LEARNING INTUITIVE PHYSICS THROUGH OBJECTS

Luis S. Piloto
Princeton Neuroscience Institute
Princeton University
{luis.s.piloto}@gmail.com

DeepMind
{piloto}@google.com

Ari Weinstein & Peter Battaglia & Matt Botvinick
DeepMind
{ariweinstein,peterbattaglia,botvinick}@google.com

ABSTRACT

From early in infancy, humans' capacity for physical reasoning is crucial to how they understand and interact with the world. By contrast, modern artificial intelligence systems struggle to realize even basic physical intuition. Here we present "Object History and Memory Interactions" (OHMI)—a deep learning approach for predicting the future of 3D physical scenes from segmented images. We explore its ability to learn a range of physical concepts, evaluating the knowledge it acquires using a standard experimental methodology from development psychology called "violation of expectations". Our results show the model benefits from object-centered representations and computations, an observation that resonates with notions of core object knowledge from developmental psychology.

1 INTRODUCTION

Understanding the world in terms of objects is key to human intelligence. From early infancy, we form rich expectations about objects' properties, how they interact with one another, and how they persist through time (Baillargeon et al. (1985); Spelke et al. (1992)), and our object-centered perceptual and cognitive faculties are instrumental to how we learn, communicate, remember, explain, and interact with our environment.

These points have been increasingly brought to bear in AI research aiming to match human intelligence in the domain of intuitive physics. The net result is an increase in the number of AI systems endowed with object-centered representations and computations (Burgess et al. (2019); Greff et al. (2019); Kipf et al. (2019); Battaglia et al. (2016)). Despite this, current artificial learning systems have failed to match the intuitive physics capabilities even of infants.

In this work, we introduce an approach termed "Object History and Memory Interactions" (OHMI) which operates by calculating interactions between the model's memory of objects and veridical object histories. To measure the knowledge which our model has learned, we adopt the "violation of expectations" (VOE) paradigm—a standard measure from human development psychology which has been used widely for assessing infants' understanding of intuitive physics.

Our model achieves strong VOE results over a diverse set of physical concepts, a currently unprecedented result for models which learn object dynamics. Furthermore, we show that these results are dependent on: 1) object-centered computation and 2) allowing the model to operate over the full object histories at each timestep.

2 MODEL

2.1 OBJECT HISTORY AND MEMORY INTERACTIONS (OHMI)

The OHMI model is a next-step predictor trained on videos depicting the physical interactions of 3D objects. OHMI has three components: an encoder Φ, a recurrent dynamics predictor Δ, and a decoder Θ. For a segmented image tuple, $X = (x, m^{1:n})$ where x is an image, and $m^{1:n}$ are ordered per-object masks, Φ produces a set of learned object codes $z^{1:n}$, such that $\Phi(X) \Rightarrow z^{1:n}$.

For a history $hist_t^{1:n} = z_{1\leq t}^{1:n}$, and object memories $cell_{t-1}^{1:n}$, Δ predicts the object codes at the next timestep: $\Delta(hist_t^{1:n}, cell_{t-1}^{1:n}) \Rightarrow \hat{z}_{t+1}^{1:n}$. The decoder reverses the encoder, with $\Theta(z^{1:n}) \Rightarrow X$.

We pretrain Φ and Θ in tandem via a Component Variational Autoencoder (ComponentVAE) (Burgess et al. (2019)) on individual images. The VAE is trained to transform X to learned codes $z^{1:n}$, and then back to X using the standard variational objective. Analysis shows that dimensions in z correspond to rough analogs of object position, color, and shape (examples in Appendix). Additionally, because our decoder outputs an image and per-object masks, we can render a per-object image, $\underset{\sim}{x}^k$ or combine them to produce a single composite image $\underset{\sim}{x}$.

