R-SQAIR:Relational Sequential Attend, Infer, Repeat

Aleksandar Stanić
Swiss AI Lab, IDSIA, USI, SUPSI
Lugano, Switzerland
aleksandar@idsia.ch

Jürgen Schmidhuber
Swiss AI Lab, IDSIA, USI, SUPSI, NNAISENSE
Lugano, Switzerland
juergen@idsia.ch

Abstract

Traditional sequential multi-object attention models rely on a recurrent mechanism to infer object relations. We propose a relational extension (R-SQAIR) of one such attention model (SQAIR) by endowing it with a module with strong relational inductive bias that computes in parallel pairwise interactions between inferred objects. Two recently proposed relational modules are studied on tasks of unsupervised learning from videos. We demonstrate gains over sequential relational mechanisms, also in terms of combinatorial generalization.

1 Introduction

Numerous studies [35, 17, 27, 1, 22] show that infants quickly develop an understanding of intuitive physics, objects and relations in an unsupervised manner. To facilitate the solution of real-world problems, intelligent agents should be able to acquire such knowledge [31]. However, artificial neural networks are still far from human-level understanding of intuitive physics.

Existing approaches to unsupervised learning about objects and relations from visual data can be categorized into either parallel [11, 12, 10] or sequential [26, 25, 7, 18, 6, 5, 36], depending on the core mechanism responsible for inferring object representations from a single image. One model from the former group is Tagger [11] which applies the Ladder Network [20] to perform perceptual grouping. RTagger [19] replaces the Ladder Network by a Recurrent Ladder Network, thus extending Tagger to sequential settings. NEM [12] learns object representations using a spatial mixture model and its relational version R-NEM [30] endows it with a parallel relational mechanism. The recently proposed IODINE [10] iteratively refines inferred objects and handles multi-modal inputs.

On the other hand, the sequential attention model AIR [7] learns to infer one object per iteration over a given image. Contrary to NEM, it extracts object glimpses through a hard attention mechanism [26] and processes only the corresponding glimpse. Furthermore, it builds a probabilistic representation of the scene to model uncertainty. Many recent models have AIR as the core mechanism: SQAIR [18] extends AIR to sequential settings, similarly does DDPAE [15]. SPAIR [6] scales AIR to scenarios with many objects and SuPAIR[28] improves speed and robustness of learning in AIR. The recent MoNET [5] also uses a VAE and a recurrent neural network (RNN) to decompose scenes into multiple objects. These methods usually model relations by a sequential relational mechanism such as an RNN which limits their relational reasoning capabilities[3].

Here we present Relational Sequential Attend, Infer, Repeat (R-SQAIR) to learn a generative model of intuitive physics from video data. R-SQAIR builds on SQAIR which we augment by a mechanism that has a strong relational inductive bias [2, 30, 21]. Our explicit parallel model of pairwise relations between objects is conceptually simpler than a sequential RNN-based model that keeps previous interactions in its memory and cannot directly model the effects of interactions of previously considered objects. Our experiments demonstrate improved generalization performance of trained models in new environments.

Workshop on Perception as Generative Reasoning, NeurIPS 2019, Vancouver, Canada.

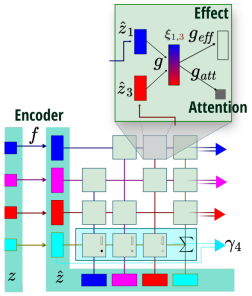

Figure 1: Interaction Network of R-NEM [30].

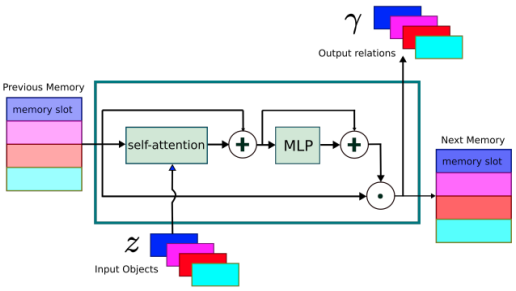

Figure 2: Relational Memory Core [21].

