Successes and Limitations of Object-centric Models at Compositional Generalisation

Milton L. Montero
Digital Design Department
IT University of Copenhagen
Copenhagen, Denmark
mlle@itu.dk

Jeffrey S. Bowers
School of Psychological Sciecne
University of Bristol
Bristol, UK
j.bowers@bristol.ac.uk

Gaurav Malhotra
Department of Psychology
SUNY Albany
Albany, NY, USA
gmalhotra@albany.edu

Abstract

In recent years, it has been shown empirically that standard disentangled latent variable models do not support robust compositional learning in the visual domain. Indeed, in spite of being designed with the goal of factorising datasets into their constituent factors of variations, disentangled models show extremely limited compositional generalisation capabilities. On the other hand, object-centric architectures have shown promising compositional skills, albeit these have 1) not been extensively tested and 2) experiments have been limited to scene composition — where models *must generalise to novel combinations of objects in a visual scene instead of novel combinations of object properties*. In this work, we show that these compositional generalisation skills extend to this later setting. Furthermore, we present evidence pointing to the source of these skills and how they can be improved through careful training. Finally, we point to one important limitation that still exists which suggests new directions of research.

1 Introduction

A hallmark of human intelligence is compositional generalisation, namely, our ability to perceive and comprehend novel combinations of familiar elements. For example, in the domain of vision, as long as we can perceive red triangles and blue squares, then we can also perceive blue triangles and green squares. This gives humans the ability to make "infinite use of finite means" [Von Humboldt et al., 1999, Chomsky, 2014, Smolensky, 1988, McCoy et al., 2021] and it is a key priority for AI to achieve human-like abilities. However, it has proven a challenge for neural network models of vision.

In the context of generative vision models, it was first hypothesized that disentangled representations based on the Variational Auto-Encoder (VAE) architecture could support such compositional generalisation abilities [Duan et al., 2020]. While this was a reasonable hypothesis, which had some preliminary experimental support [Higgins et al.], subsequent work showed that such results where not robust and they didn't extend across different levels of difficulty in generalisation [Montero et al., 2020, 2022].Furthermore these results extended to most popular architectures at the time in both supervised and unsupervised settings [Schott et al., 2021].

Recent work on the other hand has shown that models which perform perceptual grouping — i.e. which decompose images into constituent objects — exhibit increased compositional generalisation capabilities [Singh et al., 2021, Frady et al., Wiedemer et al.]. However, these results typically focused on scene compositions i.e. generalisation to novel configurations of known objects in a scene. This is however, a fundamentally different challenge to the one posed by composition of different object properties, both intrinsic (like shape and color) and extrinsic ones(like position and rotation), since the former requires manipulating the relation between objects and the later the relation between their parts and how this is translated from proximal to distal representations [Pizlo, 2001].

NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward.

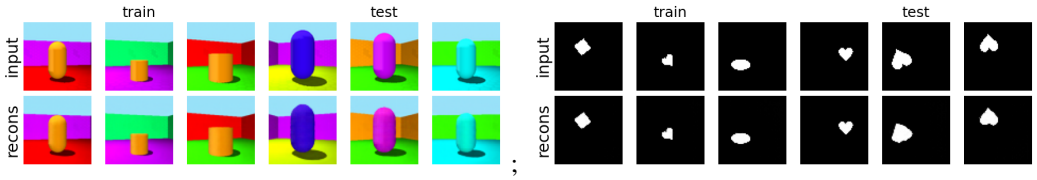

Figure 1: **Slot Attention generalisation results**: Reconstructions for a model when trained on all but some combinations of generative factors. The model is tested on said excluded combinations. Left) Generalisation results when excluding half of the combinations of the colors with the pill shape in 3DShapes. Right) Analogous test on dSprites where we exclude half of the rotations of the heart.

This has left the problem of compositional generalisation across novel combinations of *object properties* relatively unexplored. Instead most research in this area this area is focused on text-to-image (e.g. such as DALL-E) which assume the presence of language, or the use of Geometric Deep Learning approaches which exploit invariances and symmetries in the domain Bronstein et al. [2021].

In this work we turn our attention back to this question of object-level compositional generalisation (which we will refer to as simply compositional generalisation), but instead focus on exploring the capabilities of models that poses less inductive biases than the ones using either language or invariances. We will test object-centric models (specifically SlotAttention [Locatello et al., 2020] and a specific variation which we introduce later) on standard compositional generalisation tasks. Our contribution are as follows:

1. We show that SA can solve object-level compositional generalisation conditions that were found to be challenging in Montero et al. [2022] — specifically, the condition described there as recombination-to-range (R2R), where a range of combinations for two factors are removed from training.
2. We show that the model still struggles when shape needs to be combined with novel rotations, like due to this condition requiring the model to reconstruct novel local features when shapes are rotated.
3. We introduce a novel dataset to test whether this theory, and show that when novel rotations only change the global configuration of a shape but not it's local features the models then succeeds.
4. Finally we show that when trained with this novel dataset, the model even manages to solve extrapolation conditions where it must reconstruct novel shapes.

2 Object-level Compositional Generalisation in Object-centric Models

It was argued in Montero et al. [2022] that a possible cause for the failures at combinatorial generalisation described in Montero et al. [2020] and Schott et al. [2021] was that, despite learning disentangled representations, the generative models did not segment images into their constituent parts. Thus, when faced with novel combinations of properties that determine the same element of an image (e.g. novel combinations of shape and color), they could not manipulate the individual elements in an image in order to change the relevant property (e.g. the location) without affecting the representation of other properties.