Our recurrent dynamics predictor, Δ, uses a ComponentLSTM: an object-wise LSTM with shared weights, but object-specific activations. At each timestep, the LSTM for the ith object computes the following: $\hat{z}_{t+1}^i, cell_t^i, hidden_t^i = LSTM(hist_t^i, cell_{t-1}^i, hidden_{t-1}^i)$. Intuitively, we expect that the cell state of each LSTM tracks its corresponding object in the scene and thus we refer to it as the *object memory*. This computation is sufficient to predict dynamics of objects in isolation, but we also must be able to compute the influence of objects on each other, which we do by an Interaction Network (Battaglia et al. (2016)). Because we care about interactions even when objects are occluded (only represented in object memory), we compute interactions *from* the object memory to both the object memory *and* the object history inputs. For the ith object memory contained in $cell_{t-1}^i$, we compute: $int_i = IN(from = cell_{t-1}^i, to = [cell_{t-1}^1 : n; hist_t^{1:n}])$ We add the corresponding interactions as an additional input to the ith LSTM, yielding the final form of the ComponentLSTM computation: $\hat{z}_{t+1}^i, cell_t^i, hidden_t^i = LSTM(hist_t^i, cell_{t-1}^i, hidden_{t-1}^i, int_i)$ See A.1 in Appendix for an illustration of these modules.

2.2 Loss

We learn object dynamics by training Δ as a deterministic[1] next-step predictor. At each timestep, we give the the predictor Δ, $hist_t^{1:n}$. From this, Δ produces a prediction for the object states at the next timestep, $\hat{z}_{t+1}^{1:n}$. To form a prediction target, we compute the actual $z_{t+1}^{1:n} = \Phi(X_{t+1})$, where X_{t+1} is the image-mask for the next time step. During training we optimize for the loss at the level of per-object codes ($L_{codes} = ||\hat{z}_{t+1}^{1:n} - z_{t+1}^{1:n}||$). Other losses are also reasonable and are examined in evaluation: per-object images ($L_{object\ pixels} = ||\hat{\underset{\sim}{x}}_{t+1}^{1:n} - \underset{\sim}{x}_{t+1}^{1:n}||$), and per-object images excluding the background($L_{fg\ object\ pixels} = ||\hat{\underset{\sim}{x}}_{t+1}^{2:n} - \underset{\sim}{x}_{t+1}^{2:n}||$). See A.2 in Appendix for training details.

3 Dataset

3.1 Generic Physical Events

Our training dataset consists of 300,000 randomly generated scenes each containing two to four physical event building blocks. We have 14 composable building block types intended to span a wide set of physical phenomena including: rolling, collisions along the ground plane, collisions from throwing or dropping an object, occlusions (via a "curtain" that descends from the top of the screen and retracts), object stacks, covering interactions, containment events, and rolling up/down ramps. We restrict the primitive shapes in our dataset to rectangular prisms and spheres. From the rectangular prisms we build a "curtain," a ramp, an arch, and both open-top and closed-top containers. Examples can be seen here: https://bit.ly/39iyshP.

Each scene is rendered using a camera with drifting position and orientation. We generate the dataset using the Mujoco engine (Todorov et al. (2012)), where each scene is 15 frames long at 64x64 RGB resolution. For each frame, we produce a per-object mask for up to 10 objects, with a consistent ordering of object masks throughout the scene. We generate an additional 5,000 scenes for validation during hyperparameter optimization and another 5,000 scenes for a final test set.

[1]Despite inherent stochasticity in the dataset (e.g. new objects appearing from off screen), we found this model worked well enough. We leave it for future work to incorporate stochasticity into our model.

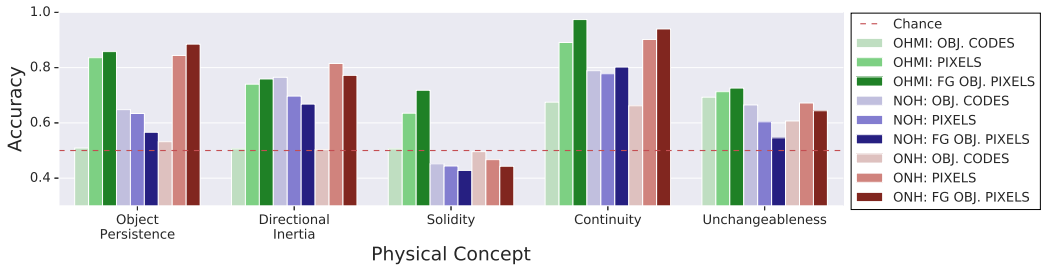

Figure 1: Comparison of model performance on VOE tests for five different physical concepts measured via classification accuracy. Models compared: **OHMI** uses interactions between object histories and memories; **NOH** *No Objects/With History*; **ONH** *Objects/No History*.

