Investigating Same-Different Concept Understanding in Generative Multimodal Models

Sunayana Rane
Department of Computer Science
Princeton University
srane@princeton.edu

Declan Campbell
Princeton Neuroscience Institute
idcampbell@princeton.edu

Thomas L. Griffiths
Departments of Psychology and Computer Science
tomg@princeton.edu

Abstract

As advanced AI systems such as generative foundation models exhibit an increasingly rich range of behaviors, a challenge for AI alignment and safety research is how to effectively systematize and characterize these behaviors in a way that helps us understand and develop safer models. One key question on the path towards this goal is whether AI systems conceptually understand the world in the same way that humans do. A classic family of tasks used to probe concept understanding in humans and non-human animals is same/different tasks, which test for an understanding of the abstract concepts of "sameness" and "difference" across different stimuli. Taking inspiration from these studies of concept learning in humans and non-human animals, we present experimental results that investigate T2I model understanding of same/different concepts. We show that while T2I models demonstrate some understanding of same/different concepts, this understanding varies significantly across different attributes of sameness and difference (such as texture, color, rotation, and size). We discuss how revealing such behavioral differences can help us design more robust model training and evaluation protocols. Finally, we explain how analogies between behavioral analyses of concept learning in humans, non-human animals, and models can help us better understand the increasingly varied and often unpredictable behaviors that models exhibit.

1 Introductionn

An important question in AI alignment research is whether AI systems understand the world in the same conceptual way that humans understand it. [7] As generative multimodal models exhibit a wider range of behavioral capabilities, we might wonder how we measure such concept-level alignment or misalignment between these models and our expectations for human behavior. While surface-level numerical performance metrics may give us a high-level picture of the model's performance, with a richer range of behaviors, such metrics no longer guarantee that a model deeply understands the underlying principle behind an abstract concept. Fortunately, cognitive science has developed tools and experimental paradigms developed to behaviorally probe for such concept understanding in humans and other intelligent organisms.

Concept learning has long been recognized as a mechanism for higher-order cognitive abilities, and concepts as a building block for thought [5]. Cognitive scientists have extensively studied how human children and adults acquire various concepts, and we can use the same principles and experimental paradigms to investigate concept understanding in foundation models. [2] Same/different concept

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

understanding has a particularly rich history of study across various human and non-human animal species [10]. Understanding sameness and difference as abstract concepts was long seen as a uniquely human ability [4], but by further investigating properties of certain concept families, cognitive scientists began to decode some form of same/different concept understanding in other animal species as well. Today, a wide variety of animal species including pigeons, honeybees, and several species of monkey have been shown to demonstrate some understanding of same/different concepts [9, 3]. Human infants have also shown sensitivity to same/different concepts at 8 months of age [1]. While some work has been done to investigate same/different relation understanding in discriminative VLMs [8], testing these models requires visual input depicting sameness and difference, often in simplified or synthetic settings. Instead, generative multimodal models allow us to analyze their naturalistic visual output for advanced concept-understanding. In this work, we adapt the extensive study of the understanding of same/different concepts in cognitive science to probe analogous concept understanding in generative multimodal foundation models.

2 Methods

In these experiments, we tested the text-to-image (T2I) model DALL-E 3 on a variety of prompts centered on sameness and difference. The DALL-E family of models [6] are large foundation models with transformer-based architectures, which take a text prompt as input and generate an image as output.

In designing our prompts for the same/different task we presented to DALL-E 3, we varied one or more of the following attributes: height, width, texture, color, rotation.

One or more of these attributes is specified in the prompt, with a prefix modifier of "the same" or "different." Prompts would take the form "Render an image in photorealistic style containing four teacups that are of the same color, different texture, the same width, different rotation, arranged against a uniform background, each distinctly separated. Include only these objects in the image and nothing else." We used five common objects that generally come in different colors, to avoid a potential color bias: balloons, teacups, pens, candles, and hats. We also varied the number of these objects requested in the prompt, from two to six.

2.1 Human evaluations

To evaluate the generated output images, we recruited human evaluators on the Prolific platform, and provided them with the images and a corresponding check-box for each potential feature.

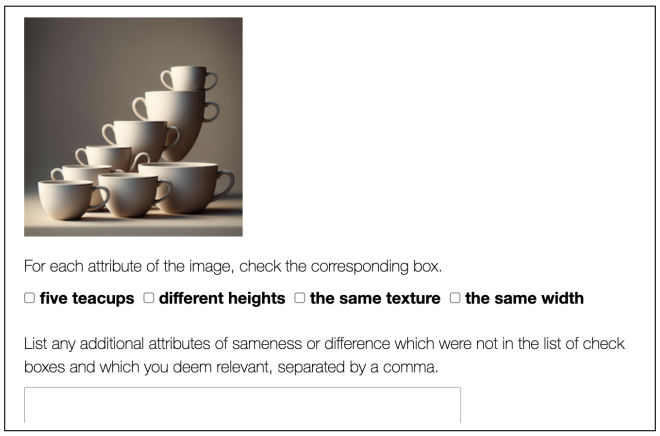

Figure 1: Human evaluation UI. Checkbox labels are created using the ground truth prompt.

