HUMAN-LIKE COMPOSITIONAL LEARNING OF VISUALLY-GROUNDED CONCEPTS USING SYNTHETIC ENVIRONMENTS

Zijun Lin[1,3] **M Ganesh Kumar**[2,3] **Cheston Tan**[3]
[1]Nanyang Technological University
[2]Harvard University
[3]Centre for Frontier AI Research, A*STAR

ABSTRACT

The compositional structure of language enables humans to decompose complex phrases and map them to novel visual concepts, showcasing flexible intelligence. While several algorithms exhibit compositionality, they fail to elucidate how humans learn to compose concept classes and ground visual cues through trial-and-error. To investigate this multi-modal learning challenge, we designed a 3D synthetic environment in which an agent learns, via reinforcement, to navigate to a target specified by a natural language instruction. These instructions comprise nouns, attributes, and critically, determiners, prepositions, or both. The vast array of word combinations heightens the compositional complexity of the visual grounding task, as navigating to a blue cube above red spheres is not rewarded when the instruction specifies navigating to "some blue cubes below the red sphere". We first demonstrate that reinforcement learning agents can ground determiner concepts to visual targets but struggle with more complex prepositional concepts. Second, we show that curriculum learning—a strategy humans employ—enhances concept learning efficiency, reducing the required training episodes by 15% in determiner environments and enabling agents to easily learn prepositional concepts. Finally, we establish that agents trained on determiner or prepositional concepts can decompose held-out test instructions and rapidly adapt their navigation policies to unseen visual object combinations. Leveraging synthetic environments, our findings demonstrate that multi-modal reinforcement learning agents can achieve compositional understanding of complex concept classes and highlight the efficacy of human-like learning strategies in improving artificial systems' learning efficiency.

1 INTRODUCTION

Humans describe the world by combining linguistic parts-of-speech (concept classes) such as adjectives and nouns. As visual complexity increases, we use more words (Sun & Firestone, 2022) and additional concept classes to enhance description accuracy (Elsner et al., 2018; Qiao et al., 2020). For example, determiners are essential for describing object quantity and ownership (Lee et al., 2023), while prepositions are critical for expressing spatial and temporal relationships between objects (Agrawal et al., 2023). In contrast, pretrained vision-language models struggle to ground visual objects to complex linguistic concepts and flexibly compose this knowledge to solve novel scenarios (Kamath et al., 2023; Shen et al., 2023). This gap restricts the alignment between human and machine communication for collaboration.

Children typically master the use of determiners and prepositions by age three (Abu-Akel et al., 2004; Brown, 1973), applying these concepts to describe novel situations and generalizing beyond their experiences (Tomasello, 1987; Washington & Naremore, 1978). They learn incrementally, beginning with simple terms and advancing to complex concepts (Valian, 1986; Gleason & Ratner, 2022; Richards, 1984). The ability to compose linguistic elements for detailed descriptions, such as "some red balls above the green box," is a skill even advanced vision-language models lack (Okawa et al., 2024). Tailoring learning curricula to individual pace and knowledge remains an ongoing chal-

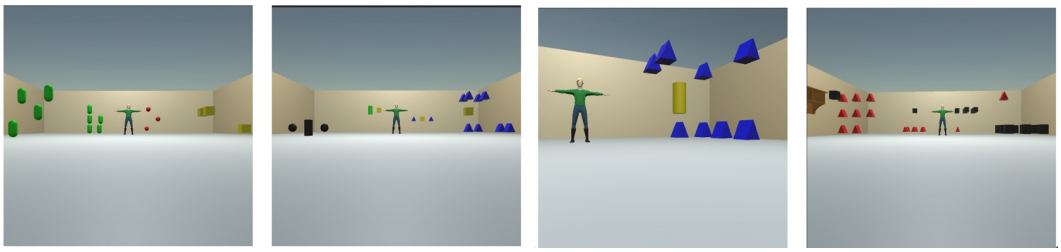

Figure 1: Example environments, each with four target options. A reward of +10 is given when the agent navigates to the target matching the instruction. Punishments of -1, -3, and -10 are incurred for hitting a wall, reaching the wrong target, or failing to reach the correct target, respectively. **A**: Agent's view of the determiner (D) environment. **B**: Agent's view of the preposition (P) environment. **C**: Close-up view of the target object in P. **D**: Agent's view of the combined determiner and preposition ($D + P$) environment.

lenge, both for humans and machine learning systems (Bengio et al., 2009; Graves, 2008; Soviany et al., 2022; Wang et al., 2021; Narvekar et al., 2020).

