

# The Objects of Our Curiosity

Intrinsic Motivation, Intuitive Physics and Self-Supervised Learning

---

NeurIPS Workshop: Modeling the Physical World

*2018.12.07*

**Daniel Yamins**

Stanford Neurosciences Institute

Stanford Artificial Intelligence Laboratory

Departments of Psychology and Computer Science

Stanford University



Our work is founded on two mutually reinforcing goals:

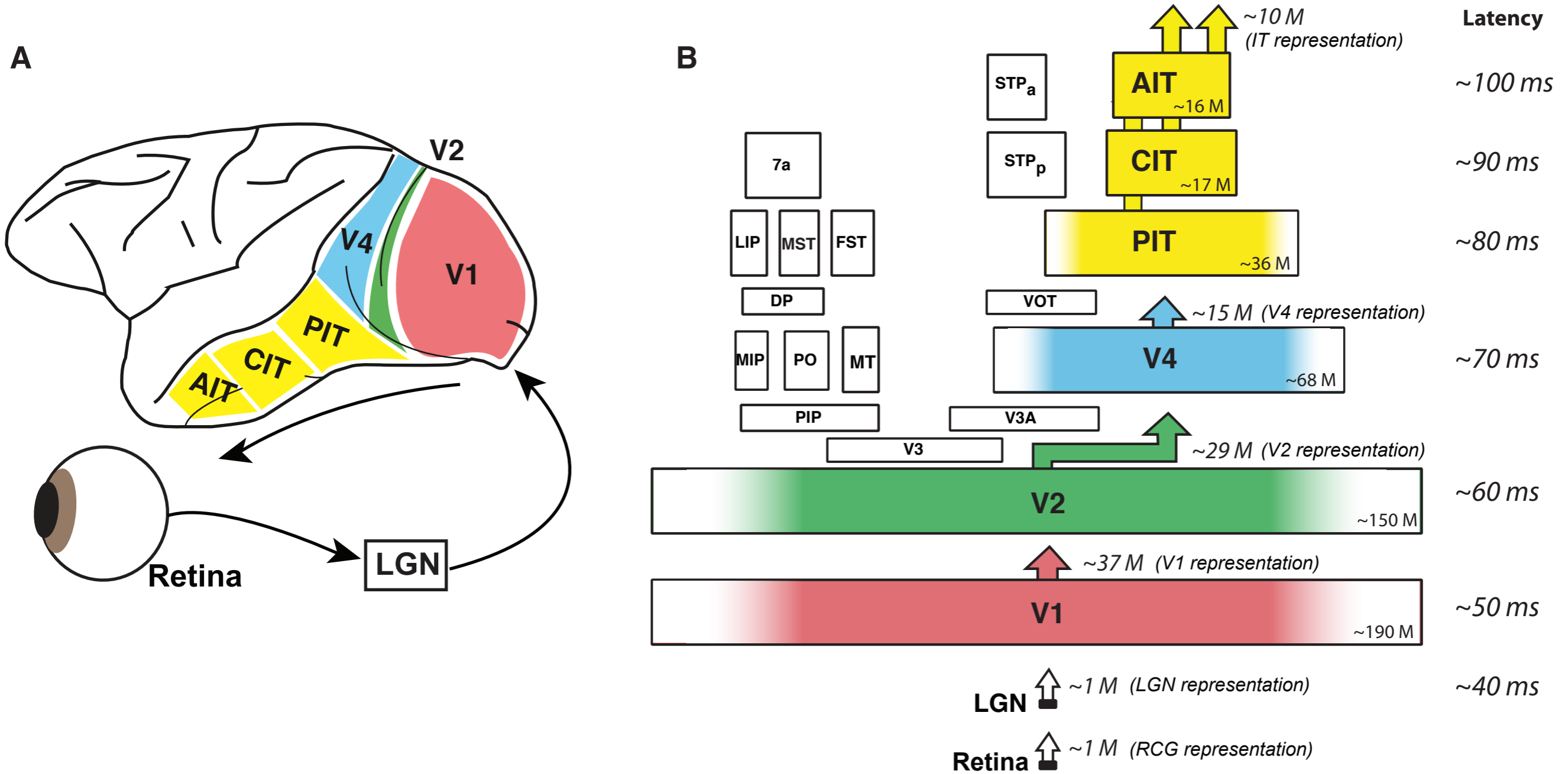
*Understanding Brains  
& Cognition*



*Better AI/ML  
techniques*

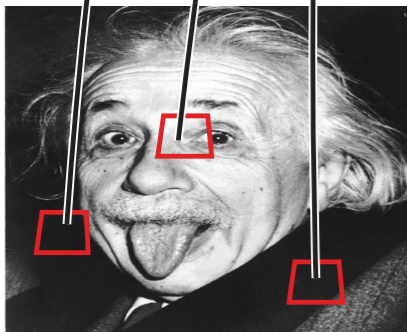
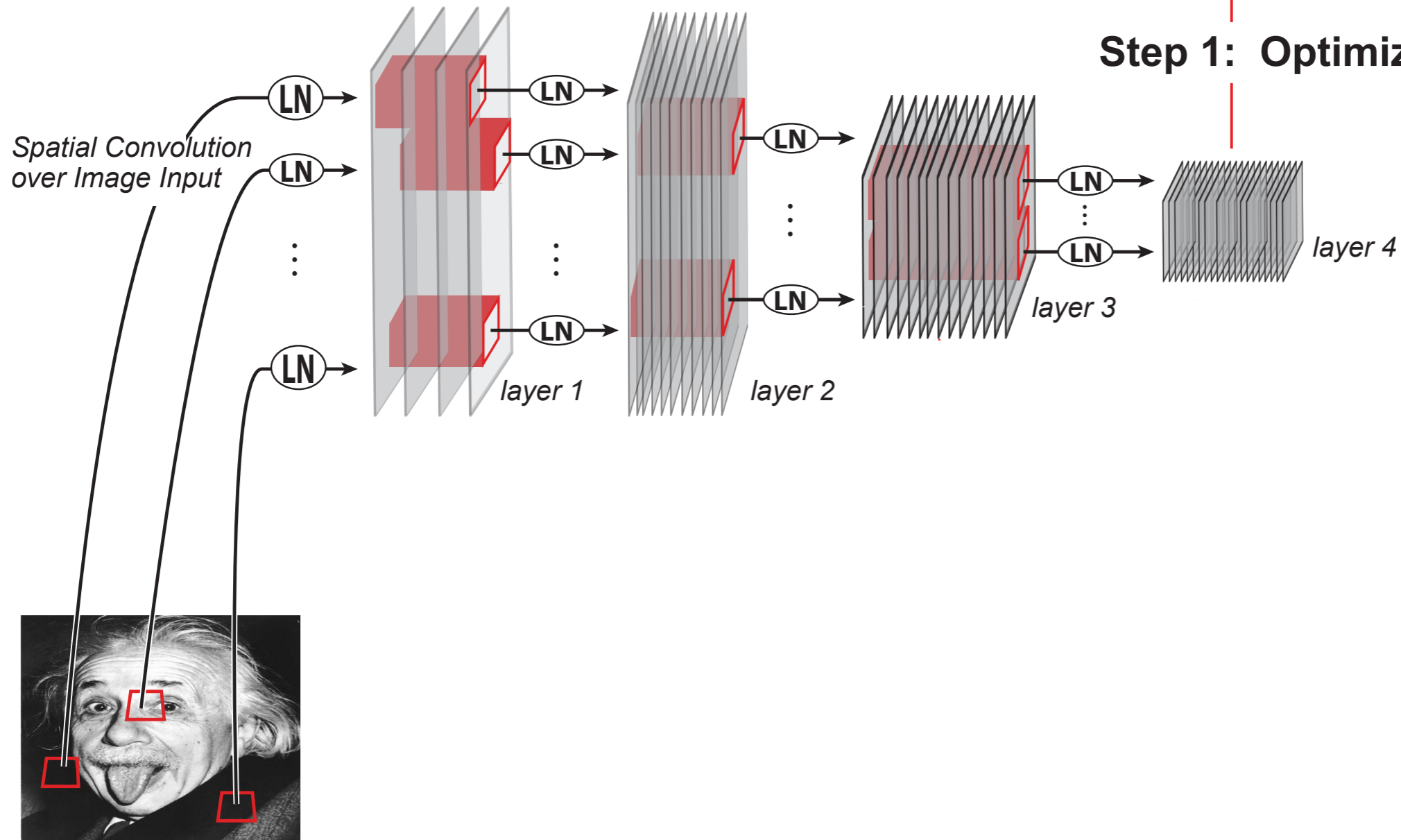
# Computational Models of the Visual System

The primate visual system as a hierarchical, convolutional neural network:



# Computational Models of the Visual System

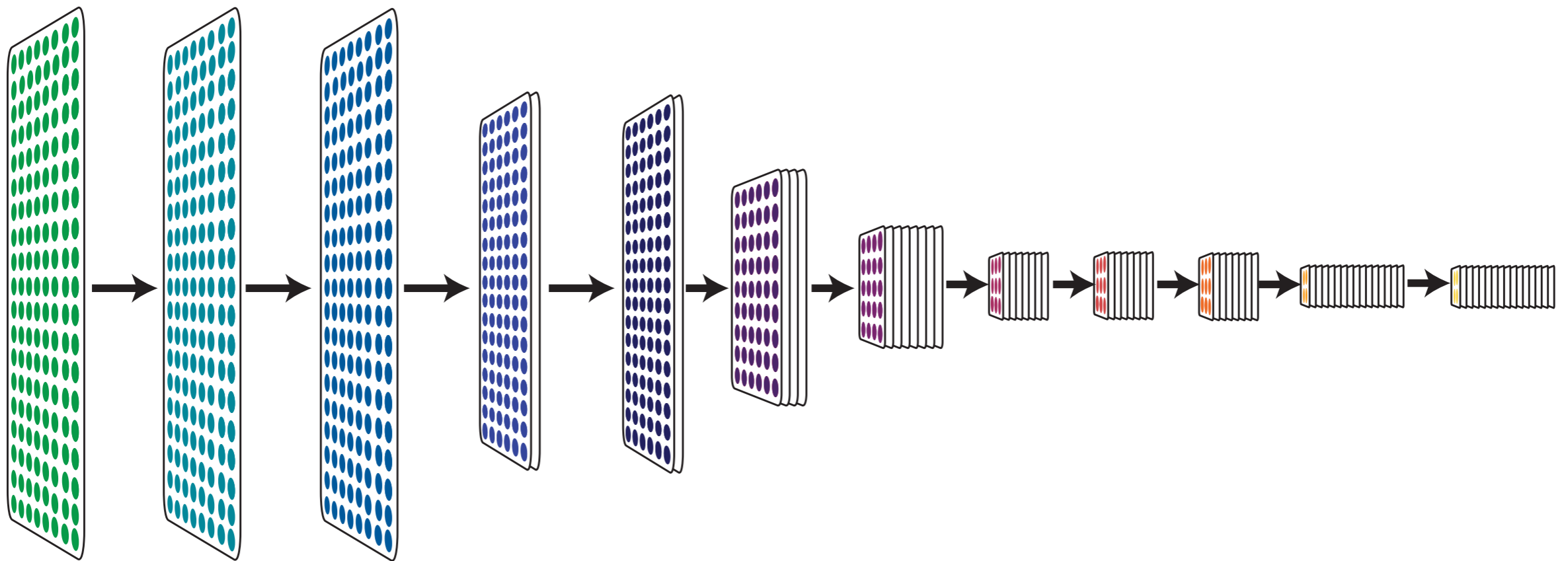
## Visual Recognition Task





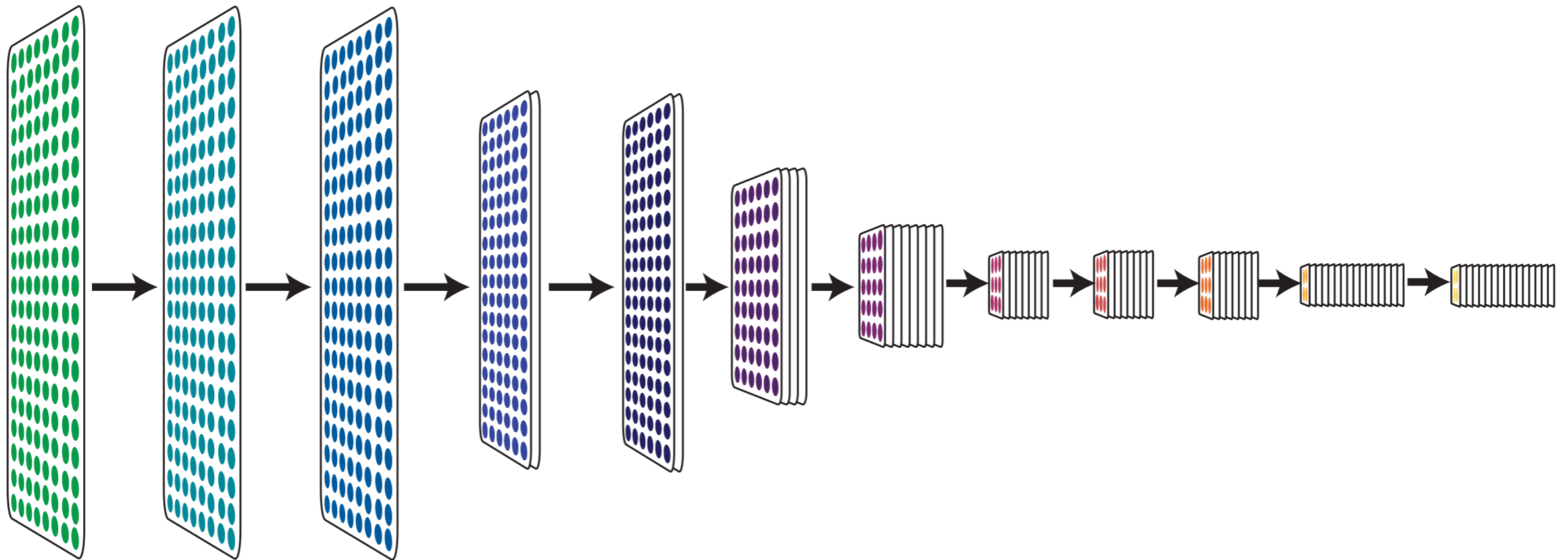
# Computational Models of the Visual System

To our knowledge best (in terms of neural prediction) feedforward model is a  $\sim 12$ -layer CNN



# Computational Models of the Visual System

To our knowledge best (in terms of neural prediction) feedforward model is a  $\sim 12$ -layer CNN



... trained on ImageNet Categorization.

# The Problem

There's just no way that these creatures receive millions of high-level semantic labels during learning.

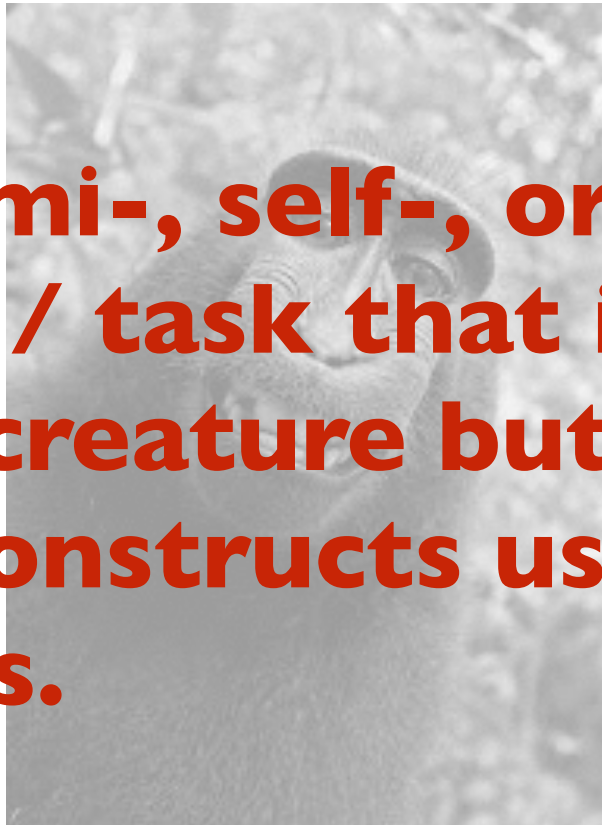



ImageNet is a pretty effective proxy, but just obviously deeply wrong.



# The Problem

There's just no way that these creatures receive millions of high-level semantic labels during learning.

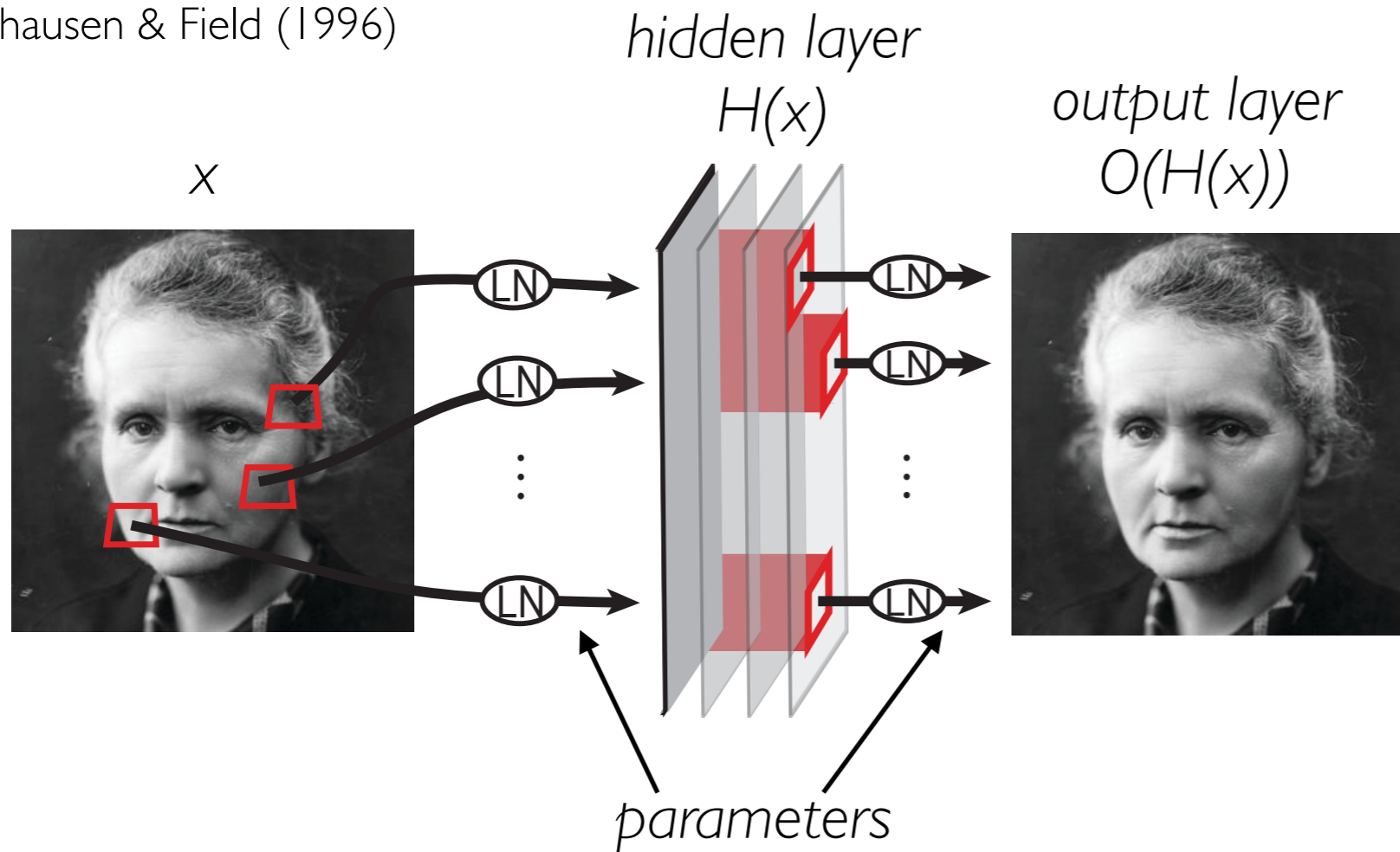


**Must find some sort of semi-, self-, or unsupervised loss function / task that is “realistically costly” to the creature but is sufficiently powerful that it constructs useful representations.**

ImageNet is a pretty effective proxy, but just obviously deeply wrong.

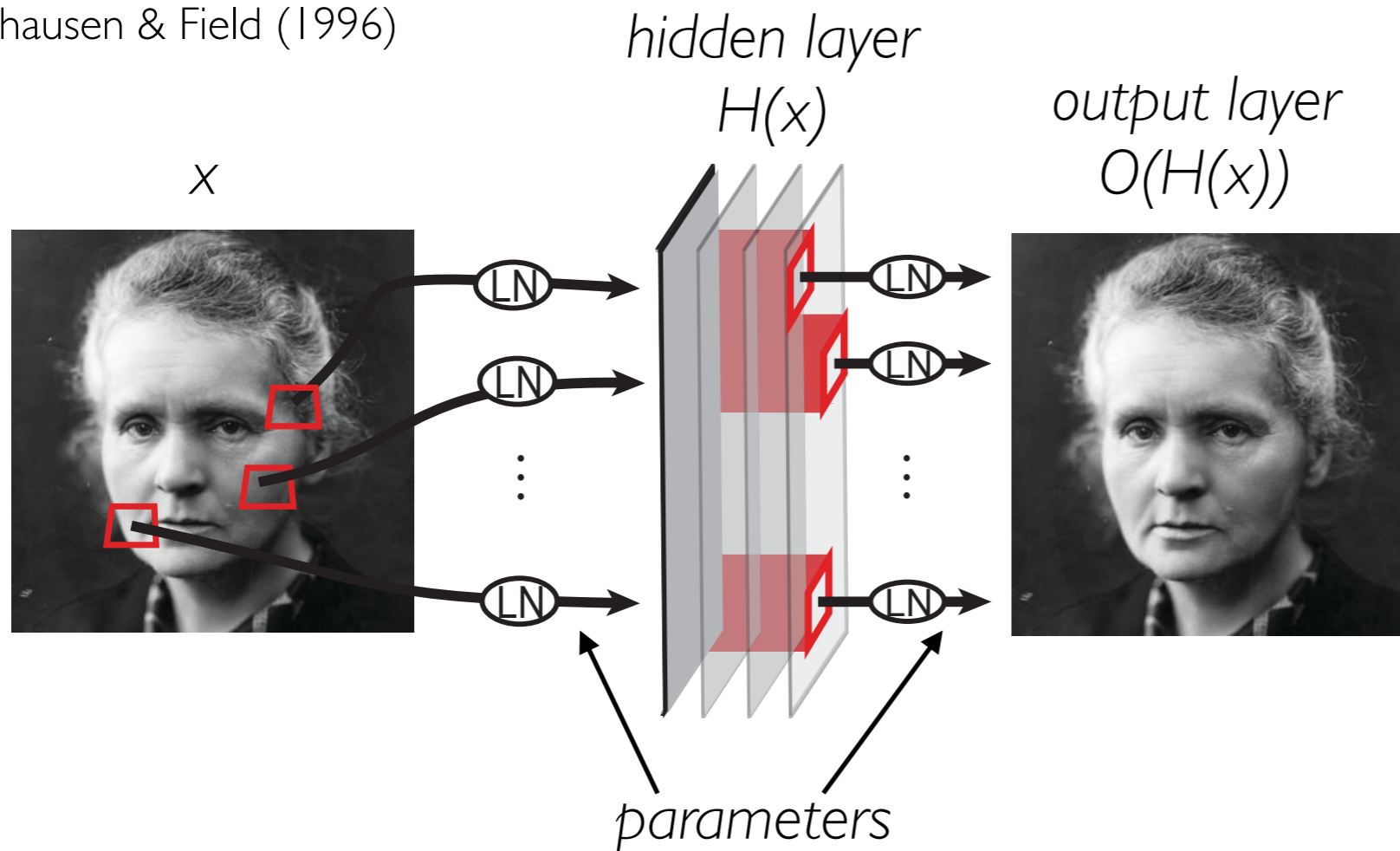
# Self-supervised learning

Olshausen & Field (1996)



# Self-supervised learning

Olshausen & Field (1996)



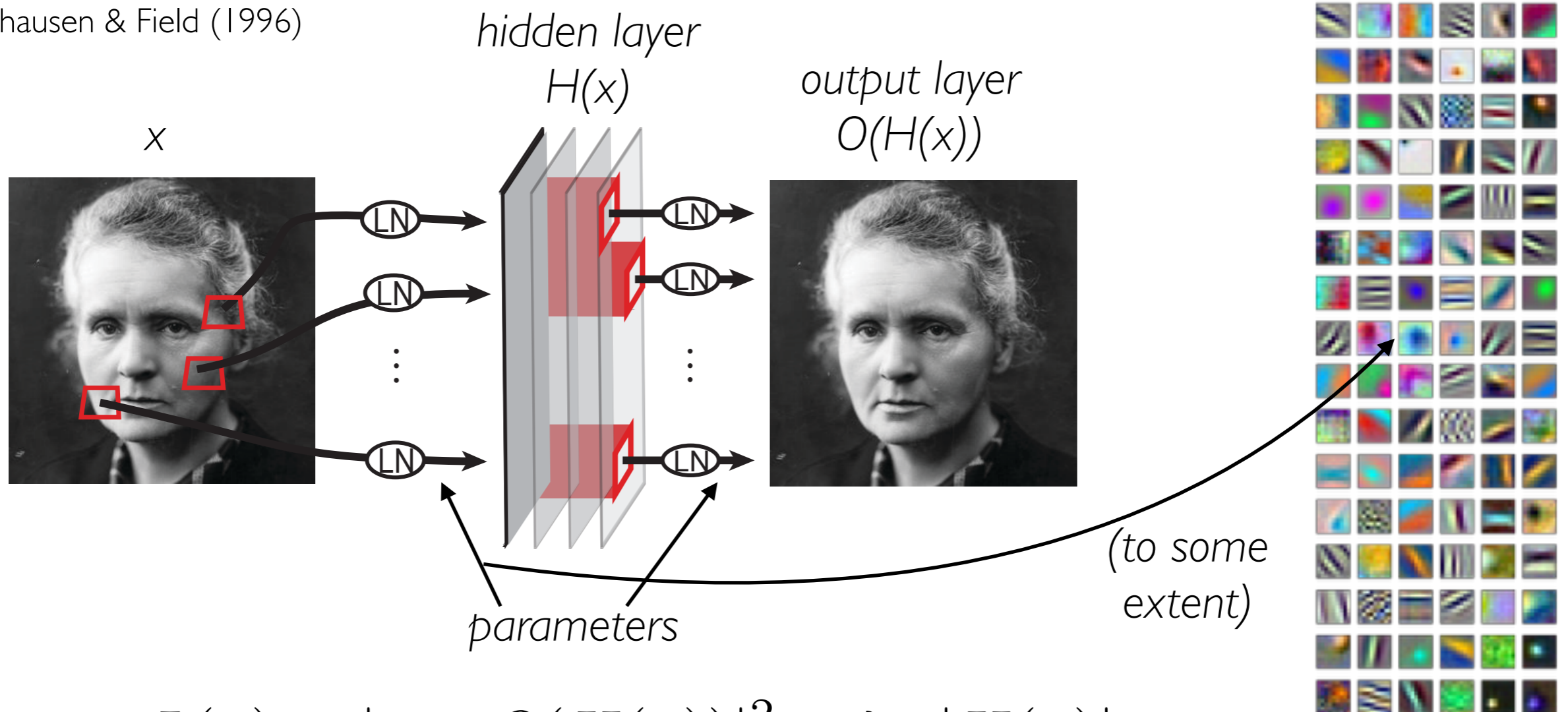
$$L(x) = |x - O(H(x))|^2 + \lambda \cdot |H(x)|$$

reconstruction  
loss

complexity  
penalty

# Self-supervised learning

Olshausen & Field (1996)



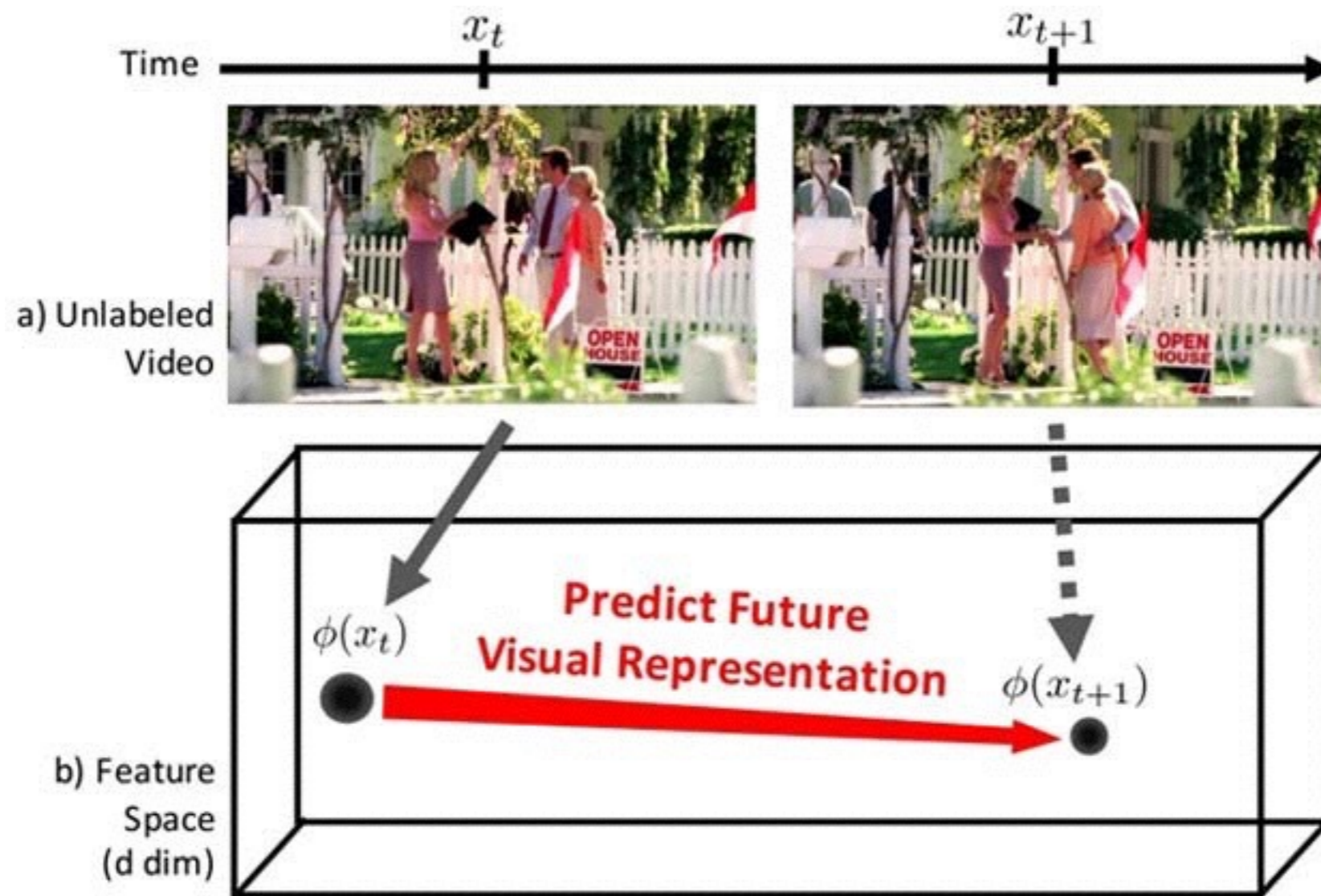
$$L(x) = |x - O(H(x))|^2 + \lambda \cdot |H(x)|$$

reconstruction  
loss

complexity  
penalty

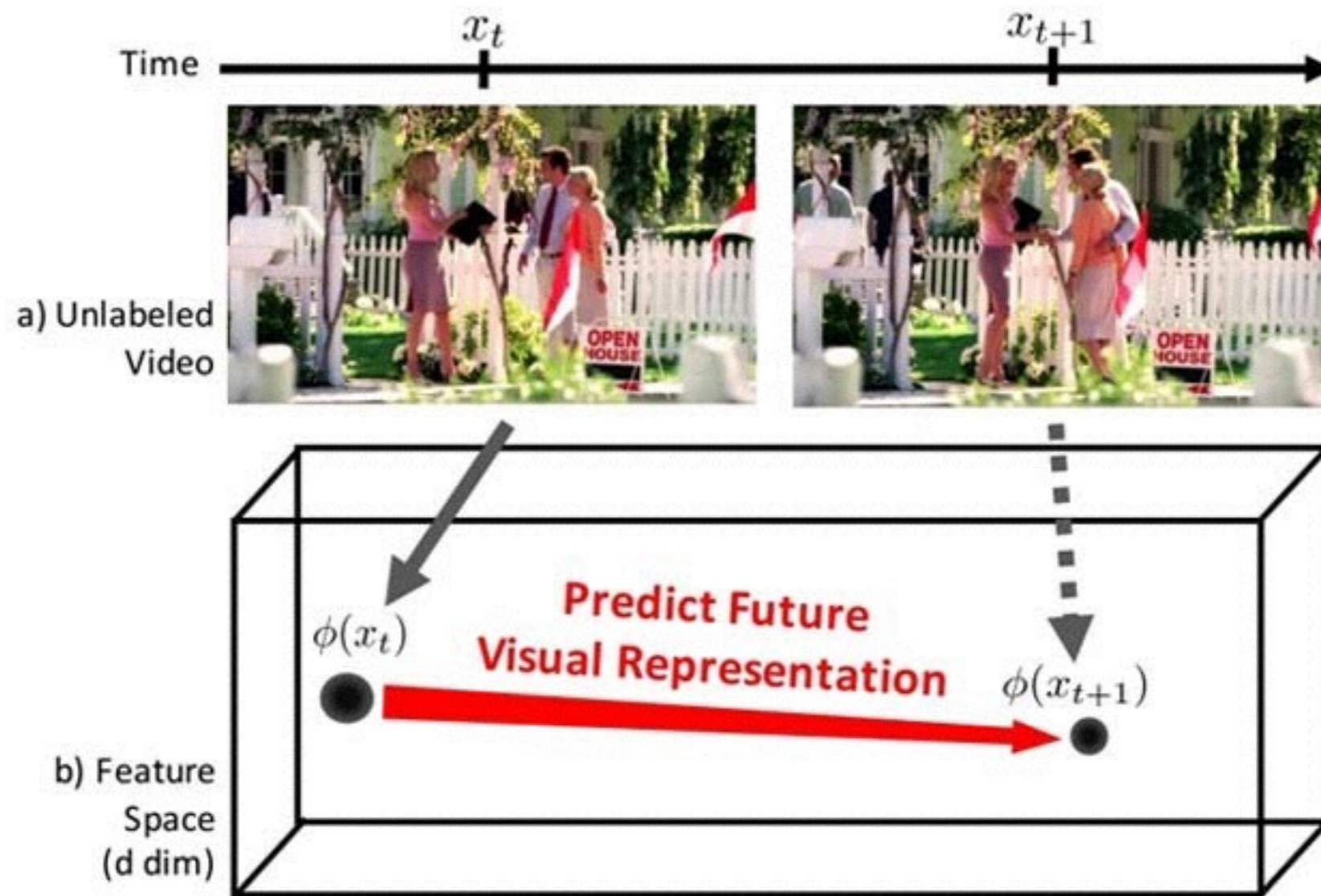
# Self-supervised learning

*Dynamics might give richer signal*



# Self-supervised learning

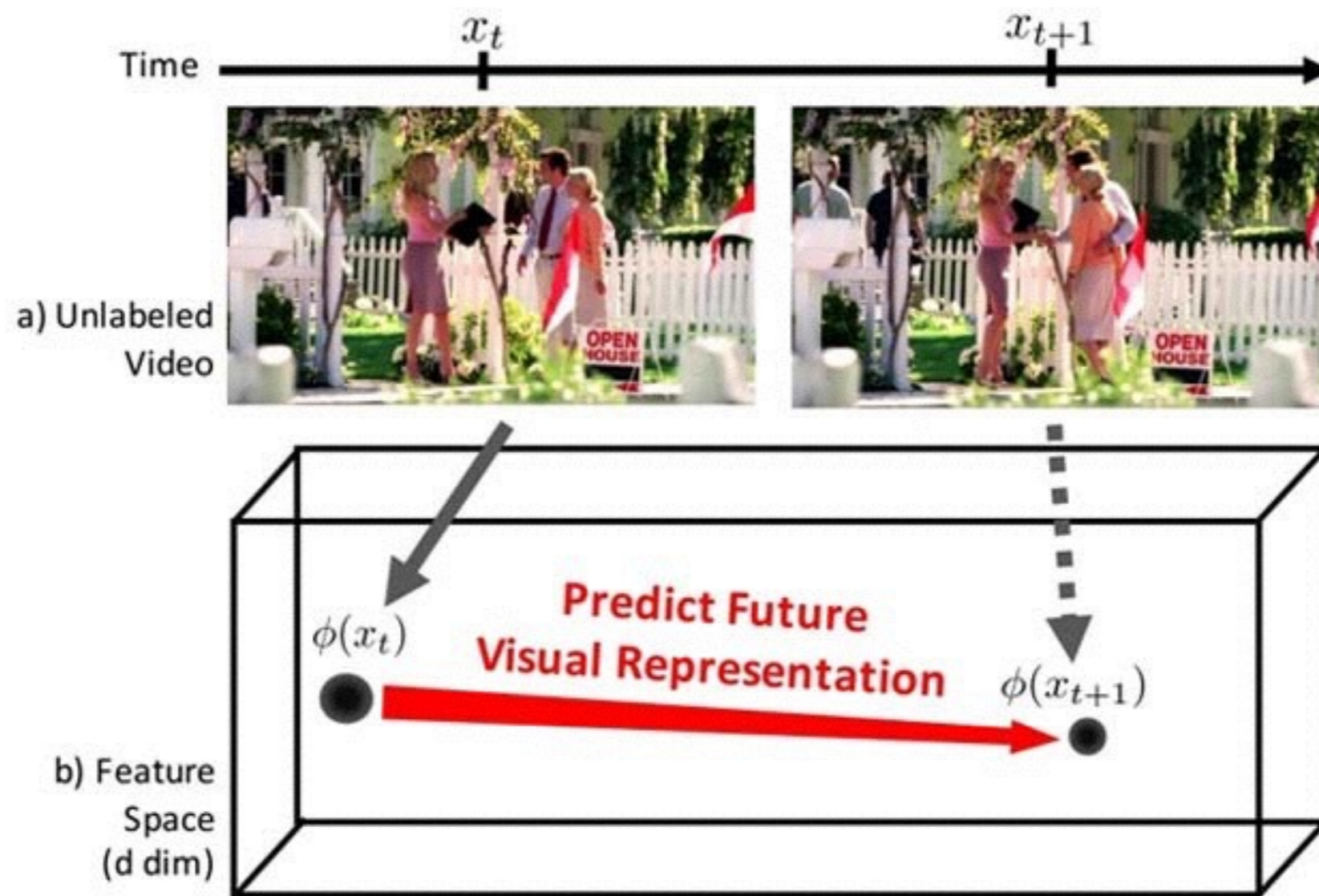
*Dynamics might give richer signal*



$$L(x) = |x_{t+1} - \text{Decode}(\text{Encode}(x_t))|^2 + \lambda \cdot \text{Penalty}(\text{Encode}(x_t))$$

# Self-supervised learning

Dynamics might give richer signal ... but most passive video sequences are **quite boring**



fairly  
trivial  
features

$$L(x) = |x_{t+1} - \text{Decode}(\text{Encode}(x_t))|^2 + \lambda \cdot \text{Penalty}(\text{Encode}(x_t))$$



Children learn through **play**.

How does this work?



