Compositional Abilities Emerge Multiplicatively: Exploring Diffusion Models on a Synthetic Task

Maya Okawa [1 2] **Ekdeep Singh Lubana** [3 2 1] **Robert P. Dick** [3] **Hidenori Tanaka** [1 2]

Abstract

Modern generative models exhibit unprecedented capabilities to generate extremely realistic data. However, given the inherent compositionality of real world, reliable use of these models in practical applications mandates they exhibit the ability to compose their capabilities, generating and reasoning over entirely novel samples never seen in the training distribution. Prior work demonstrates recent vision diffusion models exhibit intriguing compositional generalization abilities, but also fail rather unpredictably. What are the reasons underlying this behavior? Which concepts does the model generally find difficult to compose to form novel data? To address these questions, we perform a controlled study of compositional generalization in conditional diffusion models in a synthetic setting, varying different attributes of the training data and measuring the model's ability to generate samples out-of-distribution. Our results show that: (i) the compositional structure of the data-generating process governs the order in which capabilities and an ability to compose them emerges; (ii) learning individual concepts impacts performance on compositional tasks, multiplicatively explaining sudden emergence; and (iii) learning and composing capabilities is difficult under correlations. We hope our study inspires further grounded research on understanding capabilities and compositionality in generative models from a data-centric perspective.

1. Introduction

The scaling of data, models, and compute has unleashed an array of powerful capabilities in generative models, enabling controllable synthesis of realistic images (Pan et al., 2023; Saharia et al., 2022b; Yu et al., 2022), 3D scenes (Richardson et al., 2023; Lim et al., 2023; Huang et al., 2022a), videos (Mei & Patel, 2022; Ceylan et al., 2023; Shin et al., 2023), accurate image-editing (Ravi et al., 2023; Couairon et al., 2022; Brooks et al., 2022), and semantically coherent text generation (Nijkamp et al., 2022; Zheng et al., 2023; Cassano et al., 2023). With increased interest to incorporate these models in our daily lives (Vemprala et al., 2023; Globe., 2023; Riera et al., 2020; Roose, 2023), e.g., to improve robotic systems via better planning and grounding (Janner et al., 2022; Singh et al., 2022; Chi et al., 2023; Brehmer et al., 2023; Huang et al., 2022b; Liu et al., 2023), the question of their reliability is becoming crucial. For these models to be beneficial to society, we argue concerted efforts are needed to understand the limitations of capabilities already existent within them and how these capabilities can be controlled.

Motivated by the above, in this paper, we perform a study of compositional generalization in conditional diffusion models, i.e., diffusion models that are conditioned on auxiliary inputs to control their generated images (e.g., text-conditioned diffusion models (Nichol et al., 2021; Kawar et al., 2022)). Given the inherent compositionality of the real world, it is arguably difficult to ever create a training dataset that allows the model to see all possible combinations of different concepts. Correspondingly, we argue the ability to compositionally generalize can be central to a model's reliability in out-of-distribution scenarios, i.e., when the model has to reason about data distributions it has never seen before (Zhang et al., 2021; Kaur et al., 2022). Sharing this motivation, several prior works have tried to probe the compositional generalization capabilities of off-the-shelf text-conditioned diffusion models (Marcus et al., 2022; Leivada et al., 2022; Conwell & Ullman, 2022; 2023; Gokhale et al., 2022; Du et al., 2023; Liu et al., 2022; Rassin et al., 2022; Feng et al., 2022). These works demonstrate that diffusion models can often compose rather complicated concepts, producing entirely non-existent objects, but can

[1] Physics & Informatics Lab, NTT Research, Inc., CA, USA [2] Center for Brain Science, Harvard University, MA, USA [3] University of Michigan, Ann Arbor, USA. Correspondence to: Maya Okawa <maya.okawa@ntt-research.com>, Hidenori Tanaka <hidenori.tanaka@ntt-research.com>.

Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

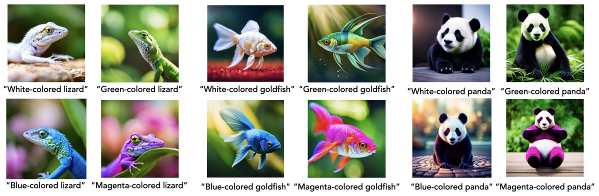

Figure 1. **(Lack of) Compositionality in text-conditioned diffusion models.** Images generated using Stable Diffusion v2.1 (AI., 2023b). (a) Diffusion models conditioned on text descriptions describing the concepts in an image often allow generation of entirely novel concepts that are unlikely to present in the training data, indicating an ability to compose learned concepts and generalize out-of-distribution. (b) However, arguably similar prompts show the model can unpredictably fail to compose its learned concepts at times, indicating its abilities to compose is dependent on precisely which concepts are being combined together. For example, generations of Panda in the above figure are difficult for the model, likely because a panda is less likely to be seen in different colors. The model seemingly chooses to alter the background or lighting to induce color alteration to some extent.

also unpredictably fail at composing arguably similarly complicated concepts (see Fig. 1). It remains unclear precisely what drives a model's ability to compositionally reason about some specific concept and yet miserably fail at another one. Indeed, what properties differentiate the concepts that the model learns to compose versus ones that it does not? We argue precisely answering these question requires a data-centric study, where the model's training data is systematically altered to observe exactly when an ability to compose a given set of concepts emerges.

In our work, we design a synthetic experimental setup that adheres to the principle of pursuing simplicity and controllability while preserving the essence of the phenomenon of interest, i.e., compositional generalization. Specifically, our data-generating process tries to abstract training data used in text-conditioned diffusion models by developing pairs of images representing geometric objects and tuples that denote which concepts are involved in the formation of a given image (see Fig. 2). We train diffusion models on synthetic datasets sampled from this data-generating process, conditioning the model on tuples denoting which concepts an object in the image should possess, while systematically controling the constitution of the dataset to alter the frequency of a given concept. Thereafter, we study the model's ability to generate samples corresponding to a novel combination of concepts by conditioning the denoising process on a correspondingly novel tuple, thus assessing the model's ability to compositionally generalize. This approach allows us to systematically investigate key configurations of a dataset that enable compositional generalization in an interpretable and controlled manner in conditioned diffusion models.

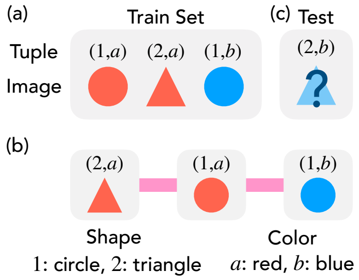

Figure 2. **Compositionality in a minimalistic conditional generation task.** (a) We train diffusion models on pairs of images and tuples, where the tuples denote which specific *concepts* compose an image (e.g., color and shape in the figure). (b) When only a single element differs between two tuples, a model can ideally learn the *capability* to recognize and alter the identifying concepts that distinguish the corresponding image pairs. (c) To test the existence of such capabilities and the model's ability to compose them, we ask the model to generate images corresponding to novel tuples that are out-of-distribution, hence requiring compositional generalization.

