# A Measure of Visuospatial Reasoning Skills: Painting the Big Picture

**Joel Michelson** JOEL.P.MICHELSON@VANDERBILT.EDU
**Deepayan Sanyal** DEEPAYAN.SANYAL@VANDERBILT.EDU
**James Ainooson** JAMES.AINOOSON@VANDERBILT.EDU
**Maithilee Kunda** MKUNDA@VANDERBILT.EDU
Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN USA

## Abstract

Visuospatial reasoning refers to a diverse set of skills that involve thinking about space and time. An artificial agent with access to a sufficiently large set of visuospatial reasoning skills might be able to generalize its reasoning ability to an unprecedented expanse of tasks including portions of many popular intelligence tests. In this paper, we stress the importance of a developmental approach to the study of visuospatial reasoning, with an emphasis on fundamental skills. A comprehensive benchmark, with properties we outline in this paper including breadth, depth, explainability, and domain-specificity, would encourage and measure the genesis of such a skillset. Lacking an existing benchmark that satisfies these properties, we outline the design of a novel test in this paper. Such a benchmark would allow for expanding analysis of existing datasets' and agents' applicability to the problem of generalized visuospatial reasoning.

## 1. Introduction

Consider the two problems in Figure 1, Bongard Problem 31 (Bongard, 1968) and a ten-frame sequence from the Moving MNIST dataset (Srivastava et al., 2015). These two problems are quite distinct, so it follows that the reasoning processes a person or artificially intelligent system would employ to solve them might differ wildly. However, compared side by side, we may observe an important similarity: both problems have to do with overlapping objects. An agent designed to solve these sorts of problems might make a similar observation, provided it possesses two features: the concept of visual occlusion and the ability to apply its understood concepts to novel scenarios and tasks. If skills like occlusion detection can be employed by multiple distinct strategies, might a high-level reasoning system be able to treat each skill as an interchangeable operation to solve a much wider array of tasks? Does there exist a minimal set of skills that is sufficient for solving all tasks in a general domain such as visuospatial reasoning?

In this paper, we examine a hypothetical basis set of skills in the domain of visuospatial reasoning (VSR). For this set of skills to function as a useful basis for artificial agents, it should cover a wide breadth of functionality and be robust to a wide range of variations. The implementation and application of such a skillset requires a deep understanding of either the abilities themselves or their acquisition, both topics which might be studied using benchmark datasets. Through the analysis

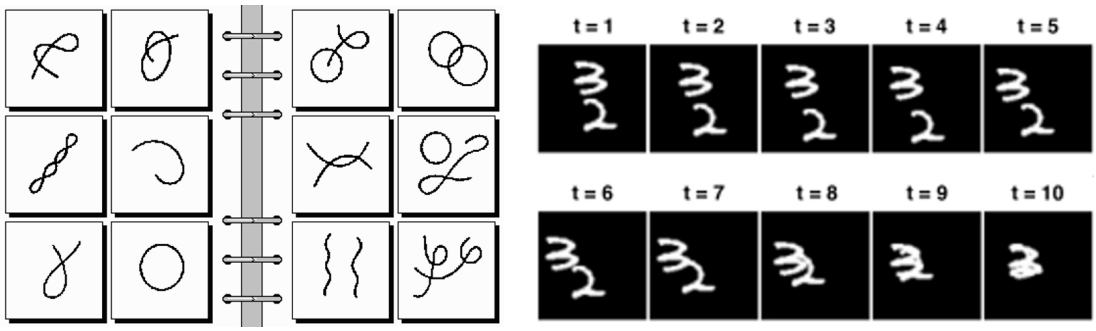



Figure 1: Left: Bongard Problem #31 (Bongard, 1968). The goal of this problem is to identify the rule that distinguishes the set of images on the left from the set on the right. Right: A 10-frame sequence from the Moving MNIST dataset (Srivastava et al., 2015). The goal of this problem might be predicting the appearance of the next frame in the sequence.

of benchmarks and the skills they beget, we intend to construct a framework for the discussion of solving VSR from a developmental perspective. The specific contributions made in this paper are:

- We define requirements for the developmental measurement of basic VSR skills, including breadth, domain specificity, depth, and explainability.
- We review existing VSR benchmarks in terms of the four requirements listed above and find that none fully and completely meet all four.
- We propose a new benchmark whose purpose is to address the problem of bestowing artificial agents with basic, expressive VSR skills. Its design aims to improve on certain characteristics we highlight in the above benchmark tests.
- We enumerate several novel applications for which our benchmark could find use, including dataset training evaluation, factor analysis in cognitive systems research, and larger-scale empirical analysis of prior lines of knowledge-based AI research, for instance qualitative spatial reasoning systems.

## 2. Visuospatial Reasoning

A meteorologist analyzes images showing the direction of wind flow and temperatures across the globe and then visualizes what might happen in the future. The curves in the dashed lines tell a foreboding story: two fronts are going to create a storm. Knowing it will be a long night at work, the meteorologist makes a cup of coffee, paying careful attention to its color while pouring in milk to obtain just the right amount of creaminess. The common thread tying these reasoning processes is that they involve visuospatial reasoning, or reasoning about space and time. In humans, visuospatial skills can be driven by a variety of modalities, including vision and touch. In both

these spatial-temporal modalities, the underlying representations are capable of encoding spatial information as experienced in those modalities.

We focus specifically on skills which preserve spatial properties as experienced by a *visual* reasoning agent. We define a VSR skill to be a function whose input and/or output spaces involve visual images, and whose mappings capture visuospatial regularities including perceptual organization and transformation. Examples of VSR skills include recognizing lines formed by multiple individual objects, rotating the mental image of an object, and isolating signal from noise from an image. Note that while the recognition of rotation would be considered a VSR skill, so too would the mental rotation of objects according to our definition.

