What is Right for Me is Not Yet Right for You: A Dataset for Grounding Relative Directions via Multi-Task Learning

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

University of Hamburg

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

Abstract

Understanding spatial relations is essential for intelligent agents to act and communicate in the physical world. Relative directions are spatial relations that describe the relative positions of target objects with regard to the intrinsic orientation of reference objects. Grounding relative directions is more difficult than grounding absolute directions because it not only requires a model to detect objects in the image and to identify spatial relation based on this information, but it also needs to recognize the orientation of objects and integrate this information into the reasoning process. We investigate the challenging problem of grounding relative directions with end-to-end neural networks. To this end, we provide GRiD-3D, a novel dataset that features relative directions and complements existing visual question answering (VQA) datasets, such as CLEVR, that involve only absolute directions. We also provide baselines for the dataset with two established end-to-end VQA models. Experimental evaluations show that answering questions on relative directions is feasible when questions in the dataset simulate the necessary subtasks for grounding relative directions. We discover that those subtasks are learned in an order that reflects the steps of an intuitive pipeline for processing relative directions.

1 Introduction

Talking about the locations of objects is a basic communicative skill for humans to manage their everyday activities [Bloom *et al.*, 1999; Coventry and Garrod, 2004]. Among several ways of talking about locations, humans often use the relative direction of the target object from the perspective of a reference object. For example, to describe the target object *laptop* in Fig. 1a one can take the perspective of the reference object *elephant* and say that "the *laptop* is *to the right* of the

(a) An example scene where an object *laptop* is *to the right* of another object *elephant*. Here, the directional relation *right* is relative and is determined by the perspective of the *elephant*.

(b) Object detection

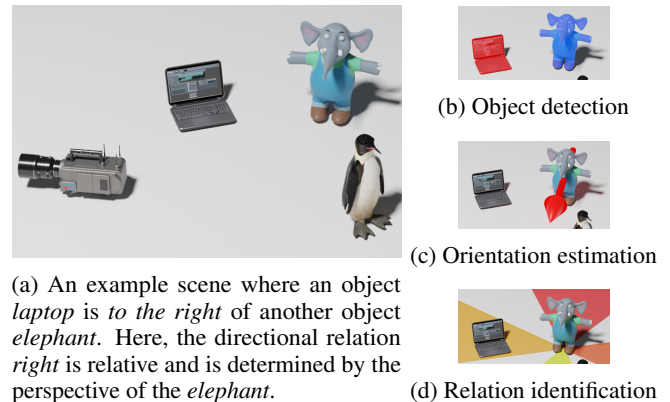(c) Orientation estimation

(d) Relation identification

Figure 1: We propose a VQA dataset for grounding relative directions; in addition to questions on relative directions, it encompasses questions regarding object detection and orientation estimation.

elephant". As relative directions are common in everyday use of language, they are important in human-robot interaction and artificial intelligence [Moratz and Tenbrink, 2006; Lee *et al.*, 2013; Hua *et al.*, 2018].

This paper is concerned with grounding relative directions with end-to-end neural networks. This task is not trivial since, following human intuition, it is composed of a sequence of three subtasks: Given a collection of objects in a scene, it first has to detect the objects by binding the visual and the linguistic representation of the target and the reference object in the scene (cf. Fig. 1b). Second, it has to estimate the orientation of the reference object (cf. Fig. 1c). Finally, it has to determine the direction of the target object based on the orientation of the reference object (cf. Fig. 1d).

Conventional approaches to process the previous three subtasks would hand-design a pipeline of modules that carry out each subtask. Such pipeline-based modular approaches often perform well, in particular when the target task is well-known such that the modules can be tailored to the subtasks of the target task. They also often facilitate interpretation and diagnosis of the results. However, they cannot be flexibly applied to a new task without either training every module on the new task or adding new modules and re-configuring the pipeline. End-to-end differentiable neural models, on the other hand, do not suffer from the latter deficiency of the modular approaches.