In this work, we developed several 3D synthetic environments to study the visuo-linguistic grounding problem in a reinforcement learning setting, similar to how children could learn these concepts. A trial-and-error based learning environment facilitates artificial agents to learn from experience, a requisite to continually improve alignment during human-machine interaction. Specifically, we investigate the influence of a curriculum in improving compositional learning to ground determiner and preposition concept classes to visual objects. **Notably, our synthetic environments allow precise control over the generation of training and held-out test sets while ensuring balanced data across classes, a challenge with real-world datasets.** Leveraging on synthetic environments, we make the following contributions:

- We demonstrate that reinforcement learning agents, when trained naively to maximize rewards, can ground descriptive attributes (determiners) to single objects but struggle to ground relational attributes between objects (prepositions).
- We show that agents cannot generalize determiners and prepositions to new combinations without incremental curriculum training from simple to complex tasks.
- We establish that agents can reason compositionally about novel I.I.D. and O.O.D. test instructions and navigate to unknown visual objects after learning determiner and preposition concepts.

2 Synthetic Environments for Grounding

This section describes the synthetic environments designed for agents to learn to ground visual objects to determiner and preposition linguistic concepts by trial-and-error learning.

2.1 Environment Design

As shown in Fig. 1, we built three synthetic 3D environments to enable the agents to learn four concepts — shape, color, determiner and preposition. While our primary emphasis is on determiner and preposition, the agents also need to acquire knowledge of color and shape throughout the learning process. Specifically, objects come in random colors (red, green, blue, yellow and black) and shapes (capsule, cube, cylinder, prism and sphere), each chosen from a set of 5.

The target object is randomly placed in one of four predetermined locations within a rectangular room, as illustrated in Fig. 1. Additionally, the positions of the individual objects within each location are randomized within a 3×3×3 block. This creates numerous visual representations that are constrained by the description of the instruction. The room layout remains constant and includes fixed visual landmarks like a door, window, shelf, and a reference man. The first-person perspective is simulated through a Unity-based camera, capturing the environmental dynamics in RGB images.

These images serve as the visual input for the agent. Additionally, the environment generates textual instructions describing the target object, which are provided to the agent as language input. Details of each environment follows below.

2.2 DETERMINER ENVIRONMENT

In the determiner environment (D), agents learn eight determiners: "A", "Few", "Some", "Many", "This", "That", "These", and "Those". Instructions follow the form "Determiner + Object(s)". For example, "*few* yellow cubes" or "*those* green cylinders" (Fig. 1A). The first four determiners describe object quantity: "A" denotes a single object, 2–3 objects as "Few", 4–6 as "Some", and 7–9 as "Many". The last four determiners depend on proximity to a reference man (Fig. 1A): "This" and "These" refer to closer objects (single or multiple 2∼9), while "That" and "Those" refer to farther objects.

Non-targets are designed to test grounding: (1) same determiner, random color/shape; (2) same color/shape, random determiner; (3) random determiner, color, and shape. For example, given the target instruction "*Many* red cylinder", non-target visual objects could be "*Few* red cylinder", "Many *blue cube*", and "*A green prism*".

2.3 PREPOSITION ENVIRONMENT

In the preposition environment (P), agents learn eight prepositions: "Above", "Below", "In front of", "Behind", "Beside", "On", "Between", and "Among". Instructions follow the form "Object A + Preposition + Object B", where Object A and Object B differ in color and shape. For the first six prepositions, Object A and Object B are single instances (e.g., "green capsule *On* red cube"). For "Between" and "Among", Object A is singular, while Object B consists of 2–9 instances, respectively.

Non-targets are designed similarly: (1) same preposition, random color/shape; (2) same color/shape, random preposition; (3) random preposition, color, and shape. For example, given the target instruction "green cube *Above* blue cylinder", non-target visual objects could be "green cube *In front of* blue cylinder", "*yellow cube* Above *black capsule*", and "*red cylinder Among green cube*".

2.4 COMBINED DETERMINER AND PREPOSITION ENVIRONMENT

The combined environment ($D + P$) integrates determiners and prepositions. Instructions follow the form "Determiner A + Object A + Preposition + Determiner B + Object B". For example, "a red cube Above Many black cubes" (Fig. 1D).