There are two broad categories of approaches that have been taken to study visuospatial reasoning in humans: top-down and bottom-up (Tversky, 2005). In the top-down approach, visuospatial reasoning is studied as an accessory to complex cognitive processes, while in developmental approaches the mechanisms and representations of the low-level skills are studied. A similar analysis of computational approaches to study VSR indicates that a large amount of focus has been given to top-down approaches while relatively less focus has been given to developmental approaches. For example, one application of visuospatial reasoning that has been extensively studied in AI is intuitive physics reasoning (Bakhtin et al., 2019; Allen et al., 2019; Wu et al., 2016). A computational system which displays competence in this domain needs to have some understanding of spatial properties of the world in which the tasks are set. These tasks do not, however, let us directly study the building blocks which drive VSR. A set of tasks which *does* can be found in Han et al. (2020), which focuses on a specific subset of VSR skills: conversion between two- and three-dimensional visual representations. The design of tasks in this dataset allows researchers to measure and understand specific skills in detail without having to worry about the various tasks in which the skills may be employed. In this paper, we examine the developmental approach to the study of VSR, i.e. we aim to study the different VSR skills which form the basis for different cognitive processes.

Our primary motivation is to discover what drives visuospatial reasoning. To find out, we must explore the ways in which VSR skills may be expressed computationally. In other words, we require both the language to describe these skills at their most basic level and the ability to measure their existence in artificial agents. In humans, there are several different ways to distinguish between cognitive abilities that are described by psychometric, cognitive, and neurological differences. For example, Linn & Petersen (1985) present three distinct spatial skills displayed by humans and present psychometric and cognitive rationale for having such distinctions. Similarly, the existence of *what* and *where* pathways of visual perception (Haxby et al., 1994) provides a method of distinguishing between different skills based on neurological organization. While dealing with computational systems, neurological differences become less important relative to cognitive differences, which are based on functionality and representation.

Various intelligence tests like the Leiter-R (Roid & Miller, 1997) allow for testing of human proficiency in different skills based on the distinctions described above. Designing computational systems for these tests could allow researchers to study and reason about different facets of cognition which contribute towards proficiency, but these tests are hardly useful for evaluating the generalizability of computational systems' skills. Such tests are simply too small and too uniform.

## 3. Benchmarking Visuospatial Reasoning Skills

Benchmarks like the ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015) have grown to be a driving force in the field of artificial intelligence. They encourage the development of new techniques and offer platforms for comparing existing techniques. Significant progress has been made towards solving important AI problems due to the existence of such benchmarks, though they admittedly introduce challenges to the field as well (Langley et al., 2011).

In this section, we describe four dimensions along which to evaluate benchmarks for measuring an artificial agent's VSR skills. In Section 4, we examine a number of existing benchmark datasets with respect to these qualities and discuss their applicability to the problem of understanding general VSR. Note that there exists a large number of universally desirable benchmark qualities that are not specific to developmental VSR and are not discussed at length in this paper. For example, test answers should not be easily found by cheating, the test should be universally applicable to any intelligent system, and its results should be meaningful, i.e. scoring well on the test is an indicator of proficiency at the given task.

### 3.1 Breadth

The breadth of a benchmark is, simply put, the number of unique skills that it tests for. Although the classification of individual skills depends upon the granularity at which one describes a problem, several distinctions of skills are useful for discussion. The most apparent sign of a dataset's breadth is its requirement of skills belonging to numerous categories.

Visuospatial reasoning skills make up a broad set of diverse functionality. Functionality is perhaps the most discrete distinction that can be made, and it has the most relevance to cognitive systems research. The functional description of a skill includes that skill's input and output types. These types could be any kind of knowledge representation, including images, videos, three-dimensional models, and symbolic descriptions. For example, the *perceived mirroring* function might determine whether part of an image is mirrored about some axis and would return a symbolic representation of that information. The *generate mirror* function might take two images: a

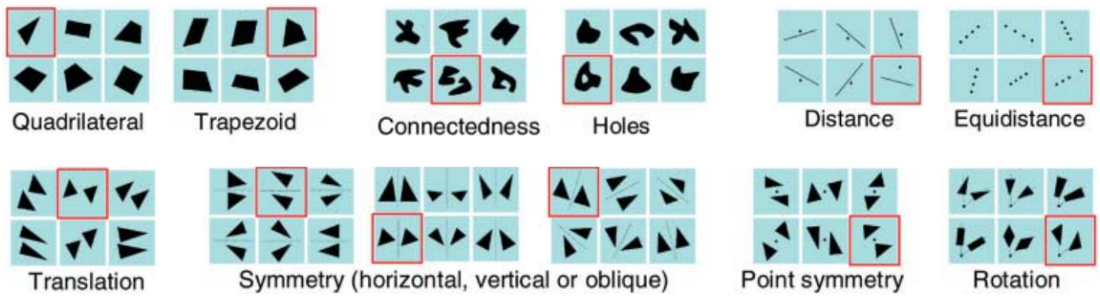



Figure 2: A number of different visuospatial skills to represent the breadth of the Core Knowledge of Geometry odd-one-out test (Dehaene et al., 2006). These skills belong to four categories. Top, from left in groups of two are geometrical figures, topology, and metric properties. The bottom row all belong to geometrical transformations. The correct answer for each item is outlined in red.

sub-image to be mirrored and an axis of symmetry, and would return a new image with the correct mirroring about the axis. The functional distinction of skills produces many such symmetrical function pairs, often with one being *receptive* and the other *expressive*. Are such skills truly unique? It is not difficult to imagine that they could share much of the same low-level machinery.

Newcombe & Shipley (2015) make an alternate distinction by which skills may be classified. They define a typology of mental imagery skills that divides the set along two axes: *intrinsic* versus *extrinsic*, and *static* versus *dynamic*. Intrinsic skills refer to the identification of features of individual objects such as shape, color and convexity. Extrinsic skills reason about relations between multiple objects, such as their spatial relationships. whereas the distinction between static and dynamic skills is based on the requirement of time-variant representation. The perception of relative sizes of two objects is thus a static skill, while mental transformation of objects following imagined motion vectors is dynamic. However, even with such typologies, the organization and taxonomy of the full breadth of VSR skills is still very much an open question. As such, a systematic review of benchmarks' breadth will inevitably be largely subjective. In our review, we tend to focus on Newcombe's typology, the apparent diversity in the benchmark's description, and the (subjectively determined) requirement for perceptive and/or generative skills.

### 3.2 Domain specificity

A caveat we must keep in mind when advocating for maximal breadth is the requirement to only measure skills in which we are interested. Given their wide variety and applicability, VSR skills tend to find use in conjunction with skills of completely different classes, rendering this requirement difficult to accomplish. An important facet of designing benchmarks is the separation of these non-VSR dependencies from the tasks themselves. This segregation is particularly important for the satisfaction of our objective of facilitating benchmarking of VSR from a developmental perspective; the required inclusion of another domain means an intelligent system's performance is contingent on its skill in the other domain.