3.2 Violation Of Expectation (VOE)

To assess the model's knowledge of specific physical concepts, we leverage the VOE paradigm from the developmental literature and adapted for artificial models by (Piloto et al. (2018); Riochet et al. (2018)). As such, we generate 5,000 probe scenarios for each of the following physical concepts adapted from developmental psychology: object persistence, continuity, 'unchangeableness', directional inertia, and solidity (which are all[2] described in Piloto et al. (2018)). Each scenario, consists of two physically plausible probe videos depicting the physical concept. Two corresponding implausible videos are generated by swapping the first and second halves of the plausible videos [3]. We only splice frames such that all adjacent frames are physically plausible, but the scene as a whole is not. We assess our model's performance by computing the classification accuracy over all 5,000 probe scenarios. A scenario is "classified" correctly when the model is more surprised by the implausible videos than the plausible videos. Examples are available here: https://bit.ly/2vu3g0B.

4 Results

For evaluation, we look at classification accuracy for five different physical concepts using the VOE paradigm. Where the VOE paradigm requires a measure of surprise, we evaluated 3 candidate losses: L_{codes}, $L_{objectpixels}$, and $L_{fg\ object\ pixels}$. We compared our model to two ablated models. The first model, (No Objects/With History), was intended to assess the importance of objects. By setting the number of "objects" in our model to 1, we remove all object-centered computation. To make a fair comparison, this model has more parameters, but the same number of activations, as OHMI. The second model, (Objects/No History), is object-based, but is aimed to assess the importance of object *histories*. Whereas the standard object-based model is fed in all previously observed frames at each timestep, this model is still recurrent, but only sees the current timestep in its input. We report our results in Figure 1

To our knowledge, this is the first time a model has learned physics from segmented images and passed VOE examination for a diverse set of physical concepts. We found that only OHMI was capable of demonstrating strong classification accuracy on all 5 datasets. Where the other models perform above chance, OHMI tends to be at least as good, if not better.

To assess whether the NOH model was bogged down by predicting the background, we also evaluated our models with $L_{fg\ pixel\ objects}$. This loss is still in pixel space, but excludes the background. Even in this regime, the NOH model struggles. Although we see the strongest results for OHMI using the $L_{fg\ pixel\ objects}$ loss which requires additional privileged information, we still get very strong results with $L_{pixel\ objects}$ which does not require information about the background. To our surprise, both the object-based models were at chance performance using L_{codes} as a measure of surprise, which is a loss that should correspond to comparing high-level object properties. Conversely, the non-object model (NOH) model, which had a single "object code" to represent the entire scene showed its best accuracy using this loss. We look forward to exploring this behavior in future work.

[2]Except for *directional inertia* which was added to specifically probe knowledge of collisions.

[3]This approach was pioneered by Riochet et al. (2018).

5 DISCUSSION

5.1 WHERE DO OUR OBJECTS COME FROM?

The ideal physical reasoning system would operate directly from perceptual inputs without any privileged or ground truth information. Currently, our model requires two pieces of privileged information: object masks and the correspondence from objects at one timestep to the next. Recent models give us the reasonable expectation that we can learn these two components directly from data. MONet (Burgess et al. (2019)) has been developed to perform unsupervised object segmentation. The AlignNet (Creswell et al. (2020)) is more recent work focused on solving the correspondence problem which, with some work, can be used to align the extracted objects from one time step to the next. Preliminary results (see Appendix) indicate that MONet is capable of extracting objects from our dataset. We leave it for future work to incorporate these components into our model.