Table 1: **Quantitative results on standard datasets.** Pixel-wise sum of squared errors for both an SA and a WAE model on the same datasets and conditions showed in Figure 1. For 3DShapes models with a score below 20 tend to be visually indistinguishable. For dSprites, the same effect tends to happen at 10.

	Shapes3D		dSprites	
Model	Train	Test	Train	Test
SlotAE	1.63	9.76	1.50	3.06
WAE	18.66	180.34	7.10	14.20

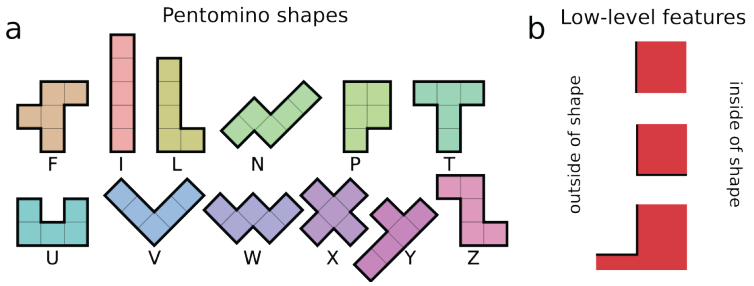

Figure 2: **Pentomino shapes** a) The twelve Pentomino shapes and their names[2]. We construct the dataset by performing affine transformations of these shapes: 5 values of scale, 40 values of rotation and 20 values of translation along each of the X and Y axis. b) The low-level features that comprise the different shapes. From top to bottom: straight lines, convex right angles and concave right angles.

We test a SlotAttention (SA) model on these same conditions for which disentangled models failed in Montero et al. [2022]: novel combinations of shape and position in dSprites, and novel combinations of shape and color in 3DShapes (see Appendix B for more details). Results can be seen in Figure 1.

The results on the left show a clear success for the 3DShapes dataset, where previous disentangled models failed catastrophically. Given the architectural properties of SA it is also easy to pinpoint to why this can be the case. First, SA possesses in-built positional knowledge in the form of position embeddings. This allows the model to capture the relation between different patches in the scene, and amongst them, the ones that encode parts of the same object. This should allow the model to map how those different patches will change when asked to produce a different rotation. Second, the SlotDecoder forces reconstructions to be performed on a per-object basis, reinforcing the innate bias of the model to limit it's representations to aggregations of patches that should be manipulated together.

The above is further illustrated with the results obtained on dSprites. On the right, the model fails when tested on novel combinations of shape and rotations on dSprites (half of the rotations of the heart where excluded). However, while the model reconstructions are worse, they do not constitute a complete failure as was the case in previous studies. In fact it seems like the model correctly identifies the required shape and rotation, but fails to properly reconstruct them. We hypothesize that in this image space, features are orientation specific which means that removing some of them prevents the model from learning how to reconstruct them. In the next section we explore how far we can push these models when we correct for said issue in the training data.

2.1 Pushing the Generalisation Capabilities of Object-centric Models

To test whether the previous failures when known shapes are presented in novel rotations are the result of models not having access to the correct local features during learning, we create a dataset where all local features are trained and test whether the model succeeds when tested on equivalent conditions (excluded shape and rotation combinations). We designed said dataset using the Pentomino shapes: sets of five blocks arranged into different shapes which we vary along different factors of variation as in dSprites ((Figure 2), see Appendix B.1 for more details).

Notice that in this dataset all low-level features are straight lines or right angles (Figure 5, panel b). Thus even when presented with a novel combination of shape and rotation as before, we can be more confident that the low-level features are not novel and the model only needs to respect the global configuration of the different parts of the shape. This is in contrast to dSprites where a novel rotation of a shape such as the heart requires reconstructing a novel local-feature (e.g. the small dip of the heart in a novel rotation is effectively a novel feature in the image frame of reference).

To facilitate our analysis we will test a simplified version of the SA architecture. In this version there is only one slot which must perform figure ground segmentation (hence we name it FgSeg). As in SA the model, it uses an attention mechanism, but instead of slots competing to explain patches of data, the latent must only integrate information about the foreground while ignoring the background. During reconstruction we use a simplified version of the SlotDecoder where instead of using a

[2]Modified from Pentomino Naming Conventions. R.A. Nonenmacher, CC BY-SA 4.0, via Wikimedia Commons.

Figure 3: **Generalization to novel shape and rotation combinations in the Pentomino dataset**. Generalization reconstructions for both FgSeg and a WAE control model. The models where trained on 11 of the 12 Pentomino shapes and tested at reconstructing a held out one in different configurations of position, rotation and scale.

Softmax to decide which slot is responsible for reconstructing a particular pixel, we use a Sigmoid activation function which determines if the latent must reconstruct a particular pixel. Thus the model is encouraged to only represent the Pentomino shape, and not the full image. This makes analysing the model easier as having slots compete to explain the Pentomino object (a.k.a. the figure) tends to split said object amongst the different slots. Additionally, it is easier to explore how the object is represented if we know it is encoded in the only available latent.

2.2 Compositional Generalisation Results

We train FgSeg on the Pentomino dataset, excluding half of the rotations for a 4 of the shapes (out of a total of 12). We then test the model on these excluded shapes as before. To control that the model performance is not only due to the qualitative differences in the dataset, we compare against a Wasserstein Auto-Encoder (WAE, Tolstikhin et al. [2017]) tested on the same generalisation condition (Figure 3).

The figure clearly shows that, while the WAE fails to reconstruct novel rotations of a known shape, the FgSeg model succeeds, which shows that a perceptual grouping model can solve previously challenging generalisation conditions given an appropriate dataset — one where novel combinations of factors do not imply the presence of novel local features in image space.

