**Naval Research Laboratory**
Washington, DC 20375-5320

NRL/7340/MR—2024/1

# Data Synthesis of Complex Spatial Relationships for Visual Reasoning

JOSHUA TYLER HUGHES

CHRISTOPHER J. MICHAEL

*Center for Geospatial Sciences Branch*
*Ocean Sciences Division*

February 15, 2024

**DISTRIBUTION STATEMENT A:** Approved for public release; distribution is unlimited.

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

**1. REPORT DATE** *(DD-MM-YYYY)*
15-02-2024

**2. REPORT TYPE**
NRL Memorandum Report

**3. DATES COVERED** *(From - To)*
17-03-2020 – 17-03-2022

**4. TITLE AND SUBTITLE**

Data Synthesis of Complex Spatial Relationships for Visual Reasoning

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**
NISE

**6. AUTHOR(S)**

Joshua T. Hughes and Christopher J. Michael

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**
N2X1

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375-5320

**8. PERFORMING ORGANIZATION REPORT NUMBER**

NRL/7340/MR--2024/1

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375-5320

**10. SPONSOR / MONITOR'S ACRONYM(S)**

NRL-NISE

**11. SPONSOR / MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**

**DISTRIBUTION STATEMENT A:** Approved for public release; distribution is unlimited.

**13. SUPPLEMENTARY NOTES**
Karles Fellowship

**14. ABSTRACT**

Visual reasoning for classifying spatial relationships has become a popular topic of research in recent machine learning studies. Simple spatial relationships (SSRs), or spatial relationships between two objects, are often well represented in visual question answering (VQA) datasets used for training visual reasoning models. Complex spatial relationships (CSRs) however, or spatial relationships that are combinations of SSRs, are not generally well represented due to the randomness of object layout during dataset generation. One such dataset, CLEVR, is the inspiration for many recent VQA datasets. By introducing a CSR called the aligned relationship, the research presented here seeks to improve on the limitation of these datasets with a model that parameterizes stochastic object placement. This model aids in VQA dataset generation by allowing control of the probability that CSRs will be expressed. Working with the CLEVR generation tool, this work also shows that datasets generated with certain distributions of these probabilities can be used to improve visual reasoning models.

**15. SUBJECT TERMS**

**16. SECURITY CLASSIFICATION OF:**

| a. REPORT | b. ABSTRACT | c. THIS PAGE | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES |
|---|---|---|---|---|
| U | U | U | U | 16 |

**19a. NAME OF RESPONSIBLE PERSON**
Joshua T. Hughes

**19b. TELEPHONE NUMBER** *(include area code)*
(228) 688-5185

**Standard Form 298 (Rev. 8-98)**
**Prescribed by ANSI Std. Z39.18**

This page intentionally left blank.

# Data Synthesis of Complex Spatial Relationships for Visual Reasoning

Hughes, J. Tyler    Michael, Chris J.

August 29, 2023

**Abstract**

Visual reasoning for classifying spatial relationships has become a popular topic of research in recent machine learning studies. Simple spatial relationships (SSRs), or spatial relationships between two objects, are often well represented in visual question answering (VQA) datasets used for training visual reasoning models. Complex spatial relationships (CSRs) however, or spatial relationships that are combinations of SSRs, are not generally well represented due to the randomness of object layout during dataset generation. One such dataset, CLEVR, is the inspiration for many recent VQA datasets. By introducing a CSR called the aligned relationship, the research presented here seeks to improve on the limitation of these datasets with a model that parameterizes stochastic object placement. This model aids in VQA dataset generation by allowing control of the probability that CSRs will be expressed. Working with the CLEVR generation tool, this work also shows that datasets generated with certain distributions of these probabilities can be used to improve visual reasoning models.

## 1 Introduction

Distinguishing spatial relationships is among one of the most important classification tasks in recent visual reasoning research. A popular visual question answering (VQA) dataset called *CLEVR* now represents a baseline for testing many visual reasoning models [1]. This dataset, generated by a custom

---

Manuscript approved January 22, 2024.

tool, comprises 3D scenes of simple but distinguishable geometric shapes and the associated questions about their attributes and spatial relationships. CLEVR, and the many datasets inspired by CLEVR, only express simple spatial relationships (SSRs), or how the location of a single object relates to another [2–9]. The randomness of the object layout allows CLEVR to express all possibilities of SSRs, but not specific complex spatial relationships (CSRs), or those being combinations of SSRs. Due to the uniform random distribution of object positions, the generator does not adequately create realistic scenarios where many CSRs consistently emerge.