2. Concept Graph: A Minimalistic Framework for Compositionality

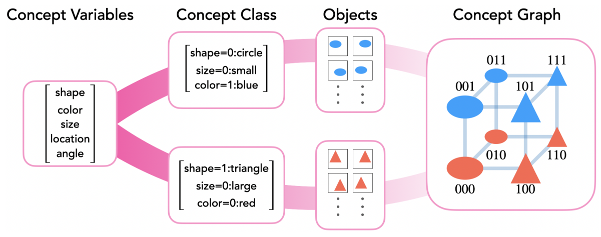

Figure 3. **Concept graphs.** We organize our study in a simple but expressive framework called *concept graphs*. The basis of a concept graph is a set of primitives called *concept variables* (e.g., `shape`, `color`, etc.). A subset of these variables are instantiated with specific values to yield a *concept class*, e.g., $\{\texttt{shape} = 0, \texttt{size} = 0, \texttt{color} = 1\}$ implies a small, blue circle. This is akin to defining a broad set of objects that share some common properties, such as all lizards of different color in Fig. 1 belong to the species of lizards. A specific *object* in this class is instantiated by filling the remaining variables; e.g., small, blue circles at different locations. Each concept class corresponds to a graph node, where nodes are connected if their concept classes differ by a *concept distance* of 1.

In this section, we present the *concept graph* framework, as illustrated in Fig. 3, which enables us to visually depict the minimal compositional structure of our synthetic data. Inspired by theories of concept learning in cognitive science (Margolis & Laurence, 1999) and object-oriented programming (Wikipedia., 2023), this framework forms our basis for generating hypotheses and designing experiments. We begin by defining the essential building blocks of our framework: concept variables and concept values. In the following, we call a specific output of our data-generating process an "object". For example, in Fig. 3, the images produced by the data-generating process form objects.

Definition 1. ***(Concept Variables.)*** *Let $V = \{v_1, v_2, ..., v_n\}$ be a set of n concept variables, where each v_i represents a specific property of an object.*

For instance, for geometric objects shown in Fig. 3, concept variables could include `shape` , `color` , `size` , `location` , and `angle`. These variables take on values from a pre-specified range, called concept values, as noted next.

Definition 2. ***(Concept Values.)*** *For each concept variable $v_i \in V$, let $C_i = \{c_{i1}, c_{i2}, ..., c_{ik_i}\}$ be the set of k_i possible values that v_i can take. Each element of the set C_i is called a concept value. Further, given an object x, $v_i(x)$ returns the value taken on by the i^{th} concept variable in that object.*

For example, in Fig. 3, the concept values for the `shape` variable can be in the set $\{\text{circle}, \text{triangle}, \text{square}\}$, while for the `color` variable can be in the set $\{\text{red}, \text{blue}, \text{green}\}$, and so on. If there are n concept variables, with each variable v_i having k_i concept values, there can be as many as $\prod_{i=1}^{n} k_i$ distinct combinations of concept values across the n concept variables. We use a unique combination of a pre-defined subset of these concept variables, $v_1, \ldots, v_p$ (where $p < n$), to define the notion of a "concept class".

Definition 3. ***(Concept Class.)*** *A concept class C is an ordered tuple $(v_1 = c_1, v_2 = c_2, ..., v_p = c_p)$, where each $c_i \in C_i$ is a concept value corresponding to the concept variable v_i. If an object x belongs to concept class C, then $v_i(x) = c_i \, \forall i \in 1, \ldots, p$.*

Note that the remaining $n - p$ concept variables are free in the above definition and, when filled, would define a specific object. That is, a concept class represents a family of objects by specifying the values of a pre-defined subset of concept variables. For example, in Fig. 1, different colored lizards instantiate images (objects) from the species (concept class) of lizards; here, color of the lizard serves as a free variable. Similarly, in the geometric objects scenario of Fig. 3, a "small red circle" would be a concept class wherein the `shape` , `color` , and `size` variables have been assigned specific values, while specific objects will be images designed by further associating a precise value with the remaining concept variables of `location` and `angle` . Next, we introduce a notion of concept distance, which serves as a proxy to succinctly describe the dissimilarity between two concept classes.

Definition 4. ***(Concept Distance.)*** *Given two concept classes $C^{(1)} = (c_1^{(1)}, c_2^{(1)}, ..., c_n^{(1)})$ and $C^{(2)} = (c_1^{(2)}, c_2^{(2)}, ..., c_n^{(2)})$, the concept distance $d(C^{(1)}, C^{(2)})$ is defined as the number of elements that differ between the two concept classes:*

$$d(C^{(1)}, C^{(2)}) = \sum_{i=1}^{n} I(c_i^{(1)}, c_i^{(2)}),$$

where $I(c_{1i}, c_{2i}) = 1$ if $c_{1i} \neq c_{2i}$ and $I(c_{1i}, c_{2i}) = 0$ otherwise.

The concept distance quantifies the dissimilarity between two concept classes by counting the number of differing concept values. It is important to note that this distance serves only as a null model, as each axis represents distinct concept variables, and each of these variables can assume various possible concept values. We are now ready to define the notion of a concept graph, which provides a visual representation of the relationships among different concept classes (see Fig. 3).

Definition 5. ***(Concept Graph.)*** *A concept graph $G = (N, E)$ consists of nodes and edges, where each node $n \in N$ corresponds to a concept class, and an edge $e \in E$ connects two nodes n_1 and n_2 representing concept classes $C^{(1)}$ and $C^{(2)}$, respectively, if the concept distance between the two concept classes is 1, i.e., $d(C^{(1)}, C^{(2)}) = 1$.*

That is, a concept graph allows us to organize different concept classes as nodes in the graph, while edges denote pairs of concept classes that differ by a single concept value. An ideal conditional diffusion model, when trained on a subset of the nodes from this graph, should learn *capabilities* that allow it to produce objects from other concept classes. We formalize this as follows.