Dataset	#Samples	#Images	Automatic generation	Non-abstract objects	Multiple tasks	Varying object counts	Relative directions
CLEVR [Johnson *et al.*, 2017a]	1M	100k	✓	✗	✓	✓	✗
PTR [Hong *et al.*, 2021]	700k	70k	✓	✓	✓	✓	✗
SpatialSense [Yang *et al.*, 2019]	17.5k	11.6k	✗	✓	✗	✓	✗
Rel3D [Goyal *et al.*, 2020]	27k	27k	✗	✓	✗	✗	✓
GRiD-3D (*ours*)	445k	8k	✓	✓	✓	✓	✓

Table 1: Comparison of GRiD-3D with existing spatial reasoning benchmarks.

Therefore, investigating how well such end-to-end neural models ground relative directions is an important and interesting open research question.

We view the relative direction grounding task from the perspective of *visual question answering* (VQA) [Wu *et al.*, 2017] and propose a dataset-based approach that allows existing end-to-end differentiable neural VQA models to implicitly learn to solve the three subtasks. As we can *control* with question inputs how the neural VQA models process images, we allow the models to solve the three subtasks merely by including additional VQA questions in the dataset pertaining to all three subtasks, *without* any modifications of the model architectures. The contributions of this paper are the following:

1. We introduce GRiD-3D[1], a novel diagnostic VQA dataset for end-to-end relative direction learning. This dataset is more realistic than existing synthetic datasets for visual reasoning. It includes objects from different real-world categories with intrinsic orientations, which allows us to evaluate the capabilities of existing neural models for grounding relative directions.

2. We conduct extensive experiments on our dataset and show that the state-of-the-art generic neural VQA models FiLM [Perez *et al.*, 2018] and MAC [Hudson and Manning, 2018] are capable of grounding relative directions. We demonstrate that the models learn the three subtasks that are intuitively necessary for grounding relative directions, i.e., (i) object detection, (ii) pose estimation, and (iii) relation identification. We also identify that the three subtasks are learned in this specific order, which suggests the presence of an emergent curriculum, i.e., the three subtasks are ordered in increasing difficulty, and easier tasks facilitate learning more difficult tasks.

2 Related Work

General relation learning. Grounding relative directions involves situated perception, such as vision, in addition to language. A prominent task in the area of vision and language integration is *visual question answering* (VQA), which is concerned with answering questions about an image [Wu *et al.*, 2017]. Several datasets have been proposed for VQA (e.g., [Goyal *et al.*, 2017; Johnson *et al.*, 2017a; Suhr *et al.*, 2019; Hudson and Manning, 2019]). The existing VQA datasets, however, typically use spatial relations in an absolute frame of reference (i.e., *left* and *right* correspond to the *left* and *right*

---
[1] https://github.com/knowledgetechnologyuhh/grid-3d

side of the input image) and do not offer a way to evaluate a model's capability in dealing with relative directions.

Relative directions. In the area of knowledge representation and reasoning, much attention has been paid to inference with relative direction information [Moratz, 2006; Lee *et al.*, 2013; Hua *et al.*, 2018; Freksa *et al.*, 2018]. Existing work in this area typically assumes that the primitives, i.e., the labels of objects and the relations between objects, are given, and does not address the implicit learning of such information based on sensory stimuli from, e.g., images.

VQA Datasets. The development and evaluation of powerful VQA models relies on the availability of well-curated datasets. Designing such datasets of real-world images is labor-intensive; for a large-scale dataset, this can only be done by using pretrained object detectors, by crowdsourcing the annotation effort [Krishna *et al.*, 2017] or by leveraging existing annotations, e.g., in the form of narrated videos [Yang *et al.*, 2021]. An alternative are synthetic datasets generated with 3D rendering environments. While artificially generated images do not provide the full complexity and noisiness of real-world images, they do offer several advantages: (i) images can be generated in an unbiased way, e.g., with regard to the distribution of relevant objects, relations and similar properties; (ii) images can be automatically generated with all required annotations; (iii) such datasets can easily be upscaled based on the demands of the task and the neural architecture. They are therefore well-suited as diagnostic datasets.

Spatial relation learning with VQA datasets. Recently, several datasets have been proposed that overcome the limitations of existing VQA datasets in dealing with spatial relations. The PTR dataset [Hong *et al.*, 2021] focuses on the part-whole relationships between entities in synthetic scenes. However, the dataset does not deal with relative directions, i.e., the directions from the perspective of the reference objects.

