Deep Neural Networks Can Learn Generalizable Same-Different Visual Relations

Alexa R. Tartaglini*^{1,†} Sheridan Feucht*^{2,3,‡} Michael A. Lepori² Wai Keen Vong¹

50

51

52 53

54

74

75

Brenden M. Lake¹ Ellie Pavlick² Charles Lovering²

> ²Brown University ³Northeastern University

Abstract

1

25

¹New York University

Although deep neural networks can achieve human-level 46 2 performance on many object recognition benchmarks, 3 prior work suggests that these same models fail to learn 4 simple abstract relations, such as determining whether 5 two objects are the same or different. Much of this 6 prior work focuses on training convolutional neural net-7 works to classify images of two same or two differ-8 ent abstract shapes, testing generalization on within-9 distribution stimuli. In this article, we comprehensively 10 study whether deep neural networks can acquire and gen-11 eralize same-different relations both within and out-of-12 distribution using a variety of architectures, forms of pre-13 training, and fine-tuning datasets. We find that certain 14 pretrained transformers can learn a same-different rela-15 tion that generalizes with near perfect accuracy to out-of-16 distribution stimuli. Furthermore, we find that fine-tuning 17 on abstract shapes that lack texture or color provides the 18 strongest out-of-distribution generalization. Our results 47 19 suggest that, with the right approach, deep neural net-48 20 works can learn generalizable same-different visual rela-49 21 tions. 22

Keywords: vision models: abstract relations: convolutional 23 neural networks: transformer models 24

Introduction

Humans and a wide variety of non-human animals can easily 55 26 recognize whether two objects are the same as each other 56 27 or whether they are different (see Figure 1; Martinho III & 57 28 Kacelnik, 2016; Christie, 2021; Gentner et al., 2021; Hespos 58 29 30 et al., 2021). The abstract concept of equality is simple—even 59 3-month-old infants (Anderson et al., 2018) and honeybees 60 31 (Giurfa, 2021) can learn to distinguish between displays of 61 32 two same or two different objects. Some researchers have 62 33 even argued that it serves amongst a number of other basic 63 34 logical operations as a foundation for higher-order cognition 64 35 and reasoning (Gentner & Goldin-Meadow, 2003; Gentner & 65 36 Hoyos, 2017). However, in contrast to humans and animals, 66 37 recent work has argued that deep neural networks struggle to 67 38 learn this simple relation (Ellis et al., 2015; Gülçehre & Ben-68 39 gio, 2016; Stabinger et al., 2016; Kim et al., 2018; Webb et 69 40 al., 2020; Puebla & Bowers, 2022). This difficulty is surprising 70 41 given that deep neural networks achieve human or superhu-71 42 man performance on a wide range of seemingly more com-72 43 plex visual tasks, such as image classification (Krizhevsky et 73 44

*Equal contribution.

[†]Correspondence to art481@nyu.edu.

al., 2012; He et al., 2016), segmentation (Long et al., 2015), and generation (Ramesh et al., 2022).



Figure 1: Same or different? For humans and a number of animal species, it is trivial to recognize that the image on the left contains two of the same objects, while the image on the right contains two different objects. Surprisingly, prior research has suggested that deep neural networks struggle to learn to discriminate between these images.

Past attempts to evaluate same-different relations in neural networks have generally used the following methodology. Models are trained to classify images containing either two of the same or two different abstract objects, such as those in Figure 1. A model is considered successful if it is then able to generalize the same-different relation to unseen shapes after training. Convolutional neural networks (CNNs) trained from scratch fail to learn a generalizable relation, and tend to memorize training examples (Kim et al., 2018; Webb et al., 2020). However, deep neural networks have been shown to successfully generalize the same-different relation in certain contexts. This generalization is either limited to in-domain test stimuli (Funke et al., 2021; Puebla & Bowers, 2022) or requires architectural modifications that build in an inductive bias towards relational tasks at the expense of other visual tasks (Kim et al., 2018; Webb et al., 2020; Webb, Frankland, et al., 2023; Webb, Mondal, & Cohen, 2023; Kerg et al., 2022; Geiger et al., 2023; Altabaa et al., 2023). Given these limited successes, an open question remains: without architectural modifications that restrict model expressivity in general, can standard neural networks learn an abstract same-different relation that generalizes to both in- and out-of-distribution stimuli?

Addressing this question requires going beyond past work in a number of ways. First, most previous studies test for indistribution generalization-that is, they use test stimuli that are visually similar to the training stimuli. We believe that out-of-distribution generalization provides much stronger evidence that a model has learned a genuine abstract relation without relying on spurious features. Second, the existing literature uses training stimuli that demonstrate the same-different relation with either closed curves (as in Figure 1) or simple

[‡]This work was done while the author was an undergraduate at ⁷⁶ Brown University. 77



Figure 2: **Example stimuli from all four datasets.** Each column represents one of the four same-versus-different datasets as indicated by the label beneath the stimuli. The top row shows an example object that is used to form the stimuli that comprise each dataset, while the second and third rows show an example "same" vs. "different" stimulus, respectively.

78 geometric shapes. It is unclear whether training on these 110

- 79 types of objects is the most helpful for learning the relation111
- 80 versus more naturalistic objects that more closely resemble112
- at a seen during pretraining. Finally, most prior work focuses
- on convolutional architectures, but Vision Transformers (ViTs)¹¹³
- 83 (Dosovitskiy et al., 2020) adapted from the language domain¹¹⁴
- ⁸⁴ (Vaswani et al., 2017) have recently emerged as a competi-¹¹⁵
- tive alternative to CNNs on visual tasks. Self-attention, a key¹¹⁶
- ⁸⁶ feature of ViTs, may provide an advantage when learning ab-¹¹⁷
- 87 stract visual relations-indeed, the ability to attend to and re-

late any part of a stimulus to any other part may be crucial for
 relational abstraction.

In this article, we address these limitations and comprehen sively investigate how neural networks learn and generalize
 the same-different relation from image data. Our main find ings are as follows:

- Fine-tuning pretrained ResNet and ViT models on the¹²³ same-different relation enables both architectures to gener-¹²⁴ alize the relation to unseen objects in the same distribution¹²⁵ as the fine-tuning set. In particular, CLIP pretraining results¹²⁶ in nearly 100% in-distribution test accuracy for ViT models,¹²⁷ and close to that for ResNet models.
- Under certain conditions, CLIP-pretrained ViTs can reliably $_{_{130}}$ 129 100 generalize the same-different relation to out-of-distribution,131 101 stimuli with nearly 100% accuracy. Furthermore, these 102 models can transfer the relation with up to 90% test accu-103 racy to a photorealistic same-different dataset of 3D objects 134 104 without any fine-tuning on the 3D setting. These results 135 105 suggest that these models acquire a generalizable $abstract_{136}$ 106 concept of equality. 107 137
- Different fine-tuning datasets lead to qualitatively differentiane patterns of generalization—fine-tuning on more visually ab-139

stract objects (which do not contain color or texture) results in stronger out-of-distribution generalization, whereas finetuning on more naturalistic objects fails to generalize.

 ViTs generally prefer to determine equality between objects by comparing their color or texture, only learning to compare shape when the fine-tuning dataset lacks color and texture information. However, we find that CLIP pretraining helps to mitigate this preference for color and texture.

Methods

We operationalize the same-different task consistently with prior work, e.g. Fleuret et al. (2011). Models are asked to perform a binary classification task on images containing either two of the same objects or two different objects (see the second and third rows of Figure 2). Models are either trained from scratch or fine-tuned on a version of this task with a particular type of stimuli (see Training and Evaluation Datasets below). After training or fine-tuning, model weights are frozen, and validation and test accuracy scores are computed on sets of same-versus-different stimuli containing unfamiliar objects. These can be either be the same type of objects that they were trained or fine-tuned on (in-distribution generalization) or different types of objects (out-of-distribution generalization). Thus, in order to attain high validation and test accuracy scores, the model must successfully generalize the learned same-different relation to novel objects. This type of generalization is more challenging than the standard image classification setting because of the abstract nature of what defines the classes-models must learn to attend to the relationship between two objects rather than learn to attend to any particular visual features of those objects in the training data.

Training and Evaluation Datasets We construct four¹⁸³
 same-versus-different datasets using four different types of¹⁸⁴
 objects (see Figure 2) ranging from abstract shapes to nat⁻¹⁸⁵
 uralistic images that are more familiar to pretrained models.¹⁸⁶
 We use the following objects to create these four datasets: ¹⁸⁷

Squiggles (SQU). Randomly generated closed shapes fol¹⁸⁸/₁₈₉
 lowing Fleuret et al. (2011).¹ Most studies in the machine₁₉₀
 learning literature on the same-different relation uses this₁₉₁
 dataset (Kim et al., 2018; Funke et al., 2021; Puebla & Bow-₁₉₂
 ers, 2022; Messina et al., 2022).

Alphanumeric (ALPH). Sampled handwritten characters¹⁹⁴
 from the Omniglot dataset (Lake et al., 2015).

152 3. Shapes (SHA). Textured and colored shapes from¹⁹⁶
 153 Tartaglini et al. (2022). Objects that match in shape, tex-¹⁹⁷
 154 ture, and color are considered the same, while objects that¹⁹⁸
 155 differ along all three dimensions are considered different. ¹⁹⁹

4. Naturalistic (NAT). Photographs of real objects on white²⁰⁰
 backgrounds from Brady et al. (2008). These stimuli are²⁰²
 the most similar to the data that the pretrained models see²⁰³
 before fine-tuning on the same-different task.

Each stimulus is an image that contains two objects ${\rm that}^{\rm 205}$ 160 are either the same or different. We select a total of $1.600^{\rm 206}$ 161 unique objects for each dataset. These objects are split into²⁰⁷ 162 disjoint sets of 1,200, 300, and 100 to form the training, val-208 163 idation, and test sets respectively. Unless otherwise speci-209 164 fied, the training, validation, and test sets each contain $6{,}400^{210}$ 165 stimuli: 3,200 same and 3,200 different. To construct a given²¹¹ 166 dataset, we first generate all possible pairs of same or different²¹² 167 objects-we consider two objects to be the same if they are²¹³ 168 the same on a pixel level.² Next, we randomly select a subset²¹⁴ 169 of the possible object pairs to create the stimuli such that each²¹⁵ 170 unique object is in at least one pair. Each object is resized to²¹⁶ 171 64x64 pixels, and then a pair of these objects is placed over²¹⁷ 172 a 224x224 pixel white background in randomly selected, non-173 overlapping positions. We consider two objects in a specific²¹⁸ 174 placement as one unique stimulus-in other words, a giveners 175 pair of objects may appear in multiple images but in different 176 positions (but with all placements of the same two objects be-177 ing confined to either the training, validation, or test set). All $_{222}^{42}$ 178 object pairs appear the same number of times to ensure that $\frac{1}{223}$ 179 each unique object is equally represented. 180

Models and Training Details We evaluate one convolu-225
 tional architecture, **ResNet-50** (He et al., 2016), and one226

Transformer architecture, ViT-B/16 (Dosovitskiy et al., 2020). We also evaluate three pretraining procedures: (1) randomly initialized, in which all model parameters are randomly initialized (Kaiming normal for ResNet-50 and truncated normal for ViT-B/16) and models are trained from scratch, (2) ImageNet, in which models are pretrained in a supervised fashion on a large corpus of images (ImageNet-1k for ResNet-50 and ImageNet-21k for ViT-B/16; Deng et al., 2009) with category labels such as "barn owl" or "airplane." and (3) CLIP (Radford et al., 2021), in which models learn an image-text contrastive objective where the cosine similarity between an image embedding and its matching natural language caption embedding is maximized. Unlike ImageNet labels, CLIP captions contain additional linguistic information beyond category information (e.g. "a photo of a barn owl in flight"). To address the difference in parameter count between ResNet-50 and ViT-B/16 (23M versus 86M parameters), we also provide results for ImageNet-pretrained ConvNeXt-B Liu et al. (2022) and DeiT-S Touvron et al. (2021) in Appendix A.4 (89M and 22M parameters respectively).

We adapt all models to the same-different task by appending a linear classifier to the output of the visual backbone. For models not trained from scratch, we directly fine-tune on training sets from Training and Evaluation Datasets. Each model is trained from scratch or fine-tuned for 70 epochs with a batch size of 128, updating all parameters. We use a binary crossentropy loss. For each architecture and pretraining combination, we perform hyperparameter tuning via grid search over the initial learning rate (1e-4, 1e-5, 1e-6, 1e-7, 1e-8), learning rate scheduler (exponential, ReduceLROnPlateau), and optimizer (SGD, Adam, AdamW). We select the best performing training configuration from the grid search according to in-distribution validation accuracy, and then train a model with those hyperparameters five times with different random seeds. We report the median test results across those five seeds.

Generalization to Unseen Objects

In-Distribution Generalization

224

We first measure the performance of each model on test data containing the same types of objects used to train or finetune the model; e.g. models fine-tuned on pairs of handwritten characters are then tested on handwritten characters that were not seen during training. We refer to this as the indistribution performance of the model. The starred (*) result in Figure 3 shows the in-distribution median test accuracy of randomly-initialized ResNet-50 models trained on the Squiggles dataset, which contains the same type of closed contours used by much of the prior work on the same-different relation (Fleuret et al., 2011; Kim et al., 2018; Funke et al., 2021; Puebla & Bowers, 2022; Messina et al., 2022). Confirming the primary findings from prior work, these models do not attain above chance level test accuracy. The same pattern holds for randomly initialized ViT-B/16 models.