Many VSR benchmarks *require* the use of either high-level reasoning or natural language in addition to visuospatial reasoning in order to solve test items. Bottou (2014) defines *reasoning* as "algebraically manipulating previously acquired knowledge in order to answer a new question". Benchmarks which require the use of symbolic reasoning, however, have limited usage in evaluating low-level skills because these skills, while necessary, are not sufficient to demonstrate proficiency. Many seemingly low-level VSR skills might involve symbolic reasoning, so we restrict our benchmark search to the latter operative word, *"new"*. Tests that may be solved by a single, sufficiently low-level solution algorithm will be described as being low-level.

To measure systems' abilities to answer new questions, datasets like ARC (Chollet, 2019) need wide varieties of tasks. VSR happens to be an excellent domain in which these tasks may be found. For many of the same reasons, datasets with wide varieties of VSR tasks often involve NLP. If an agent is meant to perform a highly expressive task, it follows that its communicative skills should be at least as expressive.

In our evaluation of existing benchmarks, domain specificity is typically easily observable, especially in the case of NLP-related tests. The low-level reasoning requirement, admittedly defined vaguely, is subjectively identifiable by a test's uniformity, notably at odds with the aforementioned

Breadth requirement. Partitioning a test into sufficiently described sub-tests alleviates this discrepancy, as a diverse test with such divisions (each being solvable by a single low-level algorithm) satisfies the low-level requirement.

### 3.3 Depth

The skills encompassed by VSR are characterized by their ubiquity and robustness, so a strong VSR system must have a *deep* understanding of each skill. When we discuss the depth of a dataset, we are referring to a breadth of factors beyond the primary array of skills that the dataset is designed to capture. Rotation, for example, may be applied to any imagined object, about any origin, in any direction. Robustness may also refer to the recognition and generation of visual features: a circle with a bite taken out of it is no longer a literal circle, but only without the gestalt reasoning required to compartmentalize the concepts of circle and bite.

Depth can be likened to generative factors in disentanglement datasets: a deep dataset contains many factors, and it also varies them as much as possible. Deep datasets are common in deep learning; a generalizable concept may be learned most easily with many varying examples along a sufficiently densely sampled target distribution. The ubiquity of VSR skills means that VSR system should have maximal transferability of its skills; its target distribution is as large as possible. Matching that target distribution is a lofty goal that any training dataset short of the human experience can only approach. We may characterize the depth of datasets and systems with two factors: the number of axes of variation and the density of the sampling in the space those axes define. Such factors

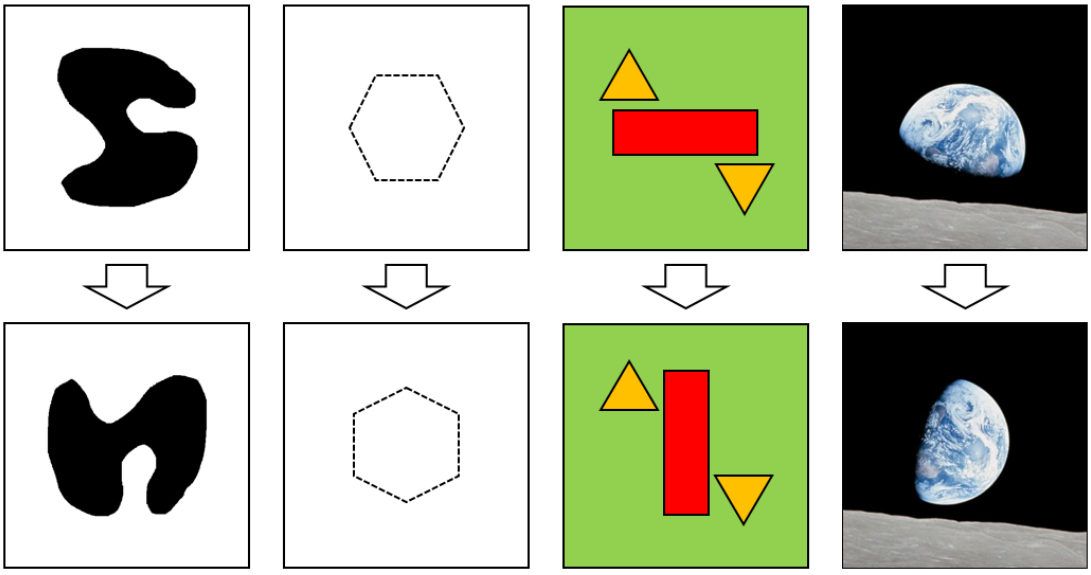


Figure 3: A demonstration of depth: one skill, 2D planar rotation of the central object, applied in four distinct scenarios. Left to right: an irregular shape, a dashed outline of a polygon, a red rectangle with two static triangles, and a modified version of *Earthrise* (Anders, 1968).

are difficult to measure, so in our review we use the number of items and test descriptions to guide subjective comparisons of tests' depth. When publishing our own benchmark, we plan to explicitly characterize its breadth for future ease of comparison, and we encourage other researchers to do the same.

### 3.4 Explainability

While human VSR can incorporate skills in all kinds of strategies, many VSR datasets end up being closely tied to a set of solution methods. Cognitive architectures designed to exhibit human-like behavior might also use varieties of strategies, so it is important that benchmarking not favor any particular solution method. Factor analysis allows test designers to enumerate unobservable *factors* that affect performance on tests based on the empirically determined relationships between items in individuals' testing. In Section 3.2, we note the importance of breadth of skills; that each *type* of skill should be represented in addition to a large number of skills. Despite the difficulty of skill classification, any amount of typological scoring has the potential to help researchers understand their systems' abilities and deficiencies. Likewise, in Section 3.3, we bring attention to the number of axes of variation, and the amount of variation along each axis in intelligence tests. Just as with skills, a dataset might display individual scores—or features such as the standard deviation of score—for each problem variant. If a test set contains many visual styles, for example, it would be useful to know when an AI system succeeds at problem solving with one style but fails with all others. This kind of feedback is particularly useful for ablation studies, in which researchers study the necessity of individual components of AI systems.