A quantitative measure is presented in Table 2, showing that the qualitative examination is supported by the scores achieved by the model. In this case, training scores refers to validation on a randomly sampled held-out dataset. In all cases the FgSeg model achieves better scores than the WAE model, though the generalisation for property prediction, while better, still doesn't match the validation performance.

Table 2: **Novel shape-rotation combination scores on the Pentomino dataset.** Scores for both FgSeg and the baseline WAE on the Pentomino dataset. Reconstruction scores are in P-MSE, while rotation prediction uses plain MSE. Classification is in accuracy.

	Reconstruction		Shape Classification		Rotation Prediction	
Model	Train	Test	Train	Test	Train	Test
WAE	5.30	10.55	48.97	25.76	0.20	0.48
FgSeg	1.11	2.15	97.7	49.92	0.022	0.42

2.3 Extrapolation results

Given this success, what happens if we now test the model on an extrapolation condition — where the model must reconstruct completely novel shapes? We show that, perhaps surprisingly, the model can also succeed when tested on reconstructing three novel shapes in the Pentomino dataset, effectively learning how to recombine the local features in the Pentomino dataset into potentially arbitrary shapes (Figure 4).

We quantify this success, showing how the FgSeg model's reconstruction scores change as we exclude more and more shapes from the training data. We see that between that there is no significant drop from 1 to three, and only after removing half of the shapes (6) do we start to see a drop in performance.

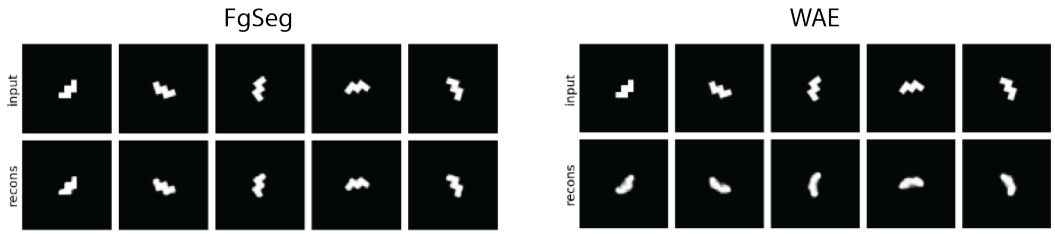

Figure 4: **New shape extrapolation** On the left, Slot Attention reconstructions of a novel shape, in this case the W. Left to right, different values of rotation sampled uniformly over the whole range of values $[0, 360)$ can be seen. On the right, the same results for WAE. It is clear that SA succeeds where WAE does not.

Table 3: **Pentomino shape-rotation generalisation.** Reconstruction scores for extrapolation conditions of increasing number of novel shapes. Reconstruction error in P-MSE.

Data Split	One Novel Shape	Three Novel Shapes	Six Novel Shapes
Train	1.65	1.29	1.25
Test	2.47	2.63	6.37

3 Discussion

We have shown that SA can solve compositional generalisation challenges that posed a significant issue for standard auto-encoder generative models. Furthermore, we have shown that failures in those datasets are likely due to the fact that some local features are effectively removed from the dataset when we exclude certain combinations, and that if we control for this fact, we can achieve generalsation for more complex factor combinations such as novel combinations of shape and rotation.

To our surprise, when trained and tested on this qualitatively different dataset, the model was even able to solve extrapolation tasks that require reconstruction of unseen shapes. This shows once a again that careful data curation can have a significant effect on model capabilities when combined with the right architecture. Our results also show that we may not require stronger inductive biases such as the ones introduce in Singh et al. [2022] to solve these tasks, and that separation into discrete tokens describing values of the properties is likely only needed for higher-level cognitive tasks (such as reasoning and planning)

Indeed our follow-up experiments suggest that the models learn abstract representations of the concepts present in said dataset. When using the learned embeddings from FgSeg to fit a classification model for the shape property, we find that a linear classifier is unable to correctly solve the problem, and instead we must use a non-linear one such as a Support Vector Machine. Even then, such a classifier is unable to correctly classify the held-out combinations from using the test embeddings without jointly training both i.e. it does not generalise in the same way the reconstruction does (see Appendix D). Extracting or learning such higher level concepts using/from the representations produced by these simpler models is thus an interesting direction for future research.

One way to accomplish this could be to design architecture that incorporate more principles from the Gestalt theory of perception, of which FgSeg and and SA only implement a subset of, namely figure-ground segmentation and perceptual-grouping [Wagemans et al., 2012, Yantis, 2001, Treisman and Gelade, 1980]. Principles such as common-fate (the idea that the visual system is biased towards grouping together features that move together in time to the same object representation), could potentially further constrain the model and induce more general representations of shape, and idea that has been preeliminary explored in Tangemann et al. [2021].

Acknowledgments

The authors would like to thank the members of the Mind & Machine Learning Group for useful comments throughout the different stages of this research. This research was supported by a ERC Advanced Grant, Generalization in Mind and Machine #741134.

References

Wilhelm Von Humboldt, Wilhelm Freiherr von Humboldt, et al. *Humboldt:'On language': On the diversity of human language construction and its influence on the mental development of the human species*. Cambridge University Press, 1999.

Noam Chomsky. *Aspects of the Theory of Syntax*, volume 11. MIT press, 2014.

Paul Smolensky. *Connectionism, constituency, and the language of thought*. University of Colorado at Boulder, 1988.

R Thomas McCoy, Jennifer Culbertson, Paul Smolensky, and Géraldine Legendre. Infinite use of finite means? evaluating the generalization of center embedding learned from an artificial grammar. 2021.

Sunny Duan, Loic Matthey, Andre Saraiva, Nicholas Watters, Christopher P. Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised Model Selection for Variational Disentangled Representation Learning. *arXiv:1905.12614 [cs, stat]*, February 2020. URL `http://arxiv.org/abs/1905.12614`. arXiv: 1905.12614.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. page 13.