However, as the rest of Figure 3 shows, pretrained models exhibit substantially improved in-distribution accuracy com-

²²⁷ ¹The original method from Fleuret et al. (2011) produces closed ²²⁸ contours with lines that are only one pixel thick. For our chosen image and object size, these shapes become very difficult to see. We²²⁹ correct this by using a dilation algorithm to darken and thicken the₂₃₀ lines to a width of three pixels.

²There is some ambiguity in how to define sameness. One could²³¹ imagine a same-different task in which two objects drawn from the²³² same category are considered the same, such as two different im-²³³ ages of the same species of parrot. Furthermore, two objects can be₂₃₄ the same in some dimensions but differ in others (see Dissociating Color, Texture, and Shape). Unless otherwise stated, we take "same"²³⁵ to mean "exactly the same."²³⁶



Figure 3: In-distribution test accuracy by architecture and pretraining method. Bars show median accuracy over 5 runs, with the bar color denoting pretraining type and the x-axis denoting the dataset used for fine-tuning. See Methods for dataset descriptions and model details, and Figure 2 for visual examples. The **starred** (*) result is a replication of findings from prior work showing that CNNs trained from scratch on stimuli like the images in Figure 1 attain chance-level test accuracy. The **double-starred** (**) result mirrors Funke et al. (2021) and Puebla & Bowers (2022), who show that ImageNet-pretrained CNNs attain substantially higher in-distribution test accuracy relative to the same architectures trained from scratch.

pared to randomly initialized models across all four datasets 263 237 In particular, models pretrained with CLIP demonstrate the264 238 largest improvements, attaining nearly 100% test accuracy 239 irrespective of fine-tuning dataset. Even without any fine 265 240 tuning, CLIP features appear to be highly useful for the 266 241 same-different task-linear probes trained to do the same-267 242 different task using CLIP ViT-B/16 embeddings of stimuli with-268 243 out any fine-tuning achieve between 80% and 100% median₂₆₉ 244 in-distribution test accuracy depending on the dataset (Table₂₇₀ 245 12. Appendix A.7). Differences in performance can also be₂₇₁ 246 observed between architectures, with ViT-B/16 models con-272 247 sistently outperforming ResNet-50 after pretraining.³ These₂₇₃ 248 differences are likely not a result of a difference in parameter₂₇₄ 249 counts, as a similar gap in performance can be observed be-275 250 tween ImageNet ConvNeXt-B and ImageNet ViT-B/16, despite276 251 ConvNeXt-B being slightly larger (Appendix A.4). 252 277

Another main finding is that the two visually abstract₂₇₈ 253 shape-based datasets (SQU and ALPH) appear to pose more₂₇₀ 254 of a challenge to models than the SHA and NAT datasets-280 255 models attain noticeably higher in-distribution accuracy on the281 256 latter two across architectures and pretraining methods (al-282 257 though the effect is small for CLIP-pretrained models). This₂₈₃ 258 difference may be due to the color and texture information₂₈₄ 259 that is available in these datasets, which provides additional 260 dimensions over which objects can be compared. We ex-261 plore the possibility that some models find it easier to eval-262

uate equality using color or texture in addition to or instead of shape information in Examination of Inductive Biases.

Out-of-Distribution Generalization

The previous section showed that pretrained models can generalize to unseen, in-distribution objects. However, if a model learns a truly abstract notion of same-different, it should be able to generalize the same-different relation to any two objects regardless of their particular visual features. Thus, model performance on stimuli that are substantially different from training stimuli is a stronger measure of abstraction. We therefore measure test accuracy for each model across all four datasets, yielding one in-distribution score and three out-ofdistribution (OOD) scores per model. Table 1 shows median test accuracy over five seeds for CLIP-pretrained models; full generalization tables for all pretraining styles and architectures can be found in Appendix A.3.

Overall, CLIP ViT-B/16 models fine-tuned on the Squiggles task exhibit the strongest OOD generalization, achieving >95% median test accuracy on the three out-of-distribution datasets.⁴ As in the previous section, models fine-tuned on objects with visually abstract shape features only (SQU and ALPH) behave differently than those fine-tuned on datasets

³ViTs also demonstrate qualitatively different training dynamics compared to CNNs, appearing to generalize the same-different relation within the first few epochs of training. Furthermore, ViTs learn more *smoothly* than ResNets. See Appendix A.2 for figures of training and accuracy curves.

 $^{^4}$ It is worth noting that both this model and CLIP ResNet-50 finetuned on the ALPH task (the model with the second best OOD generalization performance) exhibit some degree of sensitivity to the random seed used during fine-tuning: most random seeds result in nearly 100% OOD generalization for ViT or >80% for ResNet across all datasets, while some seeds result in substantially lower performance (1/5 seeds for ViT and 2/5 for ResNet). No other model configurations exhibit this bimodal behavior. See Appendix A.10 for details.

Table 1: **Out-of-distribution (OOD) test accuracy for CLIP models fine-tuned on each dataset.** Rows indicate the dataset that models are fine-tuned on, while columns indicate the test dataset. Each cell is the median performance over five random seeds. The rightmost column labeled "Avg." is the row-wise average of accuracy scores across OOD test sets (i.e. off-diagonal values), which indicates how well a model fine-tuned on a given dataset is able to generalize to other datasets. The bottom row labeled "Avg." is the column-wise average across off-diagonal values, indicating how difficult it is for models fine-tuned on other datasets to generalize to the given dataset. Note that the **bolded** diagonals are the pink bars in Figure 3. OOD generalization results for all models are in Appendix A.3; Appendix A.5 shows median AUC-ROC scores.

		CLIP Res	Net-50			CLIP VIT-B/16					
		\leftarrow Te	est $ ightarrow$			$\leftarrow Test \rightarrow$					
Train \downarrow	SQU	ALPH	SHA	NAT	Avg.	Train \downarrow	SQU	ALPH	SHA	NAT	Avg.
SQU	97.7	82.9	86.9	82.0	83.9	SQU	99.6	97.7	99.1	96.7	97.8
ALPH	82.1	97.4	92.8	91.8	88.9	ALPH	55.3	99.4	99.6	91.2	82.0
SHA	56.0	78.1	98.1	96.1	76.7	SHA	50.0	55.4	100	100	68.5
NAT	50.1	59.3	93.4	97.3	67.6	NAT	50.0	68.0	99.8	100	72.6
Avg.	62.7	73.4	91.1	90.0		Avg.	51.8	73.7	99.5	95.9	

containing objects with shape, color, and texture features324 285 (SHA and NAT). The SQU and ALPH models generally at-325 286 tain high OOD test accuracy. On the other hand, models fine-326 287 tuned on the SHA or NAT datasets generalize well to each327 288 other but struggle to generalize to the SQU and ALPH tasks 328 289 Note that some of this effect can be attributed to miscalibrated 329 290 bias, but not the entire effect—see Appendix A.5 for details. 330 291 Another way to understand the generalization pattern in Ta-331 292 ble 1 is that the more "challenging" a dataset is to generalize332 293

the same-different relation to, the more effective it is as a fine-333 294 tuning dataset for inducing out-of-distribution generalization,334 295 For example, CLIP ViT-B/16 models fine-tuned on datasets335 296 other than Squiggles attain a median test accuracy of only336 297 51.8% on the Squiggles task on average, whereas CLIP ViT-337 298 B/16 fine-tuned on Squiggles attains an average OOD tests38 299 300 accuracy of 97.8%. On the other hand, the Shapes datasets is easy for models fine-tuned on other datasets to generalize340 301 to (99.5% accuracy on average), but CLIP ViT fine-tuned on341 302 that "easier" dataset attains an average OOD test accuracy of 342 303 only 68.5%. This pattern of Squiggles being more "difficult"343 304 to generalize to persists across architectures and pretraining344 305 methods (Appendix A.3). 345 306

Out-of-Distribution Generalization to Photorealistic Stimuli 346 347 348

In the previous section, we demonstrated that CLIP-pretrained₃₄₉ 309 models fine-tuned on one set of same-different stimuli canaso 310 consistently generalize the relation to other visually distinct₃₅₁ 311 sets of stimuli. Despite these successes, one potential criti-352 312 cism is that the types of artificial stimuli we use-albeit having 313 proved significantly challenging in previous attempts to solve354 314 the same-different task-lack the additional complexities in-355 315 volved in recognizing abstract visual relations in real-worlds56 316 environments. In particular, following prior work, the objects357 317 in our "same" stimuli are the same as each other at the pixebs 318 level. It is possible that models in this conventional setting359 319 learn to generalize the same-different relation by simply rec-360 320 ognizing whether small patches of pixels or even single pix-321 els have the same values rather than comparing whole ob-322 jects. Relying on pixel-level mechanisms to adjudicate be-323

tween same and different would fail in photorealistic settings, since two instances of the same 3D object can differ at the pixel level due to differences in lighting, rotation, and depth of field. Given the successful OOD generalization of our finetuned models in an artificial setting, we wanted to test whether our findings extend to more challenging real-world settings.

To this end, we evaluate the fine-tuned models from previous sections on a dataset of 1,024 photorealistic samedifferent stimuli (see Figure 4). Each stimulus is a 224x224 pixel image depicting a pair of same or different 3D objects arranged on the surface of a table. We created these images in Blender, a sophisticated 3D modeling tool, using a set of 16 unique 3D models of different objects that vary in shape, texture and color. To construct the dataset, we first generate all possible pairs of same or different objects, then select a subset of the possible "different" pairs such that each object appears in two pairs. This ensures that all objects are equally represented and that an equal number of "same" and "different" stimuli are created. We create 32 unique stimuli for each pair of objects by placing them on the table in eight random configurations within the view of four different camera angles, allowing partial occlusions. Each individual object is also randomly rotated around its Z-axis in each image-because 11 of the objects lack rotational symmetry, these rotations provide an additional challenge, especially for "same" classifications.⁵

We evaluate the SQU, ALPH, SHA, and NAT fine-tuned models from the previous sections on the photorealistic dataset *without any additional fine-tuning*. We compute median test accuracy on the photorealistic dataset for CLIPpretrained models across the same five random seeds reported in previous sections (see Figure 22). Surprisingly, CLIP-pretrained ViT models generalize with 80-90% median test accuracy to the photorealistic stimuli despite only receiving fine-tuning on pixel-level sameness between 2D objects, indicating that their robust generalization of the same-different relation is not limited to our particular definition of the samedifferent task. On the other hand, all other pretraining and

⁵We also test models on a version of the photorealistic dataset where "same" objects are always rotated identically. We find that performance for most models improves slightly; see Appendix A.11.



Figure 4: **Examples of "same" and "different" photorealistic stimuli.** The textures of the table surface and background wall are randomly selected from a set of four options each. No two objects in an image are the same on a pixel level. See Appendix A.11 for images of all 16 3D objects.

406

architecture combinations including CLIP-pretrained ResNets
 fail to generalize consistently to the photorealistic stimuli (see
 Appendix A.11). These results suggest that, with careful
 choices of architecture and pretraining, fine-tuning on simplis tic 2D stimuli may be sufficient for learning an abstract same different relation that generalizes to 3D objects despite the ad ditional visual complexities of real-world settings.

368 Examination of Inductive Biases

What features do models use to decide whether two objects 369 in an image are the same? Since we train models without 370 explicit guidance for how to solve the task, any inductive bias 371 a given model may have likely influences how it learns the 372 same-different relation. Previous work has claimed that CNNs 373 trained on ImageNet are often biased towards texture over 374 shape (Geirhos et al., 2019; Hermann et al., 2020). This 375 may be related to results from Kim et al. (2018) that show 376 poor performance for CNNs trained from scratch on texture-377 less shapes. In this section, we investigate whether and how 378 these inductive biases influence model behavior for the same-379 different task. 380

Grayscale and Graymasked Objects We train models on 381 one of three variants of the Shapes dataset: objects are ei-382 ther kept the same (Figure 5a, "Color"), grayscaled to pre-383 serve texture but remove color (Figure 5a, "Grayscale"), or 384 completely covered in gray to remove both texture and color 385 (Figure 5a, "Masked"). If a model is biased towards color, per-400 386 formance should drop on the Grayscale and Masked datasets:401 387 if it is biased towards texture, performance should suffer on the402 388 Masked dataset. Only a model that is biased towards shape₄₀₃ 389 would generalize effectively to all three settings. We train or₄₀₄ 390 fine-tune on each of these three variants for randomly initial-405 391

Figure 6A shows the test performance of randomly initial-407 ized ViT-B/16 trained on either Color, Grayscale, or Masked408 versions of the Shapes dataset (Figure 5a) and tested on409 novel objects from each of those distributions. Figure 6A410 shows that ViT-B/16 trained from scratch only achieves411 high in-distribution accuracy for the Color Shapes dataset412 (92.9%); the hatched gray and dark gray bars representing413

ized, ImageNet-pretrained, and CLIP-pretrained models.