# Self-supervised learning

*Give agent some kind of volition to take actions*



$$L(x) = |x_{t+1}^{\text{action}} - \text{Decode}(\text{Encode}(x_t))|^2 + \lambda \cdot \text{Penalty}(\text{Encode}(x_t))$$

# Self-supervised learning

Give agent some kind of volition to take actions ... but now the agent will be **lazy**



$$L(x) = |x_{t+1}^{\text{action}} - \text{Decode}(\text{Encode}(x_t))|^2 + \lambda \cdot \text{Penalty}(\text{Encode}(x_t))$$

# Self-supervised learning

Give agent some kind of volition to take actions ... but now the agent will be **lazy**



$$L(x) = |x^{\text{action}}_{t+1} - \text{Decode}(\text{Encode}(x_t))|^2 + \lambda \cdot \text{Penalty}(\text{Encode}(x_t)) \\ + \text{Intrinsic Motivation}$$

*Environment*



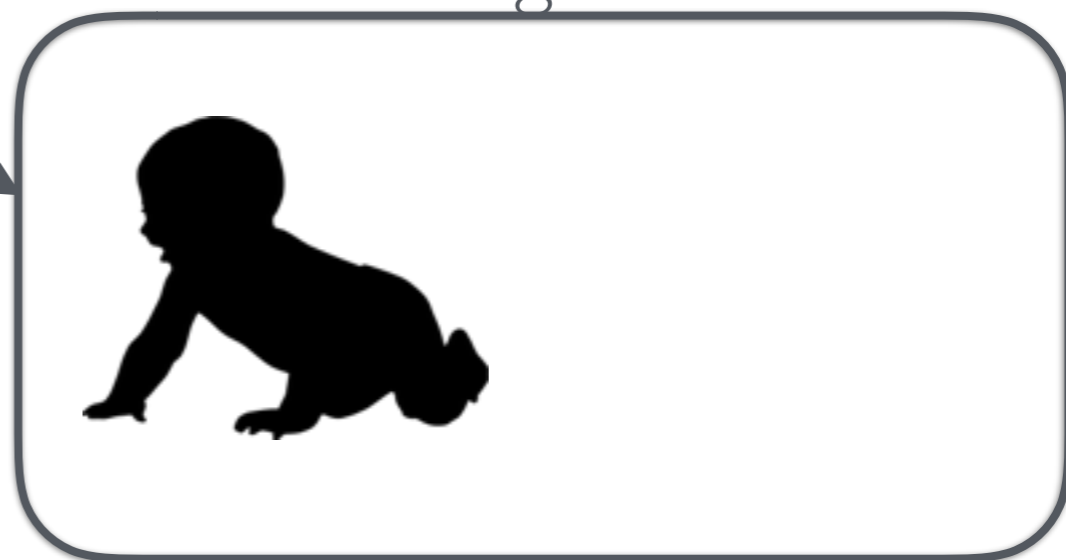
*Environment*



*Perception*



*Agent*



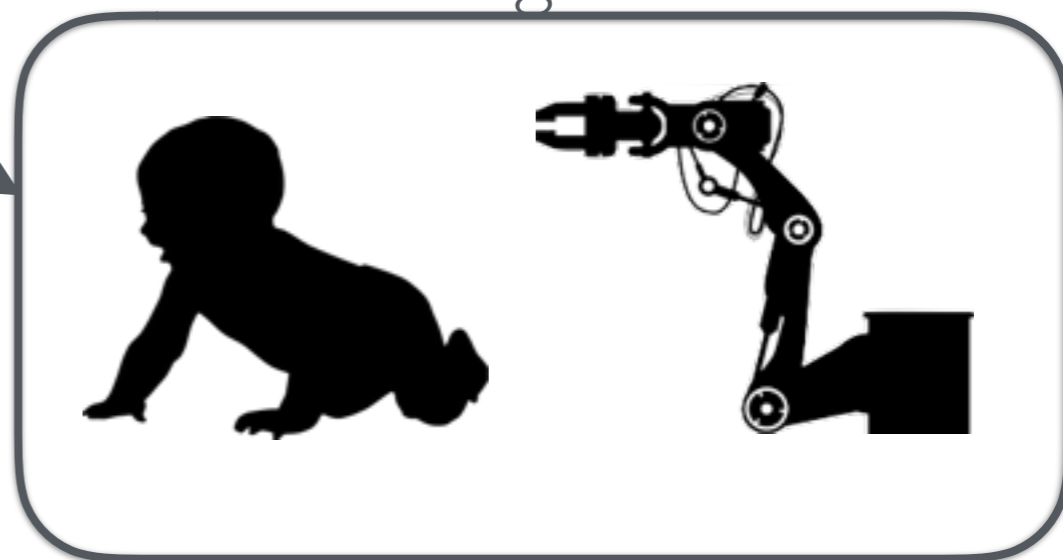
*Environment*

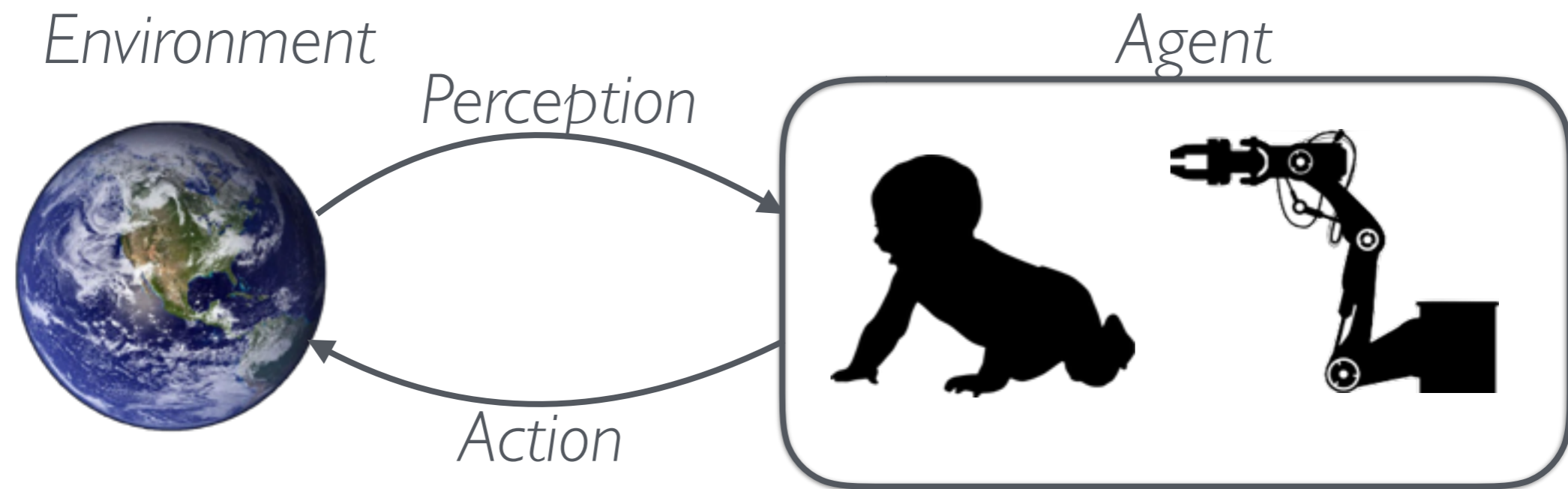


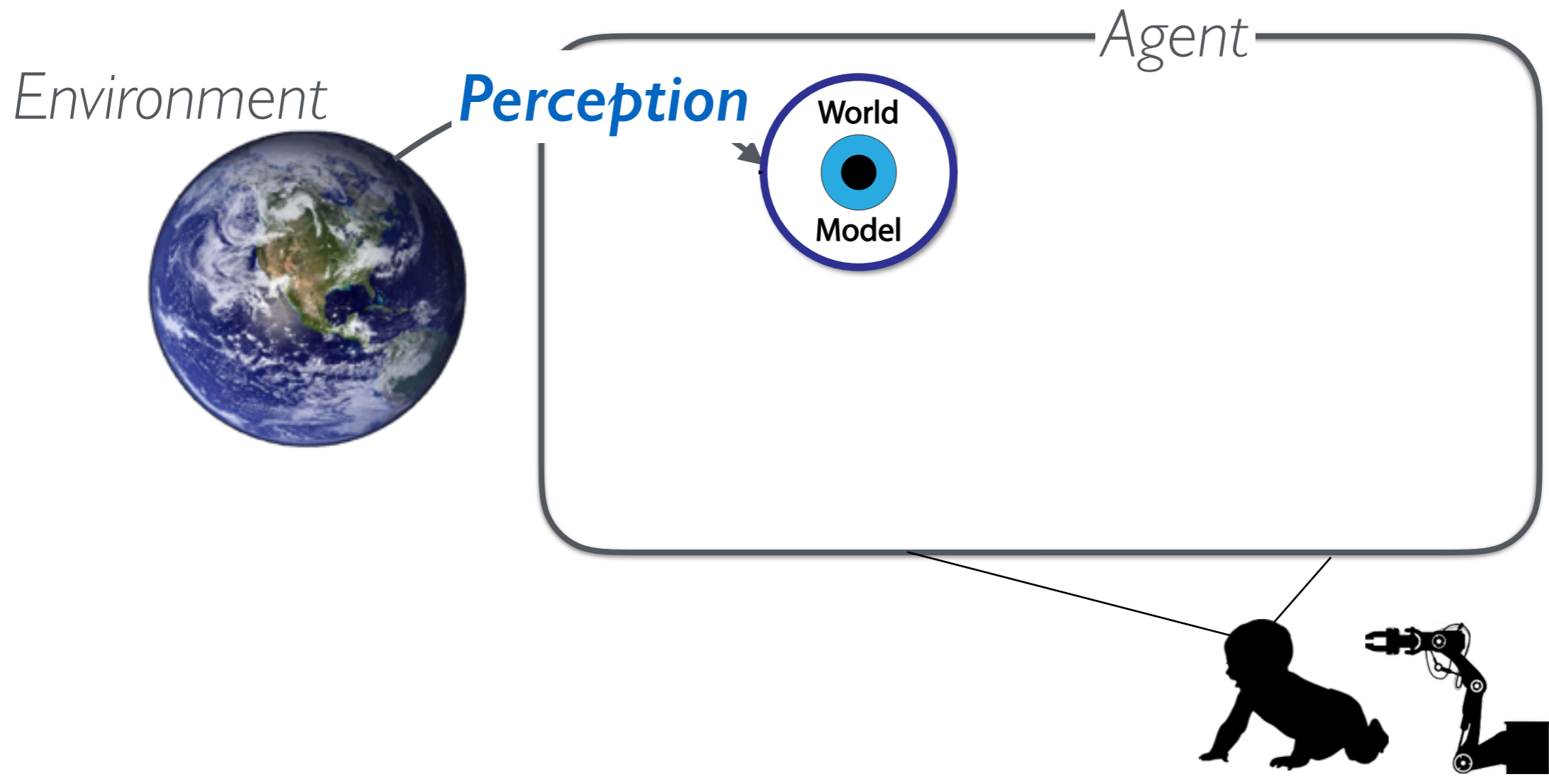
*Perception*



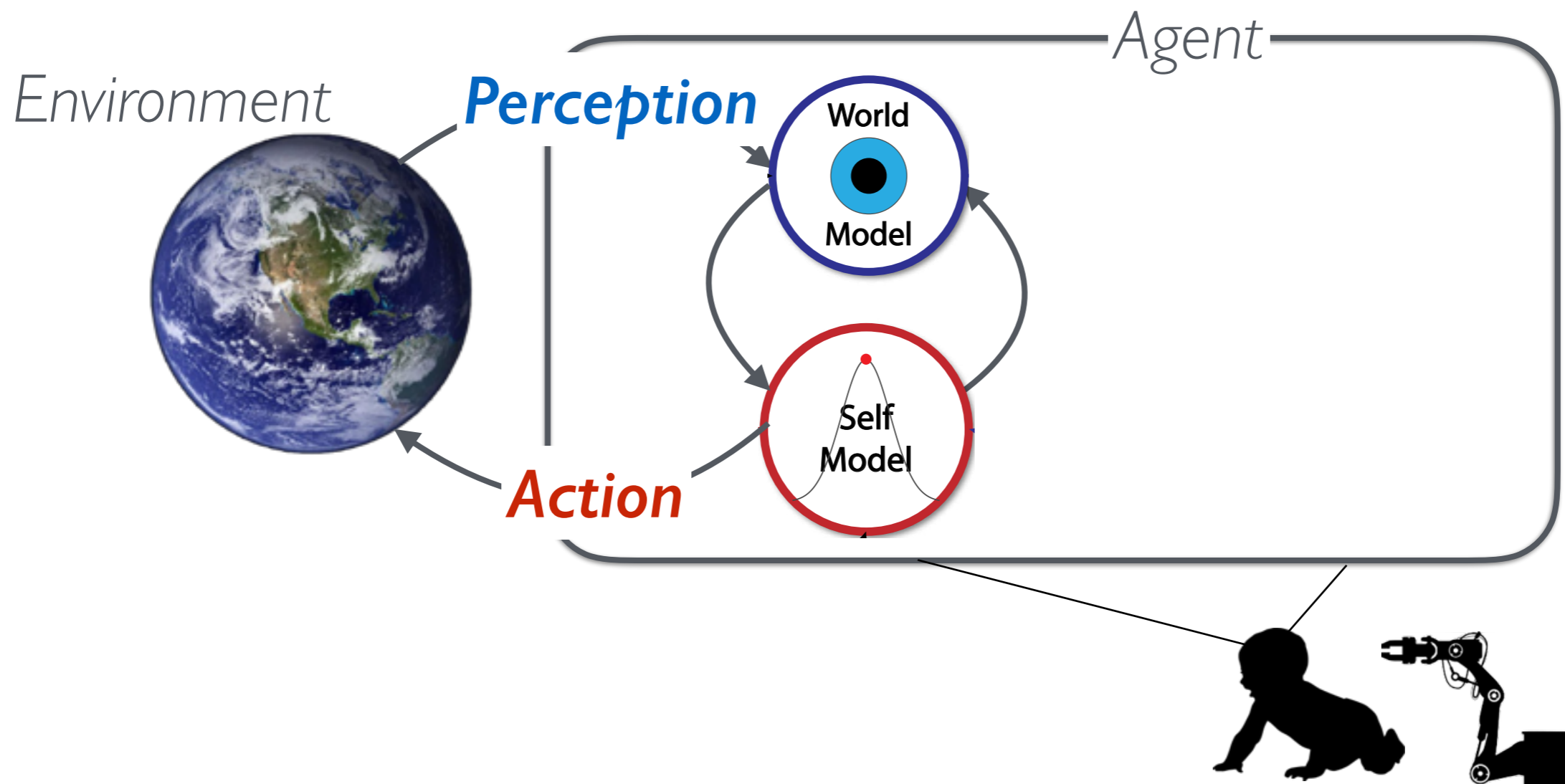
*Agent*

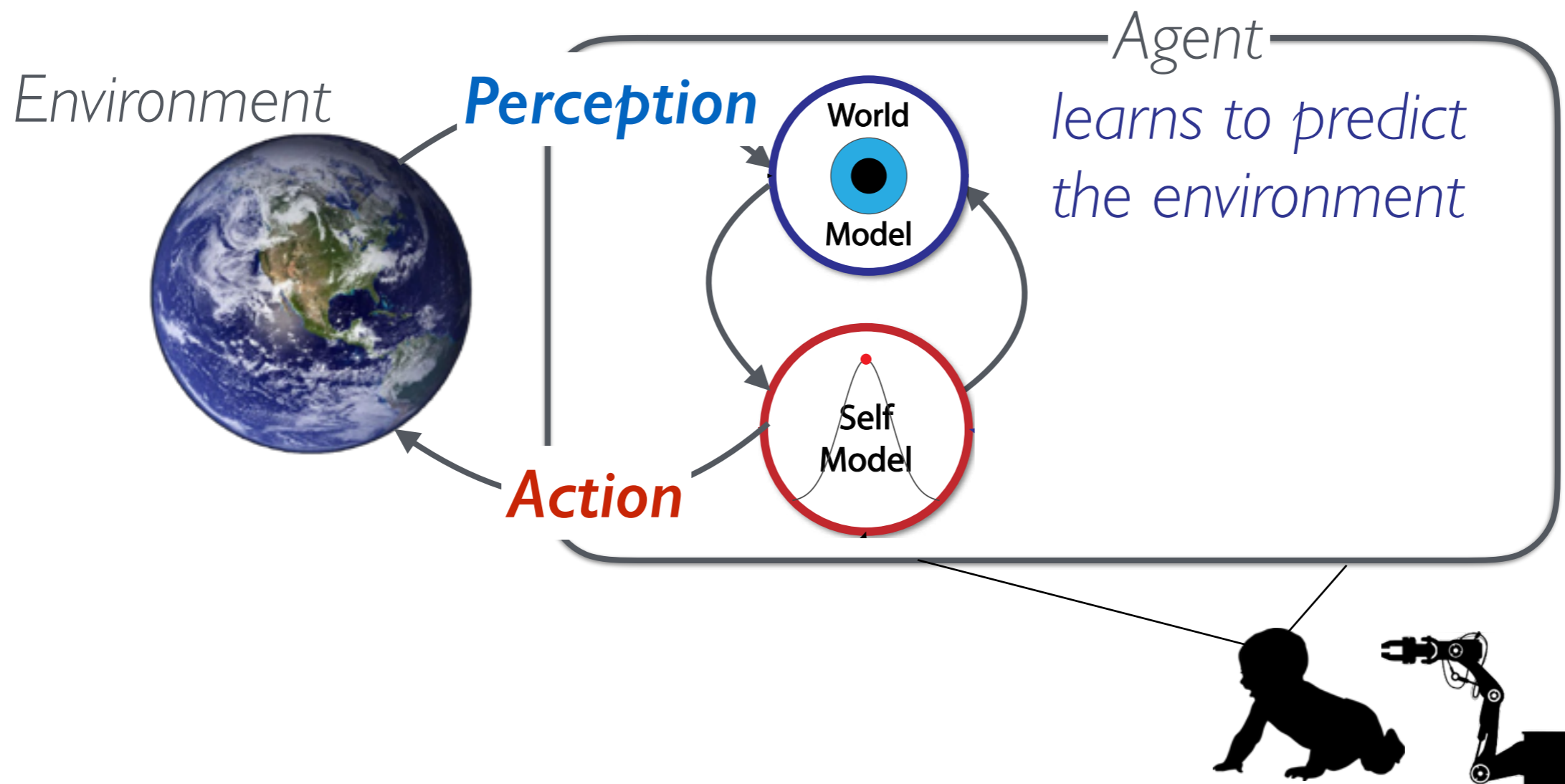


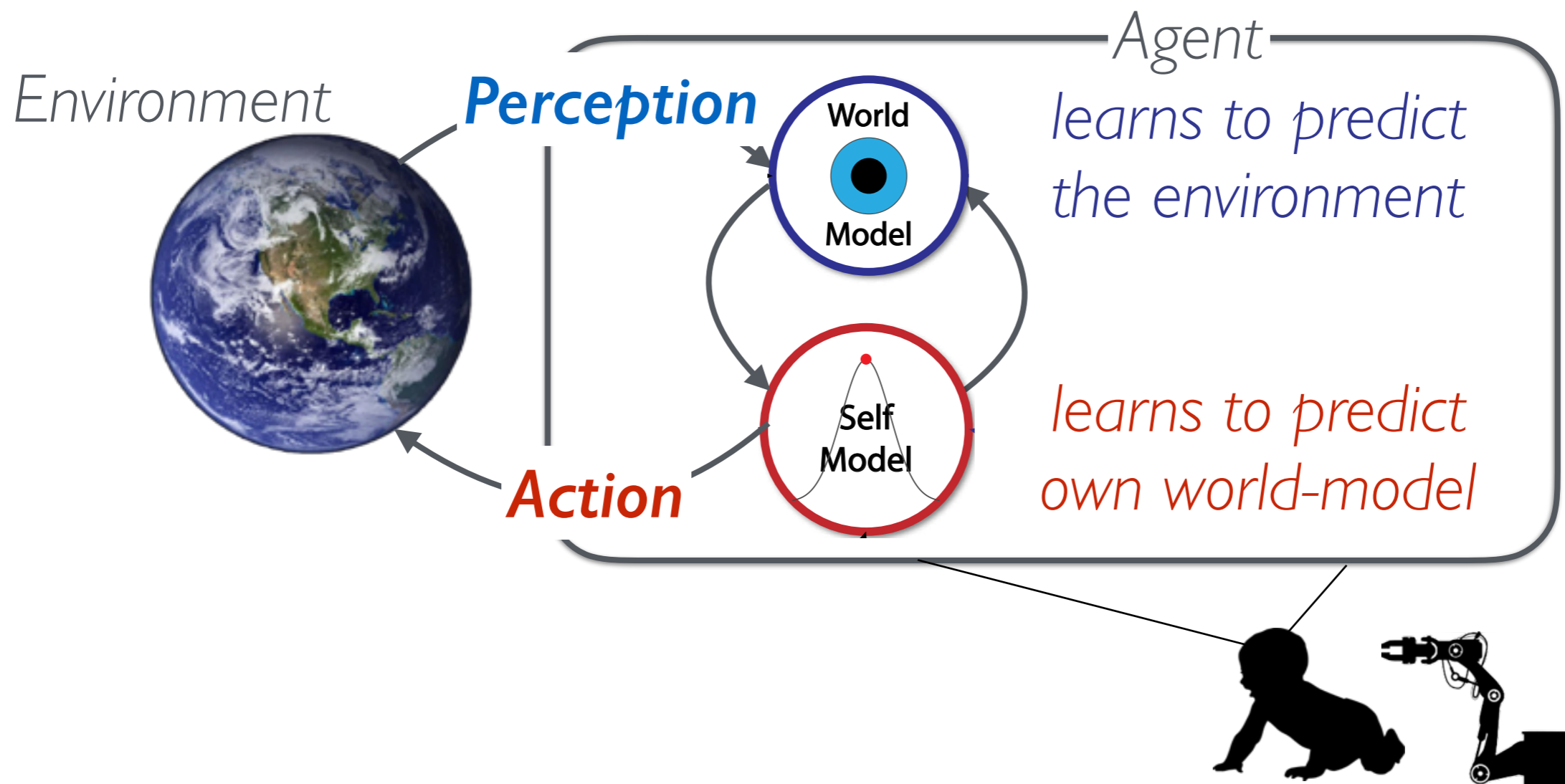






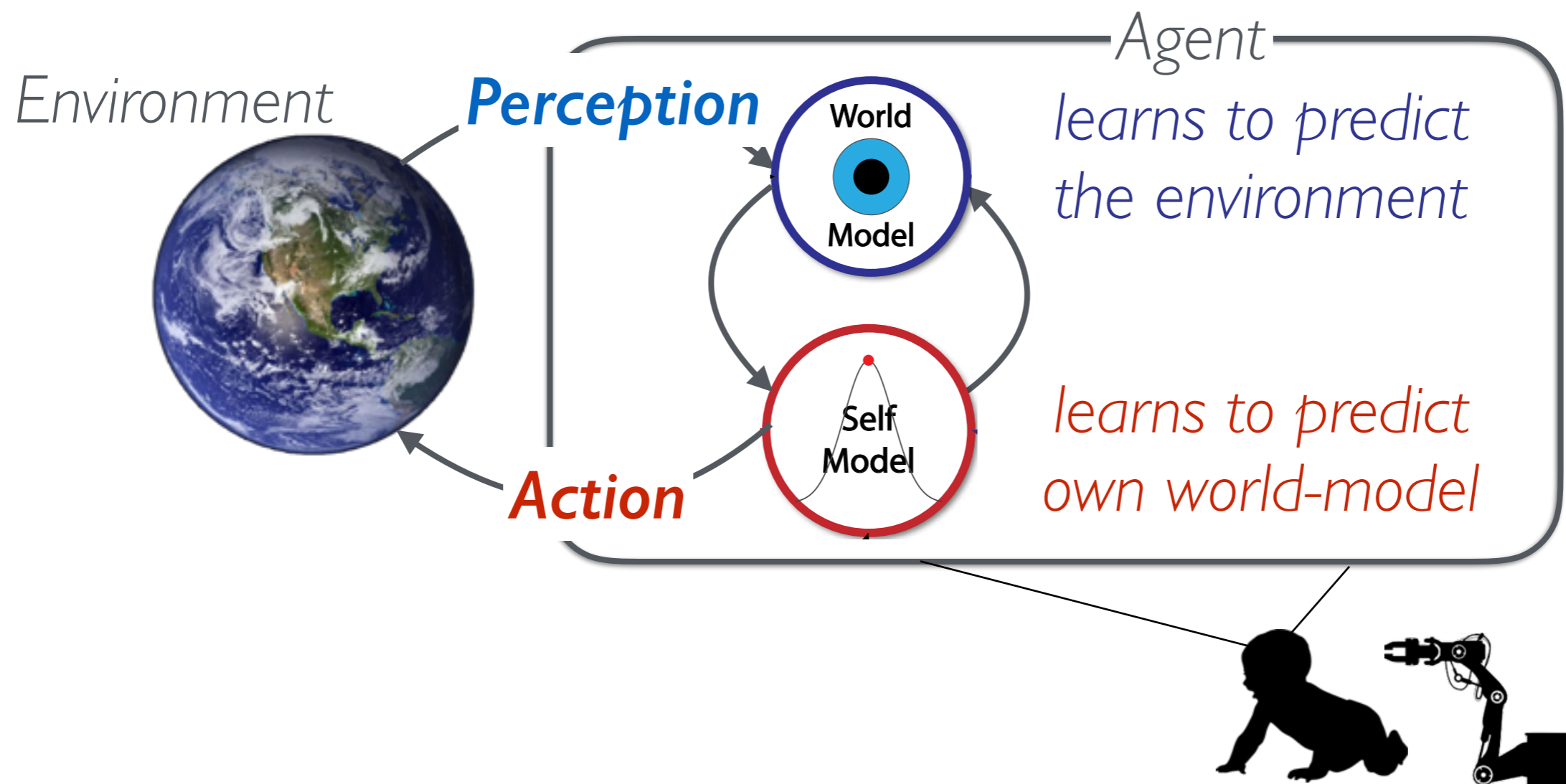






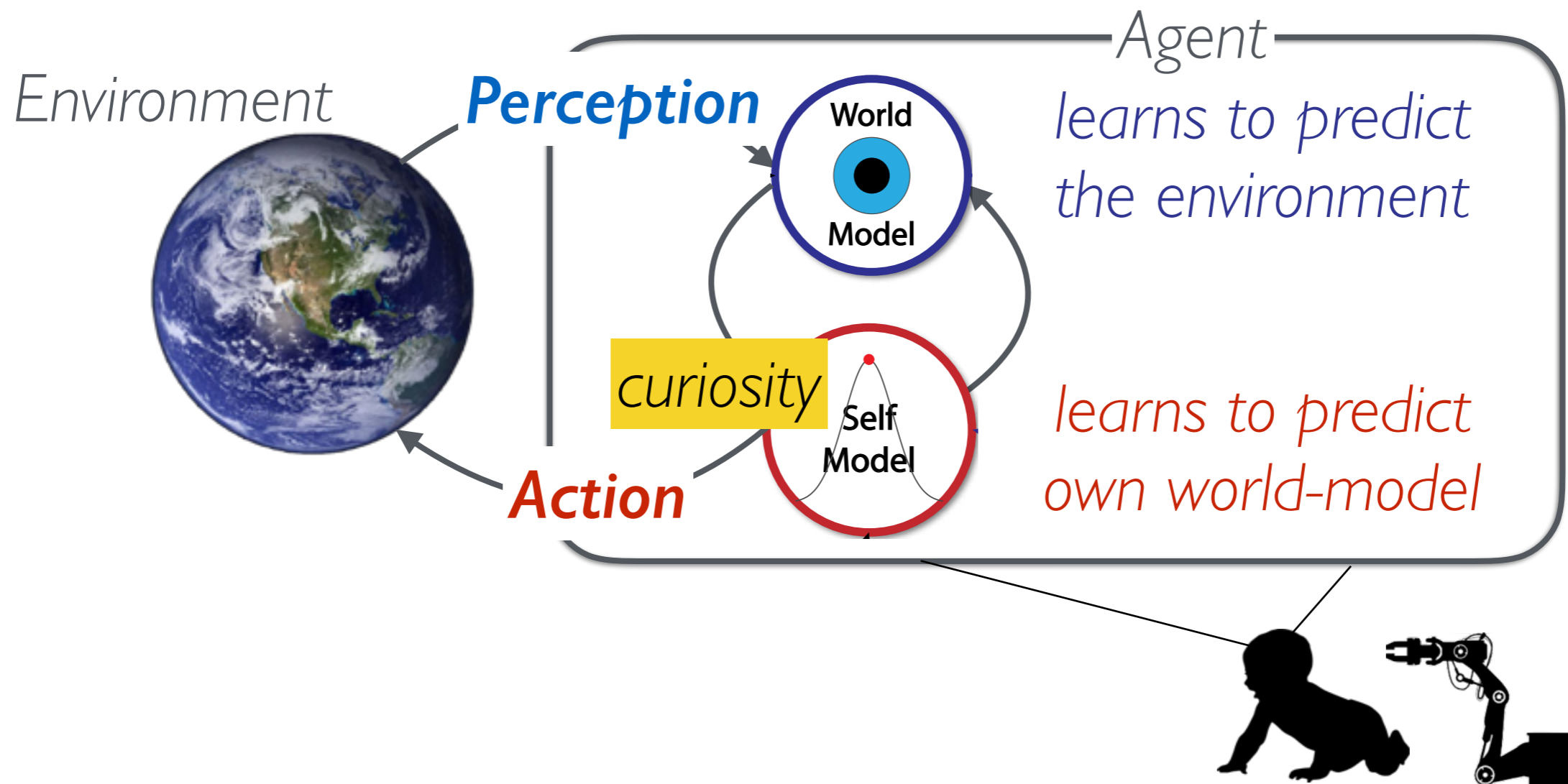
## A curiosity principle:

The **self-model** directs the agent toward **interesting** actions — the ones that the **world-model** doesn't yet fully understand



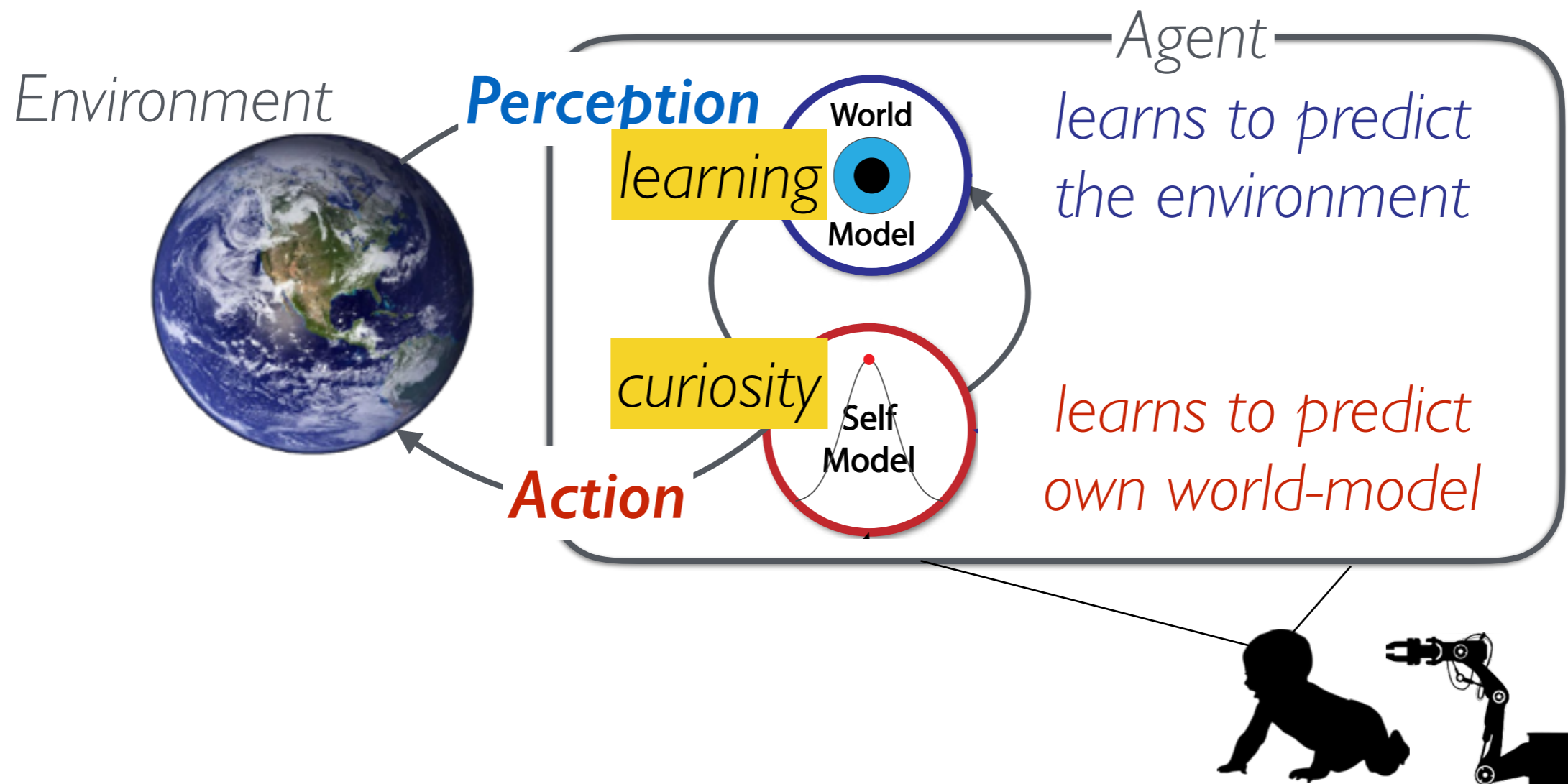
## A curiosity principle:

The **self-model** directs the agent toward **interesting** actions — the ones that the **world-model** doesn't yet fully understand



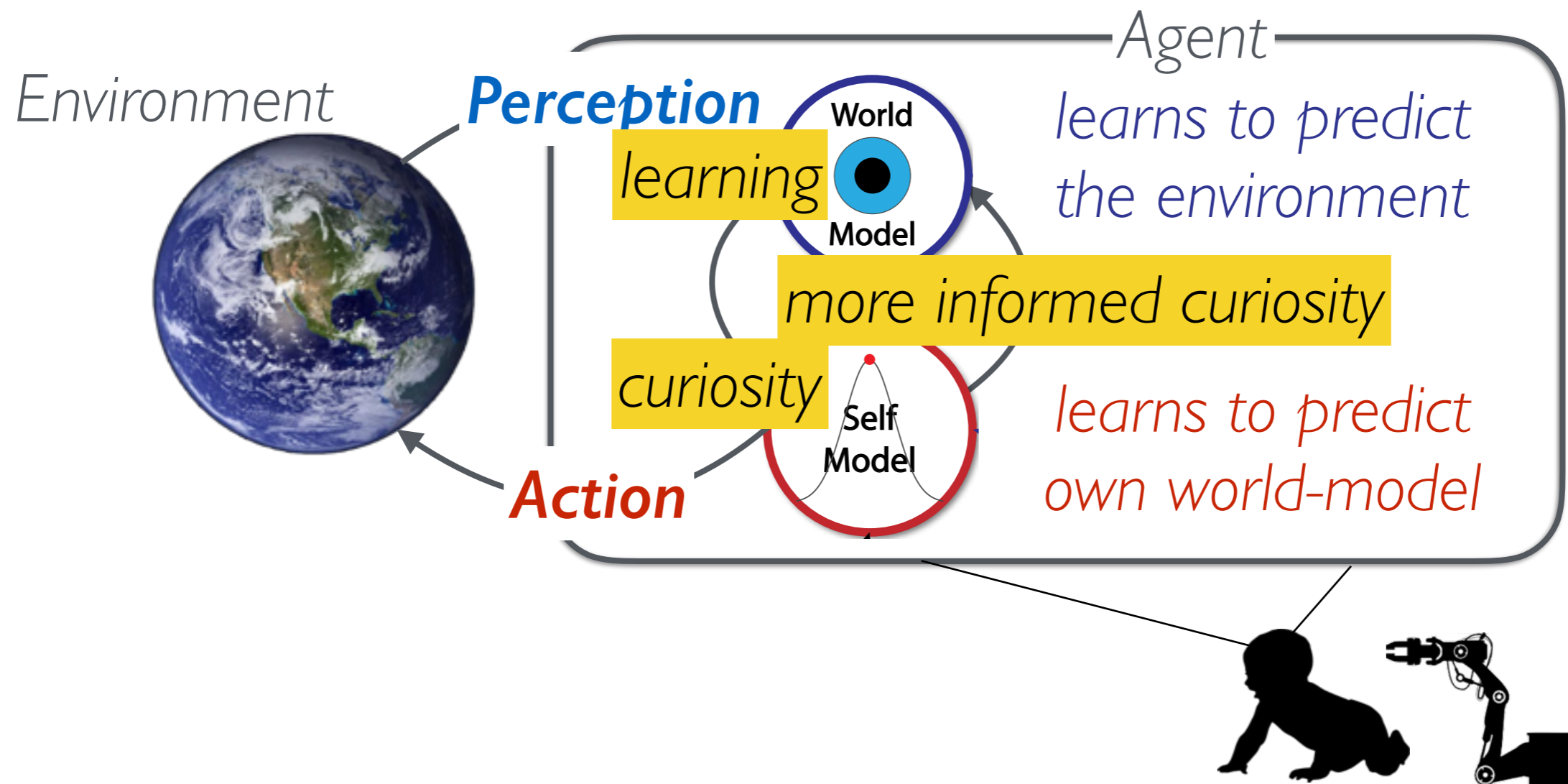
# A curiosity principle:

The **self-model** directs the agent toward **interesting** actions — the ones that the **world-model** doesn't yet fully understand



# A curiosity principle:

The **self-model** directs the agent toward **interesting** actions — the ones that the **world-model** doesn't yet fully understand



# Learning to Play



Nick Haber



Damian Mrowca



Stephanie Wang



Fei-Fei Li

NIPS 2018

---

## Learning to Play With Intrinsically-Motivated, Self-Aware Agents

---

Nick Haber<sup>1,2,3,\*</sup>, Damian Mrowca<sup>4,\*</sup>, Stephanie Wang<sup>4</sup>, Li Fei-Fei<sup>4</sup>, and  
Daniel L. K. Yamins<sup>1,4,5</sup>

Departments of Psychology<sup>1</sup>, Pediatrics<sup>2</sup>, Biomedical Data Science<sup>3</sup>, Computer Science<sup>4</sup>, and Wu  
Tsai Neurosciences Institute<sup>5</sup>, Stanford, CA 94305

{nhaber, mrowca}@stanford.edu

### Abstract

Infants are experts at playing, with an amazing ability to generate novel structured behaviors in unstructured environments that lack clear extrinsic reward signals. We seek to mathematically formalize these abilities using a neural network that implements curiosity-driven intrinsic motivation. Using a simple but ecologically naturalistic simulated environment in which an agent can move and interact with objects it sees, we propose a “world-model” network that learns to predict the dynamic consequences of the agent’s actions. Simultaneously, we train a separate explicit “self-model” that allows the agent to track the error map of its world-model. It then uses the self-model to adversarially challenge the developing world-model. We demonstrate that this policy causes the agent to explore novel and informative interactions with its environment, leading to the generation of a spectrum of complex behaviors, including ego-motion prediction, object attention, and object gathering. Moreover, the world-model that the agent learns supports improved performance on object dynamics prediction, detection, localization and recognition tasks. Taken together, our results are initial steps toward creating flexible autonomous agents that self-supervise in realistic physical environments.