Interpretability is a growing problem in the field of artificial intelligence, especially given recent trends in deep learning (Chakraborty et al., 2017). An ongoing effort attempts to interpret *black box* models' skills and strategies. We find in our review that tests tend to ignore the potential for explainablity-directed scoring, since researchers may dissect and analyze their systems' individual responses to questions. To facilitate explainable research of fundamental skills, benchmark tests with private test sets such as ARC (Chollet, 2019) and the novel benchmark proposed in Section 5 of this paper should strive to make up for the want of this ability. In our review of existing benchmarks, we identify primitive explainability with the presence of sub-tests and enumerated skills.

## 4. Survey of Benchmarks for VSR

In this section, we examine a number of VSR and VSR-adjacent datasets, tests, and environments in the domains of psychometrics, intuitive physics, dimensional projection, goal-driven, QSR, and linguistic VSR. Each subsection in this section contains a small sampling of representative tests; a more thorough review may be found in the online supplemental material[1].

1. Online appendix: `https://github.com/aivaslab/vsr-benchmark-review`

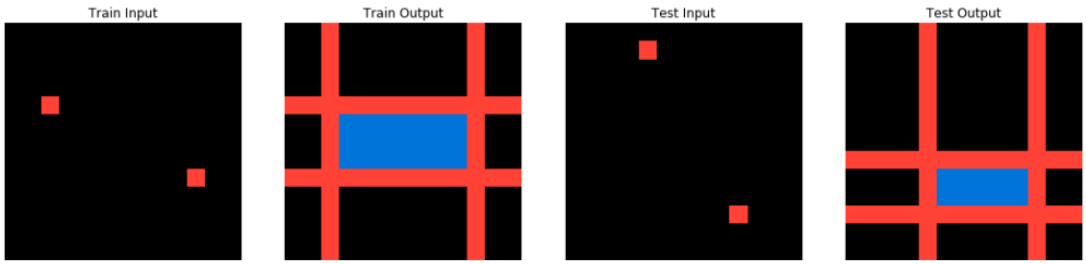



Figure 4: A problem in the ARC public validation set. Left to right: example input, example output, test input, test output generated by an AI program. For this problem, more unshown example inputs are provided. Due to ARC's all-or-nothing scoring, this AI-generated output would receive a score of zero despite missing only a single aspect of the problem's form: that red input pixels correspond to output corners, not arbitrary positions along the red lines.

### 4.1 Psychometric Tests

We would like to draw inspiration from psychometrics, or the study of measuring information about individuals such as skills and knowledge. Often such tests have impressive breadth or explainability in order to capture skills' generalizability in human lives. A number of psychometric tests study VSR skills, but since they are primarily intended for humans, they tend to make certain assumptions that hinder their use for artificial intelligence, the primary assumption being the immense amount of prior knowledge acquired over a typical human's lifetime. Many intelligence tests assume that an individual cannot take the same test twice, so for fair comparison these should be unknown to both algorithm and programmer prior to evaluation.

Raven's Progressive Matrices (RPM) is a popular intelligence test designed by John C. Raven in 1936. Each of sixty items in an RPM test features a square matrix of images with one image missing. The correct missing image is selected from a small bank of options. Items in RPM are always geometric analogies, and almost all can be solved with only a handful of image operations including set addition, subtraction, union, and conjunction (Kunda, 2013). RPM is far too narrow to qualify as a test of general VSR skills.

The Core Knowledge of Geometry test described in (Dehaene et al., 2006) is a developmental approach to the measurement of geometric concepts. Its problem format is selecting the odd one out of a set of six images, where five items in the set are linked by some geometric property like symmetry, concavity, or equidistance (of a group of points). There are only forty-five questions, but many skills have multiple variants, and skills are organized into categories like Euclidean geometry, topology, and geometrical figures. While this test has neither the breadth nor the depth to evaluate an agent's generalized VSR skills, its design—the focus on fundamental skill measurement and the hierarchical organization—is ideal for a VSR skill benchmark.

The range of VSR skills in the Abstraction and Reasoning Corpus (ARC), however, is quite broad (Chollet, 2019). ARC is an AI benchmark that mirrors the form and function of psychometric tests; it is meant to measure the generalization ability of a system across many unique problems. ARC is composed of one thousand tasks, where each task has a small set of training input-output

pairs and one test pair. Each item in a pair is an image of arbitrary (between one and thirty pixels) size and up to ten colors. Only six-hundred tasks are available for programmers; the rest are unknown for blind evaluation. Although ARC is meant as a measure of a machine's ability to generalize intelligence, it makes heavy use of visuospatial skills. Chollet assumes an intelligent problem solver has already been imbued with certain cognitive priors, including certain classes of VSR skills, basic arithmetic, and causality. As such, individual ARC items might call upon specific VSR skills, but overall ARC requires a higher reasoning component to make decisions regarding which skills should be applied and how. Although ARC's documentation in Chollet (2019) describes a number of skill *priors* a system is assumed to have, the list is relatively high-level. Additionally, skill descriptions are not given to individual problems or even groups of problems, so the test's all-or-nothing scoring (as shown in 4) results in—for our developmental purposes—a relatively uninformative test set. ARC images are also small and discrete, so skills designed for these toylike settings would likely not find general-purpose use in real-world tasks.

The current best-performing solution to ARC, after a monetized competition on Kaggle.com, notably makes use of forty-two functions or 142 function variants (icecuber, 2020). These functions, each of which uses imageistic input and output, were originally preprogrammed for use in hard-coded solutions to the first one hundred training problems. A solution to each task is generated by a search through combinations of three or four of these functions. Most other high-scoring submissions for the same competition make use of domain-specific languages with hard-coded image functions (de Miquel, 2020; Golubev, 2020).

### 4.2 Linguistic VSR and Visual Question Answering

Popularized in Hofstadter's *Gödel, Escher, Bach* (Hofstadter et al., 1979), the Bongard problems are a set of one hundred puzzles meant to test visual perception (Bongard, 1968). Each problem presents two sets of six images, and the goal is to identify the rule—expressed linguistically—that determines whether an item belongs to set $A$ or set $B$. Since solutions are often unique VSR skills, the Bongard Problems have a substantial breadth. A slight modification, adding one item to a Bongard problem that clearly belongs to one set according to the rule, solves the dependence on natural language understanding. The United Kingdom Clinical Aptitude Test (UKCAT) Abstract Reasoning Test is one such test that is used by medical schools in the United Kingdom to select applicants. Like the Bongard Problems, each task features two groups of six items that must be distinguished. Each task has five additional image items which may be assigned to fit into group A, group B, or neither.