392



Figure 5: **Examples of stimuli used to test inductive biases.** Figure (a) shows examples of objects from the three versions of the Shapes dataset used to produce results in Figure 6. Figures (b) and (c) are examples of stimuli with conflicting signals used in Dissociating Color, Texture, and Shape, where either color is the same while texture and shape differ, or color and texture are the same while shape differs.

in-distribution accuracy for Grayscale and Masked Shapes are much lower (78.8% and 53.5% respectively). Despite this high in-distribution accuracy on Color Shapes, performance drops to 66.2% and 62.6% when generalizing out-ofdistribution to Grayscale and Masked Shapes, as indicated by the two lower gray bars beside the hatched green bar. This gap suggests that ViT-B/16 only learns to compare object *color* when it is trained from scratch on the Color Shapes dataset, leading to greater errors when tested on datasets that do not contain color. Figure 6A also shows that fine-tuning on Masked Shapes allows for out-of-distribution generalization that is strong relative to in-distribution generalization, suggesting that the model learns a more generalizable shape bias in this case. Figure 6B shows that CLIP pretraining weakens



Figure 6: Test accuracy for models trained or fine-tuned on one version of the Shapes dataset (Color, Grayscale, Masked) and then tested on all three versions of the dataset. Example stimuli are shown in Figure 5a. Hatched bars indicate indistribution accuracy. Median results for the same hyperparameters trained for five different seeds are reported, with individual runs also plotted as translucent points.

ViT's bias towards color, allowing for high in-distribution accu-437 414 racy and near-perfect out-of-distribution generalization when 438 415 trained on any of the three modified Shapes datasets. 439 416 ResNet-50 does not demonstrate an inductive bias towards440 417 color or texture when the model is trained from scratch. How-441 418 ever, pretraining results in a slight bias, with a 7.3% gap₄₄₂ 419 between in-distribution and Masked OOD accuracy for CLIP443 420 ResNet-50 fine-tuned on Color Shapes (Figure 6D). 444 421

445 Dissociating Color, Texture, and Shape Results from Fig-422 ure 6 suggest that some models learn to rely on certain fea-423 tures more than others to differentiate between objects in an447 424 image. To delve deeper, we create eight test datasets based448 425 on the Shapes dataset that vary whether shape, color, and449 426 texture are the same or different between two objects in an450 427 image (examples in Appendix B.2, Figure 23). We label each451 428 set of images with letters that represent whether color (C),452 429 texture (T), or shape (S) are the same. For example, images453 430 that contain two objects with the same color, different textures,454 431 and the same shape are labeled CS. CTS and "none" repre-455 432 sent the objects being completely the same or completely dif-456 433 ferent, respectively. We then evaluate the same models from 457 434 Figure 6 on each of these test sets by measuring the propor-458 435 tion of "same" predictions for each dataset. If this proportion459 436 460

is high, the model views stimuli in those datasets as "same"; if it is low, the model views them as "different." The first rows of Table 2 show the hypothesized behavior of theoretical models with certain inductive biases when tested on each of the generated datasets. For example, if a model makes predictions by comparing object shape, then it should predict "same" whenever the shape of the two objects in an image are the same (S) and "different" otherwise. Ideally, a model that has learned our definition of "same" (i.e. pixel-level similarity) should not predict "same" for any case besides CTS.

Comparing the first row of results to predicted behavior, the "same" predictions made by Random ViT-B/16 on Color Shapes align closely with the predicted "color-biased model" behavior. This confirms our result from Figure 6, which shows that this model cannot generalize to datasets without color. If the same architecture is pretrained with CLIP and *then* finetuned on Color Shapes, its predictions become much more sensitive to texture and shape. However, these results reveal a bias towards color and texture. For example, CLIP ViT-B/16 classifies CT images as "same" 89% of the time when finetuned on Color Shapes, but only 2% of the time when finetuned on Masked Shapes. This indicates that CLIP ViT-B/16 maintains an inductive bias towards color and texture during fine-tuning; it only learns to compare object shape when there Table 2: Predicted results of dissociation experiments compared to actual results from ViT-B/16 models fine-tuned on different versions of the original SHA dataset. The proportion of "same" predictions for different types of images should change based on the inductive bias a given model is using. Even CLIP-pretrained ViT-B/16, which seemed from Figure 6 to be unbiased, is revealed to have a slight bias towards either color & texture or shape depending on its fine-tuning dataset. Median results over five seeds are reported for each row. Results for Random ViT-B/16 fine-tuned on Grayscale and Masked Shapes are not shown due to low accuracy (making the results difficult to interpret); full table is Table 16.

	Acc.		Proportion of "Same" Predictions						
Predicted \downarrow	acc.	none	S	Т	TS	С	CS	СТ	CTS
(no bias) color texture shape	1.00 1.00 1.00 1.00	0.00 0.00 0.00 0.00	0.00 0.00 0.00 1.00	0.00 0.00 1.00 0.00	0.00 0.00 1.00 1.00	0.00 1.00 0.00 0.00	0.00 1.00 0.00 1.00	0.00 1.00 1.00 0.00	1.00 1.00 1.00 1.00
ViT-B/16 (Rand) \downarrow	acc.	none	S	Т	TS	С	CS	СТ	CTS
Color Shapes	0.91	0.15	0.15	0.17	0.16	0.86	0.87	0.96	0.97
ViT-B /16 (CLIP) ↓	acc.	none	S	Т	TS	С	CS	СТ	CTS
Color Shapes Grayscale Shapes Masked Shapes	1.00 1.00 1.00	0.00 0.00 0.00	0.01 0.00 0.04	0.03 0.01 0.00	0.09 0.06 0.24	0.12 0.02 0.00	0.41 0.26 0.47	0.89 0.59 0.02	1.00 1.00 1.00

494

495

496

⁴⁶¹ are no other features available in its fine-tuning data.

objects in their visual embedding spaces, an idea supported by the results of our linear probe experiments (Appendix A.7).

Discussion and Conclusion

462

Previous work has argued that deep neural networks struggle⁴⁹⁷ 463 to learn the same-different relation between two objects in the $^{\rm 498}$ 464 same image (Kim et al., 2018; Puebla & Bowers, 2022). $\mathrm{but}^{^{499}}$ 465 the scope and nature of these difficulties are not fully under-466 stood. In this article, we tested several architectures with $a^{\rm 501}$ 467 number of pretraining methods and fine-tuning datasets in or-468 der to investigate the ability of neural networks to learn and $^{\scriptscriptstyle 503}$ 469 generalize the same-different relation. Some of our $\mathsf{model}^{^{504}}$ 470 configurations are able to generalize the relation across $\mbox{all}^{\rm 505}$ 471 of our out-of-distribution evaluation datasets; the best $\mathsf{model}^{^{506}}$ 472 is CLIP ViT fine-tuned on the Squiggles same-different task.507 473 Across five random seeds, this model yields a median test ac-508 474 curacy of nearly 100% on every evaluation dataset we use.509 475 Furthermore, this model can generalize the relation from an⁵¹⁰ 476 artificial 2D setting to a more challenging 3D setting with up511 477 to 95% test accuracy without any additional fine-tuning on 3D512 478 stimuli. The existence of such a model suggests that deep513 479 neural networks can learn generalizable representations of514 480 the same-different relation, at least for the tests we examined.515 481 There are a number of possible reasons why CLIP-516 482 pretrained Vision Transformers exhibit the strongest out-of-517 483 distribution generalization. CLIP pretraining may be helpful⁶¹⁸ 484 because of the diversity of the dataset, which Fang et al.519 485

(2022) argue is key to the robust generalization of CLIP mod-520 486 els in other settings. This success may also be due to volume 521 487 CLIP models were trained on 400 million images, an order of 522 488 magnitude greater than the 14 or 1.2 million images in Ima-523 489 geNet datasets. Another hypothesis is that linguistic supervi-524 490 sion from captions containing phrases like "same," "different,"525 491 or "two of" (which ImageNet-supervised models would haves26 492 no exposure to) helps models to separate same and differents27 493

Even with CLIP pretraining, only ViT architectures exhibit strong out-of-distribution generalization. This may be due to their larger receptive field size; CNNs can only compare distant image patches in deeper layers, whereas ViTs can compare any image patch to any other as early as the first selfattention layer. Thus, ViTs may be able to integrate complex shape information and compare individual objects to each other more efficiently than CNNs. Indeed, Lepori et al. (2024) find that ViT architectures compute this relation through direct comparisons between whole objects in early processing layers, regardless of the spatial distance separating the objects.

The success of ViT over ResNet is somewhat surprising. Although convolutional neural networks were directly inspired by models of visual perception in primates (Hubel & Wiesel, 1968; Fukushima, 1980; LeCun et al., 1995), we show that they do not necessarily learn more generalizable representations of abstract relations. Even CLIP ResNet—the strongest performing ResNet model—fails to match CLIP ViT's out-ofdistribution performance. This result suggests that Transformers may represent a competitive alternative model of human vision, despite the more biologically-inspired origins of convolutional networks. It also echoes work from Tuli et al. (2021) showing that, like humans, ViT models are consistently more biased towards shape compared to convolutional models.

Although Transformer models can learn more humanaligned representations, they may require more training data to do so. When trained from a random initialization on 6400 examples, ViT-B/16 achieves only 5% above guessing accuracy for half of our tasks (Table 9), while ResNet-50 achieves above-chance accuracy for all but one task (Table 8). Nonetheless, both models are extremely data-inefficient compared to humans: infants less than a year old require less

- 528 than ten examples to generalize the same-different relation537
- $_{\rm 529}$ (Hespos et al., 2021). Even when Transformer models ${\rm do}_{\rm 538}$
- learn robust representations of same-different, this alignment₅₃₉ seems to be a result of pretraining on large amounts of natu $_{540}$
- seems to be a result of pretraining on large amounts of naturalistic data rather than human-like inductive biases. Our findrational ratio r
- ings suggest that human-like visual processing can emerge in $_{542}^{541}$
- deep neural networks even without explicitly human-inspired $\frac{5}{543}$
- ⁵³⁵ architectural choices, which has important implications for
- ⁵³⁶ how to approach the computational modeling of vision.

References

- Altabaa, A., Webb, T., Cohen, J., & Lafferty, J. (2023). *Abstractors: Transformer modules for symbolic message passing and relational reasoning.*
- Anderson, E. M., Chang, Y.-J., Hespos, S., & Gentner, D. (2018). Comparison within pairs promotes analogical abstraction in three-month-olds. *Cognition*, *176*, 74–86.
- Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325-14329. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas .0803390105 doi: 10.1073/pnas.0803390105
- ⁵⁵⁰ Christie, S. (2021). Learning sameness: object and relational
 ⁵⁵¹ similarity across species. *Current Opinion in Behavioral Sciences*, *37*, 41–46.
- Davidson, G., Orhan, A. E., & Lake, B. M. (2023). Spatial
 relation categorization in infants and deep neural networks.
 Available at SSRN 4442327.
- ⁵⁵⁶ Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei,
 L. (2009). Imagenet: A large-scale hierarchical image
 database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D.,
 Zhai, X., Unterthiner, T., ... others (2020). An image is
 worth 16x16 words: Transformers for image recognition at
 scale. arXiv preprint arXiv:2010.11929.
- Ellis, K., Solar-Lezama, A., & Tenenbaum, J. (2015). Unsupervised learning by program synthesis. *Advances in neural information processing systems*, 28.
- Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., 567 Dave, A., & Schmidt, L. (2022, 17-23 Jul). Data de-568 termines distributional robustness in contrastive language 569 image pre-training (CLIP). In K. Chaudhuri, S. Jegelka, 570 L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), Pro-571 ceedings of the 39th international conference on machine 572 learning (Vol. 162, pp. 6216-6234). PMLR. Retrieved 573 from https://proceedings.mlr.press/v162/fang22a 574
- 575 .html

. 544

545

546

547

548

549

- Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, 576 S., & Geman, D. (2011). Comparing machines and 577 humans on a visual categorization test. Proceedings 578 of the National Academy of Sciences, 108(43), 17621-579 Retrieved from https://www.pnas.org/doi/ 17625. 580 abs/10.1073/pnas.1109168108 doi: 10.1073/pnas 581 582 .1109168108
- Fukushima, K. (1980). Neocognitron: A self-organizing neural
 network model for a mechanism of pattern recognition un affected by shift in position. *Biological cybernetics*, *36*(4),
 193–202.
- Funke, C. M., Borowski, J., Stosio, K., Brendel, W., Wallis,
 T. S., & Bethge, M. (2021). Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, *21*(3), 16–16.

- Geiger, A., Carstensen, A., Frank, M. C., & Potts, C. (2023).642
 Relational reasoning and generalization using nonsymbolic643
- neural networks. *Psychological Review*, *130*(2), 308. 644

⁵⁹⁴ Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Bren-⁶⁴⁵

del, W., Bethge, M., & Wichmann, F. A. (2020). Shortcute⁴⁶
 learning in deep neural networks. *Nature Machine Intelli*⁶⁴⁷
 gence, 2(11), 665–673.