5.2 RELATED WORK

Recent work has sought to benchmark and solve the challenge of physical reasoning by bringing the VOE paradigm to artificial intelligence. Piloto et al. (2018) examined VOE results for a wide set of physical phenomena with a non object-based VRNN. The model's generalization capabilities were modest: the model trained on different instances of the plausible VOE probes. The present work shows much stronger generalization by using training data with A) unstructured scenes quite remote from the test dataset B) a moving camera thus increasing visual diversity. Riochet et al. (2018) built a VOE dataset with parametric complexity and rich textures. They explored CNN and GAN-based models for just a single physical concept: object permanence. Riochet et al. (2020) is similar to the current work in many ways, but there are also notable differences. This model requires ground truth masks, but unlike our model also requires depth information and ground truth object properties (e.g. color, shape). Like OHMI, they predict per-object properties. However, where we use a VAE to learn the representational format of object properties (our object codes), they explicitly specify an object's state as its position and depth. Where we assume object correspondence is given as input (and hence something to work on in the future), they resolve the object correspondence problem by choosing the nearest object[4]. In the regime where they train with object masks, a depth mask, ground truth object properties and hard-code the representational format for objects, they report above-chance VOE results on a dataset analagous to our "continuity" dataset.

Finally, Smith et al. (2019) develop a VOE dataset without collisions and model human behavior on the VOE paradigm. However, they do not seek to learn physical reasoning. Similar to Riochet et al. (2020), they hard-code the representational format for their objects. They leverage this format to plug object properties into a physics engine instead of learning object dynamics. Binz & Endres (2019) model the developmental stages of physical reasoning via Bayesian Neural Networks in the domains of occlusion and numerosity, but do so in a simplified 2D environment.

5.3 CONCLUSION

In this work, we built OHMI, an object-centered model for learning physical reasoning from segmented images. We found it has unprecedented success on learning physical concepts as measured by the VOE paradigm. Our assessment revealed practical considerations for physical reasoning systems. First, we saw a clear benefit to object-centered representation and computation supporting the developmental claim that objects underpin physical reasoning in infants. Furthermore, our model demonstrates it is possible to learn a representational format for objects that enables predicting physical interactions (a critical, but under-explored, ingredient for the "object files" account of infant physical reasoning Xu (2013)). From an engineering perspective, our results show that although recurrent models are fully capable of remembering object histories, our architecture performed significantly better when allowed to operate over a full history[5]. We pave the way for future work to incorporate unsupervised segmentation and alignment into our pipeline to learn physical reasoning directly from pixels.

[4]It is unclear how this would work when the representational format is not hard-coded.

[5]This mirrors the recent trend of adopting Transformer models which eschew recurrent processing of sequences in favor of operating over a window of history. Here we do recurrent processing of the full window of history.

AUTHOR CONTRIBUTIONS

All authors helped write the paper. Ari Weinstein and Luis Piloto developed the datasets. Luis Piloto developed the model and ran all experiments.

ACKNOWLEDGMENTS

We thank the following people for helpful conversations and feedback throughout this work: Antonia Creswell, Chris Burgess, Loic Matthey, Nick Watters, Alexander Lerchner. We thank Murray Shanahan and Matt Overlan for reviewing an early draft of this paper.

REFERENCES

Renee Baillargeon, Elizabeth S Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985.

Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. *CoRR*, abs/1612.00222, 2016. URL http://arxiv.org/abs/1612.00222.

Marcel Binz and Dominik Endres. Emulating human developmental stages with bayesian neural networks. *CoRR*, abs/1902.07579, 2019. URL http://arxiv.org/abs/1902.07579.

Christopher P. Burgess, Loïc Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matthew Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *CoRR*, abs/1901.11390, 2019. URL http://arxiv.org/abs/1901.11390.

Antonia Creswell, Luis Piloto, David Barrett, Kyriacos Nikiforou, David Raposo, Marta Garnelo, Peter Battaglia, and Murray Shanahan. Alignnet: Self-supervised alignment module, 2020. URL https://openreview.net/forum?id=H1gcw1HYPr.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loïc Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. *CoRR*, abs/1903.00450, 2019. URL http://arxiv.org/abs/1903.00450.

Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. *arXiv preprint arXiv:1911.12247*, 2019.

Luis Piloto, Ari Weinstein, Dhruva TB, Arun Ahuja, Mehdi Mirza, Greg Wayne, David Amos, Chia-Chun Hung, and Matthew Botvinick. Probing physics knowledge using tools from developmental psychology. *CoRR*, abs/1804.01128, 2018. URL http://arxiv.org/abs/1804.01128.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *CoRR*, abs/1803.07616, 2018. URL http://arxiv.org/abs/1803.07616.

Ronan Riochet, Josef Sivic, Ivan Laptev, and Emmanuel Dupoux. Occlusion resistant learning of intuitive physics from videos, 2020. URL https://openreview.net/forum?id=HylfPgHYvr.

Kevin Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Josh Tenenbaum, and Tomer Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In *Advances in Neural Information Processing Systems*, pp. 8983–8993, 2019.

Elizabeth S Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowledge. *Psychological review*, 99(4):605, 1992.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012.

Nicholas Watters, Loïc Matthey, Christopher P. Burgess, and Alexander Lerchner. Spatial broadcast decoder: A simple architecture for learning disentangled representations in vaes. *CoRR*, abs/1901.07017, 2019. URL `http://arxiv.org/abs/1901.07017`.

Fei Xu. The object concept in human infants. *Human Development*, 56(3):167, 2013.

A Appendix

A.1 Architecture Details

The ComponentVAE had 10 components/objects. The ComponentVAE's encoder used a series of 4 2D convolutions with the following parameters: `output channels: (32, 32, 64, 64), kernel shapes: 2, strides: 2, paddings: 'SAME', activation: ReLU, activate final: True`. This was followed by an multi-layer perceptron (MLP) with 256 units and ReLU activations. This was projected to form a latent code with 16 dimensions. The ComponentVAE's decoder was a Broadcast Decoder (Watters et al. (2019)) to encourage disentangling. The decoder used 4 1x1 convolutional blocks with the ReLU activation. The ComponentVAE module is depicted below in Figure 2 (although we only show it for 3 components/objects.)

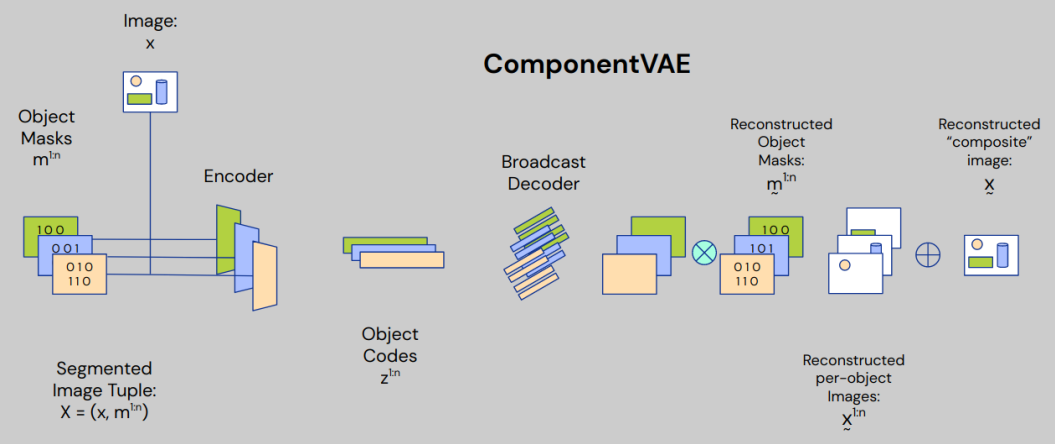

Figure 2: Depiction of ComponentVAE. Reconstructions indicated by placing a ˜ accent underneath.

The dynamics predictor takes as input its state from the previous time step and the set of object histories. Figure 3 shows how we use ComponentVAE to produce the object histories. The predictor consists of two modules: a pair of Interaction Networks (depicted in Figure 4 and a Component LSTM (depicted in Figure 5).