2 Relational Sequential Attend Infer Repeat

Attend, Infer, Repeat (AIR) [7] is a generative model that explicitly reasons about objects in a scene. It frames the problem of representing the scene as probabilistic inference in a structured VAE. At the core of the model is an RNN that processes objects one at a time and infers latent variables $\mathbf{z} = \{\mathbf{z}^{\mathrm{i}}_{\mathrm{what}}, \mathbf{z}^{\mathrm{i}}_{\mathrm{where}}, z^{\mathrm{i}}_{\mathrm{pres}}\}_{i=1}^{n}$, where $n \in \mathbb{N}$ is the number of objects. The continuous latent variable $\mathbf{z}^{\mathrm{what}}$ encodes the appearance of the object in the scene and $\mathbf{z}^{\mathrm{where}}$ encodes the coordinates according to which the object glimpse is scaled and shifted by a Spatial Transformer [16]. Given an image $\mathbf{x}$, the generative model of AIR is defined as follows:

$$p_\theta(\mathbf{x}) = \sum_{n=1}^{N} p_\theta(n) \int p^z_\theta(\mathbf{z}|n) p^x_\theta(x|\mathbf{z}) d\mathbf{z}, \tag{1}$$

where $p_\theta(n) = \mathrm{Geom}(n \mid \theta)$ represents the number of objects present in the scene, $p^z_\theta(\mathbf{z}|n)$ captures the prior assumptions about the underlying object and $p^x_\theta(x|\mathbf{z})$ defines how it is rendered in the image. In general, the inference for Equation 1 is intractable, so [7] employs amortized variational inference using a sequential algorithm, where an RNN is run for N steps to infer latent representation of one object at a time. The variational posterior is then:

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = q_\phi\big(z^{n+1}_{\mathrm{pres}} = 0 \mid \mathbf{z}^{1:n}, \mathbf{x}\big) \prod_{i=1}^{n} q_\phi\big(\mathbf{z}^i, z^i_{\mathrm{pres}} = 1 \mid \mathbf{z}^{1:i-1}, \mathbf{x}\big), \tag{2}$$

where q_ϕ is a neural network which outputs the parameters of the latent distributions: the mean and standard deviation of a Gaussian distribution for $\mathbf{z}_{\mathrm{what}}$ and $\mathbf{z}_{\mathrm{where}}$ and the probability parameter of the Bernoulli distributed z_{pres}.

Relational Sequential Attend, Infer, Repeat (R-SQAIR) augments SQAIR through a parallel relational mechanism. SQAIR extends AIR to the sequential setting by leveraging the temporal consistency of objects using a state-space model. It has two phases: Discovery (DISC) and Propagation (PROP). PROP is active from the second frame in the sequence, propagating or forgetting objects from the previous frame. It does so by combining an RNN, which learns temporal dynamics of each object, with the AIR core which iterates over previously propagated objects (explaining away phenomena). DISC phase uses the AIR core, conditioned on propagated objects, to discover new appearances of objects. For a full description of AIR and SQAIR we refer to previous work [7, 18].

R-SQAIR retains the strengths of its predecessors and improves their relational capabilities. More specifically, SQAIR relies on AIR's core RNN to model the relations. However, an RNN has only a weak relational inductive bias [3], as it needs to compute pairwise interactions between objects sequentially, iterating over them in a specific order. R-SQAIR, on the other hand, employs networks with strong relational inductive bias which can model arbitrary relations between objects in parallel. To construct conceptually simple yet powerful architectures that support combinatorial generalization, we use the following two methods: *Interaction Network* (IN) [30] and *Relational Memory Core* (RMC) [21].