This environment evaluates the influence of curriculum learning and the ability to decompose complex instructions in few-shot. Non-targets are designed to test comprehension of all concepts: (1) swapped color/shape attributes; (2) altered determiners; (3) modified prepositions. For example, given the target instruction "A black cube Above Many red prisms", non-target visual objects could be A "*red prism* Above Many *black cubes*", "*Many* black cubes Above *A* red prisms", and "A black cube *Behind* Many red prisms".

2.5 EVALUATING COMPOSITIONAL LEARNING AND GENERALIZATION

In each episode, the agent receives specific rewards based on its actions. Navigating to the target object yields a reward of +10, while collisions with non-target objects or walls result in penalties of -3 and -1, respectively. Additionally, the agent incurs a penalty of -10 if it fails to reach the target within the maximum allowed steps ($T_{max} = 500$). We define the *Performance Criterion* as the agent achieving a perfect reward of +10 in at least 800 out of 1000 episodes, corresponding to an 80% success rate. The number of training episodes required to meet this criterion in the determiner and preposition environments is detailed in Section 4.1.

Table 4 and Table 5 (supplementary material) show the train-test split setting for the D and P environments respectively. After training the agents on the training combinations, the agent's were evaluated on its zero-shot compositional generalization ability on the held-out test combinations.

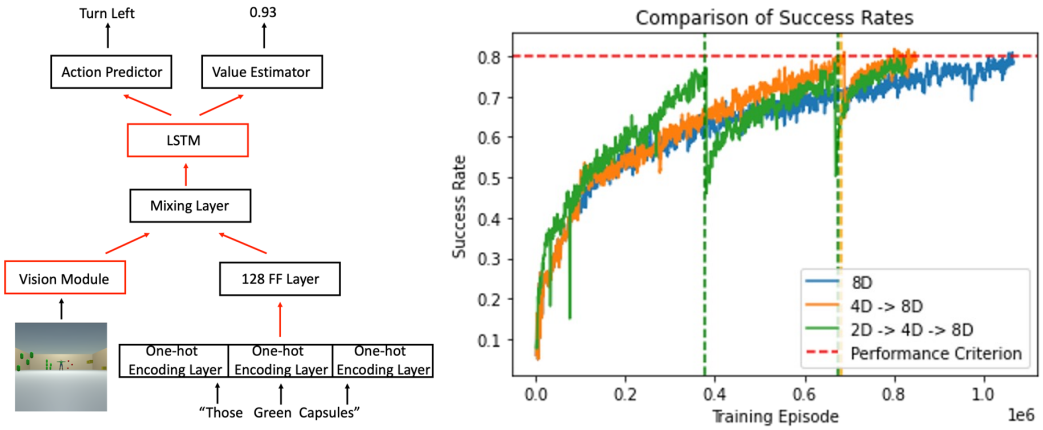

Figure 2: **Left**: Agent architecture. **Red** arrows or boxes represent **trainable** weights, while **black** arrows or boxes represent **frozen** weights. **Right**: Success rates of the agents with and without curriculum learning in D.

The results are in Section 4.2. Specifically, D comprises 200 unique instructions (160 train, 40 test), while P comprises a total of 6000 unique instructions (4800 train, 1200 test).

Naively training agents to learn the $D + P$ environment is intractable as $D + P$ contains 160,000 unique combinations (120,000 train, 40,000 test), which is 800 and 27 times more combinations than the D and P environments respectively. Hence, we used the $D + P$ environment to evaluate agents' ability to decompose novel determiner-preposition instructions. This is by first training the agents on the determiner D and prepositions P environment separately, until they reached performance criterion, and subsequently train them in the $D + P$ environment for only 100,000 episodes.

3 Agent Architecture

Our vision-language agent architecture, adapted from Hill et al. (2020); Lin et al. (2023), processes visual inputs (a $3\times128\times128$ tensor of RGB pixel values) through a vision module consisting of three convolutional layers. In the Determiner environment (D), language instructions are encoded as three one-hot vectors representing the Determiner, Color, and Shape components. For the Preposition environment (P), five one-hot vectors are used, while in the combined environment ($D + P$), seven one-hot vectors represent the instructions. These are passed to a linear layer, then concatenated with the vision module's output to form an embedding, which is processed by a Long Short Term Memory (LSTM) module, whose activity s_t informs both the action predictor (actor) and value estimator (critic). The actor derives a probability distribution $\pi(a_t|s_t)$ over the actions (move forward, move backward, turn left, turn right), while the critic estimates the state-value function $V(s_t)$.

