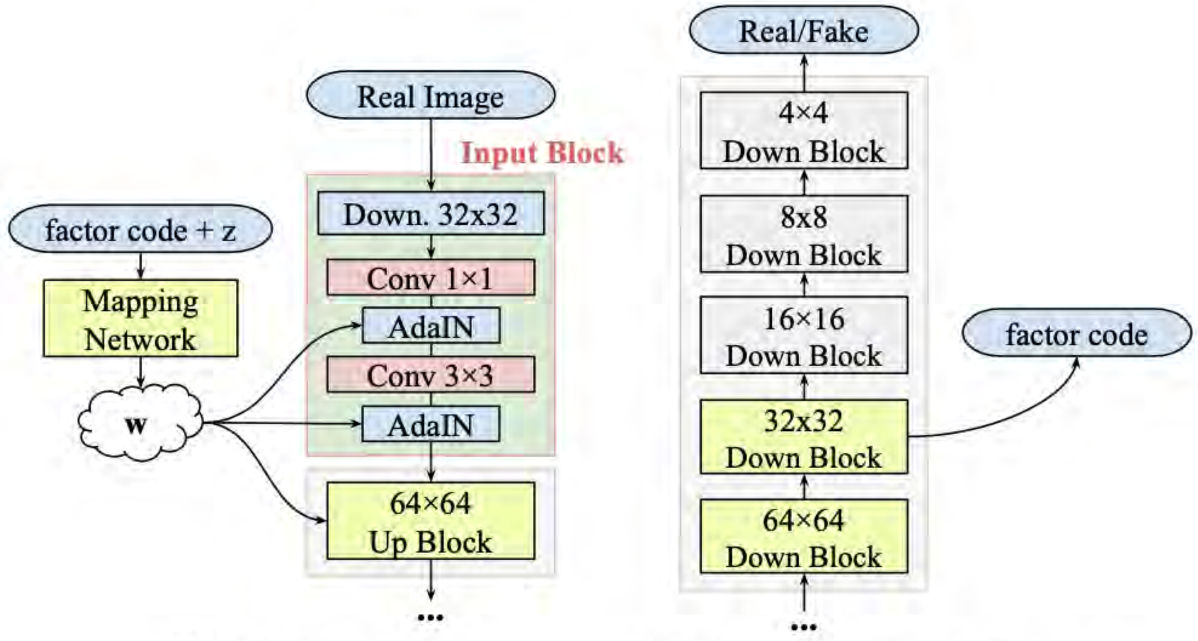
Semi-StyleGAN: Fine-Grained Tuning

New architecture with same loss model for semantic fine-grained image editing



(i) Generator (ii) Discriminator / Encoder

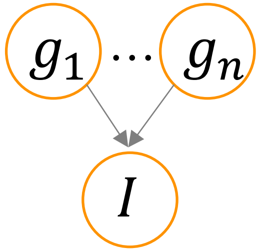
Coarse-Grained



(i) Generator (ii) Discriminator / Encoder

Fine-Grained

Semi-StyleGAN: CelebA (256x256)



I_i Data Point
eyebrow, smile

I'_i **Counterfactual** Data Point
eyebrow, grin

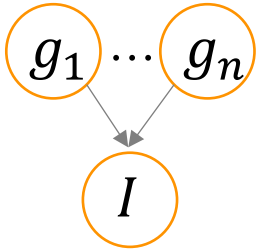
Fixed-value

Bushy Eyebrows **Slightly Open Mouth** **Pale Skin**

Semi-StyleGAN-*fine* on CelebA (1% of labeled data)

We <u>randomly</u> choose some deep learning researchers as test images

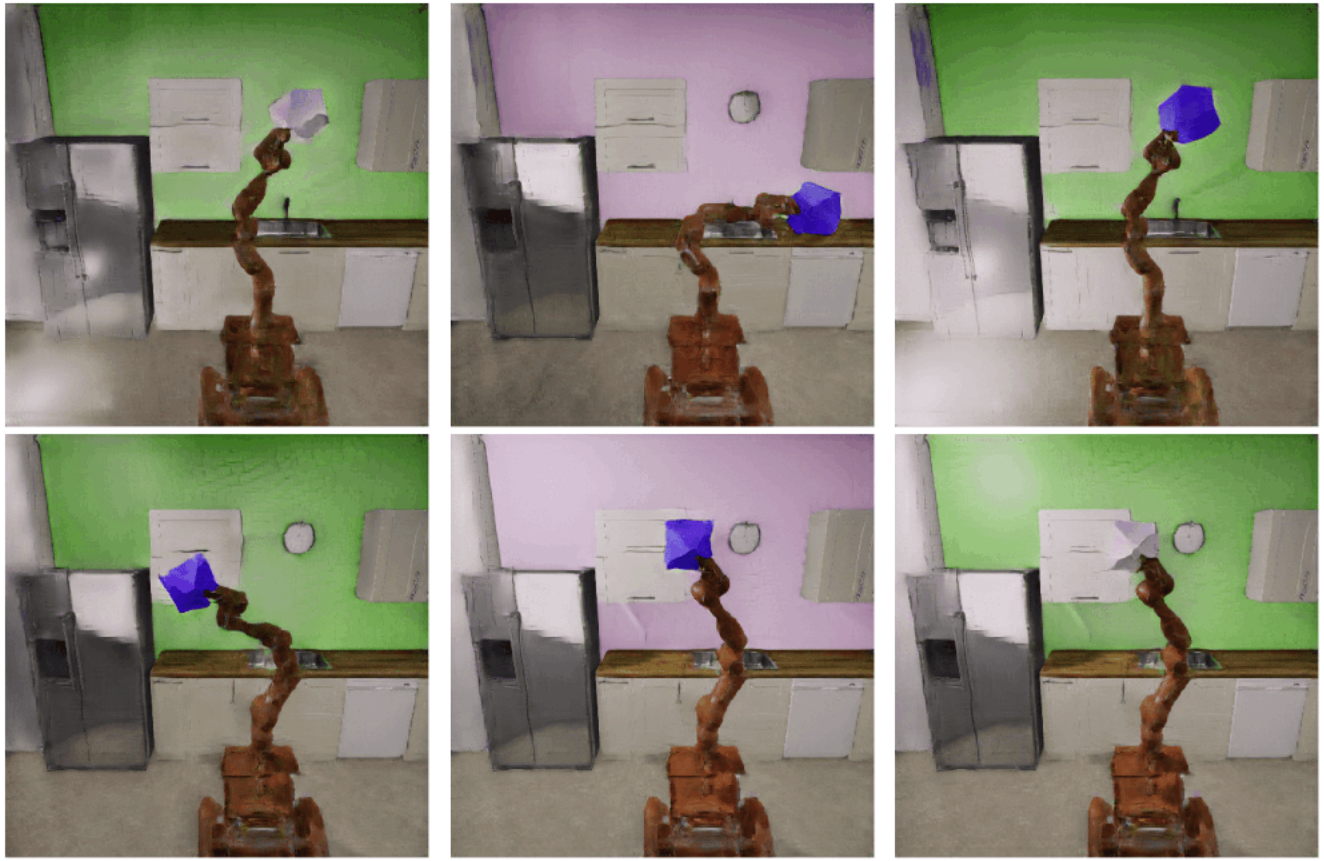
Semi-StyleGAN: CelebA (256x256)



I_i Data Point
wall_color, obj_color

I'_i **Counterfactual**
Data Point
wall_color, Obj_color

Fixed-value



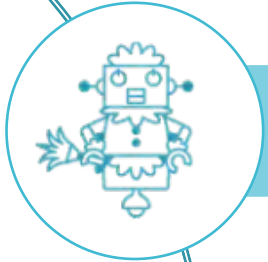
Lighting Intensity **Object Color** **Wall Color**

Semi-StyleGAN-*fine* on Isaac3D (1% of labeled data)

We shift the robot position to the right side, and attach it with an unseen object in test images

Li et al. ICML 2020.

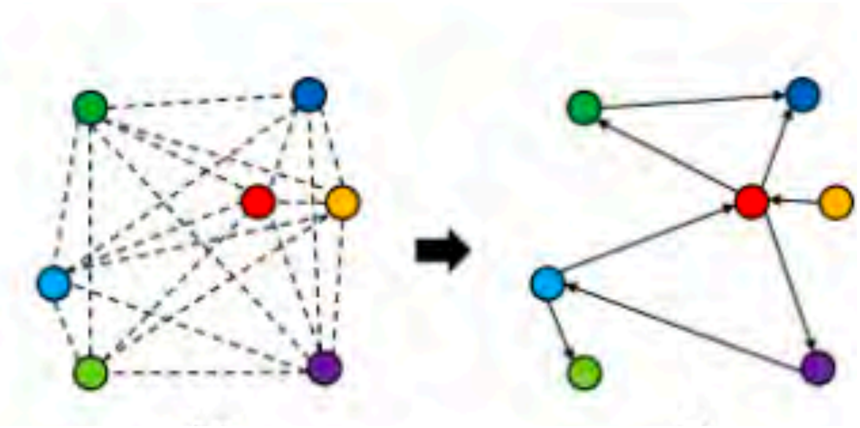
Compositional Representations

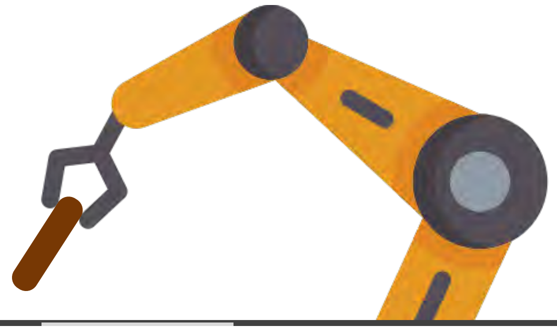


Learning Disentanglement

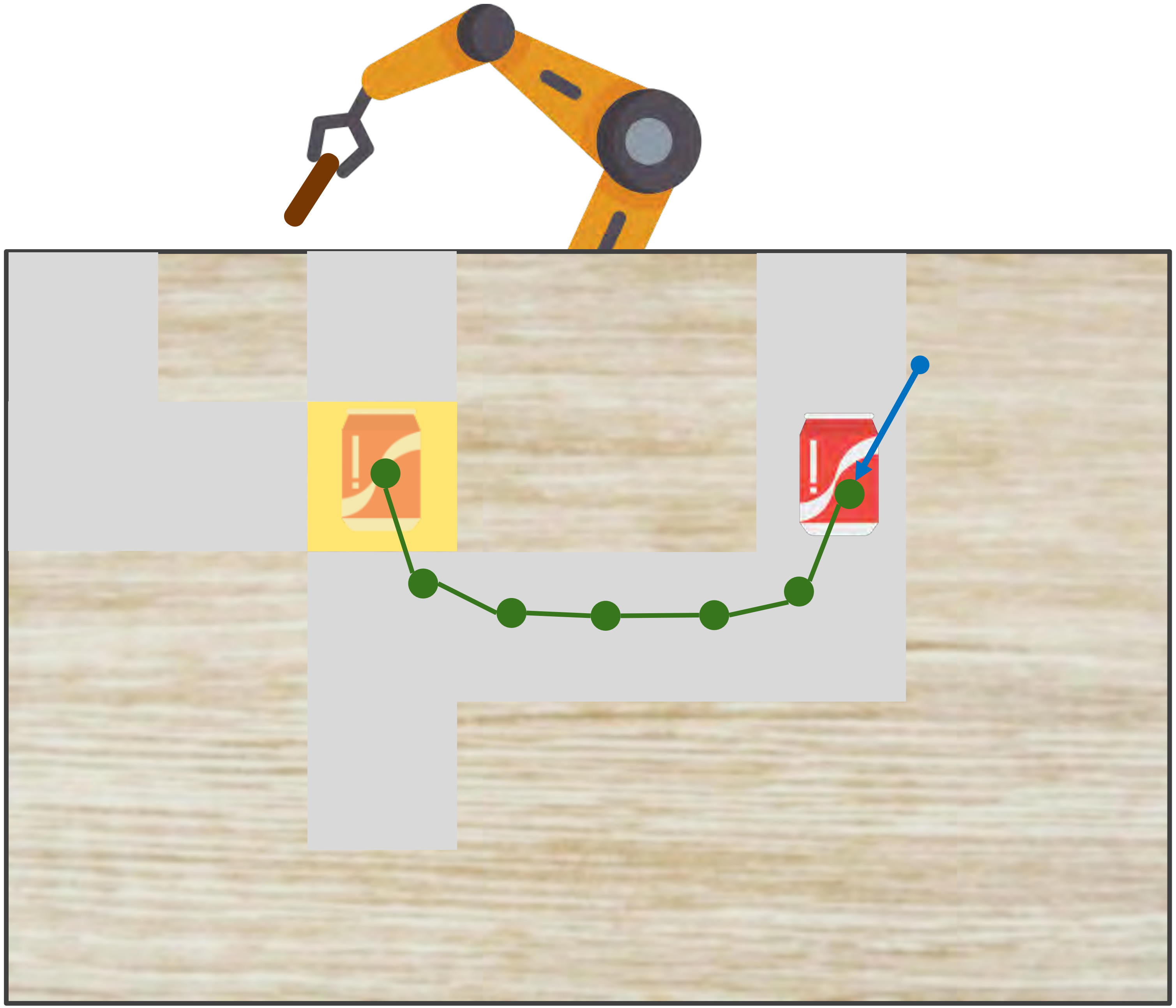
Learning Keypoints

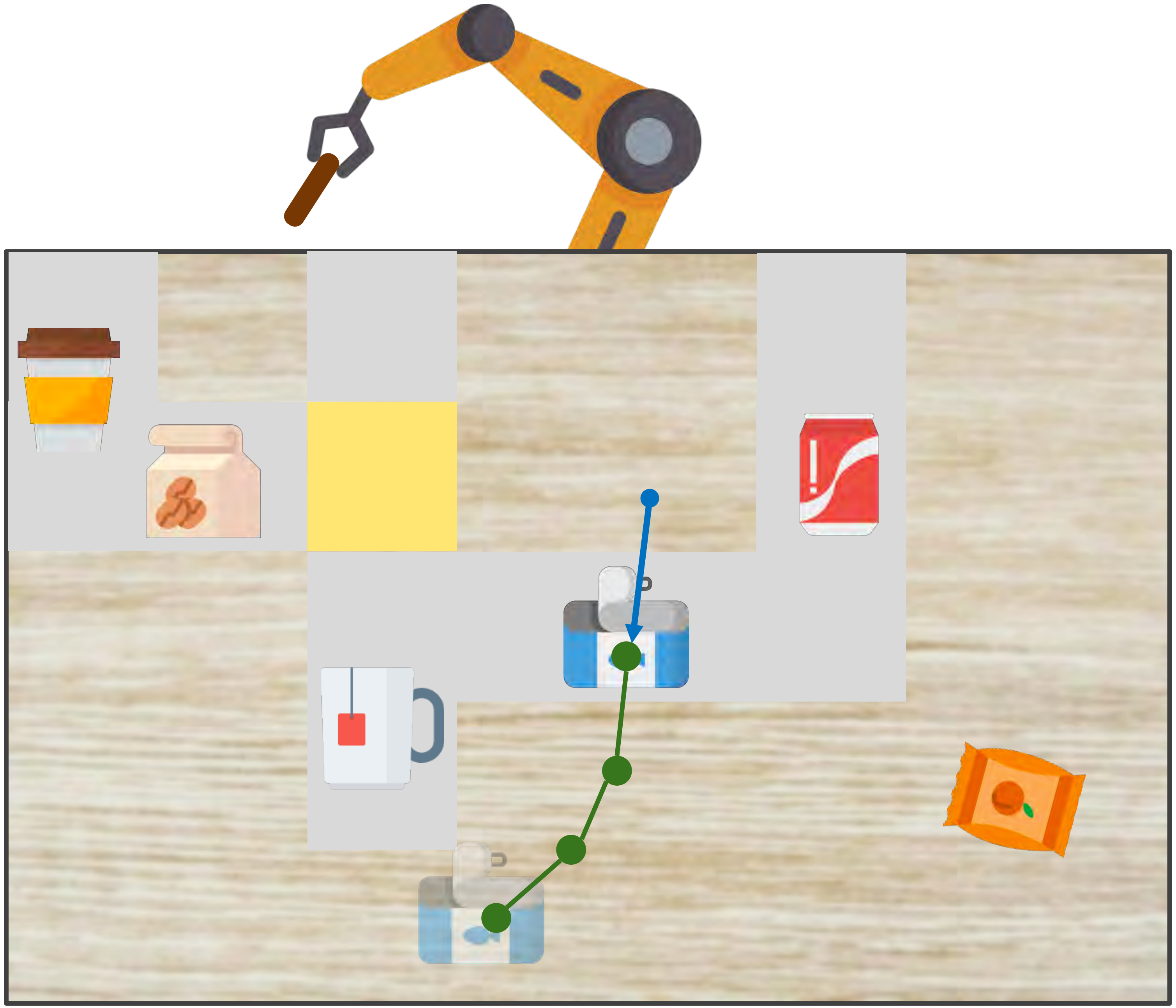
Learning Causal Graphs

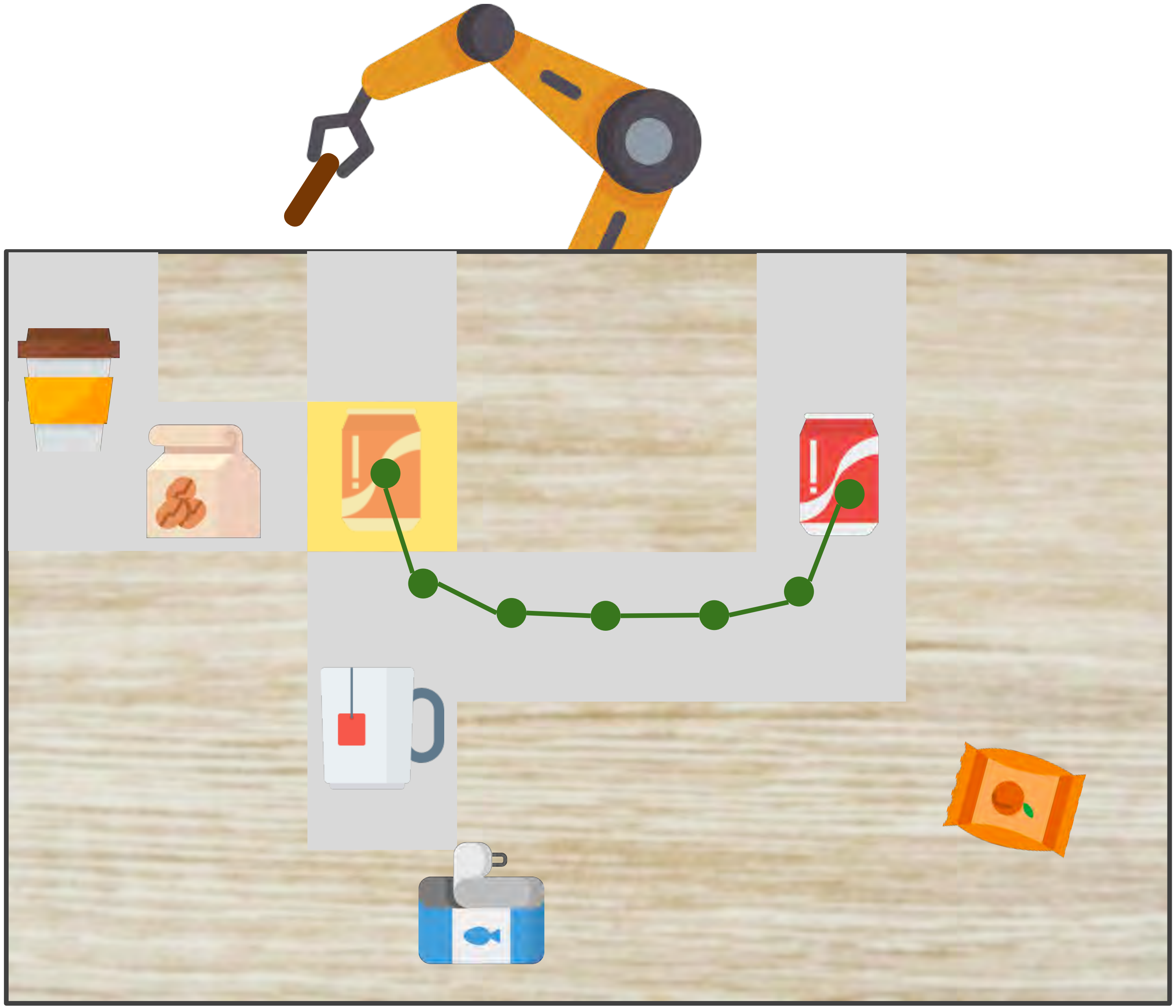




Representations for multi-step reasoning in Robotics under physical and semantic constraints