2.2 Measuring errors

Studies of same/different concept understanding in humans and non-human animals have traditionally gone beyond a simple binary answer, and instead provided an in-depth look at how various factors

affect levels of concept understanding. In this spirit, this paper aims to provide a similarly in-depth look the type of errors these models make. We focus on two main types of errors measurement: edit distance and single-clause correctness. Single-clause correctness focuses on a binary correctness measurement of a single clause such as "different heights," and does not measure the number or correctness of the remaining clauses in the same prompt. Edit distance equals the total number of clauses in the prompt that were incorrectly generated by the model. Both of these error types are also investigated as they relate to feature entropy, that is, the number of total clauses in the prompt.

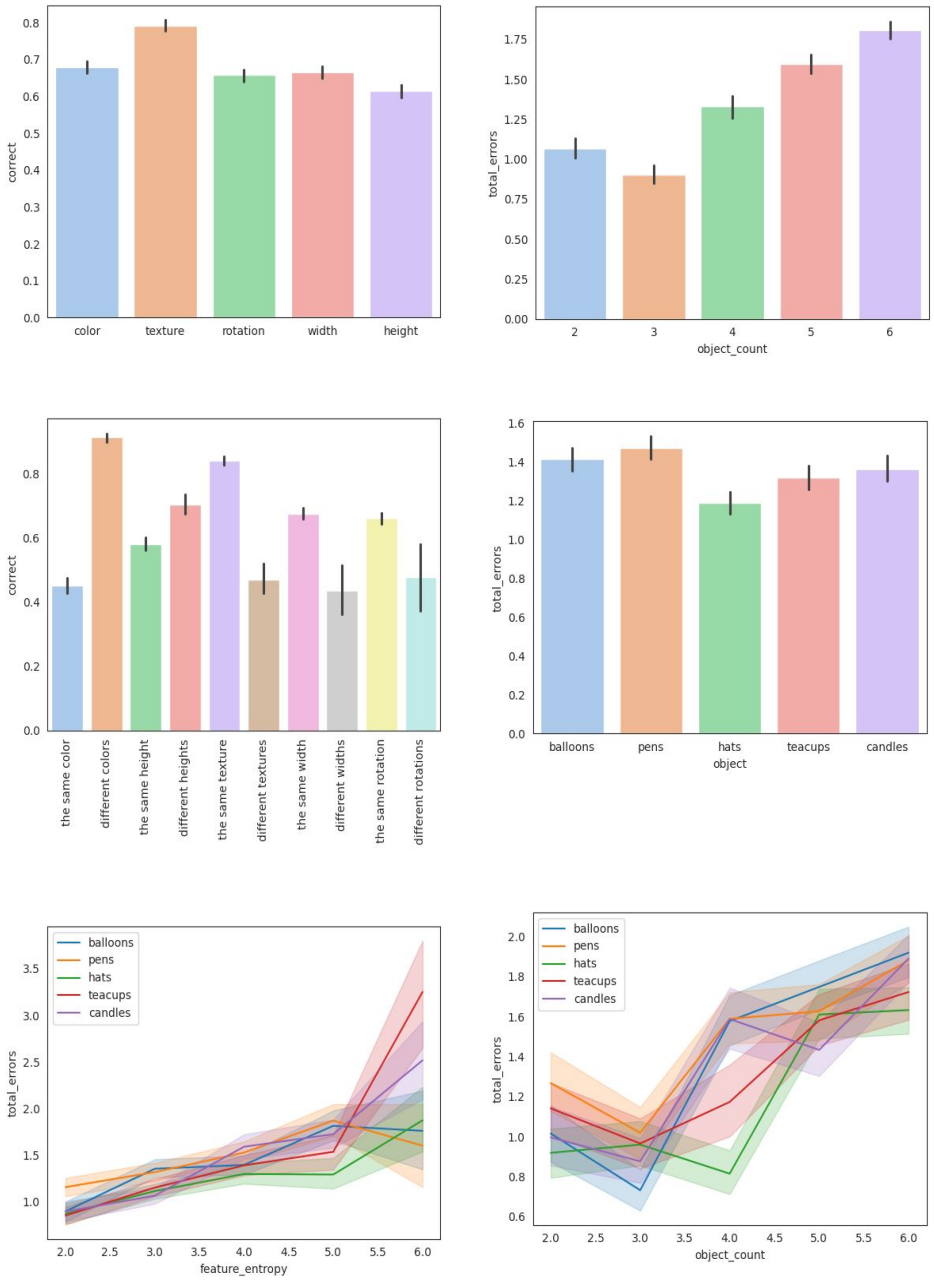

Figure 2: Numerical results for each of our error measurement types, across clause type, object type, object count, and feature entropy.