To improve on this limitation, we present a model that parameterizes stochastic object placement so that the probability of a specific CSR existing can be controlled via a single adjustable parameter called the *mix value*. The model employs an optimization technique called a genetic algorithm (GA) to locate optimal parameters for the object placement algorithm such that the mix value best approximates the probability of the existence of the CSR . To implement this model, we introduce the *aligned relationship*, a CSR that describes whether a set of objects are approximately collinear. By integrating the model with CLEVR, we also demonstrate *diversity modeling*, or the process of representing certain distributions of mix values called *demographics*. We finally show how diversity modeling can be used to improve the effectiveness of visual reasoning models.

# 2 Background

## 2.1 Complex Spatial Relationships

Formalized by Freeman in 1975, the term *spatial relations* denotes how objects in scenes are located relative to others [10]. The author gives a list of irreducible labels to describe these relationships. Examples include “left of”, “right of”, “above”, “below”, etc. The labels simply describe how a single object relates to another object. For the purpose of this study, relationships like these are referred to as *simple spatial relationships (SSRs)* to distinguish from *complex spatial relationships (CSRs)*, or combinations of these simple spatial relationships.

## 2.2 CLEVR

Visual question answering (VQA) is a visual reasoning problem in which a computer visual model answers questions associated with given images [11]. The VQA dataset CLEVR introduced by Johnson et al. is used as a baseline for many recent visual reasoning models [1]. This dataset was generated by a tool that creates 3D scenes and visual reasoning questions [12]. The authors of CLEVR also implemented visual reasoning models to test the dataset, with the best performing of these being a CNN+LSTM followed by an iteration of spatial attention (CNN+LSTM+SA) [13, 14].

To create a scene in the dataset, the generator places a number of objects represented by geometric shapes, each with random attributes, in random, non-intersecting positions on a 3D plane. Randomized object attributes include shape, color, material, and size. A sample scene is shown in Figure 1.