The Interaction Networks used consisted of 2 MLP layers with 512 units and the ELU activation function. Interactions were aggregated using a `sum_max` function which concatenated the sum of the interactions to the max of the interactions.

The ComponentLSTM had 10 components/objects. The LSTM cell and hidden states had 2056 units.

A.2 Training

Pretraining of the ComponentVAE used the RMSProp optimizer, with a learning rate of 1e-4 for 1,000,000 steps. Training the dynamics predictor used the ADAM optimizer with a learning rate of 1e-4 for 1,000,000 steps.

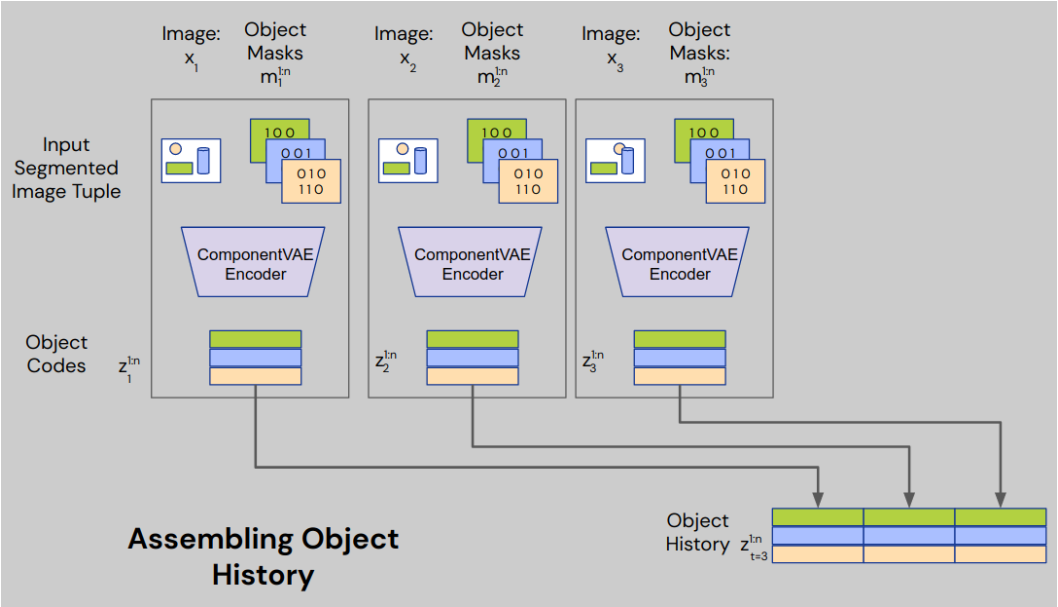

Figure 3: Here we explicitly depict how we use the ComponentVAE to encode the observations at each timestep to produce the object history. The figure specifically shows the object history for $t = 3$