The R-SQAIR generative model is built by extending the PROP module of SQAIR to include relations $\gamma_t = \Gamma(\mathbf{z}_{t-1})$, where Γ is the relational module and $\mathbf{z}_{t-1}$ are object representations from the previous timestep, defined as follows:

$$p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = p^D(\mathbf{z}^{D_1}_1) \prod_{t=2}^{T} p^D(\mathbf{z}^{D_t}_t | \mathbf{z}^{P_t}_t) p^P(\mathbf{z}^{P_t}_t | \boldsymbol{\gamma}_t) p_\theta(\mathbf{x}_t | \mathbf{z}_t), \tag{3}$$

The *discovery prior* $p^D(\mathbf{z}_t^{D_t}|\mathbf{z}_t^{P_t})$ samples latent variables $\mathbf{z}_t^{D_t}$ for new objects that enter the frame, by conditioning on propagated variables $\mathbf{z}_t^{P_t}$. The *propagation prior* $p^P(\mathbf{z}_t^{P_t}|\gamma_t)$ samples latent variables for objects that are propagated from the previous frame and removes those that disappear. Both priors are learned during training. We recover the original SQAIR model for $\gamma_t = \mathbf{z}_{t-1}$. The inference model is therefore:

$$q_\phi(\mathbf{z}_{1:T} \mid \mathbf{x}_{1:T}) = \prod_{t=1}^{T} q_\phi^D\left(\mathbf{z}_t^{D_t} \mid \mathbf{x}_t, \mathbf{z}_t^{P_t}\right) \prod_{i \in \mathcal{O}_{t-1}} q_\phi^P\left(\mathbf{z}_t^i \mid \gamma_{\mathbf{t}}{}^i, h_t^i\right), \tag{4}$$

where h_t^i are hidden states of the temporal and AIR core RNNs. Discovery q_ϕ^D is essentially the posterior of AIR. Again, the difference to SQAIR lies in the propagation module q_ϕ^P, which receives relations γ_t as the input.

Relational Module

Interaction Network Our first relational module is the Interaction Network (IN) of R-NEM [30], depicted in Figure 1, which is closely related to Interaction Networks [2, 34]. Here, the effect on object k of all other objects $i \neq k$ is computed by the relational module $\gamma_t = \Gamma^{IN}(\mathbf{z}_{t-1})$, which in the case of IN is defined as follows (for simplicity we drop time indices):

$$\hat{\boldsymbol{z}}_k = f(\boldsymbol{z}_k), \;\; \boldsymbol{\xi}_{k,i} = g([\hat{\boldsymbol{z}}_k; \hat{\boldsymbol{z}}_i]), \;\; \boldsymbol{E}_k = \sum_{i \neq k} g_{att}(\boldsymbol{\xi}_{k,i}) \cdot g_{eff}(\boldsymbol{\xi}_{k,i}), \;\; \gamma_k = [\mathbf{z}_k; \mathbf{e}_k], \tag{5}$$

where $\mathbf{z}_i = \{\mathbf{z}^{\mathrm{i}}_{\mathrm{what}}, \mathbf{z}^{\mathrm{i}}_{\mathrm{where}}, z^{\mathrm{i}}_{\mathrm{pres}}\}$ from the previous time step. First, each object $\boldsymbol{z}_i$ is transformed using an MLP f to obtain $\hat{\boldsymbol{z}}_i$, which is equivalent to a node embedding operation in a graph neural network. Then each pair $(\hat{\boldsymbol{z}}_k, \hat{\boldsymbol{z}}_i)$ is processed by another MLP g, which corresponds to a node-to-edge operation by encoding the interaction between object k and object i in the embedding $\boldsymbol{\xi}_{k,i}$. Note that the computed embedding is directional. Finally, an edge-to-node operation is performed, where the effect on object k is computed by summing the individual effects $g_{eff}(\boldsymbol{\xi}_{k,i})$ of all other objects i on the particular object k. Note that the sum is weighted by an attention coefficient $g_{att}(\boldsymbol{\xi}_{k,i})$, which allows each individual object to consider only particular interactions. This technique also yields better combinatorial generalization to a higher number of objects, as it controls the magnitude of the sum.