Another test format that couples visual cognition and linguistic skills is that of visual question answering (Antol et al., 2015; Goyal et al., 2017). The task in these tests is to answer natural language questions about images. The general format of these tests allows for testing a wide range of visual cognitive abilities. For the purpose of studying visual reasoning abilities, datasets like CLEVR (Johnson et al., 2017), SHAPES (Andreas et al., 2016) and NLVR (Suhr et al., 2017) have been developed, in which the questions are geared towards studying *primitive* aspects of visual cognition like counting, spatial relationships and attribute identification. While all these datasets focus on answering natural language questions about synthetic or real-world scenes, a related line of inquiry has been into the topic of diagram understanding. Scientists are able to incorporate and extract

a dense amount of information into and from diagrammatic images. Datasets like FigureQA (Kahou et al., 2017), FigureSeer (Siegel et al., 2016) and AI2D (Kembhavi et al., 2016) have been built to test the capabilities of diagram understanding systems. The images in these datasets vary greatly, ranging from middle-school level science diagrams to figures from scientific research publications.

The datasets mentioned above are all extremely useful in the understanding of visual reasoning. However, the solutions of all their tasks necessitate natural language skills. Thus, in a strict sense, they do not test for VSR skills alone.

### 4.3 Qualitative Spatial Reasoning

Qualitative Spatial Reasoning (QSR) involves the solution of visuospatial problems using structured abstract representations like semantic webs instead of with mental imagery (Freksa, 1991). Much of VSR is encompassed by QSR, as many VSR strategies involve re-representing imagistic data as descriptions and then reasoning about descriptions. As Sioutis & Condotta (2014) point out, "In the literature of qualitative spatial reasoning there has been a severe lack of datasets for experimental evaluation of the reasoners involved", with one exception being the Administrative Geography of Great Britain (Admingeo) dataset (Goodwin et al., 2008). Admingeo encodes geographical features of Great Britain using the RDF Schema, a data model that allows for structured resource descriptions. Lovett et al. (2007, 2010) make use of qualitative spatial representations to solve Raven's Progressive Matrices problems. Ultimately, QSR describes an important set of strategies that may be used to solve VSR problems, so a VSR benchmark could also be used as an empirical test of the general applicability of a qualitative spatial representation.

### 4.4 VSR using 2D/3D representations

Certain tests in the field of visuospatial reasoning have looked exclusively into the problem of generating 2d images of 3d objects from different perspectives. Tests for mental rotation developed by Shepard & Metzler (1971, 1988) and Vandenberg & Kuse (1978) have helped us understand human cognition much better over the decades. More recently, the Purdue Spatial Visualization Test (Guay, 1977) and its revised version developed by Yoon (2011) have been used to study 3D mental rotation abilities in individuals.

Recently, Han et al. (2020) presented SPARE3D, a dataset of line drawings of three-dimensional objects along with several tasks which test for visual reasoning abilities. The aim of the benchmark is to train machine learning models to reason about 3D objects from 2D line drawings. The dataset attempts to overcome existing bottlenecks in other spatial reasoning datasets by way of its large size, with automated generation of 2D line drawings from 3D CAD models, and the decoupling of visuospatial skills from the use of language. The tasks measure view consistency, camera pose, and shape generation, and each task includes a large number of test items which allows testing for generalization over a large set of diverse objects. Another interesting facet of SPARE3D is the task format: while some tasks are multiple choice, some require the generation of line drawings and 3D point cloud models. These tasks represent a decoupling of VSR skills from high level reasoning required to solve problems and are a useful mechanism of studying specific VSR skills in a systematic bottom-up manner.

### 4.5 Egocentric and Goal-driven AI

For all practical purposes, the real world has three dimensions of space and one of time, so it is no surprise that many systems are designed to navigate 3D environments. A massive body of research is dedicated to the creation of goal-driven, egocentric AI systems, many of which implicitly make use of VSR skills to function. This category includes mapping, navigation, and much of reinforcement learning research. Touchdown (Chen et al., 2019) is an RL environment that can be navigated similarly to Google Street View. It contains two tasks: one is navigation from highly-contextual written instructions, and the other is the identification of a specific point in a scene based on a description. Although both tasks clearly require an understanding of natural language, they also emphasize spatial reasoning, as instructions typically involve directions and distances relative to objects.

### 4.6 Video Prediction and Physics Simulation

Humans are inherently capable of reasoning about some physical phenomenon in their surroundings. Such reasoning ability, though limited is present in children and adults irrespective of their cultural and educational background. One possible reason for this could be that humans use their visual experiences from everyday life to create a model of how the physical world functions. As such, a powerful mechanism to study intuitive physics is to study how visual reasoning systems are able to learn rules of physical interaction from visual experience. This has led to the development of several datasets and computational approaches whose goal it is to develop an intuitive understanding of physics. IntPhys (Riochet et al., 2018) is a dataset which is designed to test the understanding of four concepts: object permanence, shape constancy, spatio-temporal continuity and energy conservation. The task is to look at videos showing physical interactions of synthetic objects and differentiate between videos which demonstrate one of the concepts and those that do not demonstrate them. Another similar dataset is Bounce (Purushwalkam et al., 2019), which consists of several videos of a foam ball bouncing against different surfaces. The task of predicting the trajectory of the ball after collision then requires any model to be able to compensate for different surface characteristics such as friction in order to predict accurately the future trajectory. One of the advantages of these datasets is that the number of datapoints in these systems is large. This might potentially allow the use of new deep learning techniques to try to solve the problem of physics understanding. Another related dataset is the Physics-101 dataset developed by Wu et al. (2016). It was developed with the goal of learning physical properties of interacting objects from videos.

