

Knowing Earlier what Right Means to You: A Comprehensive VQA Dataset for Grounding Relative Directions via Multi-Task Learning

Kyra Ahrens*, Matthias Kerzel*, Jae Hee Lee*, Cornelius Weber and Stefan Wermter

University of Hamburg

{kyra.ahrens, matthias.kerzel, jae.hee.lee, cornelius.weber, stefan.wermter}@uni-hamburg.de

Abstract

Spatial reasoning poses a particular challenge for intelligent agents and is at the same time a prerequisite for their successful interaction and communication in the physical world. One such reasoning task is to describe the position of a target object with respect to the intrinsic orientation of some reference object via *relative directions*. In this paper, we introduce GRiD-A-3D, a novel diagnostic visual question-answering (VQA) dataset based on abstract objects. Our dataset allows for a fine-grained analysis of end-to-end VQA models’ capabilities to ground relative directions. At the same time, model training requires considerably fewer computational resources compared with existing datasets, yet yields a comparable or even higher performance. Along with the new dataset, we provide a thorough evaluation based on two widely known end-to-end VQA architectures trained on GRiD-A-3D. We demonstrate that within a few epochs, the subtasks required to reason over relative directions, such as recognizing and locating objects in a scene and estimating their intrinsic orientations, are learned in the order in which relative directions are intuitively processed.

1 Introduction

Reasoning to solve complex spatial tasks like grounding directional relations in an intrinsic frame of reference can be decomposed into a set of subtasks that are hierarchically organized. Consider two objects o_1 and o_2 in an image, where each of the objects has a clear front side and orientation. Learning to answer whether the triple (o_1, r, o_2) holds for a given directional relation r in a frame of reference that is intrinsic to o_2 spans the following stages (see Fig. 1 for an example):

1. Both the target object and the reference object have to be recognized in the image (**existence prediction**). In other words, an agent must initially be capable of answering questions such as “Is o_1 in the image?” or “Is o_2 in the image?”.

*Equal contribution.

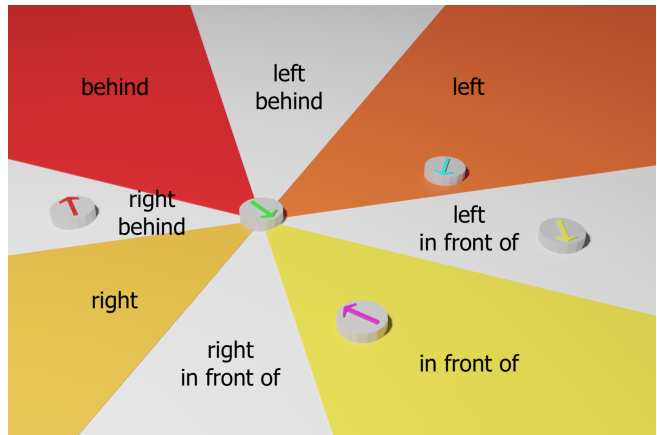


Figure 1: Example of grounding relative directions, e.g., considering the green arrow’s perspective, the yellow arrow is on the left in front of it.

2. Next, the object’s pose that defines the relative relation has to be discerned, enabling an agent to successfully respond to questions such as “What is the cardinal direction of o_2 ?” (**orientation prediction**).
3. Predicting the directional relation using the intrinsic frame of reference is learned by combining the two preceding competencies, allowing an agent to answer a question similar to “What is the relation between o_1 and o_2 from the perspective of o_2 ?” (**relation prediction**). Likewise, predicting which target object is in a specific relation to some reference object (**link prediction**) can be answered, e.g., “Taking o_2 ’s perspective, which object is in relation r to it?”.
4. Based on all previous stages, an agent can determine whether a specific directional relationship exists between the two objects (**triple classification**), thus successfully providing an answer to a question like “From o_2 ’s perspective, is o_1 left of o_2 ?”.