Relational Memory Core We compare the effects modeled by IN to the effects learned by a Relational Memory Core (RMC), $\gamma_t = \Gamma^{RMC}(\mathbf{z}_{t-1})$. RMC (Figure 2) learns to compartmentalize objects into memory slots, and can keep the state of an object and combine this information with the current object's representation $\mathbf{z}_t$. This is achieved by borrowing ideas from memory-augmented networks [29, 8, 9] and interpreting memory slots as object representations. The interactions between objects are then computed by a multi-head self-attention mechanism [32] Finally, recurrence for the sequential interactions is introduced, resulting in an architecture that is akin to a 2-dimensional LSTM[14], where rows of the memory matrix represent objects. The model parameters are shared for each object, so the number of memory slots can be changed without affecting the total number of model parameters. For a full description, we refer to previous work [21].

3 Experiments

We analyze the physical reasoning capabilities of R-SQAIR on the *bouncing balls* dataset, which consists of video sequences of 64x64 images. As done in SQAIR experiments, we crop the central 50x50 pixels from the image, such that a ball can disappear and later re-appear. Although visually simple, this dataset contains highly complex physical dynamics and has been previously used for similar studies (R-NEM [30]). The method is trained in SQAIR-like fashion by maximizing the importance-weighted evidence lower-bound IWAE [4], with 5 particles and the batch size of 32. Curriculum learning starts at sequence length 3 which is increased by one every 10000 iterations, up to a maximum length of 10. Early stopping is performed when the validation score has not improved for 10 epochs.

Qualitative evaluation of R-SQAIR is present in Figure 3. Each column represents one time step in the video. The first row is about the R-SQAIR model trained and evaluated on videos with 4 balls, with object representations highlighted by different color bounding boxes. In the second row the same model is evaluated on datasets with 6-8 balls. Note that R-SQAIR disentangles objects already in

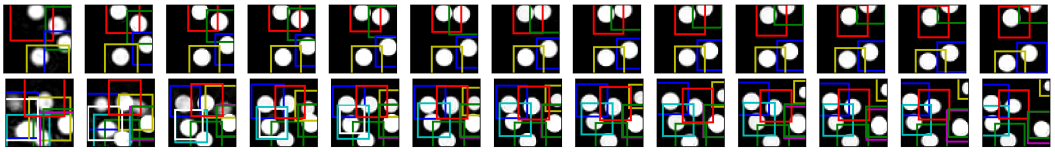

Figure 3: R-SQAIR trained on sequence of 4 bouncing balls (top rows) and evaluated on 6-8 bouncing balls.

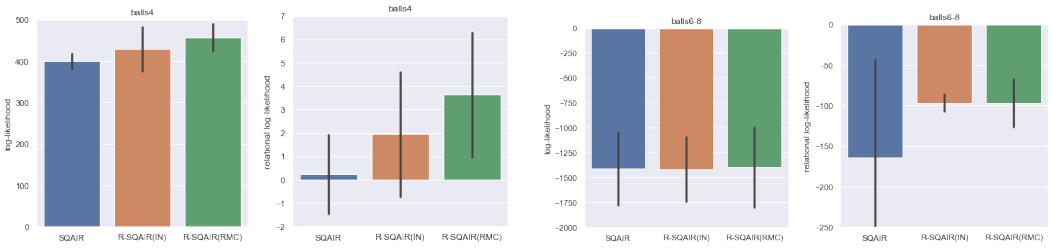

Figure 4: Log-likelihood and relational log-likelihood of R-SQAIR and SQAIR on the bouncing balls task.

the first few frames and later only refines the learned representations. At each time step, it computes up to $k = 4$ object representations, by considering objects from the previous frame and the learned dynamics.

For all SQAIR hyperparameters we use default values, except for the dimensionality of latent variable $\mathbf{z}_{\text{what}}$, which is set to 5 instead of 50. This reflects the low visual complexity of individual objects in the scene. For similar reasons, the embedding dimensionality of IN we use is also set to 5. We use a version of the IN module with attention coefficients to compute the weighted sum of the effects. In total, this adds 9 389 parameters to the 2 726 166 of the default SQAIR implementation. It also suggests that improved performance is a result of learning a better *propagation prior* instead of just increasing the number of model parameters.