Another class of physics reasoning datasets aims to study the ability of an AI agent to learn to reason about physics. The Physical Reasoning benchmark (PhyRe) is a set of two-dimensional, Newtonian physics puzzles with an image initial state and a linguistic goal state (Bakhtin et al., 2019). In each item, an agent chooses the position and size of one or two balls to be placed in the 2D environment. After the agent has positioned the ball(s), physics is activated, and the various objects in the scene interact amongst themselves, possibly achieving the goal after some time has passed. PhyRe has considerable depth, featuring a large number of minor variations for each of its problems. Similarly the TOOLS challenge (Allen et al., 2019) requires agents to position objects

in a scene so that the application of laws of physics like gravity leads to a particular goal being achieved.

Apart from these datasets, there are a few which are used for the problem of video prediction. Though not explicitly used for the purpose of understanding intuitive physics, recent research (Denton et al., 2017; Srivastava et al., 2015) has demonstrated that certain rules of physics can be learned through unsupervised learning from videos. Two datasets which are commonly used for this purpose are the moving MNIST and the Bouncing Balls (Sutskever et al., 2009) datasets.

## 5. A Proposed Benchmark Design for Isolating Diverse, Low-level Skills

Here we propose a novel VSR benchmark design whose purpose is to fill in the gaps in existing research which we have outlined above. This proposal is not meant to be interpreted as a final design, but we intend to abide by the general format and design principles, which describe an arbitrarily flexible yet developmentally-oriented test, in future work.

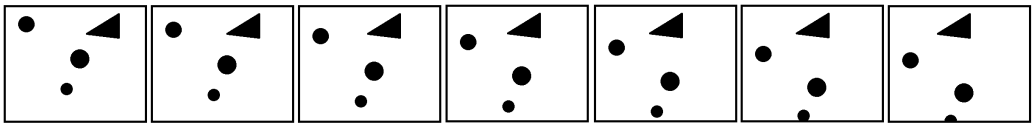
### 5.1 Video Prediction from Example

Our chosen problem format is a modified version of video prediction. Given one (or more) example video(s) and an input image, the goal is to predict the next frame(s) in a sequence following the input image. Both the example videos and the input image contain a number of *objects* with identifiable intrinsic *attributes*. The example videos also contain a number of *transformations* which are applied to the objects in the videos; in the case of the Moving subtest, all objects move according to some constant vector. Additional subtests have details that differ slightly to allow for the success of low-level algorithmic solutions. The output frame may always be inferred by applying inferred transformations to the objects in the input frame. Two output formats are provided for different modes of comparison: video prediction (imagistic), and true/false questions of the form *this image follows the input in the sequence*.

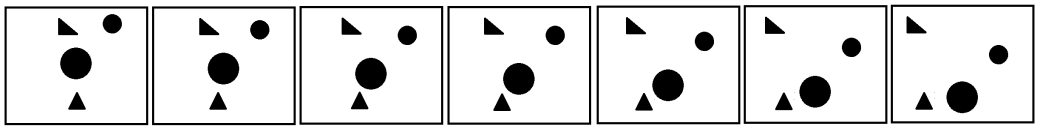
### 5.2 Satisfying design requirements

The primary goal of this test's design is to allow for the creation of problems that test an extensive number of enumerable skills. Domain specificity strongly limits such a goal, as task diversity is often paired with algorithmic flexibility or high-level reasoning, for which the kind of developmental benchmark we propose here should *not* test. To avoid high-level reasoning of this kind, we impose the constraint that all problems must be solvable with the same sufficiently low-level algorithm.

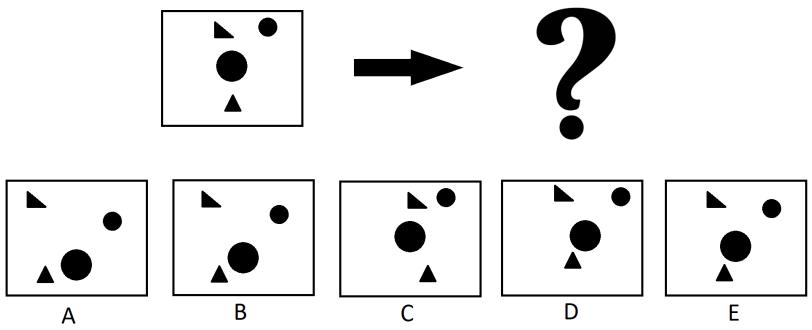
Several design choices explicitly address the breadth requirement. Let us consider the recognition and application of rotation as two different skills which correspond to the receptive and expressive functionalities discussed in Section 3.1. Our goal through the proposed design is to demonstrate flexibility to test for a wide variety of skills and functionalities. To this end, though we do not make commitments to representations and architectures, any system that is competent at this task should internally be able to display both of those mentioned skills. As an example, let us consider a system which performs this functionality in the following simple manner: it first analyses the video it sees and finds possible points of interest. Then, based on its analysis in the previous step, it



(a) Demonstration Example: The given sequence of frames demonstrate triangles translating towards the left and circles translating downwards. A solver is expected to infer these rules and subsequently apply them.



(b) Video prediction Test Frames. The first frame is provided to the system; the rest should be generated as predictions.




(c) Correctness Evaluation: Which (if any) of these frames might eventually follow the input frame? The correct answers spell a common nickname of the 16th US president.

Figure 5: A Video Prediction from Example item, with one example sequence (a) and two tasks (b) and (c). Real examples would not be restricted to black and white images such as these.

applies the required transformation to generate the subsequent frames. For this system, the test design necessitates modules with the following two functional descriptions: ImagesToDescription and ImagesAndActionToImage. The first module with functional description ImagesToDescription would then be tasked with the perception of rotation from the frames seen, while the second module (ImagesAndActionToImage) would have the task of applying the correct amount of rotation in the subsequent frames. While adhering to these functional descriptions is not mandatory, any system which is competent at this task should, either explicitly or implicitly, perform these or similar functions. In fact, even for a system which uses these functional descriptions, several sub-modules might have use different designs and we leave these entirely as the system designer's choice.