Milton L. Montero, Casimir J.H. Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey S. Bowers. The role of disentanglement in generalisation. In *International Conference on Learning Representations*, 2020.

Milton L. Montero, Jeffrey S. Bowers, Rui Ponte Costa, Casimir J.H. Ludwig, and Gaurav Malhotra. Lost in latent space: Examining failures of disentangled models at combinatorial generalisation. *Advances in Neural Information Processing Systems*, 35:10136–10149, 2022.

Lukas Schott, Julius von Kügelgen, Frederik Träuble, Peter Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. Visual representation learning does not generalize strongly within the same domain. *arXiv preprint arXiv:2107.08221*, 2021.

Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate DALL-E learns to compose. *CoRR*, abs/2110.11405, 2021. URL `https://arxiv.org/abs/2110.11405`.

E. Paxon Frady, Spencer Kent, Quinn Tran, Pentti Kanerva, Bruno A. Olshausen, and Friedrich T. Sommer. Learning and generalization of compositional representations of visual scenes. URL `http://arxiv.org/abs/2303.13691`.

Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. URL `http://arxiv.org/abs/2307.05596`.

Zygmunt Pizlo. Perception viewed as an inverse problem. *Vision research*, 41(24):3145–3161, 2001.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021. URL `https://arxiv.org/abs/2104.13478`.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *CoRR*, abs/2006.15055, 2020. URL `https://arxiv.org/abs/2006.15055`.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Gautam Singh, Yeongbin Kim, and Sungjin Ahn. Neural systematic binder. *arXiv preprint arXiv:2211.01177*, 2022.

Johan Wagemans, Jacob Feldman, Sergei Gepshtein, Ruth Kimchi, James R Pomerantz, Peter A Van der Helm, and Cees Van Leeuwen. A century of gestalt psychology in visual perception: Ii. conceptual and theoretical foundations. *Psychological bulletin*, 138(6):1218, 2012.

Steven Yantis. *Visual perception: Essential readings*. Psychology Press, 2001.

Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1): 97–136, 1980.

Matthias Tangemann, Steffen Schneider, Julius Von Kügelgen, Francesco Locatello, Peter Gehler, Thomas Brox, Matthias Kümmerer, Matthias Bethge, and Bernhard Schölkopf. Unsupervised object learning via common fate. *arXiv preprint arXiv:2110.06562*, 2021.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. *dSprites: Disentanglement testing Sprites dataset*. URL `https://github.com/deepmind/dsprites-dataset/`.

A Limitations

We have presented our results using a single model. In our preliminary tests we trained several models on the generalisation conditions and found that there was no significant difference in performance for those baseline tests. As such, we conducted the rest of the experiments using only one seed per tests. Additionally, we also tested SLATE on the same conditions of dDsprites and 3DShapes and found that it performed qualitatively the same. As such we elected to not test it further as we believe it unlikely that it would perform differently.

Another limitation is that visualising the latent space provided no useful information. This is unsatisfying since raw scores tend to not provide a full picture of how the models are performing a given task.

B Datasets

We first test SA models on hard combinatorial generalisation conditions such as the recombination-to-range ones identified in (Montero et al., 2022). Namely, we test the following conditions for the three datasets:

- **3DShapes**: Contains the 6 generative factors: `[floor hue, wall hue, object hue, scale, shape, orientation]`. Colors are defined in the HSV format, and the values correspond to the hue component. Here orientation defines the angle of point-of-view for the scene. The objects themselves do not rotate. The hard condition in this case is all images such that `[shape=pill, object-hue=>` 0.5`]`, which were excluded from the training set. Again, these are pills colored as any of the colors in the second half of the HSV spectrum. These colors (shades of blue, purple, etc) were observed on the other shapes, and the pill was observed with other colors such as red and orange. See Burgess and Kim [2018].
- **dSprites**: Contains the following generative factors: `[shape, scale, orientation, position X, position Y]`. Orientation here refers to the rotation of the shape along it's center of mass. The hard condition is all images such that `[shape=heart, rotation<` 180`]`, which were excluded from the training set. Thus, no squares rotated beyond the 180 are seen during training and the model must reconstruct them at test time. We also excluded redundant rotations for the training data since shapes such as the ellipsis and the square, unlike the heart, are less than 360° symmetrical. See Matthey et al..

Notice that excluding a combination of a shape with another factor means that said shape will not be seen the same number of times during training. For example in the dSprites condition defined above squares will be observed half as many times as the other two shapes. To ensure that shapes are observed an equal amount of times, we sample instances from the dataset with the following probabiliy:

$$p(x_i, g_i) = \frac{1}{|\{(x_j, g_j) \in \mathcal{D}_{train} \mid g_j[\texttt{shape}] == g_i[\texttt{shape}]\}|} \tag{1}$$

This ensures that images in the training set that belong to the shape that is used in the generalisation condition (e.g. the square in dSprites) are seen as frequently as other shapes. Of course there will be less variation in the factor that is used to test said generalisation condition (e.g. position along the X-axis), however these have a larger number of values so they are less likely to be the cause of over-fitting.

B.1 The Pentominos Dataset

To adjudicate between these two views we unfortunately cannot rely on the dSprites dataset since the shapes used to generate it share few low-level features amongst them. Indeed the most prominent features (the curve of the ellipsis, the right angles of the square and the dip of the heart) are unique to each of them. Thus we introduce a new dataset based on the pentomino shapes to tackle this issue.