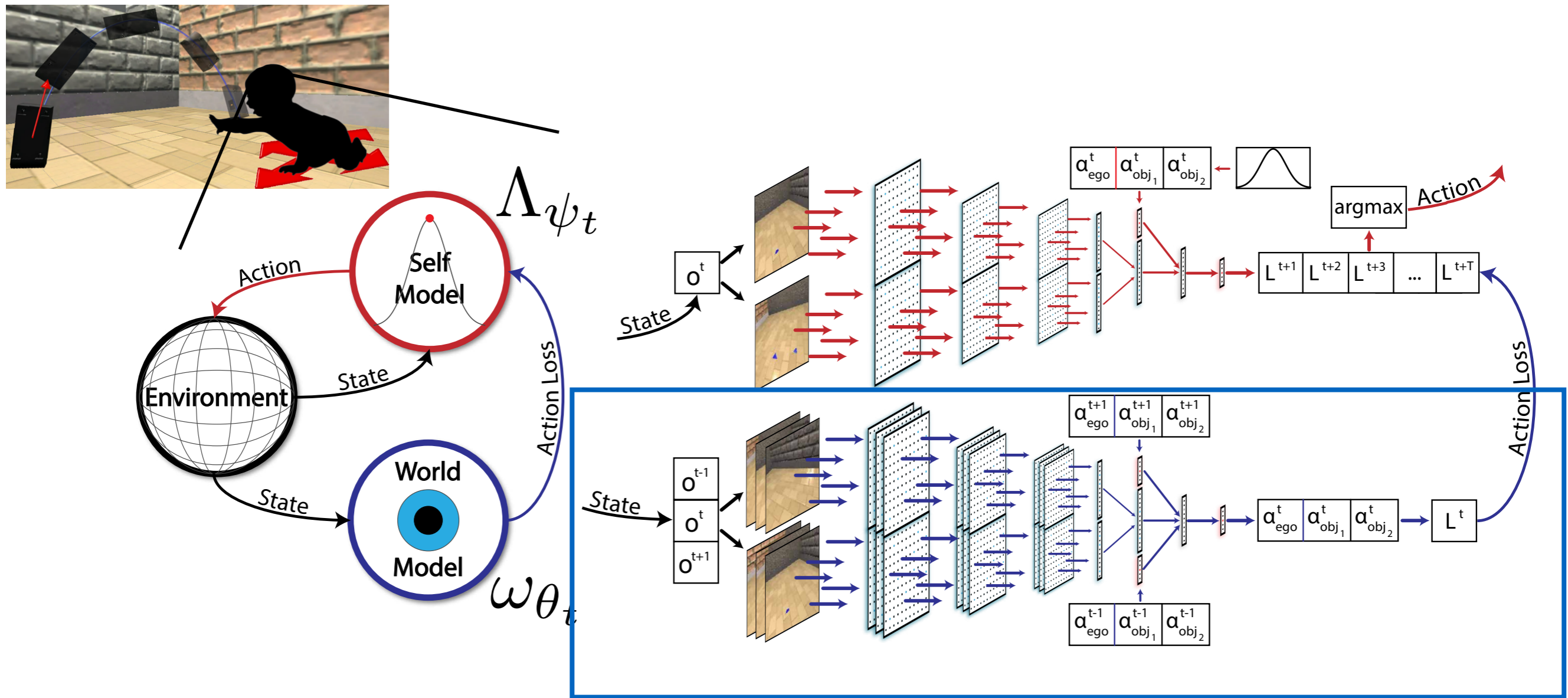
Agent (“baby”) can (a) swivel its head

(b) move around the room

(c) apply forces to objects

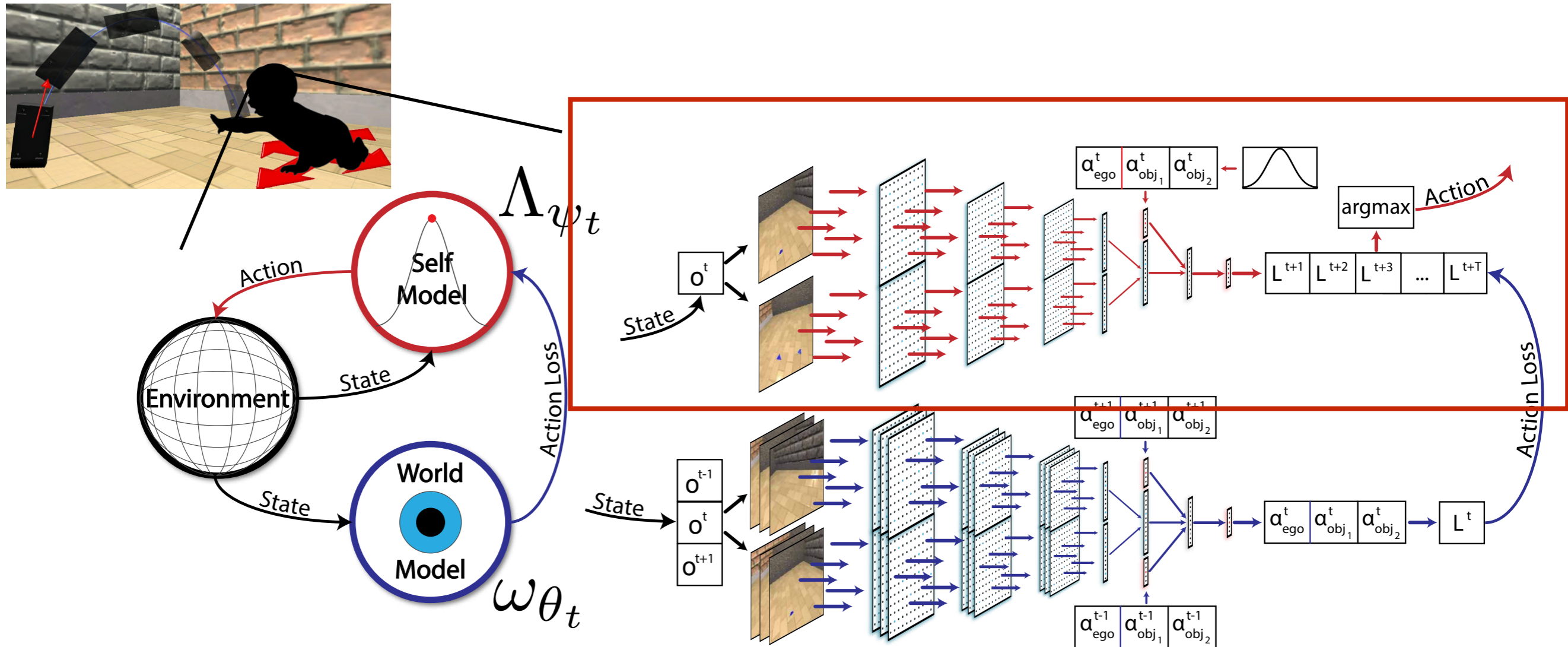
... and receives back images of what happened, given action

# Learning to Play — Overview



Model has two pieces: **(I) World-Model**

# Learning to Play — Overview

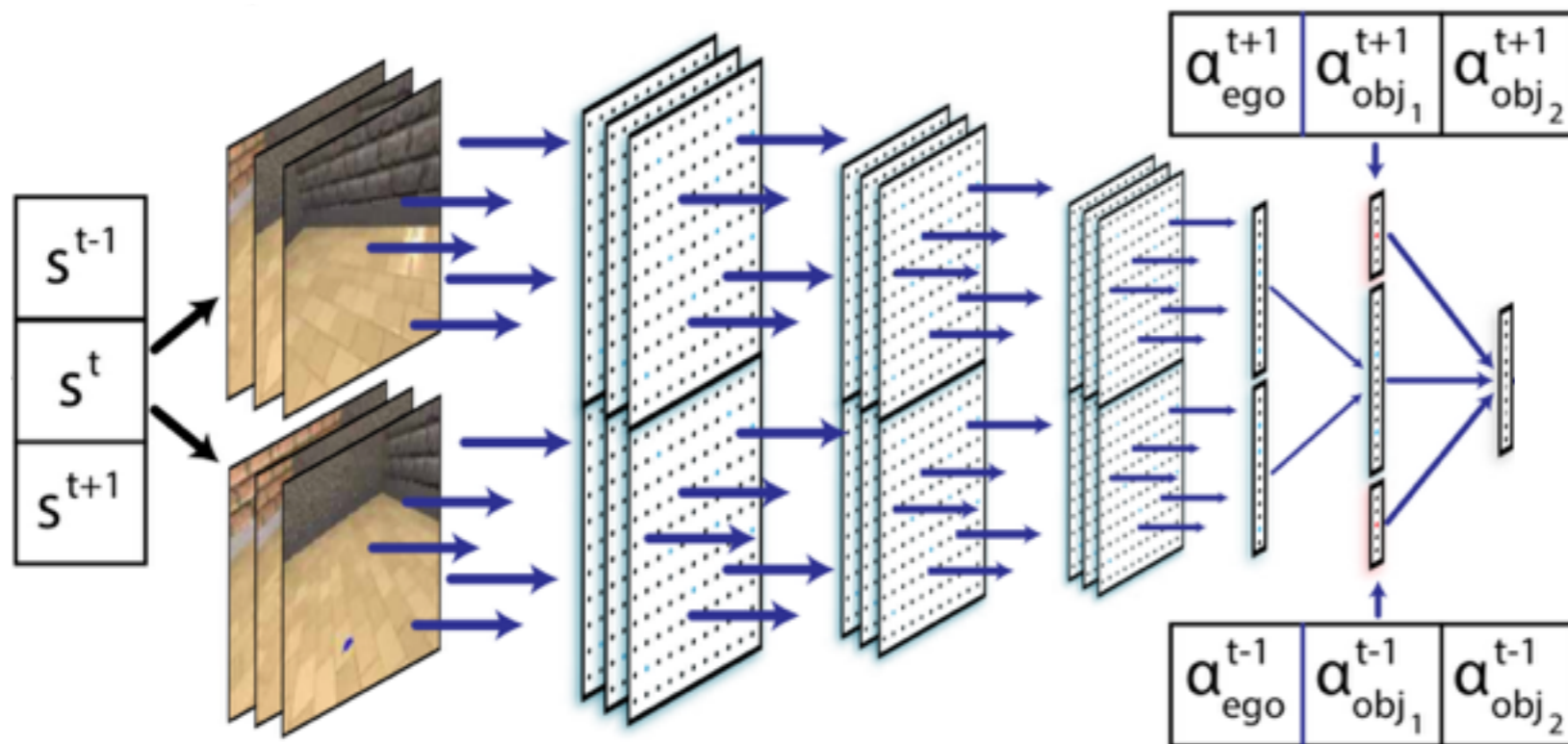


Model has two pieces: **(1) World-Model**  
**(2) Self-Model**

# Learning to Play — World-Model

Goal of **world-model**:

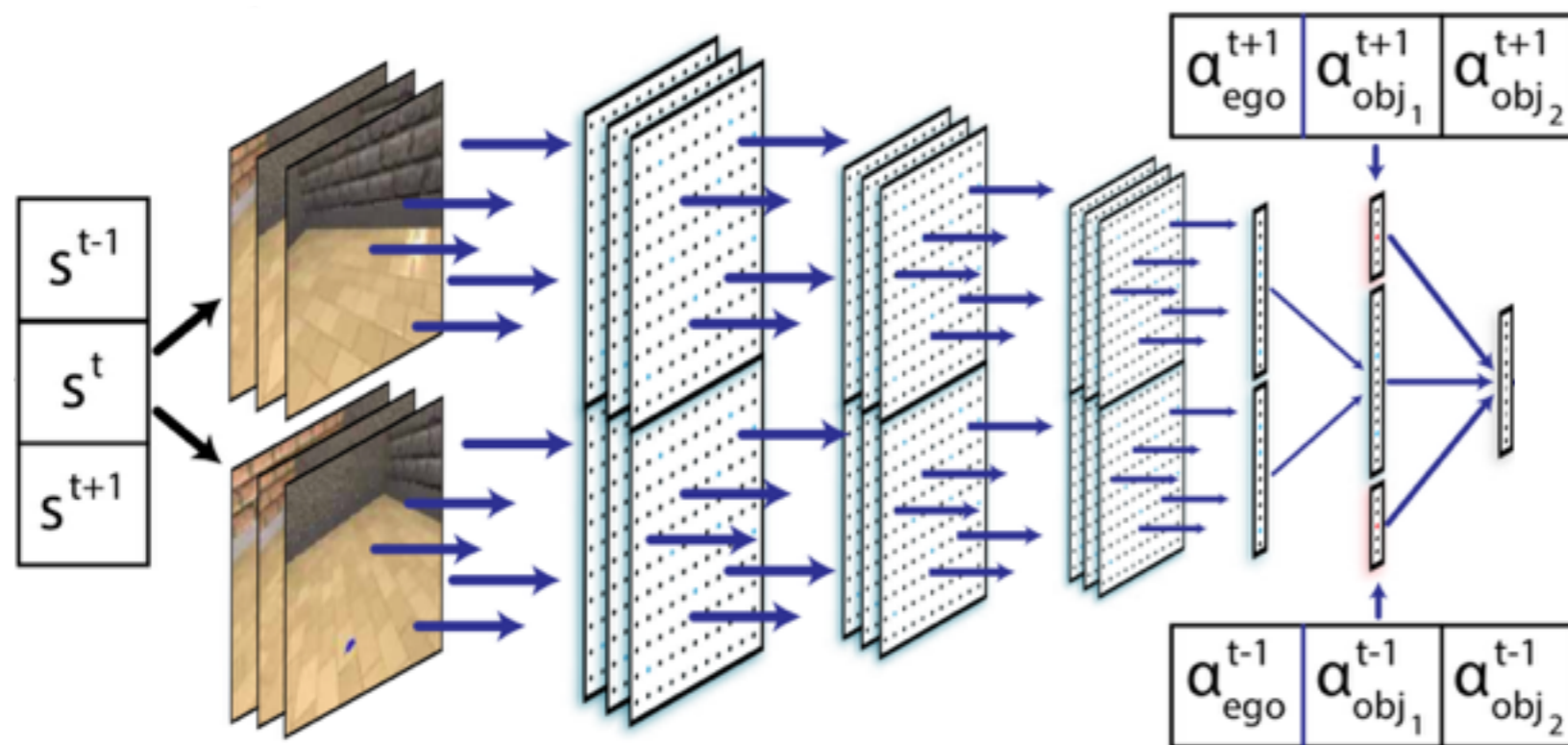
“Post-dict” the action taken given past and future states and actions



# Learning to Play — World-Model

Goal of **world-model**:

“Post-dict” the action taken given past and future states and actions

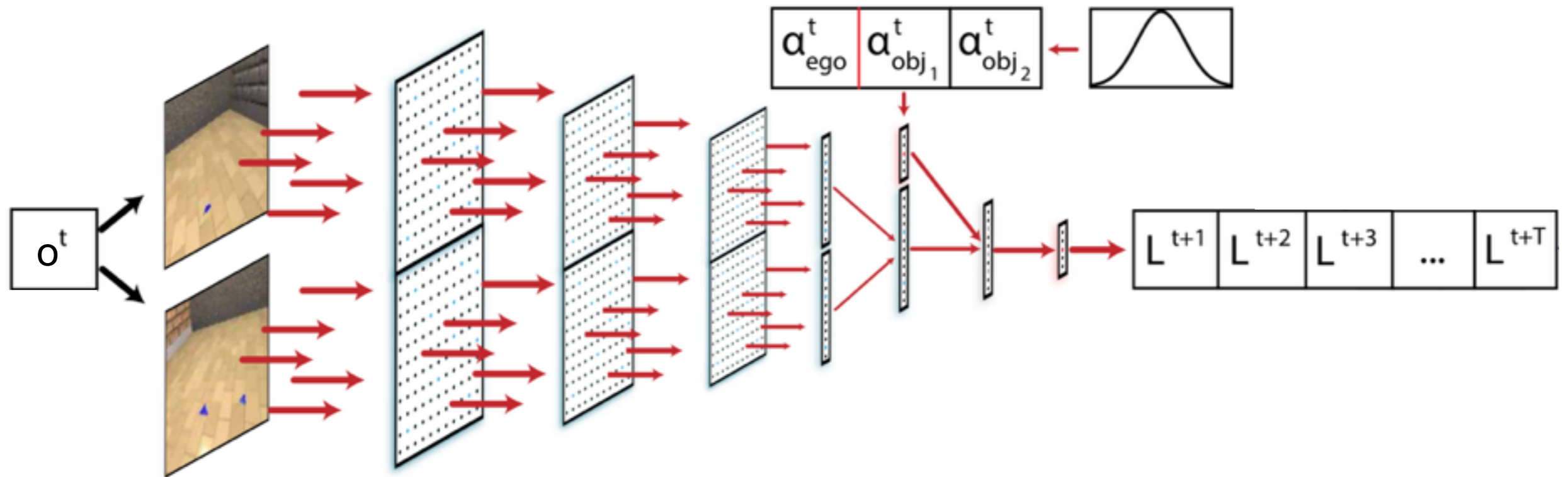


**ID**

*Inverse Dynamics*

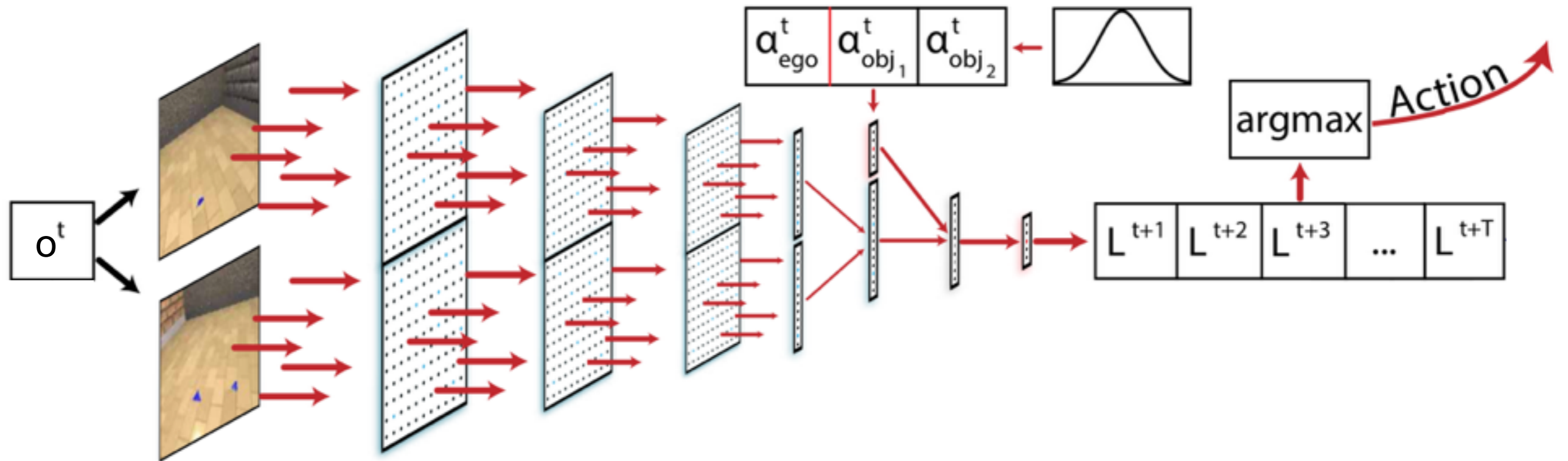
# Learning to Play — Self-Model

Goal of **self-model**: Predict errors (“loss”) of World-Model

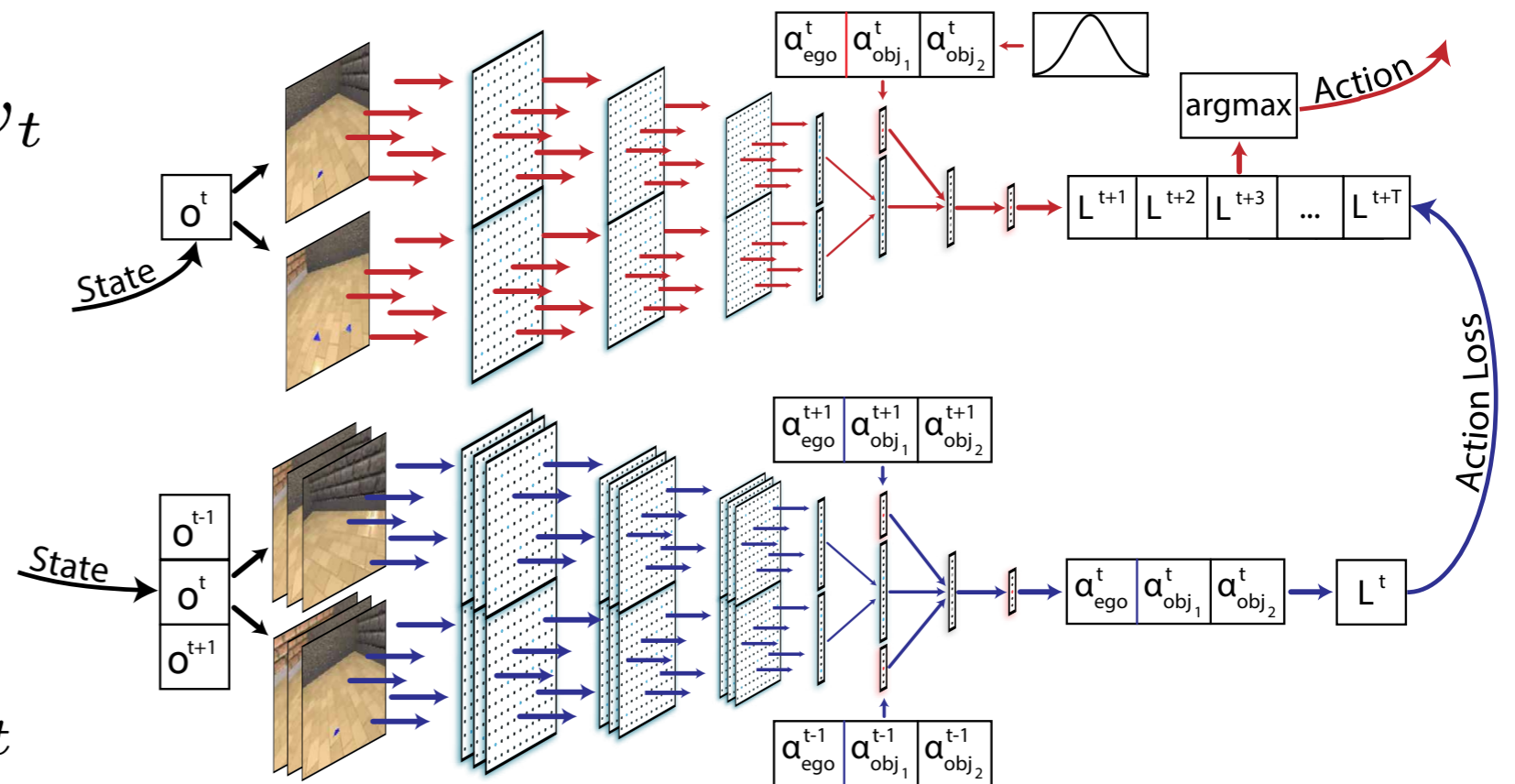
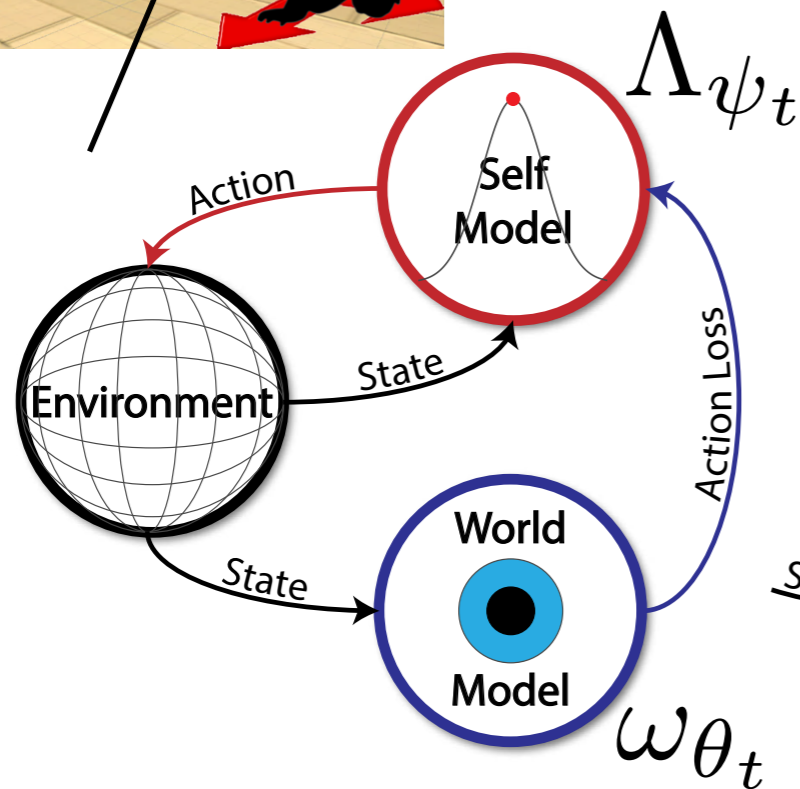
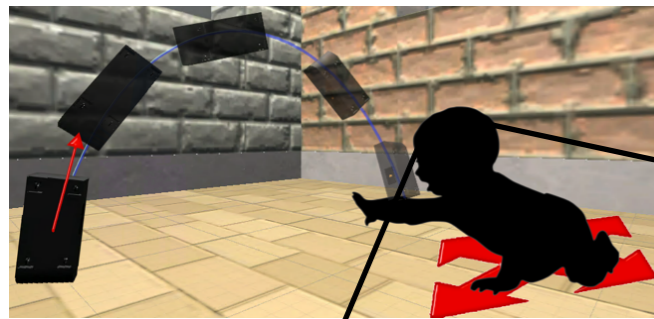


# Learning to Play — Self-Model

Goal of **self-model**: Predict errors (“loss”) of World-Model



# Learning to Play - Model overview



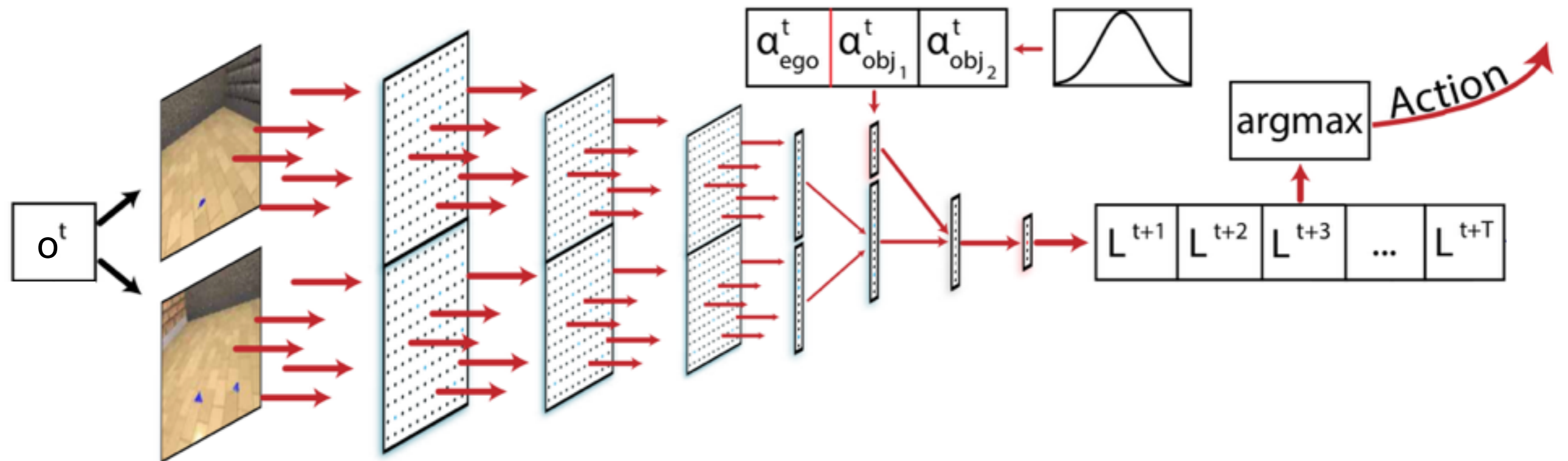
Goal of **world-model** network is to predict consequences of actions

Goal of **self-model** network is to predict errors of world-model (“self-aware”)



# Learning to Play — Self-Model

Goal of **self-model**: Predict errors (“loss”) of World-Model



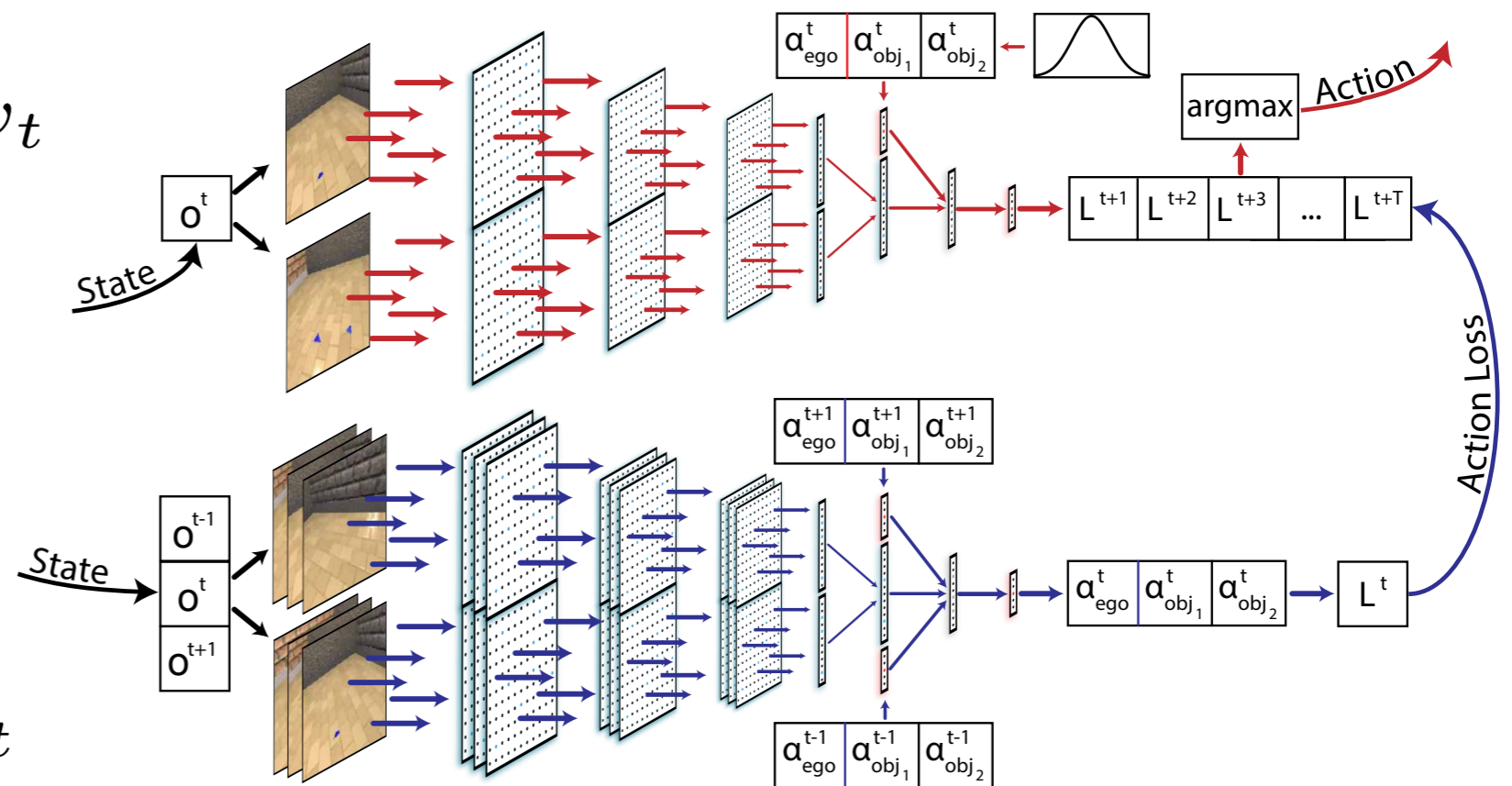
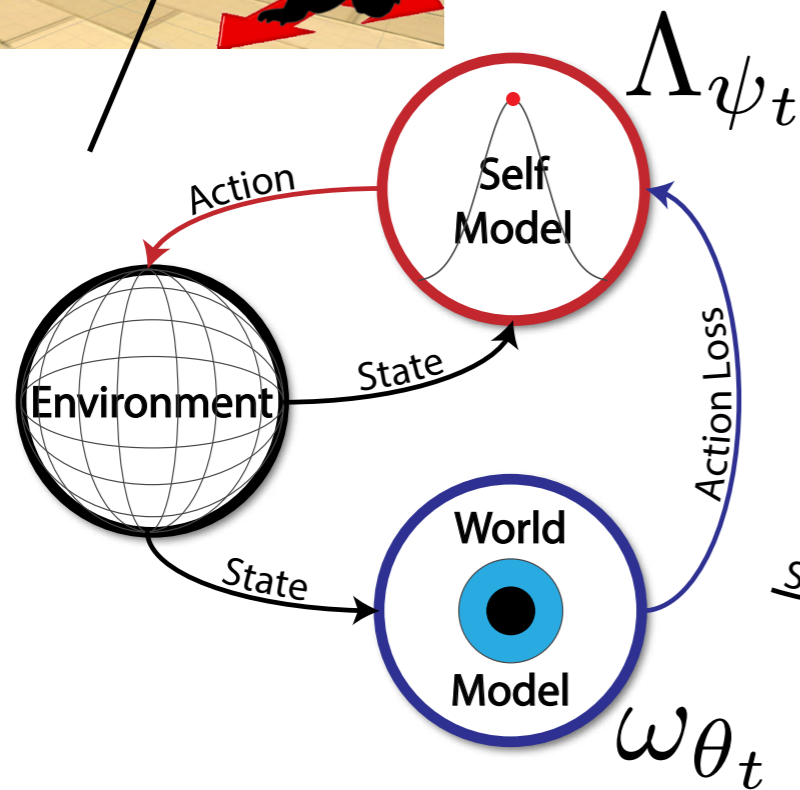
Sample 1000x actions and choose the one that maximizes the World-Model loss

$$\pi(a) \sim \exp(\beta \sigma_{\Lambda}(a))$$

**Policy mechanism**

# Learning to Play — Adversarial Policy

Action choice: self-model is **adversarial** to world-model (“curious intrinsic motivation”)



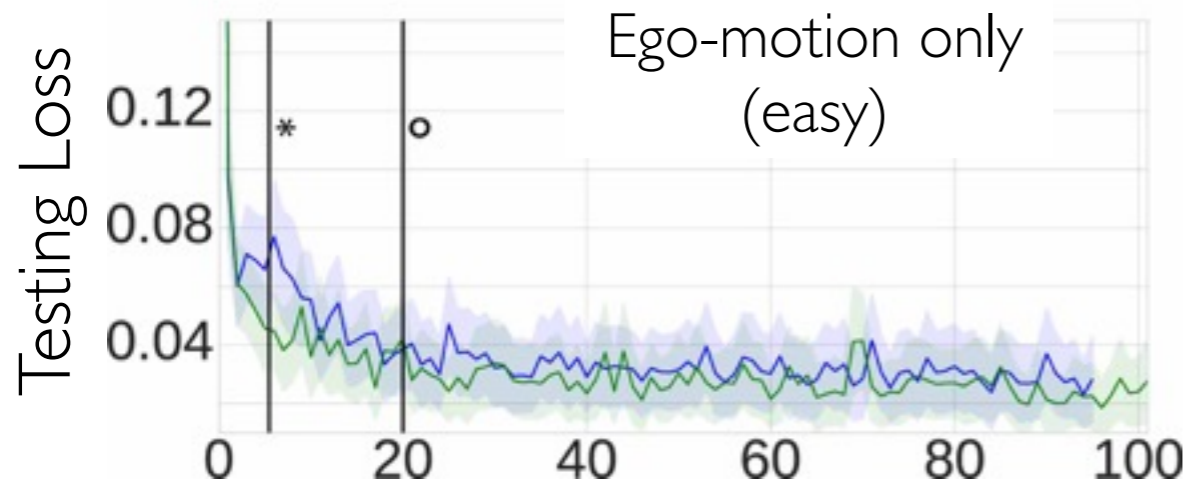
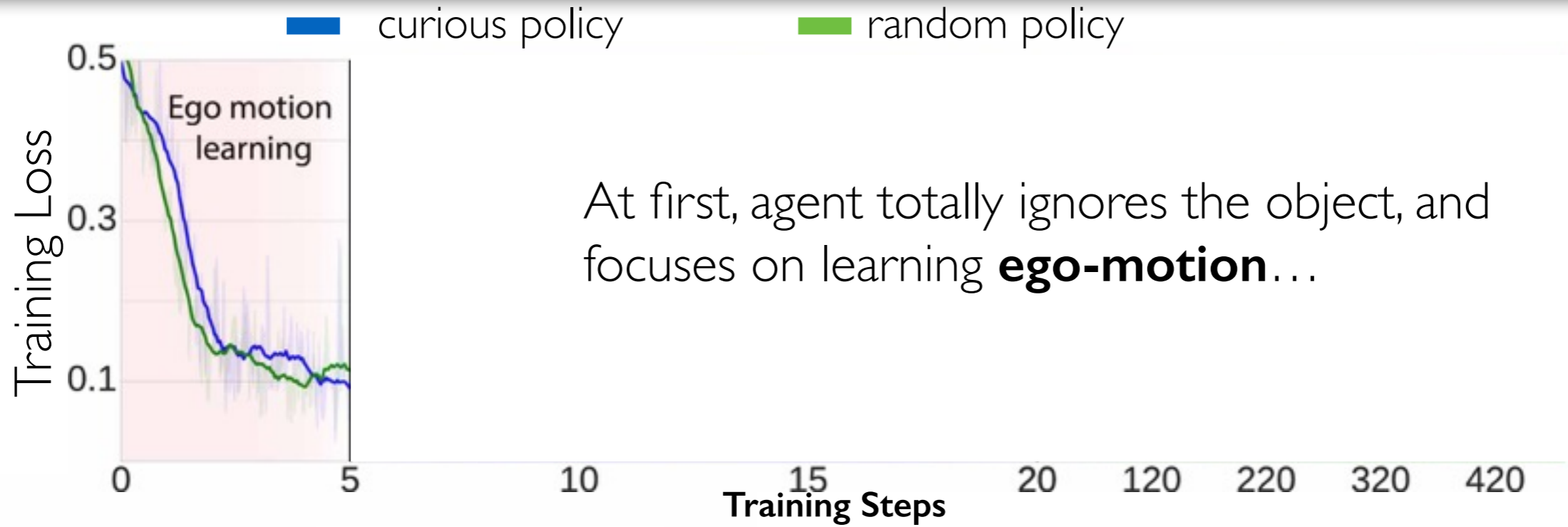
Goal of **world-model** network is to predict consequences of actions

Goal of **self-model** network is to predict errors of world-model (“self-aware”)

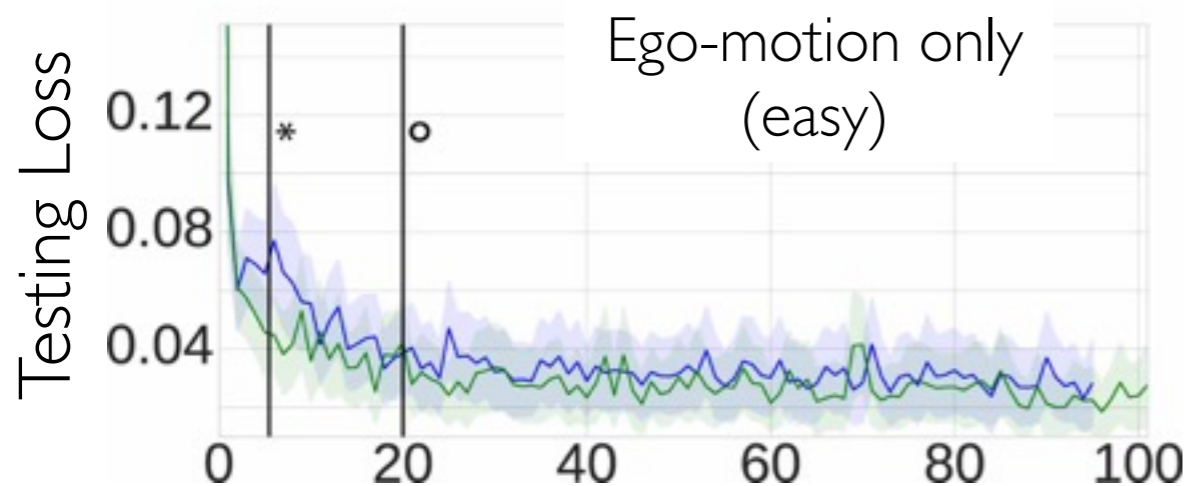
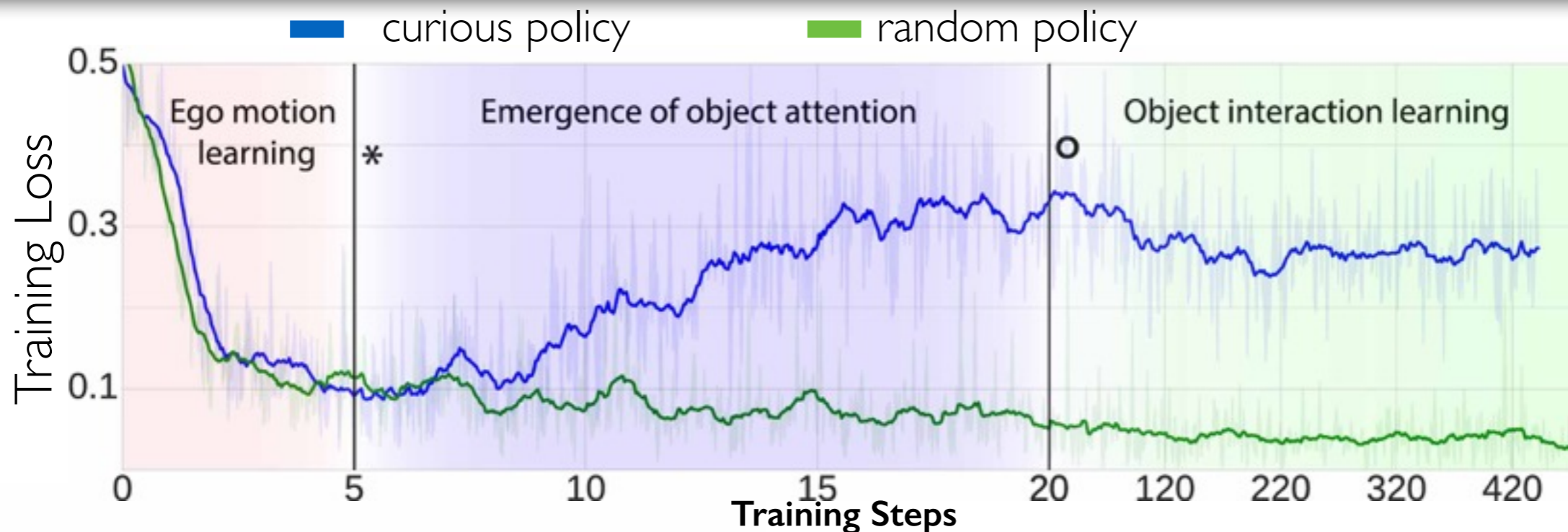
Place agent in room with a single object.