The agent's objective is to maximize cumulative discounted rewards (Sutton, 2018; Kumar et al., 2024) by navigating to correct visual targets based on language instructions (Wang et al., 2019; Kumar et al., 2022), while avoiding incorrect options incurring penalties. Training leverages the advantage actor-critic (A2C) algorithm (Mnih et al., 2016; Kumar et al., 2021a), optimized with RMSProp at a learning rate of 2.5×10^{-4} across all experiments.

4 Experiments and Results

This section evaluates the effects of a curriculum on compositional learning and generalization in grounding determiner and preposition concepts to visual objects. In experiment 1, we investigate the role of using curriculum to improve learning efficiency. Here, the agent has to ground either eight determiners or prepositions to visual objects in the Determiners or Preposition environments.

Experiment 2 examines the I.I.D. and O.O.D. generalization capabilities of agents. Agents' I.I.D. generalization capabilities are examined on held-out test instruction combinations in either the Determiners (D) or Preposition (P) environments in zero-shot (no training). O.O.D. generalization is

Determiner Environment	Training Episodes (M)	Total Episodes (M)
$8D$	0.87	0.87
$4D \rightarrow 8D$	$0.66 \rightarrow 0.11$	0.77
$2D \rightarrow 4D \rightarrow 8D$	$0.38 \rightarrow 0.33 \rightarrow 0.11$	0.82

Table 1: Curriculum learning ($4P \rightarrow 8P$ and $2P \rightarrow 4P \rightarrow 8P$) reduces the total number of episodes needed to learn the Determiner (D) environment, compared to without ($8P$) a curriculum. Values in the table indicate the number of episodes (in **millions**) needed to achieve a success rate $\geq 80\%$ (performance criterion) over 1000 episodes. Lower values indicate faster learning.

Preposition Environment	Training Episodes (M)	Total Episodes (M)
$8P$	> 3.5	> 3.5
$4P \rightarrow 8P$	$1.57 \rightarrow 1.3$	2.87
$2P \rightarrow 4P \rightarrow 8P$	$0.9 \rightarrow 0.3 \rightarrow 0.98$	2.18

Table 2: Curriculum learning ($4P \rightarrow 8P$ and $2P \rightarrow 4P \rightarrow 8P$) reduces the total number of episodes needed to learn the Prepositions (P) environment, compared to without ($8P$) a curriculum. Values in the table indicate the number of episodes (in **millions**) needed to achieve a success rate $\geq 80\%$ (performance criterion) over 1000 episodes. Lower values indicate faster learning.

evaluated by using agents pretrained on the Preposition (P) environments, to solve the Determiner and Preposition environment ($D + P$) in few-shot training. This demonstrates whether agents can rapidly adapt to an environment that contains a significantly greater number of potential combinations of instructions.

4.1 Experiment 1: Curriculum improves learning efficiency

To emulate human-like learning, we employ a curriculum, where agents progressively learn simpler concepts before advancing to more complex ones. For both the determiner and preposition environment, agents are trained under three setups: (1) learning all eight concepts directly ($8D$/$8P$), (2) learning four simple concepts first, then all eight e.g. ($4D \rightarrow 8D$), and (3) learning two, then four, and finally all eight e.g. ($2D \rightarrow 4D \rightarrow 8D$).

The results in Table 1 and Fig. 2 (right) show that curriculum learning reduces the total training episodes by 15% in the $4D \rightarrow 8D$ setup compared to learning all eight determiners from scratch. While the $2D \rightarrow 4D \rightarrow 8D$ setup does not save more episodes than $4D \rightarrow 8D$, it still achieves a modest reduction of 0.05M episodes. These improvements are significant given the relatively small number of training instruction combinations in the determiner environment (160 combinations). Simple concepts were determined based on the fastest learning curves for each concept.