⁵⁹⁷ *gence*, *2*(11), 665–673.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wich-

⁵⁹⁹ mann, F. A., & Brendel, W. (2019). Imagenet-trained⁶⁵⁰

- 600 CNNs are biased towards texture; increasing shape bias⁶⁵¹ 601 improves accuracy and robustness. In *International confer*^{£52}
- improves accuracy and robustness. In *International confer*.
 ence on learning representations. Retrieved from https://653
- openreview.net/forum?id=Byqh9j09KX 654
- Gentner, D., & Goldin-Meadow, S. (2003). Language in mind.⁶⁵⁵ Advances in the study of language and thought.
- Gentner, D., & Hoyos, C. (2017). Analogy and abstraction. *Topics in cognitive science*, *9*(3), 672–693.
- 608 Gentner, D., Shao, R., Simms, N., & Hespos, S. (2021).660
- 609 Learning same and different relations: cross-species com-661
- parisons. Current Opinion in Behavioral Sciences, 37, 84–662
 89. 663
- Giurfa, M. (2021). Learning of sameness/difference relation-664
 ships by honey bees: performance, strategies and ecolog-665
 ical context. *Current Opinion in Behavioral Sciences*, 37,666
- 615 **1–6.** 667
- Gülçehre, Ç., & Bengio, Y. (2016). Knowledge matters: Im-668
 portance of prior information for optimization. *The Journal* of Machine Learning Research, 17(1), 226–257.

⁶¹⁹ He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual⁶⁷¹ learning for image recognition. In *Proceedings of the ieee*⁶⁷²

conference on computer vision and pattern recognition (pp.⁶⁷³ 770–778).

- Hermann, K., Chen, T., & Kornblith, S. (2020). The origins and ⁶⁷⁵
- prevalence of texture bias in convolutional neural networks.⁶⁷⁶
 Advances in Neural Information Processing Systems, 33,⁶⁷⁷
 10000, 10015
- ⁶²⁶ 19000–19015. ⁶⁷⁸ ⁶²⁷ Hespos, S., Gentner, D., Anderson, E., & Shivaram, A. (2021).⁶⁷⁹
- Hespos, S., Gentner, D., Anderson, E., & Shivaram, A. (2021).⁰⁷⁹
 The origins of same/different discrimination in human in-⁶⁸⁰
 fants. *Current Opinion in Behavioral Sciences*, *37*, 69–74.
- Hochmann, J.-R. (2021). Asymmetry in the complexity of
 same and different representations. *Current Opinion in Be havioral Sciences*, *37*, 133–139.
- Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal* of physiology, 195(1), 215–243.
- Kerg, G., Mittal, S., Rolnick, D., Bengio, Y., Richards, B., &
 Lajoie, G. (2022). On neural architecture inductive biases
 for relational tasks.
- Kim, J., Ricci, M., & Serre, T. (2018). Not-so-clevr: learn-692
 ing same-different relations strains feedforward neural net-693
 works. *Interface focus*, 8(4), 20180011. 694

- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (Vol. 25, pp. 1097–1105).
- Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332-1338. Retrieved from https://www.science.org/doi/ abs/10.1126/science.aab3050 doi: 10.1126/science .aab3050
- LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.
- Lepori, M. A., Tartaglini, A. R., Vong, W. K., Serre, T., Lake, B. M., & Pavlick, E. (2024). Beyond the doors of perception: Vision transformers represent relations between objects. In A. Globerson et al. (Eds.), Advances in neural information processing systems (Vol. 37, pp. 131503– 131544). Curran Associates, Inc. Retrieved from https:// proceedings.neurips.cc/paper_files/paper/2024/ file/ed99503856b2ba85092e5413add65d86-Paper -Conference.pdf
- Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3431–3440).
- Martinho III, A., & Kacelnik, A. (2016). Ducklings imprint on the relational concept of "same or different". *Science*, *353*(6296), 286–288.
- Messina, N., Amato, G., Carrara, F., Gennaro, C., & Falchi, F. (2022). Recurrent vision transformer for solving visual reasoning problems. In *International conference on image analysis and processing* (pp. 50–61).
- Puebla, G., & Bowers, J. S. (2022). Can deep convolutional neural networks support relational reasoning in the samedifferent task? *Journal of Vision*, *22*(10), 11–11.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125.
- Stabinger, S., Rodríguez-Sánchez, A., & Piater, J. (2016). 25 years of cnns: Can we compare to human abstraction capabilities? In Artificial neural networks and machine learningicann 2016: 25th international conference on artificial neural networks, barcelona, spain, september 6-9, 2016, proceedings, part ii 25 (pp. 380–387).

Tartaglini, A. R., Vong, W. K., & Lake, B. M. (2022). A
 developmentally-inspired examination of shape versus tex ture bias in machines. *arXiv preprint arXiv:2202.08340*.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles,
A., & Jegou, H. (2021, July). Training data-efficient image transformers amp; distillation through attention. In *International conference on machine learning* (Vol. 139, pp. 10347–10357).

- Tuli, S., Dasgupta, I., Grant, E., & Griffiths, T. L. (2021). Are convolutional neural networks or transformers more like human vision? Retrieved from https://arxiv.org/abs/ 2105.07197
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L.,
 Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you
- need. Advances in neural information processing systems,
 30.
- 711 Webb, T. W., Frankland, S. M., Altabaa, A., Krishnamurthy, K.,
- Campbell, D., Russin, J., ... Cohen, J. D. (2023). *The relational bottleneck as an inductive bias for efficient abstrac-*
- tion.
- Webb, T. W., Mondal, S. S., & Cohen, J. D. (2023). Systematic
 visual reasoning through object-centric relational abstrac tion.
- Webb, T. W., Sinha, I., & Cohen, J. D. (2020). Emergent
 symbols through binding in external memory. *arXiv preprint arXiv:2012.14601*.

721 A Additional Generalization Results

A.1 Testing Models on Evaluation Sets from Puebla & Bowers (2022)

775

776

777

778

779

We test two of our models on the evaluation sets from Puebla 724 & Bowers (2022): ImageNet ResNet-50 fine-tuned on SQU,781 725 which is roughly equivalent to the models tested in Puebla $\&R_{782}$ 726 Bowers (2022), and CLIP ViT-B/16 fine-tuned on SQU, which 783 727 is our best model. We use code from Puebla & Bowers (2022)784 728 to generate test sets of 6,400 images evenly split between the $_{\rm 785}$ 729 classes, which is equal to the size of our test sets. Figure $7_{_{786}}$ 730 show all 10 evaluation datasets used in this section. Further-731 more, we report median AUC-ROC to better match Puebla₇₈₈ 732 & Bowers (2022), who report mean AUC-ROC. The rest of789 733 our methodology follows Methods. We also test models on_{790} 734 four more challenging evaluation sets from Puebla & Bowers791 735 (2022); details and results can be found in Subsection A.1.1. $_{_{792}}$ 736 We test two of our models on the 9 main evaluation $sets_{793}$ 737 from Puebla & Bowers (2022): ImageNet ResNet-50 fine-794 738 tuned on SQU, which is roughly equivalent to the models₇₀₅ 739 tested in Puebla & Bowers (2022), and CLIP ViT-B/16 fine-796 740 tuned on SQU, which is our best model. We use code from, $_{797}$ 741 Puebla & Bowers (2022) to generate test sets of 6,400 images 742 evenly split between the classes, which is equal to the size of_{7qq} 743 our test sets. Figure 7 show all 10 evaluation datasets used 744 in this section. Furthermore, we report median AUC-ROC to_801 745 better match Puebla & Bowers (2022), who report mean AUC-802 746 ROC. The rest of our methodology follows Methods. 747 803

Our ImageNet ResNet-50 results are comparable to results₈₀₄ 748 from Puebla & Bowers (2022) but not identical. Differences₈₀₅ 749 in our specific results may be due to our differing methods for₈₀₆ 750 creating our datasets. For example, the sizes of their objects₈₀₇ 751 are variable and may either be smaller or larger than our cho-ana 752 sen size of 64x64 pixels (see Figure 7 for examples). Their Im₋₈₀₉ 753 ageNet ResNet-50 model is fine-tuned on Fleuret et al. (2011)₈₁₀ 754 stimuli in which the sizes of the objects also vary, whereas our₈₁₁ 755 model is fine-tuned on objects of a fixed size. We also thicken 756 the lines of our SQU stimuli, while Puebla & Bowers (2022) do 813 757 not. Furthermore, they use more fine-tuning images than us 758 (28,000 versus 6,400), and their hyperparameters likely differ₈₁₅ 759 as well. Even still, the larger pattern of results is the same-816 760 ImageNet ResNet-50 fine-tuned on the same-different rela-817 761 tion using stimuli from Fleuret et al. (2011) (our SQU stimuli)₈₁₈ 762 attains relatively high in-distribution test accuracy but strug-819 763 gles to generalize out-of-distribution. This agrees with the re-a20 764 sults we obtain using our evaluation sets (SQU, ALPH, SHA₈₂₁ 765 & NAT); Table 6 shows that ImageNet-ResNet-50 fine-tuned₈₂₂ 766 on SQU struggles to generalize out-of-distribution. 767

In contrast, CLIP ViT-B/16 fine-tuned on our SQU dataset²²³
 achieves perfect or nearly perfect in- and out-of-distribution⁸²⁴
 generalization, with the exception of two test datasets (Lines⁸²⁵
 and Arrows). This performance is rather remarkable given⁸²⁶
 that objects in the evaluation datasets from Puebla & Bowers⁸²⁷
 (2022) vary greatly in size, whereas our CLIP ViT-B/16 modef²²⁸
 is fine-tuned on objects of a fixed size only. This suggests that⁸²⁹

CLIP ViT-B/16 fine-tuned on SQU may learn a same-different relation that is invariant to certain gualities (such as object size) without explicit fine-tuning for such invariance. This is also supported by CLIP ViT's generalization to photorealistic stimuli in Out-of-Distribution Generalization to Photorealistic Stimuli, in which objects vary in size and pose. Figure 9 shows examples of stimuli from the two more challenging datasets (Lines and Arrows) for which all five CLIP ViT-B/16 random seeds make errors. For Arrows, this lack of generalization may be due to symbols overlapping or being much closer to each other than any stimuli in our fine-tuning data. It's also possible that the model lacks the spatial reasoning required to form useful object representations for the Arrows dataset-unlike other datasets, this dataset requires the ability to reason about the direction of an identical line relative to identical triangles in order to distinguish objects. Thus, failure on Arrows may simply be due to "perceptual" errors like difficulties in segmenting the objects or failures in spatial reasoning rather than a lack of a general same-different representation. We also see a very slight decrease in test AUC-ROC for the Scrambled dataset, which is an interesting case. Errors made for this dataset were primarily due to our model misclassifying slightly scrambled and unmodified polygons as the "same." This error may offer insight into how exactly CLIP ViT-B/16 fine-tuned on SQU compares objects in an image.

However, the most surprising finding is that CLIP ViT-B/16 classifies all stimuli in the Lines dataset as "same" (see Appendix A.6), and that its ROC-AUC score is below 0.5. This is striking because all objects in the Lines dataset are actually the same under reflection. This result makes it very tempting to conclude that CLIP ViT-B/16 actually learns to generalize to reflections without ever being fine-tuned to do so. In fact, if models see the same image multiple times but flipped horizontally during pretraining-which is a common data augmentation-then pretrained models may already have reflection invariance baked in. Pretraining data augmentations have been shown to have such an effect on other abstract relational learning tasks (Davidson et al., 2023). While a proper treatment of CLIP's invariances are outside of the scope of this work, we test our intuition that CLIP ViTs learn a reflection-invariant same-different relation in Subsection A.1.3 below. We find evidence that strongly suggests our intuition. However, there are other possible contributing factors to CLIP ViT's failure on the Lines dataset. For instance, the Lines dataset consists entirely of one unique object that is scaled and flipped to create all stimuli; therefore, if our model makes a perceptual error when processing this particular object, that error could plausibly occur across the entire dataset.

A.1.1 More Challenging Datasets from Puebla & Bowers (2022) In order to better understand CLIP ViT's limitations, we further test the ability of CLIP ViT-B/16 fine-tuned on SQU (our best model from previous sections) to generalize to four additional datasets from Puebla & Bowers (2022): Rectangles, Straight Lines, Connected Squares, and Connected Circles (see Figure 10 for example stimuli). These datasets are



Figure 7: **Examples of "same" and "different" stimuli from all 10 evaluation sets in Figure 8.** The first dataset (SQU) is the in-distribution test set and is the same as our SQU dataset from the main body of this paper. The other nine datasets are generated following Puebla & Bowers (2022).



Figure 8: **Out-of-distribution test AUC-ROC for ImageNet ResNet-50 and CLIP ViT-B**/16 fine-tuned on SQU. Median AUC-ROC over five seeds is reported with individual runs also shown. The legend on the right indicates the test dataset. The two red bars in this figure show in-distribution test AUC-ROC, which is also reported for CLIP ViT-B/16 in Table 11.

somewhat more challenging than previously tested datasetses1
 because they either contain extremely minimal visual infores2
 mation (Rectangles and Straight Lines) or require models t0853
 correctly process objects consisting of two sub-objects (Con-854
 nected Squares and Connected Circles). Our methodology is855
 exactly the same as described earlier in Appendix .