In previous work [Lee *et al.*, 2022], we showed that enabling a VQA architecture to reason about relative directions is viable, provided that all of the learning stages listed above are encapsulated in corresponding subtasks as summarized in Fig. 2. Beyond that, the following two observations were made: First,

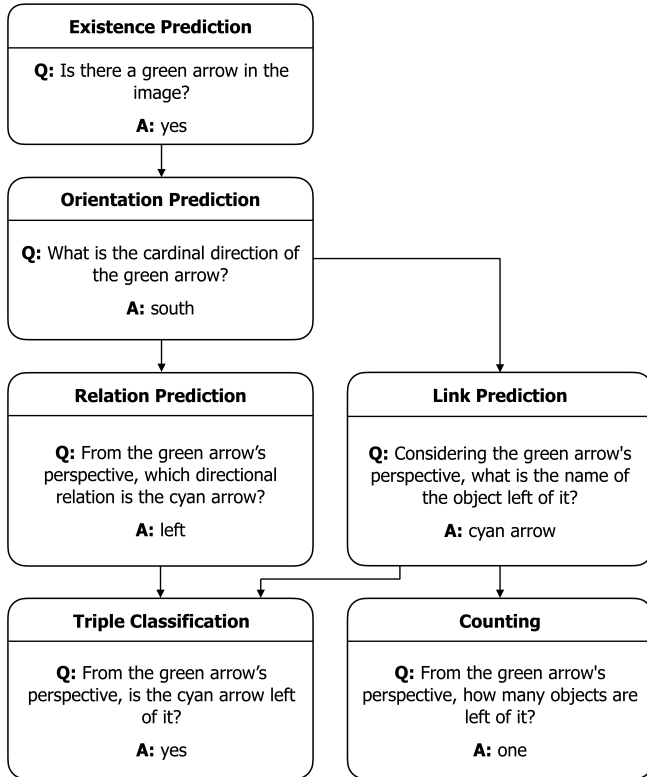
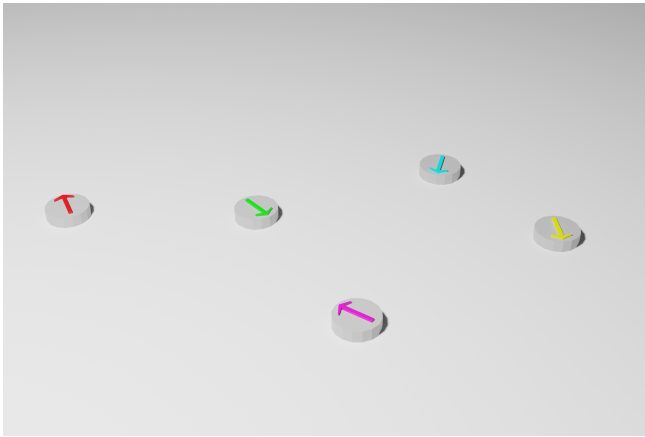


Figure 2: **Top**: Image from the GRiD-A-3D dataset. **Bottom**: Assumed hierarchy of spatial reasoning tasks to answer different question of the abstract GRiD-A-3D dataset. Arrows indicate a chronological dependency of tasks, e.g., in order to determine the orientation of an object, it first has to be recognized.

the subtasks that are found earlier in the chronology of learning stages are also learned earlier by the models, and second, this behavior is consistent for different neural end-to-end models. However, these findings are based on experiments involving images with 3D models of real objects, that may introduce a potential bias that confounds the analysis of reasoning about relative directions.

In the present work, we introduce GRiD-A-3D, a novel and simplified diagnostic VQA dataset, which allows for a more

efficient and targeted analysis of the corresponding reasoning process by removing possible biases from using real-world objects. Subsequently, we report the performance of the two established end-to-end VQA models MAC [Hudson and Manning, 2018] and FiLM [Perez *et al.*, 2018] on this dataset. With our experiments, we show that, when trained on GRiD-A-3D, both models depict a similar qualitative learning behavior compared with their replica trained on the more complex non-abstract GRiD-3D [Lee *et al.*, 2022] dataset. At the same time, training converges up to three times faster, thus allowing more efficient neural experiments.

We summarize the contributions made in this paper as follows:

- We complement our GRiD-3D benchmark suite¹ with a novel GRiD-A-3D (**G**rounding **R**elative **D**irections with **A**bstract objects in **3D**) dataset that enables a faster and less biased evaluation of spatial reasoning behavior in VQA compared with the original GRiD-3D dataset.
- We verify our previous research findings with the new dataset, thus underpinning our hypothesis that multi-task learning enables neural models to learn to ground relative directions in VQA.
- Furthermore, we add evidence to our hypothesis that during multi-task learning, spatial reasoning abilities of a neural model develop along the intuitive order of corresponding subtasks, thus forming an implicit curriculum.