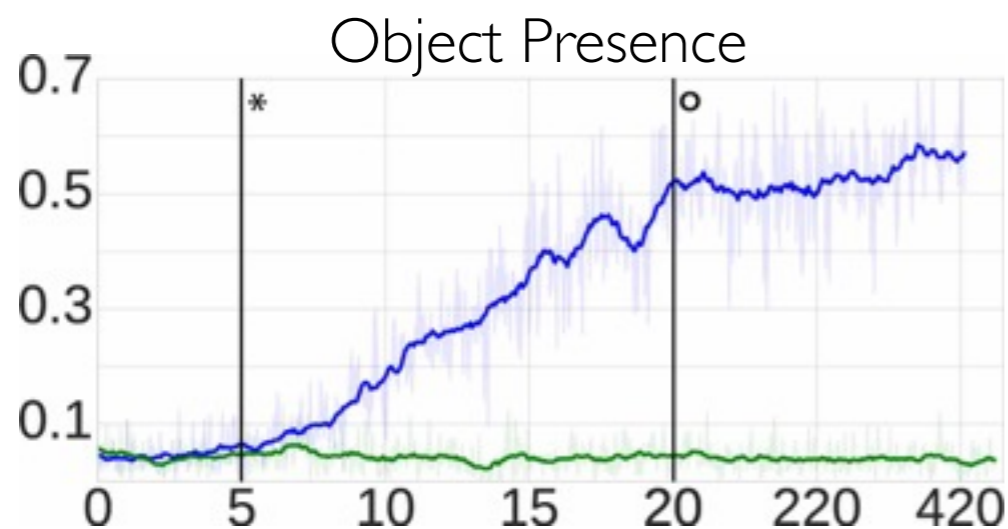
# Self-supervised learning



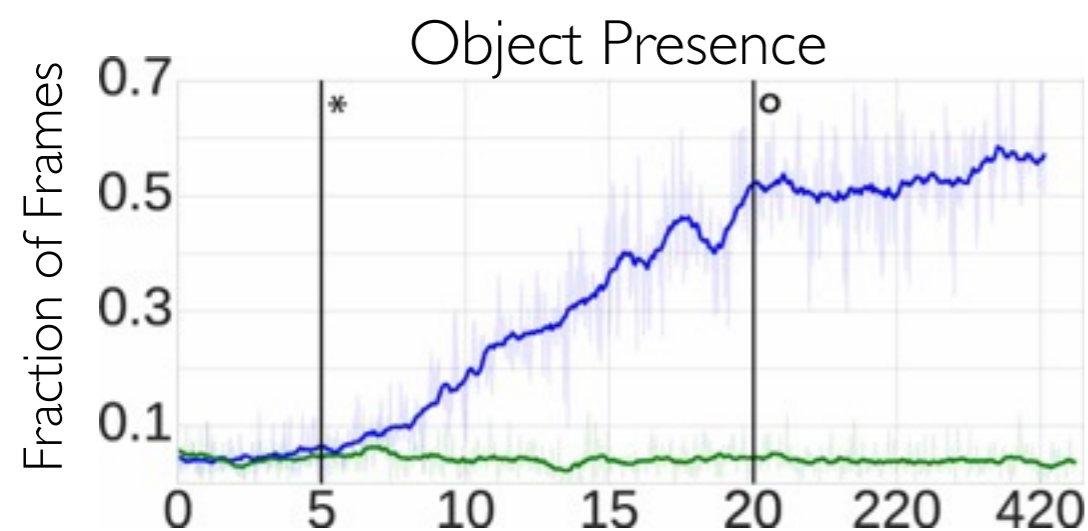
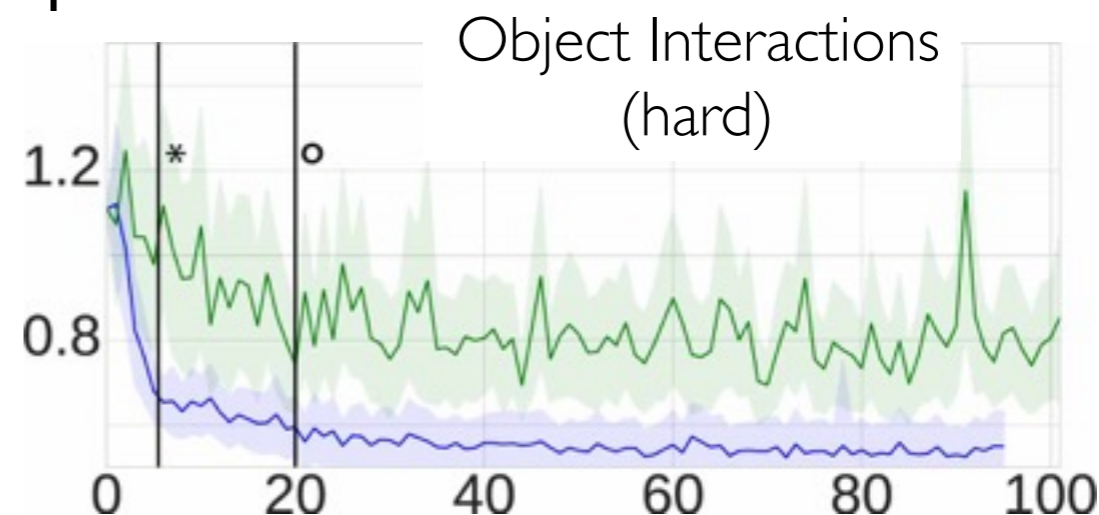
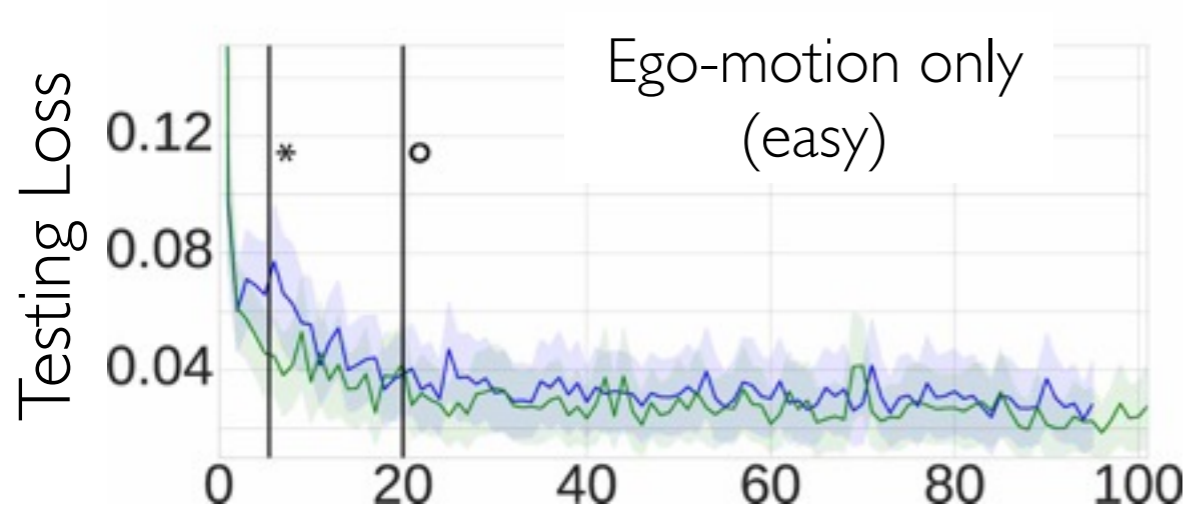
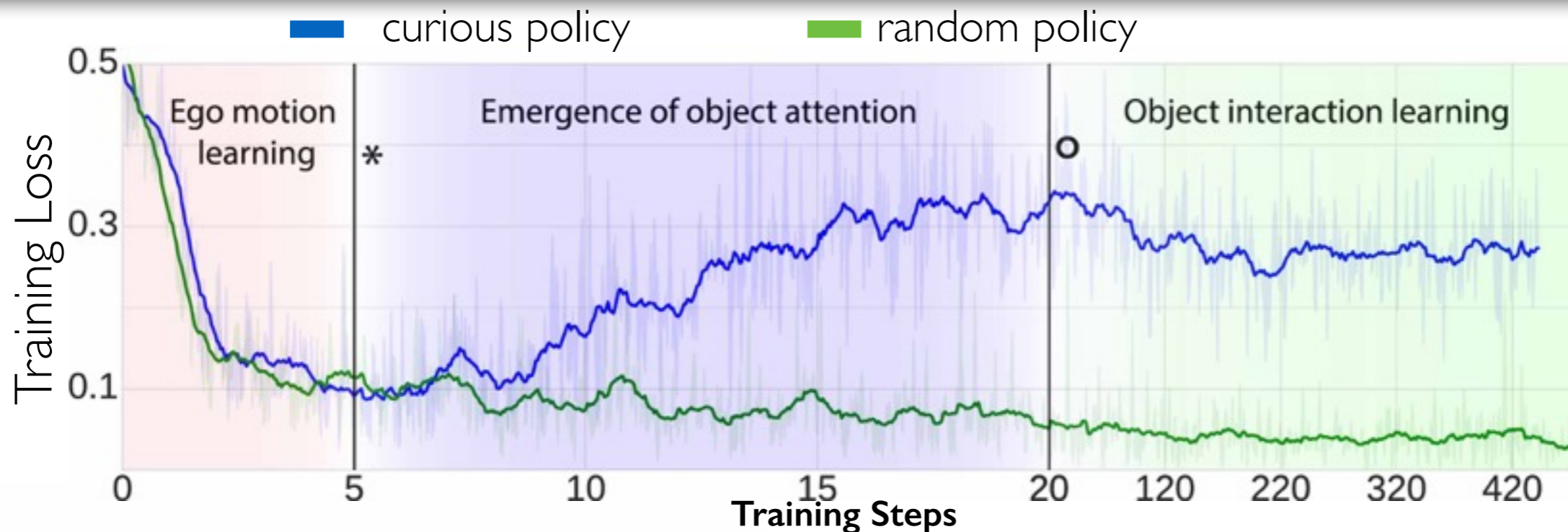
# Self-supervised learning



But the curious agent eventually “gets bored” of ego-motion prediction and starts to focus on the object!



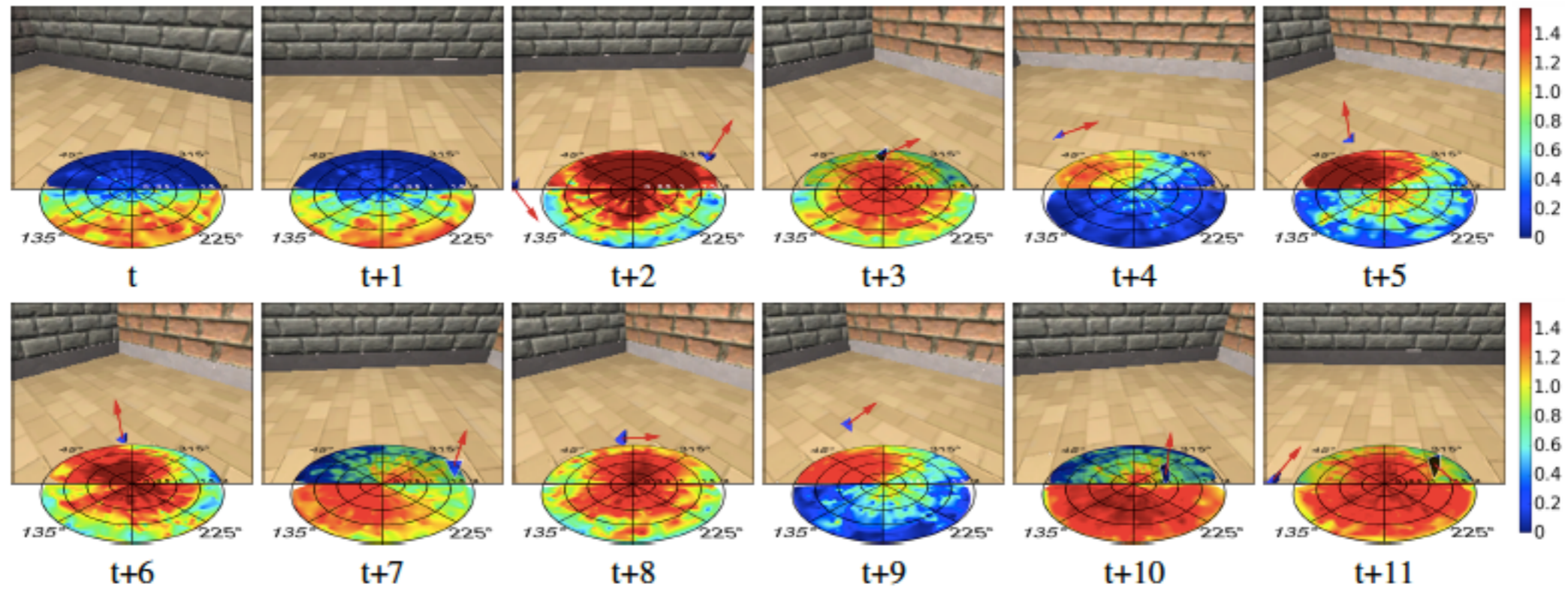
# Self-supervised learning



- Emergence of:
- (a) ego-motion understanding
  - (b) object attention
  - (c) improved world-model

# Self-supervised learning

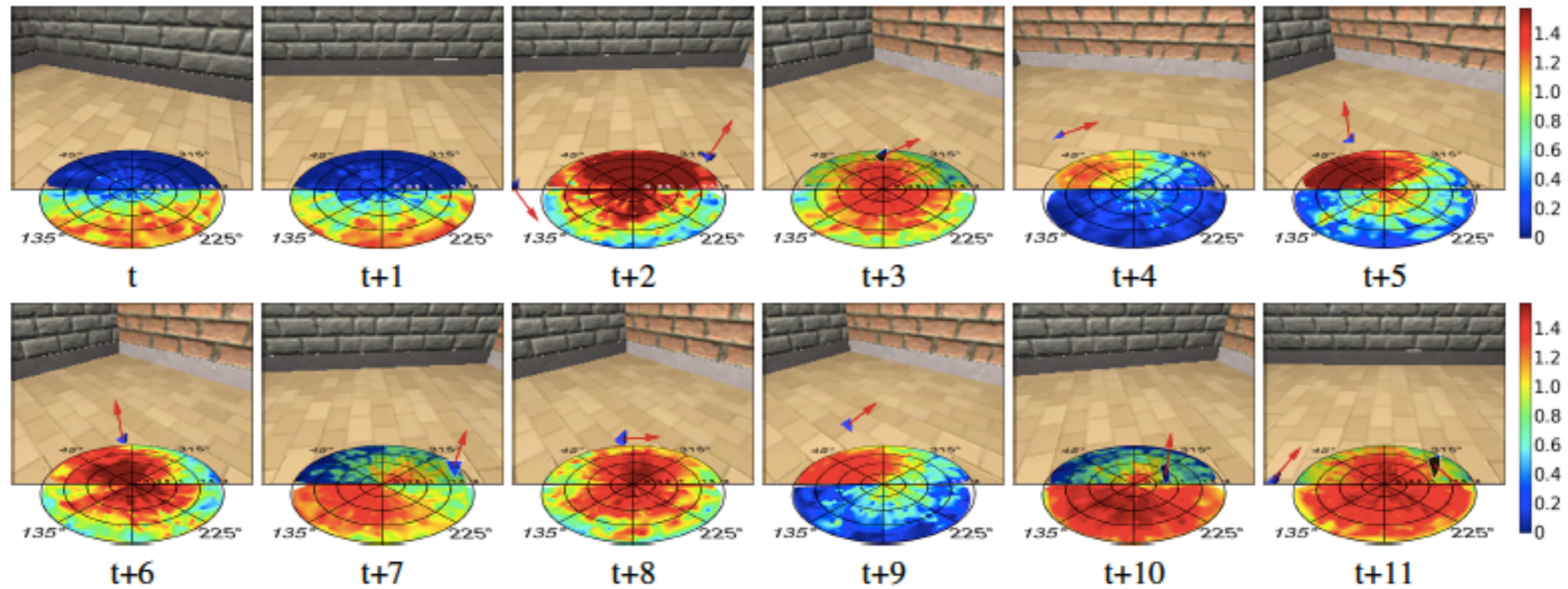
Simple navigation and planning behavior emerges ...



If an object is not in view, the agent turns to find one...

# Self-supervised learning

Simple navigation and planning behavior emerges ...



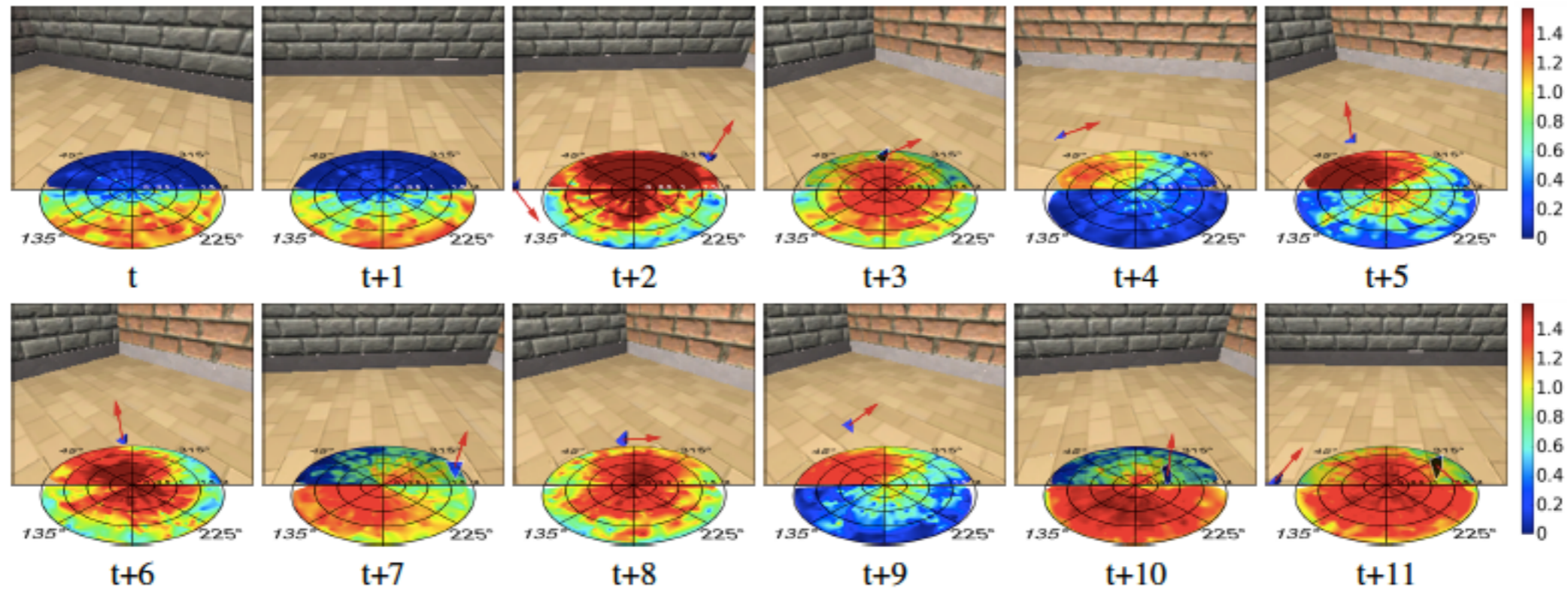
If an object is not in view, the agent turns to find one...

... if an object is too far to touch, the agent moves toward one.



# Self-supervised learning

Simple navigation and planning behavior emerges ...



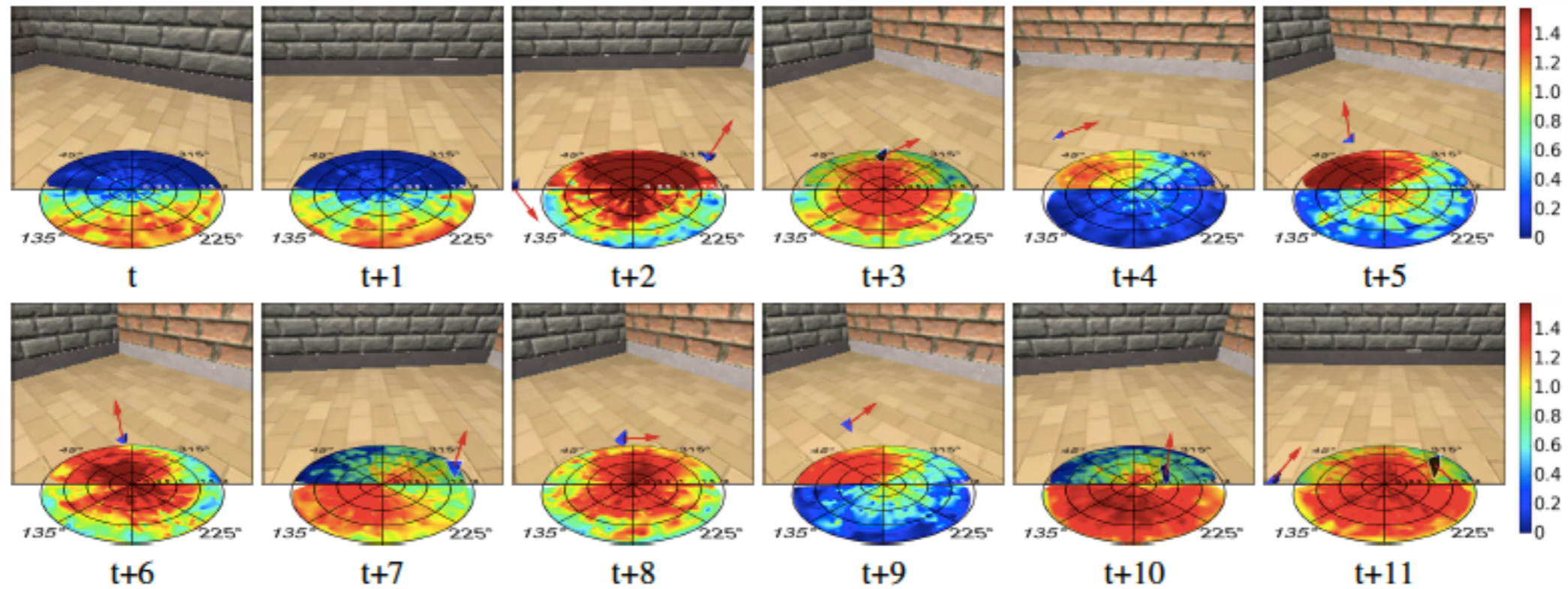
If an object is not in view, the agent turns to find one...

... if an object is too far to touch, the agent moves toward one.

... and once the agent is close to an object, it stays close and interacts with it.

# Self-supervised learning

Simple navigation and planning behavior emerges ...

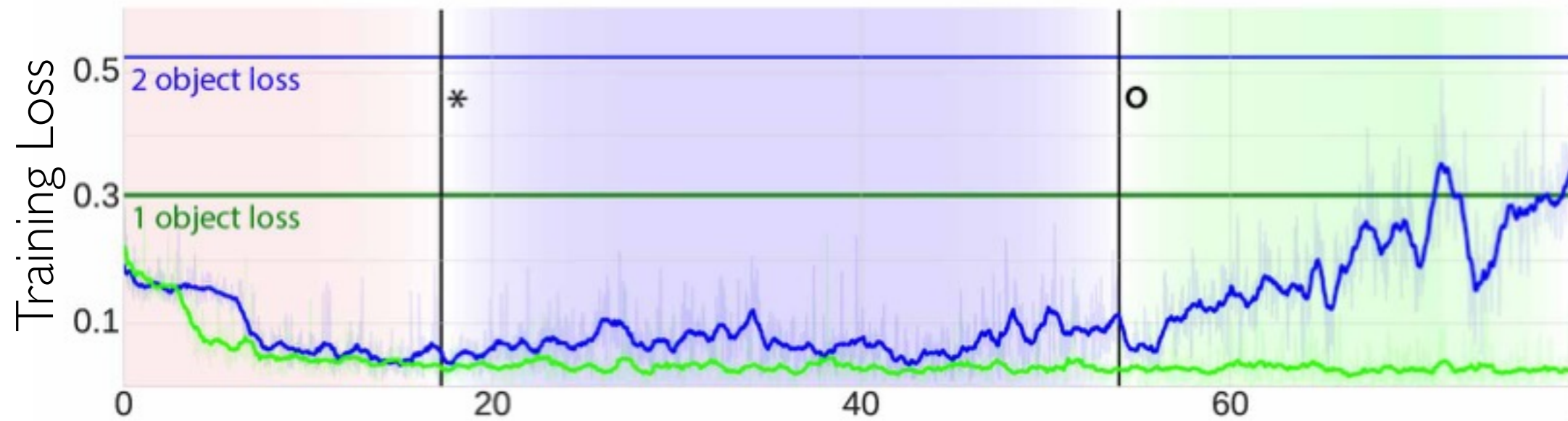


Moreover, substantially improved transfer learning accuracy:

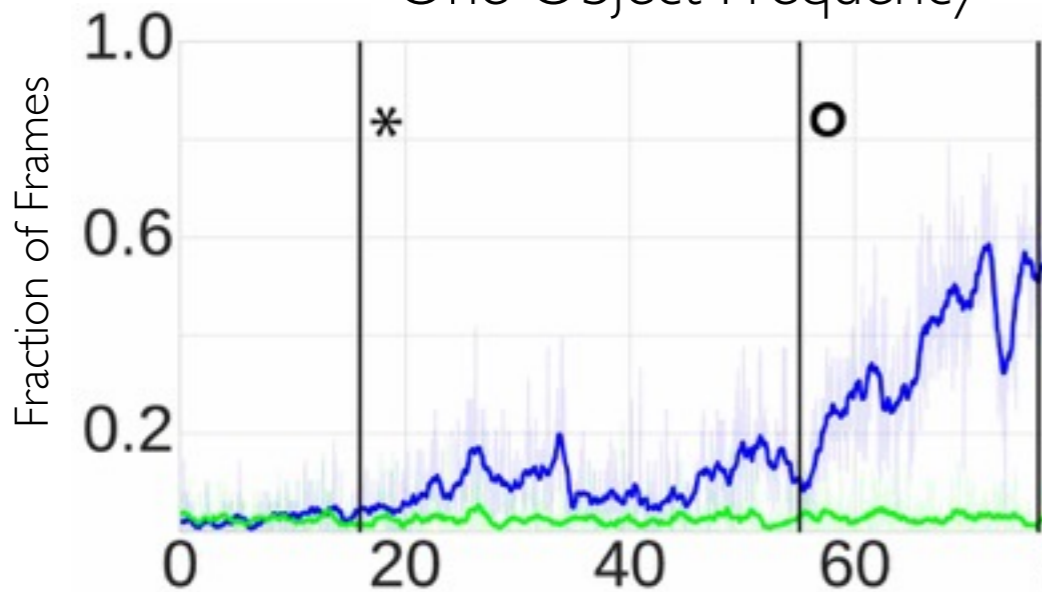
- (a) object detection (present or not):  $\sim 8\%$  vs  $\sim 40\%$  accuracy
- (b) object position:  $\sim 6\text{px}$  vs  $\sim 4\text{px}$  error
- (c) object recognition (among 16 geometries):  $\sim 12\%$  vs  $\sim 30\%$  accuracy

# Self-supervised learning

■ curious policy      ■ random policy



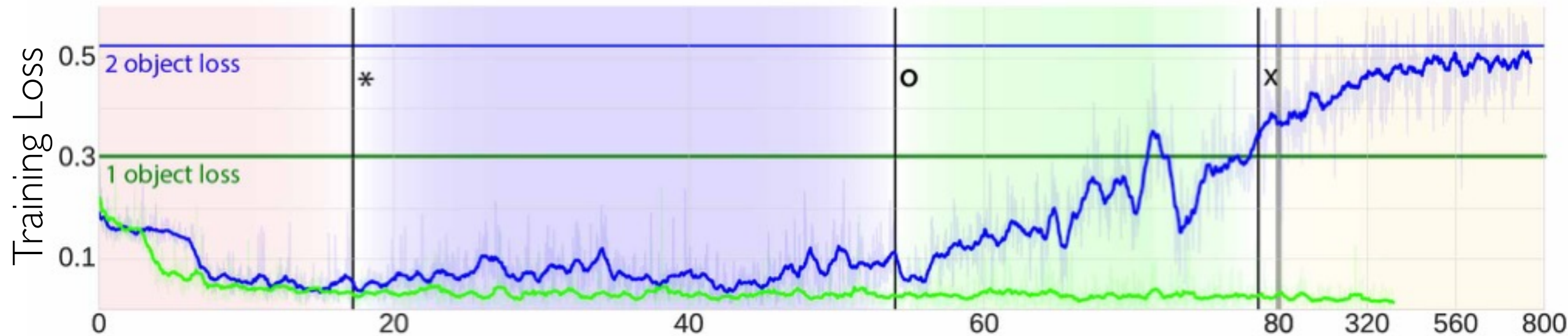
One Object Frequency



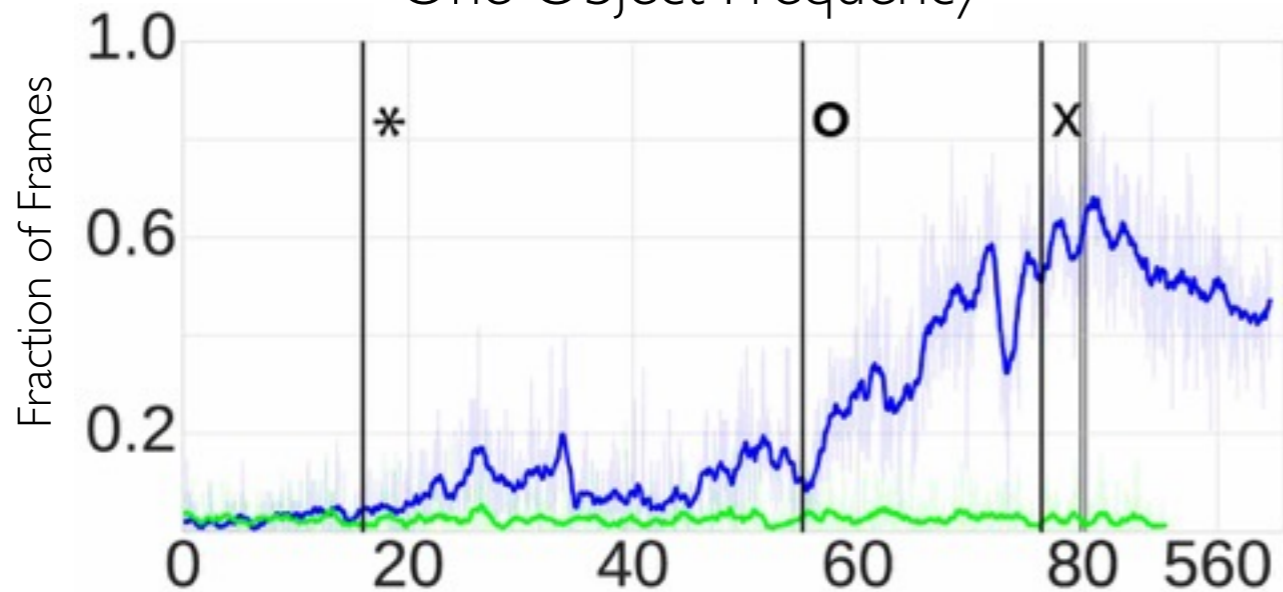
When multiple objects are present, the agent at first recapitulates its behavior with a single object ...

# Self-supervised learning

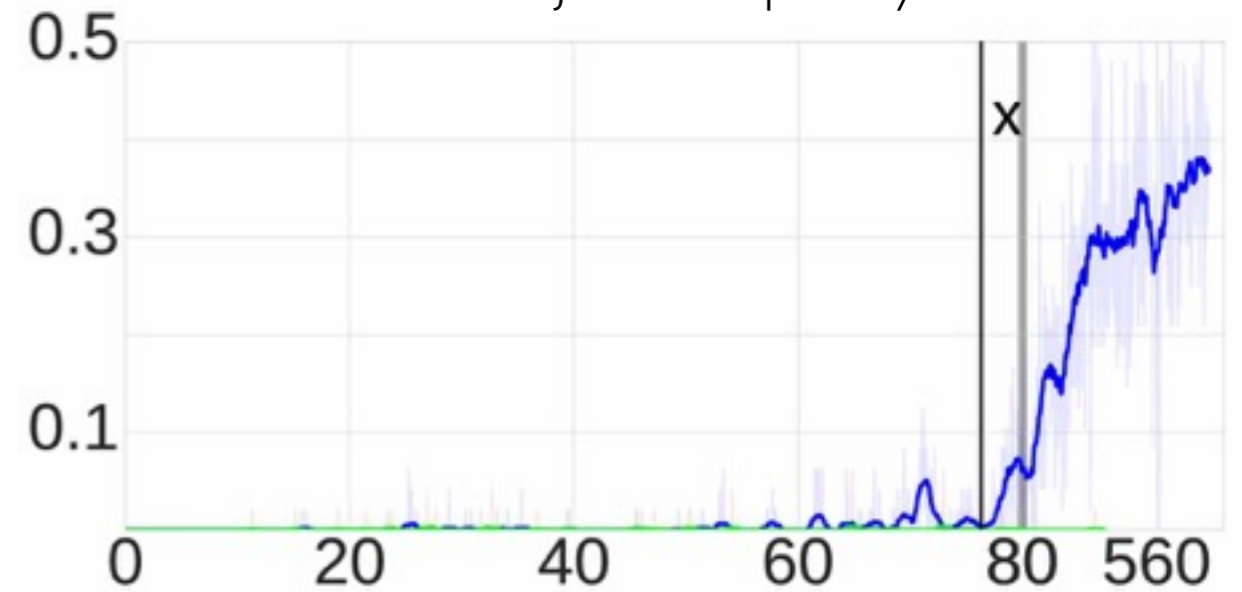
■ curious policy      ■ random policy



One Object Frequency



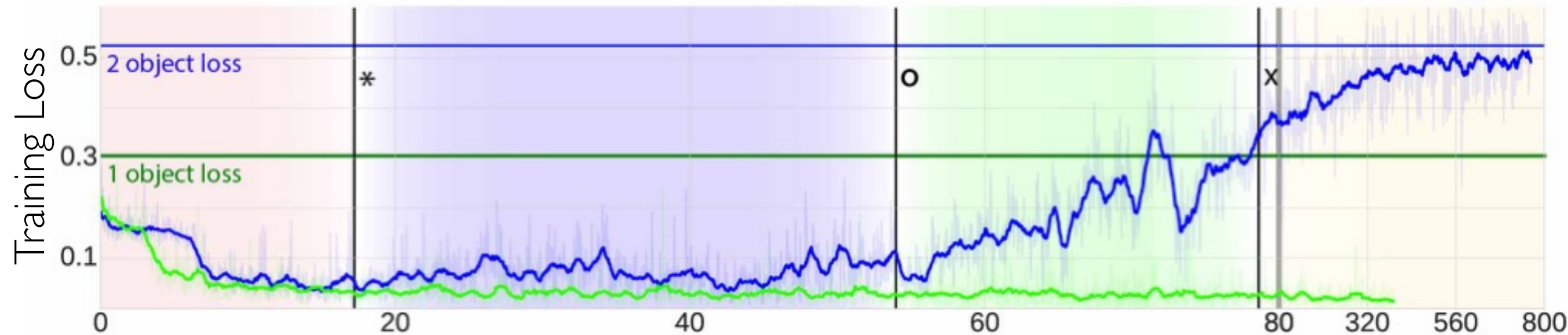
Two Object Frequency



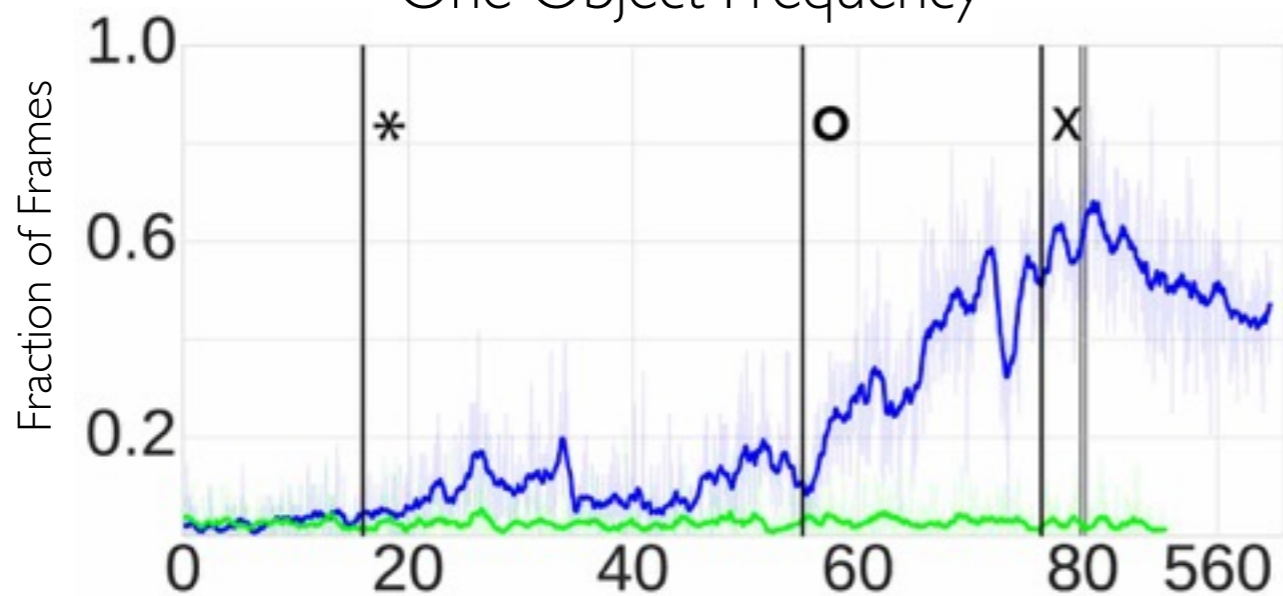
... but then discovers the interest of bringing objects together.