RMC has more hyperparameters to choose from. We use self-attention with 4 heads, each of dimensionality 10. The number of memory slots is 4 and coincides with the total number of sequential attention steps we perform. Finally, RMC can perform several computations of attention per time step, where each corresponds to one message passing phase. As we are interested only in collisions, we compute attention only once per time step. This results in 98 880 parameters. Comparing the size of the SQAIR model, we obtain a conclusion similar to the one for the case of IN.

Note that the last frames in Figure 3 are sampled from the learned propagation prior. This enables us to evaluate the role of the relational module, as it is responsible for learning the object dynamics. Moreover, as the models are stochastic, we train 5 models for each architecture and sample 5 different last frames. We compare models in terms of data log-likelihood and relational log-likelihood, which takes into account *only* the objects which are *currently colliding* (ground truth available in the dataset). The evaluation on the test set with 4 balls shows an increase in average data log-likelihood from 399.5 achieved by SQAIR (0.21 relational) to 429.2 by R-SQAIR(IN) (relational 1.95) and 457.32 by R-SQAIR(RMC) (relational 3.62). Error bars in Figure 4 represent the standard deviation of the stochastic samples from the trained models.

We test generalization of R-SQAIR by evaluating the models trained on sequences with 4 balls on a test set with videos of 6-8 balls. Both qualitative (Figure 3 bottom row) and quantitative results show that R-SQAIR is capable of generalizing, with increase in relational log-likelihood from -164.1 achieved by SQAIR to -96.7 achieved by R-SQAIR(IN) and -97 achieved by R-SQAIR(RMC). Larger margins between relational losses of R-SQAIR and SQAIR on the test set with 6-8 balls suggest higher generalization capabilities of R-SQAIR.

4 Conclusion

Graph neural networks are promising candidates for combinatorial generalization, a central theme of AI research [3, 31]. We show that a sequential attention model can benefit from incorporating an explicit relational module which infers pairwise object interactions in parallel. Without retraining, the model generalizes to scenarios with more objects. Its learned generative model is potentially useful as part of a world simulator [24, 23, 13, 33].

Acknowledgments

We would like to thank Adam R. Kosiorek, Hyunjik Kim, Ingmar Posner and Yee Whye Teh for making the codebase for the SQAIR model [18] publicly available. This work was made possible by their commitment to open research practices. We thank Sjoerd van Steenkiste for helpful comments and fruitful discussions. This research was supported by the Swiss National Science Foundation grant 407540_167278 EVAC - Employing Video Analytics for Crisis Management. We are grateful to NVIDIA Corporation for a DGX-1 as part of the Pioneers of AI Research award, and to IBM for donating a "Minsky" machine.

References

[1] R. Baillargeon, E. S. Spelke, and S. Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985.

[2] P. Battaglia, R. Pascanu, M. Lai, D. J. Rezende, et al. Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*, pages 4502–4510, 2016.

[3] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[4] Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[5] C. P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, and A. Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

[6] E. Crawford and J. Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. 2019.

[7] S. A. Eslami, N. Heess, T. Weber, Y. Tassa, D. Szepesvari, G. E. Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances in Neural Information Processing Systems*, pages 3225–3233, 2016.

[8] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[9] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471, 2016.

[10] K. Greff, R. L. Kaufmann, R. Kabra, N. Watters, C. Burgess, D. Zoran, L. Matthey, M. Botvinick, and A. Lerchner. Multi-object representation learning with iterative variational inference. *arXiv preprint arXiv:1903.00450*, 2019.

[11] K. Greff, A. Rasmus, M. Berglund, T. Hao, H. Valpola, and J. Schmidhuber. Tagger: Deep unsupervised perceptual grouping. In *Advances in Neural Information Processing Systems*, pages 4484–4492, 2016.

[12] K. Greff, S. van Steenkiste, and J. Schmidhuber. Neural expectation maximization. In *Advances in Neural Information Processing Systems*, pages 6691–6701, 2017.