SpatialSense [Yang *et al.*, 2019] is a crowdsourced dataset, where human annotators provided spatial relation labels that are difficult to predict using simple cues (e.g., 2D spatial configurations or language priors). Annotators made extensive use of relative directions, which turned out to be major failing cases for the baseline models.

Rel3D [Goyal *et al.*, 2020] is also a crowdsourced dataset, but different from SpatialSense, the images in Rel3D are synthetically generated with 3D ground-truth information. Rel3D provides examples with reduced annotation bias. The dataset is limited in that the images in the dataset contain only two objects, and it provides only one type of question, i.e., predicting the correctness of (object$_1$, relation, object$_2$) triples. By

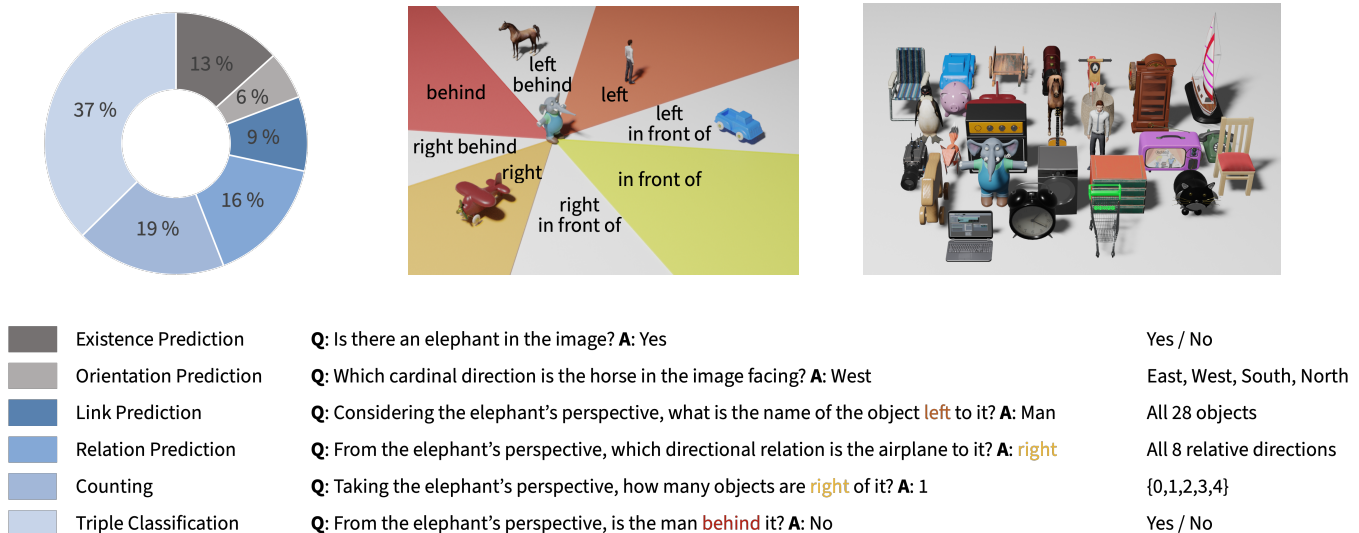Figure 2: **Top left**: Distribution of per-task questions. Tasks shown in `blue` require *understanding relative directions*. **Top center**: Overview of the 8 different relative directions from the viewpoint of the reference object *elephant*. **Top right**: All 28 3D-objects used for the synthetic dataset. Each object features a clear front side. **Bottom**: From left to right: the six tasks; sample questions and answers for each task; the answer set.

contrast, our dataset provides a varying number of objects in each scene and different categories of questions.

Table 1 shows an overview of relevant VQA datasets on relation learning. While all datasets in the table include spatial relations in their questions, GRiD-3D is the only dataset that features questions on relative directions as part of a more extensive, multi-task question inventory. With GRiD-3D, we address this gap in the state of the art of existing VQA datasets.