# Self-supervised learning

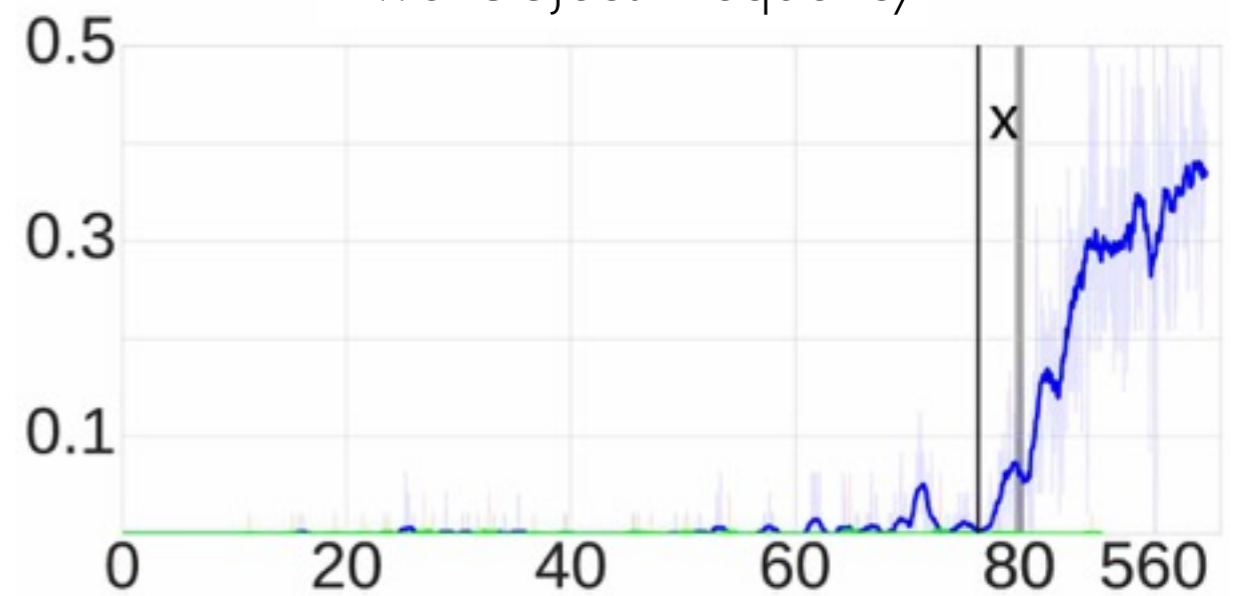
■ curious policy      ■ random policy



One Object Frequency



Two Object Frequency



Object recognition in testing (one object per image):  $\sim 16\%$  vs  $\sim 40\%$  accuracy

... especially large gain compared to training in single-obj case

Glossing over a key problem:

The above ideas rely on having the agent solve a dynamics prediction problem about the world.

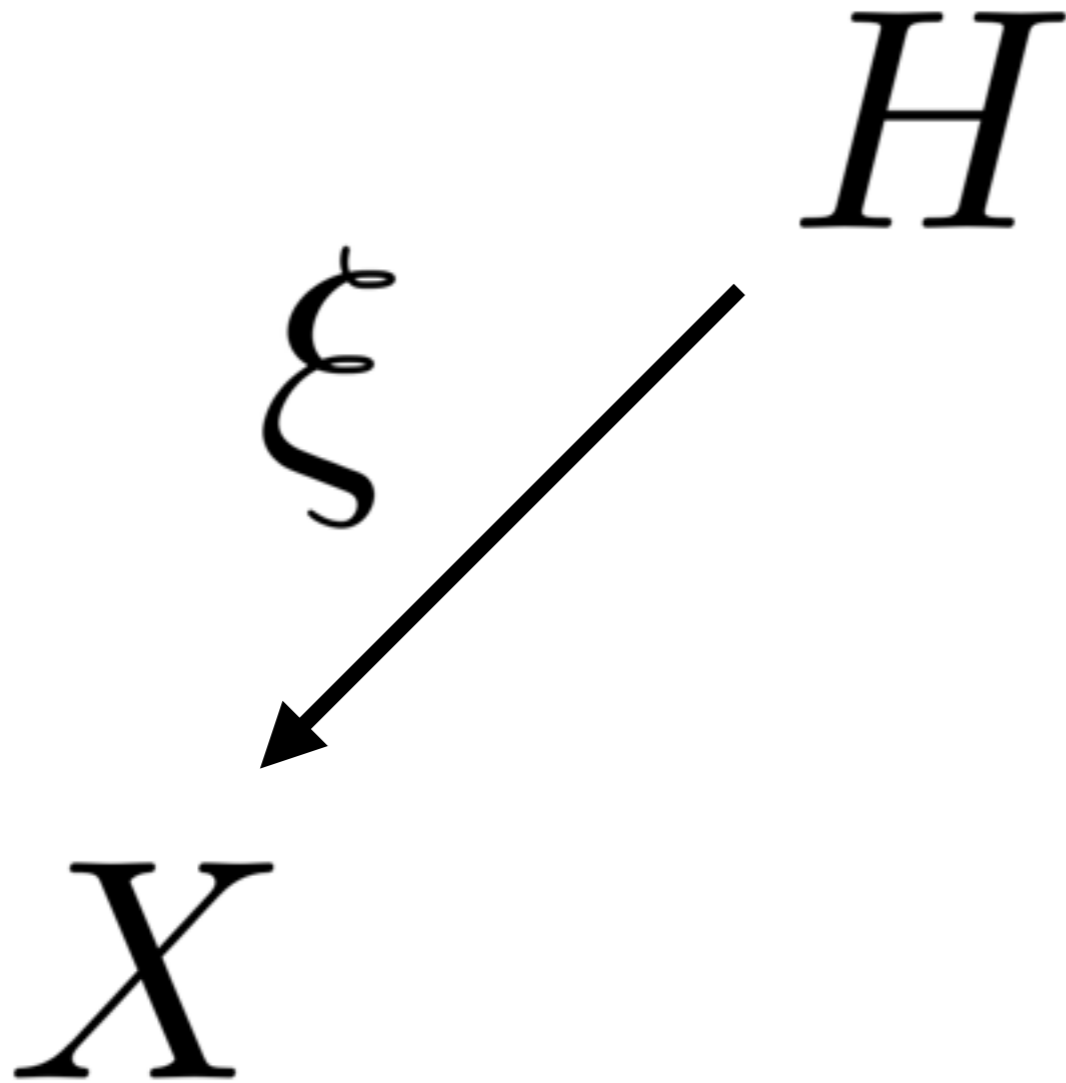
# Self-supervised learning: General Formulation

*Start with some data  $H$ ...*

*$H$*

# Self-supervised learning: General Formulation

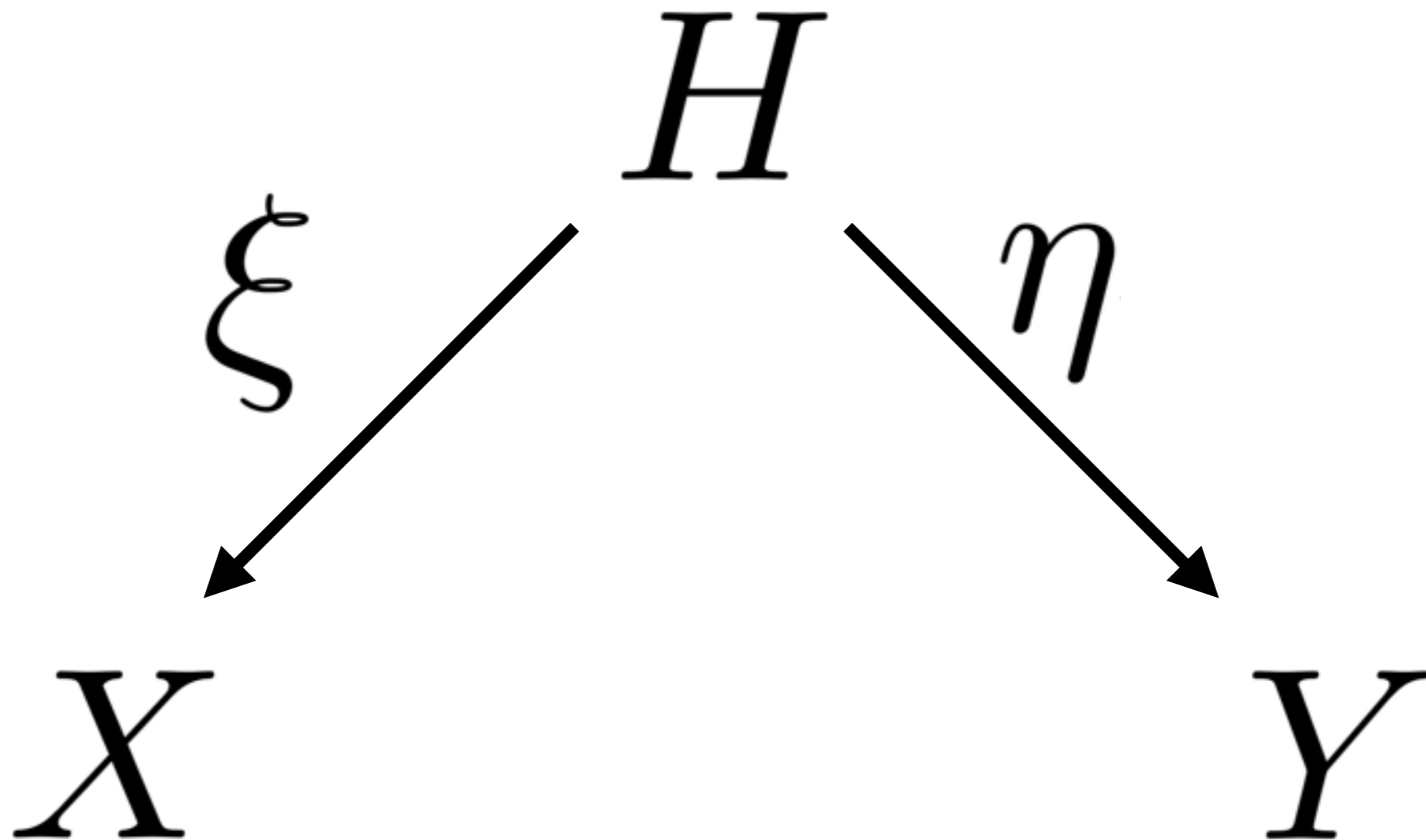
...Create input data  $X$  from  $H$ ...





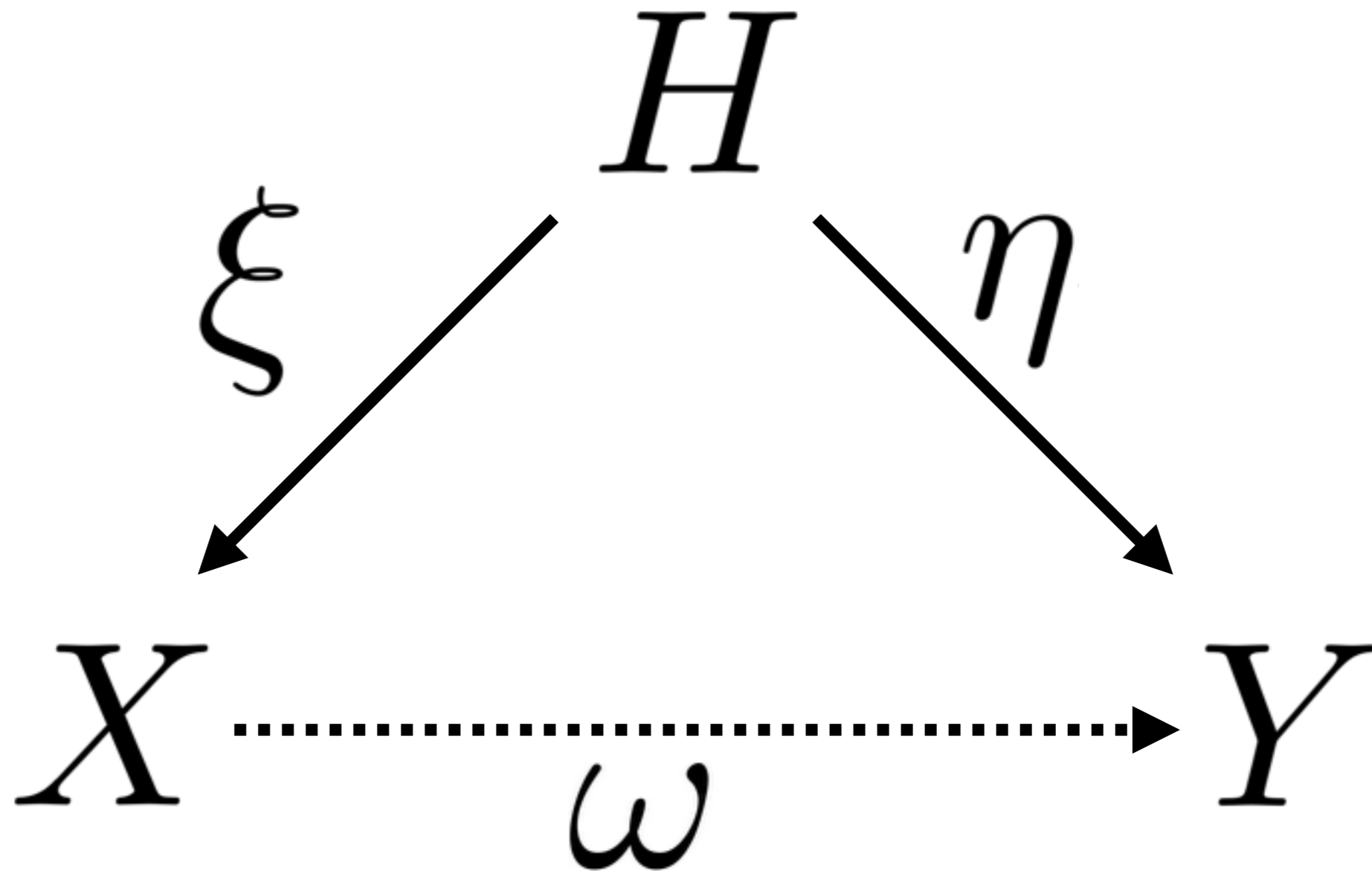
# Self-supervised learning: General Formulation

...Create output data  $Y$  from  $H$ ...

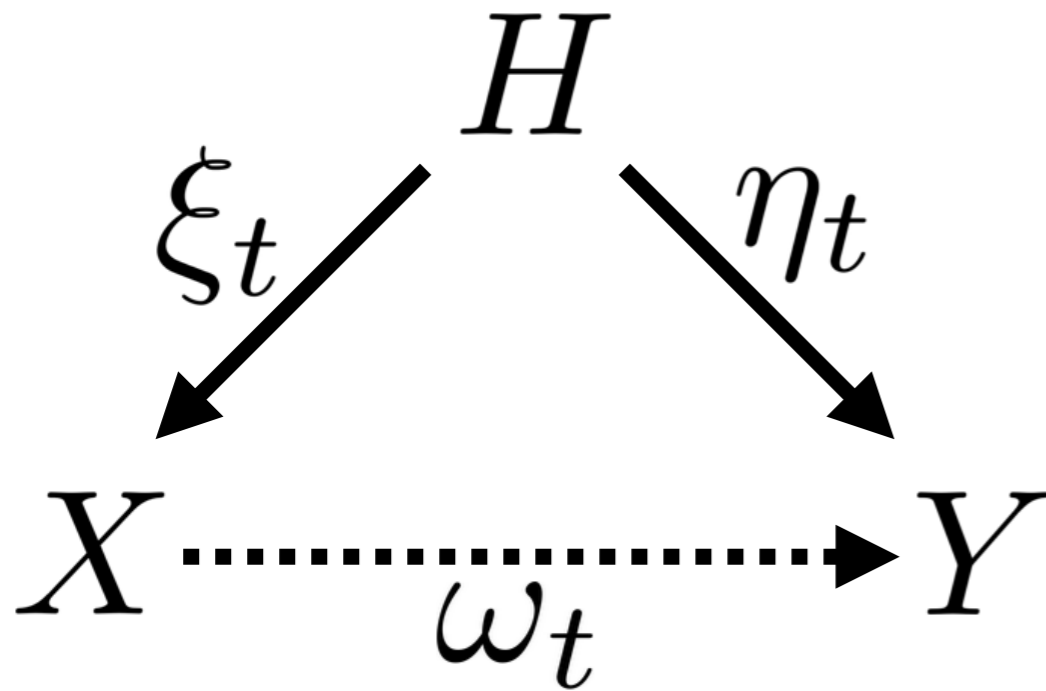


# Self-supervised learning: General Formulation

...Predict  $Y$  from  $X$ .



# Self-supervised learning: General Formulation

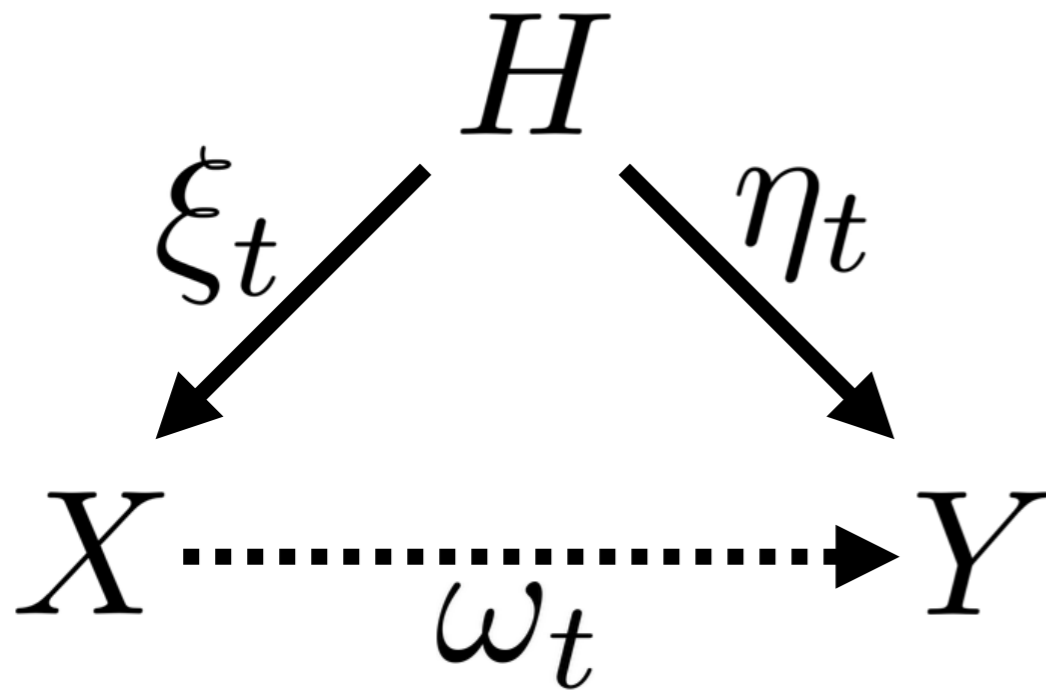


Examples:

1) Forward future prediction

$state(t), action(t) \implies state(t+1)$

# Self-supervised learning: General Formulation



Examples:

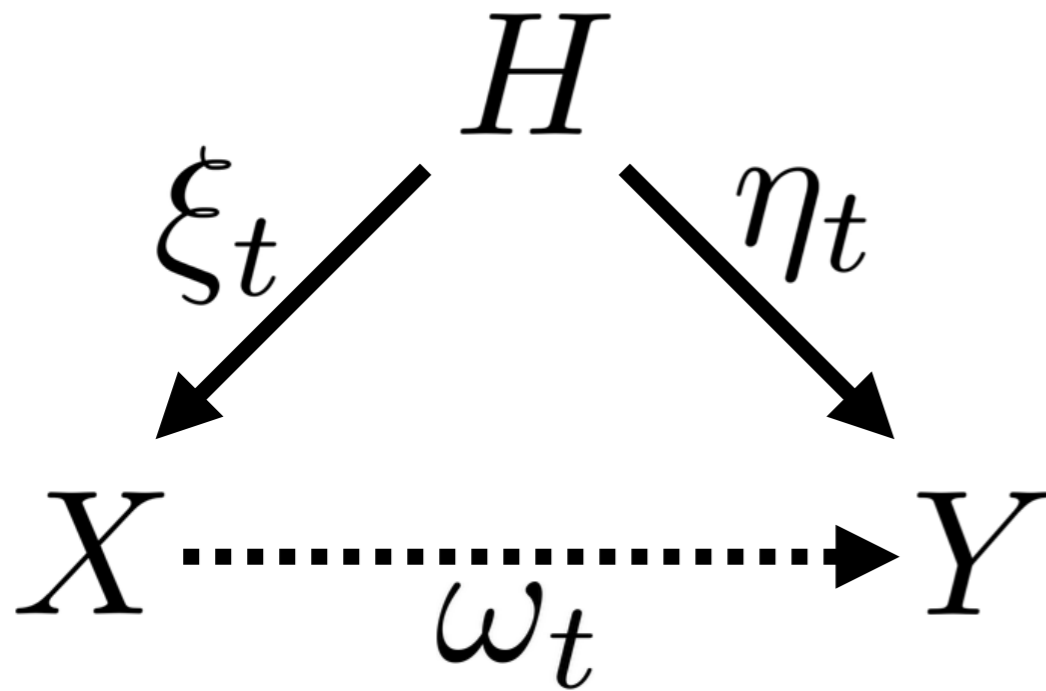
1) Forward future prediction

$$\text{state}(t), \text{action}(t) \implies \text{state}(t+1)$$

2) inverse dynamics prediction

$$\text{state}(t), \text{state}(t+1) \implies \text{action}(t)$$

# Self-supervised learning: General Formulation



Examples:

1) Forward future prediction

$$\text{state}(t), \text{action}(t) \implies \text{state}(t+1)$$

2) inverse dynamics prediction

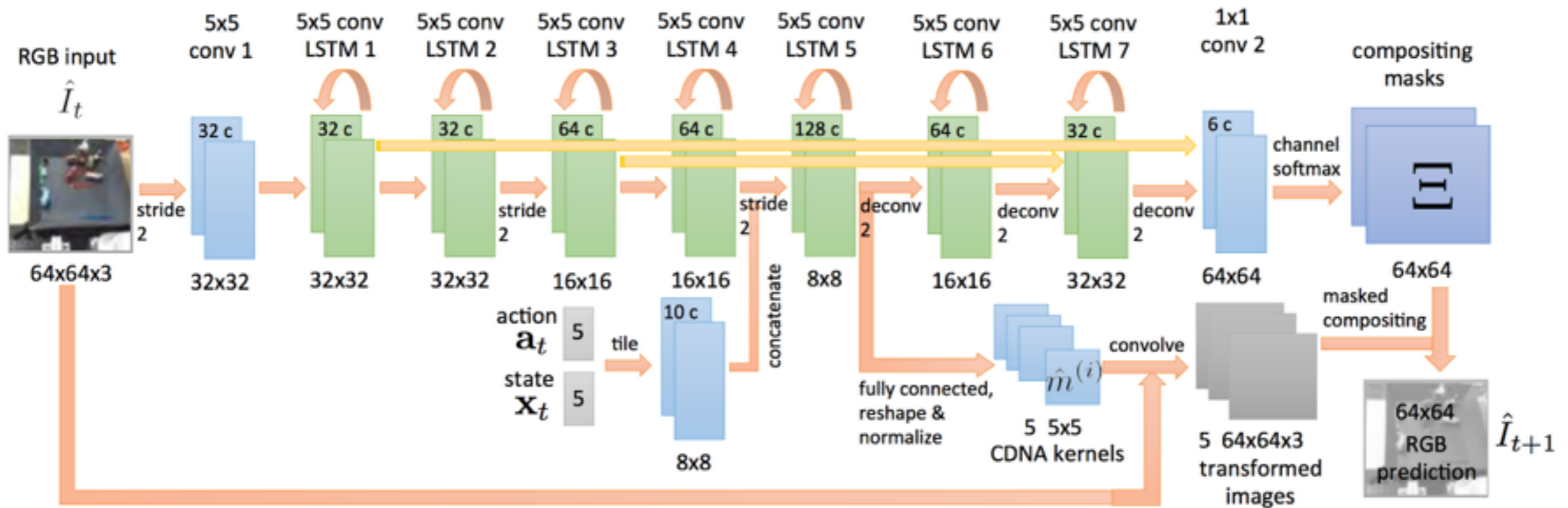
$$\text{state}(t), \text{state}(t+1) \implies \text{action}(t)$$

1) is hard, because ... pixel prediction is hard!

# Intuitive Physics as Underlying Goal

Obvious idea: just predict future pixels

Finn et. al (2016)



PredRNN(2017) ; Wang (2018) ; among many others

# Intuitive Physics as Underlying Goal

Pixel prediction is hard.

$t=1$



$t=2$



Two blue objects in a room

# Intuitive Physics as Underlying Goal

Pixel prediction is hard.

$t = 1$



$t = 2$



Two blue objects in a room

Objects acted on and camera moves



# Intuitive Physics as Underlying Goal

Pixel prediction is hard.

t=1



t=2



Two blue objects in a room

Objects acted on and camera moves

t=3



t=4



t=5



t=6



Ground  
truth

# Intuitive Physics as Underlying Goal

Pixel prediction is hard.

t=1



t=2



Two blue objects in a room

t=3



t=4



Objects acted on and camera moves

t=5



t=6

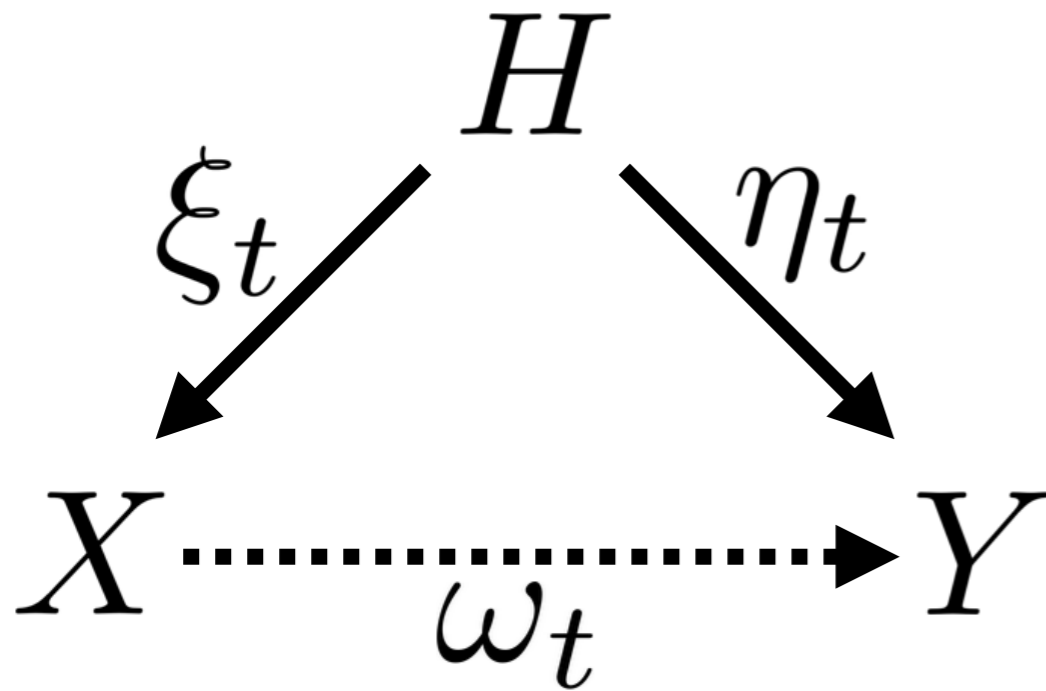


Ground truth

Prediction



# Self-supervised learning: General Formulation



Examples:

1) Forward future prediction

$$\text{state}(t), \text{action}(t) \implies \text{state}(t+1)$$

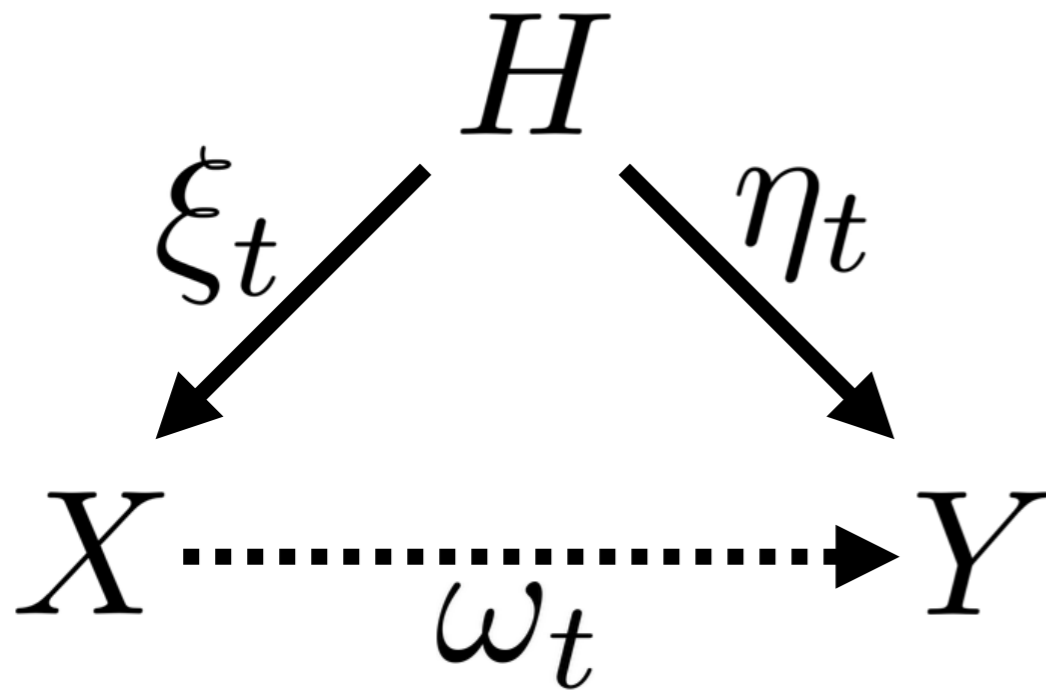
2) inverse dynamics prediction

$$\text{state}(t), \text{state}(t+1) \implies \text{action}(t)$$

1) is hard, because ... pixel prediction is hard!

2) is mostly what we did in the work described above because it's easier ...

# Self-supervised learning: General Formulation



Examples:

1) Forward future prediction

$$\text{state}(t), \text{action}(t) \implies \text{state}(t+1)$$

2) inverse dynamics prediction

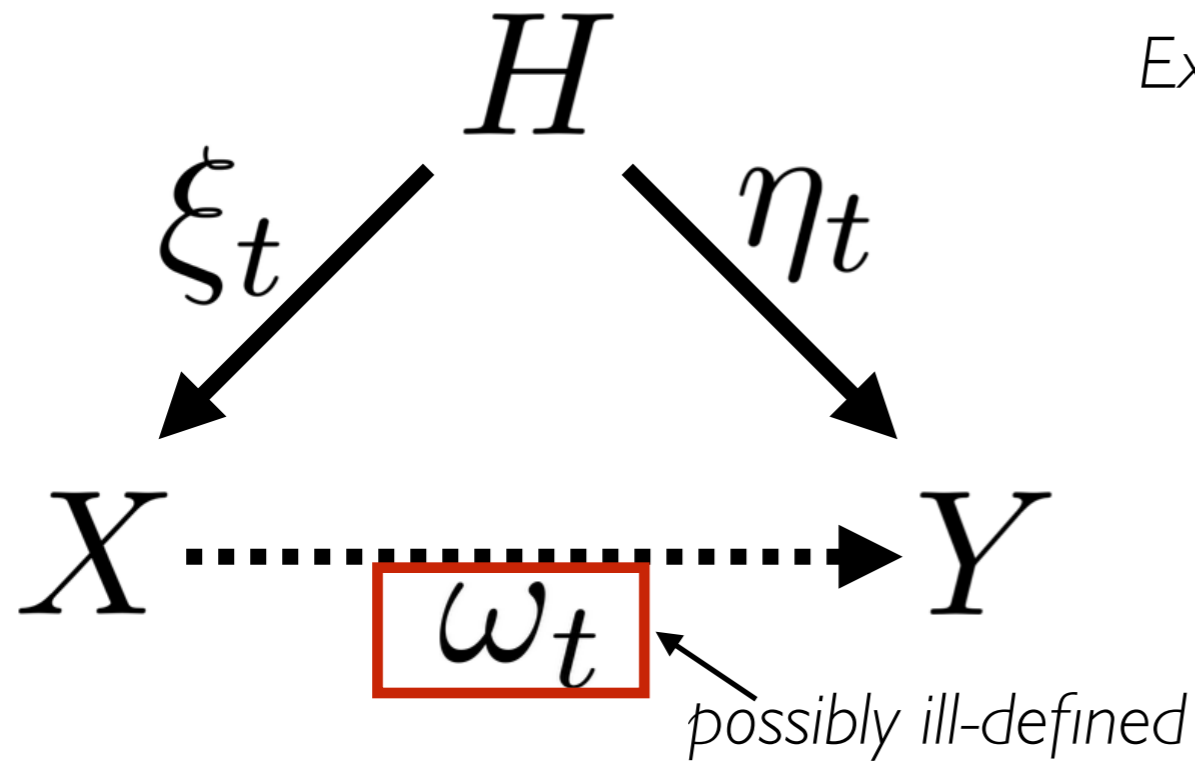
$$\text{state}(t), \text{state}(t+1) \implies \text{action}(t)$$

1) is hard, because ... pixel prediction is hard!

2) is mostly what we did in the work described above because it's easier ...

**BUT DEGENERATE!**

# Self-supervised learning: General Formulation



THE DREADED

**WHITE-NOISE PROBLEM**

Examples:

1) Forward future prediction

$$\text{state}(t), \text{action}(t) \implies \text{state}(t+1)$$

2) inverse dynamics prediction

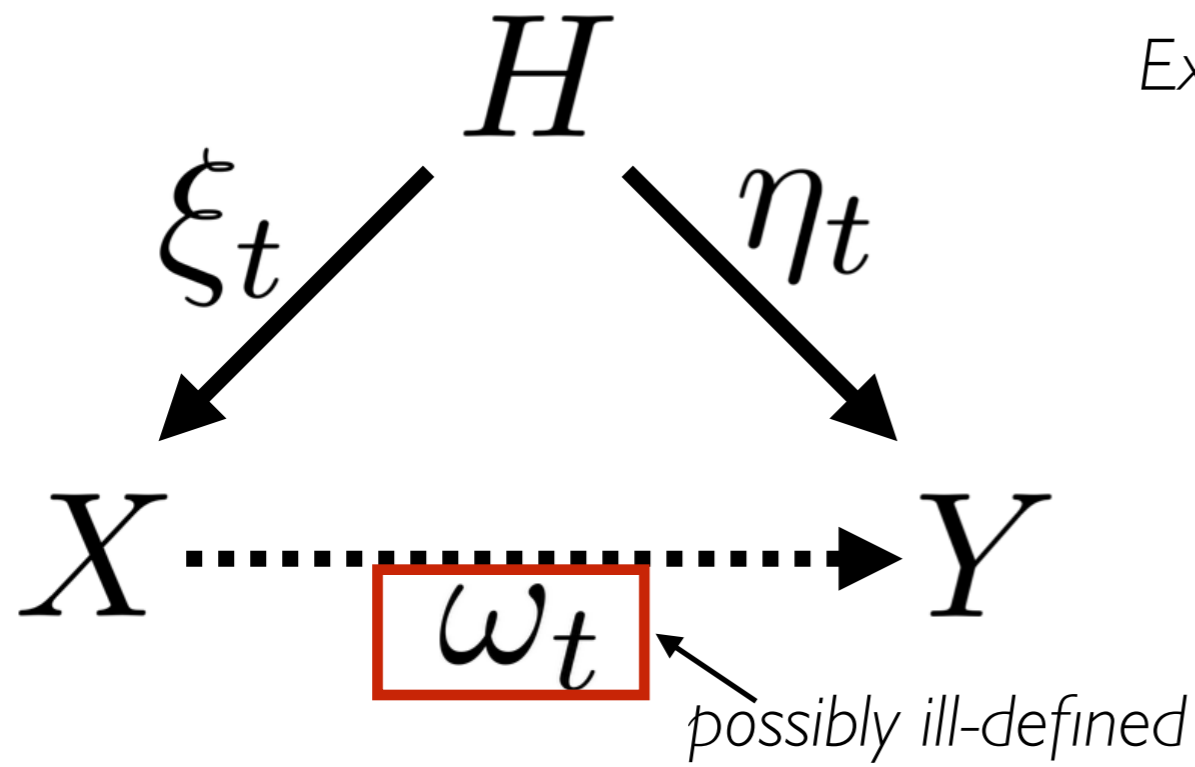
$$\text{state}(t), \text{state}(t+1) \implies \text{action}(t)$$

1) is hard, because ... pixel prediction is hard!

2) is mostly what we did in the work described above because it's easier ...

BUT DEGENERATE!

# Self-supervised learning: General Formulation



THE DREADED

**WHITE-NOISE PROBLEM**

Examples:

1) Forward future prediction

$$\text{state}(t), \text{action}(t) \implies \text{state}(t+1)$$

2) inverse dynamics prediction

$$\text{state}(t), \text{state}(t+1) \implies \text{action}(t)$$

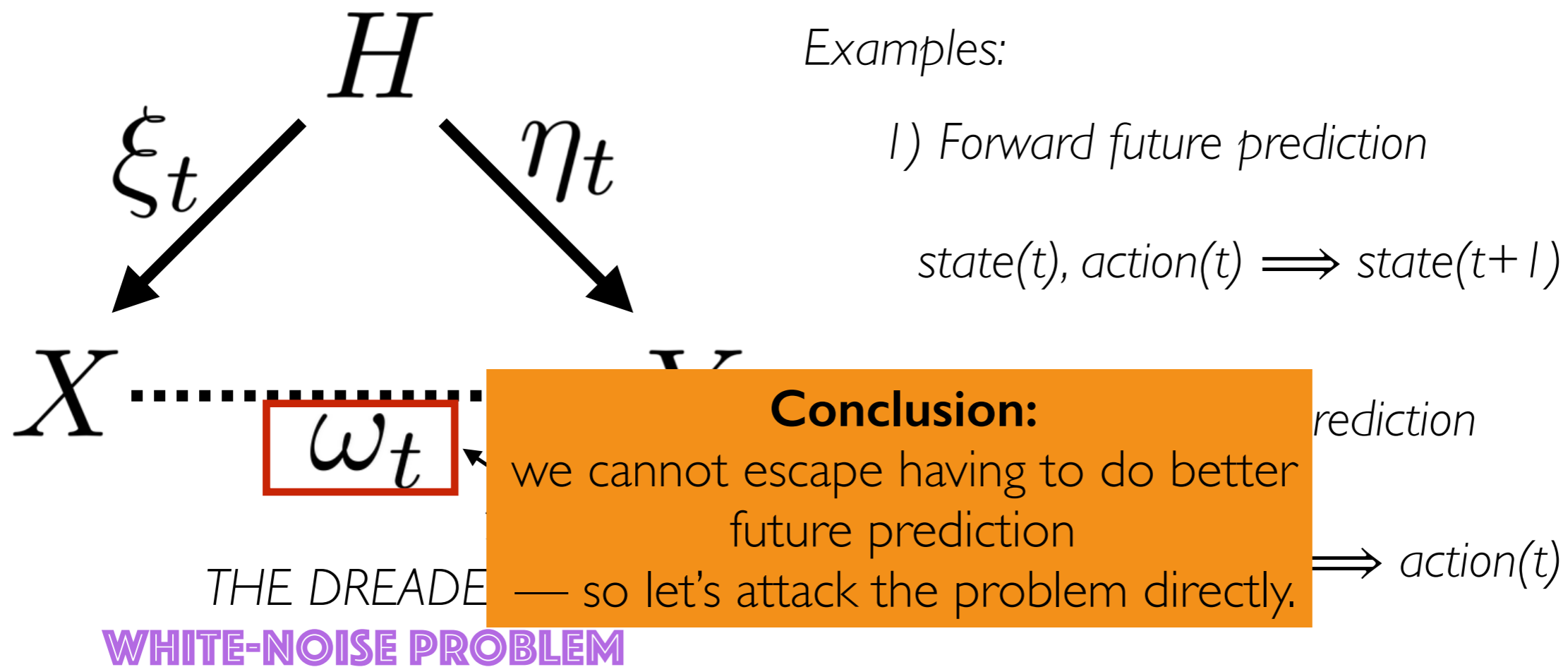
1) is hard, because ... pixel prediction is hard!

2) is mostly what we did in the previous work because it's easier ...

BUT DEGENERATE!

Ex: pushing down on an object

# Self-supervised learning: General Formulation

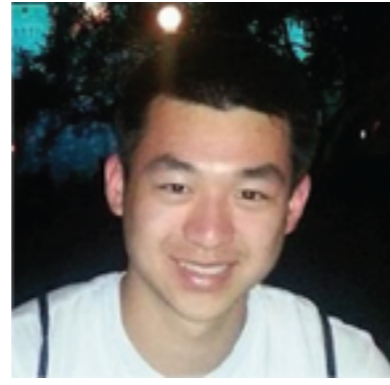


1) is hard, because ... pixel prediction is hard!

2) is mostly what we did in the previous work because it's easier ...

BUT DEGENERATE!