In contrast, the preposition environment (P) presents a much greater challenge, with 4800 instruction combinations—30 times more than the determiner environment. As shown in Table 2, agents trained directly on all eight prepositions ($8P$) fail to converge even after 3.5M episodes. This aligns with expectations, as the complexity of the $8P$ environment would require approximately 21M episodes to learn, based on the 0.87M episodes needed for the determiner environment. However, curriculum learning significantly improves learning efficiency: the $4P \rightarrow 8P$ and $2P \rightarrow 4P \rightarrow 8P$ setups reduce the required training episodes to 2.87M and 2.18M, respectively. This demonstrates that curriculum learning is essential for tackling environments with high combinatorial complexity.

4.2 Experiment 2: Compositional learning for I.I.D. and O.O.D. Generalization

We evaluated the I.I.D. zero-shot capabilities of agents on held-out test instruction combinations in the Determiner (D) and Preposition (P) environments. Although agents were exposed to all determiner or preposition instructions, they were trained only on a subset of determiner/preposition-shape combinations (Tables 4 and 5) to achieve an 80% success rate. As shown in Table 3, agents demonstrated consistent zero-shot performance on held-out test instructions, with success rates of at least 76%. In contrast, agents that did not converge during training (e.g., $8P$) exhibited chance-level performance (25%). These findings indicate that agents have acquired the ability to decompose and recompose concepts, enabling generalization to new I.I.D. instructions and visual targets during held-out tests.

Environment	Success Rate (%)
$8D$	77.5 ± 0.01
$4D \rightarrow 8D$	79.0 ± 0.06
$2D \rightarrow 4D \rightarrow 8D$	76.0 ± 0.24
$8P$	24.8 ± 0.85
$4P \rightarrow 8P$	76.1 ± 0.99
$2P \rightarrow 4P \rightarrow 8P$	76.4 ± 0.46

Table 3: Zero-shot generalization in D and P on held-out instructions. Values in the table indicate the mean and standard deviation of the success rate of the agents in the test environment over 3 iterations ($N = 3$) in percent. The RL agents are tested in the held-out instructions right after reaching the success rate of 80%, except for $8P$, which does not converge even after 3.5 million training episodes. The closer the results approach 80%, the better compositionality the agents show.

The combined Determiner and Preposition environment ($D + P$) presents a greater challenge, with 160,000 O.O.D. instruction combinations—a scale 25 times larger than the preposition environment and 750 times larger than the determiner environment. To address this complexity, we implemented two strategies: (1) pretraining the agent in the Preposition (P) environment using a $4P \rightarrow 8P$ curriculum, and (2) fine-tuning the agent for compositional learning in the $D + P$ environment for only **0.1M episodes**. Without these strategies, agent convergence was unattainable.

Remarkably, with only 2.97M training episodes, the agent achieved an average reward of 7.2 and a **success rate of 53%** over 1,000 testing episodes in the $D + P$ environment, which has 160,000 instruction combinations. This contrasts with naive training in the $8P$ environment with just 4,800 instruction combinations, where agents failed to converge even after 3.5M episodes. This demonstrates that once foundational concepts are learned, agents can rapidly adapt to new combinations and solve increasingly complex environments with significantly fewer episodes. Additionally, this highlights the importance of learning to compose concepts using a curriculum to tackle environments with extreme combinatorial complexity.

5 Conclusion, limitations and future work

We developed several 3D synthetic environments to illustrate the impact of curriculum learning on instruction based navigation tasks and demonstrated the compositional capabilities of reinforcement learning agents. Notably, we are the first to showcase the feasibility of grounding RL agents in complex instructions involving determiners and prepositions. Our findings reveal that agents can decompose and recompose instructions, akin to human intelligence, allowing them to effectively solve previously unseen I.I.D. and O.O.D. test cases in zero-shot and few-shot respectively. This work marks a significant step towards aligning human and machine interaction (e.g. collaborative robots, autonomous vehicles), as real-life referring expressions often extend beyond simple word forms or adjective-noun combinations to include numerical references and spatial relationships. The unity scripts and code for the environments and agents will be made publicly available upon acceptance.

