Compositional Learning of Visually-Grounded Concepts Using Synthetic Environment

Zijun Lin*
Nanyang Technological University
linz0048@e.ntu.edu.sg

Haidi Azaman*
National University of Singapore
e0540498@u.nus.edu

M Ganesh Kumar
Centre for Frontier AI Research, A*STAR
m_ganeshkumar@u.nus.edu

Cheston Tan
Centre for Frontier AI Research, A*STAR
cheston_tan@cfar.a-star.edu.sg

Abstract

Children can rapidly generalize compositionally-constructed rules to unseen test sets. On the other hand, deep reinforcement learning (RL) agents need to be trained over millions of episodes, and their ability to generalize to unseen combinations remains unclear. Here, we investigate the compositional abilities of RL agents, using the task of navigating to instructed color-shape targets in synthetic 3D environments, which allows better control over the train-test split and balance within the data. First, we show that when RL agents are naively trained to navigate to target color-shape combinations, they implicitly learn to decompose the instruction, allowing them to (re-)compose and succeed at held-out test instructions ("compositional learning"). Second, when agents were pretrained to learn invariant shape and color concepts ("concept learning"), the number of episodes subsequently needed for compositional learning decreased by 20×. Furthermore, only agents trained on both concept and compositional learning could solve a more complex, out-of-distribution environment in zero-shot. Finally, we demonstrate that only text encoders pretrained on image-text datasets (e.g. CLIP) reduced the number of training episodes needed for our agents to demonstrate compositional learning, and also generalized in zero-shot to five new colors unseen during training. ***Overall, our results are the first to demonstrate that RL agents can leverage synthetic data to implicitly learn concepts and compositionality, to solve more complex 3D environments in zero-shot without needing additional training episodes.***

*Equal contribution

1. Introduction

Compositionality is the ability to follow specific rules in putting together basic units of information or primitives [16]. To use an informal everyday example, following the instructions from a recipe book, one can prepare a large combination of new food dishes even with a small set of grocery items. Learning to compose basic primitives allows one to generate almost infinitely many solutions to solve complex tasks [17]. However, learning multimodal primitives and composing them to solve complex tasks in the vision-language-action domain is a significant gap in existing reinforcement learning (RL) agents [14].

Hierarchical RL agents can learn action primitives and compose them to solve complex tasks [2], if the agents are aided using symbolic methods to explore the sensorimotor space. However, naively training the parameters in RL agents end-to-end requires millions of training episodes [1] and they lack the ability to generalize to out-of-distribution tasks [9], making them undeployable in the real world [6].

For humans, children first learn individual concepts or schemas [5, 11, 15] by associating visual and verbal cues, and interacting with their environment [3, 8]. Learning of concepts in the multi-modal space facilitates subsequent rapid learning of compositional tasks.[10, 12].

Hence, we developed several synthetic 3D environments to train vision-language based navigation agents to associate visual primitives with language-based instructions, by navigating to the correctly referenced object for a reward. **Notably, the synthetic environment we've developed enables precise control over the generation of training sets and held-out test sets, while also ensuring data balance across various classes.** It facilitates demonstrating the compositionality of the agents and is hard to achieve with realistic datasets. Our contributions are as follows:

- We demonstrated the ability of RL agents to learn to de-

compose color-shape instructions and then (re-)compose them, generalizing to held-out color-shape combinations.

- We found that learning of invariant concepts enabled rapid compositional learning of concept combinations, with 100× and 20× speed-up for train and test combinations respectively.
- We showed for the first time that invariant concept learning that is followed by compositional learning enables the zero-shot ability to handle complex, out-of-distribution compositional tasks.
- We found that RL agents with certain pretrained language encoders reduced the number of training episodes needed for compositional learning, and also generalize to out-of-distribution unseen combinations in zero-shot.

2. Synthetic Environments for Grounded Learning

As shown in Figure 1, the 3D environments were developed to learn two key concepts: Shape (S) and Color (C). The environments contain objects made up of five distinct shapes, which are capsule, cube, cylinder, prism, and sphere, and five different colors: red, green, blue, yellow, and black. Hence, each object can be described using the shape and color attributes, for example "red sphere", "blue capsule", or "yellow prism". A target object will be randomly spawned at one of four predetermined locations within a rectangular room.