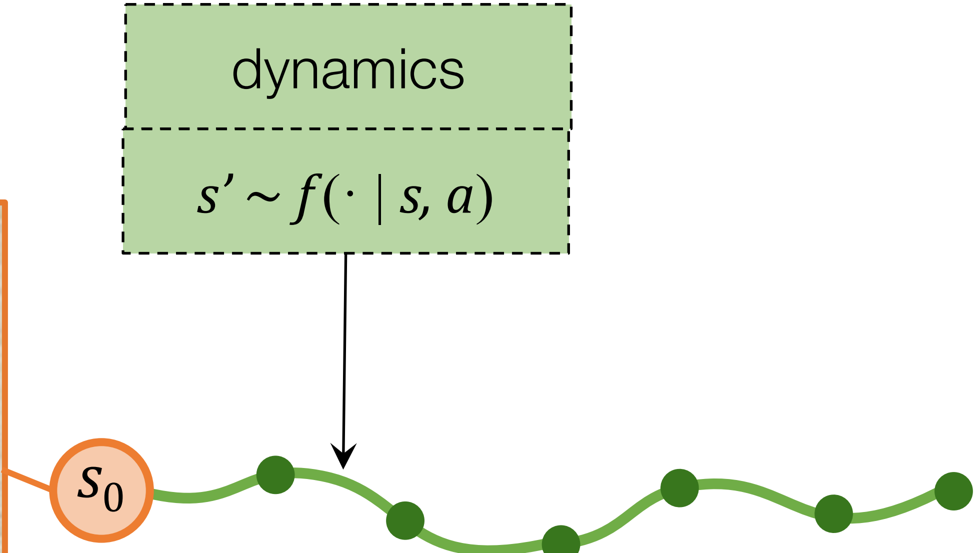




Model-based learning

choose action sequence

$a_t, \ldots\ldots, a_{t+H}$



[Deisenroth et al, RSS'07], [Guo et al, NeurIPS'14], [Watter et al, NeurIPS'15], [Finn et al, ICRA'17], ……

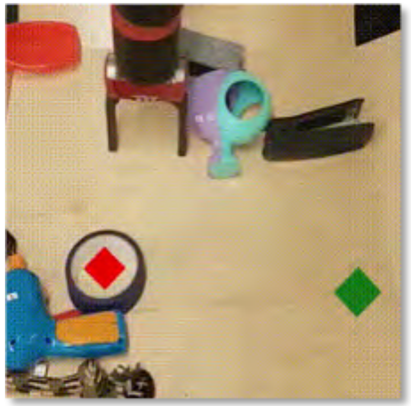
Model-based learning

?

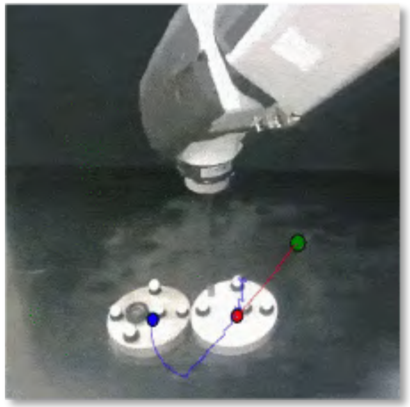
data ↑
learning ↑

[Deisenroth et al. RSS'07]

[Agrawal et al. ICRA'16]

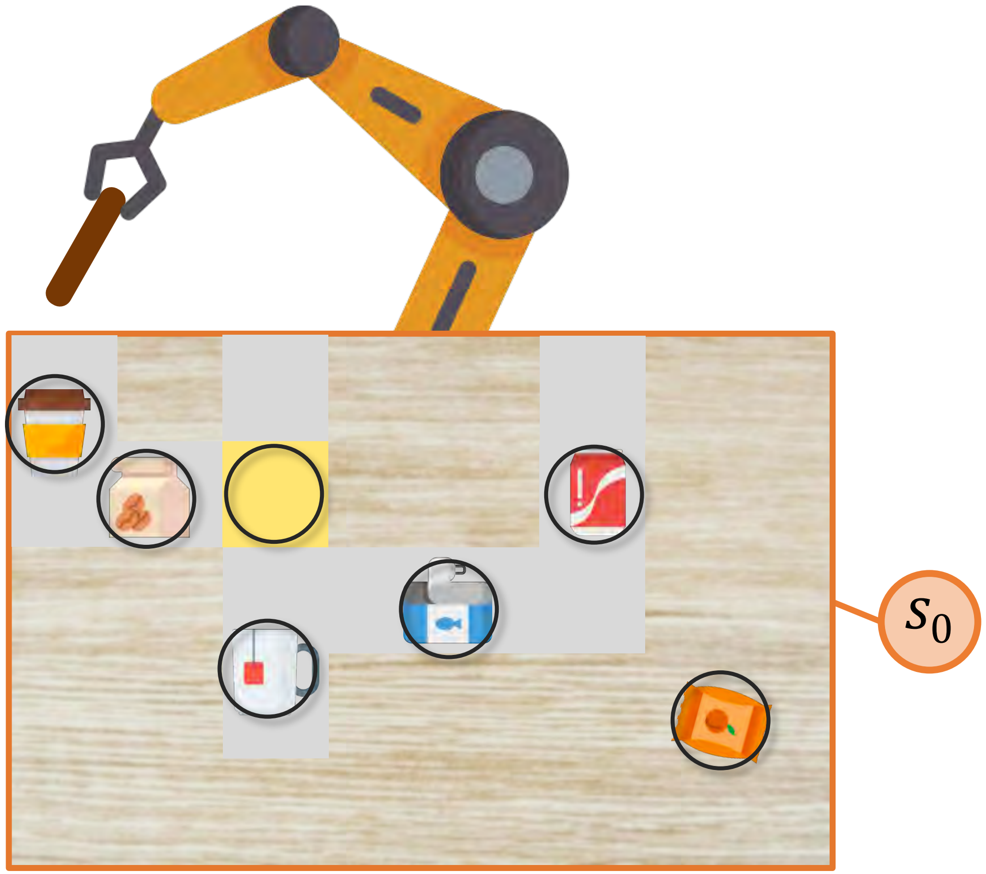


[Ebert et al. CoRL'17]



[Janer et al. ICRA'19]

CAVIN: Hierarchical planning in learned latent spaces



CAVIN Planner

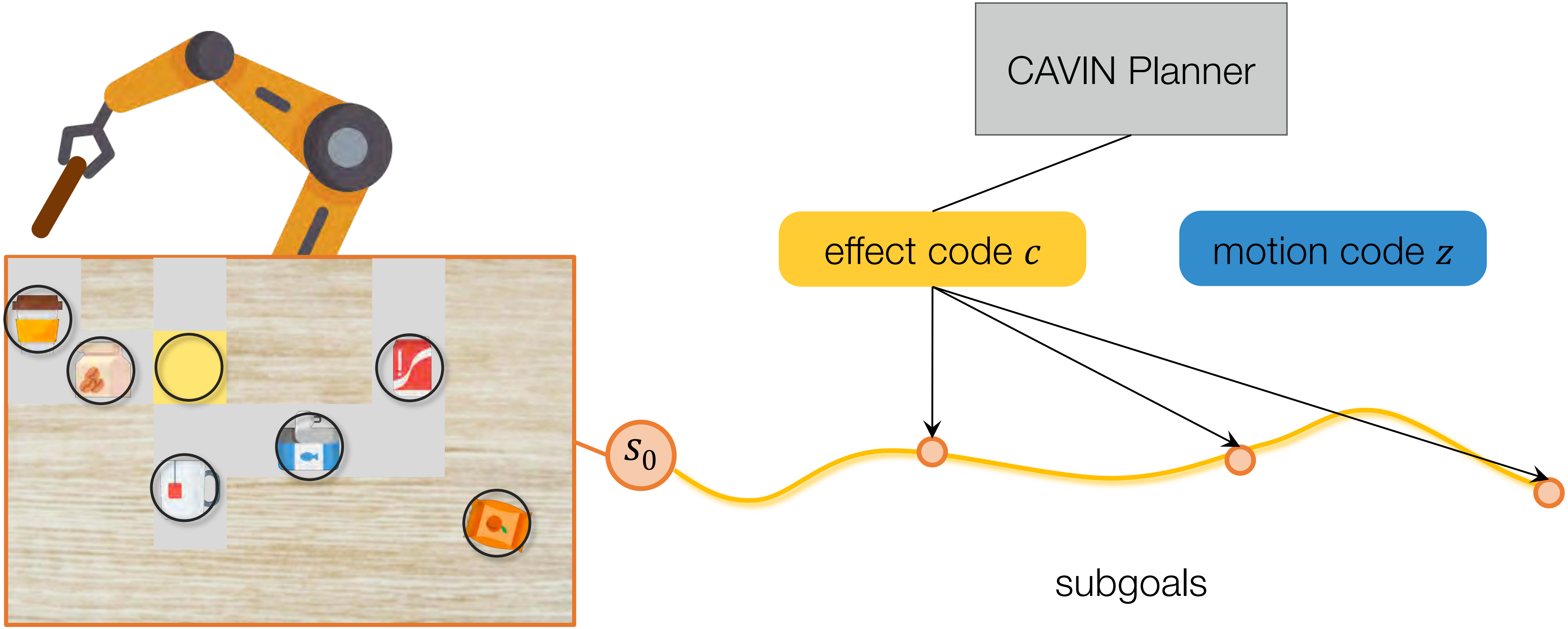
effect code c

motion code z

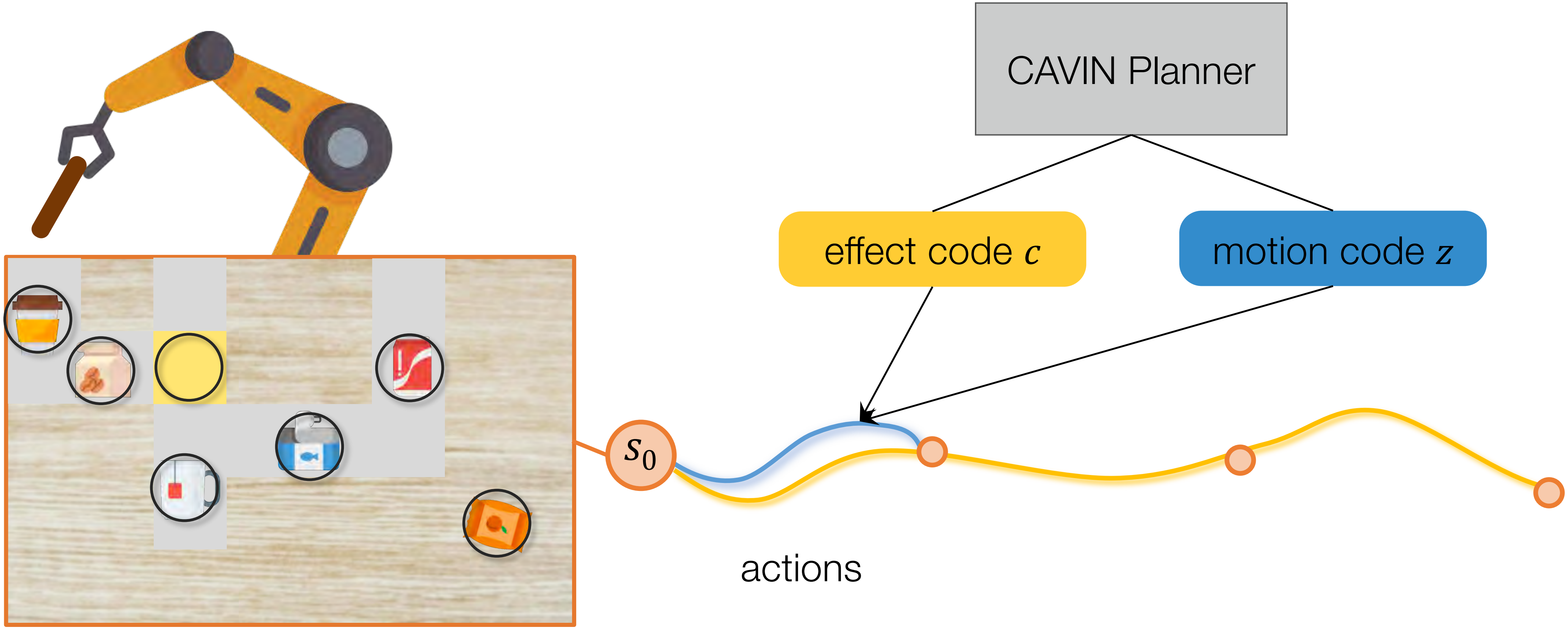
Leverage Hierarchical Abstraction in Action Space

Without Hierarchical Supervision

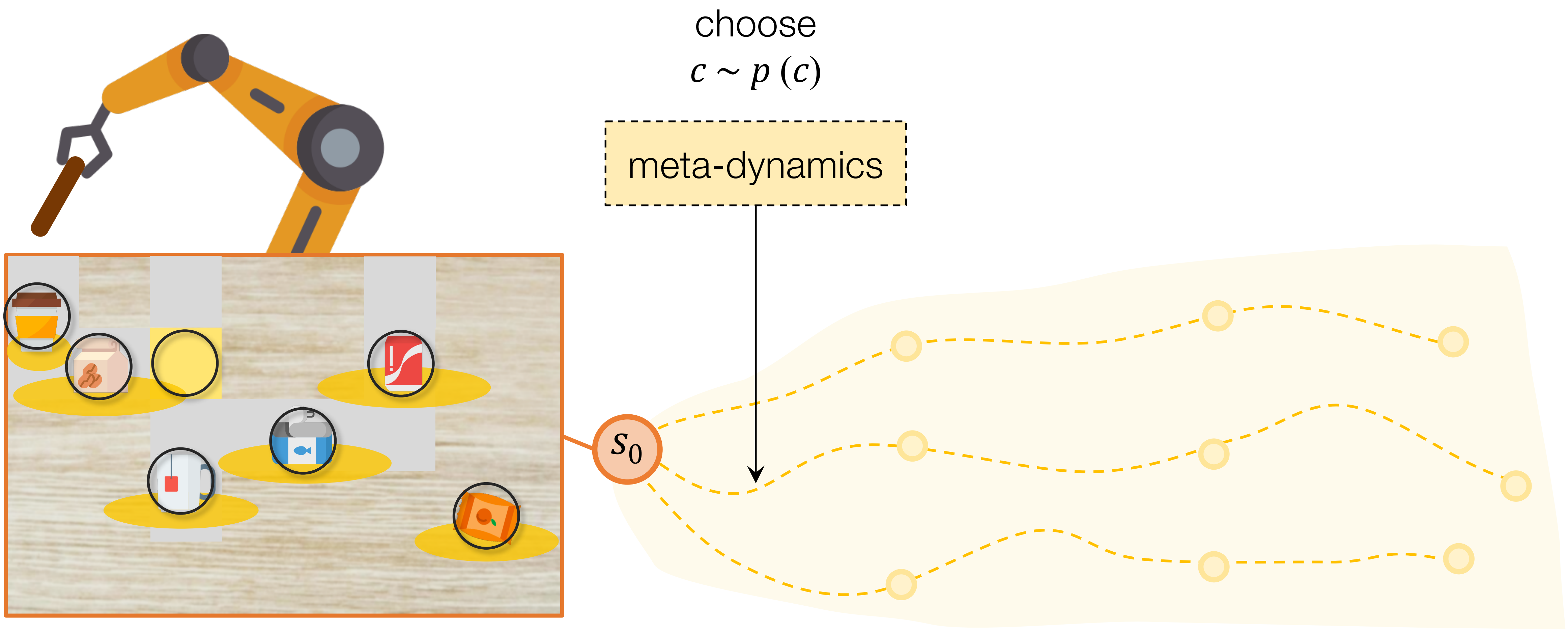
CAVIN: Hierarchical planning in learned latent spaces