Ex: pushing down on an object



Damian Mrowca\*

Chengxu Zhuang\*

Eli Wang

Nick Haber

Fei-Fei Li

Josh Tenenbaum

NIPS 2018

## Flexible Neural Representation for Physics Prediction

Damian Mrowca<sup>1,\*</sup>, Chengxu Zhuang<sup>2,\*</sup>, Elias Wang<sup>3,\*</sup>, Nick Haber<sup>2,4,5</sup>, Li Fei-Fei<sup>1</sup>,  
Joshua B. Tenenbaum<sup>7,8</sup>, and Daniel L. K. Yamins<sup>1,2,6</sup>

Department of Computer Science<sup>1</sup>, Psychology<sup>2</sup>, Electrical Engineering<sup>3</sup>, Pediatrics<sup>4</sup> and  
Biomedical Data Science<sup>5</sup>, and Wu Tsai Neurosciences Institute<sup>6</sup>, Stanford, CA 94305  
Department of Brain and Cognitive Sciences<sup>7</sup>, and Computer Science and Artificial Intelligence  
Laboratory<sup>8</sup>, MIT, Cambridge, MA 02139

{mrowca, chengxuz, eliwang}@stanford.edu

### Abstract

Humans have a remarkable capacity to understand the physical dynamics of objects in their environment, flexibly capturing complex structures and interactions at multiple levels of detail. Inspired by this ability, we propose a hierarchical particle-based object representation that covers a wide variety of types of three-dimensional objects, including both arbitrary rigid geometrical shapes and deformable materials. We then describe the Hierarchical Relation Network (HRN), an end-to-end differentiable neural network based on hierarchical graph convolution, that learns to predict physical dynamics in this representation. Compared to other neural network baselines, the HRN accurately handles complex collisions and nonrigid deformations, generating plausible dynamics predictions at long time scales in novel settings, and scaling to large scene configurations. These results demonstrate an architecture with the potential to form the basis of next-generation physics predictors for use in computer vision, robotics, and quantitative cognitive science.

t-n ...

-1 ... t+k

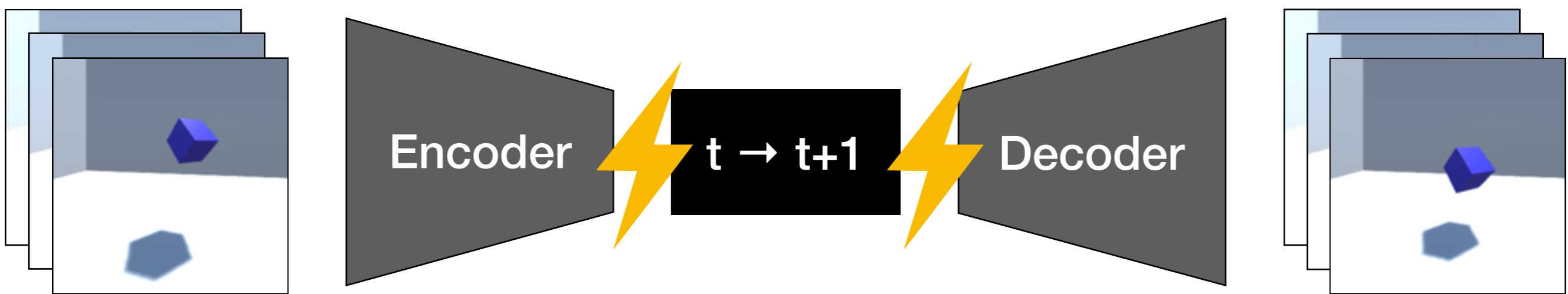


# Discovering the proper latent space for physical prediction...

*“Encoding”*

*“Physics”*

*“Rendering”*



# Intuitive Physics as Underlying Goal



Liz Spelke

Experimental results with infants: **object permanence** present very early, perhaps by 3 months.

# Intuitive Physics as Underlying Goal



Liz Spelke

Experimental results with infants: **object permanence** present very early ...



Cognition

Volume 20, Issue 3, 1985, Pages 191-208



## Object permanence in five-month-old infants ☆

Renée Baillargeon \*, Elizabeth S. Spelke \*, Stanley Wasserman \*

 [Show more](#)

[https://doi.org/10.1016/0010-0277\(85\)90008-3](https://doi.org/10.1016/0010-0277(85)90008-3)

[Get rights and content](#)

# Intuitive Physics as Underlying Goal



Liz Spelke

Experimental results with infants: **object permanence** present very early ...



Cognition  
Volume 20, Issue 3, 1985, Pages 191-208



## Object permanence in five-month-old infants ☆

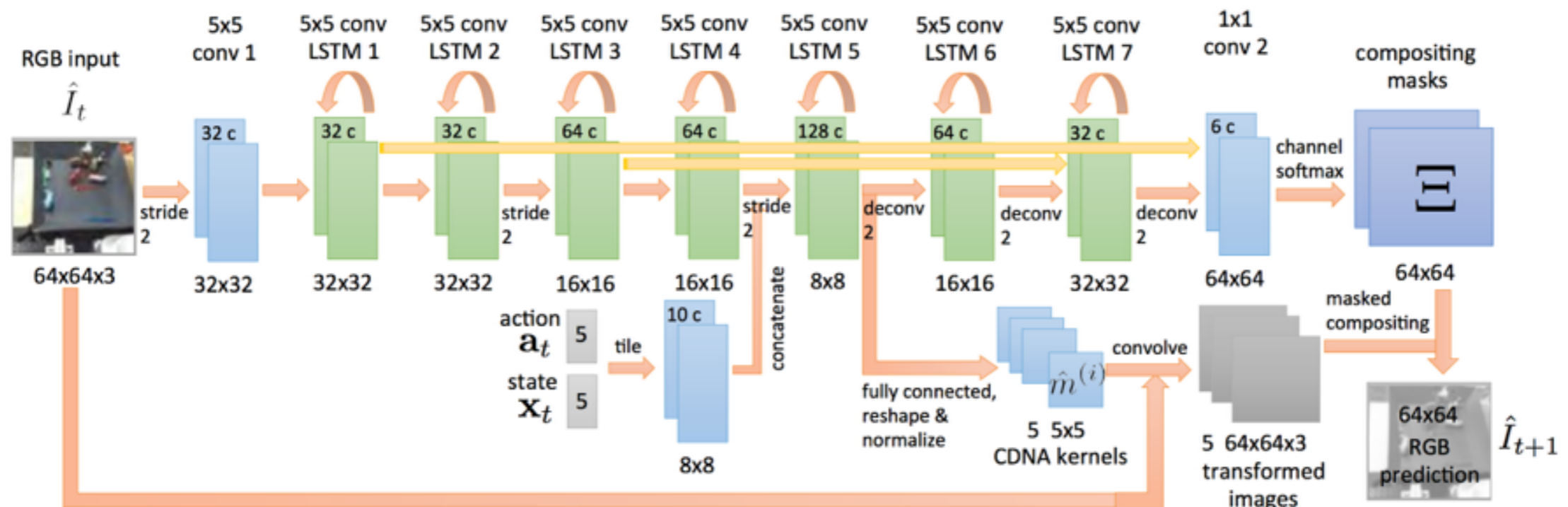
Renée Baillargeon <sup>✉</sup>, Elizabeth S. Spelke <sup>\*</sup>, Stanley Wasserman <sup>\*</sup>

Show more

[https://doi.org/10.1016/0010-0277\(85\)90008-3](https://doi.org/10.1016/0010-0277(85)90008-3)

[Get rights and content](#)

Conv2d structures, even with RNNs, have trouble with object permanence.



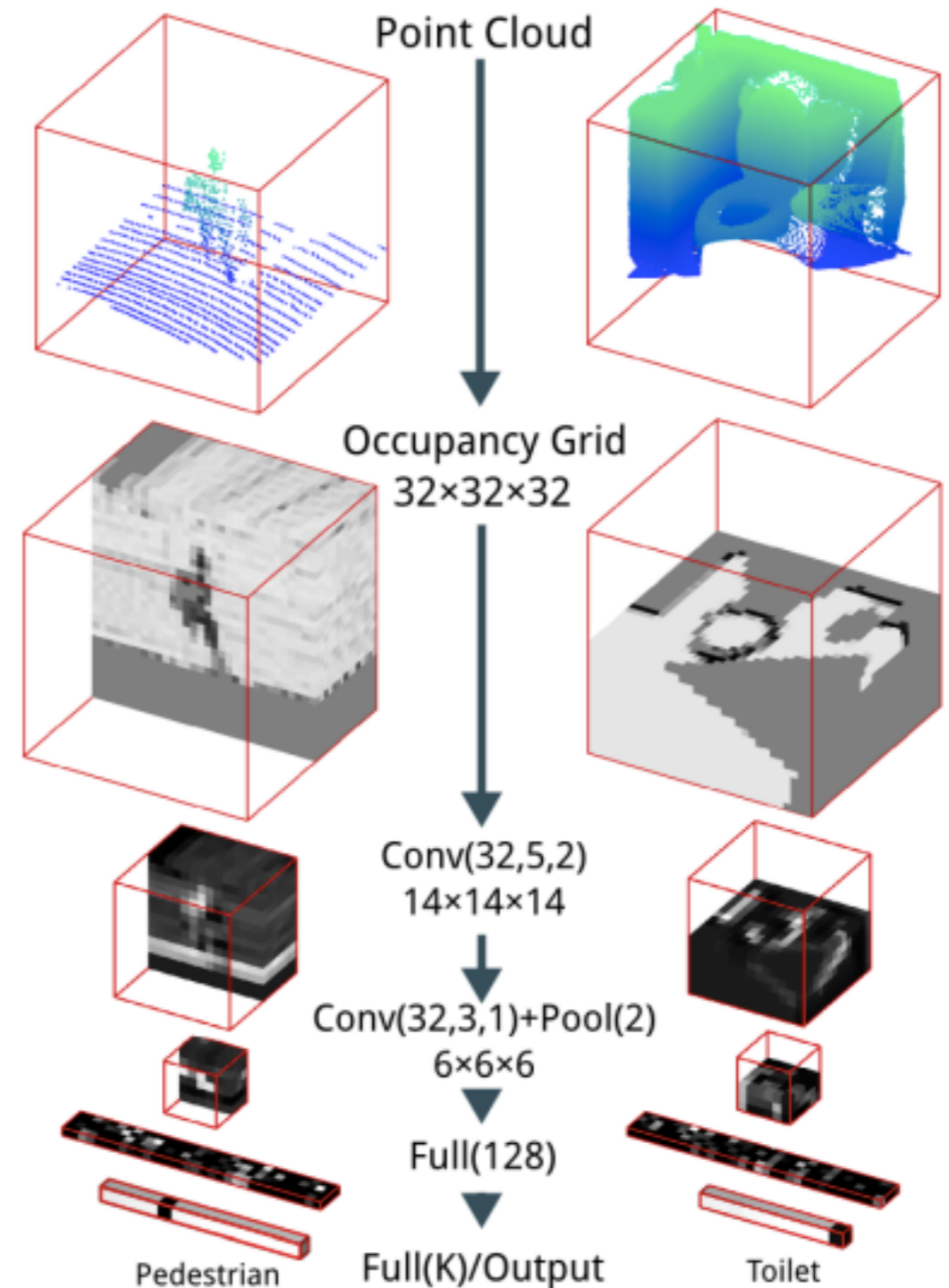
# Intuitive Physics as Underlying Goal



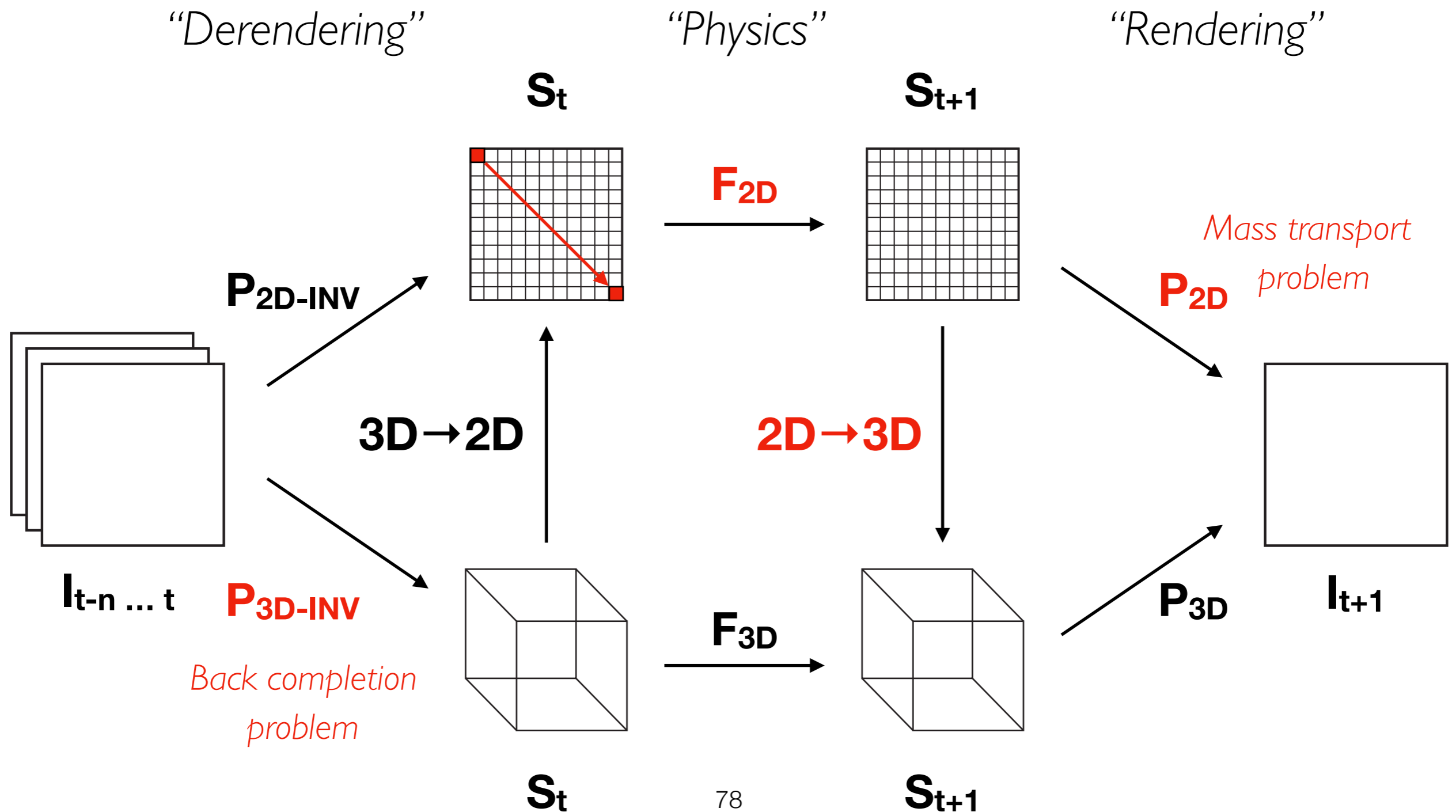
Liz Spelke

Experimental results with infants: **object permanence** present very early, perhaps by 3 months.

Conv3d structures are better for object permanence, but very inefficient: hard to achieve high resolution.



# Spatial convolutions are not ideal for physics propagation



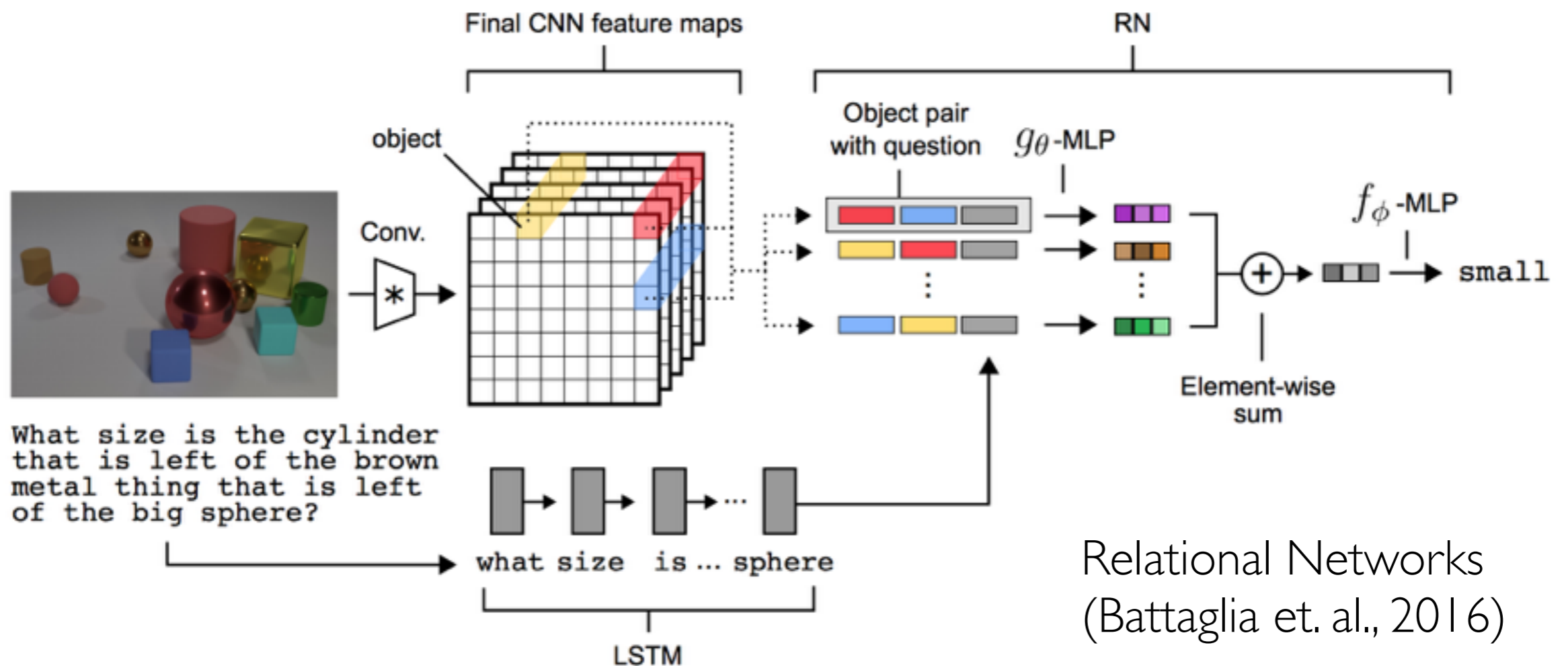
# Intuitive Physics as Underlying Goal



Liz Spelke

Experimental results with infants: **object permanence** present very early, perhaps by 3 months.

Alternative to spatially-uniform priors are **graph-based** priors



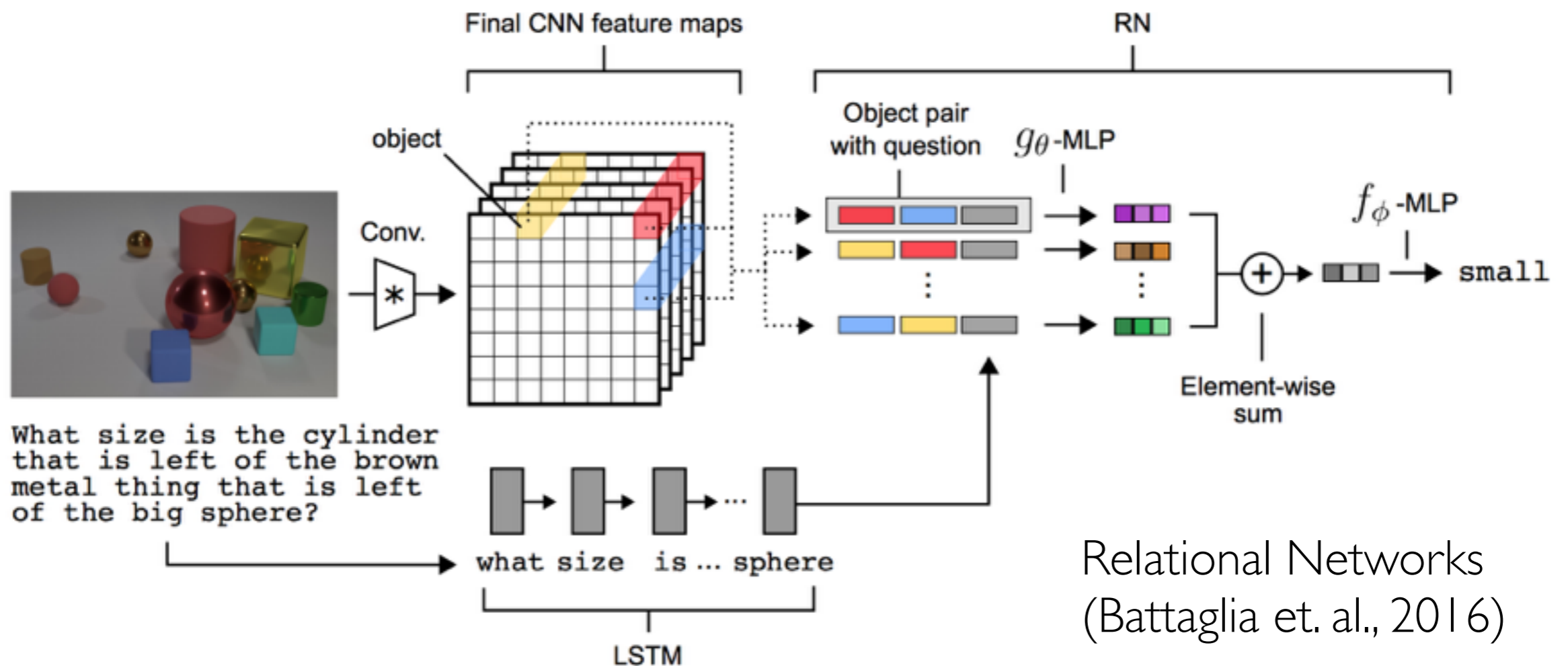
# Intuitive Physics as Underlying Goal



Liz Spelke

Experimental results with infants: **object permanence** present very early, perhaps by 3 months.

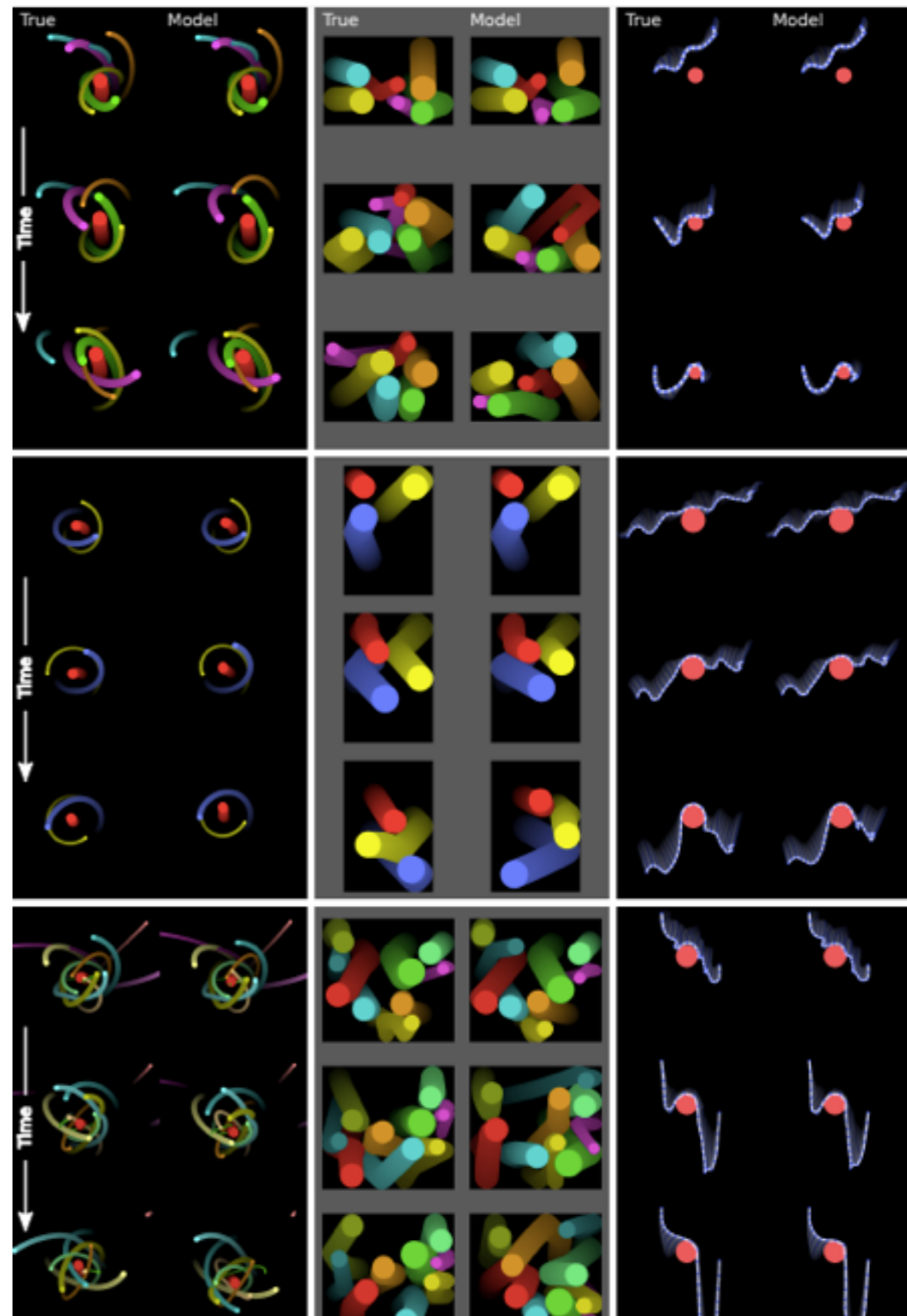
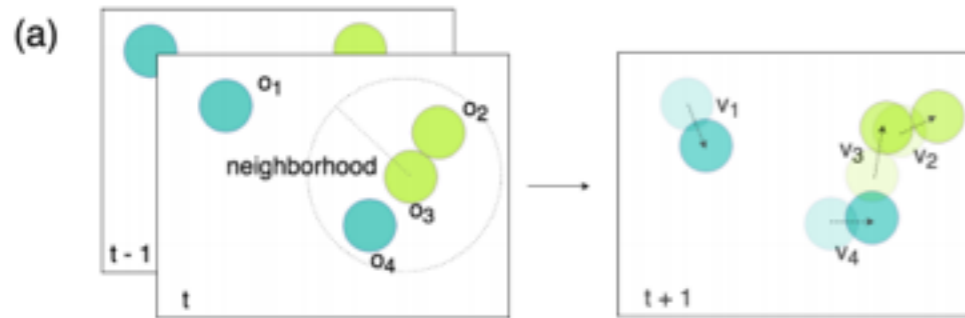
Alternative to spatially-uniform priors are **graph-based** priors  
... still local and convolutional, just on the graph.





# Relational Networks (Battaglia et. al., 2016)

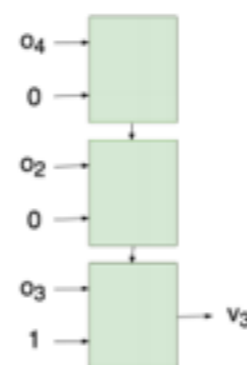
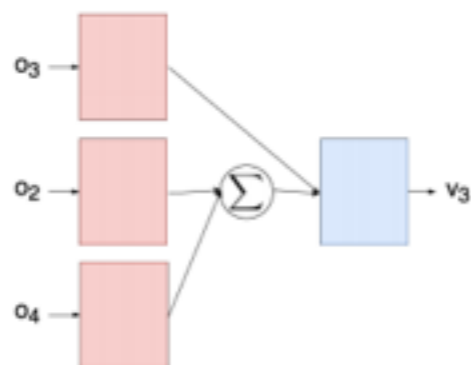
# Neural Physics Engine (Chang et. al., 2016)



(b) NPE applied on object 3

(c) NP applied on object 3

(d) LSTM applied on object 3



# Intuitive Physics as Underlying Goal

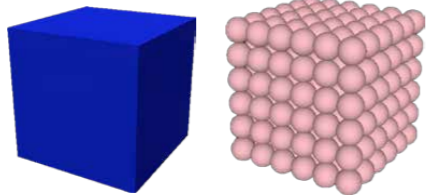
Complex Scenes



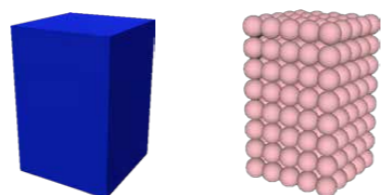
Complex Materials



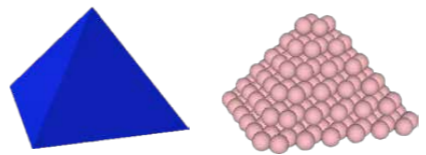
# Describe objects through complex graphs:



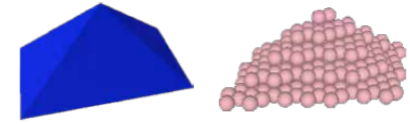
Cube



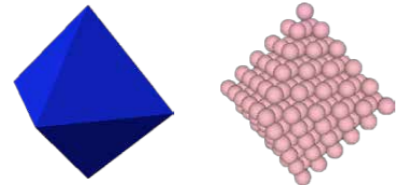
Cuboid



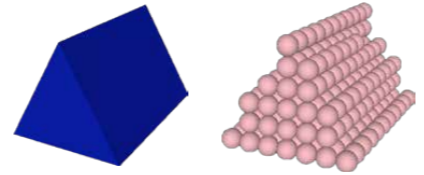
Pyramid



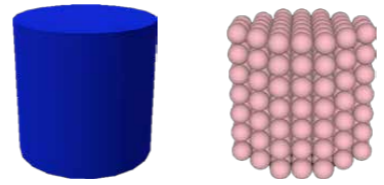
Flat Pyramid



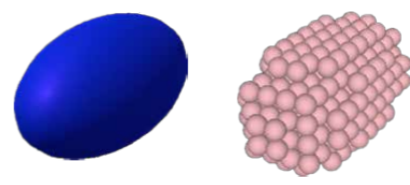
Octahedron



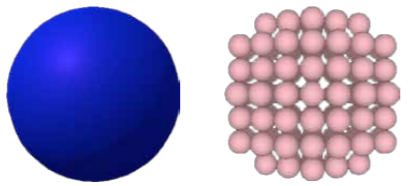
Prism



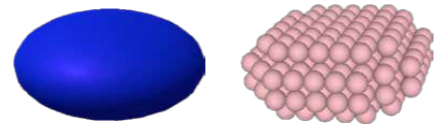
Cylinder



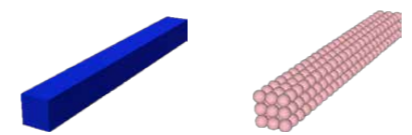
Ellipsoid



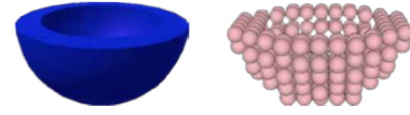
Sphere



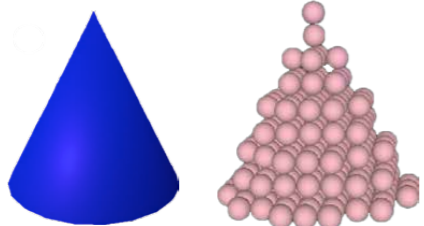
Mentos



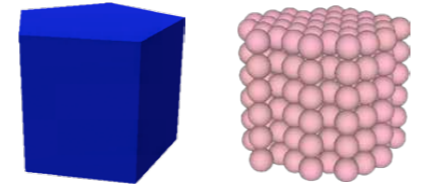
Stick



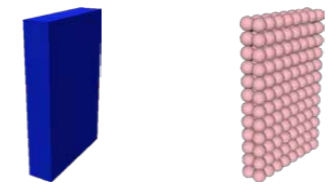
Bowl



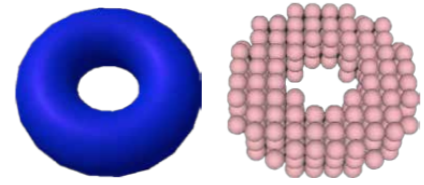
Cone



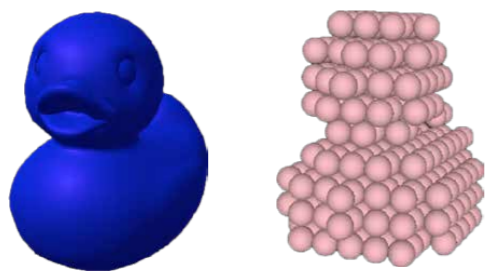
Pentagon



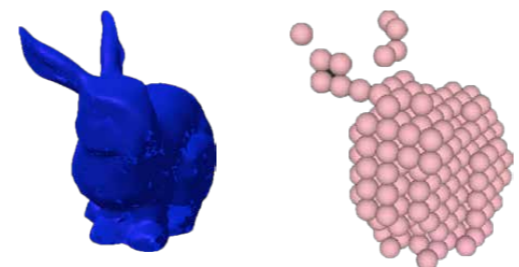
Domino



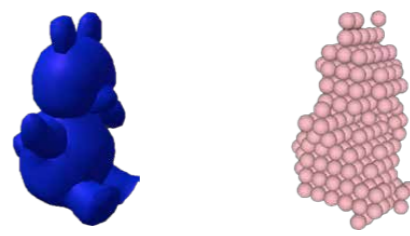
Torus



Duck

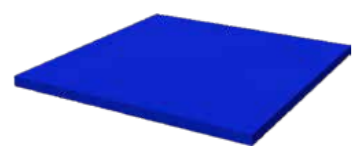


Bunny

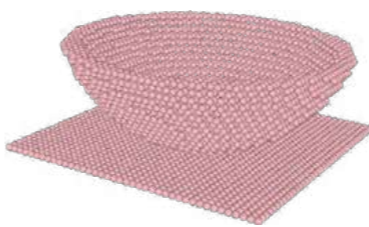
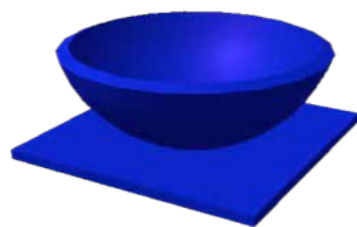


Teddy

In fact, describe whole scenes.



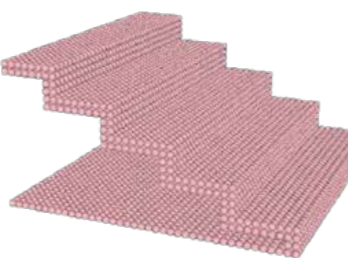
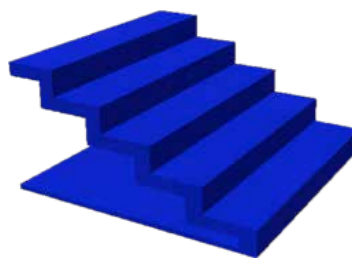
Plane



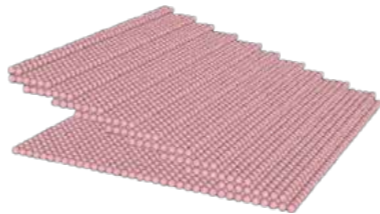
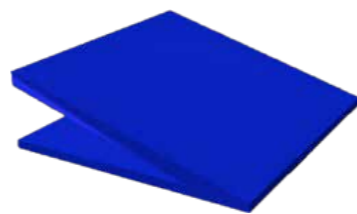
Bowl



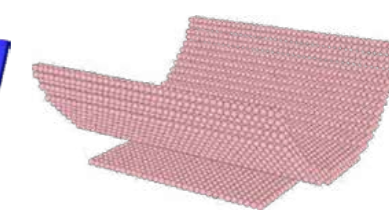
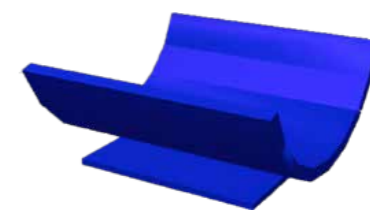
Random Plane



Stairs

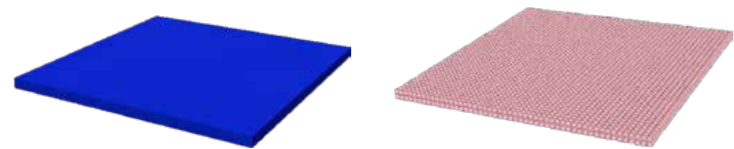


Slope

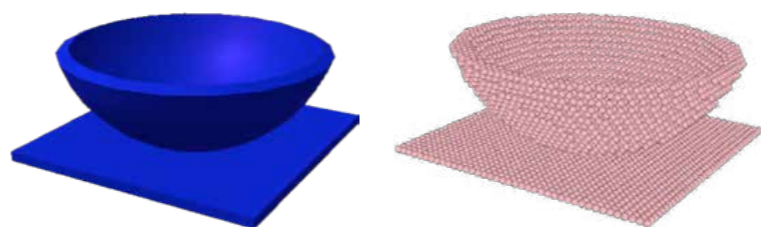


Half-Pipe

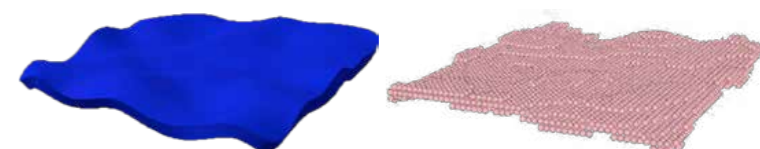
In fact, describe whole scenes.



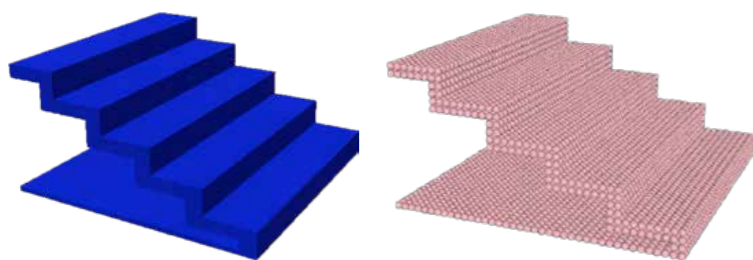
Plane



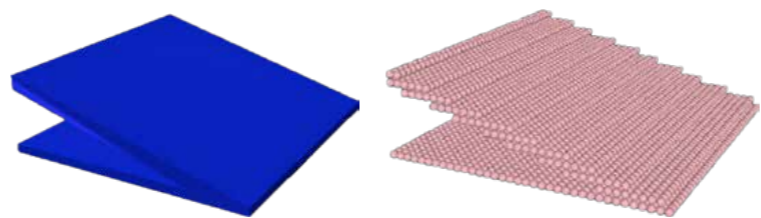
Bowl



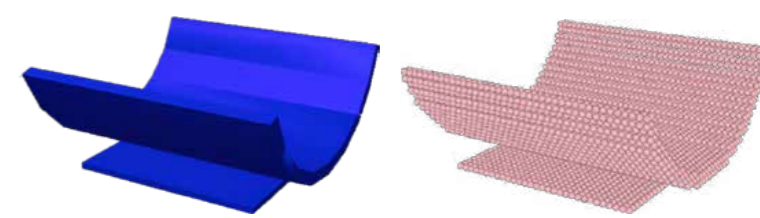
Random Plane



Stairs



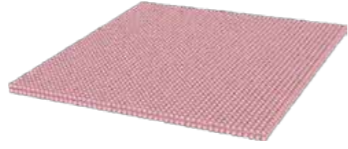
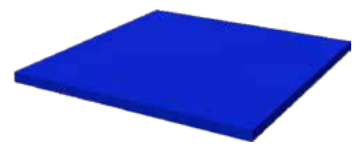
Slope



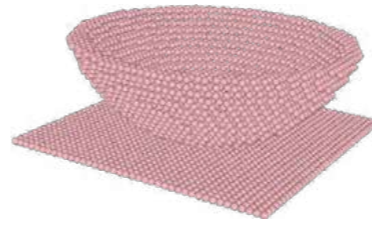
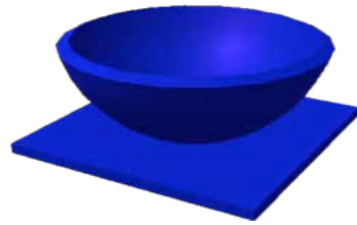
Half-Pipe

$G = \langle N, E \rangle$  scene graph

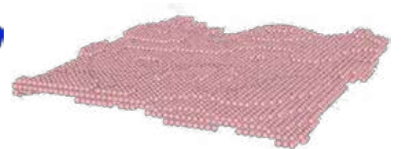
In fact, describe whole scenes.