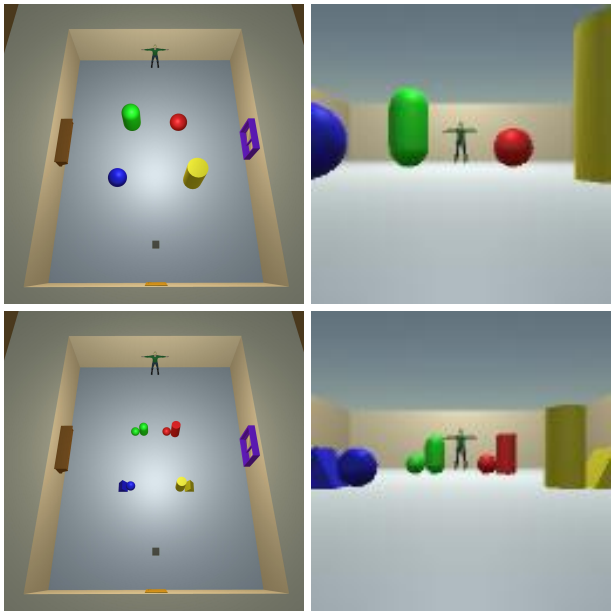

Figure 1. Two example environments. Left column shows the top-view of the environments. Right column shows the RL agent's first-person view. Top and bottom rows show the C&S environment (target instruction is "red sphere") and the C&S&S environment (target instruction is "red sphere cylinder") respectively.

The agent is built with one-hot text encoder, a CNN visual model, LSTM module, and updates its parameters by A2C algorithm. During each episode, successfully navigating to the target object yields a reward of +10, while collisions with non-target objects or walls incur penalties of -3 and -1, respectively. Additionally, the agent receives a penalty of -10 upon reaching the maximum allowed steps of 500. To ascertain the successful learning of a task by the agent, we establish a **Performance Criterion** of $+9$. The agent is deemed to have effectively learned the task when it obtains the average episode reward ≥ 9 over 100 training episodes.

3. Experiments and Results

3.1. Experiment 1: Generalization of Compositional Learning

Visually grounded agents can understand single feature instructions [7]. How navigation agents learn and compose multiple attributes is unclear. Hence, experiment 1 expands on single attribute navigation to the combination of two attributes, Color + Shape (C&S).

Shape\Color	Red	Green	Blue	Yellow	Black
Capsule	**Test**	Train	Train	Train	Train
Cube	Train	**Test**	Train	Train	Train
Cylinder	Train	Train	**Test**	Train	Train
Prism	Train	Train	Train	**Test**	Train
Sphere	Train	Train	Train	Train	**Test**

Table 1. Train-Test split for environments C|S and C&S.

To understand whether an RL agent is able to learn to decompose instructions given during training to learn each word group and recompose them to solve color-shape test combinations, the agent is trained on 20 C&S instructions and tested on 5 held-out C&S pairs in this experiment as described in Table 1. For example, the agent is trained on the instructions and visual targets "black cube" and "red sphere". After every 100 training episodes, it is tested on its ability to compose the concepts of "black" and "sphere" to accurately navigate to the visual target "black sphere" when given the held-out test combination instruction.

In this experiment, three agents were trained ($N = 3$) and their mean of the average reward are being compared. The average reward is calculated by taking the average across the most recent 100 episodes after each episode. The results show that **the agent with the one-hot text encoder requires approximately 67,000 and 95,000 episodes to achieve performance criterion ($\geq$ 9) for the 20 training and 5 held-out test combinations.** This result demonstrates that agents can gradually learn to decompose the color-shape instructions and ground them to the visual at-

tributes during training such that they can recompose the individual concepts to solve the held-out test combinations.

3.2. Experiment 2: Concept Learning Speeds Up and Generalizes Compositional Learning

Concept learning (C|S) aims to train the agents learning the individual concept first. The results in Table 3 show that the pretrained agents achieved train and held-out test performance criterion in the compositional learning environment 100× and 20× faster than naively-trained agents. This suggests that feature and policy learning requires more training episodes than compositional learning. Importantly, pretraining was not only to learn features or policy but to learn the concept of Shape and Color as well.

Concept learning turns out to be more difficult than naively learning to compose as agents need to learn shape and color invariance. For instance, when given the instruction "black", the agent needs to navigate to a black object, learning to ignore its shape. Each color or shape has five corresponding shape or color invariants respectively. However, with the instruction "black cube", the visual attribute is fixed as both color and shape are specified.

So far, agents have only been trained and tested on color and shape combinations. However, in the real-world, instructions are highly compositional, going beyond two word color-shape phrases. Furthermore, real-world objects can be composed of two or more basic shapes, such as a hammer being composed of a cylinder and a cuboid.

To investigate the generalizability of the agents, we created the C&S&S test-only environment (Figure 1, bottom) where RL agents have to compose three word instructions, one color and two shapes and navigate to the correct target which is composed of two objects, making the task more complex and difficult than C&S. These objects and instructions are grouped into Familiar Combinations (Table 2, column 5), while C&S&S evaluation combinations that do not overlap with the C&S train combinations are called Unseen Combinations (Table 2, column 6), making them truly novel and out-of-distribution composed objects and instructions.