2 Related Work

Aiming to provide a suitable setup to assess the reasoning capabilities of neural models on vision-language tasks, diagnostic datasets have been introduced [Johnson *et al.*, 2017; Hudson and Manning, 2019]. One of the major advantages of such datasets is that they provide structured and tightly controlled scenes to prevent models from circumventing reasoning by exploiting conditional biases that commonly arise with real-world images. A particular advantage of diagnostic datasets based on synthetic images is that their generation process is scalable, customizable, and therefore allows for a more fine-grained performance analysis.

The vast majority of diagnostic VQA datasets is limited to spatial reasoning tasks based on the absolute frame of reference, i.e., object positions are relative to the viewer of the image. Yet taking into account more realistic scenarios such as multi-agent dialogue in a situated environment, understanding relative directions is a prerequisite for meaningful communication. As a consequence, early models to learn symbolic reasoning with relative directions have been proposed [Moratz and Tenbrink, 2006; Lee *et al.*, 2013; Hua *et al.*, 2018]. However, they inherently assume the availability of scene annotations in terms of object labels and spatial relations instead of requiring a model to infer such information implicitly.

An early synthetic dataset providing a test bed for grounding relative directions is Rel3D [Goyal *et al.*, 2020]. Since Rel3D is restricted to two objects per scene and one single task, i.e.,

¹<https://github.com/knowledgegetechnologyuhh/grid-3d>

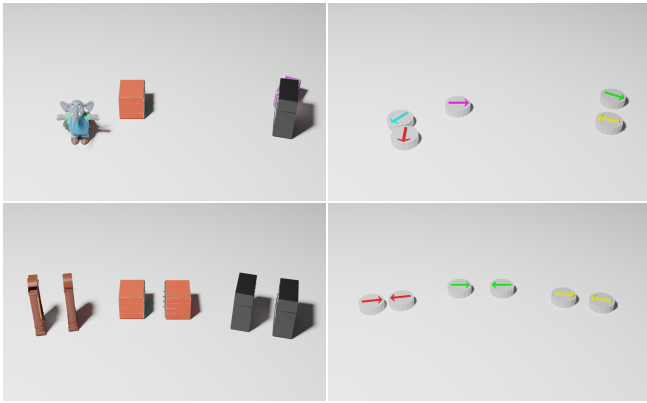


Figure 3: Common challenges in grounding relative directions arising with real objects, exemplified by objects from the original GRiD-3D dataset. **Top left:** Occlusion due to variability in heights and shapes of objects. **Bottom left:** Symmetry of objects impairs the detection of their front sides. **Top/bottom right:** Replica of the images on the left using abstract objects from the GRiD-A-3D dataset.

binary prediction of $(\text{object}_1, \text{relation}, \text{object}_2)$ triples, GRiD-3D [Lee *et al.*, 2022] was introduced, which combines the advantage of a rich number of tasks and questions as found in traditional synthetic VQA datasets with the challenge of grounding relative directions.

GRiD-3D is the first-of-its-kind to target multi-task learning of relative directions in a controlled setting. With this dataset, it was shown that, before learning how to answer the question whether a triple $(\text{object}_1, \text{relation}, \text{object}_2)$ holds, neural end-to-end VQA models rely on an implicit curriculum of related subtasks such as object detection, orientation estimation, and relation prediction [Lee *et al.*, 2022]. Objects in GRiD-3D cover a variety of categories, ranging from humanoids and animals to furniture and vehicles. Naturally, such objects differ in terms of proportions, complexity, and, most importantly, symmetry, which can be a crucial determinant of how easily a neural network can infer their orientation (and perform associated tasks).

In this work, we aim to provide a variation of the original dataset that ensures the elimination of such potential distortions (see Fig. 3 for examples), enabling a model to more quickly learn how to ground relative directions, which may be of particular value for few-shot, transfer, and curriculum learning scenarios. Accordingly, we extend the GRiD-3D benchmark suite towards another diagnostic VQA dataset with abstract objects.

3 GRiD-A-3D Abstract VQA Dataset

With the introduction of the GRiD-3D dataset [Lee *et al.*, 2022], we could show that neural VQA models are capable of grounding relative directions by implicitly deriving a curriculum of subtasks. In order to further generalize the previous findings, we extend our GRiD-3D suite towards a diagnostic dataset based on abstract objects whose cardinal direction is indicated by colored arrows.