The pentomino shapes are simple, sprite-like shapes, composed of five equal side squares in different configurations that are joined edge-to-edge (see Figure 5.a). There are twelve such shapes in total

Table 4: **Pentomino generative factors**. The generative factor values used to generate the Pentomino dataset used in our experiments.

Generative factor	No. of values	Values
Shape	12	F, I, L, N, P, T, U, V, W, X, Y, Z
Scale	5	1.5, 1.8, 2.1, 2.4, 2.7, 3.0
Rotation	40	Evenly spaced in $(0°, 351°)$
Position-X	20	Evenly spaced in (-1, 1)
Position-Y	20	Evenly spaced in (-1, 1)

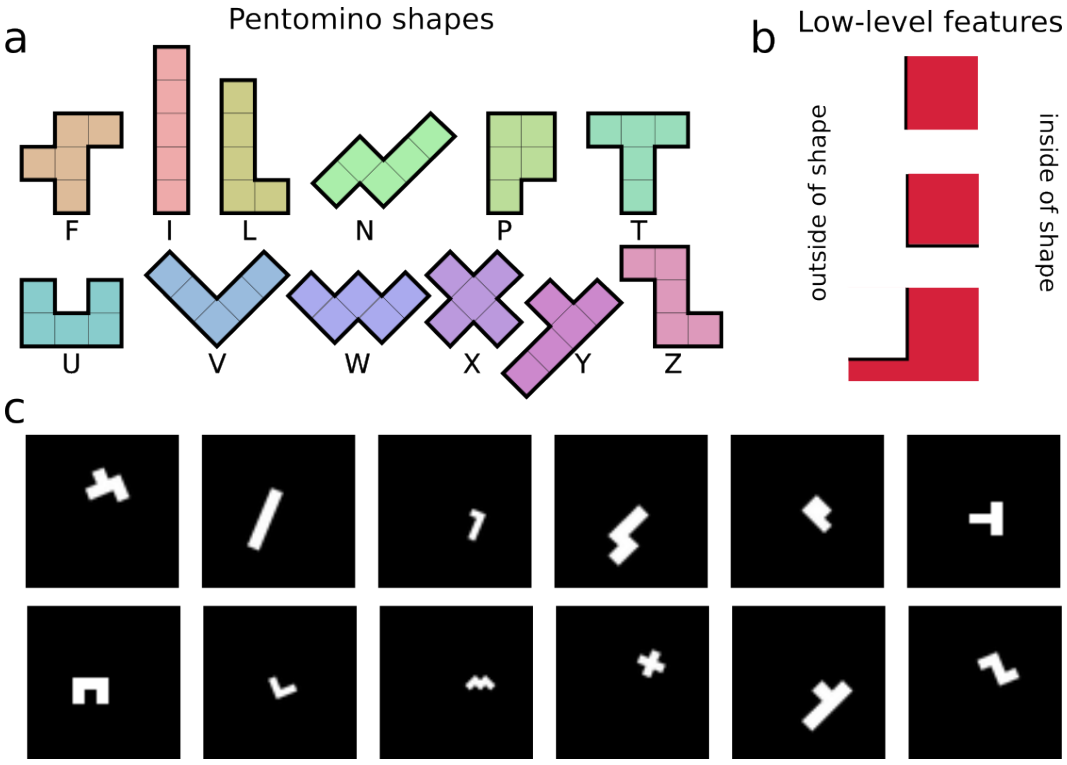

Figure 5: **Pentomino shapes** a) The twelve pentomino shapes and their names[4]. We construct the dataset by performing affine transformations of these shapes: 5 values of scale, 40 values of rotation and 20 values of translation along each of the X and Y axis. b) The low-level features that comprise the different shapes. From top to bottom: straight lines, convex right angles and concave right angles. c) Example stimuli containing different configurations of the different factors, with each shape represented once.

without taking into account rotations or mirror symmetries (the so-called free pentominos). It is then clear how these shapes solve our issue: The low-level features that compose all the shapes are straight-lines, convex right angles and concave right angles (Figure 5.b). Moreover, all shapes have at least one instance of each of these features with the sole exception of the "I" which does not have a concave right angle (notice that it is impossible to create a polygon without the other two).

To generate the dataset we use similar variations in the generative factors as in dSprites with an important caveat. Because there are 12 shapes as opposed to 3, we reduce the range of values that three of the generative factors (scale, position x, and position y) can take so that the dataset is not much larger. We preserve the 40 values of rotation since this is one of the factors we wish to test. For the rest, apart from the already established 12 values of shape, we use 5 values of scale and 20 values

for the position along each of the X and Y axis. This amounts to a total of 960000 (as opposed to the 737280 examples in dSprites). See Table 4.

C Extra results

C.1 Shape-rotation Generalisation

As in the previous section, we define a combinatorial generalisation condition that excludes some shape and rotation values:

- **Novel shape and rotation combinations**: We exclude combinations such that `[shape` $\in \{F, P, T, W\}$`, rotation >` $180°$`]`. We selected these shapes so that there are a couple of shapes that are similar to other shapes in the training data at those rotations (T and Y, W and V or U), and other two (F and P) that are very distinct. The fist pair tests if the model will confuse when the rotation value is novel and the latter if it will be able to produce a reconstruction for which it is harder to interpolate. Four shapes also allows us to maintain a similar ratio of 1:3 excluded to included during training for the shape factor (4 of 12 excluded here vs 1 of 3 in dSprites).