Definition 6. ***(Capability and Compositionality.)*** *Consider a diffusion model trained to generate samples from concept classes $\hat{C} = \{C_1, \ldots, C_T\}$. We define a capability as the ability to alter the value of a concept variable v_i to a desired value c_i. We say the model compositionally generalizes when it can compose its capabilities with the ability to generate samples from a concept class $C \in \hat{C}$ to produce samples from class $\widetilde{C}$ such that $d(C, C_i) \geq 1 \forall i \in \hat{C}$.*

The ideas above are best explained via Fig. 4(a). Specifically, assume we train a diffusion model on data from a subset of concept classes, i.e., a subset of nodes in the concept graph. To fit the training data, the model *may* learn the relevant capabilities to alter specific concept variables or might instead just memorize the training data, e.g., given samples from classes `0000` and `0001` in Fig. 4, the model may learn how to alter the fourth concept variable or just memorize the data. Models that just memorize the training data lack the capability to generate samples from out-of-distribution classes (e.g., `0010`). In contrast, capabilities would enable it to produce out-of-distribution samples starting from classes in the training data. In summary, our concept graph framework provides a systematic approach to representing and understanding a minimalistic compositional structure, allowing for an analysis and comparison of different learning

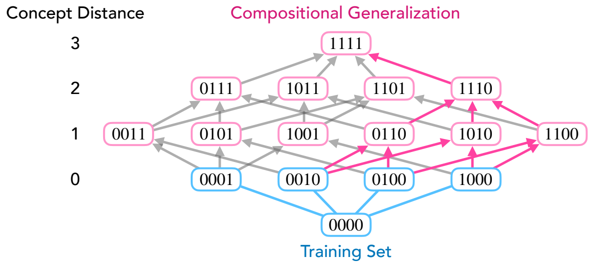

Figure 4. **Capabilities and compositionality in a concept graph.** Consider a lattice representation of a concept graph corresponding to four concept variables. Blue nodes denote classes represented in the training data; a model can either memorize data from these classes or learn capabilities to transform samples from one class to another. If it learns capabilities, it can compose them with data-generating process of samples from a concept class in the training data and produce samples that are entirely out-of-distribution, denoted as pink nodes.

algorithms' abilities to generalize across various concept classes.

3. Multiplicative Emergence of Compositional Abilities

With the concept graph framework in place, we now have the tools to systematically investigate the impact of different data-centric properties on the learning and composing of different capabilities in conditional diffusion models. Specifically, we aim to answer the following questions: (1) Can a model compositionally generalize to a concept class it has never encountered before?; (2) If so, under what circumstances does it fail?; (3) What is the order in which a model learns capabilities, and how does this process facilitate compositional generalization to out-of-distribution concept classes?

Learning dynamics respect the structure of the concept graph (Fig. 5). We first hypothesize that the ability to compositionally generalize and produce samples from out-of-distribution concept classes emerges at a rate which is inversely related to a class's concept distance with respect to classes in the training data. We empirically verify this claim in Fig. 5. Specifically, Fig. 5(a) shows the learning dynamics of the model, where lightblue nodes denote concept classes within the training dataset, while pink and darkpink nodes respectively denote classes at a concept distance of 1 and 2 from classes in the training dataset. As the model learns to fit its training data (lightblue nodes), it infers capabilities that can be composed to produce samples from concept classes entirely out of its training distribution (pink / darkpink nodes). As seen in Fig. 5 (b), we find that the learning dynamics of compositional generalization respect the concept distance from the training set: The model first memorizes the concept classes within the training dataset (lightblue lines) and then generalizes to concept classes with a concept distance of 1 from the training dataset (pink lines). Thereafter, the model suddenly acquires the capability to compositionally generalize to a concept class with a concept distance of 2 from the training dataset (darkpink line). Fig. 5 (c) shows the images generated by the model over time. We observe that rough shapes and sizes are learned relatively early in training, by the 4th epoch, while the color is dominantly biased to be red, the majority color in the training dataset, up to the 10th epoch. Then, around the 20th epoch, the model learns to generate the minority color (blue) for concept classes with a concept distance of 1. Finally, around the 40th epoch, the model learns to generate the minority color (blue) for the class at concept distance 2, showing a sudden emergence of capability to generate samples from that class.

Delayed emergence of abilities to generate minority colors for compositional generalization (Fig. 6). To better understand how the capability to generate minority colors is learned, we plot the accuracy of generated colors over training in Fig. 6. First, as expected, we observe that the model is capable of generating the majority color (red) much earlier than generating the minority color (blue). Importantly, as the concept distance of the given concept class for compositional generalization increases, the timing of generalization during training is further delayed. This observation provides important insights for training models with fairness in their design. Specifically, even once generalization for in-distribution concept classes is achieved, stopping the training of a model will likely lead to a failure in generating minority concepts, particularly for compositional generalization.

Multiplicity drives the sudden emergence of compositional abilities (Fig. 7). We leverage the interpretability of our experimental setup to illustrate how *multiplicity* is the critical mechanism behind the sudden emergence of compositional abilities. Fig. 7(a) depicts the learning dynamics of accuracy for generating concept class $\{111$, (triangle, small, blue)$\}$, which has a concept distance of 2 from the training data. We first observe a rather sudden occurrence of strong compositional generalization. To better comprehend this compositional ability, in Fig. 7(b), we plot the accuracy of a linear probe predicting each concept variable (shape, size, color). From the plot, we notice that the model struggles to acquire color transformation capabilities until the final stage of training, effectively bottlenecking compositional generalization to the '111'. This leads to the following intuitive understanding of *emergent behaviors in generative models*.

Remark. *(Multiplicative Emergence.) The nonlinear increase in capability observed in large neural networks as size and computational power scale up is driven by the task's compositionality. Models must learn all required*

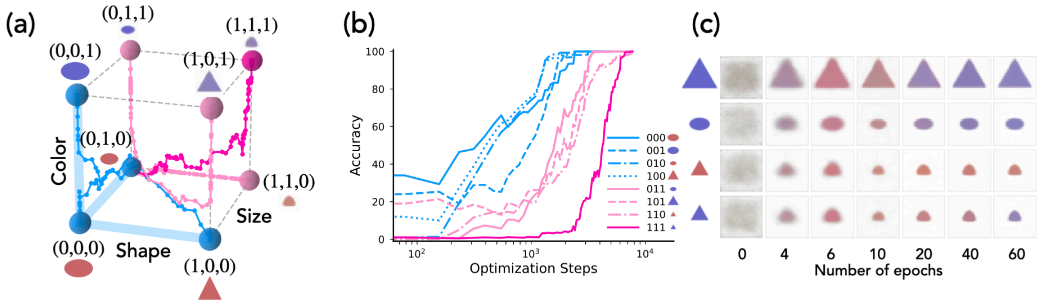

Figure 5. **Concept distance from the training set govern the order in which compositional capabilities emerge.** (a) Concept graph (cube) depicting training data points (blue nodes) and concept distances for test data points, where pink nodes represent distance = 1, and darkpink nodes represent distance = 2. Each trajectories represents a learning dynamics of generated images given each tuple prompt. Each trajectory represents the learning dynamics of generated images based on each tuple prompt. During every epoch of training, 50 images are generated, and binary classification is performed to predict each concept, including color, shape, and size. (b) Compositional generalization happens in sequence, starting with concept distance = 1 and progressing to concept distance = 2. The x-axis represents the number of epochs, and the y-axis represents the progress of compositional generalization. (c) Images generated as a function of time clearly show a sudden emergence of capability to change color for small, red triangles.