The 3D environments crafted for this study utilized simple geometric shapes, such as "capsule" and "prism", raising concerns about the model's generalization to more realistic objects (Hill et al., 2020). Additionally, the absence of obstacles results in straightforward navigation, suggesting that agents trained in these environments may struggle with more complex navigation scenarios that involve obstacles (Anderson et al., 2018; Gu et al., 2022; Kumar et al., 2021b). Future research directions include exploring diverse model architectures and integrating pre-trained text and vision encoders (Shah et al., 2023). Investigating the optimal combination of determiners or prepositions in the curriculum, such as using 2 or 4 determiners to expedite learning, is also worthwhile. A thorough analysis of the model's representations is expected to provide insights into how concepts are embedded (Kumar et al., 2024; Lee et al., 2023), enhancing the model's generalization capabilities across diverse scenarios. Moreover, evaluating whether pretrained large language models can effectively ground visual objects to concept classes using these 3D environments offers another avenue of exploration. This could involve using LLMs to make inferences at each time step (Wang et al., 2021) or generating code to predict actions based on visual frames (Cloos et al., 2024).

REFERENCES

Ahmad Abu-Akel, Alison L Bailey, and Yeow-Meng Thum. Describing the acquisition of determiners in english: A growth modeling approach. *Journal of Psycholinguistic Research*, 33:407–424, 2004.

Palaash Agrawal, Haidi Azaman, and Cheston Tan. Stupd: A synthetic dataset for spatial and temporal relation reasoning. *arXiv preprint arXiv:2309.06680*, 2023.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3674–3683, 2018.

Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, pp. 1140–1149, 2021.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Roger Brown. *A first language: The early stages*. Harvard University Press, 1973.

Nathan Cloos, M Ganesh Kumar, Adam Manoogian, Christopher J Cueva, and Shawn A Rhoads. Generating and validating agent and environment code for simulating realistic personality profiles with large language models. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*, 2024.

Joachim De Beule and Benjamin K Bergen. On the emergence of compositionality. In *The Evolution of Language*, pp. 35–42. World Scientific, 2006.

Micha Elsner, Alasdair Clarke, and Hannah Rohde. Visual complexity and its effects on referring expression generation. *Cognitive science*, 42:940–973, 2018.

Jean Berko Gleason and Nan Bernstein Ratner. *The development of language*. Plural Publishing, 2022.

Kathleen Graves. The language curriculum: A social contextual perspective. *Language teaching*, 41(2):147–181, 2008.

Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2203.12667*, 2022.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

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. URL `https://api.semanticscholar.org/CorpusID:231793206`.

Ruixiang Jiang, Lingbo Liu, and Changwen Chen. Clip-count: Towards text-guided zero-shot object counting. *arXiv preprint arXiv:2305.07304*, 2023.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023.

Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 234–251, 2018.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

M Ganesh Kumar, Cheston Tan, Camilo Libedinsky, Shih-Cheng Yen, and Andrew Yong-Yi Tan. One-shot learning of paired associations by a reservoir computing model with hebbian plasticity. *arXiv preprint arXiv:2106.03580*, 2021a.

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:2106.03580*, 2021b.

M Ganesh Kumar, Cheston Tan, Camilo Libedinsky, Shih-Cheng Yen, and Andrew YY Tan. A nonlinear hidden layer enables actor–critic agents to learn multiple paired association navigation. *Cerebral Cortex*, 32(18):3917–3936, 2022.

M Ganesh Kumar, Blake Bordelon, Jacob A Zavatone-Veth, and Cengiz A Pehlevan. A model of place field reorganization during reward maximization. *bioRxiv*, pp. 2024–12, 2024.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Clarence Lee, M Ganesh Kumar, and Cheston Tan. Determinet: A large-scale diagnostic dataset for complex visually-grounded referencing using determiners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20019–20028, 2023.

Martha Lewis, Nihal V. Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H. Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models. arXiv:2212.10537 [cs.LG], 2023. URL `https://arxiv.org/abs/2212.10537`.

Zijun Lin, Haidi Azaman, M Ganesh Kumar, and Cheston Tan. Compositional learning of visually-grounded concepts using reinforcement. *arXiv preprint arXiv:2309.04504*, 2023.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*, 2023.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *The Journal of Machine Learning Research*, 21(1):7382–7431, 2020.

Maya Okawa, Ekdeep S Lubana, Robert Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *Advances in Neural Information Processing Systems*, 36, 2024.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. *arXiv preprint arXiv:2302.12066*, 2023.

Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3593–3602, 2019.

Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2020.

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, 2021.

Jack C Richards. Language curriculum development. *RELC journal*, 15(1):1–29, 1984.

Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pp. 492–504. PMLR, 2023.

Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts. *arXiv preprint arXiv:2303.07226*, 2023.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022.