Table 2 shows the zero-shot evaluation performance on the C&S&S environment of agents trained either on concept learning (C|S), compositional learning (C&S) or both (C|S → C&S). For the agent exclusively trained on compositional learning, two training checkpoints were selected based on training episodes. The first is the 67.4K checkpoint at which the agent achieved performance criterion. The second checkpoint is at 168.6K which is an agent that is trained for the same number of training episodes as agents undergoing both concept and subsequently compositional learning (last row).

The best performing agents were the ones that were trained on C|S first and then C&S, achieving rewards of 5.5 on both the familiar and unseen C&S&S combinations in zero-shot. None of the other agents demonstrated equivalent zero-shot proficiency in the C&S&S environment, even if the agent was trained for the same number of training episodes on the compositional learning task. Interestingly, even though the agent was pretrained on C|S and then trained on C&S, it maintained its zero-shot performance on the C|S combinations, implying that its foundational conceptual knowledge of colors and shape was not forgotten. These results suggest that out-of-distribution generalization cannot be obtained by simply learning features, composition and navigation policy, but rather foundational concepts need to be learned first before learning to compose them to solve more complex tasks.

Furthermore, Table 2 shows that compared to the agent trained on C&S for 67.4K episodes, agents trained only on C|S perform slightly worse on the familiar C&S&S combinations, but better on the unseen C&S&S combinations. While these agents' zero-shot generalization performance are nowhere close to the sequentially-trained agent's performance (C|S→C&S), they perform much better than the agents that were over-trained in C&S for 168.6K episodes, as well as the agents that were not trained in any environment so as to show chance performance ("Nil").

3.3. Experiment 3: Comparison of Text Encoders

Thus far, we have seen an agent with a vanilla visual module and one-hot encoded text module can be trained end-to-end using a reinforcement learning objective. We next hypothesized that substituting vanilla one-hot text module with pretrained but frozen text encoders such as CLIP [13] and BERT [4] might reduce the number of training episodes needed to achieve performance criterion. Additionally, we constructed vanilla text encoders utilizing the CLIP tokenizer, and the tokens are passed through an embedding layer, a pooling layer and a feed-forward layer. Table 4 shows the average number of training episodes needed for three agents ($N = 3$) initialized with different text encoders to achieve performance criterion in the C&S environment on both the train and test combinations.

The vanilla text encoder used CLIP's tokenizer to encode the various colors and shapes into orthogonal tokens, similar to the one-hot encoder, making it easier for the vanilla text encoder to disentangle the embedding even before training. However, training from scratch required about two times the training episodes of the one-hot encoder to achieve performance criterion for the train and test combinations. The longer training episodes could be due to the larger number of parameters in the embedding layer, which is absent in the one-hot text encoder.

Although the agent with the BERT text encoder was trained for a maximum of 200,000 episodes, the agent only achieved an average reward of 8.5, failing to reach performance criterion for testing combinations. The CLIP

		Average reward for zero-shot Evaluation in Environments			
Training Environment	**Training Episodes (K)**	**Familiar C|S combo**	**Unseen C|S combo**	**Familiar C&S&S combo**	**Unseen C&S&S combo**
Nil	0.0	-24.42 ± 1.29	-23.42 ± 2.57	-29.15 ± 3.07	-36.48 ± 3.60
C&S	67.4	0.37 ± 1.50	3.08 ± 0.32	2.84 ± 0.92	-5.10 ± 2.59
C&S	168.6	-8.02 ± 3.01	-2.05 ± 1.57	-8.27 ± 3.75	-23.2 ± 9.97
C|S	168			1.19 ± 1.24	-4.02 ± 2.44
C|S → C&S	168 → 0.6	$\mathbf{8.74 \pm 0.29}$	$\mathbf{7.58 \pm 0.32}$	$\mathbf{5.49 \pm 0.26}$	$\mathbf{5.55 \pm 0.39}$

Table 2. Summary of zero-shot evaluation experiments. The values are the mean and standard deviation of the rewards obtained by ($N = 3$) agents over 100 episodes in the novel environments with color and shape combinations previously trained on (familiar) and untrained on (unseen). Higher reward values indicate better performance.

	Episodes (K) for performance criterion	
Training Environment	**Train combinations**	**Held-out Test combinations**
C&S	67.4 ± 7.2	94.8 ± 3.7
C|S → C&S	0.6 ± 0.1	5.5 ± 2.9

Table 3. Mean and standard deviation of the number of episodes (in **thousands**) required to achieve performance criterion over three repeats. These experiments were run in the C&S environments, and the results compare agents trained from scratch (i.e. row C&S), versus with pretraining on C|S (i.e. row C|S → C&S). Lower values indicate faster learning.