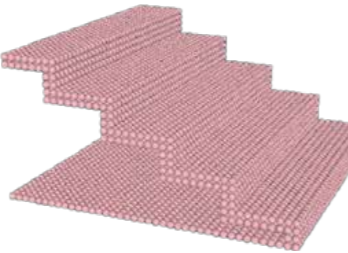
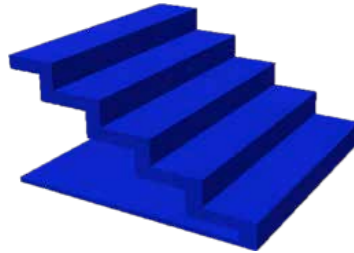
Plane



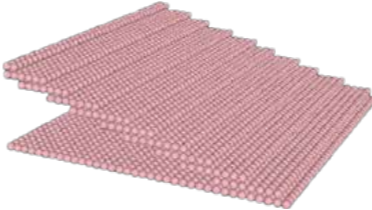
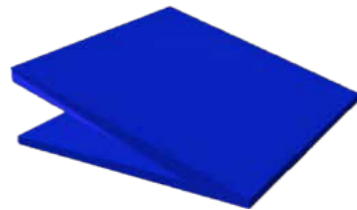
Bowl



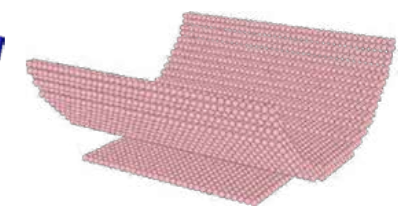
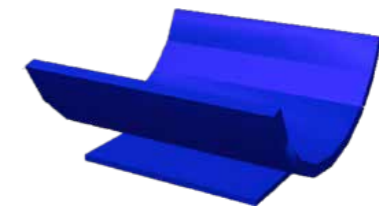
Random Plane



Stairs



Slope



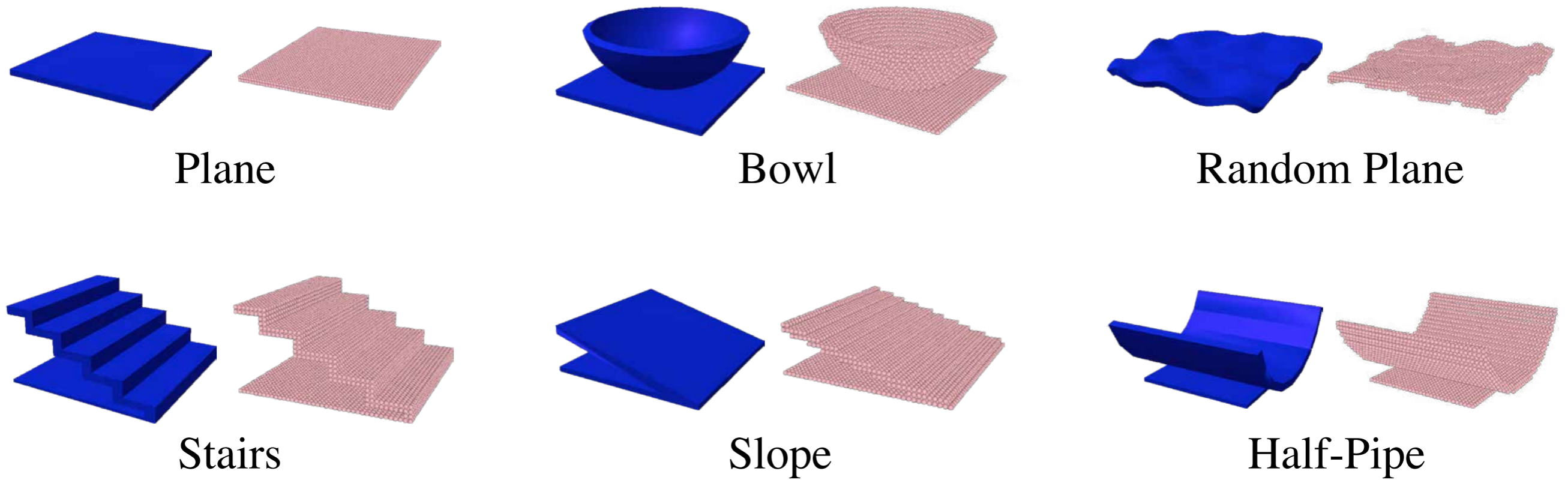
Half-Pipe

$G = \langle N, E \rangle$  scene graph

$N$  = nodes corresponding to particles comprising objects

$E$  = edges corresponding to relationships between particles

In fact, describe whole scenes.



$$G = \langle N, E \rangle \text{ scene graph}$$

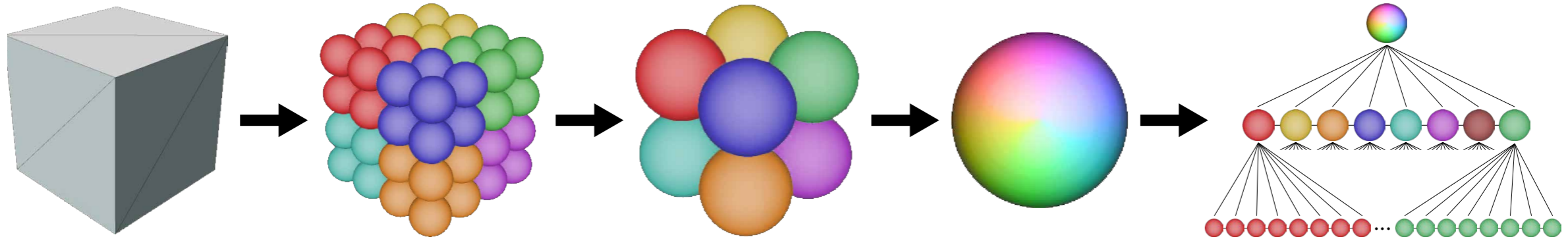
$N$  = nodes corresponding to particles comprising objects

$E$  = edges corresponding to relationships between particles

edges are labelled by vector capturing bond characteristics

# Intuitive Physics as Underlying Goal

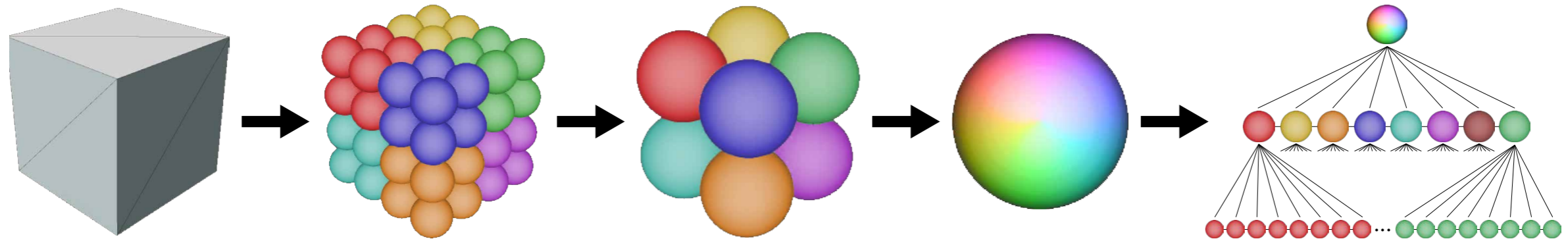
Of course, humans don't think about all the particles at once all the time.





# Intuitive Physics as Underlying Goal

Of course, humans don't think about all the particles at once all the time.

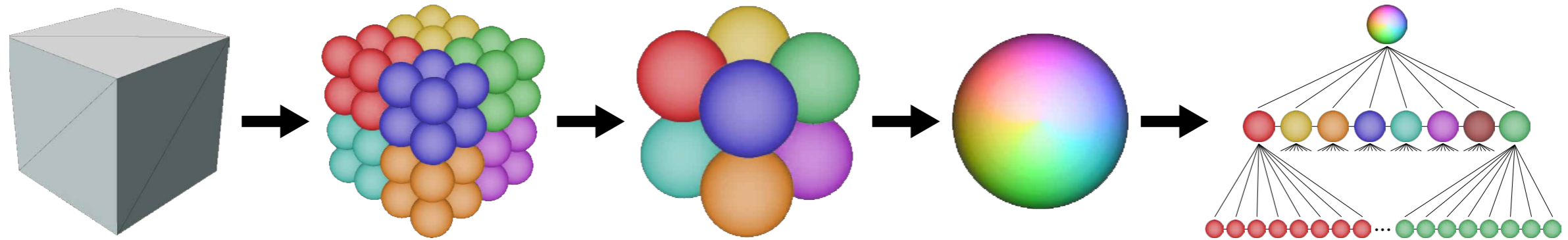


$$G \mapsto G_H$$

$G_H$  = dynamic “hierarchicalization” of underlying scene graph  
(right now computed via k-means)

# Intuitive Physics as Underlying Goal

Of course, humans don't think about all the particles at once all the time.

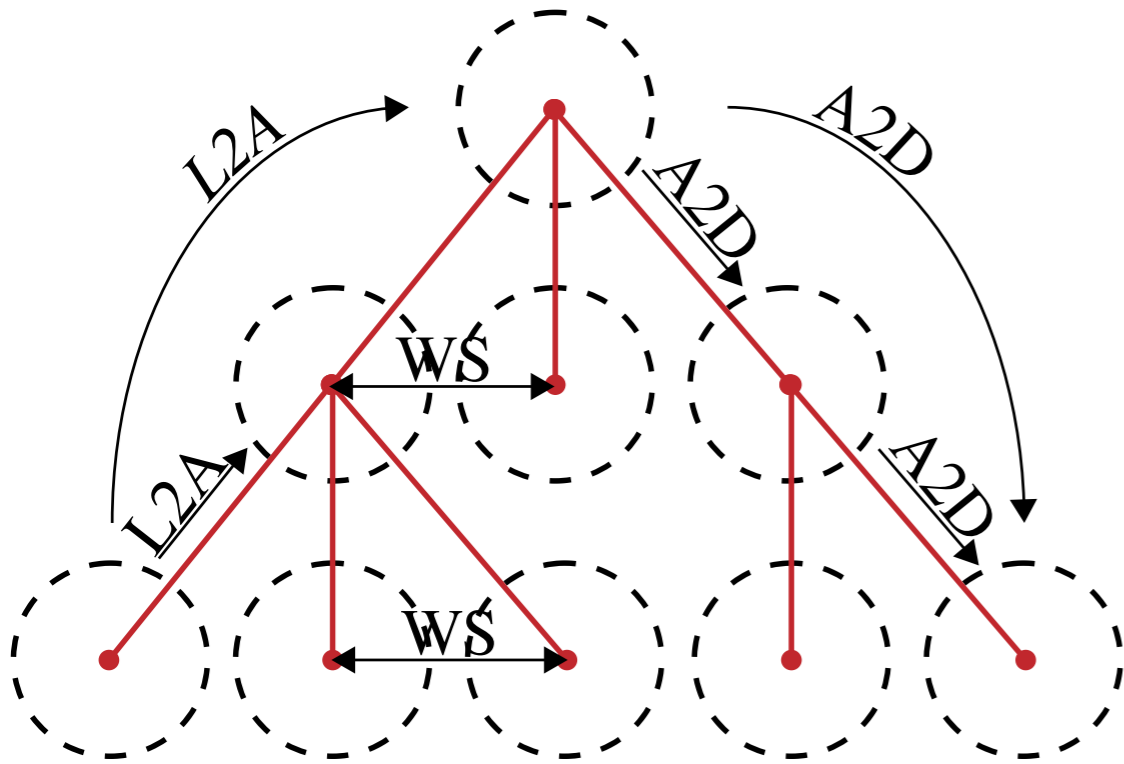


$$G \mapsto G_H$$

$G_H$  = dynamic “hierarchicalization” of underlying scene graph  
(right now computed via *k-means*)

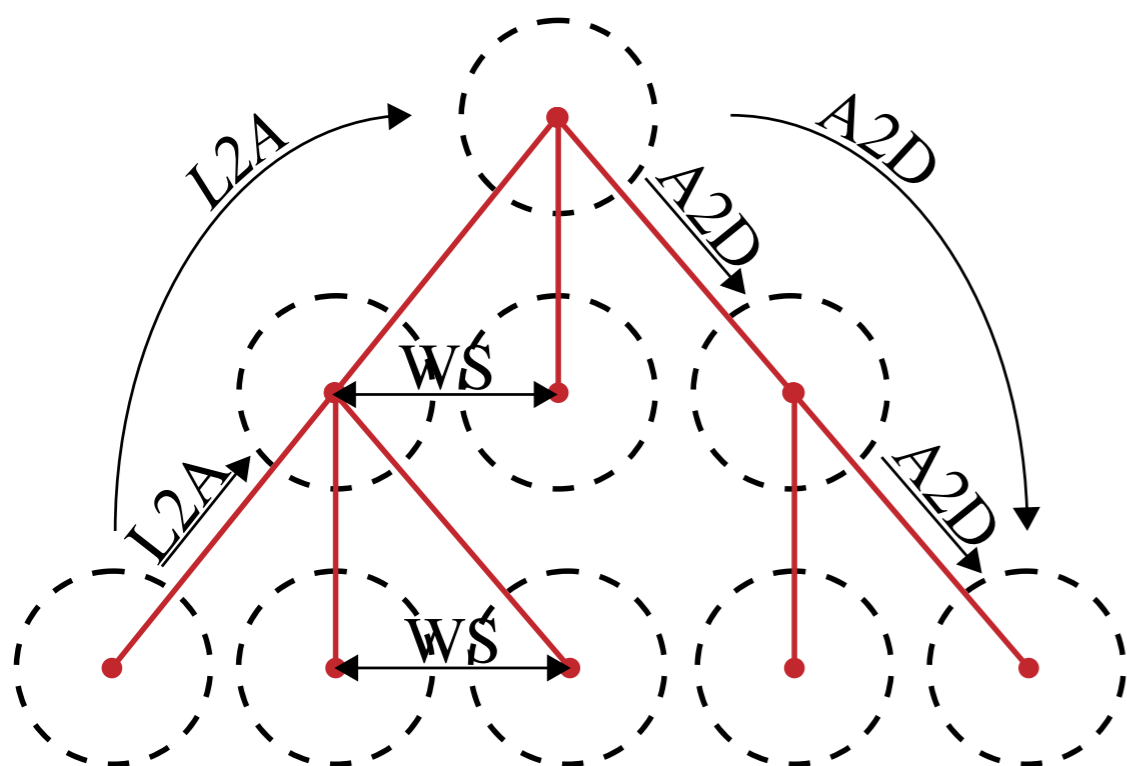
graph convolution  $\rightarrow$  hierarchical graph convolution

# Intuitive Physics as Underlying Goal

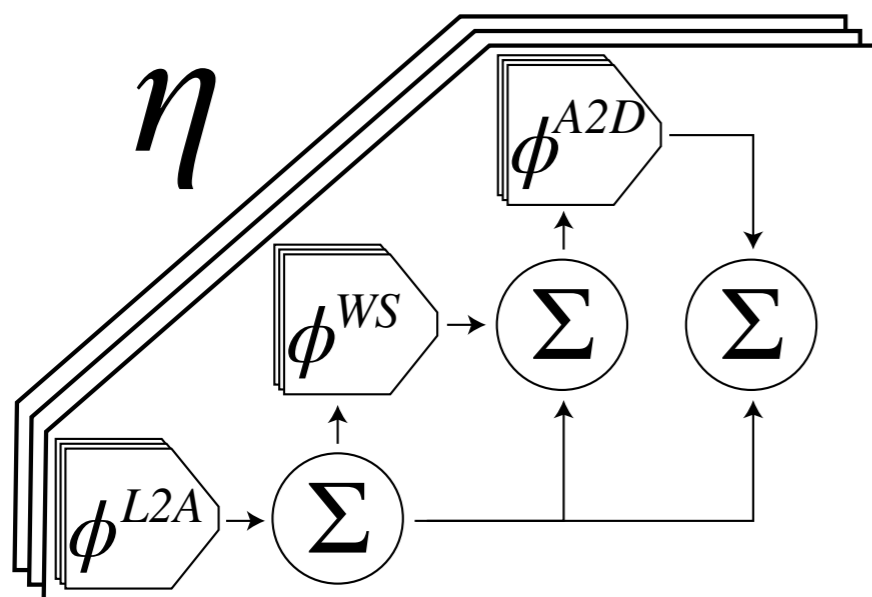


- $\phi^{L2A}$  graph conv. leaves to ancestors
- $\phi^{WS}$  graph conv. with siblings
- $\phi^{A2D}$  graph conv. ancestors to descendants

# Intuitive Physics as Underlying Goal

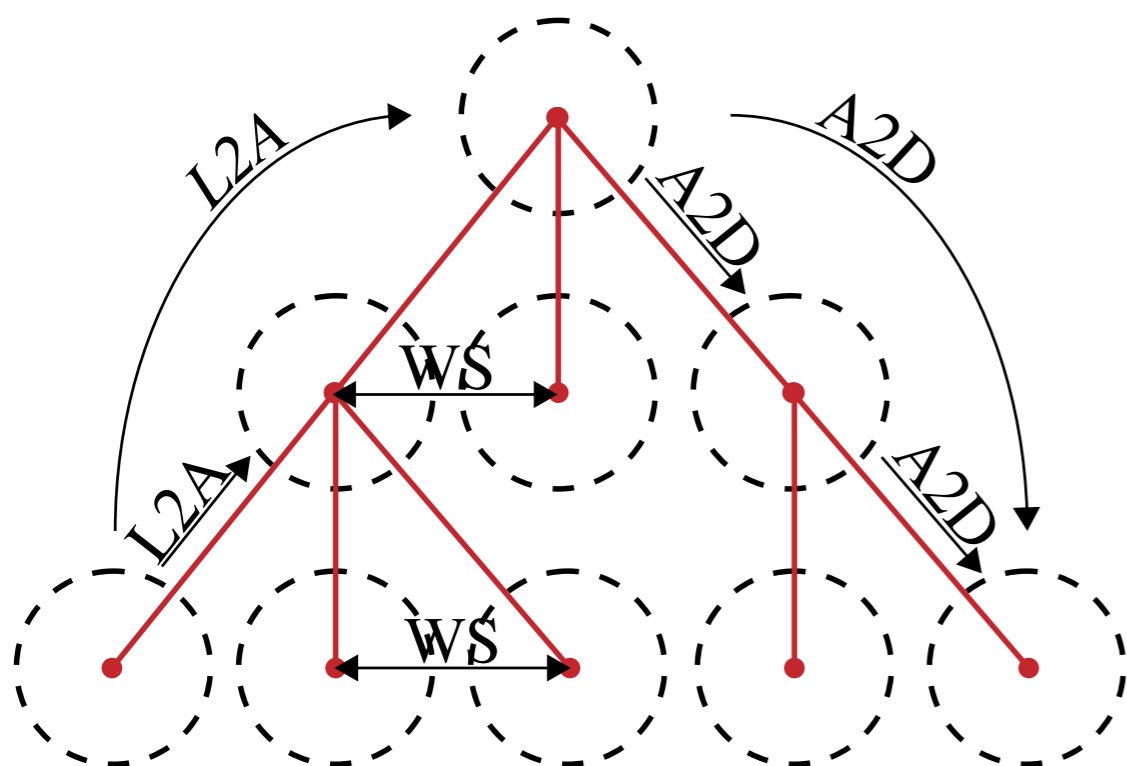


- $\phi^{L2A}$  graph conv. leaves to ancestors
- $\phi^{WS}$  graph conv. with siblings
- $\phi^{A2D}$  graph conv. ancestors to descendants

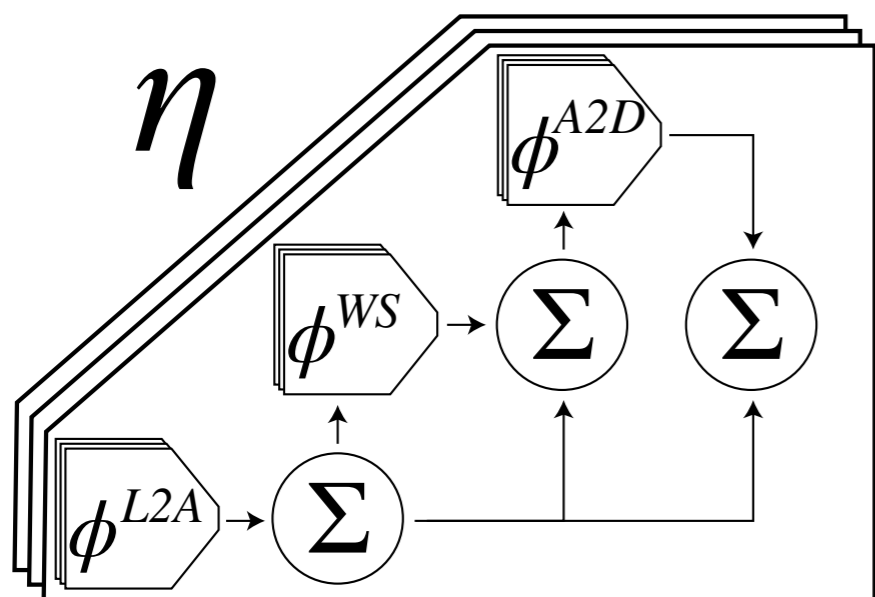


- $\eta$  module composing these three operations from one up-down cycle, adding physical effects

# Intuitive Physics as Underlying Goal



- $\phi^{L2A}$  graph conv. leaves to ancestors
- $\phi^{WS}$  graph conv. with siblings
- $\phi^{A2D}$  graph conv. ancestors to descendants

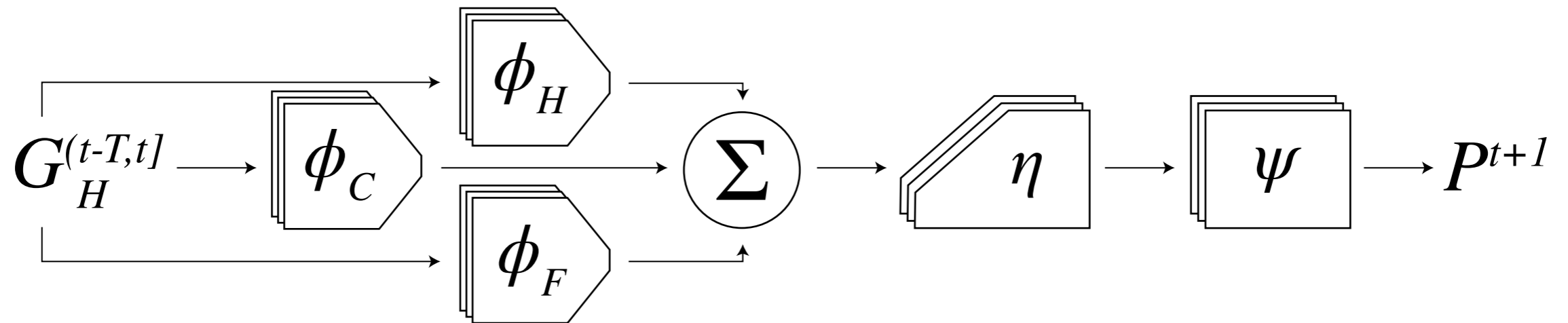


- $\eta$  module composing these three operations from one up-down cycle, adding physical effects

Hierarchical graph convolution propagates interactions efficiently

# Intuitive Physics as Underlying Goal

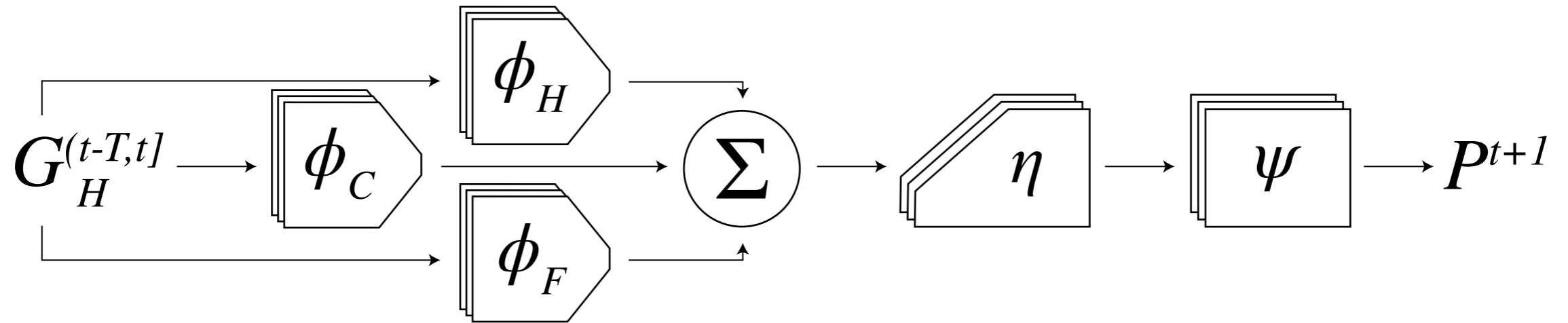
Hierarchical Relational Network (HRN):



... generates momentum updates ( $\mathbf{P}$ ) from hierarchical graph state ( $\mathbf{G}$ ).

# Intuitive Physics as Underlying Goal

Hierarchical Relational Network (HRN):

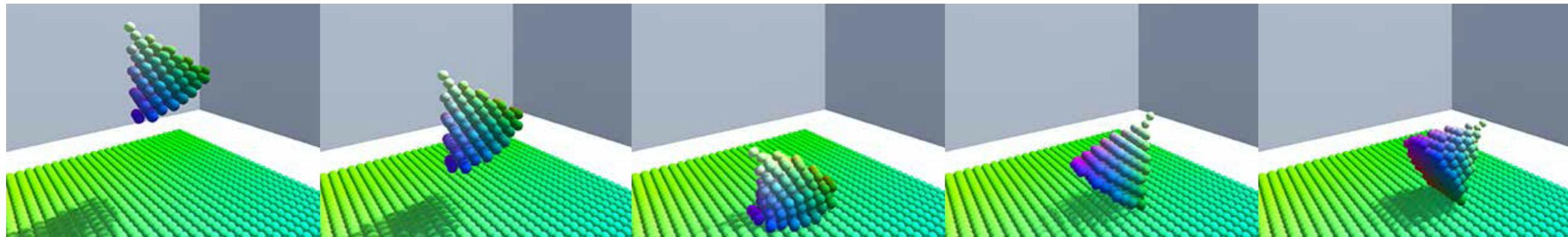


... generates momentum updates ( $\mathbf{P}$ ) from hierarchical graph state ( $\mathbf{G}$ ).

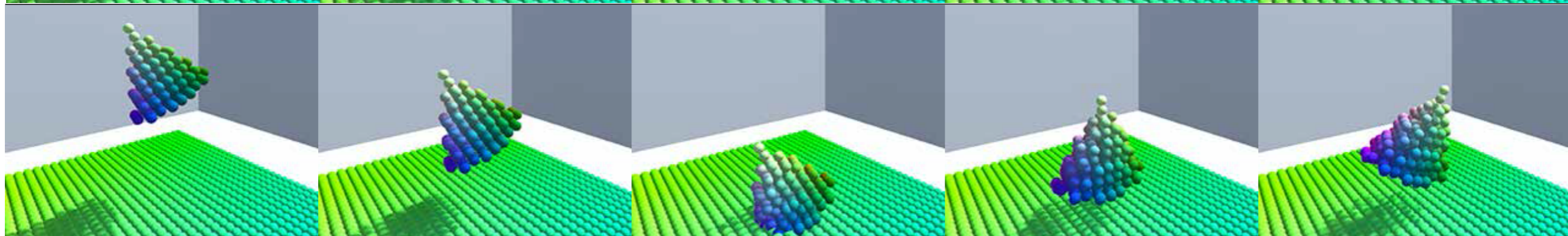
Network learns to interpret graph structure (including meaning of material-vector edge labels)...