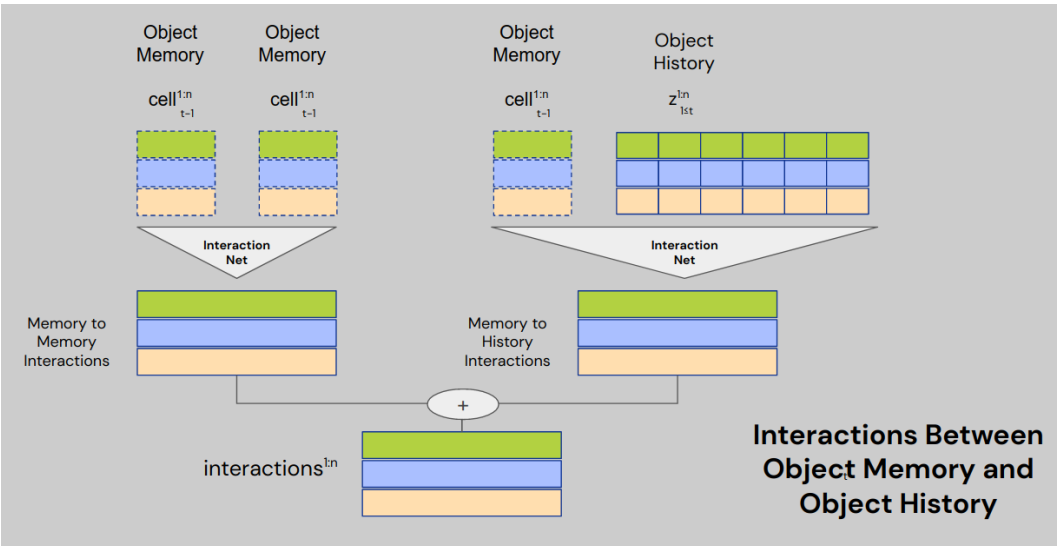

Figure 4: Here we explicitly depict the interactions that give rise to the name of our model, Object History and Memory Interactions. We compute the interactions *from* the object memory to both the object memory and the object history via two separate interaction networks. We then aggregate the outputs to yield the interactions for each object in memory. Importantly, the object memory and object history are different entities. As shown above, the history is the encoding of all the observed inputs - as such it is veridical. The object memory is taken to be the cell states of the ComponentLSTM and is not necessarily veridical.

A.3 Learned Representational Format For Object Codes

In Figure 6 an example of the representational format learned by a ComponentVAE when trained on an earlier version of our dataset. We haven't produced such plots for the most recent version of the dataset, but have no reason to expect it is qualitatively different than the results presented here.

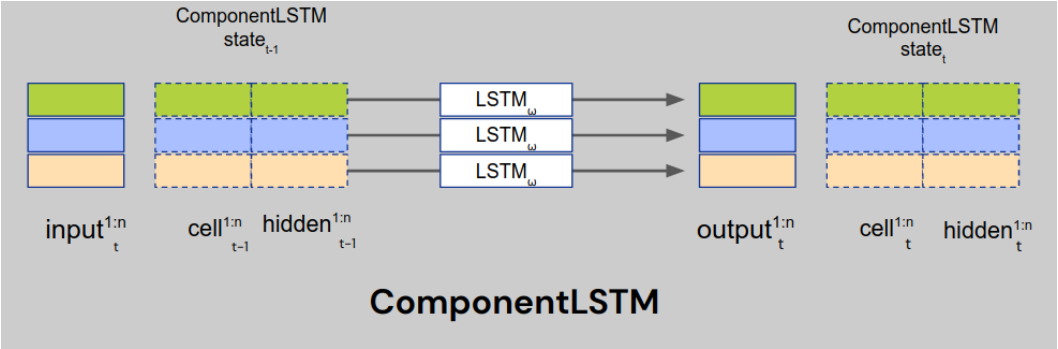

Figure 5: Depiction of generic ComponentLSTM. At each timestep, it receives a set of n inputs and previous LSTM states and runs them in parallel through a set of LSTMs with shared weights ω to produce a set of n outputs and corresponding LSTM states for the next time step. In our dynamics predictor, the cell state is called the object memories. The input consists of the object histories concatenated to the corresponding interactions (computed in Figure 4). The outputs are the predictions for the object codes at the next timetep.

A.4 Unsupervised Object Segmentation With MONet

In Figure 7 we show the result of applying MONet to our dataset to produce unsupervised object segmentations. The model deals fairly well with cluttered objects. A point for improvement is that the model tends to group shadows as a unique object, although it is unlikely this over-segmentation would prevent us from using MONet in our pipeline.