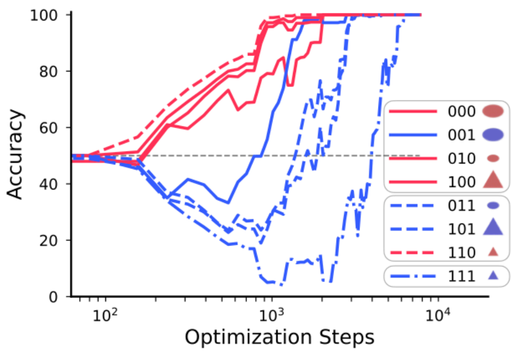

Figure 6. **Delayed emergence of abilities to generate minority colors for distant classes.** Prediction accuracy of color based on generated samples at each epoch during training. We observe that the ability to generate minority colors emerges significantly later as the concept distance of a concept class increases, highlighting the need for extended training beyond the point of achieving in-distribution generalization. This prolonged training enables effective composition of the minority concept and leads to improved generalization.

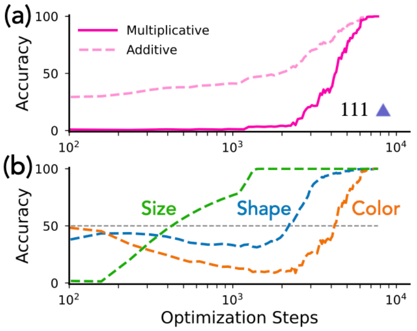

Figure 7. **Multiplicity underlies the sudden emergence of compositional capabilities.** (a) Accuracy of producing samples from the concept class {111, (triangle, blue, small)}, which has a concept distance of 2 from the training data. A multiplicative metric (solid line) assigns a score of 1 when all concept variables of shape, color, and size are correctly predicted. Conversely, an additive score (dashed line) provides partial credit for accurately predicting each of the concept variables independently, deceptively showing smooth progress (cf. (Barak et al., 2022; Nanda et al., 2023)). (b) Learning dynamics of accuracies for predicting each of the three concept variables: shape (blue), color (orange), and size (green).

concepts, but compositional generalization is hindered by the multiplicative, rather than additive, impact of learning progress on each concept. This results in a rather sudden emergence of capabilities to produce or reason about data not seen during training.

4. Conclusion

We introduce an abstraction of the training pipeline involved in training conditional diffusion models by designing a simple, interpretable, and yet powerful framework, titled *concept graphs*, that allow us to infer precisely when a model can learn the ability to compositionally generalize, producing novel, out-of-distribution samples. We show compositionality emerges in a sequence that respects the geometric structure of the concept graph, eliciting a *multiplicative emergence* effect that manifests as sudden increase in the model's performance to produce out-of-distribution data well after it has learned to produce samples from its training distribution. This behavior is reminiscent of the recently observed phenomenon of grokking in language modeling-like objectives (Power et al., 2022; Nanda et al., 2023; Barak et al., 2022). We further study settings where a model can fail to generalize compositionally, seeing phenomenon such as need for a critical amount of data to learn relevant capabilities.

References

AI., M. *Midjourney*, 2023a. https://docs.midjourney.com.

AI., S. *Stable Diffusion v2.1*, 2023b. https://beta.dreamstudio.ai/generate.

Andreas, J. Measuring compositionality in representation learning. *arXiv preprint arXiv:1902.07181*, 2019.

Barak, B., Edelman, B., Goel, S., Kakade, S., Malach, E., and Zhang, C. Hidden progress in deep learning: Sgd learns parities near the computational limit. *Advances in Neural Information Processing Systems*, 35:21750–21764, 2022.

Brehmer, J., Bose, J., De Haan, P., and Cohen, T. Edgi: Equivariant diffusion for planning with embodied agents. *arXiv preprint arXiv:2303.12410*, 2023.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Bugliarello, E. and Elliott, D. The role of syntactic planning in compositional image captioning. *arXiv preprint arXiv:2101.11911*, 2021.

Cassano, F., Gouwar, J., Nguyen, D., Nguyen, S., Phipps-Costin, L., Pinckney, D., Yee, M.-H., Zi, Y., Anderson, C. J., Feldman, M. Q., et al. Multipl-e: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Transactions on Software Engineering*, 2023.

Ceylan, D., Huang, C.-H. P., and Mitra, N. J. Pix2video: Video editing using image diffusion. *arXiv preprint arXiv:2303.12688*, 2023.

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., and Chan, W. Wavegrad: Estimating gradients for waveform generation. *ICML*, 2021.

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., and Song, S. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.

Conwell, C. and Ullman, T. Testing relational understanding in text-guided image generation. *arXiv preprint arXiv:2208.00005*, 2022.

Conwell, C. and Ullman, T. A comprehensive benchmark of human-like relational reasoning for text-to-image foundation models. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

Couairon, G., Verbeek, J., Schwenk, H., and Cord, M. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

Desai, K., Kaul, G., Aysola, Z., and Johnson, J. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Du, Y., Li, S., Sharma, Y., Tenenbaum, J., and Mordatch, I. Unsupervised learning of compositional energy concepts. *Advances in Neural Information Processing Systems*, 34: 15608–15620, 2021.

Du, Y., Durkan, C., Strudel, R., Tenenbaum, J. B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., and Grathwohl, W. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. *arXiv preprint arXiv:2302.11552*, 2023.

Eastwood, C. and Williams, C. K. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X. E., and Wang, W. Y. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

Frankland, S. M. and Greene, J. D. Concepts and compositionality: in search of the brain's language of thought. *Annual review of psychology*, 71:273–303, 2020.

Franklin, N. T. and Frank, M. J. Compositional clustering in task structure learning. *PLoS computational biology*, 14(4):e1006116, 2018.

Globe., B. *A Boston Dynamics robot can now be run using ChatGPT. What could go wrong?*, 2023. https://www.bostonglobe.com/2023/05/03/business/robots-can-now-be-run-using-chatgpt-what-could-go

Goel, V., Peruzzo, E., Jiang, Y., Xu, D., Sebe, N., Darrell, T., Wang, Z., and Shi, H. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv preprint arXiv:2303.17546*, 2023.

Gokhale, T., Palangi, H., Nushi, B., Vineet, V., Horvitz, E., Kamar, E., Baral, C., and Yang, Y. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022.

Goodman, N. D., Tenenbaum, J. B., Griffiths, T. L., and Feldman, J. Compositionality in rational analysis: Grammar-based induction for concept learning. *The probabilistic mind: Prospects for Bayesian cognitive science*, 2008.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. *In Proc. Int. Conf. on Learning Representations (ICLR)*, 2017.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Huang, I., Achlioptas, P., Zhang, T., Tulyakov, S., Sung, M., and Guibas, L. Ladis: Language disentanglement for 3d shape editing. *arXiv preprint arXiv:2212.05011*, 2022a.

Huang, W., Abbeel, P., Pathak, D., and Mordatch, I. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pp. 9118–9147. PMLR, 2022b.

Hupkes, D., Dankers, V., Mul, M., and Bruni, E. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67: 757–795, 2020.

Janner, M., Du, Y., Tenenbaum, J. B., and Levine, S. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Kaur, J. N., Kiciman, E., and Sharma, A. Modeling the data-generating process is necessary for out-of-distribution generalization. *arXiv preprint. arXiv:2206.07837*, 2022.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022.