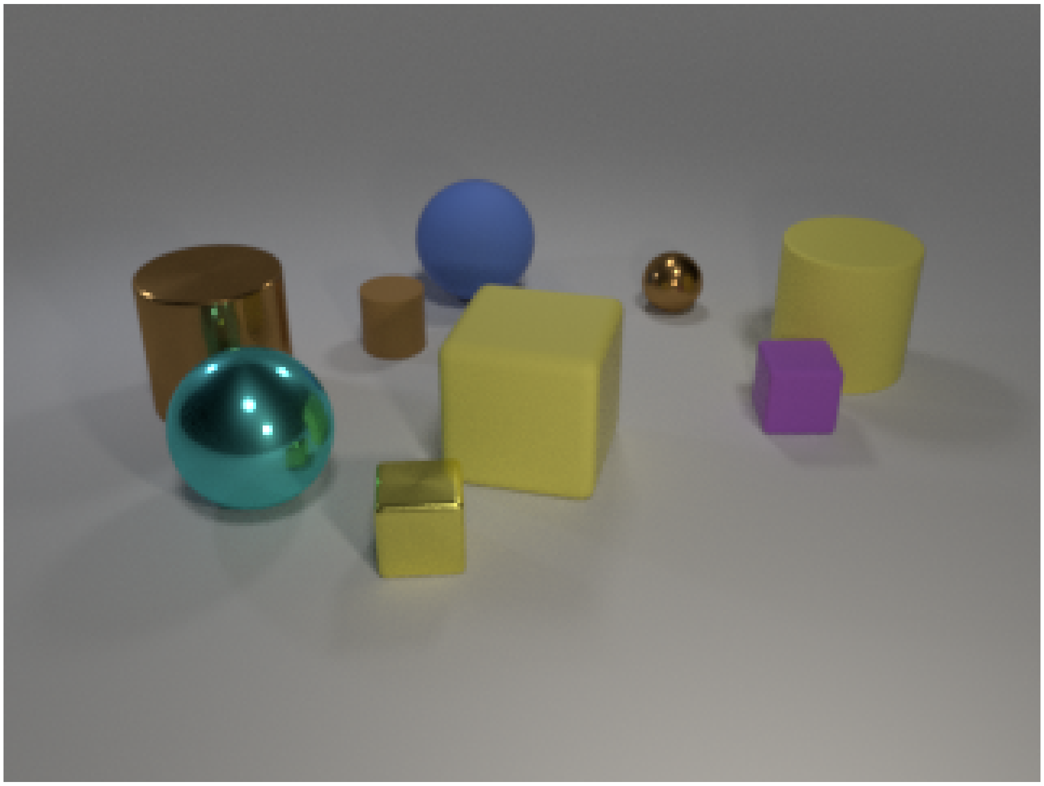


**Figure 1:** A sample scene created with the CLEVR generator tool.

The generator then creates questions and answers about the scene from a template based on these attributes and the relationships between the objects. It generates questions of the following types: simple queries, comparisons, existence queries, and counts (e.g. “How many metal things are there?”, “What shape is the object to the left of the blue sphere?”). All questions involving spatial relationships only refer to SSRs. The tool then converts the questions to machine-readable representations for model training.

## 2.3 Genetic Algorithms

The genetic algorithm (GA) is a heuristic approach designed to optimize a set of parameters to approximate a solution for a given problem [15]. It loosely simulates biological evolution through natural selection and genetic drift. A GA begins with an initial population of individuals, often called chromosomes, each containing information that corresponds to a set of parameters. The population is usually generated at random. The chromosomes are quantified with a *fitness value* according to how well they solve the problem. Constraints, usually related to the fitness values, are set to stop the GA from continually running. Processes such as genetic crossover and mutation are repeated on the population until a resulting chromosome satisfies the constraints.

A basic iteration of the GA is illustrated as follows: Pairs of *best fit* chromosomes, or those having the highest fitness, are selected to be parents of offspring. Each parent passes on a portion of its information in a process called genetic crossover. The resulting offspring replace the *least fit* chromosomes. Some chromosomes are then subjected to mutation, randomly changing a portion of their information. This all repeats until the constraints are reached.

Continually replacing the least fit chromosomes causes an overall loss of information in the population, and subsequently, a decrease in diversity. This can result in premature convergence to a local minimum. Mutation helps to mitigate this, but is often insufficient. To further preserve diversity, a modified GA called the island model [16] maintains multiple populations representing islands, and allows gene flow to occur between pairs of populations. This is done by migrating a random chromosome from each island to a neighboring island after each GA iteration.

# 3 Implementation

With the existing visual reasoning datasets, the positions of objects in scenes assume a uniform random distribution. This makes it difficult to sufficiently capture CSRs since they are not often expressed in these datasets. The approach presented in this paper demonstrates the capability to generate CSRs in the following way. First, the *aligned relationship* CSR is implemented that describes whether a set of objects is approximately collinear. Then a model is presented to reduce aligned relationships into a single scalar parameter called the *mix value*. Finally, the CLEVR generator tool is modified to create *demographics*, or datasets generated from distributions of mix values to improve visual reasoning model effectiveness.

## 3.1 Aligned Relationships

For this work, we introduce a specific CSR called the *aligned relationship* that describes some level of collinearity of a set of objects. We define a set of at least three objects as *aligned* if all objects are collinear within a specified threshold distance. That is, the set must contain at least three objects and must contain objects that are all within the threshold distance of a straight line, called the *benchmark line*.

To determine if all objects are within the threshold distance, connecting lines are formed between every pair of objects. From each connecting line, the perpendicular distance from every object not on the line must be less than the threshold distance. An example line is drawn connecting the two blue objects in Figure 2.

## 3.2 Model for Complex Spatial Relationship Representation

The CLEVR generator positions objects in a scene according to a uniform random distribution. The likelihood of aligned relationships emerging out of uniform randomness is very small. Ideally, the probability that a generated sample contains aligned relationships is controlled via an adjustable parameter, the *mix value* $x \in [0.0, 1.0]$. The mix value represents a stochastic parameterization of object layout with lower values leading to placement that tends less towards misalignment and higher values tend more towards misalignment. Thus, the mix value $x = 0.0$ (Figure 3a) represents the highest