Austin Stone, Huayan Wang, Michael Stark, Yi Liu, D Scott Phoenix, and Dileep George. Teaching compositionality to cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5058–5067, 2017.

Zekun Sun and Chaz Firestone. Seeing and speaking: How verbal "description length" encodes visual complexity. *Journal of Experimental Psychology: General*, 151(1):82, 2022.

Richard S Sutton. Reinforcement learning: An introduction. *A Bradford Book*, 2018.

Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. Galip: Generative adversarial clips for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14214–14223, 2023.

Open Ended Learning Team, Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, et al. Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*, 2021.

Bahareh Tolooshams, Andrew Song, Simona Temereanca, and Demba Ba. Convolutional dictionary learning based auto-encoders for natural exponential-family distributions. In *International Conference on Machine Learning*, pp. 9493–9503. PMLR, 2020.

Michael Tomasello. Learning to use prepositions: A case study. *Journal of child Language*, 14(1): 79–98, 1987.

Virginia Valian. Syntactic categories in the speech of young children. *Developmental psychology*, 22(4):562, 1986.

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6629–6638, 2019.

Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2021.

Dora S Washington and Rita C Naremore. Children's use of spatial prepositions in two-and three-dimensional tasks. *Journal of Speech and Hearing Research*, 21(1):151–165, 1978.

Guangyue Xu, Parisa Kordjamshidi, and Joyce Y Chai. Zero-shot compositional concept learning. *arXiv preprint arXiv:2107.05176*, 2021.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

A AGENT ARCHITECTURE

For vision input, RGB pixel values are passed into the vision module, which contains three convolutional layers, and the output flattened into a 3136 (a 64×7×7 tensor) dimensional embedding. In D, the language module takes in three one-hot vector embedding of the instructions, each representing the Determiner, Color and Shape. For example, an instruction in D such as "Those Green Capsules" is represented by three one-hot vectors respectively. Followed by the same rule, instructions in P are represented by five one-hot vectors considering that the instructions contain five words and $D + P$ adopts seven one-hot vectors. These one-hot vectors are fully connected to a 128 unit linear embedding layer. The 3136-D vector from the vision module and the 128-D vector from the language module are concatenated and fed into a 256-D linear mixing layer.

A Long Short Term Memory (LSTM) module takes the 256-dimensional embeddings as input. Its activity s_t is passed to both the action predictor (actor) and value estimator (critic). The action predictor maps the LSTM's activity to a probability distribution $\pi(a_t|s_t)$ over four possible actions, i.e., move forward, move backward, turn left and turn right. Meanwhile, the value estimator computes a scalar approximation of the agent's state-value function $V(s_t)$.

B ENVIRONMENTAL DESIGN

As shown in Table 4 and 5, 75% of color-shape combinations were used to train the agents on the relevant tasks and 25% held-out test combinations were used to evaluate the agents' zero-shot ability to use the rules learned for unseen combinations. For example, during the training phase, the agents are trained on the instructions such as "Few blue cylinder" and "A black cube". Once the agents reach the performance criterion in training environment, they are tested whether they could combine the concept "Few" and "black cube" to accurately navigate to "Few black cube".

Shape\Determiner	A	Few	Some	Many	This	That	These	Those
Capsule	**Test**	Train	Train	**Test**	Train	Train	Train	Train
Cube	Train	**Test**	Train	Train	**Test**	Train	Train	Train
Cylinder	Train	Train	**Test**	Train	Train	**Test**	Train	Train
Prism	Train	Train	Train	**Test**	Train	Train	**Test**	Train
Sphere	Train	Train	Train	Train	**Test**	Train	Train	**Test**

Table 4: Train-Test split for D.

Shape\Preposition	Above	Below	In front of	Behind	Beside	On	Between	Among
Capsule	**Test**	Train	Train	**Test**	Train	Train	Train	Train
Cube	Train	**Test**	Train	Train	**Test**	Train	Train	Train
Cylinder	Train	Train	**Test**	Train	Train	**Test**	Train	Train
Prism	Train	Train	Train	**Test**	Train	Train	**Test**	Train
Sphere	Train	Train	Train	Train	**Test**	Train	Train	**Test**

Table 5: Train-Test split for P.