Overview and statistics With our new GRiD-A-3D dataset, we address the following six tasks: Existence Prediction, Ori-

entation Prediction, Link Prediction, Relation Prediction, Counting, and Triple Classification. All 8 000 rendered images are split without overlap into 6 400 for training, 800 for validation, and 800 for testing. The 432 948 corresponding input questions follow largely the same 80:10:10 ratio, yielding 346 984, 43 393, and 42 571 questions for each set, respectively. The GRiD-A-3D dataset has an order of magnitude comparable with the GRiD-3D dataset, both in terms of image and question counts.

Image generation For each image, we generate a scene by randomly placing three to five distinct objects onto a plane and render the corresponding image with 480x320 pixel resolution via Blender.² We choose a consistent lighting setup across all images, add shadows to each object, and restrict the image generation to a fixed camera angle, thus obtaining one image per scene.

Our object set comprises gray-coloured polygonal prisms approximating a cylinder shape, each marked with an arrow in one of the six different colours: three primary colours (red, blue, and green) and three additive secondary colours (yellow, cyan, and magenta). The tip of each arrow depicts the object’s front side, allowing for distinct relative directions between objects in the image. An overview of all six objects can be found in Fig. 4. Note that the overall object count in the original GRiD-3D dataset is 28, whereas GRiD-A-3D is restricted to six different objects.

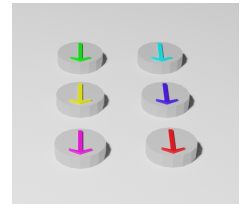


Figure 4: The six abstract objects used in the GRiD-A-3D dataset.

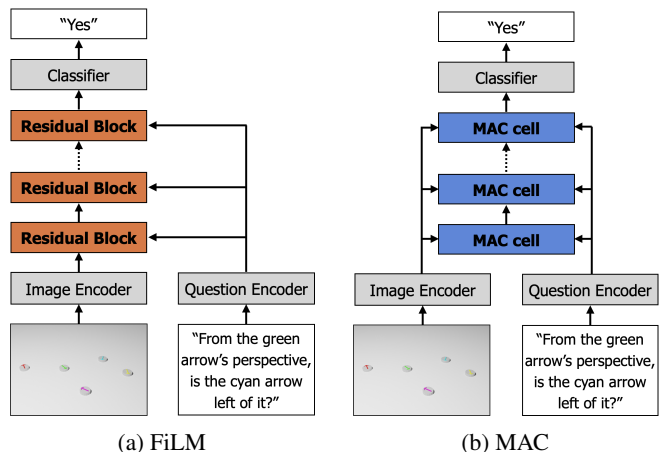


Figure 5: Neural end-to-end VQA models FiLM and MAC used for our experiments. The generic units (here colored in orange and blue, respectively) control how the question and image features are being processed.

Question generation In addition to rendering the images from our sampled scenes, we obtain scene graphs equipped with ground truth information such as absolute position, orientation, and relative directions of objects, that we use to