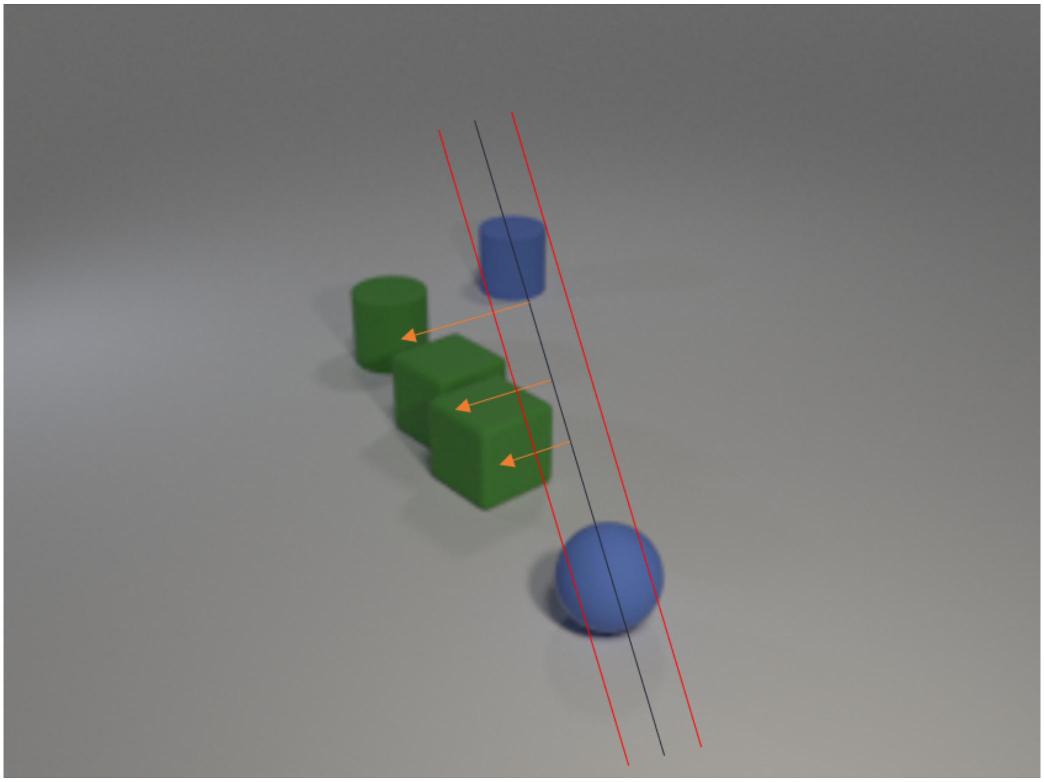
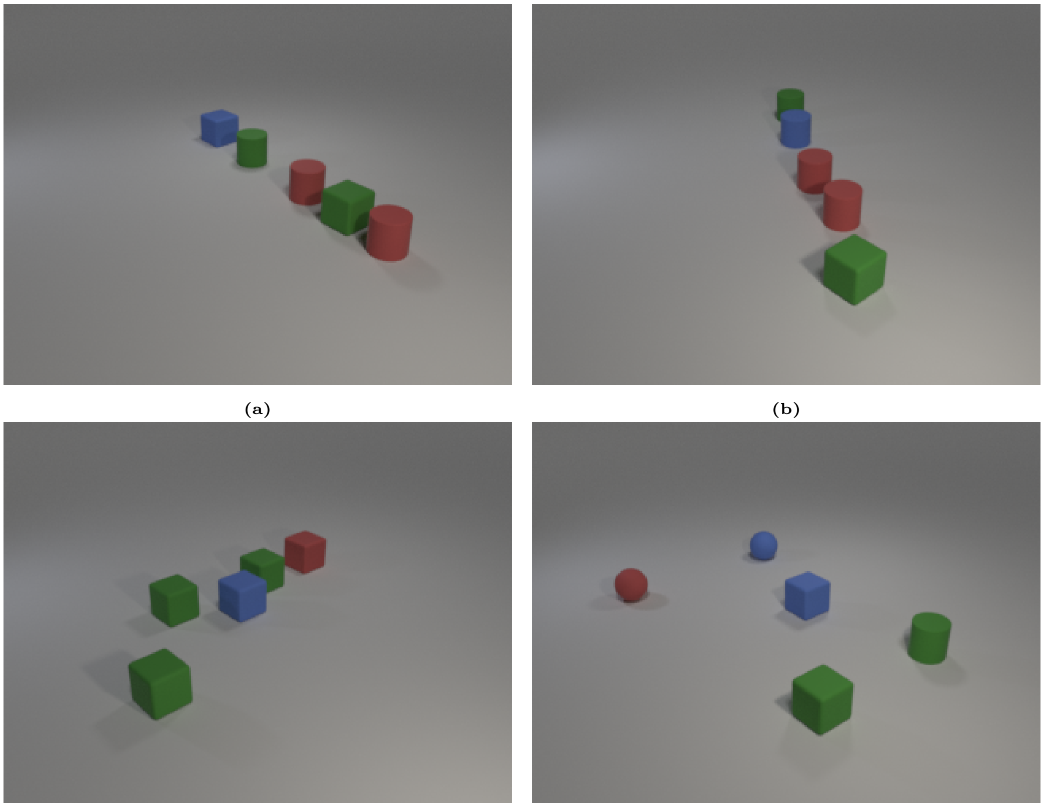