C RELATED WORK

C.1 DETERMINER AND PREPOSITION

Large vision-language models, excel in zero-shot classification (Radford et al., 2021) and text-to-image generation (Tao et al., 2023), yet they face documented challenges in object counting (Paiss et al., 2023) and spatial relationship understanding (Lewis et al., 2023). Despite attempts to address these limitations through additional loss functions or fine-tuning (Jiang et al., 2023), performance improvements remain modest and fall short of perfection.

In the realm of agent navigation, instructions incorporating determiners (e.g., Few, Many) or prepositions (e.g., Above, Below) have received limited attention. Although these words are prominently used by humans since age three and play a fundamental role in realising human-AI interaction, there is a scarcity of studies in this area. Our research seeks to explore the generalization capabilities of agents trained by reinforcement learning to successfully reach target objects when provided with

instructions containing determiners or prepositions. Our environment dataset and analysis aims to provide deeper insights into how agents learn spatial relationships and acquire object counting capabilities in a human-like learning setting.

C.2 Curriculum Learning for Reinforcement learning

Curriculum learning can be divided the into two categories namely model-level curriculum learning and data-level curriculum learning (Soviany et al., 2022). The former is to dynamically adjust the model's complexity or structure as it learns, potentially making it more adept at handling progressively challenging tasks. The latter focuses on defining a difficulty criterion for the training data or task. The model begins learning from simpler tasks and gradually progress to more challenging ones. This entails organizing the training dataset in a way that facilitates a smooth transition from easy to difficult examples. The underlying principle is to guide the learning process by presenting the model with tasks of increasing complexity over time.

Inspired by how human learns and the non-convex optimization properties of reinforcement learning, curriculum learning has been proposed to improve policy convergence if the task order is well curated (Soviany et al., 2022; Wang et al., 2021; Narvekar et al., 2020). However, optimizing the dataset or task sequence is still not well understood and hence a specific curriculum may not always result in successful learning outcome.

Hill et al. (2020) demonstrated that the data-level curriculum learning approach reduced the number of training episodes needed for a reinforcement learning agent to ground 40 visual objects to its attribute-noun combination. However, a relatively small task as such still required about 600,000 training episodes for convergence. Furthermore, compositional generalization to held-out instruction was not evaluated. Comparatively, the number of instruction combinations involving either determiners and prepositions in our grounding task are 200 and 4800 respectively. We seek to establish if curriculum learning can expedite the concept grounding process and ensure compositional generalization with a task complexity that are orders of magnitude higher.

C.3 Models for Compositional Learning

Compositional learning involves breaking down information into fundamental elements or concepts and then integrating these elements to address new and unfamiliar combinations in a few-shot or zero-shot setting. (Lake et al., 2015; Xu et al., 2021; De Beule & Bergen, 2006).

Compositional learning has primarily been investigated in the realm of object detection, where visual models are trained on pairs of object and attribute information. These models leverage learned invariances to effectively handle unseen test sets (Kato et al., 2018; Purushwalkam et al., 2019; Anwaar et al., 2021). Another approach involves using the loss function to encourage networks to break down information into generalized features (Stone et al., 2017; Tolooshams et al., 2020). Recent advancements include models that can recognize or parse objects in images using bounding boxes (Lee et al., 2023) or segmentation masks (Kirillov et al., 2023), enabling them to tackle novel tasks.

Recent multi-modal models learn to align visual inputs to language inputs (Radford et al., 2021; Ma et al., 2023; Yuksekgonul et al., 2022) to solve the task of compositional reasoning (Lu et al., 2023) such as Visual-Question Answering (VQA) (Johnson et al., 2017), Referring Expressions (Lee et al., 2023), or augment images using instructions (Gupta & Kembhavi, 2023). There are few models that integrate compositional learning across vision, language, and action domains. An example is to train reinforcement learning agents that are grounded to visual inputs and language queries. Although these agents require millions of training episodes in diverse simulated environments, they demonstrate impressive proficiency in solving instruction-based tasks (Team et al., 2021).

Nevertheless, how these models ground vision-language-action representations for compositional learning, what the individual concepts are, and how these concepts are recomposed to solve novel combinations remains elusive. Only recently, has the compositional generalization abilities of reinforcement learning agents trained to ground visual objects to nouns and attributes have been explored (Lin et al., 2023). In this work, we aim to demonstrate the compositionality of agents in the 3D navigation task given the complex language instructions containing determiners and prepositions.