Kim, H. and Mnih, A. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Kumari, N., Zhang, B., Wang, S.-Y., Shechtman, E., Zhang, R., and Zhu, J.-Y. Ablating concepts in text-to-image diffusion models. *arXiv preprint arXiv:2303.13516*, 2023.

Lake, B. and Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pp. 2873–2882. PMLR, 2018.

Leivada, E., Murphy, E., and Marcus, G. Dall-e 2 fails to reliably capture common syntactic processes. *arXiv preprint arXiv:2210.12889*, 2022.

Lepori, M. A., Serre, T., and Pavlick, E. Break it down: Evidence for structural compositionality in neural networks. *arXiv preprint arXiv:2301.10884*, 2023.

Lewis, M., Yu, Q., Merullo, J., and Pavlick, E. Does clip bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*, 2022.

Lim, J. H., Kovachki, N. B., Baptista, R., Beckham, C., Azizzadenesheli, K., Kossaifi, J., Voleti, V., Song, J., Kreis, K., Kautz, J., et al. Score-based diffusion models in function space. *arXiv preprint arXiv:2302.07400*, 2023.

Liu, A. Z., Logeswaran, L., Sohn, S., and Lee, H. A picture is worth a thousand words: Language models plan from pixels. *arXiv preprint arXiv:2303.09031*, 2023.

Liu, N., Li, S., Du, Y., Torralba, A., and Tenenbaum, J. B. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pp. 423–439. Springer, 2022.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proc. int. conf. on machine learning (ICML)*, 2019.

Marcus, G., Davis, E., and Aaronson, S. A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*, 2022.

Margolis, E. E. and Laurence, S. E. *Concepts: core readings*. The MIT Press, 1999.

Mei, K. and Patel, V. M. Vidm: Video implicit diffusion models. *arXiv preprint arXiv:2212.00235*, 2022.

Nanda, N., Chan, L., Liberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., Savarese, S., and Xiong, C. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*, 2022.

Pan, Z., Zhou, X., and Tian, H. Arbitrary style guidance for enhanced diffusion-based text-to-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4461–4471, 2023.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Phillips, S. and Wilson, W. H. Categorial compositionality: A category theory explanation for the systematicity of human cognition. *PLoS computational biology*, 6(7): e1000858, 2010.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rassin, R., Ravfogel, S., and Goldberg, Y. Dalle-2 is seeing double: flaws in word-to-concept mapping in text2image models. *arXiv preprint arXiv:2210.10606*, 2022.

Ravi, H., Kelkar, S., Harikumar, M., and Kale, A. Preditor: Text guided image editing with diffusion prior. *arXiv preprint arXiv:2302.07979*, 2023.

Reverberi, C., Görgen, K., and Haynes, J.-D. Compositionality of rule representations in human prefrontal cortex. *Cerebral cortex*, 22(6):1237–1246, 2012.

Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., and Cohen-Or, D. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023.

Riera, K., Rousseau, A.-L., and Baudelaire, C. *Doctor GPT-3: hype or reality?*, 2020. https://www.nabla.com/blog/gpt-3/.

Roose, K. *A Conversation With Bing's Chatbot Left Me Deeply Unsettled*, 2023. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022a.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022b.

Schott, L., Von Kügelgen, J., Träuble, F., Gehler, P., Russell, C., Bethge, M., Schölkopf, B., Locatello, F., and Brendel, W. Visual representation learning does not generalize strongly within the same domain. *arXiv preprint arXiv:2107.08221*, 2021.

Shin, C., Kim, H., Lee, C. H., Lee, S.-g., and Yoon, S. Edit-a-video: Single video editing with object-aware consistency. *arXiv preprint arXiv:2303.07945*, 2023.

Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., and Garg, A. Progprompt: Generating situated robot task plans using large language models. *arXiv preprint arXiv:2209.11302*, 2022.

Spilsbury, S. and Ilin, A. Compositional generalization in grounded language learning via induced model sparsity. *arXiv preprint arXiv:2207.02518*, 2022.

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Valvoda, J., Saphra, N., Rawski, J., Williams, A., and Cotterell, R. Benchmarking compositionality with formal languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6007–6018, 2022.

Van Steenkiste, S., Locatello, F., Schmidhuber, J., and Bachem, O. Are disentangled representations helpful for abstract visual reasoning? *Adv. in Neural Information Processing Systems (NeurIPS)*, 2019.

Vemprala, S., Bonatti, R., Bucker, A., and Kapoor, A. Chatgpt for robotics: Design principles and model abilities. *2023*, 2023.

Wikipedia. *Composition over Inheritance*, 2023. https://en.wikipedia.org/wiki/Composition_over_inheritance.

Xu, G., Kordjamshidi, P., and Chai, J. Prompting large pre-trained vision-language models for compositional concept learning. *arXiv preprint arXiv:2211.05077*, 2022.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. When and why vision-language models behave like bag-of-words models, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Yun, T., Bhalla, U., Pavlick, E., and Sun, C. Do vision-language pretrained models learn primitive concepts? *arXiv preprint arXiv:2203.17271*, 2022.

Zhang, D., Ahuja, K., Xu, Y., Wang, Y., and Courville, A. Can subnetwork structure be the key to out-of-distribution generalization? In *International Conference on Machine Learning*, pp. 12356–12367. PMLR, 2021.

Zheng, Q., Xia, X., Zou, X., Dong, Y., Wang, S., Xue, Y., Wang, Z., Shen, L., Wang, A., Li, Y., et al. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568*, 2023.

A. Related Work

Diffusion models. Diffusion models are the current state-of-the-art in generating extremely realistic visual data (Dhariwal & Nichol, 2021; AI., 2023a; Ho et al., 2020; Ho & Salimans, 2022; Nichol & Dhariwal, 2021; Nichol et al., 2021; Ramesh et al., 2022; 2021; Pan et al., 2023; Saharia et al., 2022b; Yu et al., 2022). Often these models are trained using image-text pairs, where the text is processed using a large language model to produce semantically rich embeddings that can allow controlled generation of novel images or even editing of existing ones (Kawar et al., 2022; Saharia et al., 2022a; Goel et al., 2023; Ravi et al., 2023; Couairon et al., 2022; Brooks et al., 2022). Such conditional diffusion models are easy to probe for compositional generalization, as one can directly specify a text description that requires composition of concepts the model is likely to know (e.g., avocado and chair) to produce images that are unlikely to exist in the model's training data (e.g., avocado chair; see (Dhariwal & Nichol, 2021; Nichol et al., 2021)). Such results demonstrate the model's ability to compose and generalize out-of-distribution. However, the use of text-conditioning also implies a possible failure to generalize compositionally can involve the text model being unable to properly represent desired concepts in the text-embedding space. To avoid this failure mode and only focus on the abilities of the diffusion process for image generation, in this work, we prefer to use ordered tuples that denote without ambiguity precisely which concepts are involved in a scene's composition.