	Training Episodes (K)	
Text Encoder	**Train combinations**	**Test combinations**
One-hot	67.4 ± 7.2	94.8 ± 3.7
Vanilla	116.2 ± 15.4	185.9 ± 15.5
BERT	109.0 ± 9.1	≥ 200
CLIP	$\mathbf{56.2 \pm 5.3}$	$\mathbf{72.6 \pm 6.0}$

Table 4. Init → C&S: Learning to decompose instructions for compositional performance. Values in the table indicate the mean of episodes (in **thousands**) needed to achieve performance criterion across three repeats in training and testing environments. Lower values indicate faster learning.

text encoder achieved performance criterion on both the train and test combinations 1.3 times faster than the one-hot text encoder agent. The expedited learning implies that CLIP's prelearned word embedding is useful to increase the training efficiency for a reinforcement learning agent. This result demonstrates that it is possible to use vision-language grounded models to improve the training efficiency of multi-modal reinforcement learning agents, especially for compositional learning.

To assess the agent's ability to generalize to out-of-distribution instructions not encountered during training, the agent with CLIP text encoder was tested in the C*&S environment, where five new colors (Orange, Cyan, Pink, Purple, and White) were introduced. Example instructions were "cyan prism" or "purple cube", but "cyan" and "purple" were never presented during training. **The agent achieves a noteworthy zero-shot result, averaging 6.9 over three repeats (100 test episodes at each repeat)**, demonstrating that agents trained on the original five colors, equipped with a CLIP text encoder, successfully adapted to the five novel colors in the C*&S environment. Thus, the abilities of these agents are not limited to the 25 color and shape combinations, and can comprehend words not encountered previously.

4. Conclusion

We developed several 3D synthetic environments to demonstrate the compositional abilities of reinforcement learning agents. Specifically, we found that agents can learn to decompose and recompose instructions to solve held-out test instructions. Furthermore, we showed that invariant concept learning accelerates compositional learning. Finally, we tested various text encoders, with the CLIP showing the ability to speed up learning.

Acknowledgments

This work was supported by A*STAR through a CRF award (C.T.), as well as ARIA (H.A.) and CIARE (Z.L.) internships.

References

[1] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017. 1

[2] Andrew G Barto and Sridhar Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 2003. 1

[3] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Expe-

rience grounds language. *arXiv preprint arXiv:2004.10151*, 2020. 1

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, 2019. 3

[5] Susan A Gelman and Ellen M Markman. Categories and induction in young children. *Cognition*, 23(3):183–209, 1986. 1

[6] Theophile Gervet, Soumith Chintala, Dhruv Batra, Jitendra Malik, and Devendra Singh Chaplot. Navigating to objects in the real world. *Science Robotics*, 2023. 1

[7] Felix Hill, Stephen Clark, Phil Blunsom, and Karl Moritz Hermann. Simulating early word learning in situated connectionist agents. In *Annual Meeting of the Cognitive Science Society*, 2020. 2

[8] Jana M Iverson. Developing language in a developing body: The relationship between motor development and language development. *Journal of Child Language*, 37(2):229–261, 2010. 1

[9] Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76:201–264, 2023. 1

[10] M Ganesh Kumar, Cheston Tan, Camilo Libedinsky, Shih-Cheng Yen, and Andrew Yong-Yi Tan. One-shot learning of paired association navigation with biologically plausible schemas. *arXiv preprint arXiv:2205.09710*, 2023. 1

[11] Jason F Macario. Young children's use of color in classification: Foods and canonically colored objects. *Cognitive Development*, 6(1):17–46, 1991. 1

[12] James L McClelland. Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General*, 142(4):1190, 2013. 1

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021*, 2021. 3

[14] Nicholas Roy, Ingmar Posner, Tim Barfoot, Philippe Beaudoin, Yoshua Bengio, Jeannette Bohg, Oliver Brock, Isabelle Depatie, Dieter Fox, Dan Koditschek, et al. From machine learning to robotics: challenges and opportunities for embodied intelligence. *arXiv preprint arXiv:2110.15245*, 2021. 1

[15] David E Rumelhart and Andrew Ortony. The representation of knowledge in memory. *Schooling and the Acquisition of Knowledge*, 99:135, 1977. 1

[16] Zoltán Gendler Szabó. Compositionality. In *Stanford Encyclopedia of Philosophy*. 2008. 1

[17] Emanuel Todorov and Zoubin Ghahramani. Unsupervised learning of sensory-motor primitives. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1750–1753. IEEE, 2003. 1