Results for the same five model seeds in Figure 8 (and the857 836 main body of the paper) are presented in the bar chart in Fig-858 837 ure 10. Performance is slightly above chance for the Rect-859 838 angles and Straight Lines datasets, exactly chance for Con-860 839 nected Squares, and well above chance for Connected Cir-861 840 cles (although not near perfect or excellent). At first, these 862 841 results appear to contradict our main claim: that CLIP ViT-863 842 B/16 fine-tuned on SQU learns a generalizable representation864 843 of same-different. However, we believe that the failure of CLIP₈₆₅ 844 ViT to generalize to these datasets is actually a result of the 845 model's "fuzzy" same-different computation rather than an ab-867 846 ject failure to generalize the relation. Instead of computing a868 847 perfect equality between each pixel of the two objects, CLIP₈₆₉ 848 ViT appears to use an embedding similarity threshold to de-870 849 termine sameness. This can lead to model errors when the 871 850

learned threshold is too low for a new OOD dataset.

Our first line of evidence that this is the case is CLIP ViT's strong performance on the photorealistic dataset in Outof-Distribution Generalization to Photorealistic Stimuli. The "same" objects in these images are not the same on a pixel level, yet CLIP ViT can still accurately classify them. This strong performance could only be enabled by a "fuzzy" samedifferent computation whereby exact pixel-level details are disregarded. Note that the objects in the photorealistic dataset can vary greatly in size and pose; the objects in the Rectangles and Straight Lines datasets (Figure 10) are more or less the same entity but varied in size (and in the case of Rectangles, "pose," due to the slightly different height-width ratios of the rectangles). Thus, this size and pose invariance could explain the model's poor generalization to Rectangles and Straight Lines.

Our second line of evidence is the distribution of CLIP ViT logits on the Rectangles, Straight Lines, and Connected Circles datasets. We examine the logits of the median seed of CLIP ViT fine-tuned on SQU (i.e. the seed corresponding to the bars in Figure 10) on these four datasets. See Table 3

Lines		Г. J			л Г					
Arrows	∆_ ↑	e b	4	Ž ‡	Α	d d	4	& √	4 P	R

Figure 9: "Different" images misclassified by CLIP ViT-B/16 as "same" from Puebla & Bowers (2022)'s Lines and Arrows datasets. These stimuli are randomly sampled from the set of stimuli misclassified by all five seeds. Nearly 100% of model errors across evaluation datasets and seeds are mistaking "different" stimuli for "same" stimuli, so we only show mistakes of this kind. Note that the "different" Lines stimuli (middle row) are actually the same under reflection. Confusion matrices computed on these two datasets for the models tested in this section can be found in Appendix A.6.



Figure 10: Example stimuli (left) and median test AUC-ROC scores for CLIP ViT-B/16 fine-tuned on SQU (right) on four more challenging evaluation sets from Puebla & Bowers (2022). Left: "same" stimuli are displayed in the top row, while "different" stimuli are in the bottom row. Note that "different" stimuli in the Connected Squares and Connected Circles datasets are actually the same under reflection. Right: CLIP ViT's generalization to these four datasets is notably worse than the other datasets tested in Appendix . There are a number of possible explanations; see Subsections A.1.2 and A.1.3.

Table 3: Model classifications and average logits by ground-truth class for CLIP ViT-B/16 fine-tuned on SQU on the four datasets in Figure 10. The % Pred. "Same" column indicates the percentage of all stimuli (which are evenly split between "same" and "different") are predicted "same." Nearly 100% of stimuli for each dataset receive a "same" classification. The GT "Same" and "Diff" Logit columns indicate the model's average "same" logit for ground truth "same" and "different" images respectively. Images that are actually the same receive reliably stronger "same" judgements except in the case of Connected Squares.

Dataset ↓	% Pred. "Same"	GT "Same" Logit	GT "Diff" Logit
Rectangles	99.97	3.82	3.45
Straight Lines	99.98	3.81	3.61
Connected Squares	100.0	3.73	3.76
Connected Circles	99.95	4.0	3.2

for results. First, we note that the model predicts "same" for 872 nearly 100% of the images in each dataset (% Pred. "Same" 873 in Table 3). However, the strength of these "same" classifi-874 cations differs reliably between ground truth "same" and "dif-875 ferent" images for three of the four datasets. The average 876 "same" logit for truly "same" images (GT "Same" Logit in Ta-877 ble 3) in the Rectangles. Straight Lines, and Connected Cir-878 cles datasets is higher than the average "same" logit for "dif-879 ferent" images (GT "Diff" Logit in Table 3). This indicates that 880 the model does in fact discriminate between "same" and "dif-881 ferent" to some extent for these datasets. 882

Our third and final line of evidence is the relationship of 883 the average cosine similarity between "different" object em-884 beddings in a given dataset and CLIP ViT SQU's performance 885 on the dataset. Objects that are considered on average much 886 more similar according to CLIP ViT SQU than the objects in 887 the SQU dataset predict poor generalization performance; this 888 is likely because these objects exceed the model's learned 889 threshold for judging "same," which is calibrated for SQU stim-890 uli. See Subsection A.1.2. 891

Separately, note that CLIP ViT appears to learn a same-892 different relation that is invariant to reflection; in other words, 893 the same object reflected is still considered "same" when 894 compared to the non-reflected version. The "different" im-895 ages in the Lines, Connected Squares, and Connected Cir-027 896 cles datasets are in fact the same object reflected. In fact, the gas 897 two most difficult datasets for CLIP ViT-Lines and Connected 898 Squares-both feature the same type of reflection: reflection, 330 899 across the y-axis of the objects. We test CLIP ViT SQU on at 900 a version of our fine-tuning datasets where objects are re-932 901 flected across the y-axis, finding that this drops model gener-933 902 alization to these datasets from near perfect to near chance 934 903 This strongly suggests reflection-invariance in this model. See $_{_{935}}$ 904 Subsection below. 905 936

A.1.2 Object Embedding Similarity Predicts Generaliza-937 906 tion for CLIP VIT SQU We seek to measure how visu-938 907 ally distinct the OOD objects in the Puebla & Bowers (2022)939 908 datasets are according to a CLIP ViT fine-tuned on SQU com-940 909 pared to the objects in the model's fine-tuning dataset (SQU).941 910 We hypothesize that because of the model's "fuzzy" same-942 911 different computation, it will perform worse on datasets that⁹⁴³ 912 contain objects that are more visually similar to each other944 913 945 compared to SQU objects. 914

We measure inter-object similarity by creating separate in-946 915 put images for each individual object. These images are947 916 equivalent to the "same" and "different" stimuli, except only⁹⁴⁸ 917 one object is present. The single object is randomly placed in949 918 the image. In the case of the datasets from Puebla & Bowers950 919 (2022), we generate 1,000 unique object images in this way⁹⁵¹ 920 for each dataset. In the case of our SQU fine-tuning set, we₉₅₂ 921 source 1,000 unique objects not seen during fine-tuning. For₉₅₃ 922 each dataset, we embed each single-object image using thease 923 image encoder from CLIP ViT fine-tuned on SQU. We then use 924 compute pairwise cosine similarity between the embeddings₉₅₆ 925 of all different objects. Finally, we report the average of the₉₅₇ 926



Figure 11: Average inter-object CLIP embedding similarity for each evaluation dataset (green) vs. CLIP ViT generalization performance (magenta). The vertical green lines represent the variance in inter-object similarity for each dataset. If different objects are overly similar to each other compared to the fine-tuning data (SQU; green star), model generalization performance drops significantly due to misclas-

pairwise object cosine similarity scores as the solid green line in Figure 11.

sifying all "different" images as "same."

The green star in Figure 11 marks the average interobject embedding similarity for unseen SQU objects, while the dashed green line indicates the maximal average interobject embedding similarity for which CLIP ViT SQU performs well (this corresponds to the Wider dataset; see Figure 7). The datasets with an average inter-object embedding similarity above this threshold are Lines, Connected Squares (C. Squares), Straight Lines (S. Lines), Rectangles, and Connected Circles (C. Circles). Plotted in magenta in Figure 11 are the median test AUC-ROC scores on each evaluation dataset. Model performance is near perfect for all datasets with object similarities below the threshold marked by the dashed green line (with the exception of Arrows). Model performance drops precipitously for datasets with object similarities above the threshold. Essentially, the objects in these datasets are significantly more similar to each other compared to the objects in the fine-tuning data (SQU); thus, because CLIP ViT learns a "fuzzy" same-different computation, it considers the objects in these high-similarity datasets to be the "same" according to the lower threshold learned on SQU. Model predictions for these high-similarity datasets accord with this interpretation; nearly all "different" images are misclassified as "same" (see Table 3 as well as Appendix A.6).

A.1.3 Reflection Invariance in CLIP ViT The poor generalization performance of CLIP ViT on the Lines, Connected Squares, and Connected Circles datasets from Puebla & Bowers (2022) (see Figures 7 and 10) suggest an interesting possibility: that CLIP ViT learns a reflection invariant same-different relation, despite not being trained to do so (see the



Figure 12: Example stimuli from each "flipped" dataset. 1007 1008

1009

end of Appendix). To test this, we evaluate CLIP ViTs finemon 958 tuned on SQU on "flipped" versions of our SQU, ALPH, and 959 NAT fine-tuning datasets. We skip the SHA dataset since 960 many of the shapes have bilateral symmetry. The "same" 961 stimuli in the flipped datasets are the same as the regular 962 datasets; the "different" stimuli however are created by re-963 flecting a copy of a given object about its y-axis. See the 964 right side of Figure 12 for example stimuli. Note that this 965 matches the definition of "different" used by the Lines, Con-966 nected Squares, and Connected Circles datasets from Puebla 967 & Bowers (2022). 968

We create flipped SQU, ALPH, and NAT datasets contain-969 ing 6,400 images each, evenly divided between "same" and 970 "different." We then compute test accuracy on these datasets 971 for CLIP VIT-B/16 fine-tuned on SQU using the same five 972 seeds used elsewhere in the paper. We find that median 973 model performance drops significantly for the flipped datasets 974 due to models predicting "same" for "different" images, indi-975 cating that the model considers reflected versions of the same 976 object to be the same. This effect is more severe for the two 977 OOD datasets (ALPH and NAT). Decreases in median test 978 accuracy as well as the percentage of all stimuli predicted 979 "same" for each dataset are the following: 99.6% (original) 980 to 77.1% (flipped) test accuracy for SQU, with 72.9% pre-981 dicted "same;" 97.7% (original) to 51.6% (flipped) test accu-982 racy for ALPH, with 98.4% predicted "same;" 96.7% (original) 983 to 51.8% (flipped) for NAT, with 98.2% predicted "same." This 984 invariance likely helps to explain why model generalization 985 suffers on Lines, Connected Squares, and Connected Circles 986 from Puebla & Bowers (2022). 987

988 A.2 In-Distribution Learning Curves

For each architecture and pretraining method, we plot loss and in-distribution validation accuracy per epoch of fine-tuning or training on each dataset. Lines show averages for the same set of hyperparameters (for that model & dataset) across five seeds.

A.3 Out-of-distribution Generalization Tables

We report median test accuracy over five random seeds for each pretraining method, architecture, and fine-tuning dataset. The tables below include the four main fine-tuning datasets (SQU, ALPH, SHA, NAT: see Figure 2), the gravscale and masked versions of the SHA dataset (SHA-G and SHA-M; see Figure 5a), and grayscale and masked versions of the NAT dataset (NAT-G and NAT-M). As in Table 1, rows indicate the dataset that models are fine-tuned on, while columns indicate the test dataset. The rightmost column labeled "Avg." is the row-wise average of accuracy scores across OOD evaluation sets (i.e. off-diagonal values), which indicates how well a model fine-tuned on a given dataset is able to generalize to other datasets. The bottom row labeled "Avg." is the columnwise average across off-diagonal values, indicating how difficult it is for models fine-tuned on other datasets to generalize to the given dataset.





Epoch

0.6

0.4

0.2

Epoch

0.6

0.4

0.2



Figure 14: Average loss curves for ImageNet-pretrained ResNet-50 and ViT-B/16 fine-tuned on each dataset. Models converge substantially faster than in Figure 13. ImageNet ViT-B/16 models fine-tuned on SHA and NAT already attain nearly 100% validation accuracy after only one epoch.



Figure 15: Average loss curves for CLIP-pretrained ResNet-50 and ViT-B/16 fine-tuned on each dataset. In-distribution generalization to color-containing datasets SHA and NAT seem much more difficult for CLIP ResNet-50 than CLIP ViT-B/16 (or any other model configuration). CLIP ViT-B/16 attains nearly 100% validation accuracy after only one epoch of fine-tuning on all datasets except SQU.

Table 4: **OOD test accuracy for CLIP ResNet-50 models fine-tuned on each dataset.** The model fine-tuned on NAT-G exhibits the strongest average OOD generalization, although it fails to generalize to the SQU stimuli.