²<https://www.blender.org/>

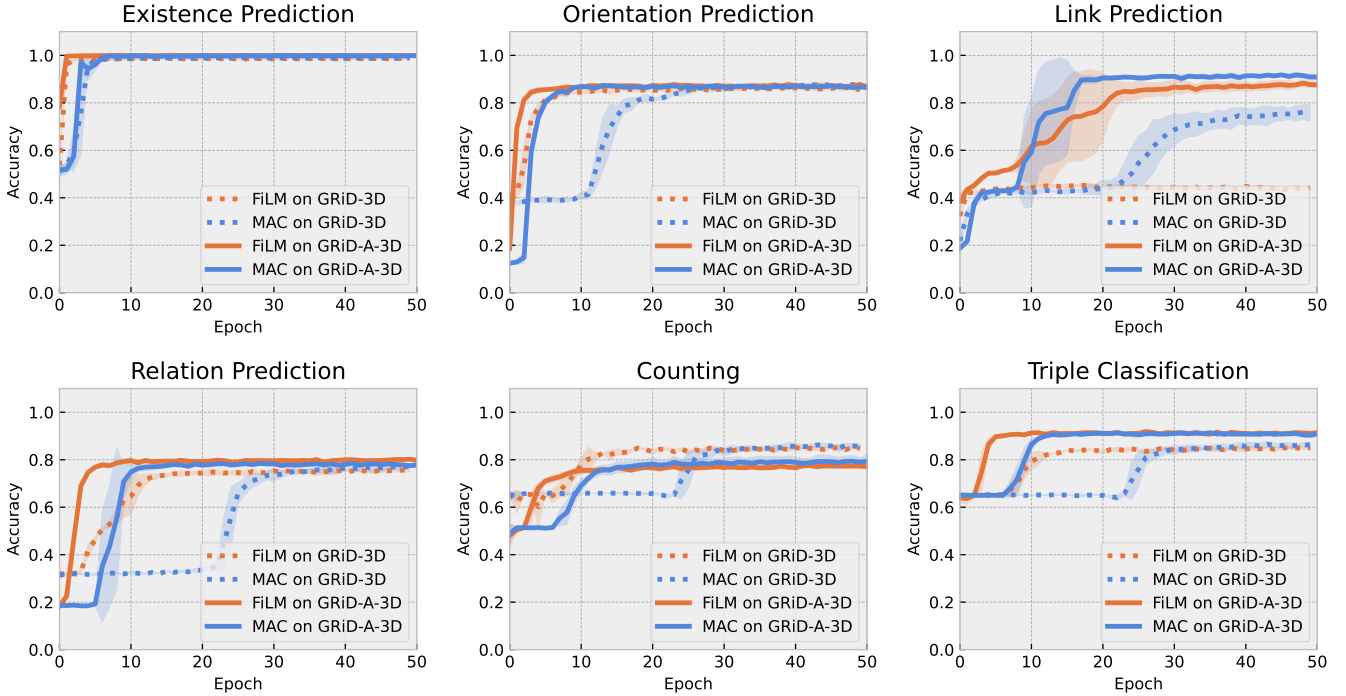


Figure 6: Multi-task learning results of FiLM (orange lines) and MAC (blue lines) on each of the six reasoning tasks of the GRiD-A-3D dataset (solid lines) vs. training the same models on the original GRiD-3D dataset (dotted lines).

generate questions related to the six tasks contained in GRiD-A-3D. Our question generation builds upon the framework provided with CLEVR [Johnson *et al.*, 2017], whose question templates, synonym, and metadata files we tailor to our dataset. Likewise, our question generation pipeline is expressed as a template-based functional program executed on each scene graph.

We follow the depth-first search strategy to determine and instantiate question-answer pairs that comply with the scene information and can therefore be considered valid. We set additional constraints to make sure that answers are uniformly distributed for each task. To ensure a wide variety of natural language questions, we sample from a rich set of differently phrased question templates for each reasoning task and randomly omit utterances or replace words with suitable synonyms.

4 Evaluations

For our experiments, we train MAC [Hudson and Manning, 2018] and FiLM [Perez *et al.*, 2018], two state-of-the-art neural end-to-end VQA architectures, on our new GRiD-A-3D dataset (cf. Fig. 5). Both architectures take raw RGB images and plain text question-answer pairs as input for training. Image features are extracted by a pretrained ResNet101 [He *et al.*, 2016] for both models, while questions are encoded by a GRU [Chung *et al.*, 2014] (FiLM) or a bidirectional LSTM [Hochreiter and Schmidhuber, 1997] (MAC), respectively. Subsequently, image and question features are fed to special neural units called *residual blocks* (FiLM) or *MAC*

cells (MAC). A chain of such units provides the core of the reasoning process.

We use existing PyTorch³ implementations of FiLM and MAC with their default hyperparameters for the published CLEVR [Johnson *et al.*, 2017] dataset evaluations, except for the number of MAC cells that we reduce to four to prevent overfitting. All experiments are run for 100 epochs and repeated three times with different seeds to reduce the impact of the random initialization of the models on the results. Fig. 6 shows the *mean* and the *standard deviation* of the evaluations.

We interpret our results in the following way: Existence and Orientation Prediction are learned earlier than other tasks. We explain this observation with the fact that these tasks only require a model to focus on one single object. For the most straightforward task of Existence Prediction, we observe similar behavior for the two datasets: Both converge to an accuracy of almost 100% at nearly the same time. For the Orientation Prediction task, we observe convergence to an accuracy of over 80% for both datasets. Noticeably, the learning happens faster for the abstract GRiD-A-3D dataset. The shorter learning time can be attributed to the more unequivocal identification of front and back sides of the abstract objects due to the lack of symmetry related noise as shown in Fig. 3. The fact that the accuracy on Orientation Prediction is capped at about 85% can be explained by objects placed close to the border between two cardinal directions, as such cases are difficult for the models to learn and classify.