**Figure 2:** An example connecting line (black) joining the two blue objects, showing perpendicular lines (orange) and an example threshold distance (red). No objects fall within the threshold distance.

confidence that objects will be aligned, while $x = 1.0$ (Figure 3d) represents the highest confidence that objects will be not aligned. Subsequently, $x = 0.5$ (Figure 3b) represents the boundary case between aligned and not aligned, where small disturbances of object positions could change classifications. Scenes generated with other mix values, i.e. $x = 0.75$ shown in Figure 3c, are generally more or less aligned depending on whether $x$ is closer to 0.0 or 1.0, respectively.



**(a)** **(b)** **(c)** **(d)**

**Figure 3:** Scenes showing an aligned relationship generated with mix values $x = 0.0$ (a), $x = 0.5$ (b), $x = 0.75$ (c), and $x = 1.0$ (d).

To represent aligned relationship strength with a single scalar parameter, the model first creates a perfectly aligned scene called the *initial state*. In this scene, the centers of all objects lie on a random benchmark line. Each scene object is then translated by a random variable $Y$ called the *disturbance amount* in a direction perpendicular to the benchmark line. Examples of this are represented by the orange arrows in Figure 2. Disturbing each object by any amount increases the likelihood of the scene not being aligned. With $Y$ being large enough, the object positions should approximate a uniform random distribution, resulting in a scene that is almost always not aligned, called the *uniform random state*.

Notice that $x$ is directly proportional to the likelihood of scenes not possessing the alignment relationship. For example, with $x = 0.75$, a generated dataset should contain not aligned scenes 75% of the time and aligned for the remaining 25%. When $x = 0.0$, scenes from the dataset should resemble the initial state. Therefore, $Y$ should be minimal. Conversely, scenes generated from $x = 1.0$ should approximate the uniform random state, so $Y$ should be very large. The relationship between $x$ and $Y$ is modeled by the *parameterized normal distribution* $N$:

$$Y \sim N(0.0, v(x)), \tag{1}$$

where $Y$ is the disturbance amount and $N$ is a normal probability distribution with mean 0.0 and a parameterized variance $v(x)$. The variance $v(x)$ is a function of the mix value $x$. With a very small variance, i.e. near zero, generated scenes should be aligned, while a much larger variance should produce not aligned scenes. Thus, $x$ should be proportional to $v(x)$ and subsequently, the probability of scenes being not aligned. In our testing, the following nonlinear variance function performed best:

$$v(x) = p_0 x^{p_1} + p_2 x^{p_3}, \tag{2}$$

where $p_0, p_1, p_2, p_3$ are unknown constants called the *variance parameters*.

## 3.3 Genetic Algorithm for Placement Parameterization

The variance parameters in Equation 2 are selected such that a given $x$ approximates the probability of scenes being not aligned. Let there be a set $P = \{p_0, p_1, p_2, p_3\}$ that best approximates this probability for any $x$. Locating these values necessitates an optimization technique.

To find $P$, we used the island model genetic algorithm described in Section 2.3. This was run for 2,000 iterations over nine populations, each containing 400 chromosomes. There are a total of $9 * 400 = 3600$ chromosomes, therefore each one corresponds to a set of variance parameters $P_n$, where