CLIP ResNet-50									
$\leftarrow Test \rightarrow$									
Train ↓	SQU	ALPH	SHA	SHA-G	SHA-M	NAT	NAT-G	NAT-M	Avg.
SQU	97.7	80.8	82.9	81.9	73.6	82.0	86.6	82.6	81.5
ALPH	83.5	97.4	88.9	90.1	92.9	90.7	78.2	83.8	86.9
SHA	51.3	69.3	98.1	96.2	90.8	95.2	76.3	86.6	80.8
SHA-G	65.5	80.7	98.1	98.2	95.1	93.7	95.8	91.5	88.6
SHA-M	55.9	68.1	94.7	92.1	76.1	79.6	100	86.4	82.4
NAT	53.4	76.0	95.2	96.1	96.1	97.3	87.0	94.3	85.4
NAT-G	55.6	81.3	95.4	97.3	98.0	95.7	89.7	92.7	88.0
NAT-M	59.8	80.6	90.6	91.4	90.1	94.8	94.3	95.0	85.9
Avg.	60.7	76.7	92.3	92.2	90.9	90.2	88.3	88.2	

Table 5: **OOD test accuracy for CLIP ViT-B/16 models fine-tuned on each dataset.** It is interesting to note the different patterns of generalization between models fine-tuned on SHA, SHA-G, and SHA-M. Models fine-tuned on the SHA dataset (which contains color and texture) do not generalize very well to NAT-G and NAT-M datasets; models fine-tuned on SHA-G (which removes color) generalize somewhat better to NAT-G and NAT-M; and models fine-tuned on SHA-M (which removes color) and texture) attain 100% or near 100% accuracy on NAT-G and NAT-M. The same pattern holds for models fine-tuned on NAT, NAT-G, and NAT-M tasks.

CLIP ViT-B/16									
$\leftarrow Test \rightarrow$									
Train \downarrow	SQU	ALPH	SHA	SHA-G	SHA-M	NAT	NAT-G	NAT-M	Avg.
SQU	99.5	97.7	99.1	98.9	94.8	95.5	95.0	98.1	97.0
ALPH	59.5	99.4	99.9	99.9	98.8	99.7	95.1	97.5	92.9
SHA	50.0	56.0	100	98.6	98.2	100	60.6	77.7	77.3
SHA-G	50.2	63.5	100	99.9	99.9	100	85.5	95.6	85.0
SHA-M	55.6	93.3	100	100	99.8	100	100	97.8	92.4
NAT	50.0	68.4	99.8	97.8	99.3	100	63.0	83.7	80.3
NAT-G	50.2	70.6	99.9	98.9	100	100	71.5	93.9	87.6
NAT-M	60.2	92.7	100	99.9	100	100	94.3	98.5	92.4
Avg.	53.7	77.5	99.8	99.1	98.7	99.3	84.8	92.0	

Table 6: **OOD test accuracy for ImageNet ResNet-50 models fine-tuned on each dataset.** Unlike CLIP-pretrained models, ImageNet ResNet-50 fine-tuned on SQU actually exhibits the weakest OOD generalization.

ImageNet ResNet-50									
$\leftarrow Test \rightarrow$									
Train \downarrow	SQU	ALPH	SHA	SHA-G	SHA-M	NAT	NAT-G	NAT-M	Avg.
SQU	84.8	57.4	59.3	52.6	65.1	62.9	50.2	60.8	58.3
ALPH	61.3	83.7	60.4	69.0	78.5	70.2	68.0	73.9	68.8
SHA	51.2	66.7	94.4	90.1	78.4	84.0	64.1	66.5	71.6
SHA-G	53.9	72.6	70.8	94.6	84.2	74.2	90.1	78.9	75.0
SHA-M	56.2	68.9	73.7	92.4	79.4	68.7	99.3	79.8	77.0
NAT	50.3	58.3	80.4	69.5	78.6	90.5	62.4	70.7	67.2
NAT-G	50.8	72.2	70.0	82.8	89.8	78.2	69.1	81.1	75.0
NAT-M	50.1	74.9	66.9	76.2	84.0	74.2	78.9	88.4	72.2
Avg.	53.4	67.3	68.8	76.1	79.8	73.2	73.3	73.1	

Table 7: **OOD test accuracy for ImageNet ViT-B/16 models fine-tuned on each dataset.** Interestingly, models fine-tuned on SHA exhibit strong generalization to the grayscale and masked versions of that dataset (but still don't generalize to SQU or ALPH).

ImageNet ViT-B/16									
$\leftarrow Test \rightarrow$									
Train \downarrow	SQU	ALPH	SHA	SHA-G	SHA-M	NAT	NAT-G	NAT-M	Avg.
SQU	95.4	65.8	57.6	53.3	59.7	60.5	51.8	66.3	59.3
ALPH	81.7	97.0	50.5	51.0	59.1	52.1	52.1	67.0	59.1
SHA	50.0	50.1	100	96.2	99.3	99.4	55.8	82.5	76.2
SHA-G	50.0	61.2	100	99.8	99.8	99.9	73.9	84.8	81.4
SHA-M	57.7	88.0	99.9	99.8	99.6	97.5	99.9	97.1	91.4
NAT	50.0	50.4	97.3	80.4	97.8	100	50.4	71.8	71.2
NAT-G	50.0	50.2	98.3	91.7	99.7	99.9	54.0	87.8	82.5
NAT-M	52.2	72.3	99.8	99.3	99.9	100	91.7	98.4	87.9
Avg.	55.9	62.6	86.2	81.7	87.9	87.1	68.0	79.6	

Table 8: **OOD test accuracy for randomly-initialized ResNet-50 models trained on each dataset.** Models attain surprisingly high in-distribution test accuracy for certain datasets, such as SHA and SHA-G. Models trained on SQU appear to learn nothing even though their loss curves diminish (see Figure 13). This indicates that models are memorizing training examples, which is consistent with results from prior work (e.g. Kim et al. (2018)).

Randomly Initialized ResNet-50										
\leftarrow Test $ ightarrow$										
Train \downarrow	SQU	ALPH	SHA	SHA-G	SHA-M	NAT	NAT-G	NAT-M	Avg.	
SQU	49.8	49.7	49.5	48.2	49.1	48.3	49.6	50.0	49.2	
ALPH	53.1	69.2	58.9	59.0	55.2	58.6	50.0	50.4	55.0	
SHA	51.2	69.3	82.6	80.4	82.3	82.9	53.8	61.8	68.8	
SHA-G	50.6	67.2	85.0	85.5	87.5	84.0	59.8	67.4	71.6	
SHA-M	50.0	57.0	77.3	77.0	77.0	75.0	78.3	74.3	69.9	
NAT	52.9	69.5	81.6	80.4	80.3	80.2	55.4	68.0	69.7	
NAT-G	51.1	64.2	77.3	83.6	82.8	82.5	61.3	72.5	73.4	
NAT-M	50.0	59.4	77.2	79.1	80.3	79.2	69.2	74.4	70.6	
Avg.	51.3	62.3	72.4	72.5	73.9	72.9	59.5	63.5		

Table 9: **OOD test accuracy for randomly-initialized ViT-B/16 models trained on each dataset.** Given their larger receptive field size, randomly initialized ViTs somewhat surprisingly perform worse overall than randomly initialized ResNets (Table 8).

Randomly Initialized ViT-B/16									
$\leftarrow Test \rightarrow$									
Train \downarrow	SQU	ALPH	SHA	SHA-G	SHA-M	NAT	NAT-G	NAT-M	Avg.
SQU	51.7	51.8	51.0	53.5	52.7	53.8	51.4	53.9	52.6
ALPH	49.9	54.8	51.7	51.9	56.8	52.0	49.9	51.4	51.9
SHA	50.0	50.1	92.9	66.2	62.6	73.8	50.7	56.0	58.5
SHA-G	50.0	50.5	74.2	78.8	66.5	66.4	55.7	56.4	60.0
SHA-M	50.0	50.0	56.5	59.6	53.5	55.5	81.3	63.5	59.5
NAT	50.1	51.7	75.7	58.7	62.2	76.7	53.4	62.8	59.2
NAT-G	50.2	51.8	64.1	69.9	70.7	67.0	55.9	65.9	62.8
NAT-M	50.0	50.5	56.4	56.8	58.4	58.2	55.0	66.2	55.0
Avg.	50.0	50.9	61.4	59.5	61.4	61.0	56.8	58.6	

1011 A.4 ImageNet Models with Comparable Parameters

One explanation for the difference in performance between 1012 ResNet-50 and ViT-B/16 is the fact that ResNet-50 consists of 1013 23M parameters, whereas ViT-B/16 has a total of 86M pa-1014 rameters. To explore this possibility, we fine-tune and test 1015 ConvNeXt-B Liu et al. (2022) as an example of a convolutional 1016 model with 89M parameters (comparable to ViT-B/16), as well 1017 as DeiT-S Touvron et al. (2021), a 22M parameter transformer 1018 model of similar size to ResNet-50. 1019

1020 Results for ConvNeXt-B are shown on the left in Table 10. This model was pre-trained on ImageNet-22k, the same 1021 dataset as ImageNet ViT-B/16, and also has a similar number 1022 of parameters as ViT-B/16. Comparing the left-hand side of 1023 Table 10 to Table 7, we can see that despite ConvNeXt's com-1024 petitive edge in parameter count, ViT-B/16 still seems to have 1025 slightly stronger OOD generalization between SHA and NAT 1026 as well as between SQU and ALPH. However, ConvNeXt does 1027 generalize with a 71.7% accuracy to SHA when trained on 1028 ALPH, which is better than any other non-CLIP model tested. 1029 Looking at the right-hand side of Table 10, we see that 1030 compared to ResNet-50 (Table 6), DeiT-S has more success 1031 generalizing within SQU and ALPH as well as SHA and NAT. 1032 These two smaller models were pre-trained on ImageNet-1033 1k, but despite this DeiT-S is still guite competitive with the 1034 larger ConvNeXt, which was pre-trained on significantly more 1035 data (ImageNet-22k). Taken as a whole, these additional re-1036 sults seem to suggest that the biggest differentiator between 1037 ResNet-50 and ViT-B/16 is not parameter count or size of pre-1038 training data but architectural design. 1039

1040 A.5 Area Under the ROC Curve for CLIP Models

In addition to reporting median test accuracy across seeds,
 we report median area under the ROC curve for CLIP ResNet 50 and CLIP VIT-B/16. Table 11 below mirrors Table 1 from the
 main paper.

Models fine-tuned on the Shapes and Naturalistic datasets 1045 attain rather high AUC across all OOD test datasets, notably 1046 including the Alphanumeric task (which does not contain color 1047 or texture). CLIP ResNet-50 in particular attains > 0.8 AUC 1048 across all fine-tuning conditions and test datasets. This is in 1049 contrast to median accuracy results reported in Table 1, which 1050 shows more of a dramatic "upper triangular" pattern. This in-1051 dicates that some of the models that achieve poor OOD test 1052 accuracy may perform much more strongly with a correctly 1053 calibrated bias. Even still, the "upper triangular" pattern is 1054 still evident here-models fine-tuned on SQU and ALPH tasks 1055 demonstrate stronger generalization than models fine-tuned 1056 on SHA and NAT tasks. Furthermore, ViT still outperforms 1057 ResNet, achieving perfect AUC across all test datasets when 1058 fine-tuned on SQU. 1059

1060 A.6 Out-of-distribution Test Confusion Matrices

We consider the pattern of errors produced by two of our models: ImageNet ResNet-50 fine-tuned on SQU, which is the most similar to models tested in some prior work (Funke et Table 10: **OOD test accuracy for ImageNet ConvNeXt-B and DeiT-S fine-tuned on the main four datasets.** Performance is competitive between a small transformer model of similar size to ResNet-50 and a large convolutional model comparable to ViT-B/16, suggesting that the difference mainly comes down to architecture.

ImageNet-22k ConvNeXt-B										
	\leftarrow Test $ ightarrow$									
Train \downarrow	SQU	ALPH	SHA	NAT	Avg.					
SQU	87.0	55.3	52.8	52.5	61.9					
ALPH	64.0	93.6	71.7	61.7	72.7					
SHA	50.0	50.2	98.4	86.9	71.4					
NAT	50.0	50.6	78.5	98.2	69.3					
Avg.	62.8	62.4	75.4	74.8						
	Ima	ageNet-1	k DeiT-S		•					
		\leftarrow Te	est $ ightarrow$							
Train \downarrow	SQU	ALPH	SHA	NAT	Avg.					
SQU	85.6	61.4	54.0	53.9	63.7					
ALPH	62.2	88.7	63.1	71.3	71.3					
SHA	50.1	52.2	97.9	94.5	73.7					
NAT	50.0	52.7	85.2	98.4	71.6					
Avg.	62.0	63.7	75.0	79.5						

Table 11: **Out-of-distribution test AUC for CLIP models fine-tuned on each dataset.** Rows indicate the dataset that models are fine-tuned on, while columns indicate the test dataset. Each cell is the median performance over five random seeds. The rightmost column labeled "Avg." is the rowwise average of accuracy scores across OOD evaluation sets (i.e. off-diagonal values), which indicates how well a model fine-tuned on a given dataset is able to generalize to other datasets. The bottom row labeled "Avg." is the column-wise average across off-diagonal values, indicating how difficult it is for models fine-tuned on other datasets to generalize to the given dataset.