Compositional generalization. Compositionality is an inherent property of the real world (Peters et al., 2017), wherein some primitive such as color can be composed with another primitive such as shape to develop or reason about entirely novel concepts that may not have been witnessed before (Zhang et al., 2021). It is especially hypothesized to play an integral role in human cognition, enabling humans to operate seamlessly in novel scenarios (Goodman et al., 2008; Phillips & Wilson, 2010; Frankland & Greene, 2020; Reverberi et al., 2012; Franklin & Frank, 2018). Inspired by this, several works in machine learning have focused on developing (Du et al., 2021; 2023; Liu et al., 2022; Xu et al., 2022; Yuksekgonul et al., 2022; Bugliarello & Elliott, 2021; Spilsbury & Ilin, 2022; Kumari et al., 2023) and benchmarking (Thrush et al., 2022; Andreas, 2019; Lewis et al., 2022; Lake & Baroni, 2018; Yun et al., 2022; Lepori et al., 2023; Johnson et al., 2017; Conwell & Ullman, 2022; Yuksekgonul et al., 2022; Schott et al., 2021; Gokhale et al., 2022; Valvoda et al., 2022) systems to respectively improve and analyze a system's ability to compositionally generalize. We note that a thorough formalization of compositionality in generative models is relatively lacking, though a noteworthy work includes the paper by Hupkes et al. (Hupkes et al., 2020).

B. Experimental and Evaluation Setup

Experimental Setup. The detailed setup is presented in Appendix B. In brief, we train conditional diffusion models that follow the U-Net pipeline proposed by Dhariwal and Nichol (Dhariwal & Nichol, 2021). Our dataset involves concept classes defined using three concept variables, each with two values; specifically, `shape` = $\{$circle, triangle$\}$, `color` = $\{$red, blue$\}$, and `size` = $\{$large, small$\}$. Tuples, which stand as an abstraction for text-conditioning, are used for conditioning the diffusion model's training and defined using binary numbers. For example, the tuple 000 implies a large, red circle is present in the image. To sample images from this process, we simply map the size and color axes to the range $[0, 1]$ and sample points in between to develop a training dataset of 5000 samples (the precise samples depend on which concept classes are allowed in the data-generating process). In this setup, the minimal required set for learning capabilities to alter concepts is just four pairs of tuples and images, each drawn from one of the following four concept classes: $\{000, (\text{circle}, \text{red}, \text{large})\}$, $\{100, (\text{triangle}, \text{red}, \text{large})\}$, $\{010, (\text{circle}, \text{blue}, \text{large})\}$, $\{001, (\text{circle}, \text{blue}, \text{small})\}$. By comparing the first elements of the sets $\{000, (\text{circle}, \text{red}, \text{large})\}$ and $\{100, (\text{triangle}, \text{red}, \text{large})\}$, we can observe that the concept of shape is encoded in the first element. Here, 0 represents a circle and 1 represents a triangle. Similar arguments can be made for the remaining elements in the sets.

Evaluation Metric. Evaluating whether a generated image corresponds to the desired concept class can require a human-in-the-loop. To circumvent this issue, we propose to follow literature on disentanglement which trains classifiers to test whether a generated image possesses some property of interest (Higgins et al., 2017; Kim & Mnih, 2018; Eastwood & Williams, 2018; Chen et al., 2018; Kumar et al., 2017; Van Steenkiste et al., 2019; Locatello et al., 2019). Specifically, we use the data used for training the diffusion model for training three linear classifiers that perform binary classification for each of the three concept variables, i.e., `shape` , `color` , and `size` . We define a model's accuracy for generating images of a given concept class as the product of the probabilities outputted by the three classifiers that each concept variable matches the value defined by the concept class. We also note that a random classifier of just one concept variable will predict the correct result with a 0.5 probability. We report this random baseline with dotted, gray lines in our plots whenever necessary.

B.1. Synthetic dataset

The dataset consists of 5,000 rendered images of 2D geometric shapes, along with corresponding concept classes. These synthetic images are generated using Blender[1] by creating a scene graph and rendering it. Each scene contains a single object placed on a blank background of size 28 × 28. The objects have three types of attributes: size, color, and shape. There are two shapes (circle and triangle), three colors (red, blue), and two sizes (large, small), resulting in up to eight different combinations of attributes. Each image is annotated with the corresponding object attributes, which can be utilized as conditional features for image generation. This enables us to directly evaluate the text-to-image generation capability of the diffusion model against ground truth images. Fig. 8 depicts example images with the corresponding concept classes.

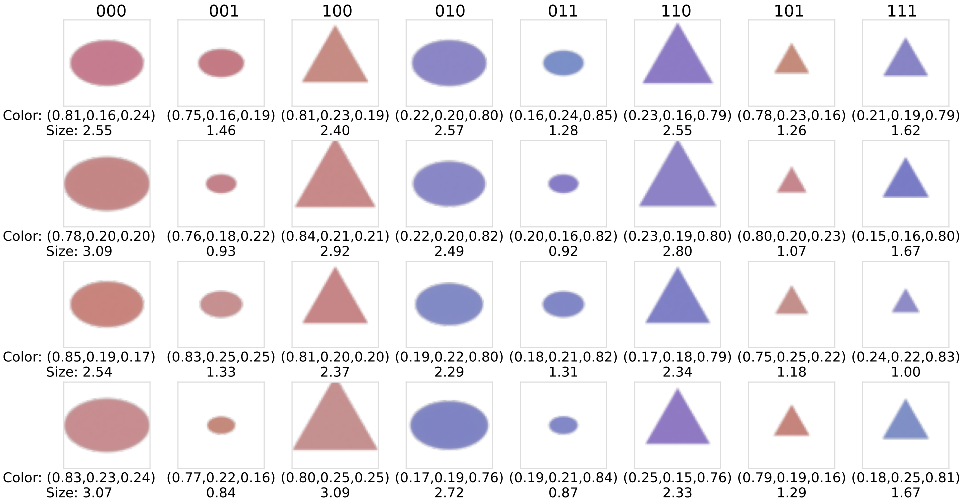

Figure 8. Examples of data samples with pairs of images and corresponding concept classes.

B.2. Loss function

Diffusion models convert Gaussian noise into samples from a data distribution through an iterative denoising process. The sampling process starts with a noisy input $\mathbf{x}_T$, and denoised samples are generated through gradual iteration, $\mathbf{x}_{T-1}$, $\mathbf{x}_{T-2}$, until the original input $\mathbf{x}_0$ is obtained. Conditional diffusion models (Chen et al., 2021; Saharia et al., 2022b) allow for a denoising process conditioned on texts or class labels. In all the experiments, we used the conditional diffusion model with the form $p(\mathbf{x}|V)$, where $\mathbf{x}$ denotes an image and $V = \{v_1, v_2, ..., v_n\}$ denotes a set of n concept variables. To predict the noise ϵ at each timestep $t \in [0, T]$, we follow the approach proposed in (Ho et al., 2020) by training a neural network. Specifically, we construct a neural network $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{a})$ and minimize the mean squared error (MSE) between the predicted Gaussian noise and the true noise:

$$\mathcal{L} = \mathbb{E}_{t \in [0,T], \mathbf{x}_0 \sim q(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})} \left[\|\epsilon - \epsilon_\theta(\mathbf{x}_0, t, V)\|^2 \right], \tag{1}$$

where $q(\mathbf{x}_0)$ denotes the distribution of input image $\mathbf{x}_0$, and $\mathcal{N}(\mathbf{0}, \mathbf{I})$ denotes the standard Gaussian distribution.