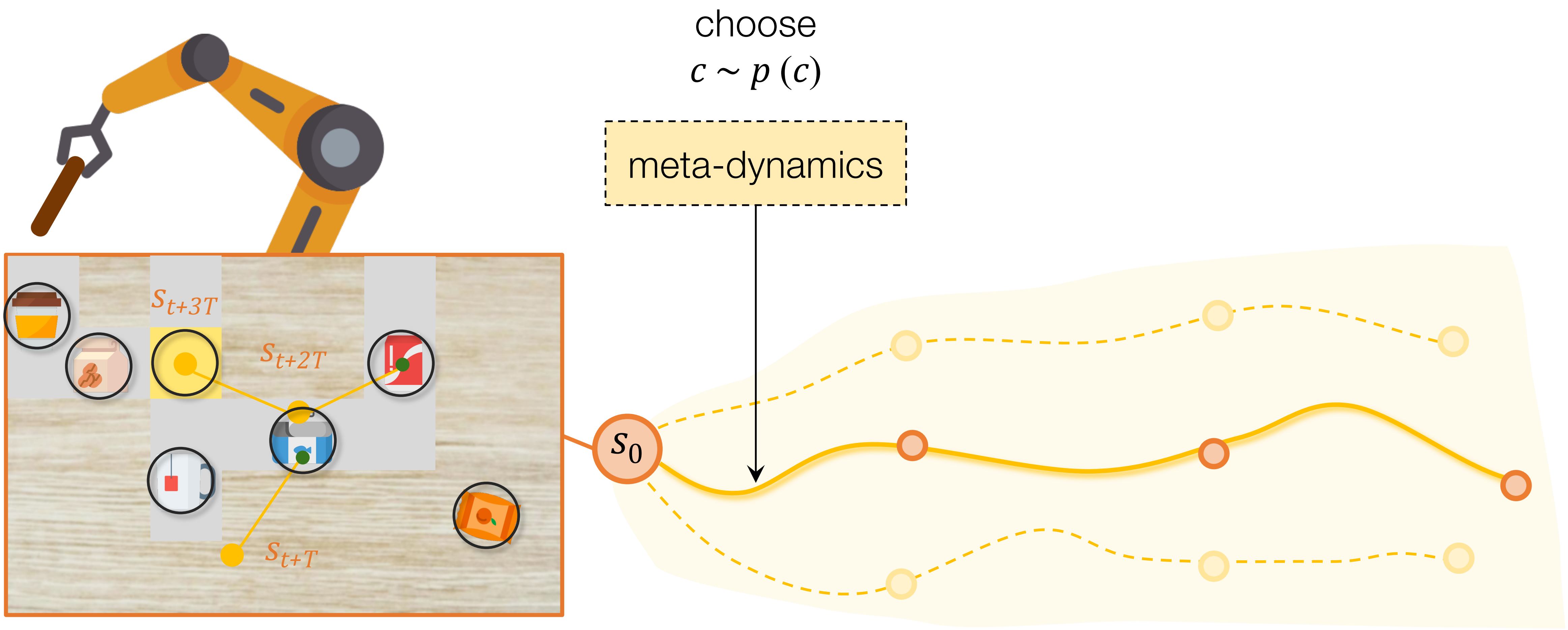
CAVIN: Hierarchical planning in learned latent spaces



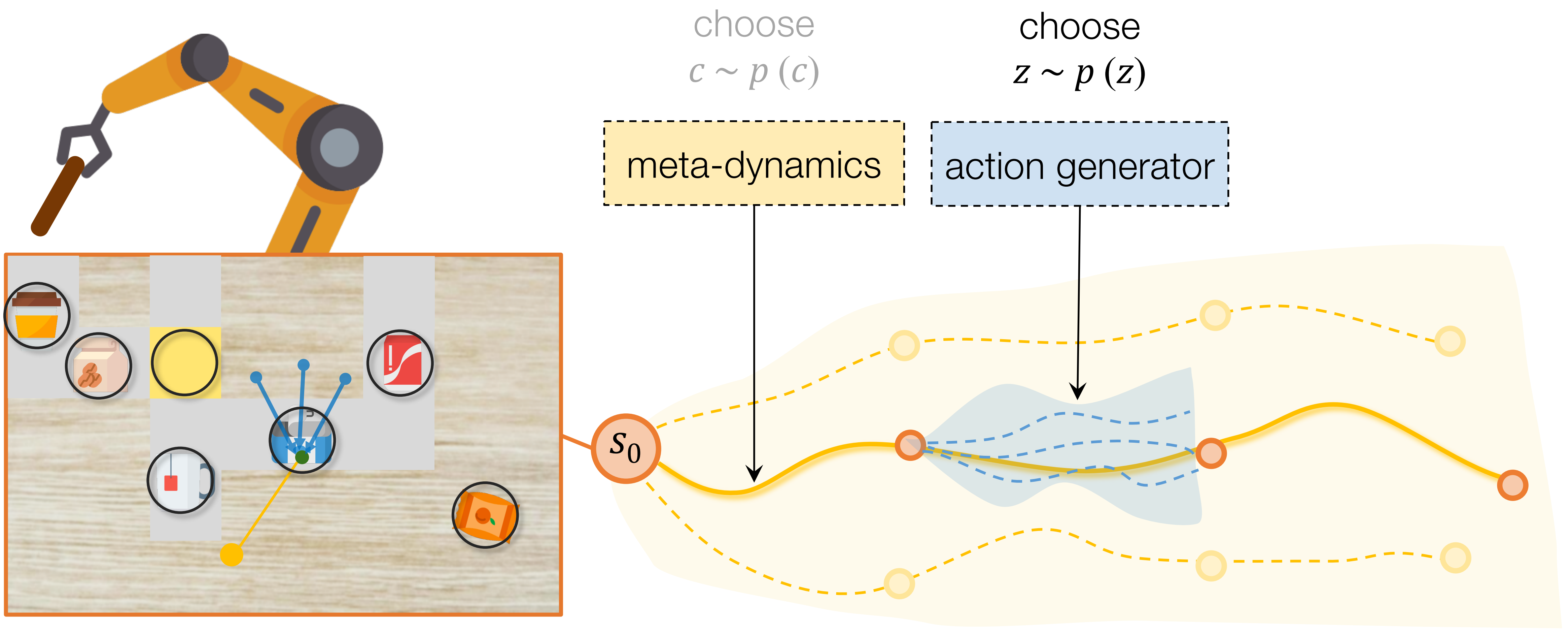
CAVIN: Hierarchical planning in learned latent spaces



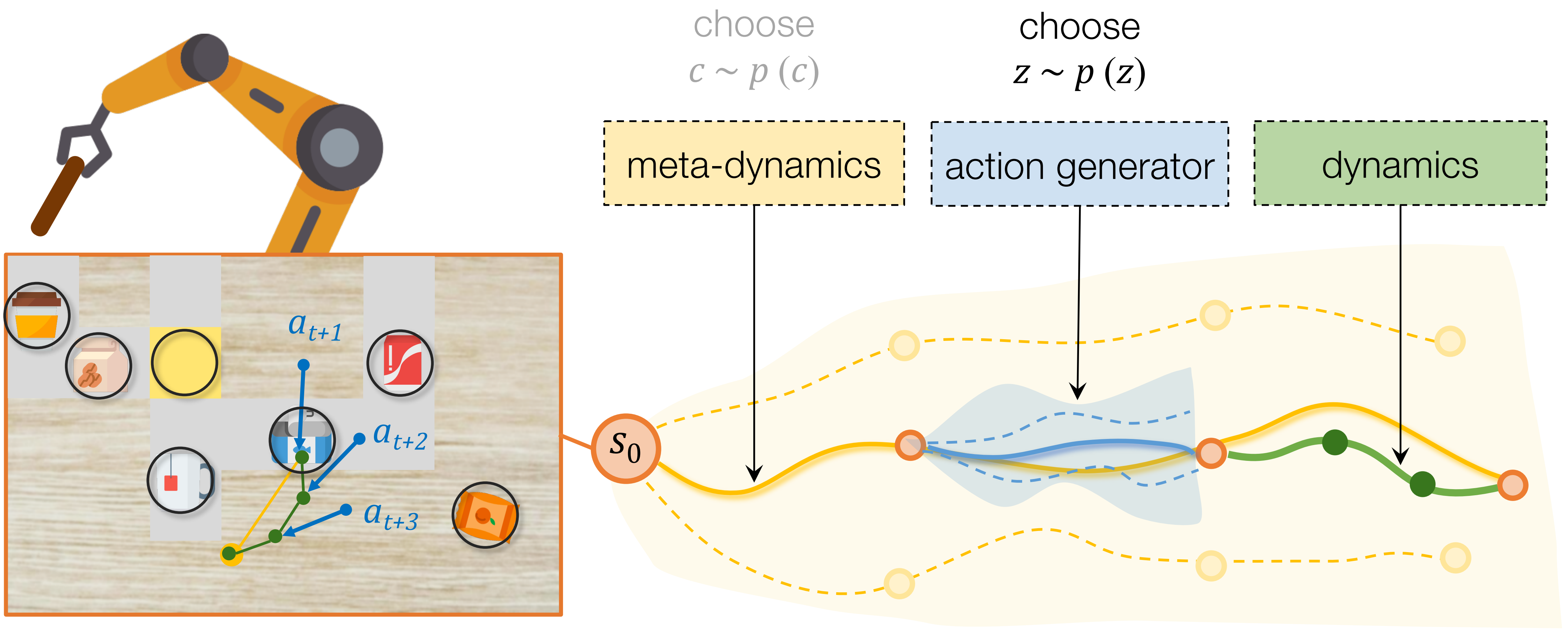
CAVIN: Hierarchical planning in learned latent spaces



Hierarchical planning in learned latent spaces

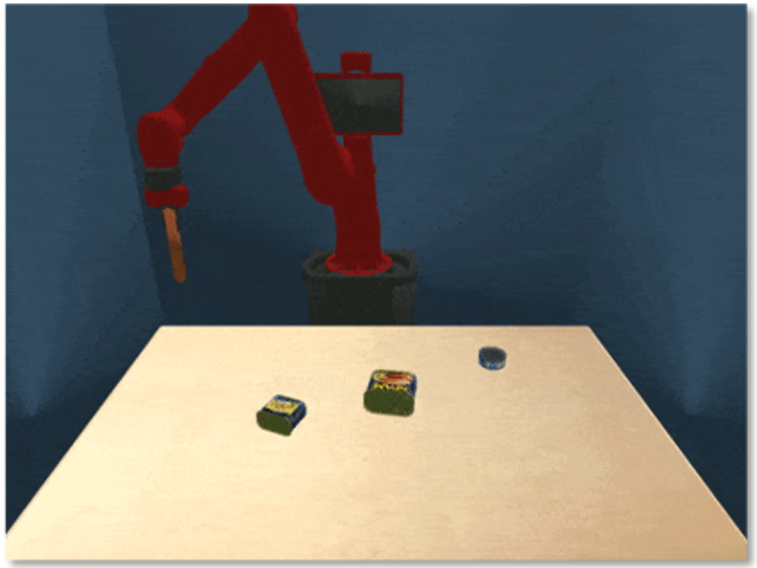


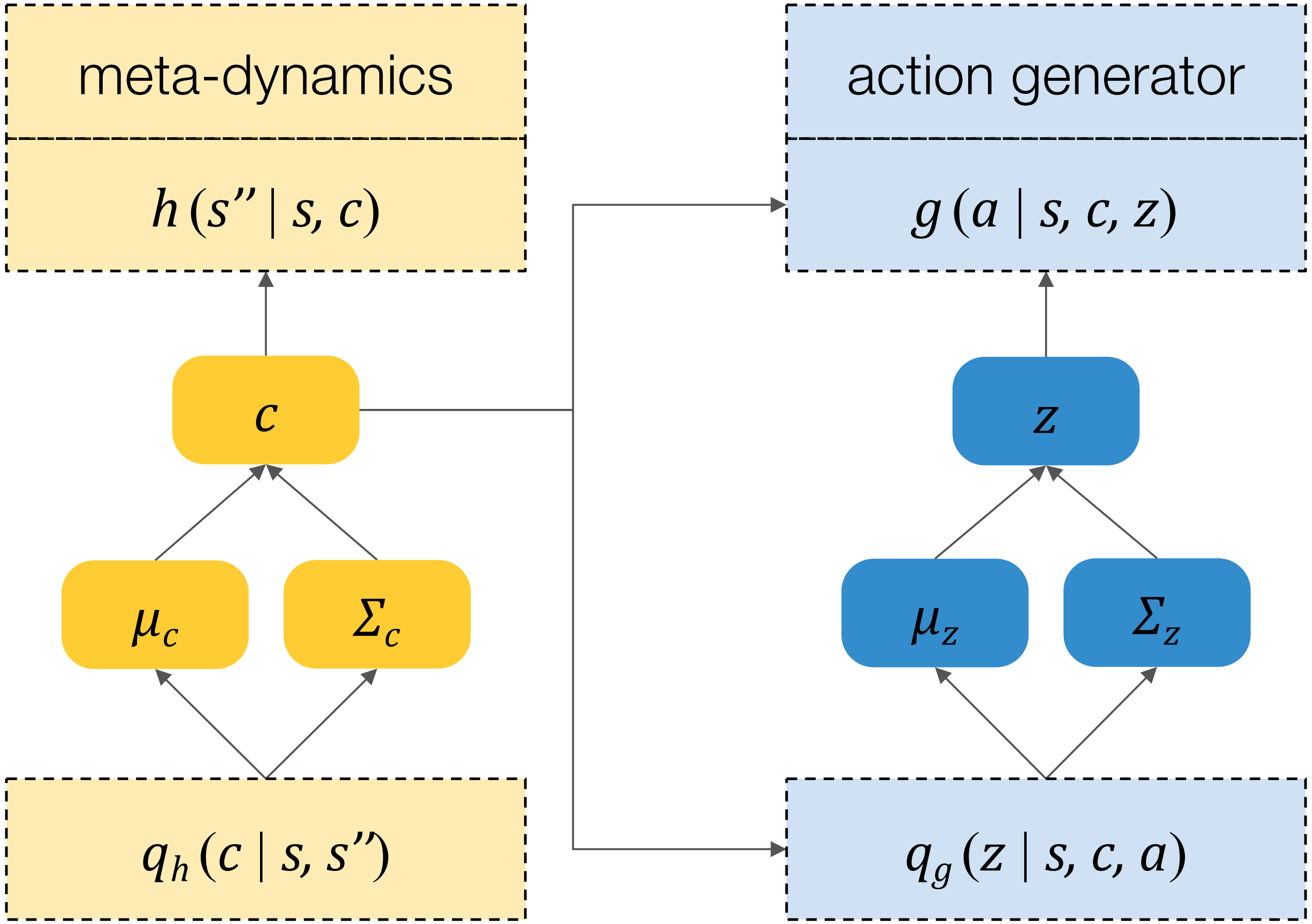
CAVIN: Hierarchical planning in learned latent spaces

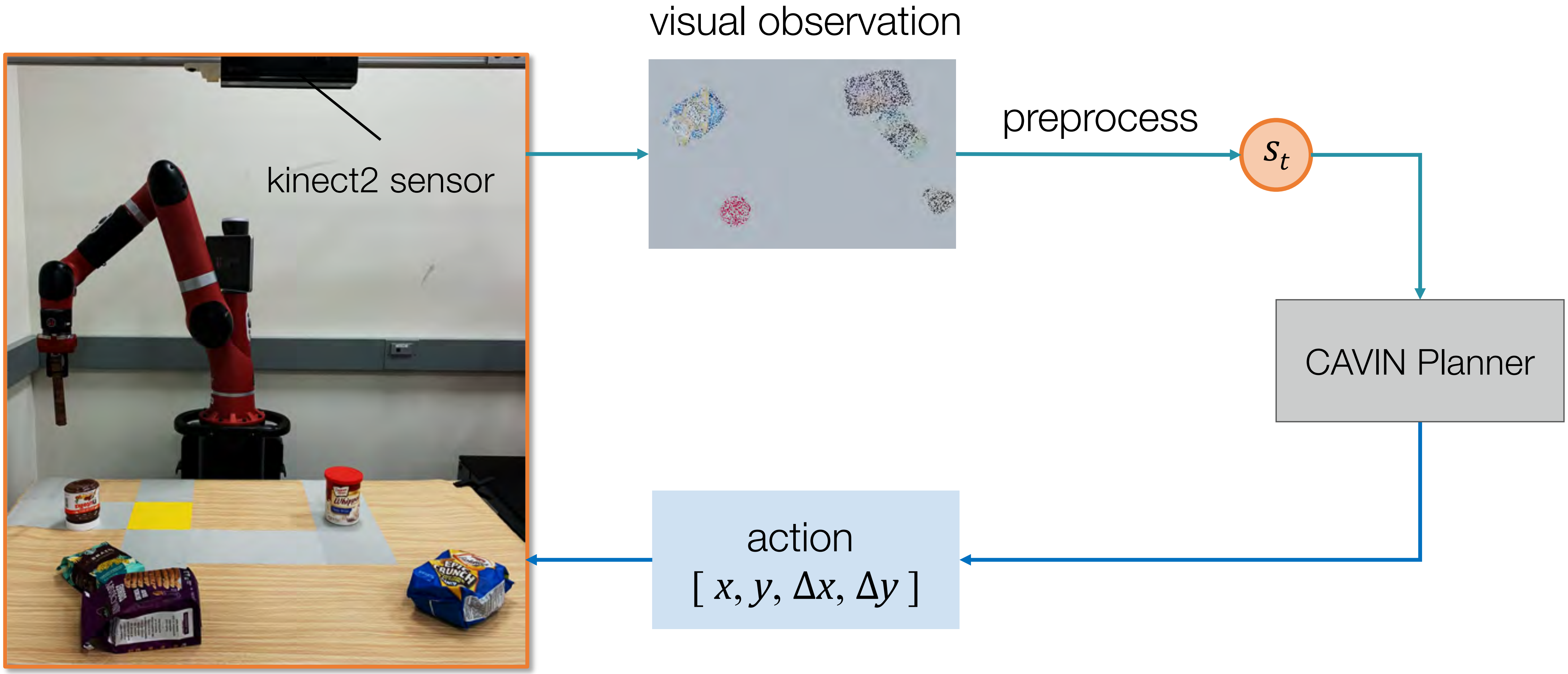


Learning with cascaded variational inference

task-agnostic interaction







visual observation
kinect2 sensor
preprocess
s_t
CAVIN Planner
action
$[\, x, y, \Delta x, \Delta y \,]$

Tasks

clearing

Clear all objects within the area of blue tiles.