[13] D. Ha and J. Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, pages 2450–2462, 2018.

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[15] J.-T. Hsieh, B. Liu, D.-A. Huang, L. F. Fei-Fei, and J. C. Niebles. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, pages 517–526, 2018.

[16] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu. Spatial transformer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2017–2025. Curran Associates, Inc., 2015.

[17] P. J. Kellman and E. S. Spelke. Perception of partly occluded objects in infancy. *Cognitive psychology*, 15(4):483–524, 1983.

[18] A. R. Kosiorek, H. Kim, I. Posner, and Y. W. Teh. Sequential attend, infer, repeat: Generative modelling of moving objects. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 8615–8625, USA, 2018. Curran Associates Inc.

[19] I. Prémont-Schwarz, A. Ilin, T. Hao, A. Rasmus, R. Boney, and H. Valpola. Recurrent ladder networks. In *Advances in Neural Information Processing Systems*, pages 6009–6019, 2017.

[20] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko. Semi-supervised learning with ladder networks. In *Advances in neural information processing systems*, pages 3546–3554, 2015.

[21] A. Santoro, R. Faulkner, D. Raposo, J. Rae, M. Chrzanowski, T. Weber, D. Wierstra, O. Vinyals, R. Pascanu, and T. Lillicrap. Relational recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 7299–7310, 2018.

[22] R. Saxe and S. Carey. The perception of causality in infancy. *Acta psychologica*, 123(1-2):144–165, 2006.

[23] J. Schmidhuber. Making the world differentiable: On using fully recurrent self-supervised neural networks for dynamic reinforcement learning and planning in non-stationary environments. Technical Report FKI-126-90 (revised), Institut für Informatik, Technische Universität München, November 1990. (Revised and extended version of an earlier report from February.).

[24] J. Schmidhuber. An on-line algorithm for dynamic reinforcement learning and planning in reactive environments. In *Proc. IEEE/INNS International Joint Conference on Neural Networks, San Diego*, volume 2, pages 253–258, 1990.

[25] J. Schmidhuber. On decreasing the ratio between learning complexity and number of time-varying variables in fully recurrent nets. In *Proceedings of the International Conference on Artificial Neural Networks, Amsterdam*, pages 460–463. Springer, 1993.

[26] J. Schmidhuber and R. Huber. Learning to generate artificial fovea trajectories for target detection. *International Journal of Neural Systems*, 2(1 & 2):135–141, 1991.

[27] E. S. Spelke, R. Kestenbaum, D. J. Simons, and D. Wein. Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British Journal of Developmental Psychology*, 13(2):113–142, 1995.

[28] K. Stelzner, R. Peharz, and K. Kersting. Faster attend-infer-repeat with tractable probabilistic models. In *International Conference on Machine Learning*, pages 5966–5975, 2019.

[29] S. Sukhbaatar, J. Weston, R. Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.

[30] S. van Steenkiste, M. Chang, K. Greff, and J. Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *International Conference on Learning Representations*, 2018.

[31] S. van Steenkiste, K. Greff, and J. Schmidhuber. A perspective on objects and systematic generalization in model-based rl. *arXiv preprint arXiv:1906.01035*, 2019.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[33] N. Watters, L. Matthey, M. Bosnjak, C. P. Burgess, and A. Lerchner. Cobra: Data-efficient model-based rl through unsupervised object discovery and curiosity-driven exploration. *arXiv preprint arXiv:1905.09275*, 2019.

[34] N. Watters, D. Zoran, T. Weber, P. Battaglia, R. Pascanu, and A. Tacchetti. Visual interaction networks: Learning a physics simulator from video. In *Advances in neural information processing systems*, pages 4539–4547, 2017.

[35] F. Xu and S. Carey. Infants' metaphysics: The case of numerical identity. *Cognitive psychology*, 30(2):111–153, 1996.

[36] J. Yuan, B. Li, and X. Xue. Generative modeling of infinite occluded objects for compositional scene representation. In *International Conference on Machine Learning*, pages 7222–7231, 2019.