3 GRiD-3D VQA Dataset

We introduce GRiD-3D (**G**rounding **R**ela**t**ive **D**irections in **3D**), a novel diagnostic VQA dataset that enables learning relative directional relations between objects. Our dataset comprises 8 000 synthetic images and 445 080 questions addressing six different reasoning tasks: Existence Prediction, Orientation Prediction, Link Prediction, Relation Prediction, Counting, and Triple Classification. Exemplary input questions and answers per reasoning task are shown in Fig. 2.

The images are split in an 80:10:10 ratio without overlapping images between training, validation and test sets (i.e., 6 400, 800, and 800, respectively). The 445 080 questions are split into similar proportions (i.e., 357 839, 45 030, and 42 211, respectively).

All images have a 480x320 pixel resolution and are rendered using Blender[2] by randomly placing 2–5 objects onto a plane in a non-overlapping manner and following a uniform distribution. We choose a consistent lighting setup that provides shadows for the sake of more realistic scenes. In addition, we restrict the image generation to a fixed camera angle that allows for partial but not complete occlusion of objects. Consequently, we have one image per scene, and we will therefore use these two terms interchangeably in this work.

Scene objects are randomly selected from a set of 28 different non-abstract 3D models that all have a distinctive front side, thus allowing to describe a spatial layout with respect to the object's intrinsic frame of reference, i.e., via relative directions. All 3D models of the objects were obtained from BlenderKit[3], released under "Royalty-Free" or "CC0" licenses. They are depicted in Fig. 2.

The dataset obtains ground-truth annotations regarding absolute position, orientation, and relative directions of objects for each generated scene. Following the strategy of Johnson *et al.* [2017a], questions are expressed as a functional program on each scene's ground-truth information. Each task's questions, instantiated and validated based on depth-first search, follow a uniform answer distribution. For example, the body of questions addressing the object Existence Prediction task has a 50/50 ratio of *yes* or *no* as the target answer. For each reasoning task, we vary question phrasing by using different question skeletons as well as by omitting or replacing terms by corresponding synonyms with some predefined probability.

The ratio between per-task question counts, as illustrated in Fig. 2, is drawn from the ratio of the total number of possible questions per task. For example, the number of uniquely instantiable questions on Orientation Prediction per scene equals the total number of its objects, whereas, for the same scene, we can generate twice as many questions on Existence Prediction (one negative and one positive sample per object).

4 Language-Controlled Multi-Task Learning

Solving VQA requires different capabilities of a model, as the questions in a VQA dataset involve multiple tasks, e.g., object detection, matching, comparison, counting, relation identification. For example, to answer the following question

---

[2]https://www.blender.org/

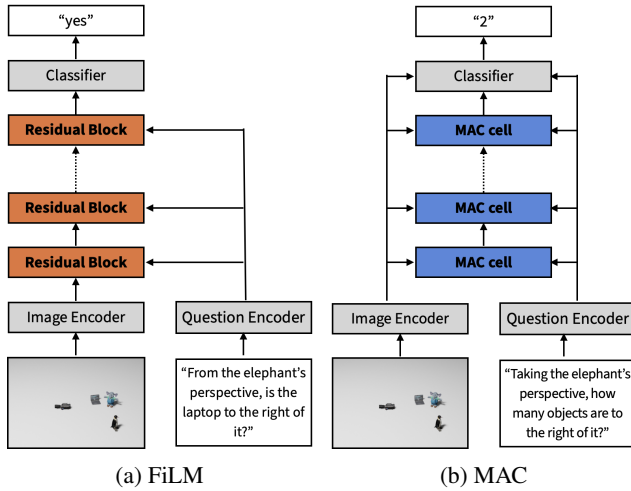[3]https://www.blenderkit.com/

(a) FiLM   (b) MAC

Figure 3: An abstract overview of the FiLM and MAC architectures. In both cases, the encoded question is passed to each generic unit (colored in orange and blue, respectively) and affects how the encoded image is processed by the models.

in the CLEVR dataset [Johnson *et al.*, 2017a]: "There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?" a model needs to be able to perform object detection, matching, and comparison.