³<https://pytorch.org/>

A similar learning behavior can be observed for the more complex tasks of Relation Prediction, Triple Classification and Link Prediction, where both models converge faster when trained on GRiD-A-3D and also reach slightly higher accuracy. Similarly to the results on the Orientation Prediction task, the main reason for these observations may lie in the facilitated learning conditions due to the lack of front-back symmetries or strong occlusions with the abstract objects. This effect is most pronounced for Link Prediction, i.e., predicting which target object is in a given relation to some reference object. We attribute this observation to the smaller set of objects in the GRiD-A-3D dataset.

Finally, we observe a mixed result for the Counting task: While learning of both VQA models converges faster for the GRiD-A-3D dataset, higher accuracy is reached for the GRiD-3D dataset. We hypothesize that this higher accuracy stems from the more diverse-looking objects in the GRiD-3D dataset, facilitating the models to distinguish and thus count multiple objects in close proximity.

In summary, our results suggest the following two facts: First, the abstract GRiD-A-3D dataset leads to faster learning and can thus enable more computationally efficient experimentation while achieving comparable results to the original GRiD-3D dataset. Second, the results support our assumption of a chronology of subtasks, as Existence Prediction and Orientation Prediction are learned before the models can reason about relative directions.

5 Conclusions

This work is an extension to previous work on grounding relative directions with end-to-end neural VQA architectures. We provide a comprehensive, simplified GRiD-A-3D dataset with abstract objects that shows similar behavior to the original GRiD-3D dataset when learned by the two established VQA models FiLM and MAC. With our experiments, we show that the learning of tasks that focus on a single object like object recognition and orientation prediction happens prior to learning to ground relative directions and object counting.

The abstract nature of the dataset eliminates approximate front-back object symmetries that can have a negative impact on object orientation prediction and all reasoning tasks about directional relations that build upon it. Furthermore, the simplification of the object set allows for conducting experiments with a more comprehensive dataset. In future work, this will allow us to conduct fast pilot studies on curriculum and transfer learning based on the intuitive dependency of the different spatial reasoning tasks on one another and the observed implicit curriculum.

Acknowledgments

The authors gratefully acknowledge support from the German Research Foundation DFG for the projects CML TRR169, LeCAREbot and IDEAS.

References

- [Chung *et al.*, 2014] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning*, 2014.
- [Goyal *et al.*, 2020] Ankit Goyal, Kaiyu Yang, Dawei Yang, and Jia Deng. Rel3D: A Minimally Contrastive Benchmark for Grounding Spatial Relations in 3D. *Advances in Neural Information Processing Systems*, 33, 2020.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [Hua *et al.*, 2018] Hua Hua, Jochen Renz, and Xiaoyu Ge. Qualitative Representation and Reasoning over Direction Relations across Different Frames of Reference. In *Sixteenth International Conference on Principles of Knowledge Representation and Reasoning*, September 2018.
- [Hudson and Manning, 2018] Drew A. Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*, February 2018.
- [Hudson and Manning, 2019] Drew A. Hudson and Christopher D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [Johnson *et al.*, 2017] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [Lee *et al.*, 2013] Jae Hee Lee, Jochen Renz, and Diedrich Wolter. StarVars: Effective Reasoning About Relative Directions. In *Twenty-Third International Joint Conference on Artificial Intelligence*. AAAI Press, 2013.
- [Lee *et al.*, 2022] Jae Hee Lee, Matthias Kerzel, Kyra Ahrens, Cornelius Weber, and Stefan Wermter. What is Right for Me is Not Yet Right for You: A Dataset for Grounding Relative Directions via Multi-Task Learning. In *Thirty-First International Joint Conference on Artificial Intelligence*, 2022.
- [Moratz and Tenbrink, 2006] Reinhard Moratz and Thora Tenbrink. Spatial Reference in Linguistic Human-Robot Interaction: Iterative, Empirically Supported Development of a Model of Projective Relations. *Spatial Cognition & Computation*, 6(1), March 2006.
- [Perez *et al.*, 2018] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, April 2018.