$n \in [0, 3599]$. Since we know that the fraction of negative samples generated by the model should approximate a given mix value, the fitness of each chromosome can be quantified by this approximation. To do this, a set of mix values $x \in M$ is tested by generating a number of samples for each $x$ and comparing against the resulting fraction of negative samples $f$. It is computationally expensive to test a large number of mix values, so we limit the number tested. If we let the set $M = \{0.0, 0.01, 0.02, 0.03, ..., 1.0\}$, then the fitness corresponds to the mean accuracy for $f$ approximating all $x \in M$. To calculate this, we first define the mean relative error for $f$ approximating all $x \in M$ as

$$E = \frac{1}{101} \sum_{x \in M} \frac{|f - x|}{x}, \tag{3}$$

where $\frac{1}{101}$ is derived from the cardinality of $M$. It follows then that the accuracy, and thus the fitness, is

$$F = 1 - E = 1 - \frac{1}{101} \sum_{x \in M} \frac{|f - x|}{x}. \tag{4}$$

Selecting the highest fit chromosome gave us the set of variance parameters $P = \{0.273, 10.911, 0.531, 29.320\}$. Substituting in for Equations 1 and 2, a disturbance amount $Y$ sampled from the resulting parameterized normal distribution $N$ is given as:

$$Y \sim N(0.0, 0.273x^{10.911} + 0.531x^{29.320}), \tag{5}$$

with mean 0.0 and parameterized variance $v(x) = 0.273x^{10.911} + 0.531x^{29.320}$.

## 3.4 Data Diversity Model for Training Complex Spatial Relationships

Since a uniform random distribution of objects in a synthetic dataset does not adequately express aligned relationships, we show how certain distributions of mix values allows us to better train visual learning models through a process called *diversity modeling*. This involves selecting a *demographic*, or a dataset $D$ generated from a set of mix values $x \in D$, where each $x$ is sampled from a probability distribution. We hypothesize that for a particular visual learning model, an optimal training demographic exists. Some example demographics are presented here, being named according to how each $x$ is sampled. For example, each $x$ in the *uniform demographic*, denoted as $D_{uniform}$, is sampled from a uniform random distribution. Here, all mix values are nearly equally expressed. Similarly, the *normal demographic*, denoted as $D_{normal}$, is sampled from a normal distribution. The *binary demographic*, $D_{binary}$, models the strongest representations of aligned and not aligned by sampling near $x = 0.0$ and $x = 1.0$. A comparison between $D_{uniform}$ and $D_{normal}$ is illustrated in Figure 4.

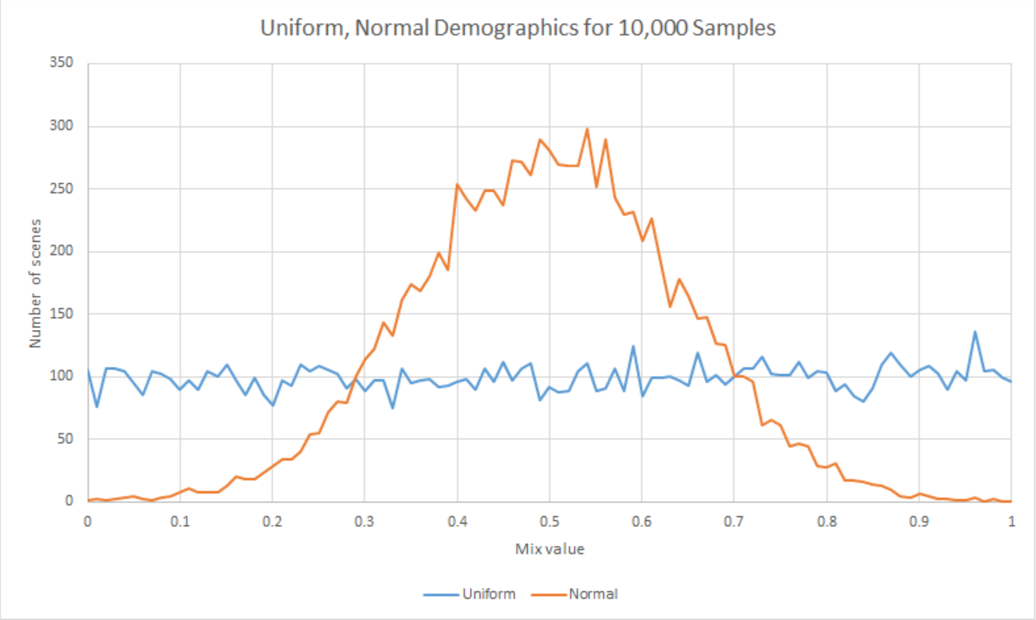


**Figure 4:** Example demographics of 10,000 scenes for uniform and normal.

# 4 Experimental Design

To test the diversity model, the CLEVR generator tool was updated to create aligned relationships using the model presented in Sections 3.2 and 3.3. The question generation template was changed to questions that only concern object alignment. Some restrictions were set to simplify the visual reasoning model scope: (1) five objects per scene, (2) all objects are small, (3) all objects are comprised of the rubber material, and (4) three color possibilities. Sample scenes generated from the modified tool are shown in Figures 2 and 3.