Modular VQA models, such as the ones proposed by Johnson *et al.* [2017b], Hu *et al.* [2017], Yi *et al.* [2018], achieve multi-task learning by providing a module for each subtask comprising a target task and automatically building pipelines based on those modules. The modules are hand-crafted (sometimes with trainable parameters) and require detailed knowledge about the target task. It is, therefore, difficult, if not impossible, to apply them to new tasks without designing new such modules.

By contrast, end-to-end differentiable VQA models, such as FiLM [Perez *et al.*, 2018] and MAC [Hudson and Manning, 2018], do not need specific modules designed for specific subtasks. Instead, those models feature generic units called *residual blocks* (FiLM) or *MAC cells* (MAC), which are controlled by the language input features.

As can be seen in the high-level overview of the model architectures in Fig. 3, each generic unit receives as input the output from the previous unit together with the encoded question (and, in the case of MAC, the encoded image). Therefore, the encoded question basically controls each generic unit and modifies their inner workings such that the whole architecture can appropriately process the image input and accomplish the target task.[4]

---

[4]There are two main differences between FiLM and MAC: (i) in FiLM the encoded image is passed to the first residual block only, whereas in MAC the encoded image is passed to every MAC cell; (ii) the residual blocks do not share weights, while the MAC cells share weights; thus MAC can be regarded as a recurrent model with a fixed depth of recurrence.
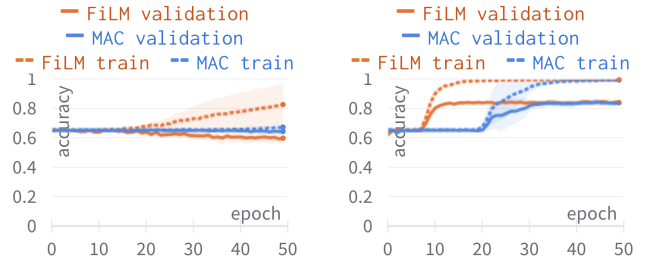
(a) Trained on Triple Classification only.   (b) Trained on all tasks in GRiD-3D.

Figure 4: Mean validation and training accuracies of FiLM and MAC on Triple Classification.

5 Evaluations

In this section, we evaluate the two state-of-the-art end-to-end differentiable VQA models, FiLM and MAC. As both models have their generic units directly controlled by the input question, each unit's function can be adjusted with respect to the target task and can process the input image accordingly.

Following our question of whether a model can ground relative directions, we pay particular attention to model performance on those tasks targeting such capability, i.e. Link Prediction, Relation Prediction, Counting, and Triple Classification. We choose the accuracy on Triple Classification, which comprises the largest part of the dataset (cf. Fig. 2), as a representative evaluation criterion for grounding relative directions and answer the following questions:

1. Is multi-task learning necessary for Triple Classification?
2. Is multi-task learning with the tasks in GRiD-3D sufficient for Triple Classification?
3. Is there an order in learning the tasks?
4. What tasks in GRiD-3D are necessary for Triple Classification?

Our evaluation framework is implemented using the PyTorch[5] library and existing implementations of FiLM and MAC with their default parameter configurations used for the evaluations on the CLEVR dataset, except for the number of MAC cells that is set to four instead of the default value 12, as we observed an improved model performance. We run each experiment three times for 50 epochs. Each graph of an experiment shows the *mean* and the *standard deviation* of the three runs.