CLIP ResNet-50									
\leftarrow Test $ ightarrow$									
Train ↓	SQU	ALPH	SHA	NAT	Avg.				
SQU	0.99	0.95	0.93	0.86	0.91				
ALPH	0.96	0.99	0.96	0.97	0.96				
SHA	0.8	0.91	1.0	0.99	0.9				
NAT	0.83	0.94	0.99	0.99	0.92				
Avg.	0.86	0.93	0.96	0.94					
-		CLIP Vi	T-B/16						
		← Te	est $ ightarrow$						
Train \downarrow	SQU	ALPH	SHA	NAT	Avg.				
SQU	1.00	1.00	1.00	1.00	1.00				
ALPH	0.93	1.00	1.00	1.00	0.98				
SHA	0.62	0.91	1.00	1.00	0.84				
NAT	0.63	0.93	1.00	1.00	0.85				
Avg.	0.73	0.95	1.00	1.00					

al., 2021; Puebla & Bowers, 2022), and CLIP ViT-B/16 fine-1064 tuned on SQU, which is our best model. We compute confu-1065 sion matrices for both of these models on our four main test 1066 sets (SQU, ALPH, SHA, & NAT) as well as the Lines and Ar-1067 rows test sets from Puebla & Bowers (2022), which our CLIP 1068 ViT-B/16 model finds challenging (see Appendix for visual ex-1069 amples and results). We report matrices for the random seed 1070 that yields the median in-distribution test accuracy (i.e. the run 1071 that corresponds to the bars in Figure 3). 1072

In general, both ImageNet ResNet-50 and CLIP ViT-B/16 1073 models tend to mistake "different" stimuli for "same" stimuli 1074 more frequently than the converse. However, this is not al-1075 ways the case for ImageNet ResNet-50-as the top row of 1076 Figure 16 shows, ResNet makes the opposite error (mistak-1077 ing "same" for "different") much more frequently when tested 1078 on SHA and NAT datasets. This is never the case for CLIP 1079 ViT-B/16 (bottom row of Figure 16). Furthermore, the differ-1080 ence in frequency between the two types of errors is much 1081 more stark for CLIP ViT-B/16; the vast majority of errors made 1082 by this model across all test datasets are mistaking "different" 1083 stimuli for "same" stimuli. Hochmann (2021) argues that ${\rm much}^{\!\!\!\!\!120}$ 1084 of the studies on same-different relation learning in children¹¹²¹ 1085 and animals can actually be accounted for by subjects learn¹¹²² 1086 ing a concept of "same" without learning a symmetric concept t^{1123} 1087 of "different;" in other words, a subject can achieve high per-1088 formance on many same-different tasks used in the cognitive 1089 science literature by only recognizing when two objects $ar e^{126}$ 1090 the same as each other (without explicitly representing "dif- $^{\!\!\!127}$ 1091 ferent"). This seems to align with the errors made by $\mathsf{CLIP}^{^{128}}$ 1092 ViT-B/16. It is possible that this model learns a stronger or^{1129} 1093 more coherent concept of sameness and thus decides to out¹¹³⁰ 1094 . 1131 put "same" whenever it is less certain. 1095

Another notable result is the CLIP ViT-B/16 confusion $\mathrm{ma}^{^{1132}}$ 1096 trix for the Lines dataset from Puebla & Bowers (2022). The $^{1133}_{\mbox{\scriptsize h}}$ 1097 model assigns the label "same" to 100% of the "different" stim¹¹³⁴ 1098 uli with relatively high confidence (as indicated by the $<0.5^{\rm \scriptscriptstyle 135}$ 1099 AUC-ROC score on this dataset in Appendix). This is in ${\rm con}^{\!\!\!\!\!^{1136}}$ 1100 trast to ImageNet ResNet-50, which appears to assign cate¹¹³⁷ 1101 gory labels at random for the Lines dataset. As extrapolate $d^{1^{138}}$ 1102 in Appendix , the "different" stimuli in this dataset are actually $\ensuremath{\mathsf{I}}^{139}$ 1103 the same under reflection, suggesting that CLIP ViT-B/16 fine $\stackrel{i_{140}}{-}$ 1104 tuned on SQU may learn a reflection-invariant same-different $\!\!\!\!\!\!\!^{1141}$ 1105 1106 though this is speculative). 1107 1143

1108 A.7 Probing CLIP Embeddings

In order to determine the degree to which CLIP pretrain1145 1109 ing alone encodes useful information for learning the same1146 1110 different relation, we perform a linear probe on the CLIP¹⁴⁷ 1111 ResNet-50 and CLIP ViT-B/16 models. As in our main ex1148 1112 periments, we append a linear binary classifier to the visual149 1113 backbone of each model. However, in this experiment, we¹⁵⁰ 1114 freeze the pretrained weights in the backbone of each model151 1115 and train only the parameters of the classifier on the fixed em1152 1116 beddings given by the backbone. Results are displayed in153 1117 1154 Table 12. 1118

Table 12: Out-of-distribution test accuracy for the best linear probe trained on CLIP embeddings of each dataset.

	CLI	P ResNe	t-50 Pro	be	
		\leftarrow Te	est $ ightarrow$		
Train \downarrow	SQU	ALPH	SHA	NAT	Avg.
SQU	62.4	50.0	50.0	50.0	50.0
ALPH	50.0	72.7	50.1	49.8	49.9
SHA	50.0	50.0	85.6	50.3	50.1
NAT	50.0	49.9	52.5	85.6	50.8
Avg.	50.0	50.0	50.8	50.0	
	CL	IP ViT-B	16 Prob	e	
		\leftarrow Te	est $ ightarrow$		
Train \downarrow	SQU	ALPH	SHA	NAT	Avg.
SQU	81.9	51.1	55.8	52.7	53.2
ALPH	50.0	94.4	53.1	58.5	53.9
SHA	50.0	50.0	99.9	90.4	63.5
NIAT					
INAI	50.0	50.1	70.6	100	56.9

We find that the linear probe can generally exhibit rather high in-distribution generalization. CLIP embeddings of Naturalistic stimuli produce the highest in-distribution test accuracy, followed closely by Shapes. CLIP embeddings of ALPH and SQU datasets are more difficult to learn from. This mirrors the ordering observed in Out-of-Distribution Generalization in which the two same-different tasks containing color and texture features tend to be easier to learn, while the shape-based tasks tend to be more difficult. The fact that Alphanumeric and Squiggles probes are unable to generalize OOD, however, is odd considering the fact that the solutions to both of these datasets should be the same (based on shape); this implies there is some other signal that linear probes are picking up on in order to separate "same" and "different" stimuli in these cases.

In the case of CLIP ResNet-50, the linear probe does not generalize to any OOD stimuli. On the other hand, CLIP ViT-B/16 probes trained on Shapes or Naturalistic stimuli generalize somewhat well to each other (90.4% generalization from Shapes to Naturalistic; 70.6% from Naturalistic to Shapes). Somewhat surprisingly, the CLIP ViT-B/16 probe trained on the Squiggles dataset does not generalize the relation to other datasets despite the impressive generalization performance of the fully fine-tuned model.

A.8 CLIP Embedding Cosine Similarity Distributions

1144

Interestingly, Table 13 shows that ViT-B/16's embeddings seem to become more distinct during fine-tuning whereas ResNet-50's become closer together. This is likely *not* due to differences in generalization performance given that the median difference between ViT-B/16 and ResNet-50 for withindistribution generalization is only 1.9%, and the median difference in out-of-distribution generalization is 13.9%. We do not have a clear explanation for this phenomenon, and also concede that it may be a methodological problem resulting from calculating cosine similarity between CLIP embeddings after



Figure 16: Confusion matrices for ImageNet ResNet-50 (top row) and CLIP ViT-B/16 (bottom row) fine-tuned on SQU. Each column gives confusion matrices for a given test set as indicated by the labels above. The rows of the confusion matrices are the true labels (TD means "true different"; TS means "true same"), while the columns of the matrices are the predicted classes (PD means "predicted different"; PS means "predicted same"). Each cell in the matrix shows the number of test images with a given true label and a predicted label as assigned by each model.



Figure 17: Distribution of cosine similarities between CLIP ResNet-50 representations of the Squiggles, Alphanumeric, Shapes, and Naturalistic datasets. These cosine similarities are calculated *before* fine-tuning. n = 6,400 for each dataset, 20.48M pairs calculated per dataset.

1155 extensive fine-tuning.

1167 Given our hypothesis that generalization accuracy should₁₆₈ 1156 correlate with greater cosine similarity of representations $be_{\overline{1169}}$ 1157 fore fine-tuning, it is odd that the masked versions of Shapes,170 1158 and Naturalistic sometimes have greater average cosine sim-1159 ilarity measures than Alphanumeric, despite having worse171 1160 generalization accuracy (Tables 4-8). However, this is likely₁₇₂ 1161 due to the fact that masking shapes greatly decreases the $ef_{f_{173}}$ 1162 fective number of unique tokens in the dataset. For example_{p174} 1163 the Shapes dataset only has 16 unique shapes, so masking₁₇₅ 1164 those objects results in only 16 unique objects in total. Ap₁₁₇₆</sub> 1165 pendix C shows that training on a dataset with so few tokens is177 1166



Figure 18: Distribution of cosine similarities between CLIP ViT-B/16 representations of the Squiggles, Alphanumeric, Shapes, and Naturalistic datasets. These cosine similarities are calculated *before* fine-tuning. n = 6,400 for each dataset, 20.48M pairs calculated per dataset.

detrimental to in- and out-of-distribution generalization. Thus, datasets with a high average cosine similarity seemingly only improve generalization in the cases where they also include a diversity of unique training objects (like the Squiggles dataset).

A.9 Fine-tuning on Noise

We initially calculated average pairwise cosine similarity for CLIP representations of random Gaussian noise as a baseline for measuring visual diversity within our datasets (Table 14). However, after observing a pattern in which more closelyembedded datasets induce stronger out-of-distribution generalization, we decided to see whether models perform even



Figure 19: **Examples of stimuli used when fine-tuning on noise.** From left to right: a single example object; a stimulus labeled as "same;" a stimulus labeled as "different." All noise stimuli were sampled from a Gaussian distribution with $\mu = 0$ and $\sigma = 1$.

Table 13: Average pairwise cosine similarity between CLIP embeddings of training stimuli within each dataset. Because n = 6,400 for each dataset, averages are computed over 20.48M pairs. We extract CLIP embeddings *before* fine-tuning on the same-different task and *after* fine-tuning on the Squiggles task (median across five seeds).

	Before Fin	e-tuning	Fine-tuned on SQU		
Dataset \downarrow	ResNet-50	ViT-B/16	ResNet-50	ViT-B/16	
noise	0.992	0.993	0.983	0.997	
SQU	0.929	0.940	0.992	0.283	
ALPH	0.881	0.889	0.984	0.634	
SHA	0.855	0.861	0.949	0.548	
NAT	0.788	0.805	0.937	0.568	
SHA-G	0.868	0.873	0.938	0.538	
SHA-M	0.900	0.904	0.948	0.579	
NAT-G	0.845	0.850	0.944	0.513	
NAT-M	0.882	0.879	0.940	0.407	

better when they are fine-tuned on a version of the same-1178 different task where they must label two same-versus-different 1179 64x64 squares of random Gaussian noise (see Figure 19). 1180 Theoretically, if models fine-tuned on this task are forced to 1181 compare objects on the level of individual pixels, they should 1182 be able to generalize to any same-different dataset in which 1183 objects are the same on a pixel level (the definition of same-1184 ness we employ in this work). 1185

We use the same methodology as described in Methods. That is, we fine-tune CLIP ResNet-50 and CLIP ViT-B/16 on this task, sweeping over the learning rates (1e-4, 1e-5, 1e-6, 1e-7, 1e-8) and two learning rate schedulers (Exponential,²⁰⁰ ReduceLROnPlateau). We report results for the best models²⁰¹ trained for 70 epochs with a batch size of 128 in Table 15. ¹²⁰²

As shown in Table 15, models fine-tuned on noise largel 1192 fail to generalize. One likely explanation for this lack of gen1204 1193 eralization is that models fine-tuned on noise learn to attend²⁰⁵ 1194 to small regions in both objects (e.g. two adjacent pixels in 206 1195 the corner of each object) and calculate whether those small^{p07} 1196 regions are equivalent. This might help explain why CLIP ViTL208 1197 B/16 fine-tuned on noise generalizes quite strongly to the SHA²⁰⁹ 1198 and NAT datasets-these two datasets contain textures, so 1199

Table 14: Average pairwise cosine similarity between CLIP embeddings of training stimuli within each dataset. Because n = 6,400 for each dataset, averages are computed over approximately 20M pairs. We extract CLIP embeddings *before* fine-tuning on the same-different task. For similarities afterwards, see Appendix .

Dataset \downarrow	ResNet-50	ViT-B/16
noise	0.992	0.993
SQU	0.929	0.940
ALPH	0.881	0.889
SHA	0.855	0.861
NAT	0.788	0.805

Table 15: **Out-of-distribution test accuracy for CLIP models fine-tuned on noise.** Rows indicate model architecture and number of epochs, while columns indicate the test dataset. Each cell is the median performance over five random seeds. The rightmost column labeled "Avg." is the rowwise average of accuracy scores across OOD evaluation sets (i.e. not including the NOISE column), which indicates how well a model is able to generalize to other datasets. The bottom row labeled "Avg." is the column-wise average, indicating how difficult it is for models fine-tuned on noise to generalize to that given dataset.