Datasets for the uniform, normal, and binary demographics described in Section 3.4 were created using the modified generator. For each demographic, 11 cumulative training datasets and one testing dataset were generated. The cumulative training datasets were created with an initial dataset of 4,000 samples, and then repeatedly concatenating 2,000 samples for each subsequent dataset, up to 24,000 samples. Each testing dataset contained 8,000 samples. Using the CNN+LSTM+SA visual reasoning model implemented by the CLEVR authors, we trained eleven models for each demographic, one on each of the cumulative training datasets [13]. Each model was trained for 75,000 epochs, and then evaluated on all testing data.

# 5 Results

Data from the experiment detailed in Section 4 are shown in Table 1. The accuracy for each model is listed by the model's training demographic (Column 1), the testing demographic (Column 2), and the number of training samples used for the model (Columns 3–13). For example, the model trained on 4000 normal demographic samples and tested on the normal demographic had an accuracy of 71%. Note that at sample sizes of 8000, 10000, and 18000, some accuracies are not shown. These models did not converge for a reason unknown at this time, so they were not included.

| Train Demo. | Test Demo. | Number of Samples | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 4000 | 6000 | 8000 | 10000 | 12000 | 14000 | 16000 | 18000 | 20000 | 22000 | 24000 |
| Normal | Normal | 0.7122 | 0.7444 | 0.7486 | 0.7668 | 0.7754 | 0.7826 | 0.8018 | 0.7979 | 0.8071 | 0.8087 | 0.8101 |
| Uniform | | 0.6677 | 0.7232 | | | 0.7594 | 0.7737 | 0.78 | 0.7808 | 0.7854 | 0.794 | 0.7953 |
| Binary | | 0.6904 | 0.7317 | 0.735 | 0.7297 | 0.7438 | 0.7383 | 0.7562 | | 0.753 | 0.7498 | 0.7391 |
| Normal | Uniform | 0.7625 | 0.7583 | 0.7832 | 0.8094 | 0.8152 | 0.8313 | 0.8344 | 0.8482 | 0.835 | 0.8507 | 0.8549 |
| Uniform | | 0.7168 | 0.7734 | | | 0.8091 | 0.8164 | 0.8294 | 0.8312 | 0.842 | 0.8444 | 0.8448 |
| Binary | | 0.7511 | 0.7845 | 0.7927 | 0.7983 | 0.8133 | 0.814 | 0.8261 | | 0.8209 | 0.8211 | 0.8218 |
| Normal | Binary | 0.7862 | 0.8133 | 0.8319 | 0.8406 | 0.8495 | 0.8636 | 0.8735 | 0.8589 | 0.8812 | 0.8863 | 0.8853 |
| Uniform | | 0.7639 | 0.8081 | | | 0.8443 | 0.8584 | 0.8606 | 0.8614 | 0.8839 | 0.8769 | 0.8885 |
| Binary | | 0.77 | 0.8307 | 0.8516 | 0.8563 | 0.8826 | 0.888 | 0.8827 | 0.9033 | 0.8978 | 0.8869 | |

**Table 1:** Percent of correct classification of aligned or not aligned objects for each demographic evalued on all demographics.

Of the models tested on the normal demographic, the normal models consistently resulted in the highest accuracies. In these experiments, the normal models tended to reach higher accuracies with less training data than the others. For many sample sizes, the normal models perform as well as others trained with more samples. When tested on the uniform demographic, most of the normal models performed best, except for those at sample sizes 6000, 8000, and 22000. When tested on the binary demographic, the binary models were the most effective in all comparisons but 4000 and 24000.

Intuitively, a model of a specific demographic would perform best on testing data of the same demographic. However, this is not necessarily the case as shown with the uniform testing demographic. This may be explained by the normal demographic having different proportions of samples near certain mix values. Compared to uniform, it has significantly smaller sample sizes near $x = 0.0$ and $x = 1.0$, while it has many more boundary cases near $x = 0.5$. This can be seen in Figure 4. In general, boundary cases are more difficult to classify since these CSRs can be more ambiguous. The normal models train with more boundary cases than the other demographics, suggesting that boundary cases may carry more weight for training.