5.1 Multi-Task Learning is Necessary

The first intuitive question we can ask is whether multi-task learning is necessary at all to solve the Triple Classification task from GRiD-3D. To answer this question, we train and evaluate FiLM and MAC on Triple Classification only. The training and validation learning curves of both models of this experiment are shown in Fig. 4a. While the initial mean validation accuracies of the two models are around 65% after the first epoch, they drop to 60% and 64% after 50 epochs of training. This is a rather poor performance considering that a random baseline achieves 50% accuracy.
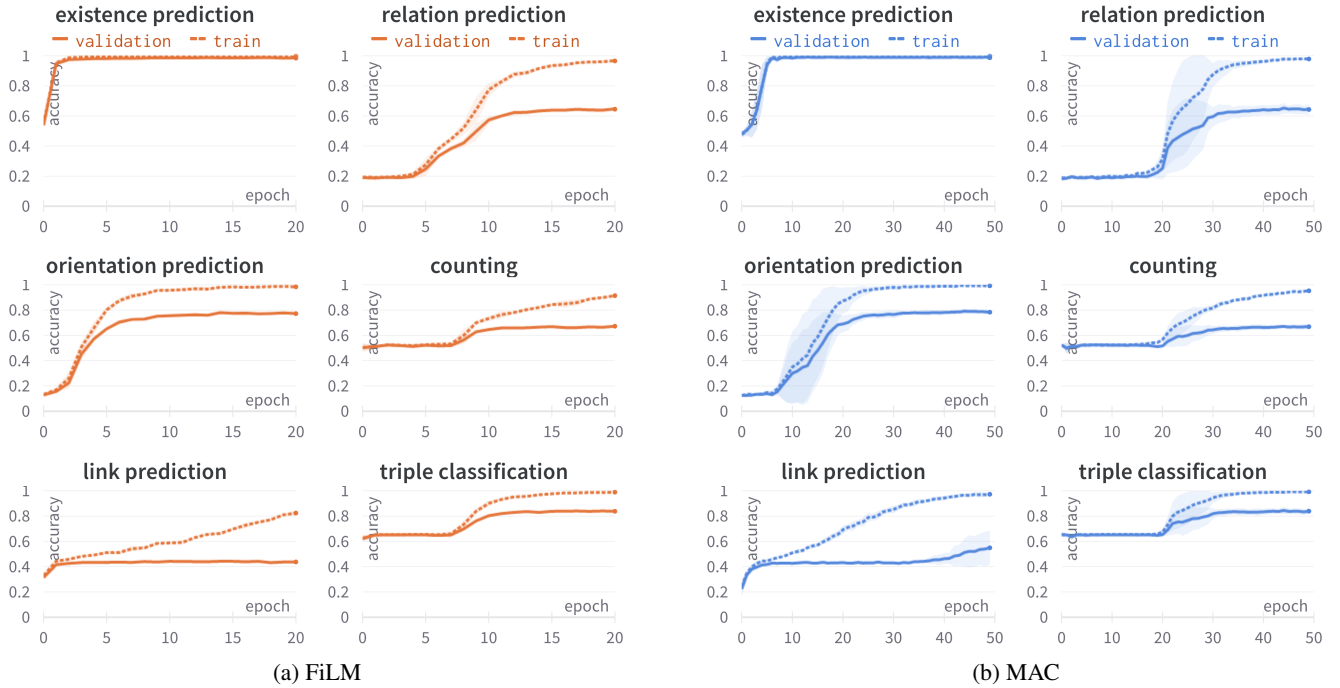
---

[5]https://pytorch.org/

Figure 5: Multi-task learning results on all six tasks of the GRiD-3D dataset. For FiLM, only the first 20 of the 50 total training epochs are shown, to allow for comparison of the two models in terms of convergence behavior.

We can make two further observations: First, the mean validation accuracies are about 10% above the chance level, indicating some spurious correlations in the dataset that the models can exploit. Second, both models even failed to consistently learn the training data for Triple Classification, as the mean training accuracies after 50 epochs are 83% and 67%. This hints at the fact that Triple Classification is a complex task, and merely memorizing some correlations between the inputs and the outputs is not sufficient. Overall, the result shows that single-task learning is not sufficient and hence multi-task learning is necessary for Triple Classification with GRiD-3D.

5.2 GRiD-3D Tasks are Sufficient

Based on our previous finding that multi-task learning on GRiD-3D could facilitate Triple Classification, a natural follow-up question would be whether the tasks included in the GRiD-3D dataset are sufficient to obtain a meaningful estimate of the Triple Classification capabilities of a neural VQA model.

To answer this question, we train FiLM and MAC on all six tasks and evaluate their performances on the Triple Classification task. The results are shown in Fig. 4b. In this experiment, both FiLM and MAC achieve 84% mean validation accuracy after 50 epochs, respectively, improving their mean validation accuracies on the single task experiment by more than 31%.