3 Results

See Figure 2 for an in-depth look at the relationships between each of our error types (single-clause correctness and edit distance/total errors) across clause type, object type, object count, and feature entropy. The total number of errors per prompt (edit distance) was positively correlated with feature entropy ($r(5888) = 0.250$, $p<.0001$) indicating that correctly rendering multiple same/different concepts simultaneously is a difficult task for our T2I model. Total errors also increased as object count increased, indicating that it is difficult to render same/different concepts correctly as the number of objects in the prompt increases ($r(5888) = 0.285$, $p<.0001$). Our results also show a significant effect of clause type (color, texture, rotation, width, height) on the single-clause error ($\chi^2(4) = 251.035$, p<.0001), and an even more significant variation when the granularity is increased to individual clauses such as "same texture" and "different rotations" ($\chi^2(9) = 1348.093$, p<.0001) (second row, first subplot of Figure 2). However, which type of concept (same or different) is easier for the model to learn varies across the clause types. For example, "the same color" is more difficult to learn than "different color," but "the same texture" is easier to learn than "different textures" (Figure 2).

4 Discussion

Our results show a high variability in the T2I model's demonstrated understanding of same/different concepts, and highlight the lack of a universally robust understanding of these concepts. For example, there is significant variation in single-clause correctness between and among model types. Of the same/different clause types, texture seems to be the easiest for the model to understand, while color, width, rotation, and height are more difficult. However, a more granular look at the subplots in Figure 2 show us that the performance also varies within subcategories of the clause types – for example, "different colors" is the easiest for the model to understand, while "the same color" is one of the most difficult. Conversely, "the same texture" shows excellent performance, while "different textures" does not. While further investigation is required to fully understand the reasons for this remarkable variation in ability across types of same/different concepts, it helps us begin to design better experiments to probe these questions. For example, if training data includes more examples of different colors than the same colors, or of the same textures than different textures, this is perhaps one easy way to begin to remedy the variation in performance. More broadly, however, it is important that models should demonstrate an understanding of the abstract concepts of sameness and difference in a way that can be disentangled from the particulars of the *type* of clause and the statistics of the dataset – our results make it clear that this has not yet happened.

While studies in animal cognition literature have found certain animal species demonstrate a fairly robust concept understanding of a single same/different relation at a time, it is less clear how robust that understanding would be when tested on multiple same/different attributes at the same time [10]. Furthermore, studies have found that some species such as pigeons are easily confused by the introduction of multiple different objects within a 16-object same/different task instead of just one type of object [10]. Although our results don't demonstrate precisely the same effect, it is worth considering if the mechanism might be similar, and if we can learn something from this failure mode to improve model performance and concept understanding. After all, models theoretically have fewer computational limits than pigeons, and yet they exhibit their own related failure models related to having many different same/different clauses in the same prompt. Conversely, perhaps the lack of pressures from such computational constraints contribute to the lack of learning an abstraction of a concept, because it allows models to focus on many statistical clues instead of distilling something more fundamental about sameness and difference. Looking back at the foundations for the experimental paradigms that led to same/different becoming a classic task in concept understanding can thus help us better design and implement experimental paradigms of our own for model behavior. This work takes a first step in that direction, by empirically demonstrating this principle through an analysis of same/different concept understanding in multimodal foundation model behavior.

References

[1] Caspar Addyman and Denis Mareschal. "The perceptual origins of the abstract same/different concept in human infants". In: *Animal Cognition* 13 (2010), pp. 817–833.

[2] Susan Carey. "The Origin of Concepts". In: *Journal of Cognition and Development* 1.1 (2000), pp. 37–41.

[3] Martin Giurfa. "Learning of sameness/difference relationships by honey bees: performance, strategies and ecological context". In: *Current Opinion in Behavioral Sciences* 37 (2021), pp. 1–6.

[4] Jeffrey S Katz and Anthony A Wright. "Issues in the comparative cognition of same/different abstract-concept learning". In: *Current Opinion in Behavioral Sciences* 37 (2021), pp. 29–34.

[5] Stephen Laurence and Eric Margolis. *Concepts: Core Readings*. MIT Press, 2000.

[6] Aditya Ramesh et al. "Zero-shot text-to-image generation". In: *International Conference on Machine Learning*. 2021, pp. 8821–8831.

[7] Sunayana Rane et al. "Concept Alignment". In: *arXiv preprint arXiv:2401.08672* (2024).

[8] Alexa R Tartaglini et al. "Deep neural networks can learn generalizable same-different visual relations". In: *arXiv preprint arXiv:2310.09612* (2023).

[9] Anthony A Wright and Jeffrey S Katz. "Mechanisms of same/different concept learning in primates and avians". In: *Behavioural Processes* 72.3 (2006), pp. 234–254.

[10] Thomas R Zentall et al. "Concept learning in animals". In: *Comparative Cognition & Behavior Reviews* 3.1 (2008), pp. 13–45.