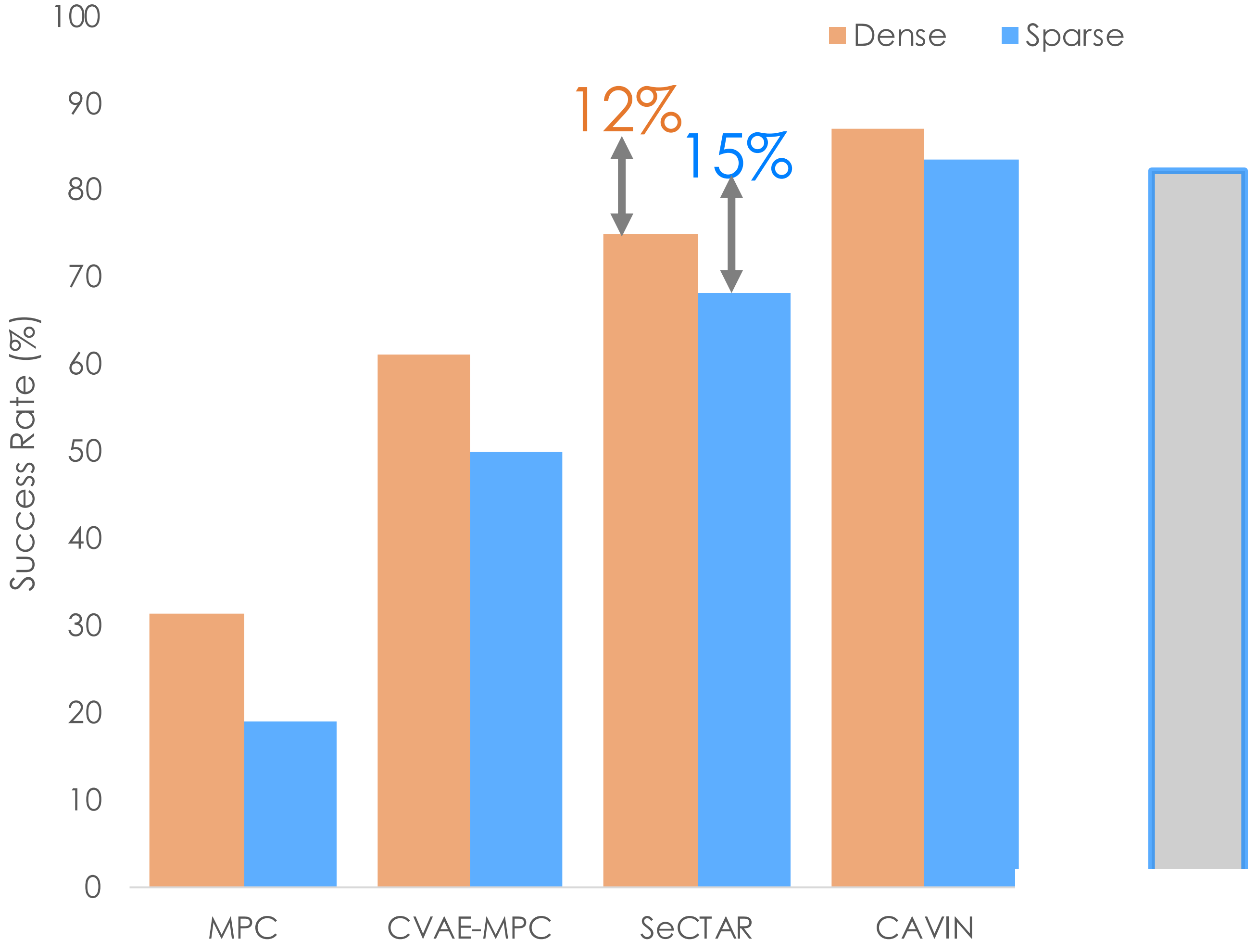
insertion

Move the target to the goal without traversing red tiles.

crossing

Move the target to the goal across grey tiles

Quantitative Evaluation



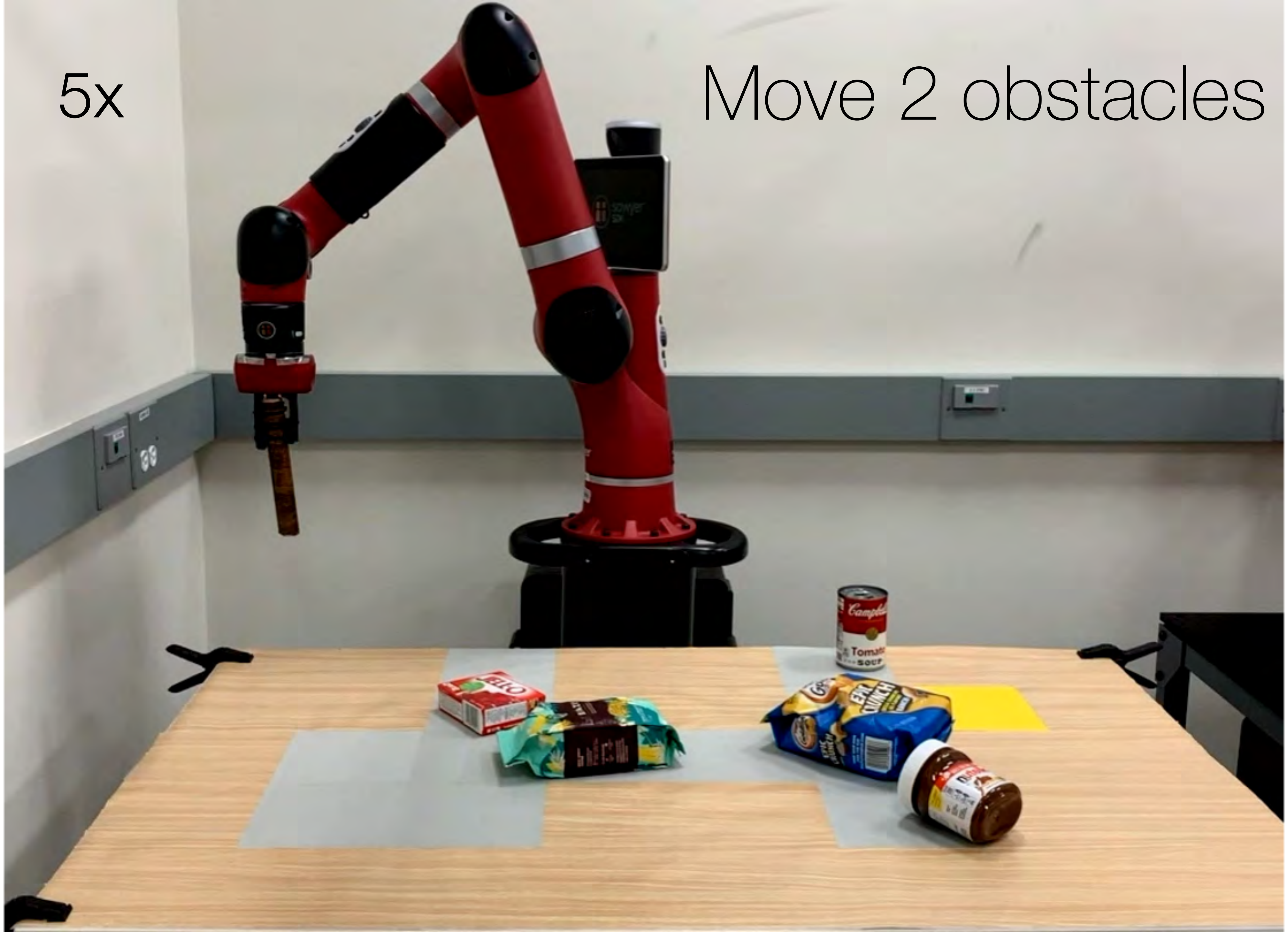
Hierarchical Latent space dyn.

↓

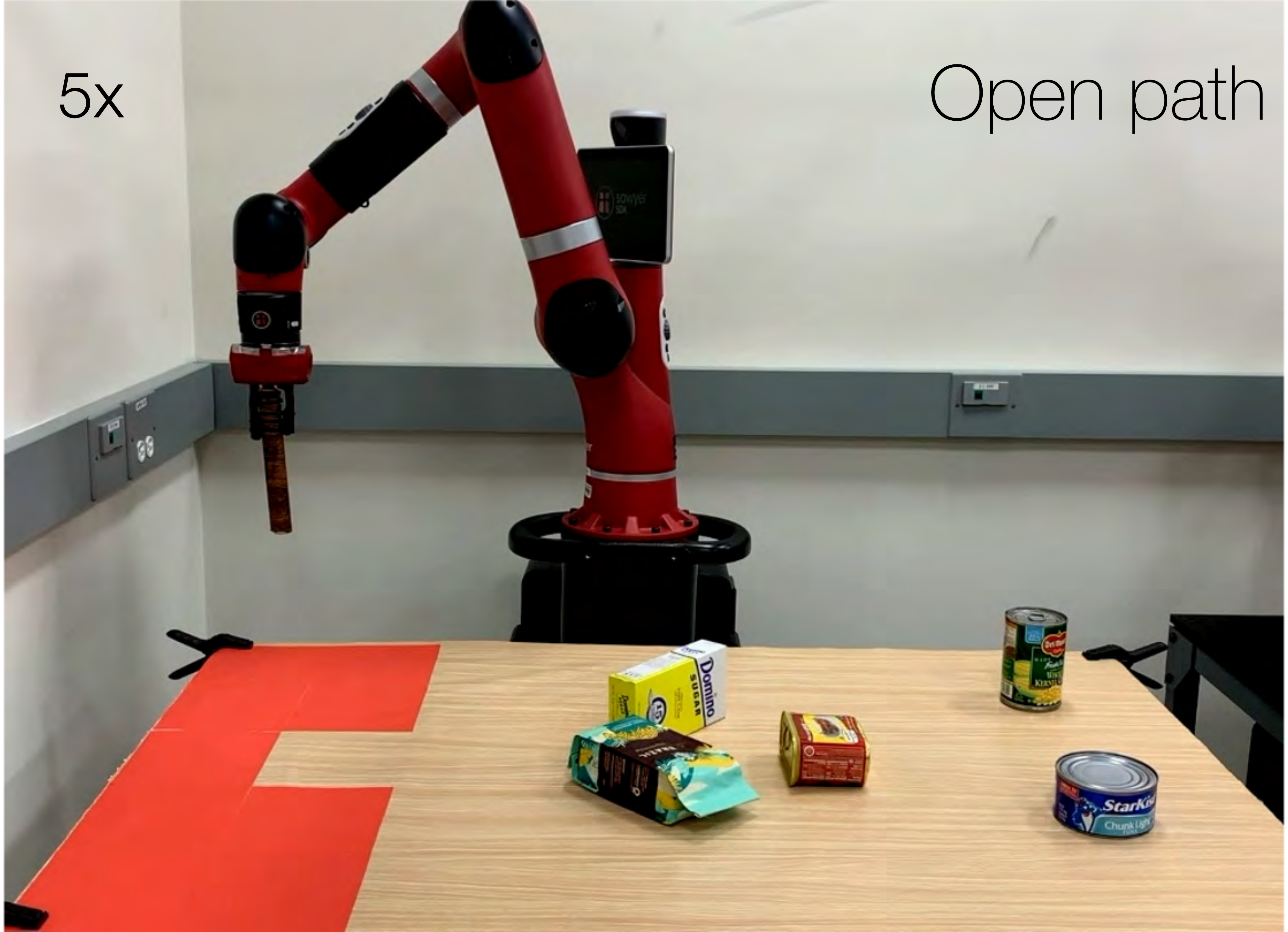
Better performance with sparse reward signal

Averaged over 3 Tasks
with 1000 test instances each

MPC (Guo et al. ‘14, Agrawal et al. ‘16, Finn et al. 17); CVAE-MPC (Ichter et al. 18), SeCTAR (Co-Reyes et al ‘18)


5x
Move 2 obstacles


5x
Get around
nutella

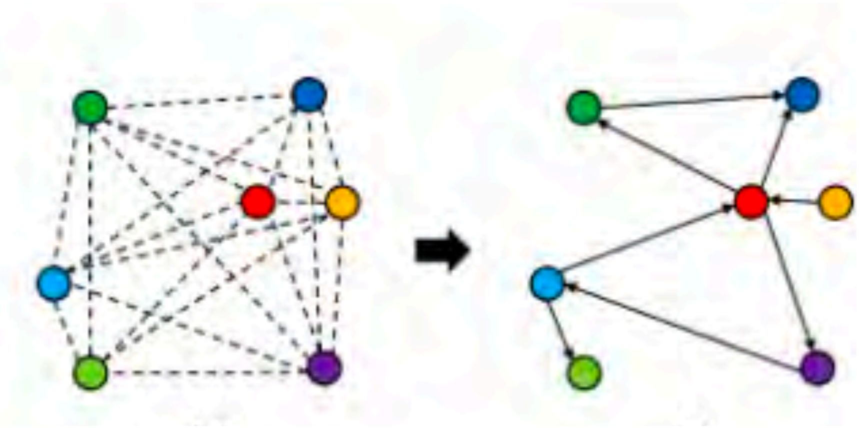

5x
Open path
Domino
SUGAR
StarKist
Chunk Light

Compositional Representations

Learning Disentanglement

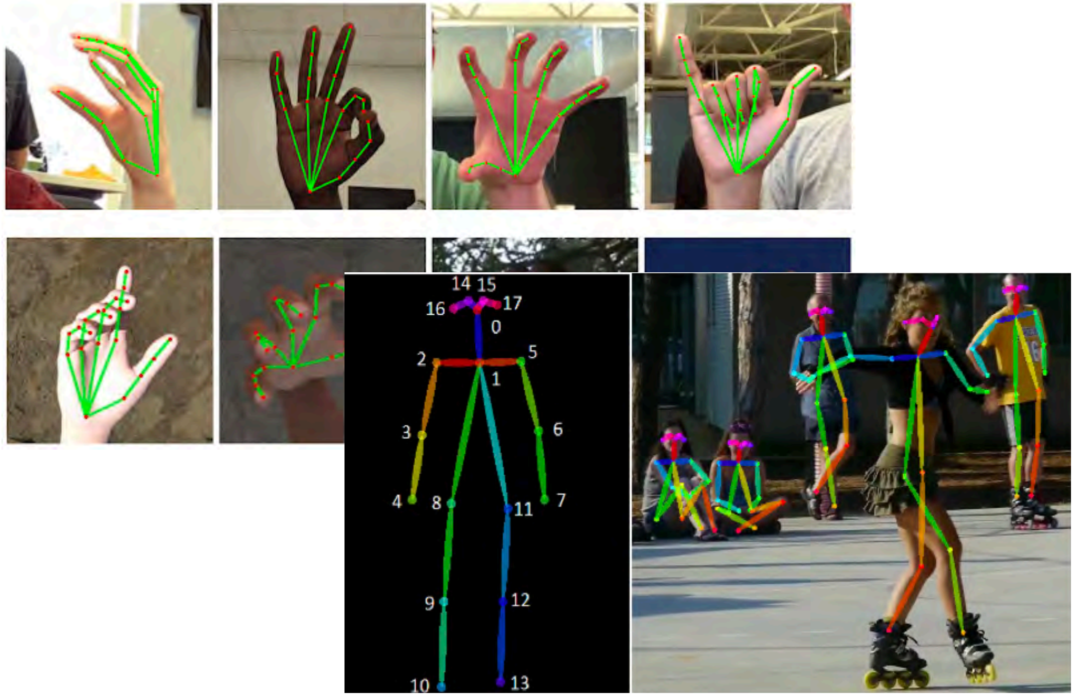
Learning Keypoints

Learning Causal Graphs



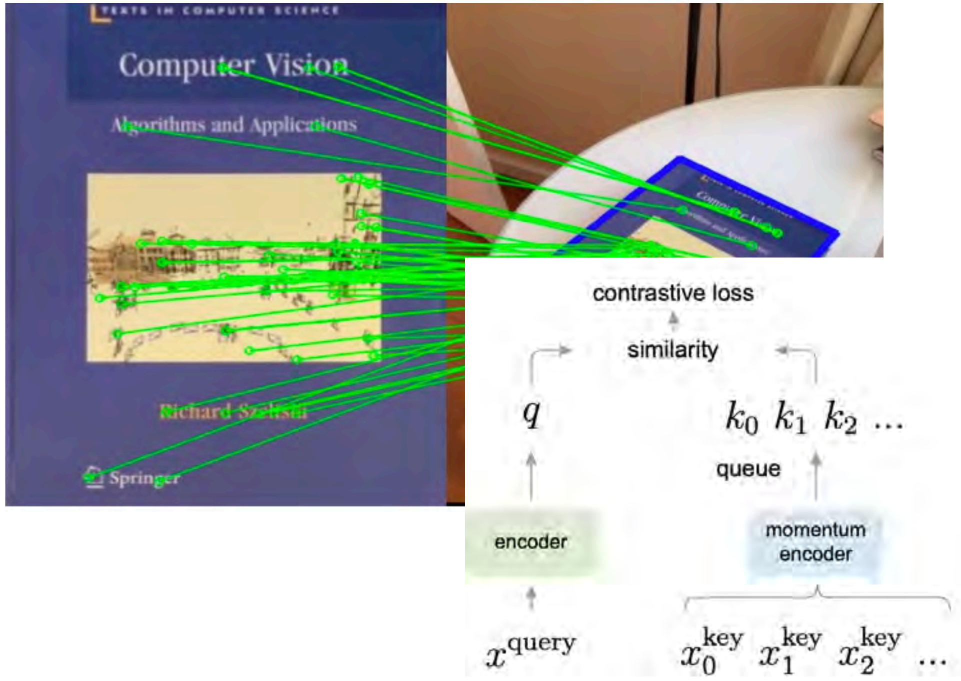
Composition through Keypoints

Interpretable



He et al 2017, Kreiss et al 2019, Lin et al 2020, Sprurr et al 2020

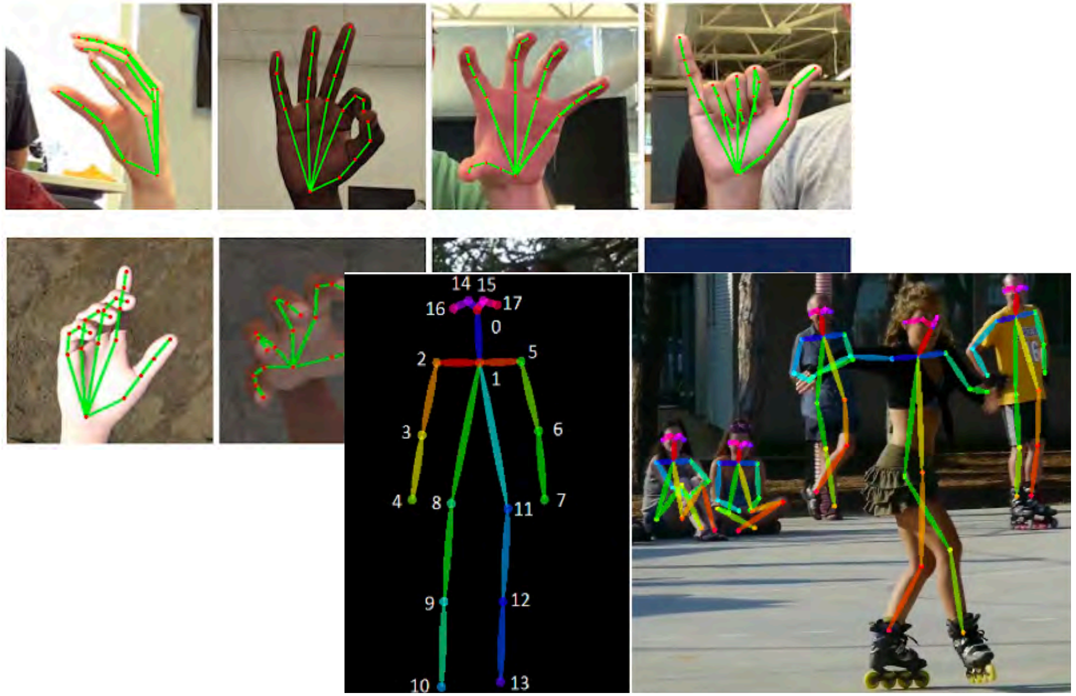
Unsupervised



Tang et al 2019, Christiansen et al 2019, Bian et al 2019, He et al 2019, Chen et al 2019