Given that there are difficulties in detecting objects (e.g., due to occlusions), estimating orientations (e.g., due to almost symmetric objects such as *closet*) and identifying relations (e.g., target objects being close to the decision boundary between two relations), further improving the performance would be challenging without injecting additional prior knowl-

edge to the models. Consequently, we can conclude that the tasks in GRiD-3D are sufficient to help FiLM and MAC solve the Triple Classification task.

5.3 The Order of Learning the Tasks

Multi-task learning with the tasks in GRiD-3D allows FiLM and MAC to solve Triple Classification subject to the intrinsic difficulties of the task mentioned previously. It is, however, not clear *why* and *how* adding different question-answer pairs (e.g., questions about orientations), while keeping the same set of images in the training set, leads to the sudden improvement both in the training and the validation performances. We can partially answer this question by observing the validation performances on all six tasks that are presented in Fig. 5. We observe that there is an order of the tasks that governs the dynamics of learning, which reflects the pipeline of the subtasks for grounding relative directions as suggested in the introduction of the paper (cf. Fig. 1):

1. Existence Prediction is learned, which corresponds to learning object detection (cf. Fig. 1b)
2. Orientation Prediction is learned, which corresponds to learning orientation estimation (cf. Fig. 1c),
3. The remaining four tasks that involve grounding relative directions are learned (with the exception of link prediction in the case of FiLM), which correspond to learning relation identification (cf. Fig. 1d).

This intuitive but surprising emergence of an implicit order of tasks suggests that curating a VQA dataset with well-chosen questions can already facilitate existing end-to-end differentiable VQA models such as FiLM and MAC to solve new tasks

that they are not initially designed for and are difficult to solve in isolation, all *without* including any additional images.

One noticeable result in Fig. 5 is the discrepancy between the mean validation and training performance on Link Prediction. After 50 epochs, FiLM and MAC achieve 44% and 55% mean validation accuracies despite their mean training accuracies being as high as 97% (not shown in the figure for FiLM). Furthermore, the mean validation accuracies of both models quickly reach 42%, which is very high considering the fact that there are 28 candidate target objects the models can choose from. Because their performances on Existence Prediction peak about the same time, it is likely that the models have learned to first *detect* the reference object and reduce the number of candidate objects from 27 to the remaining 2.5 ones in each scene (there are on average 3.5 objects in each scene), leading to 40% random chance to guess the target object correctly. These observations show that (i) Link Prediction is intrinsically more difficult than other tasks, (ii) MAC is more robust in learning Link Prediction, and (iii) not all tasks in GRiD-3D are necessary for solving Triple Classification.

5.4 Essential Tasks for Grounding Relative Directions

In the preceding subsection, we identified that training FiLM and MAC on all tasks from GRiD-3D is *not* necessary to solve Triple Classification. This raises the question about what tasks are essential. To answer this question, we train FiLM and MAC on all but one GRiD-3D tasks to determine the relevance of the one task to Triple Classification. The results are presented in Fig. 6.

In the figure, we observe that removing Counting, Link Prediction, or Existence Prediction from the six tasks for training does not lead to much performance drop, whereas removing Orientation Prediction or Relation Prediction from the tasks reduces the mean validation accuracy dramatically. On the other hand, training only on the latter two tasks in addition to Triple Classification already facilitates both models to perform well on Triple Classification, both achieving 84% mean accuracies. These findings suggest that Orientation Prediction and Relation Prediction are essential for Triple Classification while the other three tasks are less relevant.

6 Conclusions

Grounding relative directions is a relevant and challenging task for visual question answering (VQA). In contrast to absolute directions that require recognizing and localizing objects and determining their relations, relative directions additionally require reasoning about the objects' orientations. Existing VQA datasets either only address absolute directions or they do not include the subtasks that facilitate grounding relative direction in a multi-task learning setting. We address this gap by introducing the novel synthetic VQA dataset GRiD-3D, which focuses on relative directions and features related subtasks of object detection and orientation estimation.