Additionally, a small number of task variants are included in our test. Each task variant is an example of the same functional appearance and high-level task, video prediction from example, but they differ in lower-level distinctions for ease of defining algorithmic solutions. The Motion-3D task addresses 2D-3D image understanding. Motion-R focuses on the ability to recognize extrinsic features. Grouping, Objectness, and Occlusion all approach the robust ability of humans to segregate

| Subtest | Description |
|---|---|
| Motion | Objects move continuously based on intrinsic or extrinsic features |
| Motion-R | Like Motion, but Motion is relative to objects or features |
| Motion-3D | Like Motion, but objects move out of plane |
| Grouping | Like Motion, but moving "objects" are made of smaller objects |
| Objectness | Like Motion, but objects collide and stop upon collision |
| Occlusion | Like Motion, but objects occlude each other |
| Geometric Sequences | Simple geometric patterns, including analogies |
| Repeating Patterns | Repetition of a simple rule as in a fractal or tiling |
| Mixed | All of the above |

Table 1: Nine proposed Video Prediction from Example task variants, designed to cover a wide breadth of skill types outside of those demanded by the high-level task structure. This table is meant to demonstrate how the test can be partitioned into sub-tests to flexibly measure a broad variety of skills without adding a higher-reasoning requirement, since each sub-test has a relatively specific algorithmic solution.

scenes into object-level descriptions. Geometric Sequences and Repeating Patterns both contain discrete transformations. Finally, a large number of explicitly enumerated skills will be built into the problem generator, drawing inspiration from all the tests listed in Section 4.

In addition to task variety, our benchmark will also have consistent visual style variety across all tasks. In all subtests, objects may be silhouettes, outlined cartoons, or even 3D shapes (even for 2D tasks) with varied lighting. As in the task variant distinction, scores will be reported for each of these categories.

Explainability is partially achieved by the aforementioned score reports, which will give researchers detailed reports on algorithms' skills at different task variants, style variants, and individual visuospatial skills. The video prediction will undoubtedly shed some insight as to agents' reasoning capabilities, but we leave the rest of the problem of understanding artificial agents in the hands of their designers.

### 5.3 Weaknesses

Our proposed benchmark is far from an ideal measure of VSR skills. As with all benchmark tests, there are likely unforeseen methods of "cheating," or finding the correct answers in a way that bypasses the intended reasoning process. An explicit detail of our benchmark's design is a desire to require generative, or expressive, skills, yet its primary answer format is multiple choice. Intuitively, its answers seem to require video prediction abilities, but a system that describes image differences linguistically might also succeed.

Procedurally generating the dataset with software has many upsides, including uniform sampling of spaces, guaranteed representation of each skill, and the dataset's size; but generation also has downsides. Most obviously, since the data is generated on a computer, it will not contain photorealistic images. Our synthetic dataset cannot be compared to the incredibly noisy, varied, and high

resolution inputs of humans' and animals' everyday sensory experiences, so we cannot claim its solutions will generalize nearly as well. That said, we expect that it will retain much of its use as a test set, so systems trained with real-world data might still be evaluated with sufficient skill transfer.

Aside from the richness of the human sensory experience, we also possess a number of specialized skills that this dataset makes no attempt to measure. Among these are large-scale object recognition, face detection, and the ability to recognize foot trails in forests, as well as a menagerie of abilities that fail the domain specificity requirement.

The breadth and depth of our proposed benchmark is limited by the creativity of its designers. A seemingly limitless vocabulary of VSR skills may be enumerated with enough thought. Axes of variation added for the purpose of depth are also arbitrarily expressive. Since we wish for a dense sampling of their space, the size of our dataset grows exponentially for each added axis.

### 5.4 Similarity to Existing Tests

Although a union of existing tests could certainly help with either breadth or depth, simple unioning of existing tests would not be sufficient for a developmental VSR benchmark, since most existing tests lack adequate breadth, depth, explainability, or domain specificity. Combining two tests as their union would essentially miss out on the rich space of problems interpolated between the two tests. For example, imagine a union of the tasks in the Spare3D dataset with the Bongard Problems. The goal of such a union might be to have Bongard-like problems with the depth and explainability of the Spare3D representations, but a simple union of the two datasets would not produce this outcome, without giving additional effort into generating new problems that combine properties of each test.

The test we have outlined in Sections 5.1—5.3 bears much resemblance to the task presented by Moving MNIST and similar tests from the video prediction and physics understanding literature (Srivastava et al., 2015; Hsieh et al., 2018; Denton et al., 2017; Allen et al., 2019; Bakhtin et al., 2019; Riochet et al., 2018), the primary distinction being that the test sequence for each section has only one starting frame. Salient features in the initial sequence of our proposed test must be extracted just like features in Bongard Problem sets and ARC example pairs. In the future, datasets like Moving MNIST might be used more often as toy datasets like how MNIST is used today. Unlike the Bongard Problems and ARC, each enumerated skill tested by our benchmark is intended to be recruited in multiple examples under many circumstances.

## 6. Future work

No small dataset can capture the breadth and depth of skills required for VSR. For this reason, we are interested in the applicability of egocentric video datasets as repositories of prior experience for artificially intelligent agents solving our (and others') VSR tasks. Likewise, we expect that training on other VSR benchmark datasets will transfer specific universal skills that we hope to measure. Can we empirically characterize any visuospatial dataset in terms of its teaching power for different machine learning systems? Can machine learning systems be characterized by their ability to transfer learned VSR abilities, in terms of both breadth (inter-skill) and depth (intra-skill)?

As mentioned in Section 3.4, factor analysis is a useful tool for the design and interpretation of psychometric tests. To our knowledge, the same principles have not been applied to cognitive systems. After all, the thought processes of such systems are often known, so characterizing them in terms of "latent" variables seems pointless. As hypothesized factors guide the development of human psychometric tests, so too might they improve the design of future versions of AI benchmarks. Latent factors have the potential to describe individual questions, benchmark sub-categories, and even complete benchmarks more thoroughly than trivial concepts like average score.

We also mentioned the proximity of our research to that of QSR and other lines of knowledge-based AI research that may have been or may be hampered by a lack of comprehensive and large-scale benchmark datasets. A future version of our benchmark dataset could be generated including schematic descriptions of each image. This way, it can be used for multiple QSR tasks including translation from images to schemas, translation from schemas to images, and a novel dataset of schematic questions to be answered by QSR systems.

## Acknowledgements

We would like to thank David Crandall, Linda Smith, and Chen Yu for a discussion of this research, and also the anonymous reviewers for their helpful feedback. This work was supported in part by NSF BCS Award #1730044 and NSF DGE Award #1922697 through the Neurodiversity Inspired Science and Engineering (NISE) NSF Research Traineeship (NRT) program.

## References

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2019). The tools challenge: Rapid trial-and-error learning in physical problem solving. *arXiv preprint arXiv:1907.09620*.