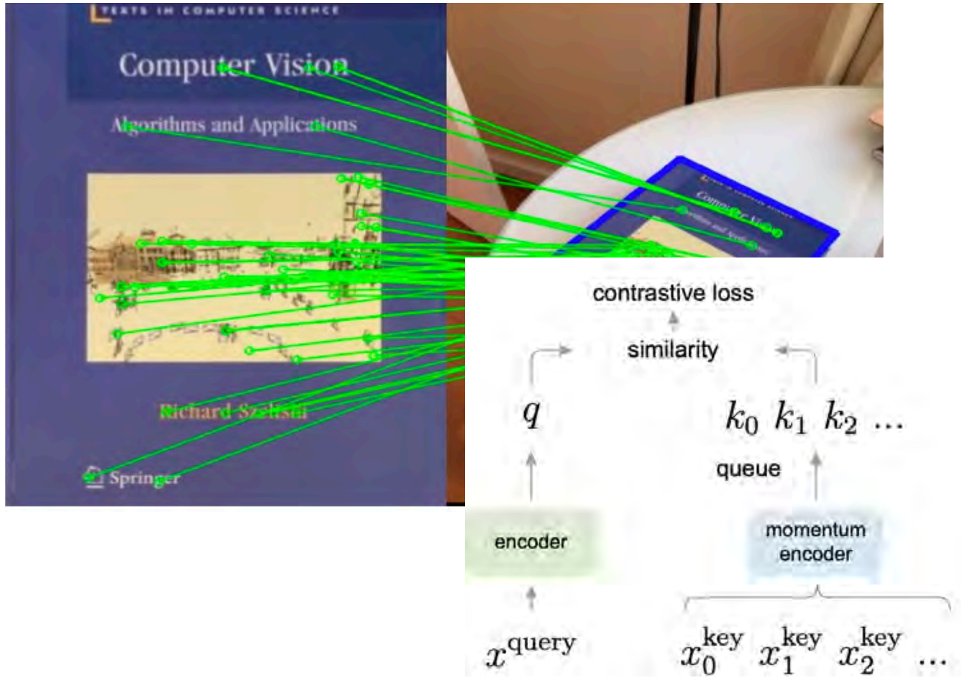
Composition through Keypoints

Interpretable



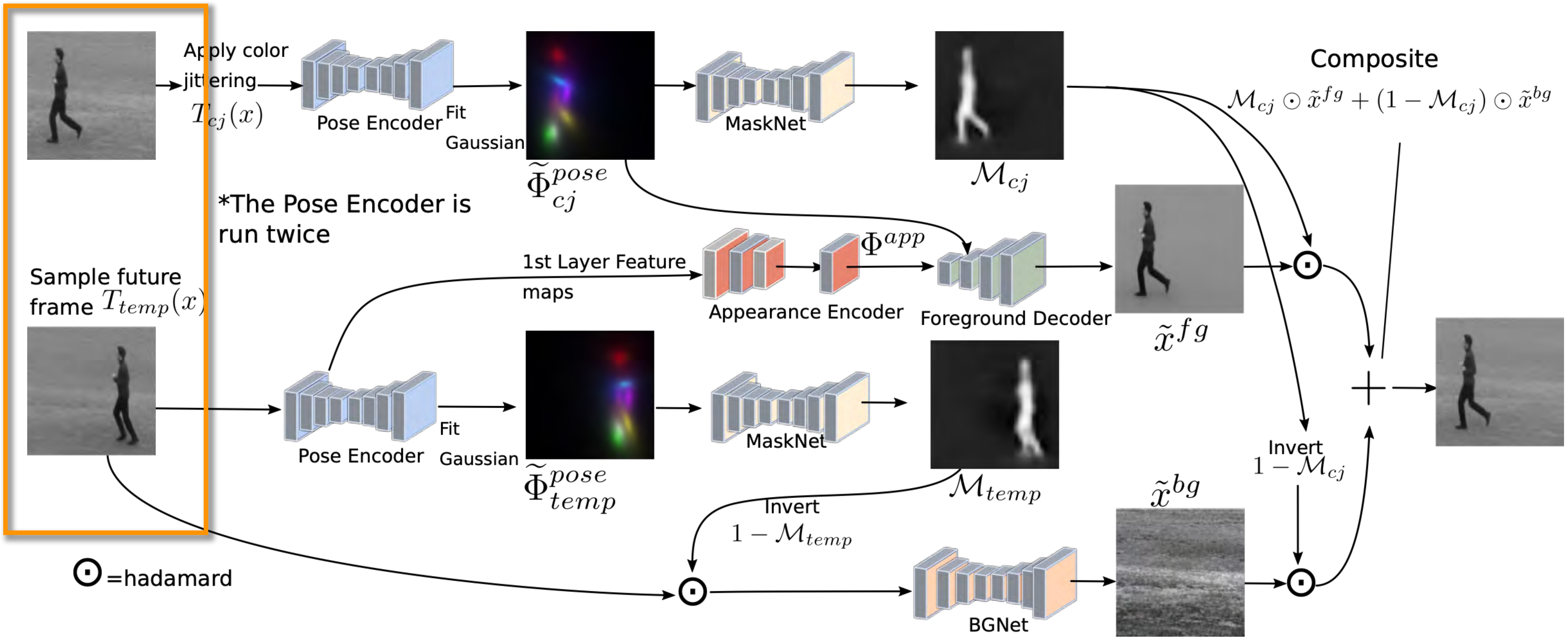
He et al 2017, Kreiss et al 2019, Lin et al 2020, Sprurr et al 2020

Unsupervised

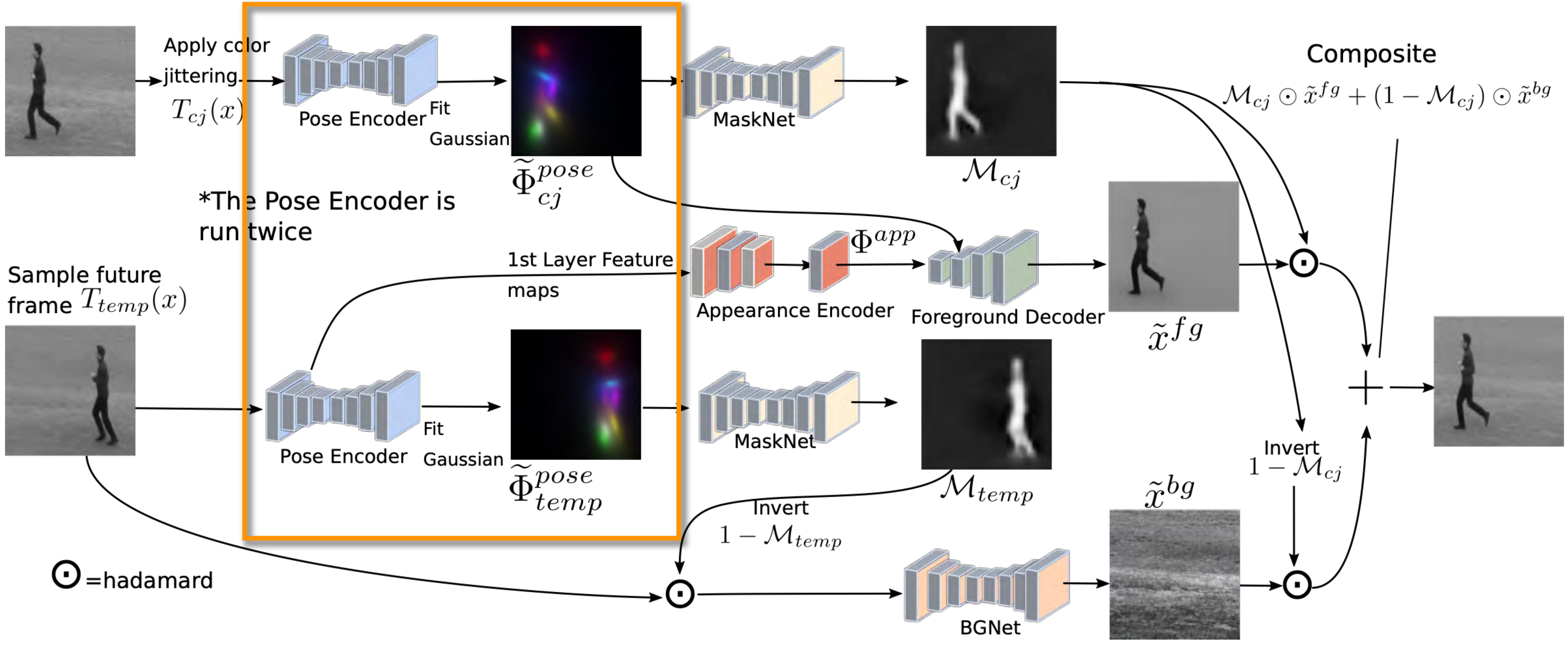


Tang et al 2019, Christiansen et al 2019, Bian et al 2019, He et al 2019, Chen et al 2019