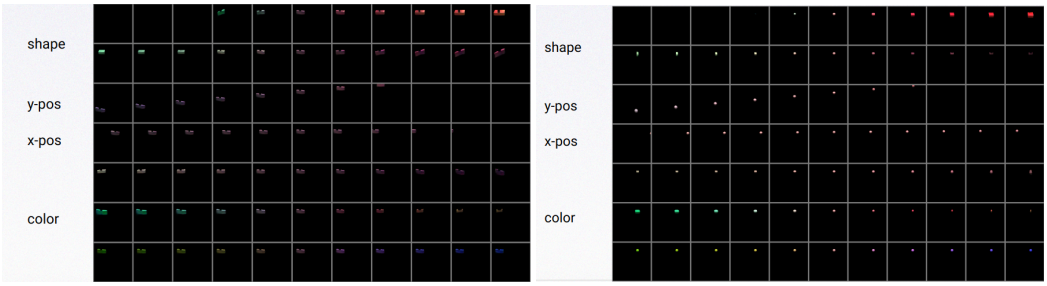

Figure 6: Latent traversals for two separate objects. We run an object, seen in the center column of each image, through the ComponentVAE's encoder to produce the object code: z^i. The jth row shows the result of perturbing the jth dimension of z_i and running the perturbed code through the ComponentVAE's decoder. In this way, we can visualize the properties encoded by each dimension of the object code. The first two dimensions seem to code for a rough analog of shape, the next two for position, and the last three for color.

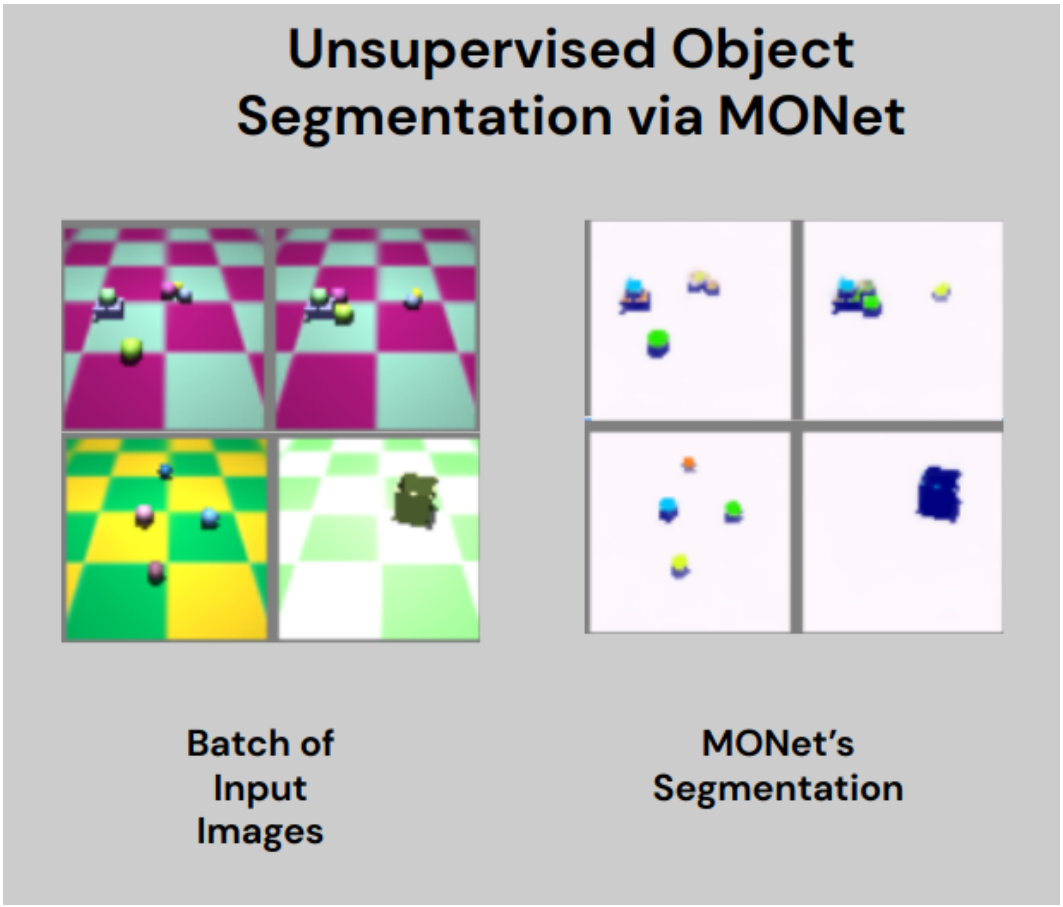

Figure 7: Here we show the result of applying MONet to our dataset to yield unsupervised object segmentations. Each unique color in the righthand images correspond to a segmented object.