B.3. Architecture

We use the conditional U-Net architecture (Dhariwal & Nichol, 2021), as in (Ho et al., 2020), for our neural network $\epsilon_\theta(\cdot)$. Our architecture comprises two down-sampling and up-sampling blocks, with each block consisting of 3×3 convolutional layers, GELU activation, the global attention, and pooling layers. The conditional information V are fed through an embedding layer and concatenated with the image feature maps at each stage of the up-sampling blocks. We illustrate our network architecture in Fig. 9.

[1] http://www.blender.org

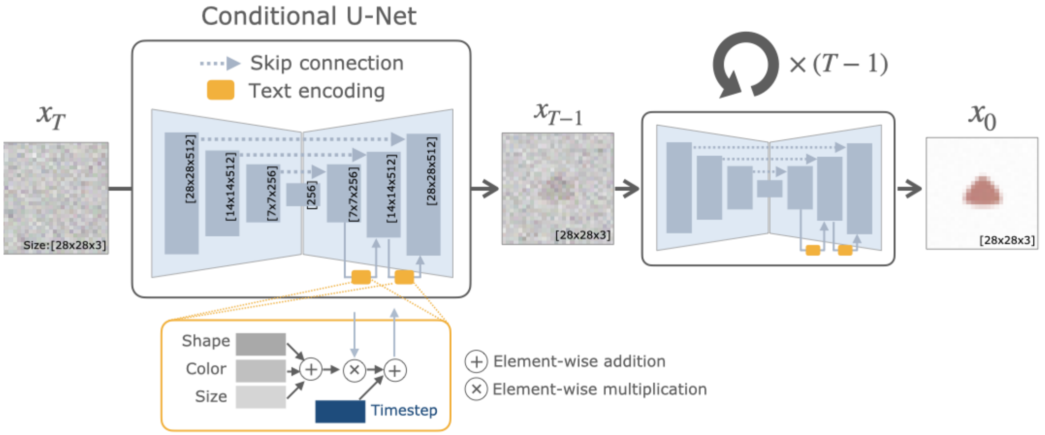

Figure 9. **The architecture of the conditional diffusion model.** The architecture of the conditional diffusion model involves an iterative process comprising noise addition and denoising steps. The model leverages conditioning information, specifically concept classes, to guide the transformation of the input image towards a desired state. In our implementation, we utilize a U-Net to parameterizethe denoising process. The U-Net architecture consists of three upsampling convolutional layers and three downsampling convolutional layers, which are connected through skip connections. Each layer within the U-Net includes a pooling layer, a global attention mechanism, and a GELU activation function.

B.4. Optimizer

We implemented the diffusion model using PyTorch and trained it on four Nvidia A100 GPUs. We performed a hyperparameter search based on a validation set. We tested batch sizes ranging from 32 to 256, the number of channels in each layer from 64 to 512, leaning rate from 10^{-4} and 10^{-3}, the number of steps in the diffusion process from 100 to 400. We employed the Adam optimizer (Kingma & Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and weight decay of 10^{-5}.

B.5. Evaluation metric

For the evaluation, we follow a probing protocol popularly used in several prior works on learning disentangled representations (Higgins et al., 2017; Kim & Mnih, 2018; Eastwood & Williams, 2018; Chen et al., 2018; Kumar et al., 2017; Van Steenkiste et al., 2019; Locatello et al., 2019). To evaluate the attributes of the images generated by our conditional diffusion model, we trained linear classifiers on these images for three specific attributes: shape, color, and size. For each attribute, we developed a dedicated classifier: $f_0(\hat{x}_0)$ for shape, $f_1(\hat{x}_0)$ for color, and $f_2(\hat{x}_0)$ for size. Here $\hat{x}_0$ denotes the image generated by the conditional diffusion model. We utilized a cross-entropy loss function to train these classifiers. The output of each classifier fell into one of two categories: for shape, the categories were circle or triangle; for color, blue or red; and for size, large or small. We then calculated the accuracy for each attribute using the corresponding classifier outputs. To quantitatively assess the accuracy of predicted attributes aligning with their corresponding ground-truth concept classes, we utilize a multiplicative measure. This measure gauses the accuracy of all attributes, and defined by the product of individual accuracies for each attribute as follows:

$$\text{Accuracy} = \frac{1}{N_t} \sum_{n=1}^{N_t} \nVdash\big(f_0(x_0^{(n)}), v_0^{(n)}\big) \cdot \nVdash\big(f_1(x_0^{(n)}), v_1^{(n)}\big) \cdot \nVdash\big(f_2(x_0^{(n)}), v_2^{(n)}\big), \tag{2}$$

where $\nVdash(\cdot)$ is the indicator function, n is the index of the test samples, and N_t is the total number of samples used for evaluation. For our experiments, we generated $N_t = 50$ images for each input of concept classes. $v_0^{(n)}$, $v_1^{(n)}$, and $v_2^{(n)}$ denote the actual (ground truth) concepts classes for shape, color, and size, respectively. We trained them over 50 epochs using the training dataset comprising 5,000 pairs of concept classes and images. The trained linear classifiers achieved an accuracy rate of 100% on the test set drawn from the original synthetic dataset.

C. Additional experimental results

C.1. Challenges for Compositional Generalization

We have demonstrated that, given a well-structured training dataset, a conditional diffusion model can learn to compose its capabilities to generate novel inputs not encountered in the training set. Next, we systematically investigate adversarial setups under which the model fails to learn relevant capabilities or an ability to compose them, failing to generalize compositionally. Understanding these limitations is as crucial as recognizing the successes, as they carry significant implications for limiting and controlling the emergence of harmful capabilities while preserving beneficial ones.

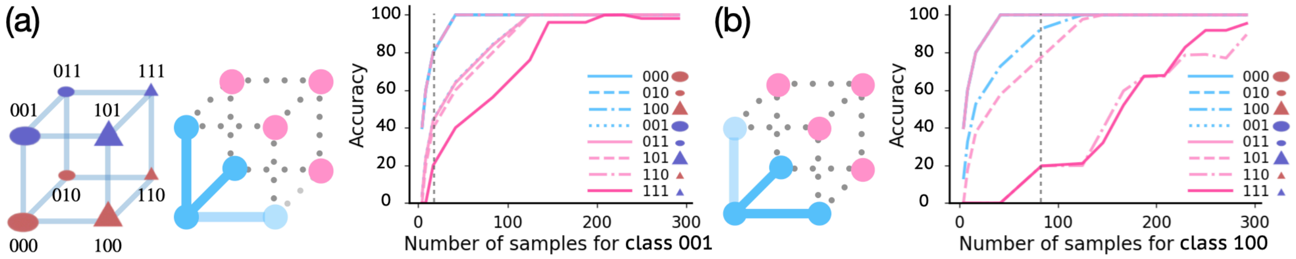

Figure 10. **How does the frequency of data samples impact the learning of capabilities?** We systematically control the frequency of a specific concept class in the training dataset and observe how it affects the model's learning of capabilities. (a) Capability to alter colors is quickly learned after introducing approx. 10 samples with a concept class of large blue circle, '001'. (b) In contrast, a critical threshold (marked with dotted vertical lines) exists for learning a capability to alter the shape: as we gradually introduce samples, with a concept class of large red triangle, '100'.