# Deformable cone bouncing off a flat floor

Ground Truth



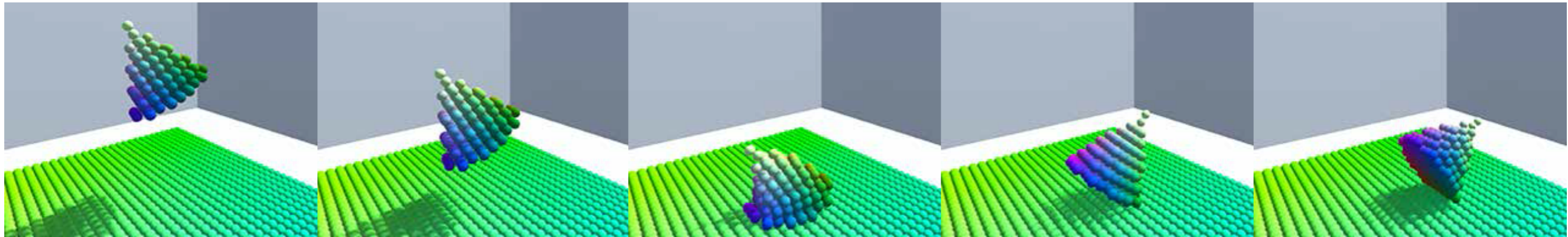
Prediction



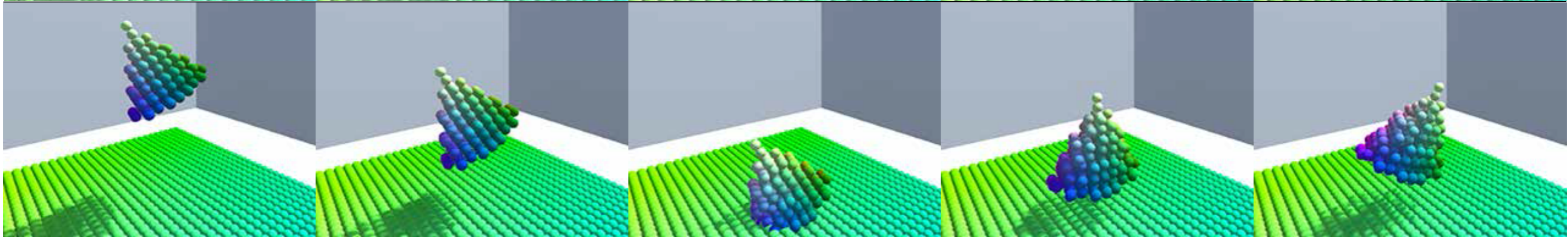


# Deformable cone bouncing off a flat floor

Ground Truth

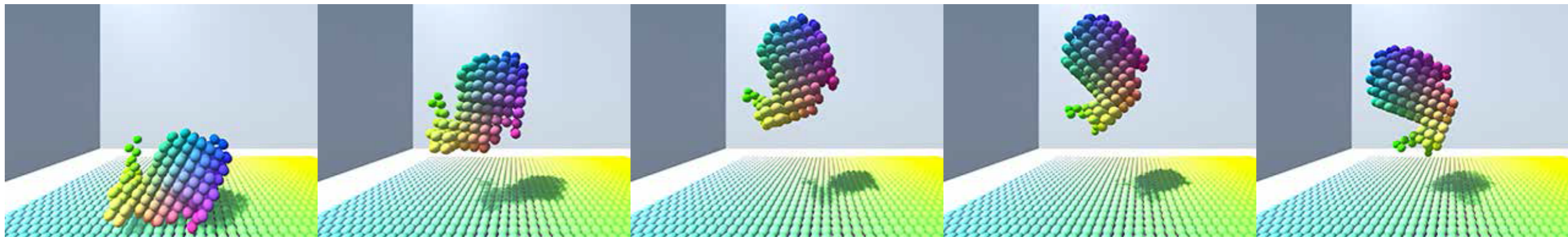


Prediction

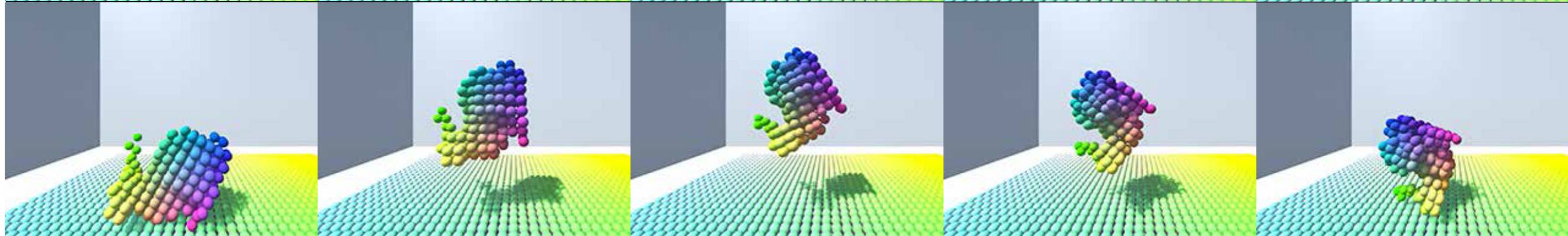


# Stanford bunny

Ground Truth

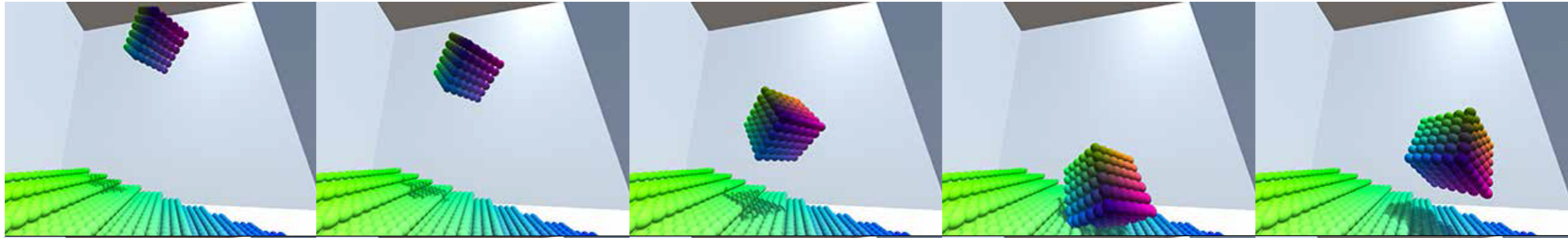


Prediction

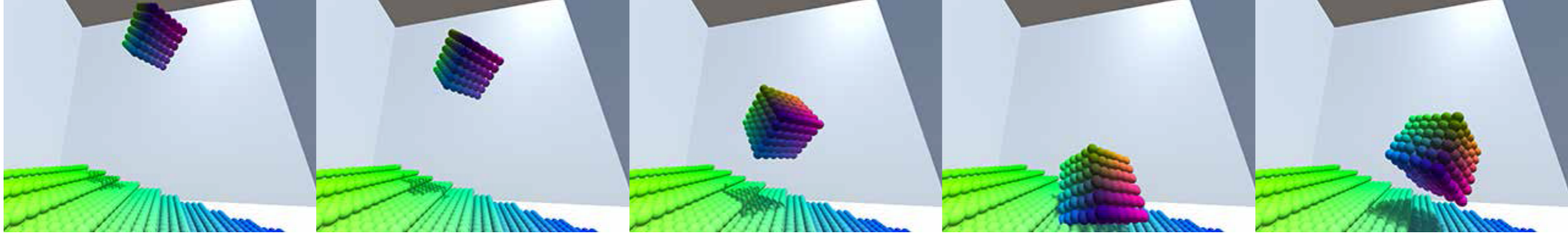


# Deformable box bouncing off an incline

Ground Truth

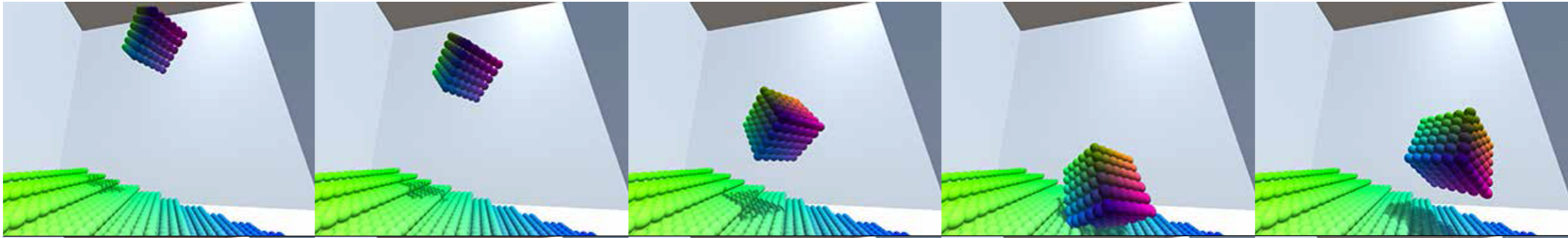


Prediction

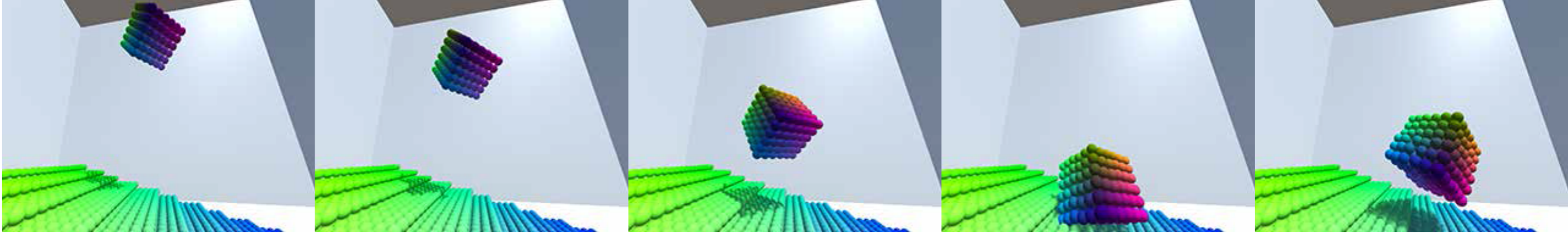


# Deformable box bouncing off an incline

Ground Truth

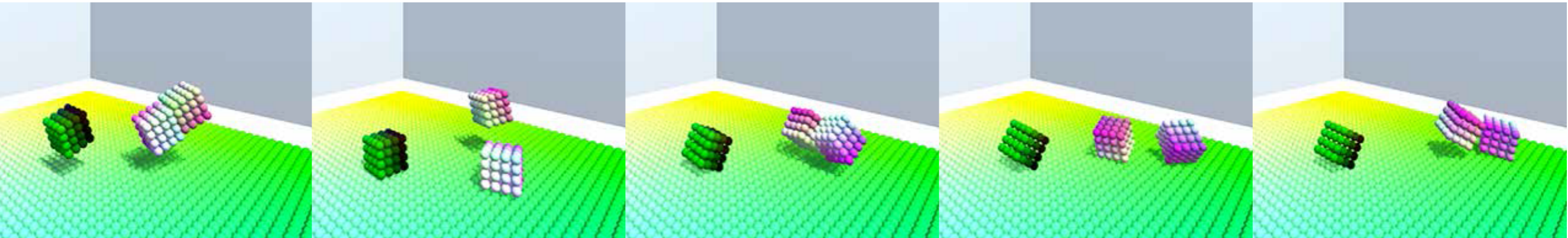


Prediction

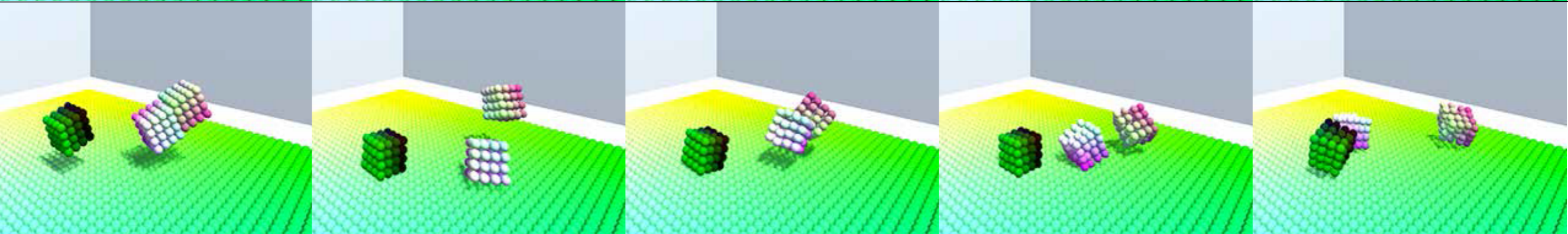


# Multiple rigid objects colliding

Ground Truth

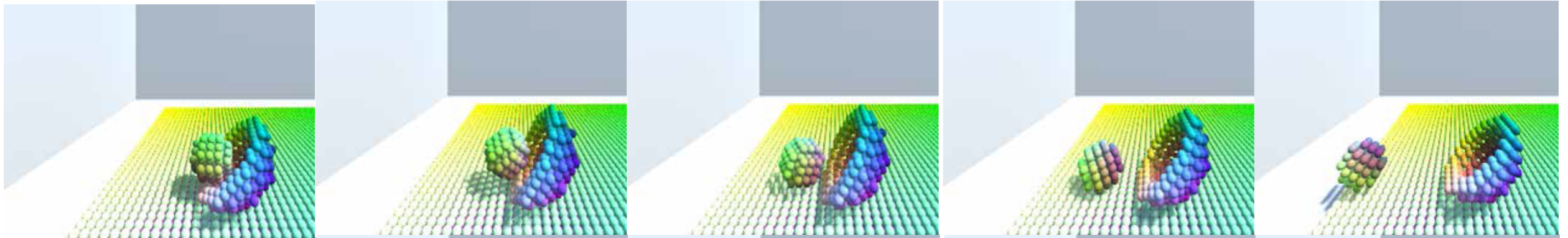


Prediction

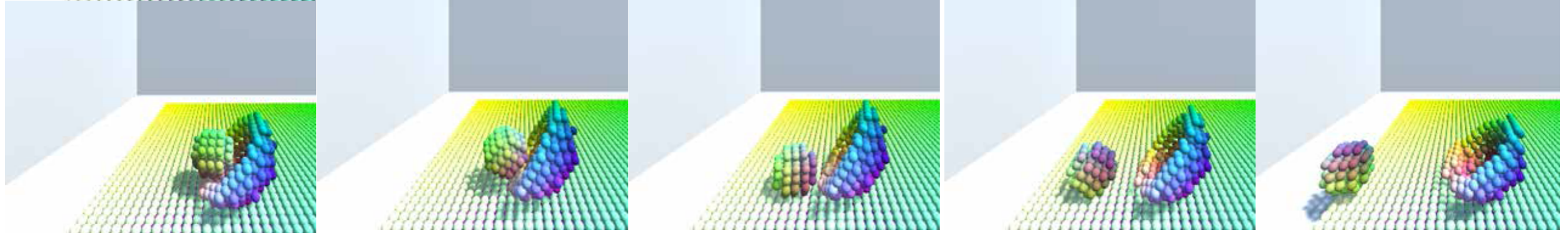


# rigid sphere rolling out of rigid bowl

Ground truth

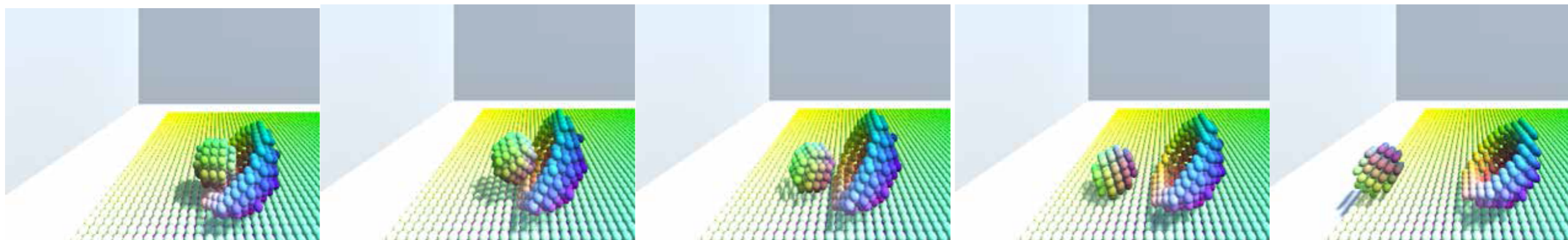


Prediction

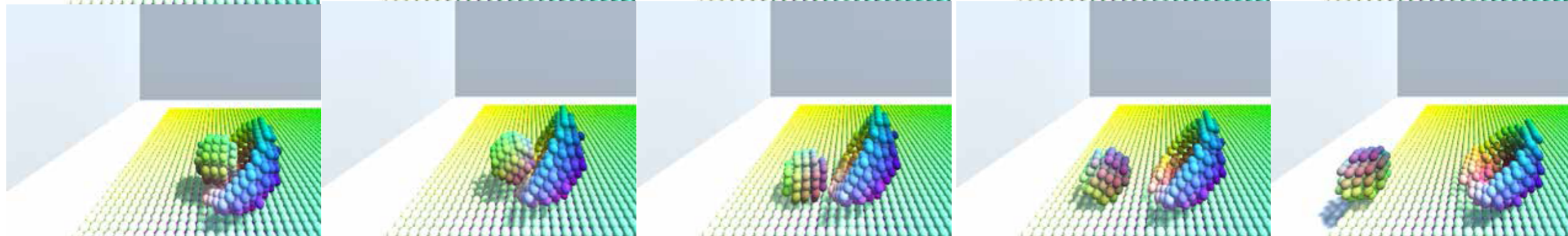


rigid sphere rolling out of rigid bowl

Ground truth

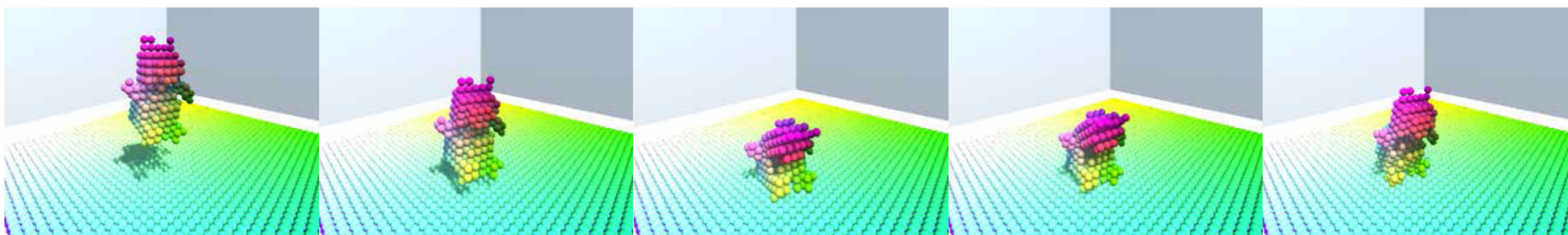


Prediction

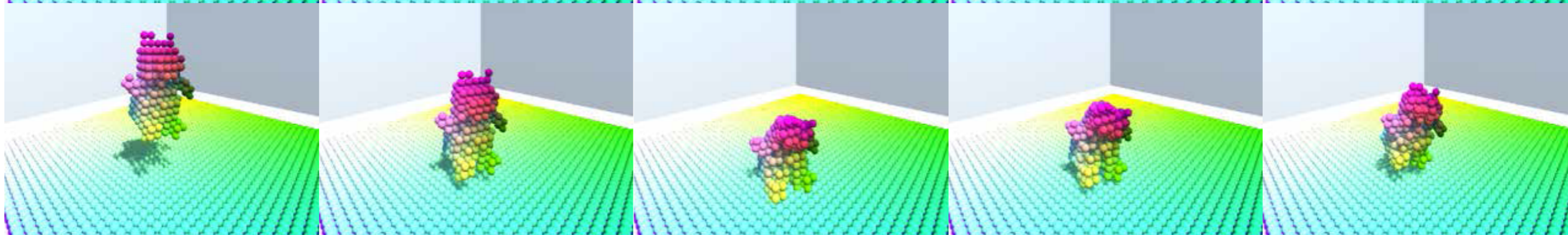


floppy teddybear bouncing off floor and recovering

Ground truth

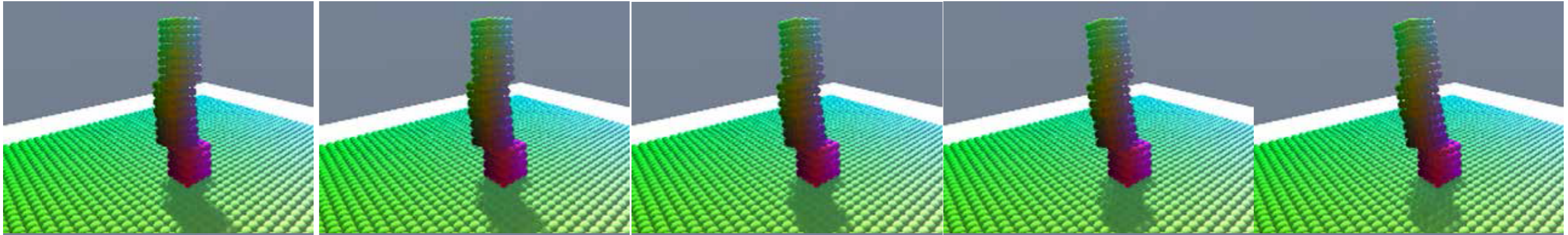


Prediction

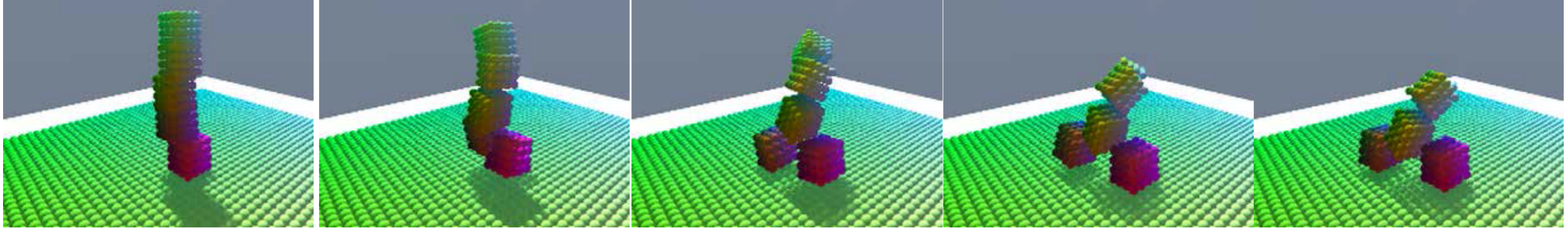


# knocking over an unstable block tower

Ground truth



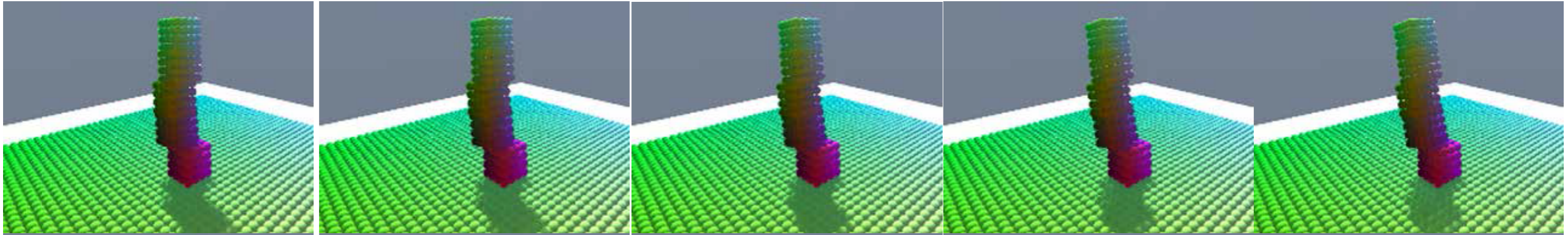
Prediction



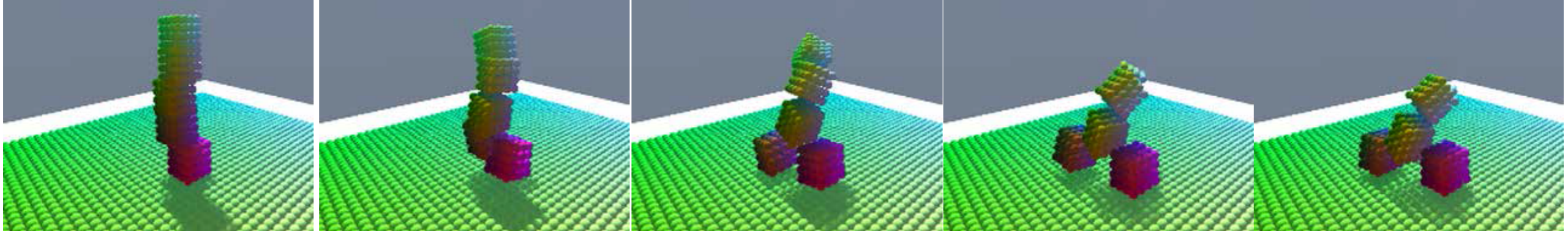
*in GT the tower does fall, but prediction falls too fast ...*

# knocking over an unstable block tower

Ground truth

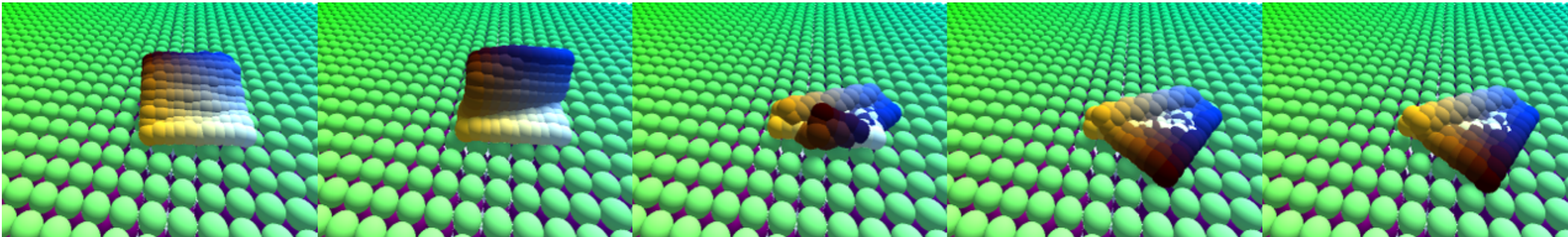


Prediction

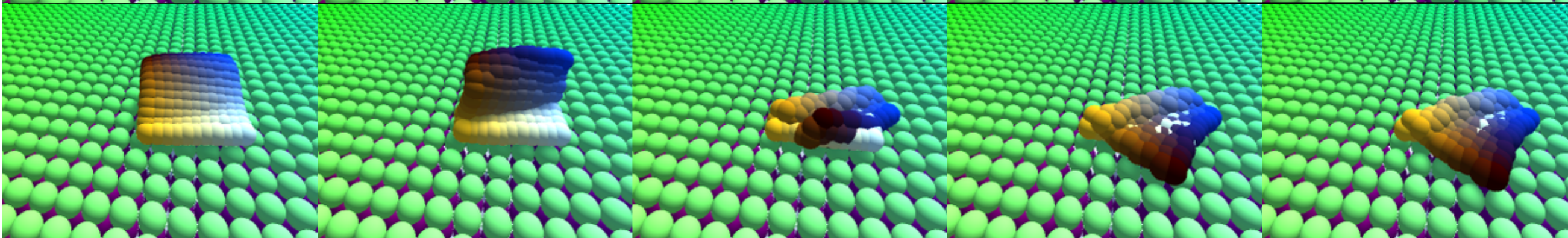


# Folding Cloth

Ground Truth



Prediction



t+1

t+3

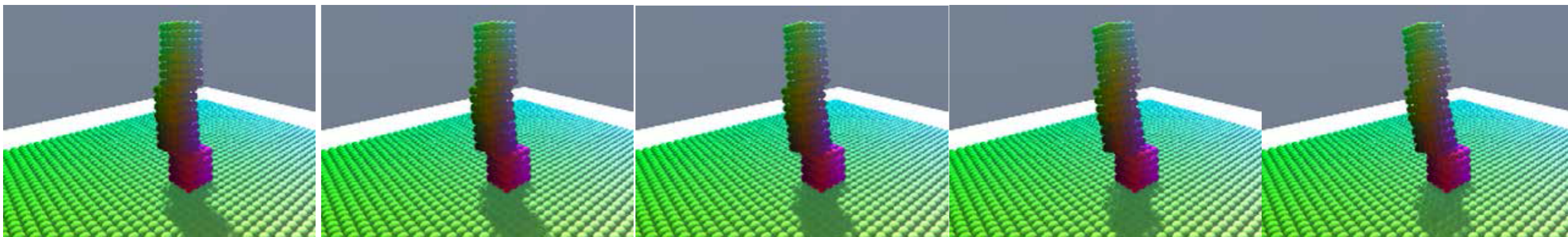
t+5

t+7

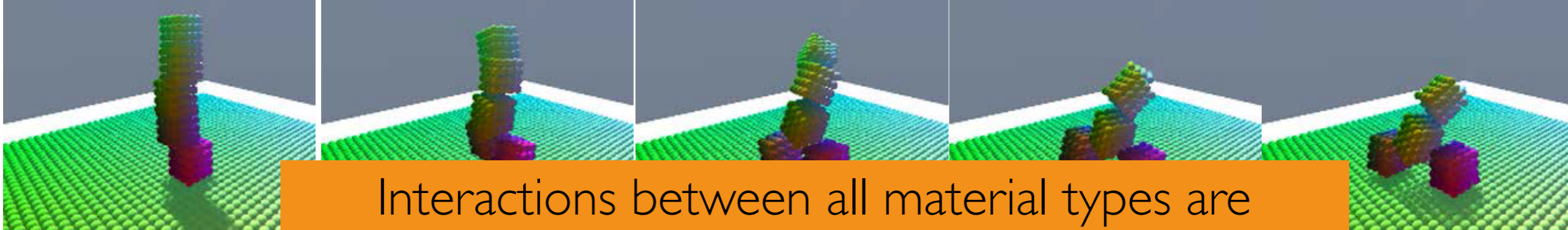
t+9

# knocking over an unstable block tower

Ground truth

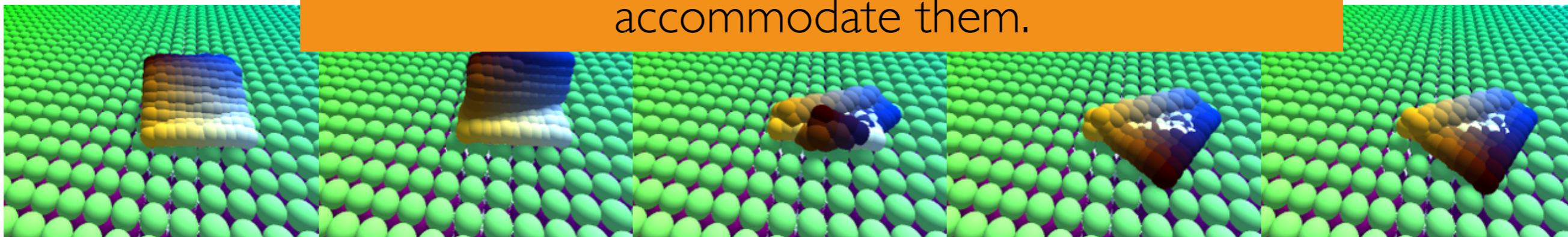


Prediction

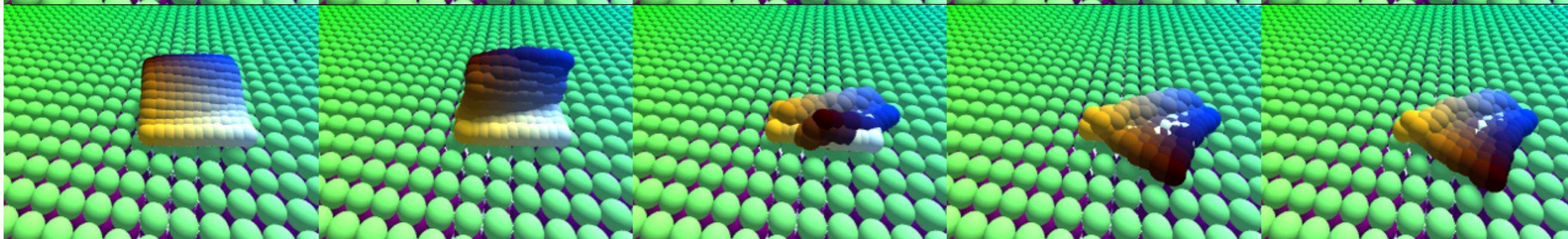


Interactions between all material types are possible — as well as non-uniform materials — since the edge-labelled graph structure can accommodate them.

Ground Truth



Prediction



t+1

t+3

t+5

t+7

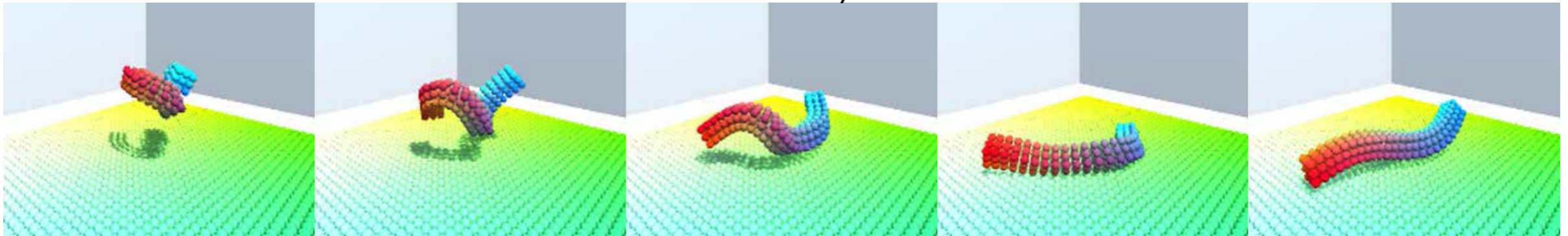
t+9



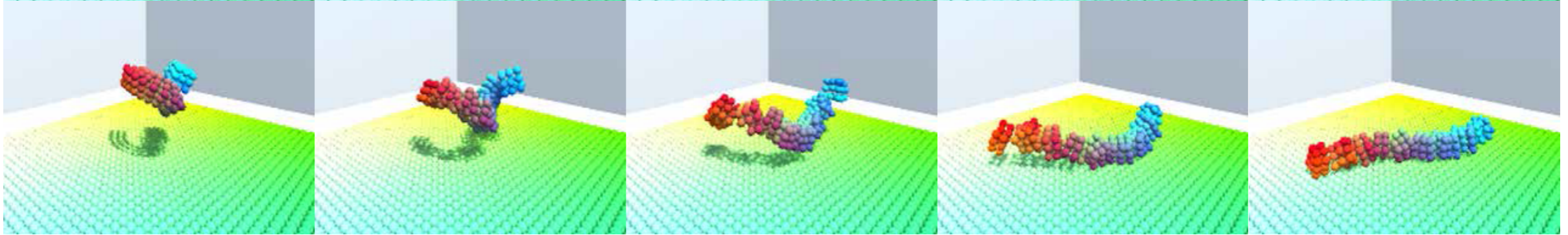
# Challenges:

slinky

Ground truth



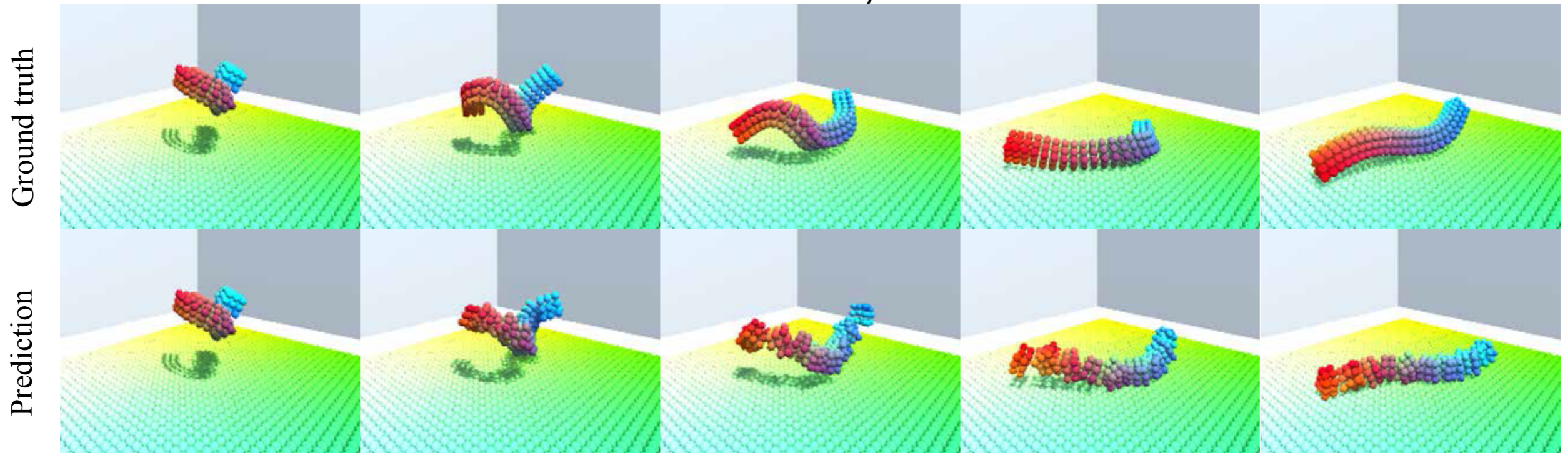
Prediction



*shape is not preserved super well over long rollouts...*

# Challenges:

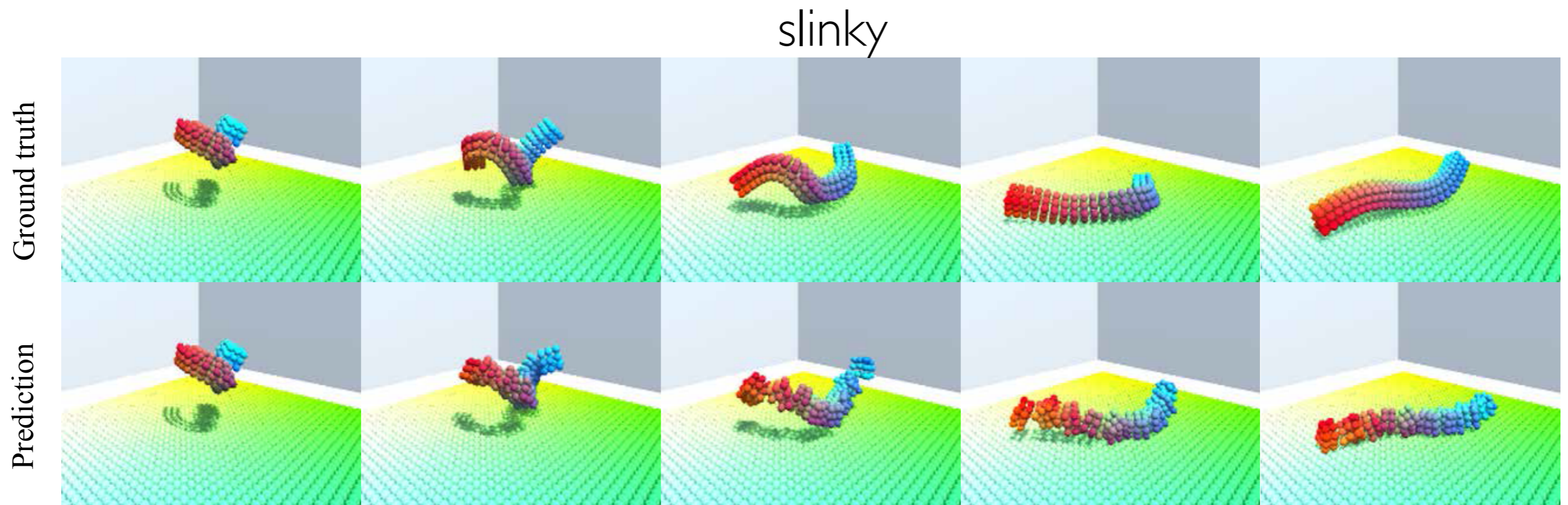
slinky



*shape is not preserved super well over long rollouts...*

Easy to impose simple shape conversation rules — in a “per material” way. (e.g. rigid different than cloth different than soft-body)

# Challenges:



*shape is not preserved super well over long rollouts...*

Easy to impose simple shape conversation rules — in a “per material” way. (e.g. rigid different than cloth different than soft-body)

... less easy to understand how to do this in material-agnostic way.

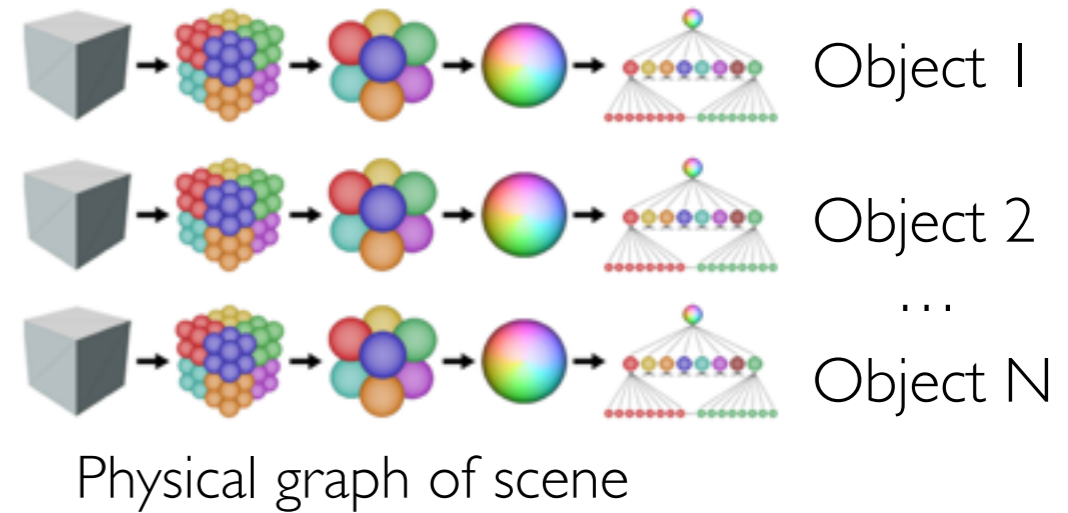
# Challenges:

Extracting the graph description from video.



Scene at time  $T = 0, 1, \dots, t$

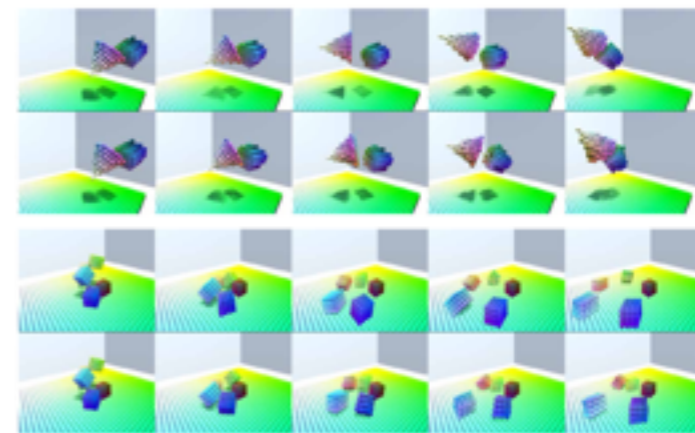
ConvRNN



Hierarchical  
Relation  
Network



Rendering



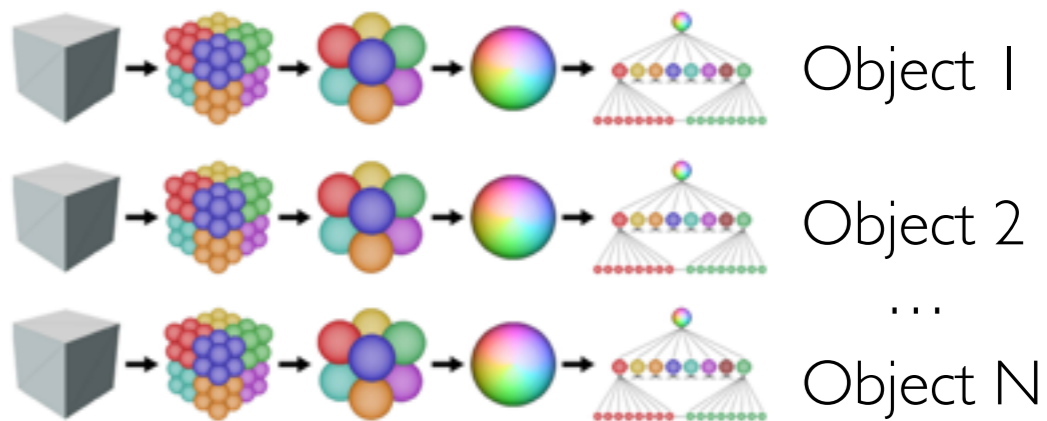
t+1 t+2 t+3 t+4 t+5

Predictions of graph in future

Scene at time  $T = t+1, t+2, \dots$

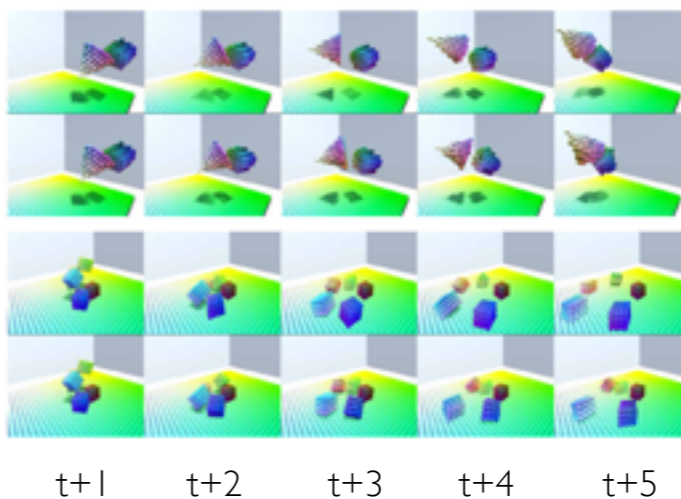


ConvRNN

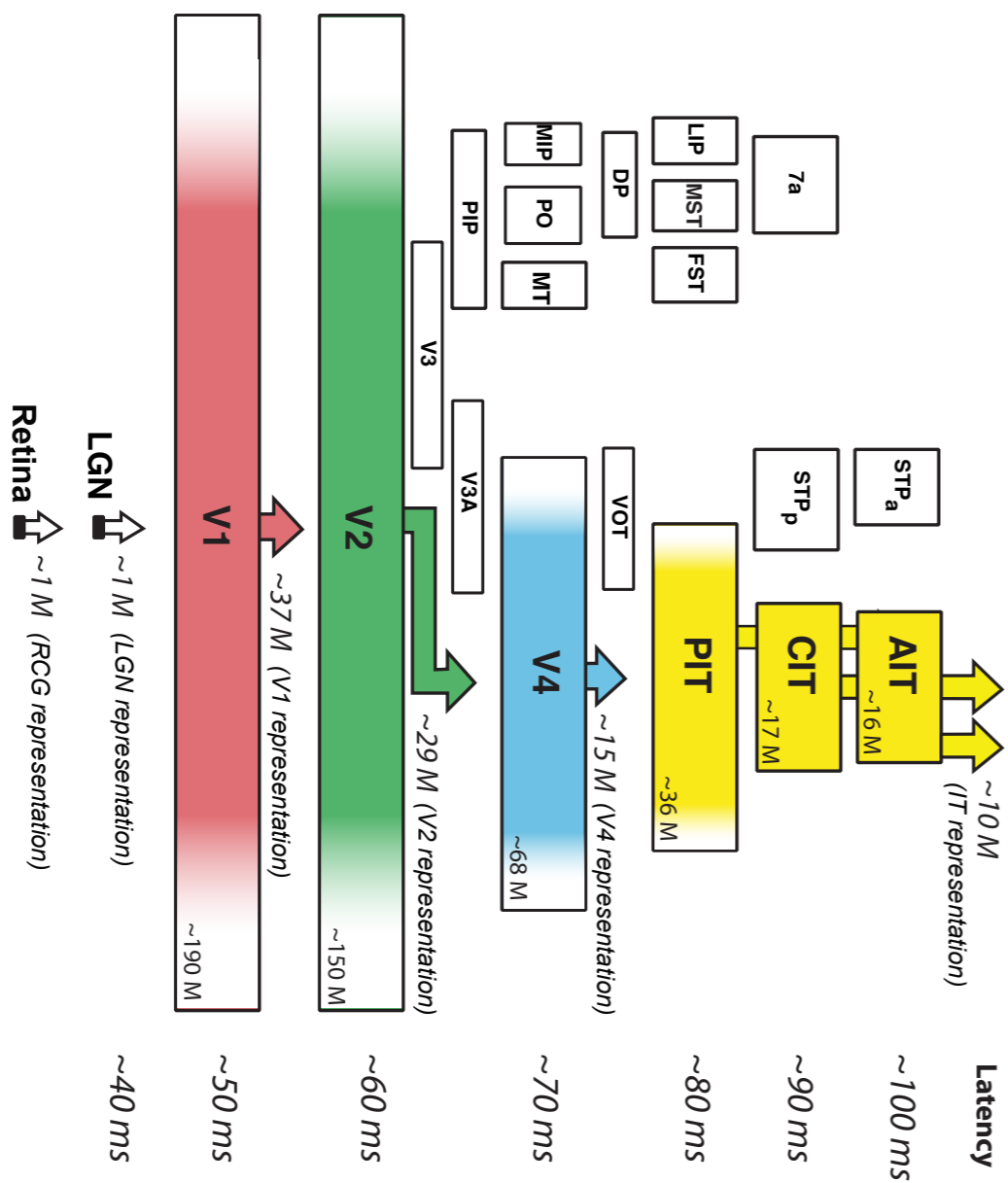


Physical graph of scene

Hierarchical  
Relation  
Network



Predictions of graph in future



# Human-centered feedback loop

## Artificial Intelligence Algorithms

1.  $\mathbf{A}$  = architecture class  
e.g. **CNNs**
2.  $\mathbf{L}$  = loss function     $\mathbf{D}$  = dataset  
e.g. **Object Categorization**    "task"

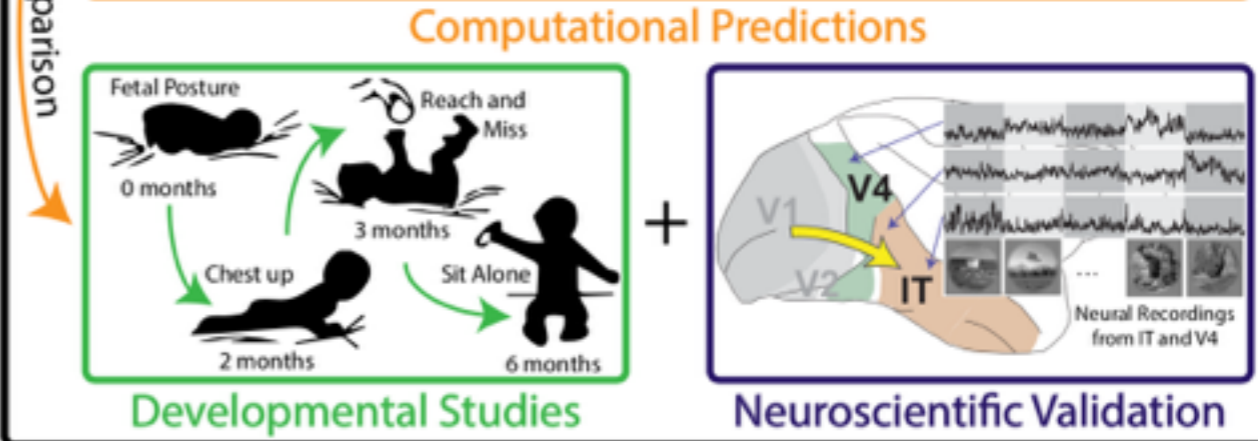
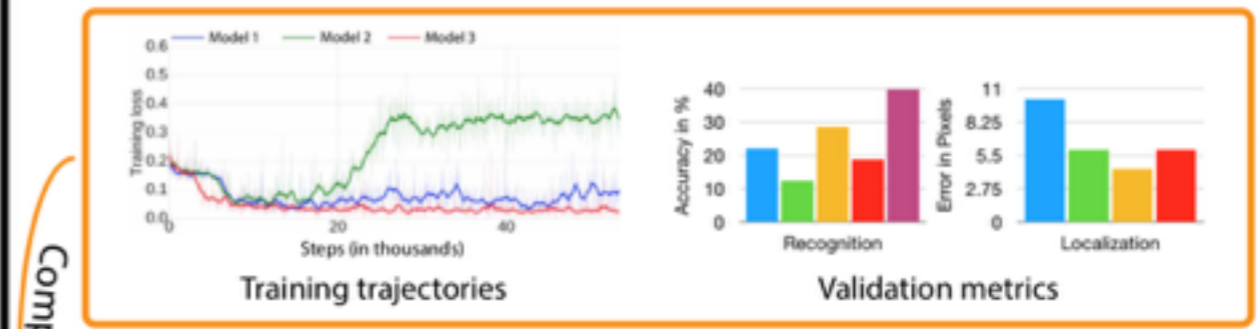
3. **Learning Rule**

$$\operatorname{argmin}_{a \in \mathcal{A}} [L(p_a^*)]$$

where  $p^*$  is result of

$$\frac{dp_a}{dt} \stackrel{\text{backprop}}{=} -\lambda(t) \cdot \langle \nabla_{p_a} L(x) \rangle_{x \in \mathcal{D}}$$

## Experimental Testing



AI Improvement

Comparison

Model Testing

Refinement

Thanks!