In two cases with the uniform training demographic, the normal models do fall behind: sample sizes 6000 and 20000. These particular models are also exceptions to the overall positive trend in accuracy for the normal demographic. During training, they converged much slower than the other normal models. As with the non-converging models, the cause is unknown.

The binary testing demographic only shows extreme examples of aligned or not aligned cases, while the normal and uniform demographics have significantly less of these examples. The inclusion of boundary cases during training does not seem to be as beneficial for classifying this demographic.

# 6 Future Work

The models presented in this paper demonstrate that some demographics may better train visual learning models. The next step is compare these demographics against a baseline dataset, or one that has uniform random object positions. We also need to identify the source of the convergence issue. Furthermore, we will run new tests with these models to determine whether our claims are supported when applied to real-world situations.

# References

[1] **Johnson, Justin et al.** "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning". In: 2017, pp. 1988–1997.

[2] **Yi, Kexin et al.** "CLEVRER: CoLlision Events for Video REpresentation and Reasoning". In: *CoRR* abs/1910.01442 (2019). arXiv: 1910.01442. URL: http://arxiv.org/abs/1910.01442.

[3] **Santoro, Adam et al.** "A simple neural network module for relational reasoning". In: *CoRR* abs/1706.01427 (2017). arXiv: 1706.01427. URL: http://arxiv.org/abs/1706.01427.

[4] **Sampat, Shailaja Keyur et al.** "CLEVR_HYP: A Challenge Dataset and Baselines for Visual Question Answering with Hypothetical Actions over Images". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021, pp. 3692–3709. URL: https://aclanthology.org/2021.naacl-main.289.

[5] **Kottur, Satwik et al.** "CLEVR-Dialog: A Diagnostic Dataset for Multi-Round Reasoning in Visual Dialog". In: *CoRR* abs/1903.03166 (2019). arXiv: 1903.03166. URL: http://arxiv.org/abs/1903.03166.

[6] **Arras, Leila, Ahmed Osman, and Wojciech Samek**. "CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations". In: *Information Fusion* 81 (2022), pp. 14–40. ISSN: 1566-2535. URL: https://www.sciencedirect.com/science/article/pii/S1566253521002335.

[7] **Liu, Runtao et al.** "CLEVR-Ref+: Diagnosing Visual Reasoning with Referring Expressions". In: *CoRR* abs/1901.00850 (2019). arXiv: 1901.00850. URL: http://arxiv.org/abs/1901.00850.

[8] **Bahdanau, Dzmitry et al.** "CLOSURE: Assessing Systematic Generalization of CLEVR Models". In: *CoRR* abs/1912.05783 (2019). arXiv: 1912.05783. URL: http://arxiv.org/abs/1912.05783.

[9] **Beckham, Christopher et al.** *Visual Question Answering From Another Perspective: {CLEVR} Mental Rotation Tests*. 2021. URL: https://openreview.net/forum?id=aYbCpFNnHdh.

[10] **Freeman, John**. "The modelling of spatial relations". In: *Computer Graphics and Image Processing* 4.2 (1975), pp. 156–171. ISSN: 0146-664X. URL: https://www.sciencedirect.com/science/article/pii/S0146664X75800074.

[11] **Kafle, Kushal and Christopher Kanan**. "Visual question answering: Datasets, algorithms, and future challenges". In: *Computer Vision and Image Understanding* 163 (2017), pp. 3–20. ISSN: 1077-3142. URL: https://www.sciencedirect.com/science/article/pii/S1077314217301170.

[12] **Johnson, Justin et al.** "CLEVR Dataset Generation". In: 2017. URL: https://github.com/facebookresearch/clevr-dataset-gen.

[13] **Johnson, Justin et al.** "Inferring and Executing Programs for Visual Reasoning". In: *ICCV*. Vol. abs/1705.03633. 2017. URL: http://arxiv.org/abs/1705.03633.

[14] **Johnson, Justin et al.** "inferring-and-executing". In: *ICCV*. 2017. URL: https://github.com/facebookresearch/clevr-iep.

[15] **Holland, John**. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, 1975.

[16] **Belding, Theodore C.** "The Distributed Genetic Algorithm Revisited". In: *International Conference on Genetic Algorithms*. 1995. URL: https://arxiv.org/pdf/adap-org/9504007.pdf.