As in the previous section, we remove the redundant rotations of the I, X and Z shapes and correct for the unbalance of presentations of the shapes as defined by Equation 1. We use the same architecture and training configuration that we used for dSprites for both FgSeg and WAE. We use the latter as a baseline. We also use this baseline to control for the increase in the amount of shapes with respect to dSprites. If FgSeg was to succeed in this new dataset, it could be argued that this is because having twelve shapes gives the model more opportunity to learn a proper representation of what constitutes a shape. A success at generalisation by FgSeg and failure from a baseline model would rule out this possibility, which means the former's success is much more interesting than if both succeed.

o

The results can be seen in Figure 6. Reconstructions are plotted along the circumference according to the ground truth rotation value. For simplicity we keep the other generative factors fixed at the midpoint value. In the case of FgSeg we see that the results are very impressive. The model shows no sign of confusing either to rotation or the shape of unseen combinations. Thus they clearly support the second view over the first one described in the introduction to this section: errors committed on novel shape-rotation combinations are due to decoder errors.

It is also clear that said improvement in generalization is not due to the increase in the number of shapes as the results for WAE show clear and systematic failures at generalisation, especially for rotations that are far from the ones observed during training, as should be expected.

C.2 Additional Extrapolation Results

We also test the model when excluding more than one shape from traning to test the degree to which variety in the training examples influences the quality of the learned representations/genertive mechanism:

- **Three novel shapes**: We exclude all instances where `[shape` $\in \{P, T, W\}$`]`. We include P and T because when we tested WAEs these were routinely confused with F and Y respectively. Thus it may increase the likelihood of SA failing by making the same mistakes.
- **Half novel shapes**: We further increase the number of shapes by excluding inputs such that `[shape` $\in \{F, P, N, T, V, W\}$`]`. We just add F, V, Z to make it 6 out of 12 shapes excluded.

We also including raw reconstruction errors for all three conditions, including the one in the main text (Figure 4).

On the other hand, removing more shapes does produce significant degradation in the reconstruction quality. The results for this condition can be observed in Figure 8. For some shapes such as W we

[4]Modified from Pentomino Naming Conventions. R.A. Nonenmacher, CC BY-SA 4.0, via Wikimedia Commons.

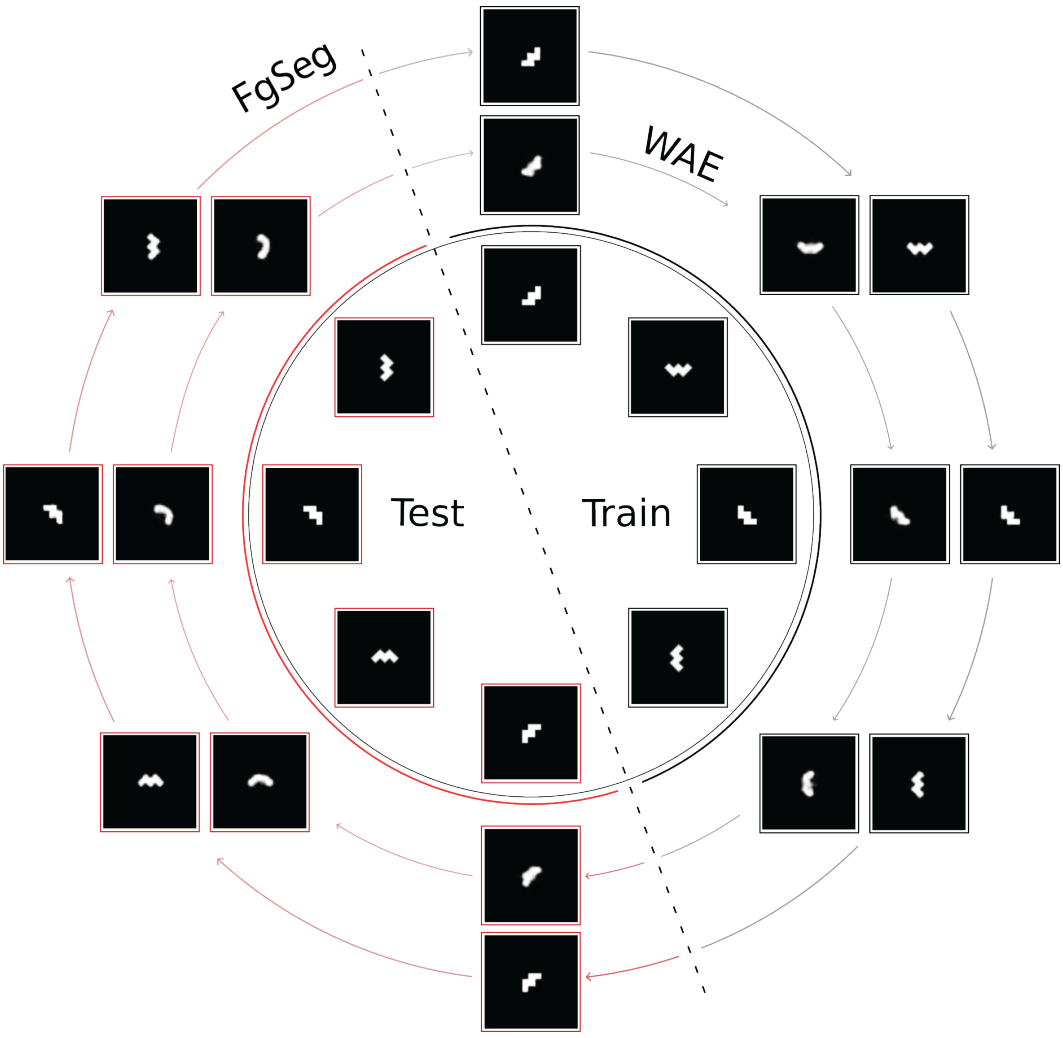

Figure 6: **Generalization to novel shape and rotation combinations in the Pentomino dataset**. Generalization reconstructions for both FgSeg and a WAE control model. Example rotations of the excluded shape are plotted inside their circumference according to their rotation value. Samples highlighted in red are novel combinations only seen during testing. The rest (highlighted in black are used during training. Reconstructions for FgSeg and WAE are plotted on the outside in the corresponding angle. FgSeg is always outside and WAE inside.

can see that the corners are not as well defined as before. For others like the T, the deformations are more pronounced. In the case of the V it seems to even confuse it with the W on some examples. Nonetheless, these results are still better than what we obtained with a WAE when excluding only one shape. Thus we can conclude that while there is an effect of data diversity, the model is fairly good at generalization even when a large number of examples are excluded form the training data.