Anders, W. (1968). *Nasa image as08-14-2383*. NASA. From `http://www.hq.nasa.gov/office/pao/History/alsj/a410/AS8-14-2383HR.jpg`.

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016). Neural module networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 39–48).

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual Question Answering. *International Conference on Computer Vision (ICCV)*.

Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L., & Girshick, R. (2019). Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems* (pp. 5083–5094).

Bongard, M. M. (1968). *The recognition problem*. Technical report, Foreign technology div Write-Patterson AFB Ohio.

Bottou, L. (2014). From machine learning to machine reasoning. *Machine learning*, *94*, 133–149.

Chakraborty, S., et al. (2017). Interpretability of deep learning models: a survey of results. *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (pp. 1–6). IEEE.

Chen, H., Suhr, A., Misra, D., Snavely, N., & Artzi, Y. (2019). Touchdown: Natural language navigation and spatial reasoning in visual street environments. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 12538–12547).

Chollet, F. (2019). The measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Dehaene, S., Izard, V., Pica, P., & Spelke, E. (2006). Core knowledge of geometry in an amazonian indigene group. *Science*, *311*, 381–384. From `https://science.sciencemag.org/content/311/5759/381`.

Denton, E. L., et al. (2017). Unsupervised learning of disentangled representations from video. *Advances in neural information processing systems* (pp. 4414–4423).

Freksa, C. (1991). Qualitative spatial reasoning. In *Cognitive and linguistic aspects of geographic space*, 361–372. Springer.

Golubev, V. (2020). 3rd place[short preview+code]. From `https://www.kaggle.com/c/abstraction-and-reasoning-challenge/discussion/154305`.

Goodwin, J., Dolbear, C., & Hart, G. (2008). Geographical linked data: The administrative geography of great britain on the semantic web. *Transactions in GIS*, *12*, 19–30.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6904–6913).

Guay, R. (1977). Purdue spatial visualization test-visualization of rotations. *W. Lafayette, IN. Purdue Research Foundation*.

Han, W., Xiang, S., Liu, C., Wang, R., & Feng, C. (2020). Spare3d: A dataset for spatial reasoning on three-view line drawings. *arXiv preprint arXiv:2003.14034*.

Haxby, J. V., Horwitz, B., Ungerleider, L. G., Maisog, J. M., Pietrini, P., & Grady, C. L. (1994). The functional organization of human extrastriate cortex: a pet-rcbf study of selective attention to faces and locations. *Journal of Neuroscience*, *14*, 6336–6353.

Hofstadter, D. R., et al. (1979). *Gödel, escher, bach: an eternal golden braid*, volume 13. Basic books New York.

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

icecuber (2020). 1st place solution. From `https://www.kaggle.com/c/abstraction-and-reasoning-challenge/discussion/154597`.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kahou, S. E., Michalski, V., Atkinson, A., Kádár, Á., Trischler, A., & Bengio, Y. (2017). Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., & Farhadi, A. (2016). A diagram is worth a dozen images. *European Conference on Computer Vision* (pp. 235–251). Springer.

Kunda, M. (2013). *Visual problem solving in autism, psychometrics, and ai: the case of the raven's progressive matrices intelligence test*. Doctoral dissertation, Georgia Institute of Technology.

Langley, P., et al. (2011). The changing science of machine learning. *Machine Learning*, *82*, 275–279.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child development*, (pp. 1479–1498).

Lovett, A., Forbus, K., & Usher, J. (2007). Analogy with qualitative spatial representations can simulate solving raven's progressive matrices. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Lovett, A., Forbus, K., & Usher, J. (2010). A structure-mapping model of raven's progressive matrices. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

de Miquel, A. (2020). Our contribution to the 2nd place solution. From `https://www.kaggle.com/c/abstraction-and-reasoning-challenge/discussion/154391`.

Newcombe, N. S., & Shipley, T. F. (2015). Thinking about spatial thinking: New typology, new assessments. In *Studying visual and spatial reasoning for design creativity*, 179–192. Springer.

Purushwalkam, S., Gupta, A., Kaufman, D. M., & Russell, B. (2019). Bounce and learn: Modeling scene dynamics with real-world bounces. *arXiv preprint arXiv:1904.06827*.

Riochet, R., Castro, M. Y., Bernard, M., Lerer, A., Fergus, R., Izard, V., & Dupoux, E. (2018). Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*.

Roid, G. H., & Miller, L. J. (1997). Leiter international performance scale-revised (leiter-r). *Wood Dale, IL: Stoelting*.

Russakovsky, O., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*, 211–252.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*, 701–703.

Shepard, S., & Metzler, D. (1988). Mental rotation: effects of dimensionality of objects and type of task. *Journal of experimental psychology: Human perception and performance*, *14*, 3.

Siegel, N., Horvitz, Z., Levin, R., Divvala, S., & Farhadi, A. (2016). Figureseer: Parsing result-figures in research papers. *European Conference on Computer Vision* (pp. 664–680). Springer.

Sioutis, M., & Condotta, J.-F. (2014). Tackling large qualitative spatial networks of scale-free-like structure. *Hellenic Conference on Artificial Intelligence* (pp. 178–191). Springer.

Srivastava, N., Mansimov, E., & Salakhudinov, R. (2015). Unsupervised learning of video representations using lstms. *International conference on machine learning* (pp. 843–852).

Suhr, A., Lewis, M., Yeh, J., & Artzi, Y. (2017). A corpus of natural language for visual reasoning. *55th Annual Meeting of the Association for Computational Linguistics (2)* (pp. 217–223).

Sutskever, I., Hinton, G. E., & Taylor, G. W. (2009). The recurrent temporal restricted boltzmann machine. *Advances in neural information processing systems* (pp. 1601–1608).

Tversky, B. (2005). Visuospatial reasoning. *The Cambridge handbook of thinking and reasoning*, (pp. 209–240).

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and motor skills*, *47*, 599–604.

Wu, J., Lim, J. J., Zhang, H., Tenenbaum, J. B., & Freeman, W. T. (2016). Physics 101: Learning physical object properties from unlabeled videos. *BMVC* (p. 7).

Yoon, S. (2011). Revised purdue spatial visualization test: Visualization of rotations (revised psvt: R). *Texas A&M University, College Station, TX*, *7*.