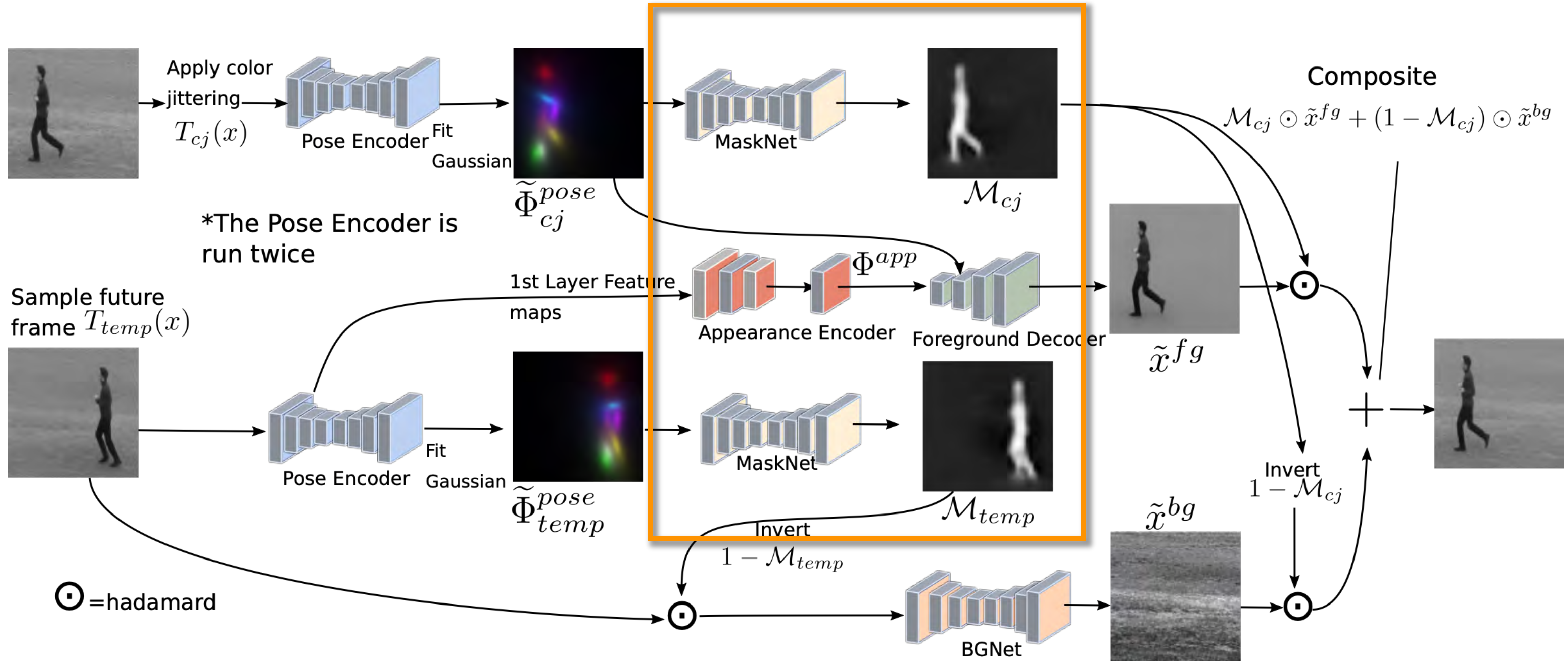
Learning Keypoints From Video



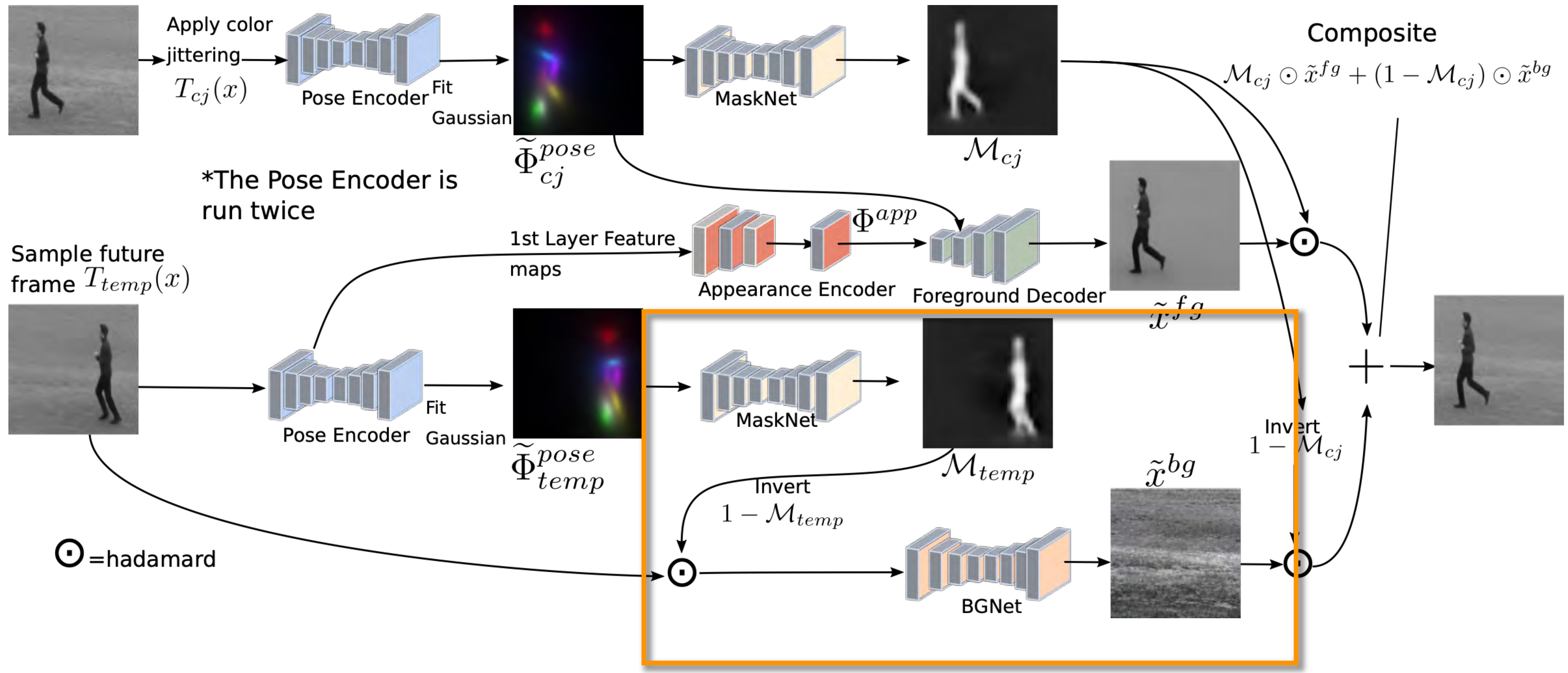
Learning Keypoints From Video



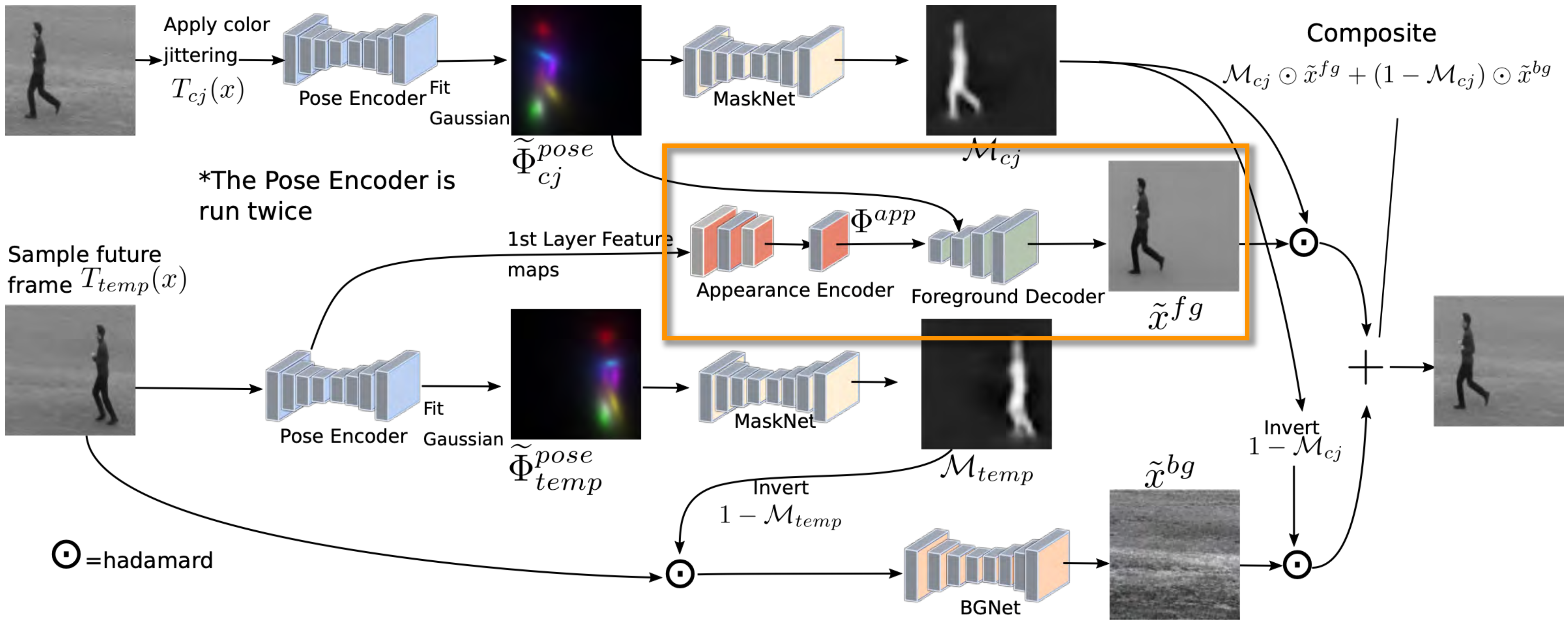
Learning Keypoints From Video



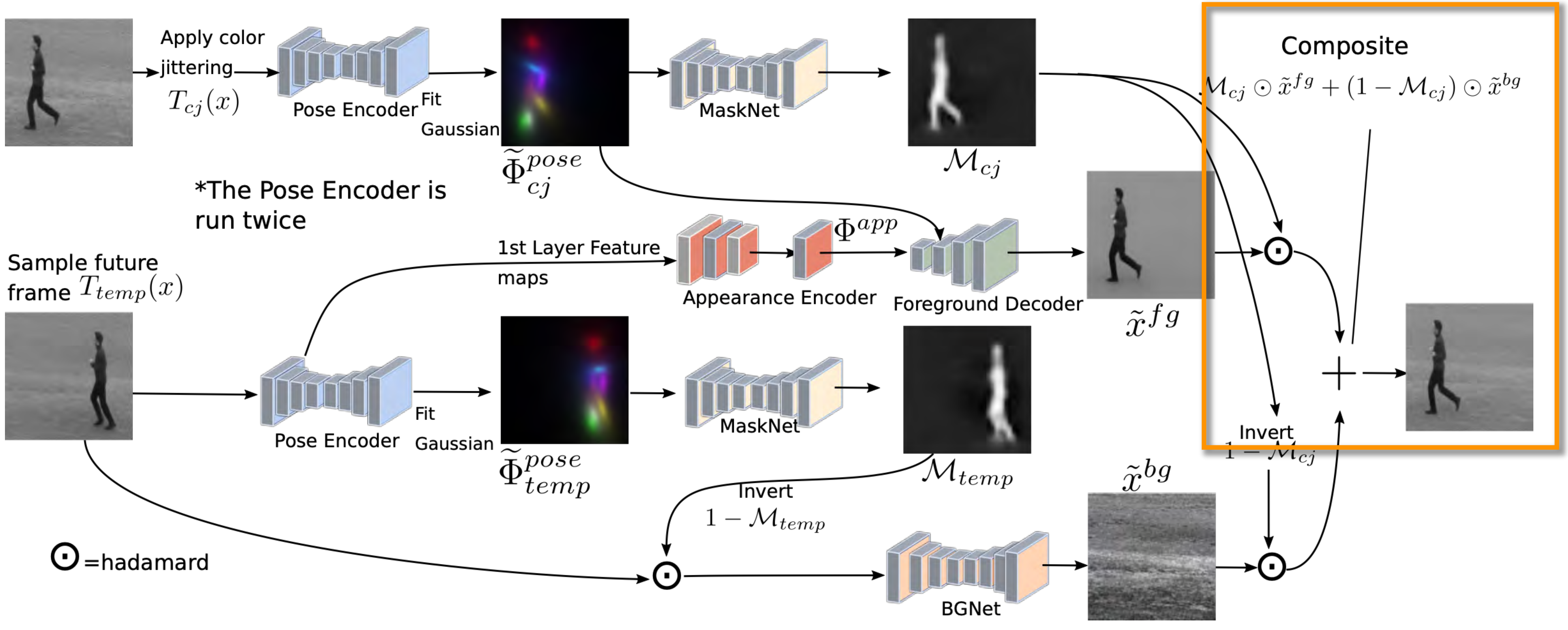
Learning Keypoints From Video



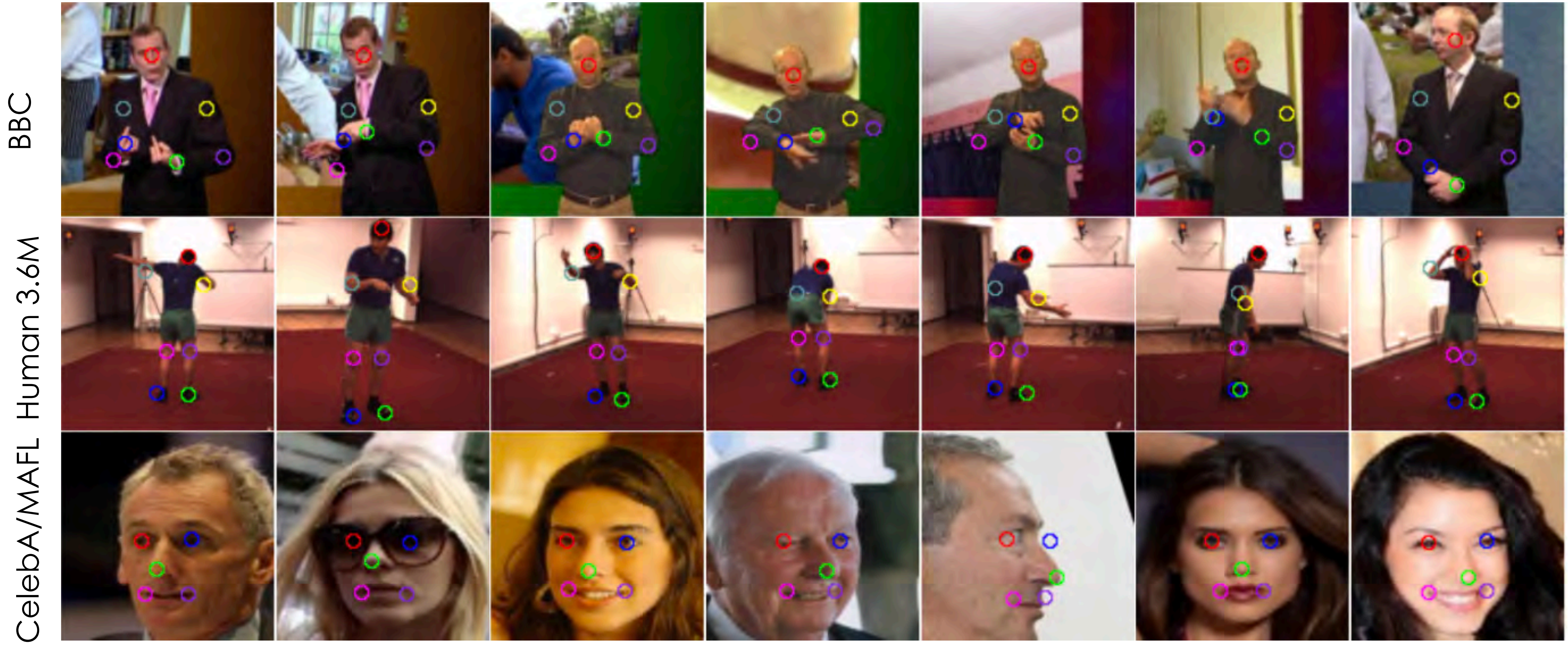
Learning Keypoints From Video



Learning Keypoints From Video



Learning Keypoints From Video



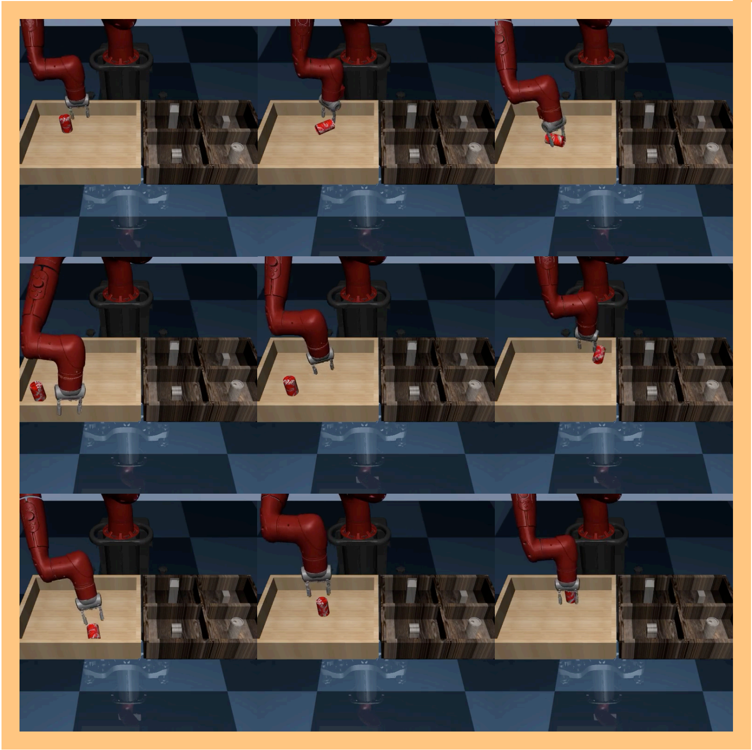
Learning Keypoints From Video

BBC Pose		Acc.
supervised	Charles et al. [3]	79.9%
	Pfister et al. [21]	88.0%
unsupervised	Jakab et al. [8]	68.4%
	Lorenz et al. [14]	74.5%
	Baseline (temp)	73.3%
	Baseline (temp,tps)	73.4%
	Ours	78.8%

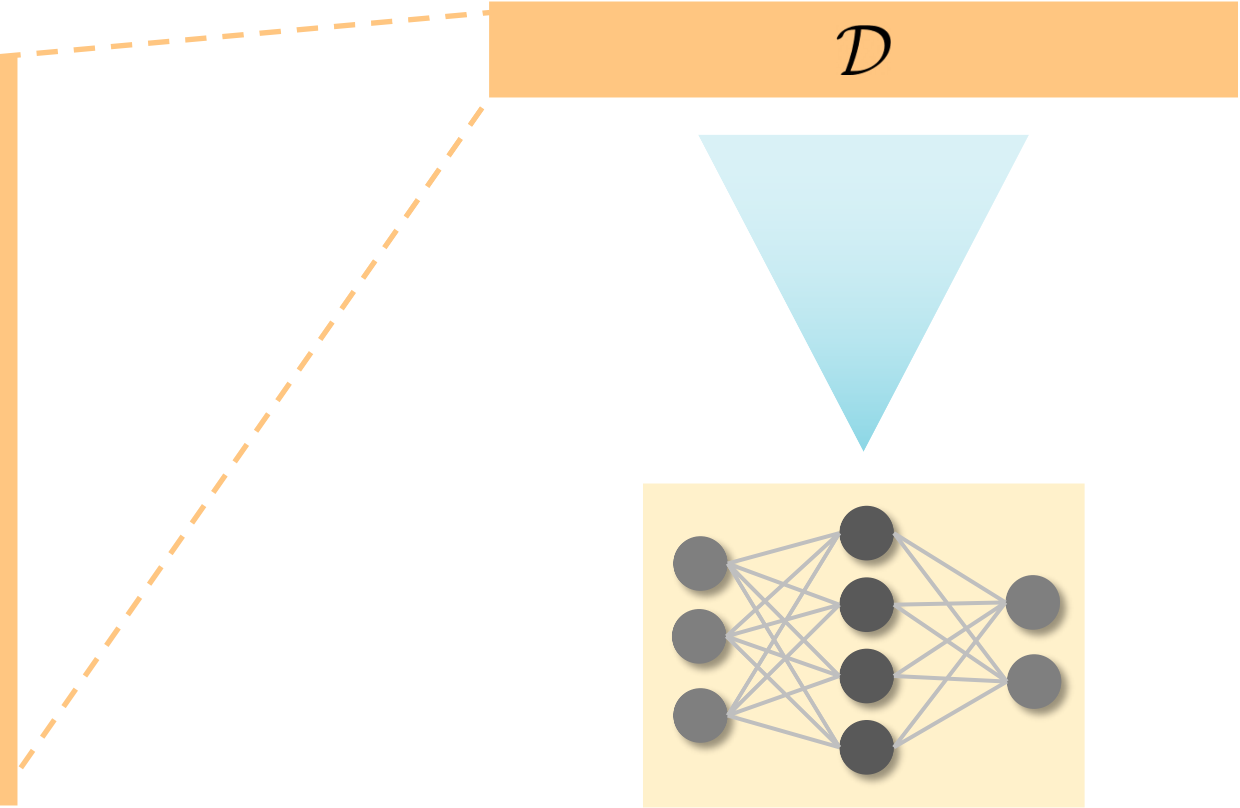
Human3.6M		Error
supervised	Newell et al. [18]	2.16
unsupervised	Thewlis et al. [33]	7.51
	Zhang et al. [41]	4.91
	Lorenz et al. [14]	2.79
	Baseline (temp)	3.07
	Baseline (temp,tps)	2.86
	Ours	2.73

MAFL		Error
unsupervised	Thewlis et al. [33]	6.32
	Zhang et al. [41]	3.46
	Lorenz et al. [14]	3.24
	Jakab et al. [8]	3.19
	Baseline (tps)	4.34
	Ours (No Mask)	2.88
	Ours	2.76