D Testing predictivity of learned representations

Are the learned representations high level? In other words does the model learn concepts related to each shape and color or does it solve the task using simpler, lower level representations? Unlike disentangled VAEs it is not easy to directly visualise the latent representations of these models. Instead we turn to prediction of the different concepts (such as the shape classes) as a way to assess this. If models learn representations that are abstract, we should be able to use them to predict the classes for both training and held out samples. We test this below. First of all, we see that models

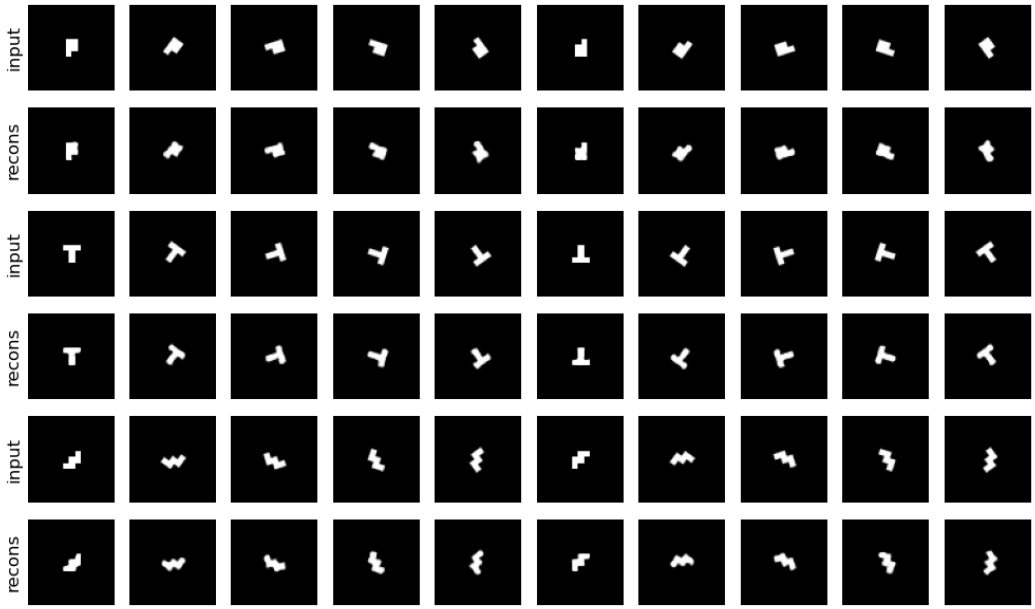

Figure 7: **Extrapolation to three new shapes**. Figure-ground Segmentation model reconstructions when three shapes are excluded from training (P, T, W). Every pair of consecutive rows contains images of one of the shapes at 10 different rotation values. These are taken at evenly spaced angels from 180° to 360° and plotted increasingly from left to right. The rest of the generative factors are kept constant. We can observe that the model achieves good quality reconstructions in spite of not having seen these shapes at all.

do learn representations that are decodable using linear probes since we need at least a two-layer MLP to achieve 100% accuracy on the training data. An even these models cannot achieve significant prediction accuracy in the test data, which shows that the model's representations are not abstract after all.