We positively evaluated GRiD-3D with two established neural end-to-end VQA models, FiLM and MAC. We show that these models can learn to correctly answer questions on relative directions when trained with a set of questions pertaining
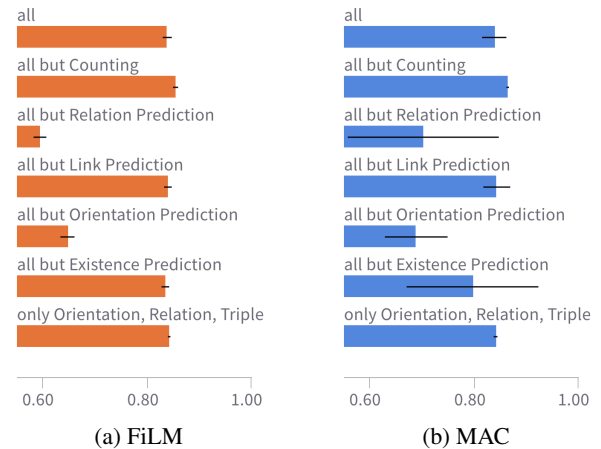

Figure 6: Mean validation accuracies of FiLM and MAC on Triple Classification when trained on selected tasks (see labels).

not only to relative directions but also to the subtasks, i.e., object detection and orientation estimation. Furthermore, an analysis of the learning process of both models shows how first the subtasks object detection and orientation estimation are learned before questions on relative directions are answered correctly. This incremental learning process based on starting with partial or simplified tasks can be compared to learning during child development and linked to research from cognitive science and deep reinforcement learning (cf. [Kerzel *et al.*, 2018]). These results support the hypothesis that multi-task learning with an implicit curriculum of subtasks can be beneficial for tackling more difficult VQA tasks. It would be interesting to see the applicability of this learning approach to other tasks that are conventionally tackled with approaches that feature dedicated and hand-designed modules.

GRiD-3D can serve as a diagnostic dataset for the further development of VQA models that tackle challenging VQA problems and for closing the performance gap for tasks on relative directions in comparison to less complex question classes. GRiD-3D will allow investigating further the mechanisms of how multiple tasks support each others' learning, or whether alternative mechanisms can learn relative directions in isolation, independent of other tasks. In future work, we will extend the dataset by adding more tasks to allow an even more fine-grained analysis of the curricular learning of models.

References

[Bloom *et al.*, 1999] Paul Bloom, Merrill F. Garrett, Lynn Nadel, and Mary A. Peterson, editors. *Language and Space*. The MIT Press, 1999.

[Coventry and Garrod, 2004] Kenny R. Coventry and Simon C. Garrod. *Saying, Seeing and Acting: The Psycholog-*

ical Semantics of Spatial Prepositions. Psychology Press, July 2004.

[Freksa *et al.*, 2018] Christian Freksa, Jasper van de Ven, and Diedrich Wolter. Formal representation of qualitative direction. *International Journal of Geographical Information Science*, 32(12), December 2018.

[Goyal *et al.*, 2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[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.

[Hong *et al.*, 2021] Yining Hong, Li Yi, Joshua B. Tenenbaum, Antonio Torralba, and Chuang Gan. PTR: A benchmark for part-based conceptual, relational, and physical reasoning. *arXiv:2112.05136 [cs]*, December 2021.

[Hu *et al.*, 2017] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*, Venice, October 2017. IEEE.

[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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702, 2019.

[Johnson *et al.*, 2017a] 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.

[Johnson *et al.*, 2017b] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *IEEE International Conference on Computer Vision (ICCV)*, October 2017.

[Kerzel *et al.*, 2018] Matthias Kerzel, Hadi Beik Mohammadi, Mohammad Ali Zamani, and Stefan Wermter. Accelerating deep continuous reinforcement learning through task simplification. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2018.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1), May 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*, Beijing, China, 2013. AAAI Press.

[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.

[Moratz, 2006] Reinhard Moratz. Representing relative direction as a binary relation of oriented points. In *17th European Conference on Artificial Intelligence*. IOS Press, May 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.

[Suhr *et al.*, 2019] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv:1811.00491 [cs]*, July 2019.

[Wu *et al.*, 2017] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163, October 2017.

[Yang *et al.*, 2019] Kaiyu Yang, Olga Russakovsky, and Jia Deng. SpatialSense: An adversarially crowdsourced benchmark for spatial relation recognition. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[Yang *et al.*, 2021] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021.

[Yi *et al.*, 2018] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018.