			\leftarrow Test –	÷		
Model \downarrow	NOISE	SQU	ALPH	SHA	NAT	Avg.
ViT-B/16	95.3	50.3	65.1	97.1	96.9	77.4
ResNet-50	94.9	50	50	61.2	59.3	55.1
Avg.	95.1	50.2	57.6	79.2	78.1	

this potential strategy of computing equality based on highly localized features would work well. On the other hand, this strategy would likely fail for stimuli in the Squiggles and Alphanumeric tasks, which consist of primarily empty space and require the integration of more global shape information. Although the idea of training on noise for abstract-relations is promising in theory (since there should not be spurious, nongeneralizing visual features), it would require careful design to counteract such undesirable local "shortcuts" (Geirhos et al., 2020).



Figure 20: **Out-of-distribution test accuracy for CLIP models for each fine-tuning dataset across all five random seeds.** The top row shows test accuracy for CLIP ResNet-50, while the bottom row shows test accuracy for CLIP ViT-B/16. The columns indicate the fine-tuning dataset (from left to right: SQU, ALPH, SHA, & NAT), while the legend indicates the test dataset. Each individual plot point is the test accuracy for a given random seed. Stars represent the median test accuracy, which are equivalent to the values reported in Table 1.

A.10 Sensitivity of OOD Generalization to Random²³⁶ Seed

In Table 1, we report median out-of-distribution test accur238 racy across five random seeds for CLIP ResNet-50 and CLIP239 VIT-B/16. Here, we extend this table by reporting out-ofree distribution test accuracy for all five random seeds.

All model configurations demonstrate some sensitivity to242 1216 random seed. However, the two best generalizing models-1243 1217 CLIP ResNet-50 fine-tuned on ALPH (Figure 20B) and CLIP244 1218 VIT-B/16 fine-tuned on SQU (Figure 20E)-demonstrate a dis+245 1219 tinct bimodal distribution across seeds. While some seeds at1246 1220 tain high test accuracy across all three OOD test sets. one247 1221 (CLIP ViT) or two seeds (CLIP ResNet) perform substantially248 1222 worse across all three sets. This creates a visible gap be+249 1223 tween points that persists across all three OOD test sets im250 1224 panels B and E in Figure 20. Other configurations demon+251 1225 strate such a gap for one or two test sets (e.g. panel A and²⁵² 1226 panel D in Figure 20), but no other configurations demonstrate²⁵³ 1227 such a gap for all three OOD sets. 1254 1228

1229It is interesting to consider the fact that the only randomnes\$2551230in our setup for these models is in the data batching (sinc1231models are initialized with deterministic, pretrained weights)1232This indicates that the order in which models see particular2581233examples from the training set is important for abstraction and1234determines whether or not models discover the generalizing1235solution.

A.11 Additional Photorealistic Evaluation Results

Using the objects depicted in Figure 21, we create two conditions of the photorealistic evaluation dataset described in Outof-Distribution Generalization to Photorealistic Stimuli: one in which individual objects in a given image are randomly and independently rotated, and one in which objects are given the same random rotation. The first condition presents a more challenging out-of-distribution task for our fine-tuned models than the second since it introduces additional and substantial variation between "same" objects.

We evaluate our SQU. ALPH. SHA. & NAT fine-tuned models as described in the Out-of-Distribution Generalization to Photorealistic Stimuli section on both conditions of the photorealistic dataset. None of the models receive any additional fine-tuning on the photorealistic dataset. Results for all pretraining and fine-tuning combinations are displayed in Figure 22-the hatched bars indicate the easier identical rotation condition, while unhatched bars indicate the more difficult individual rotation condition. CLIP ViT models demonstrate impressive generalization to the photorealistic stimuli across all fine-tuning datasets. ImageNet pretrained ViT models that are fine-tuned on the SHA and NAT datasets demonstrate some generalization to the photorealistic setting. All other models fail to generalize. In particular, although CLIP ResNet-50 demonstrates a similar generalization pattern to CLIP VIT-B/16 in the Out-of-Distribution Generalization sec-



Figure 21: Images of all 16 3D objects used to create the photorealistic evaluation set in Out-of-Distribution Generalization to Photorealistic Stimuli. Note that many of the objects lack rotational symmetry, e.g. the rubber duck (top row, third image) or the mug (bottom row, first image)—thus, different views of these objects can appear substantially different.



Figure 22: Median test accuracy (top row) and AUC-ROC (bottom row) for models fine-tuned on SQU, ALPH, SHA, & NAT and tested on the photorealistic dataset. The two plots on the left show results for ResNet models, while the two on the right show results for ViT. The bars are grouped by fine-tuning dataset, as indicated by the labels along the x-axis. The colors indicate the pretraining method. Hatched bars indicate model performance on the version of the photorealistic dataset in which objects are given identical random rotations; unhatched bars indicate model performance on the version in which individual objects are rotated independently. Individual seeds are also shown over each bar; these seeds are identical to those used in the In-Distribution Generalization and Out-of-Distribution Generalization sections.

tion as shown in Table 1, none of the ResNet models generizes
 alize robustly to the photorealistic dataset. This suggests that
 ResNet models may be prone to relying on pixel-level heurisize
 tics.

Performance improves slightly for most models when obt273 jects are rotated identically. However, models perform nearly274 just as well when objects are rotated individually. This is impressive in the case of CLIP-pretrained ViT, seeing as models were not fine-tuned for rotational invariance. Evaluation results on the Lines dataset from Puebla & Bowers (2022) in Appendix seem to support the possibility that CLIP ViT models acquire a same-different relation that is also reflection invariant despite receiving no signal to do so.



Figure 23: Examples of training images from every Table 2 and Table 16 testing dataset. Every test dataset contained 6400 images and 300 unique objects.

1275 B Inductive Bias Experiment Details

1276 B.1 Grayscale and Mask Details

Because training datasets are constructed by sampling random objects, the exact objects used between the original,
Grayscale, and Masked datasets are not the same. Details
on the training datasets are as follows:

Grayscaled Shapes Images were taken from the Shapes dataset (Training and Evaluation Datasets) and converted to grayscale using the PIL ImageOps.grayscale method.

Masked Shapes Images were taken from the Shapes dataset. Because the background was already white, we selected RGB pixels that were \leq (250, 250, 250) and replaced them with pixels of the value (100, 100, 100). Extra pixels with any values greater than 250 that are not equal to the background color (255, 255, 255) were also converted to (100, 100, 100).

1291 B.2 Dissociating Color, Texture, and Shape

1292 C Diversity of Training Data Heatmaps

1293 See Figure 24.

Table 16: Predicted results of dissociation experiments, along with actual results from all models trained on different versions of the original Shapes dataset. Ideally, the proportion of "same" predictions for different types of images should change based on the inductive bias a given model is using. Median results over five seeds are reported for each row. SHA=Color Shapes, GRAY-SHA=Grayscale Shapes, MASK-SHA=Masked Shapes.

	Acc.		Р	roportio	on of "Sa	ame" Pro	ediction	s	
Predicted ↓	acc.	none	S	Т	TS	С	CS	СТ	CTS
(no bias)	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
color	1.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
texture	1.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	1.00
shape	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
ViT-B/16 ↓	acc.	none	S	Т	TS	С	CS	СТ	CTS
SHA (Rand)	0.91	0.15	0.15	0.17	0.16	0.86	0.87	0.96	0.97
GRAY-SHA (Rand)	0.77	0.33	0.35	0.45	0.50	0.41	0.48	0.80	0.87
MASK-SHA (Rand)	0.61	0.52	0.65	0.55	0.66	0.59	0.68	0.63	0.73
SHA (ImageNet)	1.00	0.00	0.02	0.01	0.06	0.34	0.81	0.82	1.00
GRAY-SHA (ImageNet)	1.00	0.00	0.01	0.00	0.06	0.05	0.40	0.47	1.00
MASK-SHA (ImageNet)	1.00	0.00	0.15	0.00	0.28	0.00	0.82	0.03	1.00
SHA (CLIP)	1.00	0.00	0.01	0.03	0.09	0.12	0.41	0.89	1.00
GRAY-SHA (CLIP)	1.00	0.00	0.00	0.01	0.06	0.02	0.26	0.59	1.00
MASK-SHA (CLIP)	1.00	0.00	0.04	0.00	0.24	0.00	0.47	0.02	1.00
ResNet-50 ↓	acc.	none	S	Т	TS	С	CS	СТ	CTS
SHA (Rand)	0.83	0.25	0.29	0.34	0.35	0.43	0.44	0.71	0.90
GRAY-SHA (Rand)	0.84	0.27	0.29	0.39	0.41	0.38	0.40	0.81	0.96
MASK-SHA (R)	0.79	0.26	0.36	0.37	0.48	0.34	0.47	0.47	0.85
SHA (ImageNet)	0.93	0.15	0.49	0.17	0.59	0.39	0.94	0.43	1.00
GRAY-SHA (ImageNet)	0.79	0.41	0.64	0.44	0.65	0.59	0.96	0.61	0.99
MASK-SHA (ImageNet)	0.84	0.17	0.49	0.17	0.45	0.27	0.83	0.28	0.85
SHA (CLIP)	0.98	0.04	0.11	0.05	0.15	0.20	0.60	0.47	1.00
GRAY-SHA (CLIP)	0.98	0.04	0.42	0.07	0.54	0.06	0.53	0.15	1.00
MASK-SHA (CLIP)	0.98	0.04	0.90	0.05	0.95	0.05	0.92	0.07	1.00



Figure 24: Validation accuracies for a ViT-B/16 ImageNet model fine-tuned on different numbers of unique objects and different amounts of Squiggles stimuli. Hyperparameters chosen correspond with the best-performing Squiggles model from Figure 3. Each cell is averaged over five different seeds. ImageNet ViT-B/16 must be fine-tuned on at least 25,600 images containing at least 1,024 unique tokens to achieve high out-of-distribution accuracy.



Figure 25: Examples of stimuli in the Aligned condition (a) and (b) and the Unaligned condition (c) and (d). Objects used are from the same Naturalistic dataset in the main paper (Brady et al., 2008).

1294 D Patch Alignment Experiment

One intuition is that ViT models may be able to more easily compute same-different due to their ability to directly compare image patches using attention. This implies that if objects were aligned with ViT image patches, it might be easier for ViT models to implement the same-different relation (since segmentation would effectively already be done for the model).

We consider whether aligning objects with ViT patches al-1301 lows for quicker convergence or more robust in-distribution 1302 generalization. Figure 25a and 25b show stimuli under the 1303 Aligned condition, where objects are aligned within the grid of 1304 tokens used by ViT models to process images. For ViT-B/16, 1305 each object takes up a 4x4 sub-grid of tokens (16 total); for 1306 ViT-B/32, objects take up 2x2 sub-grids (4 tokens total). The 1307 sub-grids in which the objects are placed are randomly chosen 1308 for each stimulus. The number of possible spatial configura-1309 tions is exactly 36 for same stimuli (9 choose 2) and 72 for 1310 different stimuli. 1311

On the other hand, Figure 25c and 25d show stimuli un-1312 der the Unaligned condition. In this case, stimuli are randomly 1313 placed and do not have to align with ViT tokens (just as in the 1314 rest of our experiments). The result is that the objects span 1315 a larger number of tokens, and the number of configurations 1316 that the objects can occupy from the point of view of the ViT 1317 is combinatorially much larger than in the Aligned condition. 1318 Thus, ViT models trained on these stimuli must integrate in-1319 formation across a larger and much less predictable set of 1320 tokens. The number of possible spatial configurations is on 1321 the order of 100 million. 1322

In all experiments, models are trained to classify images as same or different with cross entropy loss and a batch size of 64 for 30 epochs. Each experiment uses an initial learning rate of 2e-6, a ReduceLROnPlateau learning rate scheduler (patience=2), and an AdamW optimizer (weight decay=1e-2). Models are fine-tuned on 6400 stimuli (with 1920 unique training objects, disjoint from 240 other validation objects). ⁶

Table 17 shows results for ImageNet models fine-tuned on these datasets. Contrary to our hypothesis, it seems that ViT models do not benefit from having objects aligned to their token patches. In fact, the Unaligned condition provides slightly better generalization, likely because there is more variability in the training data.

Table 17: **Results for ImageNet models fine-tuned on stimuli from Figure 25.** Training accuracy, in-distribution validation accuracy, and out-of-distribution generalization to the Shapes dataset is shown.

Aligned \downarrow	Train Acc.	Val. Acc.	SHA Acc.
ViT-B/16	100	100	85.6
ViT-B/32	100	99.7	82.4
ResNet-50	87.2	68.9	53.9
ResNet-152	99.5	89.2	74.6
Unaligned \downarrow	Train Acc.	Val. Acc.	SHA Acc.
Unaligned↓ ViT-B/16	Train Acc.	Val. Acc.	SHA Acc. 91.0
Unaligned↓ ViT-B/16 ViT-B/32	Train Acc. 100 100	Val. Acc. 100 99.5	SHA Acc. 91.0 96.9
Unaligned ↓ ViT-B/16 ViT-B/32 ResNet-50	Train Acc. 100 100 85.7	Val. Acc. 100 99.5 66.9	SHA Acc. 91.0 96.9 55.9

⁶The setup of this experiment is slightly different from the main paper is because it was an early exploratory result.