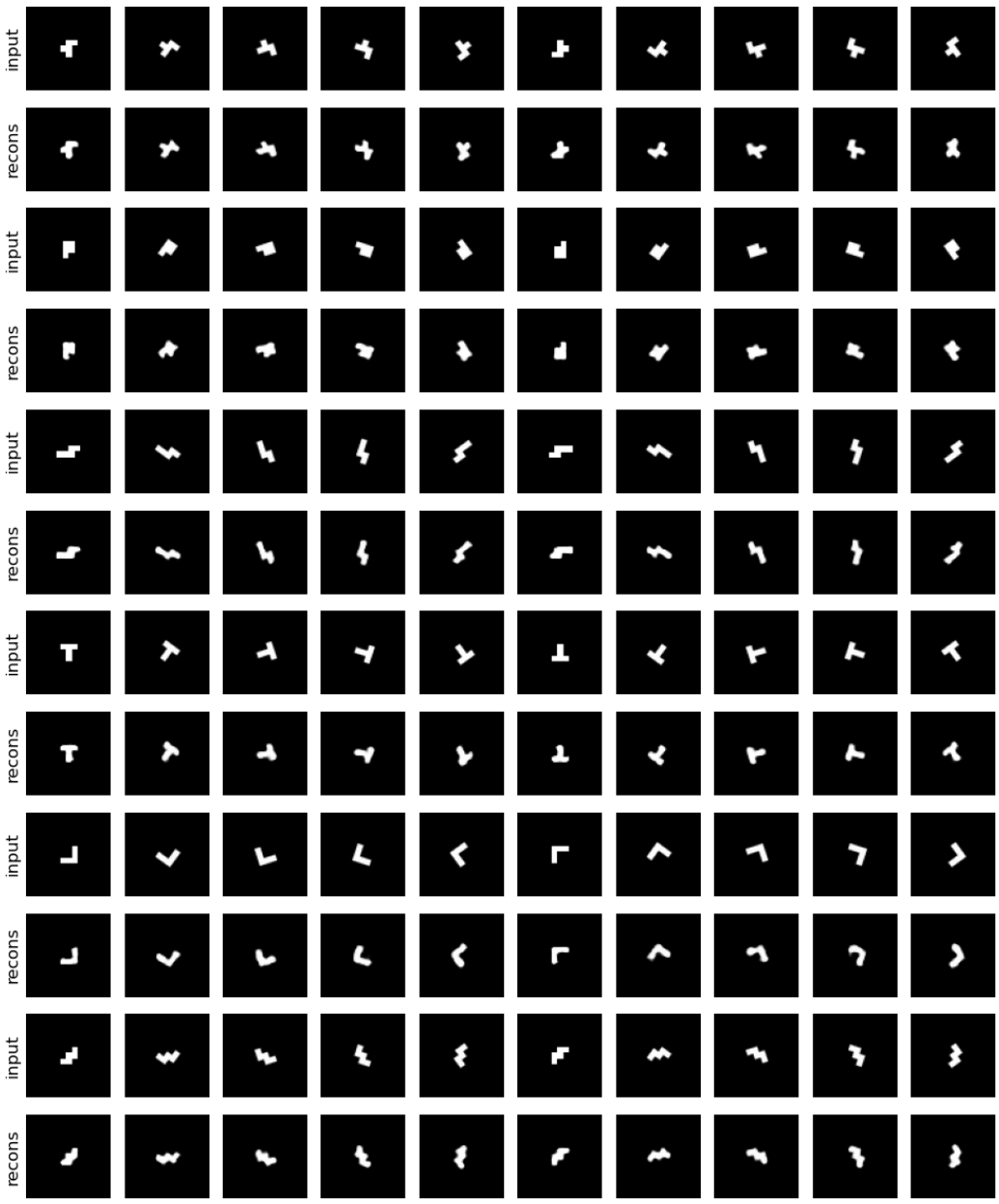

Figure 8: **Extrapolation to six new shapes**. Figure-ground Segmentation model reconstructions when six shapes are excluded from training (F, P, N T, V, W). Every pair of consecutive rows contains images of one of the shapes at 10 different rotation values. These are taken at evenly spaced angels from 180° to 360° and plotted increasingly from left to right. The rest of the generative factors are kept constant. Unlike the previous example, model reconstructions start to show significant degradation.

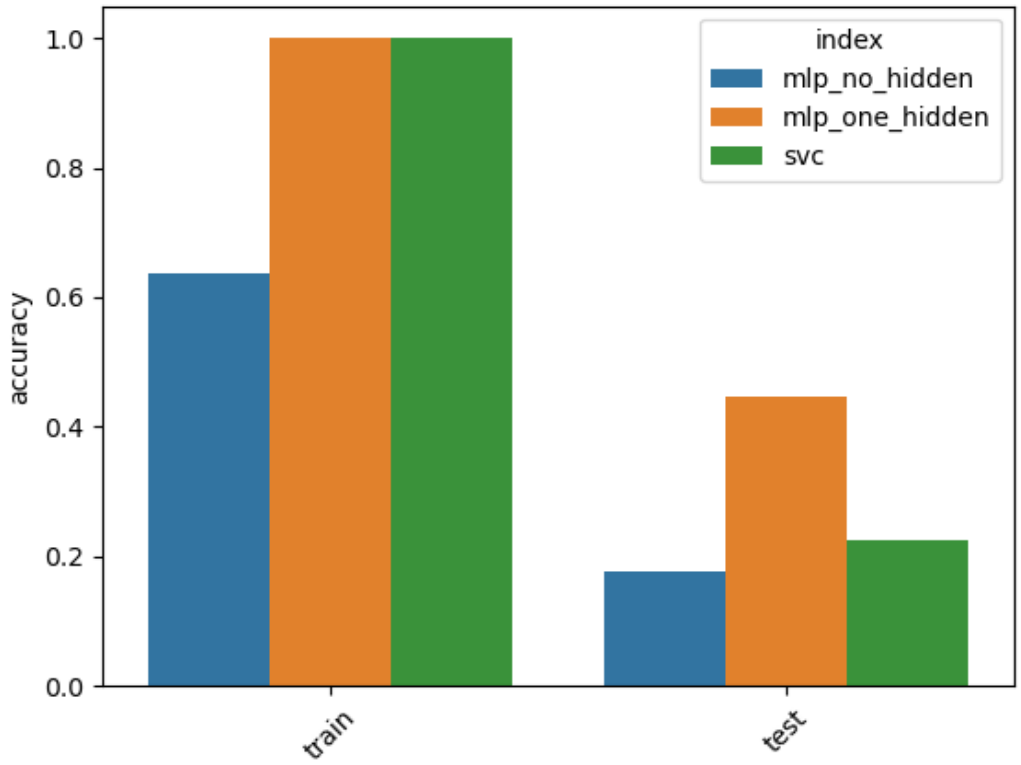

Figure 9: **Testing abstract representations**. On the left accuracy of three different probing models when prediction the shape of training images using the representations learned by a Slot Attention model: A simple Linear classifier trained with SGD, an MLP with one hidden layer and a Support Vector Machine. On the right, the same models now tested on prediction shapes for novel images using the slot embeddings obtained from the model.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: [Yes]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: [Yes] see Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: No code has been released.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will release the code upon acceptance at this or another venue.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [No]

   Justification: Method to generate some of the data is not disclosed. Model definition and training parameters are extracted from previous articles.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Preemiliminary results showed models achieved similar performance, so most other tests we ran once per model.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: There are no unreasonable training times.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: Will be released along with the code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.