Critical frequency for learning capabilities (Fig. 10). We first systematically probe the effect of changing frequency of samples from different concept classes in the training data and examine how this affects the model's ability to learn and compose capabilities involving that concept class. Results are shown in Fig. 10 and demonstrate how the frequency of color and size concept in the training data impacts the generalization capabilities of the diffusion model. Specifically, we change the number of samples in the training data from 0 to 300 for concept class '001' (Fig. 10 (a)) and class '100' (Fig. 10 (b)). As can be seen, low frequencies of concepts degrade the accuracy of the model in both settings. Notably, training for out-of-distribution concept classes (pink lines) require more samples than that for in-distribution ones (lightblue lines). This suggests that as the sample size grows, memorization occurs first, and generalization is achieved beyond a certain threshold of data frequency. More importantly, we observe a critical number of samples are required before we can see the onset of capabilities to alter a concept. Specifically, in Fig. 10 (a), we can see that the model rapidly learns the color concept after being provided with 10 samples for a concept of large blue circle, '001'. In contrast, in Fig. 10 (b), the model learns the shape concept only after reaching a certain threshold in the number of samples with a concept of large red triangle, '100'.

We believe the results above are especially interesting because an often used strategy to prevent a generative model from learning harmful capabilities, such as the ability to generate images involving sensitive concepts like pornographic images, involves cleaning the dataset to filter images corresponding to such concepts (Desai et al., 2021; Brown et al., 2020; Radford et al., 2019). The hope is that this hinders the model's ability to generate samples corresponding to it. However, such dataset filtering can not only be expensive, but arguably statistically impossible to achieve to perfection, i.e., a few samples corresponding to the sensitive concept are likely to remain in the data. Our results above imply that perhaps one need not filter the data to an extreme zero presence of such sensitive concepts: if there presence in the training data is below the relevant critical threshold of frequency, that can be sufficient to deter the model from learning a capability to generate samples related to that concept.

Diffusion models struggles when concept variables are strongly correlated (Fig. 11). We next evaluate a setting where concept classes present in the training data are not neighboring, i.e., their concept distance is greater than 1, but nonetheless represent all concept variables to allow a model to learn relevant capabilities. Specifically, as shown in Fig. 11 (a), we use only the following four pairs of tuples and images for training: $\{000, (\text{circle}, \text{red}, \text{large})\}$, $\{100, (\text{triangle}, \text{red}, \text{large})\}$, $\{001, (\text{circle}, \text{blue}, \text{small})\}$, $\{111, (\text{triangle}, \text{blue}, \text{small})\}$. Arguably, capabilities corresponding to shape change (000 to 100) and color change (000 to 001) should be easy to learn since the setup for these is similar to our prior experiments. However, size change is observed only via samples from the class 111, and is necessarily observed in tandem with change in

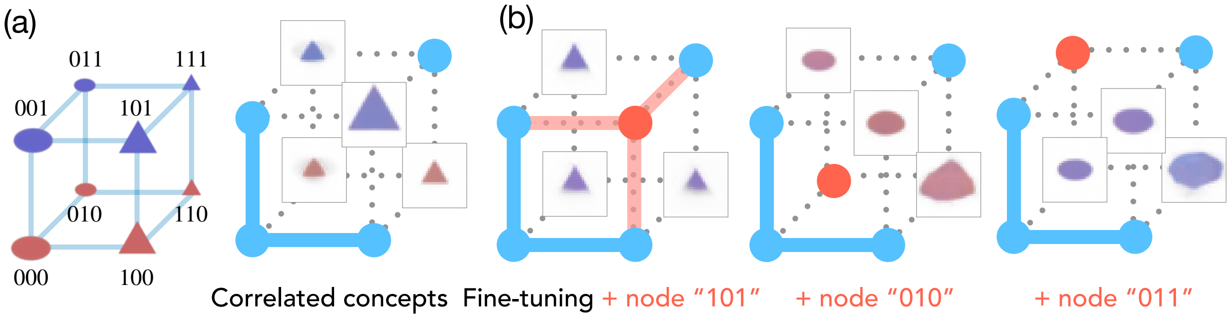

Figure 11. **Correlation in concept variables makes learning of capabilities difficult; fine-tuning is not a remedy.** (a) The model struggles when two concept variables, such as size and color, are perfectly correlated. In this example, large objects are red, and small objects are blue. In the example shown in (a), it must first compare 000 (circle, large, red) and 100 (triangle, large, red) to deduce that the first element specifies shape, and compare 000 (circle, large, red) and 001 (circle, large, blue) to deduce that the third element specifies color. Based on these findings, the model should then infer that 101 represents (triangle, large, blue) and by contrasting this with a training data point 111 (triangle, small, blue), learn that the second element specifies size. However, as seen in the plot, the model incorrectly associates the second element with the triangle shape and produces triangles for both 011 and 010, even though they should be circles. (b) The model faces difficulty learning new concepts through fine-tuning. We add node 101 (triangle, large, blue) to the dataset and attempt to fine-tune the model. However, even with a large learning rate equal to the one used for training, the model fails to learn the capability to alter the concept of size.

another concept (e.g., change in color and size co-occur as we move from 100 to 111). Correspondingly, the model has to perform an extra step of reasoning, disentangling size from other concept variables to learn the relevant capability. Our results show an interesting failure model of the model in this setting: Fig. 11 (a) shows the model struggles to dissociate the second element being 1 with the shape of a small triangle, and this strong, misinterpreted bias causes the model to generate small triangles for both '011' and '010', which should have been a small circle. This finding demonstrates that correlation in the data can be hard to disentangle for the model. The potential bias in the training data poses a significant challenge when applying the model in practical applications. If specific concepts are missing, the conditional diffusion models can have stereotypes and discrimination in the generated images. Given this clear failure mode, we now investigate whether it can be fixed via fine-tuning in Fig. 11 (b). To test this, we fine-tuned the trained model on a dataset that includes the concept class of '101' (left), '010' (middle), and '010' (right). However, as shown in Fig. 10 (b), the model still generates small triangles for '011' and '010' even after fine-tuning. When the concept classes '010' (middle) and '010' (right) are added, the newly introduced concepts overwrite all existing concepts, causing the previously learned concepts (e.g., the color and shape concept for '101') to be forgotten. These results present a further challenge in addressing the learned bias.