Unsupervised Keypoints: Batch RL

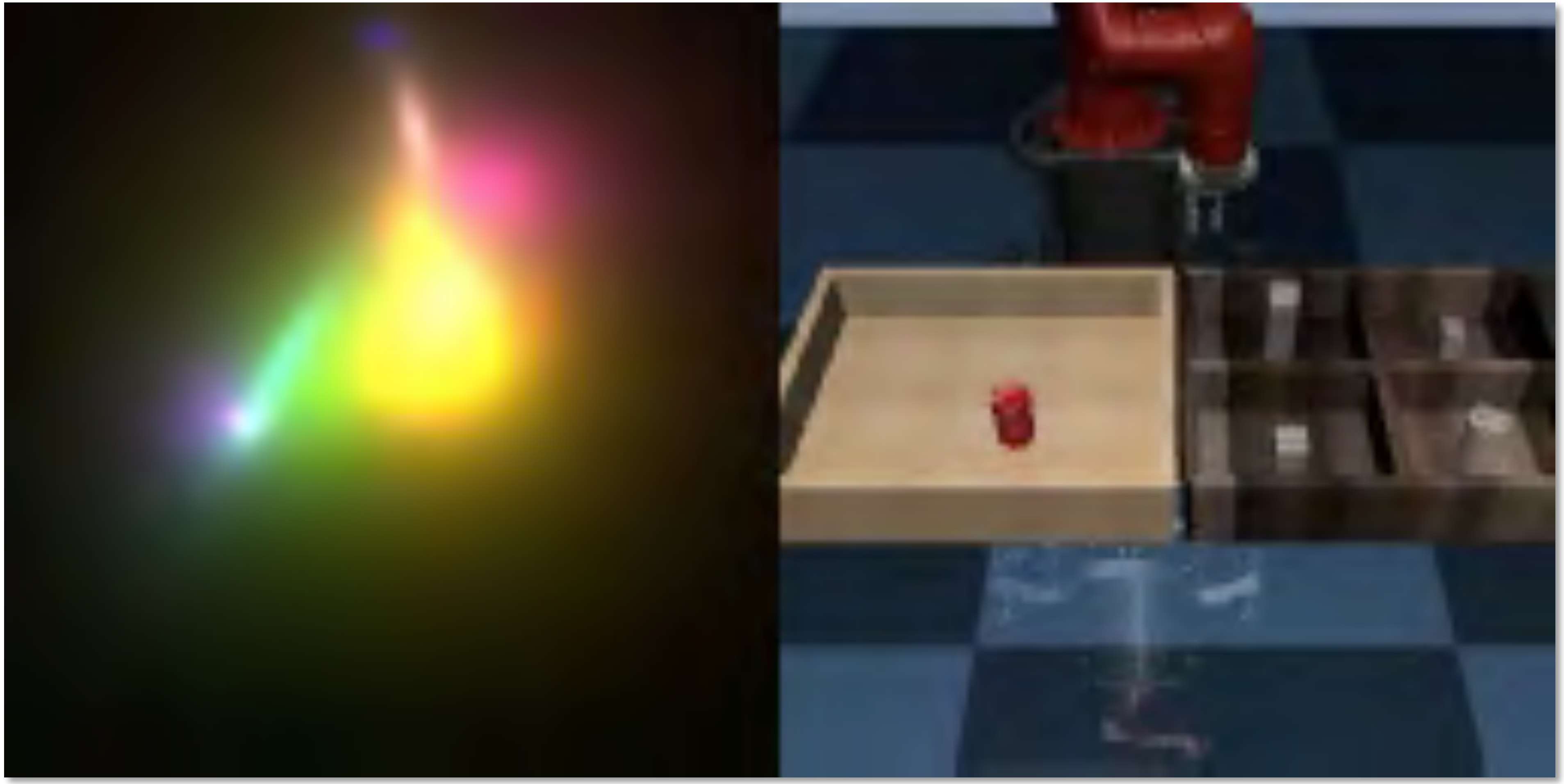


Large Set of
Task Demonstrations



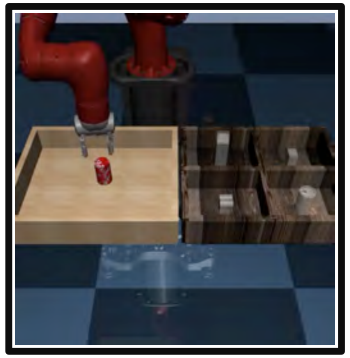
Policy Learning
without Interaction

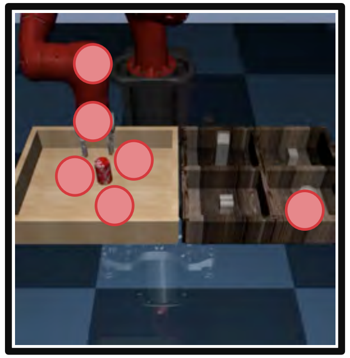
Unsupervised Keypoints: Batch RL

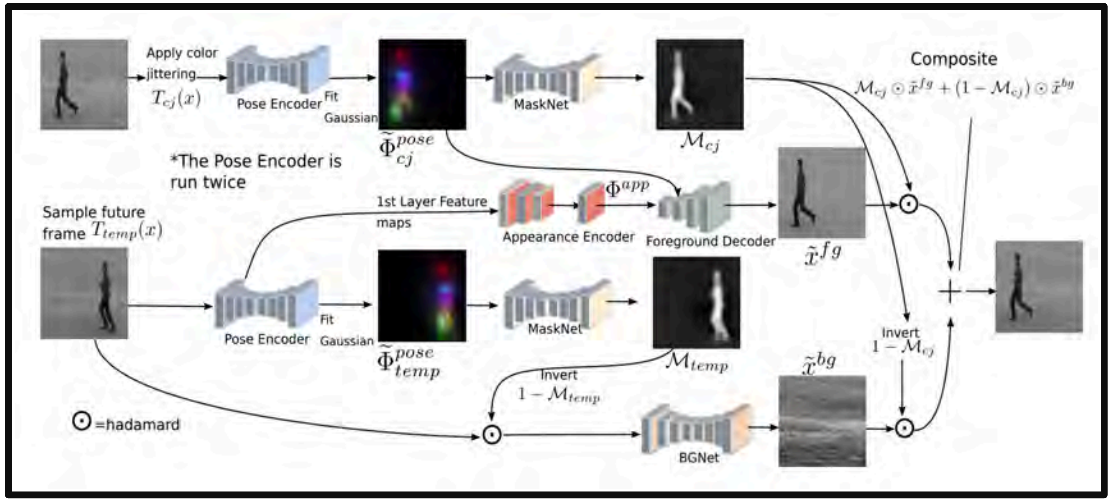


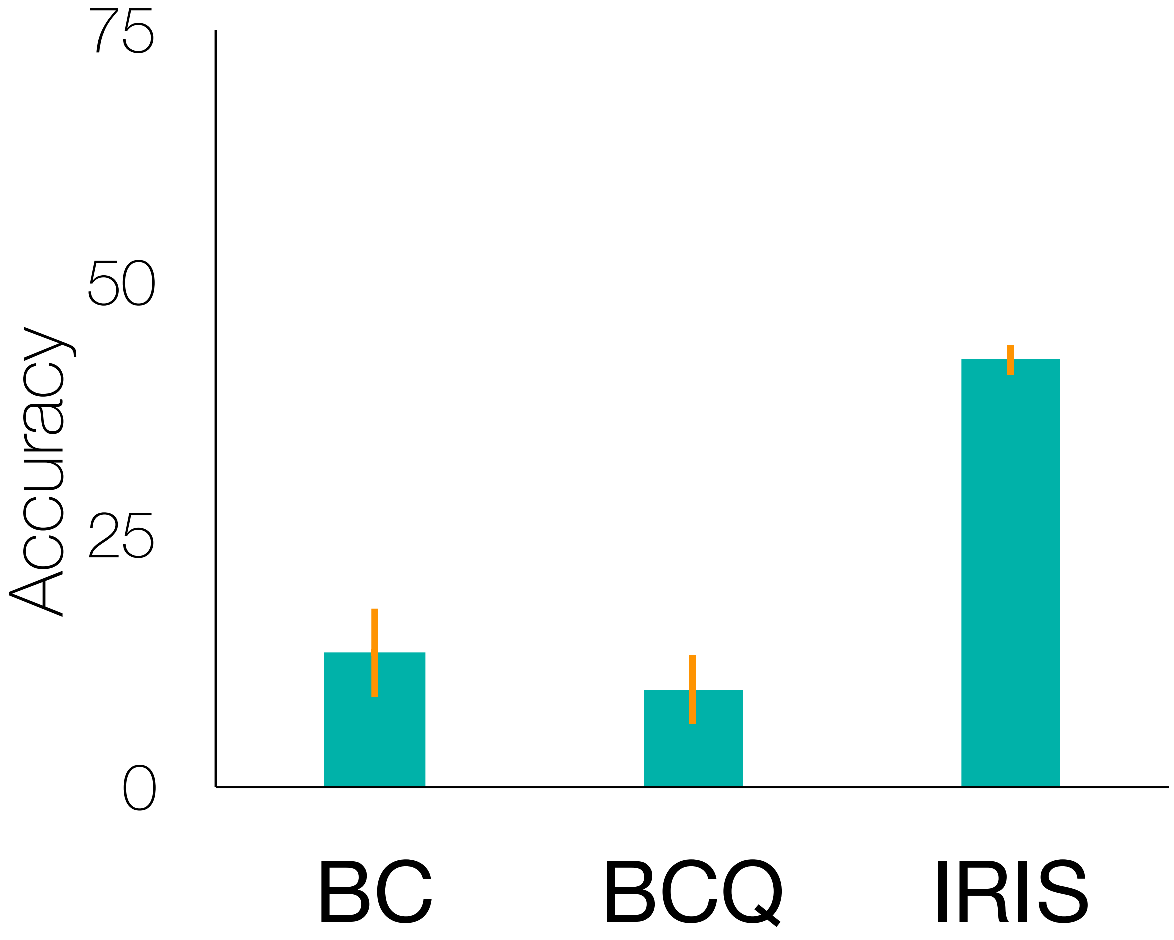
Unsupervised Keypoints: Batch RL

Unsupervised Representation Learning



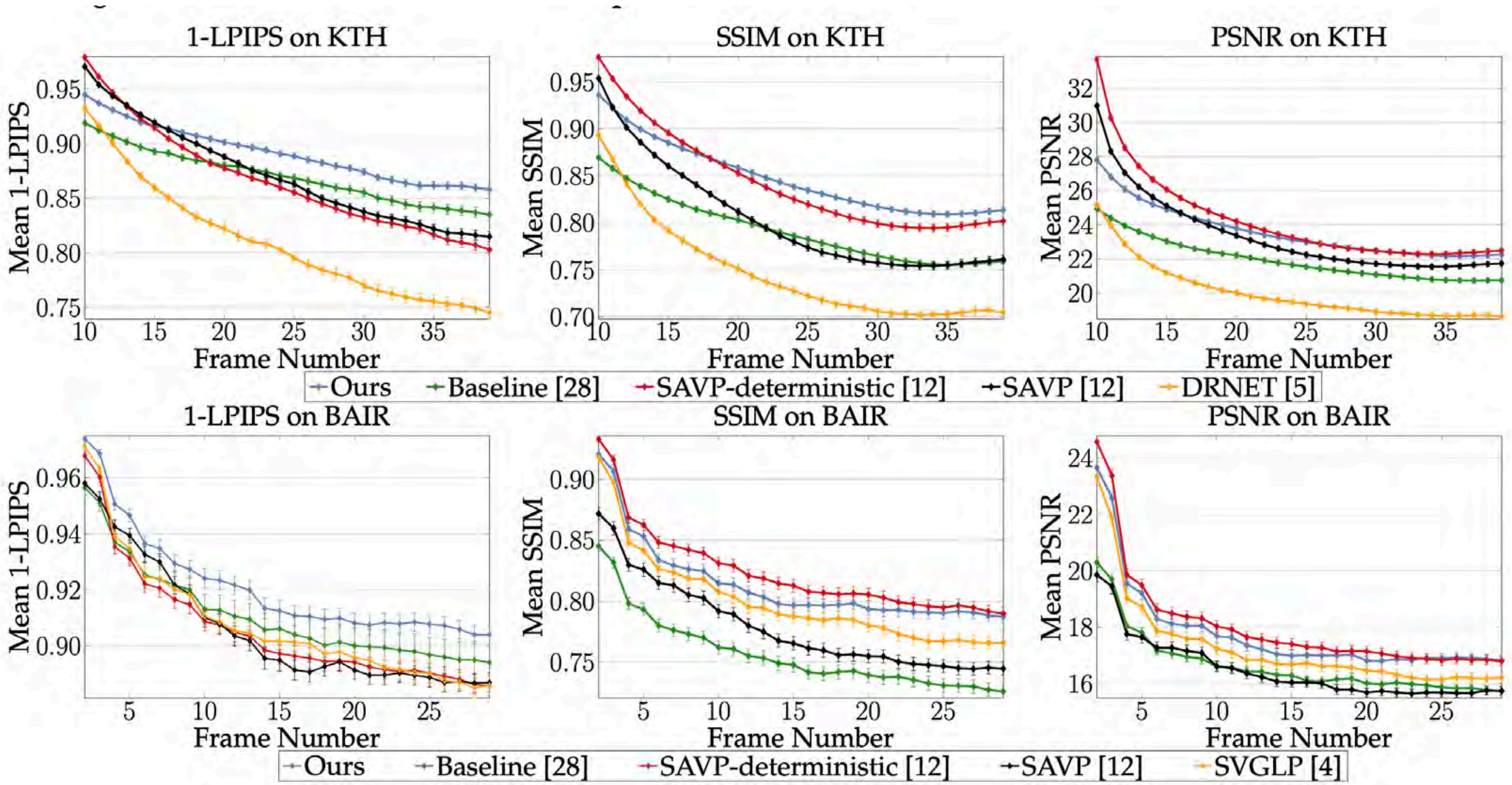




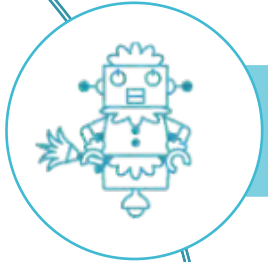


Unsupervised Keypoints: Video Prediction

Unsupervised Keypoints: Video Prediction



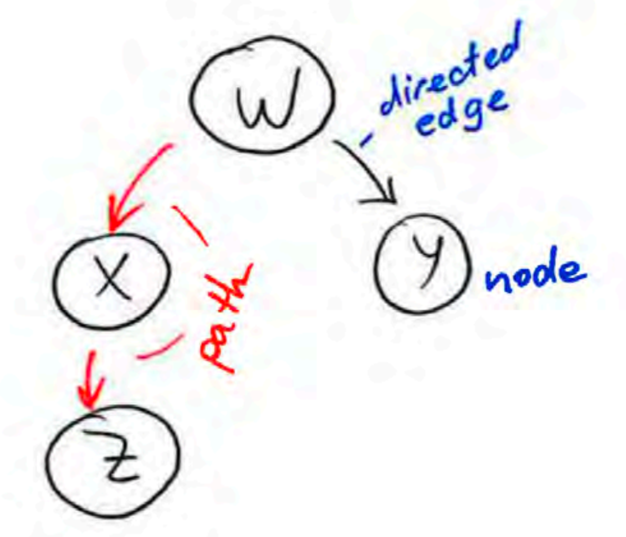
Compositional Representations



Learning Disentanglement

Learning Keypoints

Learning Causal Graphs



Learning Causality

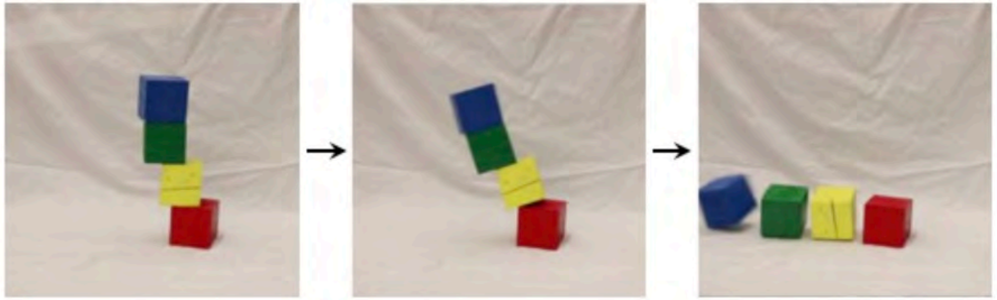
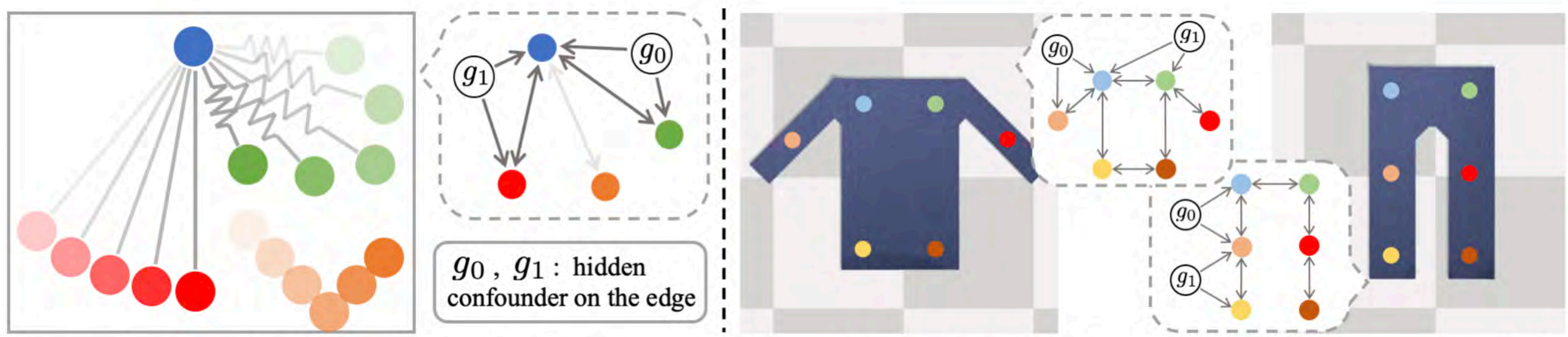


Photo from Wu et al., Learning to See Physics via Visual De-animation

- Intuitive Physics

- vs Reinforcement Learning
  - Generalization
  - Goal specification
  - Sample efficiency
- vs Analytical Physics Model
  - Underlying dynamics is uncertain or unknown
  - Simulation is too time consuming
  - Partial observation

Learning Causality



Learning Causality

- Intuitive Physics

- State Representation?
    - Keypoints

    - vs. 6 DoF pose

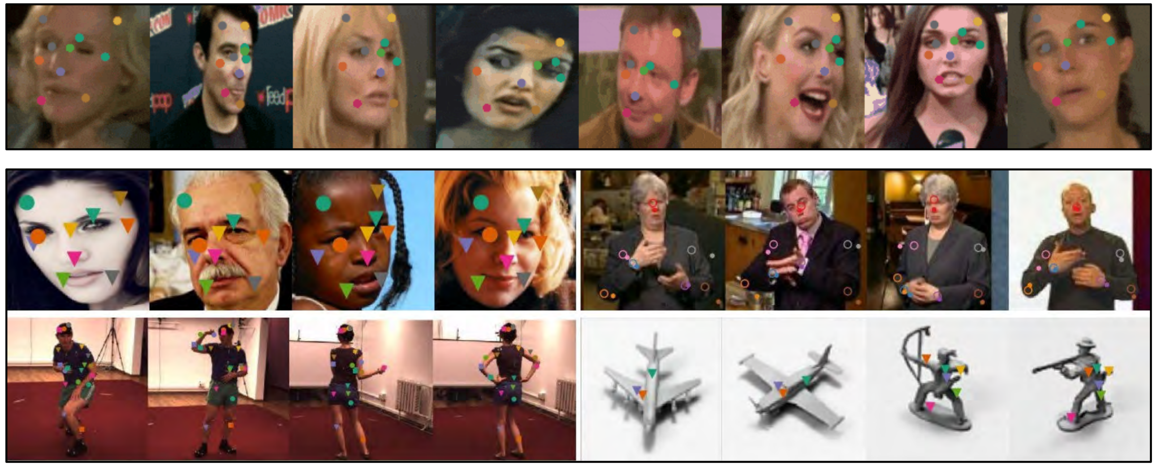


Photo from Jakab et al., Unsupervised Learning of Object Landmarks through Conditional Image Generation

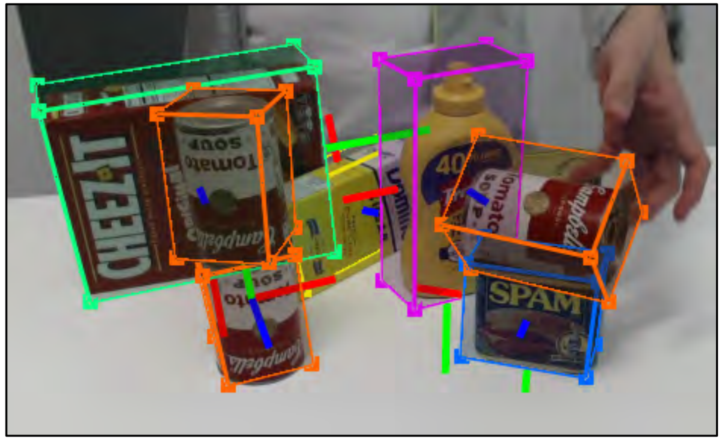
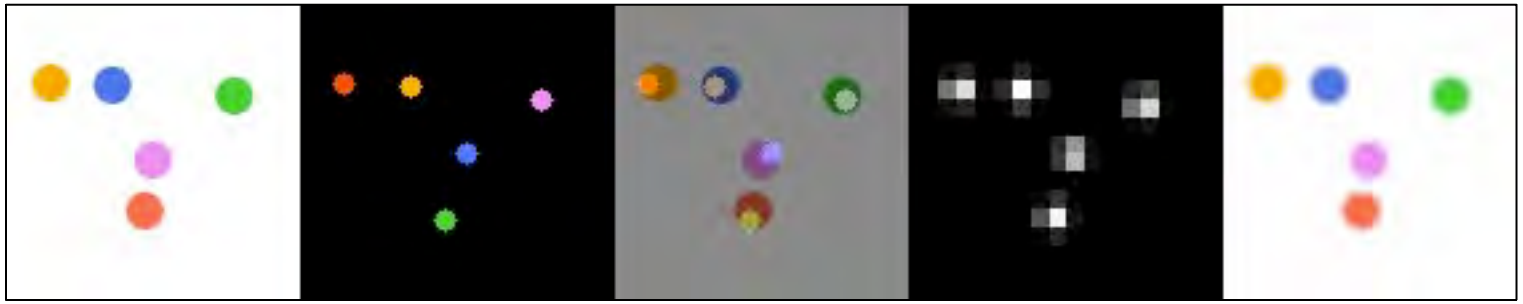
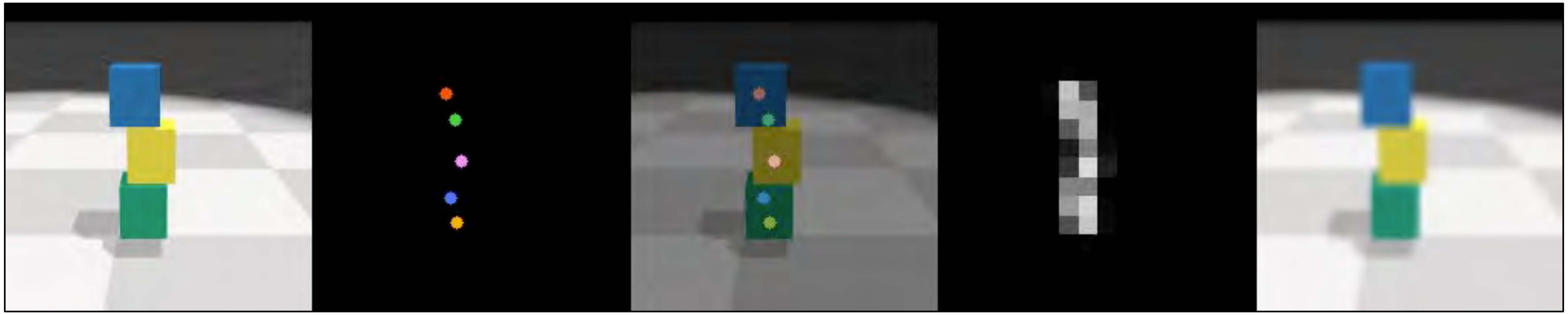


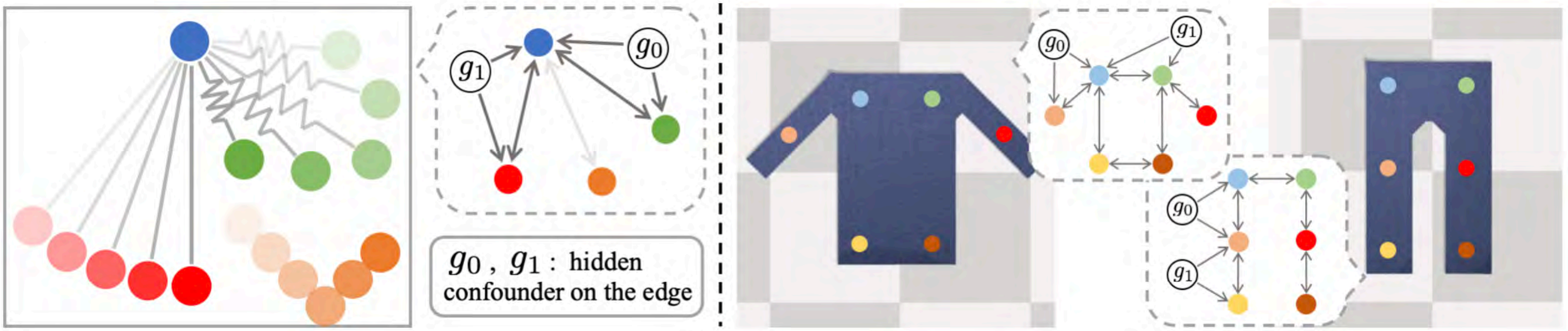
Photo from Tremblay et al., Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects

From left to right:
(1) Input image
(2) Predicted keypoints
(3) Overlay
(4) Heatmap from the keypoints
(5) Reconstructed target image

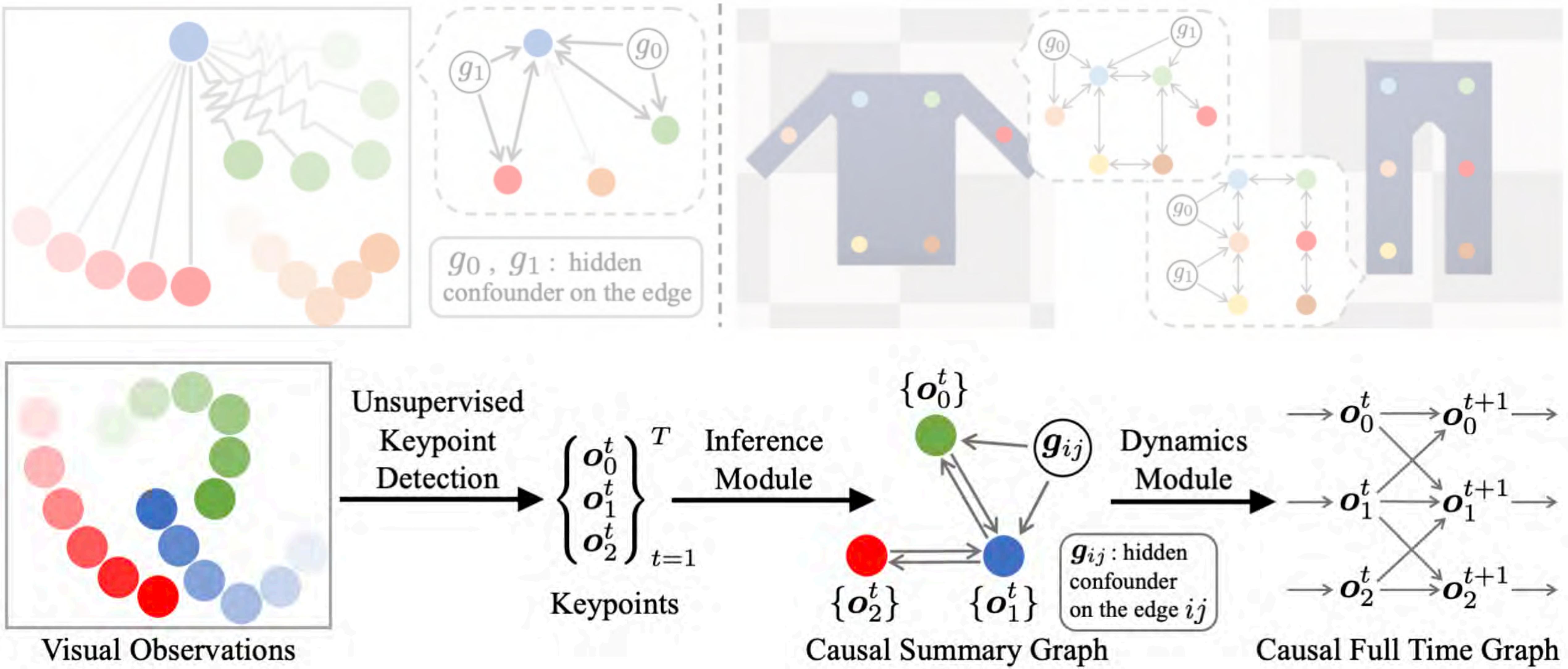




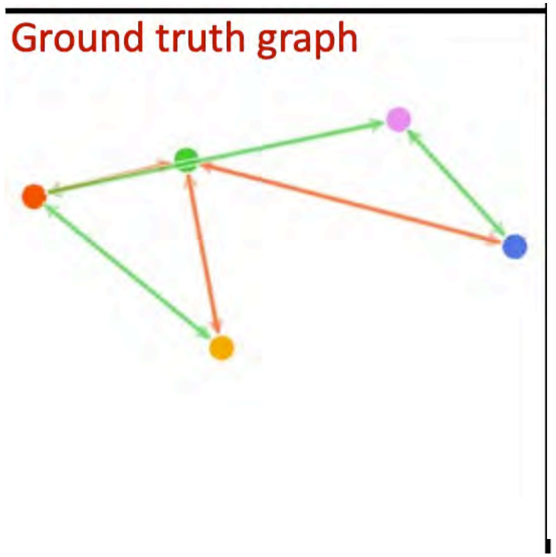
Learning Causality

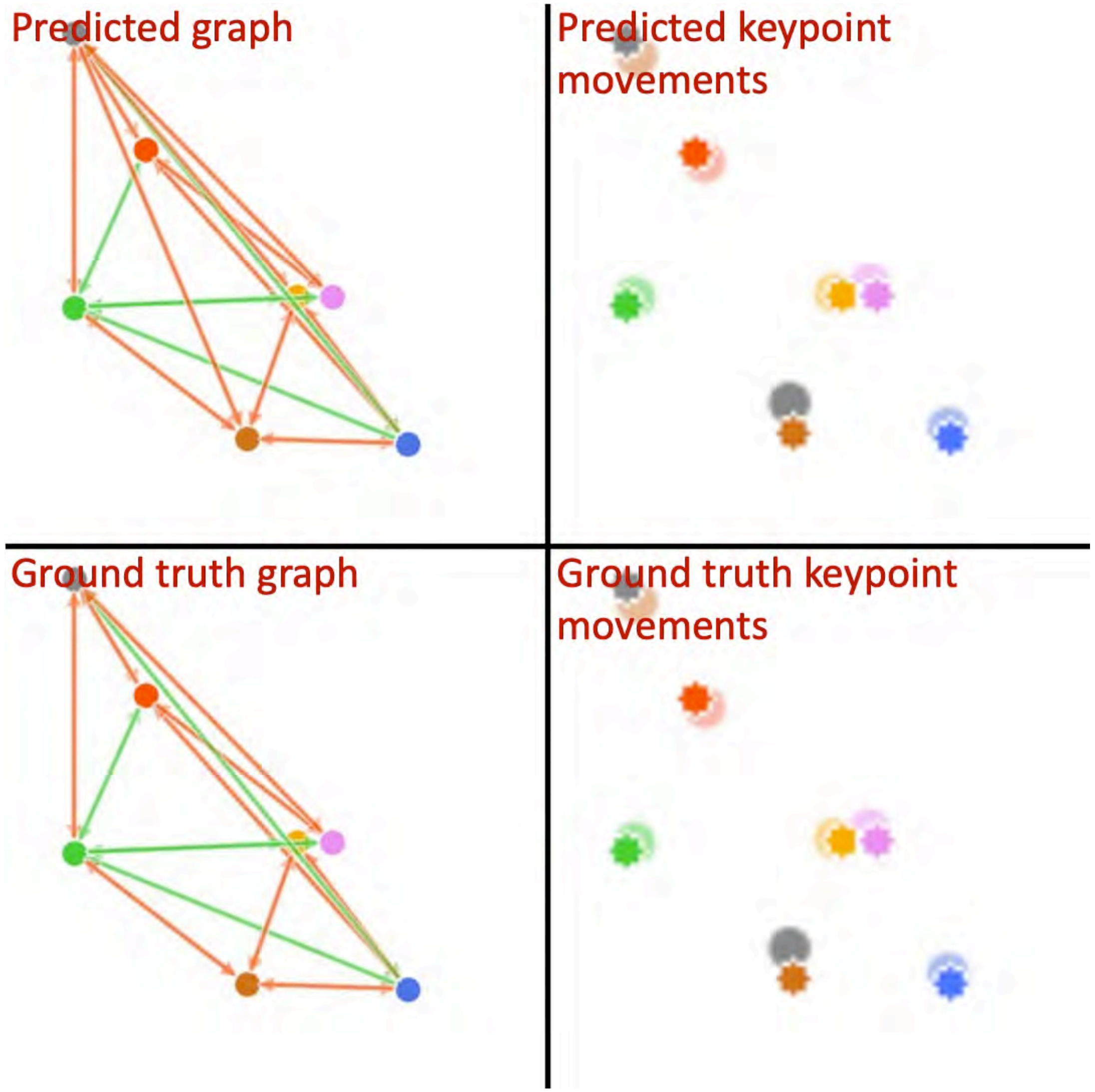


Learning Causality

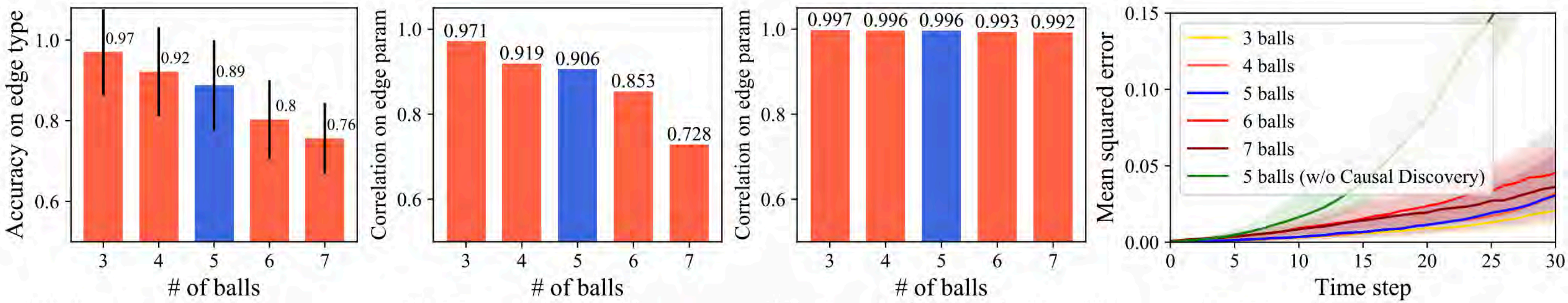


Learning Causality





Learning Causality: Extrapolation



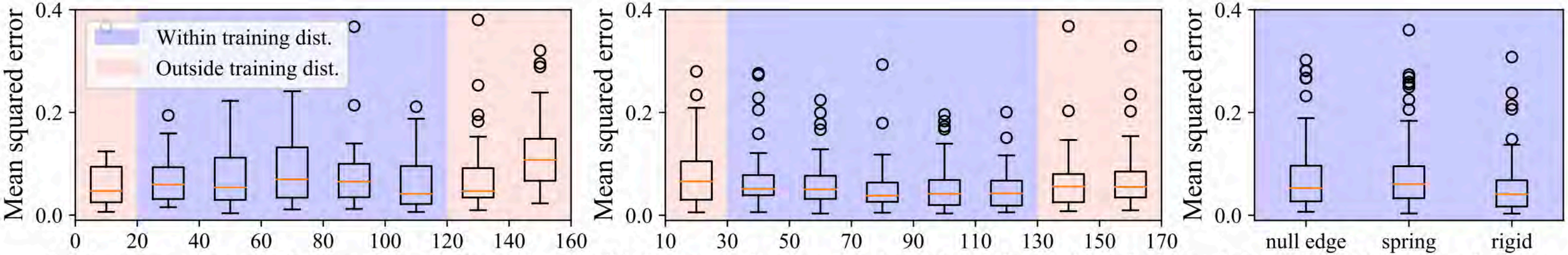
(a) Accuracy on edge type: {null edge, spring, rigid}

(b) Correlation on the rest length of the spring relation

(c) Correlation on the length of the rigid relation

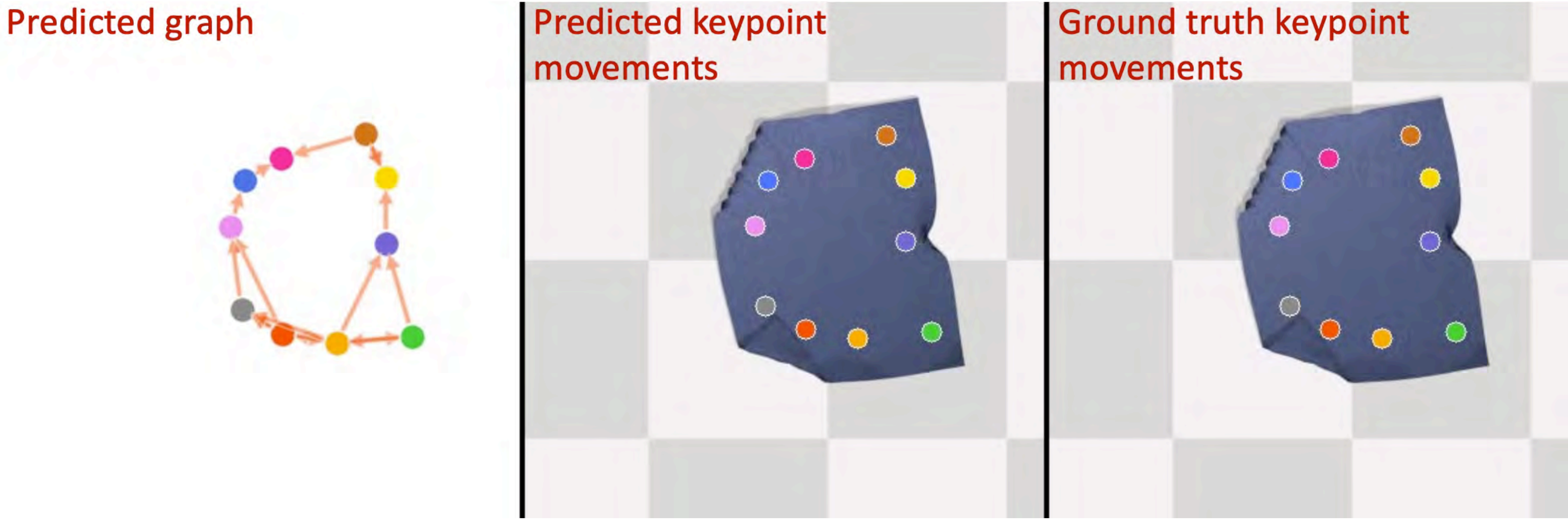
(d) Mean squared error on future prediction

Learning Causality: Counterfactual

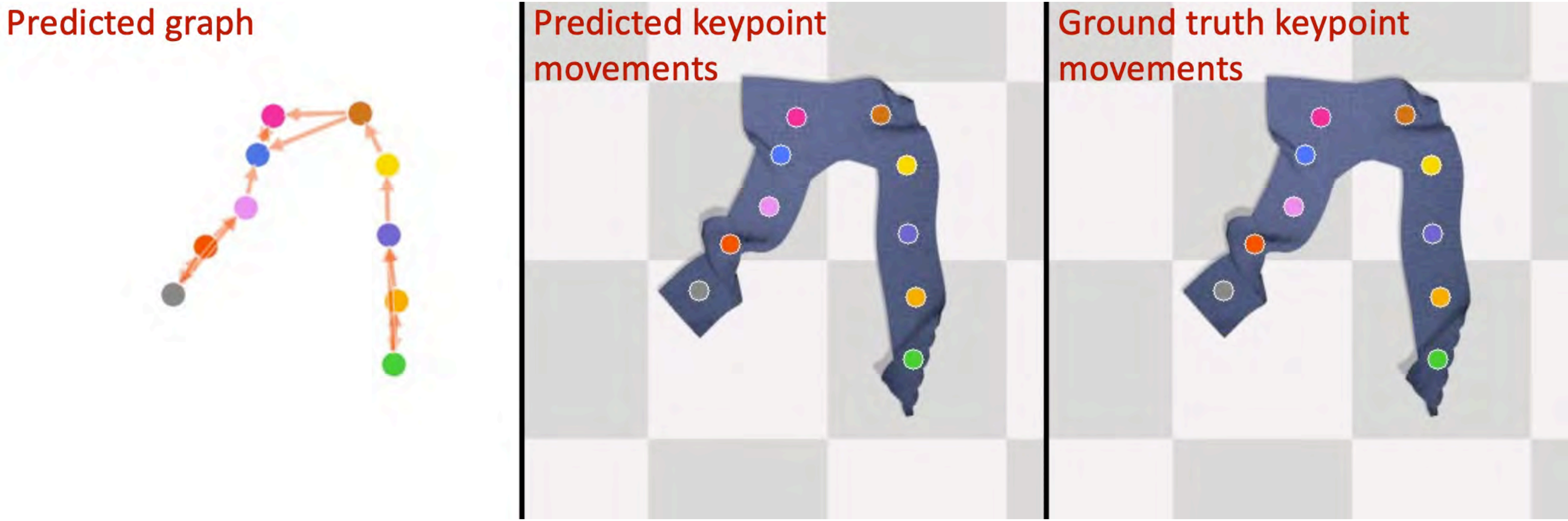


(a) Intervention on the rest length in spring (b) Intervention on the length of the rigid relation (c) Intervention on edge type

Learning Causality

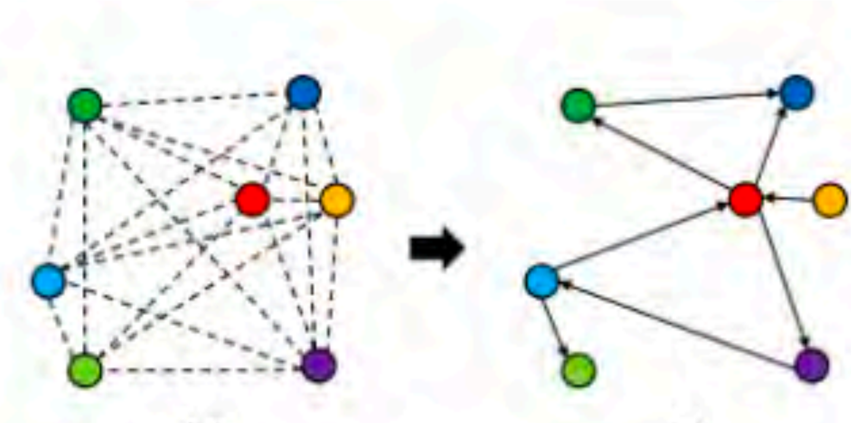


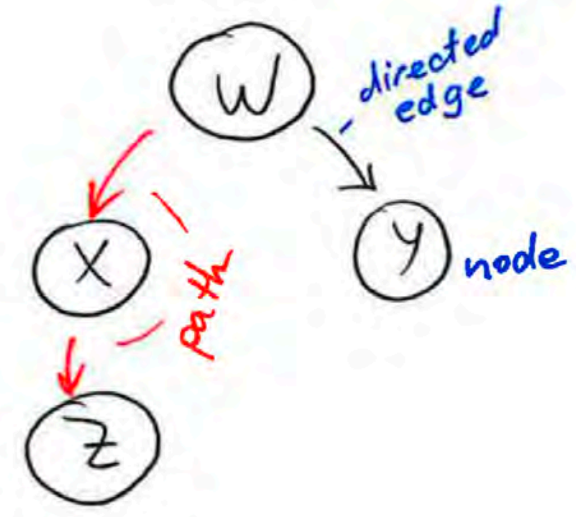
Learning Causality



Learning Causality

Unsupervised Representations towards Counterfactual Predictions





garg@cs.toronto.edu // pair.toronto.edu // Twitter: @Animesh_Garg

Animesh Garg | UNIVERSITY OF TORONTO VECTOR INSTITUTE