A COMPREHENSIVE BENCHMARK OF HUMAN-LIKE RELATIONAL REASONING FOR TEXT-TO-IMAGE FOUNDATION MODELS

Colin Conwell & Tomer Ullman
Harvard University, Department of Psychology
{conwell,tullman}@{g,fas}.harvard.edu

ABSTRACT

Relations are basic building blocks of human cognition. Classic and recent work suggests that many relations are early developing, and quickly perceived. Machine models that aspire to human-level perception and reasoning should reflect the ability to recognize and reason generatively about relations. We report a systematic empirical examination of a recent text-guided image generation model (DALL-E 2), using a set of 15 basic physical and social relations studied or proposed in the literature, and judgements from human participants (N = 169). Overall, we find that only $\sim$22% of images matched basic relation prompts. Based on a quantitative examination of people's judgments, we suggest that current image generation models do not yet have a grasp of even basic relations involving simple objects and agents. We examine reasons for model successes and failures, and suggest possible improvements based on computations observed in biological intelligence.

1 INTRODUCTION

Consider the line 'the flooben was on the demaglis'. Even if you don't know what a *flooben* or *demaglis* are, you know something is *on* something[1]. This is because *on* is a basic relation. Our understanding of basic relations is general, early developing (18), and fundamental to our reasoning (49). There is also growing evidence that basic relations are perceived as directly as basic object properties (15). Machines that attempt to capture elements of human reasoning would do well to accurately perceive such relations in images, and produce accurate images from such relations as input. Here, we propose a core set of relational reasoning prompts, and evaluate the accuracy of a recent text-to-image foundation model (DALL-E 2) on its ability to generate images that match them.

Recent advances in image synthesis have achieved seemingly remarkable success in producing arbitrary images from arbitrary text (e.g. 40; 37). A prompt such as 'a robot-cat wearing cool glasses, gazing at a supernova' produces images that look somewhat like a robot-cat, wearing cool glasses, gazing at a supernova. Such successes lead to the impression that these models understand the input as a human would, as a compositional combination of objects, properties, and relations.

Despite their successes, these models are not without their limitations: early analyses drew attention to failures (amongst others) in common sense reasoning, feature binding, and text generation (27; 8; 23), and the literature on this topic has since grown at a precipitous rate. To aid in addressing the failures of text-to-image models, the machine learning community has since proposed a number of 'visuolinguistic' benchmarks (31; 39; 51; 4; 13; 30; 23; 29) that test all sorts of multimodal competences. (A more comprehensive, up-to-date review of related work may be found in Section 5.)

In the current work, our intent is to provide a benchmark inspired less by automated machine learning methods, and more directly by human psychology: Specifically, we focus on a set of 15 basic relations previously described in the cognitive, developmental, or linguistic literature. The set contins both grounded spatial relations (e.g. 'X on Y'), and more abstract agentic relations (e.g. 'X helping Y'). The prompts are intentionally simple, without attribute complexity or elaboration.

Rather than rely on our own intuition for whether an image matches a given relation prompt, we surveyed the intuitions of 169 participants. The use of multiple relations and many participants allows

[1]As Alice remarks after reading the nonsense poem, Jabberwockey: *"Somehow it seems to fill my head with ideas – only I don't exactly know what they are! However, somebody killed something: that's clear, at any rate"*

a more nuanced examination of model performance than pass/fail judgements. The stimulus set, prompts, images, and participant data are all openly available at `https://osf.io/sm68h`.

2 EXPERIMENT

We designed our experiment to assess the fit between basic relations and the images generated by DALL-E 2, by presenting images and sentences to human respondents and asking them whether an image and sentence matched.

Based on core empirical work in the domains of cognitive, linguistic, and developmental psychology (3; 26; 45; 21; 43; 54; 11; 17; 52; 53; 12; 6; 14; 9; 16), we created a set of 15 relations (8 physical, 7 agentic). The physical relations were: *in*, *on*, *under*, *covering*, *near*, *occluded by*, *hanging over*, and *tied to*. The agentic relations were: *pushing*, *pulling*, *touching*, *hitting*, *kicking*, *helping*, and *hindering*. We created a set of 12 entities (6 objects, 6 agents) to engage in these relations. The objects were: *box*, *cylinder*, *blanket*, *bowl*, *teacup*, and *knife*. The agents were: *man*, *woman*, *child*, *robot*, *monkey*, and *iguana*. The objects were simple bodies or common items used in previous data-sets that study relations (e.g. 7; 20), or in psychophysics tasks (16), or both. The agents were human, human-like, or of interest to the AI community. The iguana was a novel, visually distinct subordinate category we included as a treat.

For each relation, we created 5 different prompts, by randomly sampling two entities five times. This resulted in 75 prompts total (15 relations x 5 samples). For some relations, we restricted the set of allowable entities as follows: (i) Physical relations involved two physical objects, (ii) *Covering* had *blanket* as the first entity, (iii) *In* had *box* or *bowl* as the second entity, (iv) Agentic relations had an agent as the first entity, and either an object or an agent as the second entity, (v) The relations *helping* and *hindering* exclusively involved two agents.

We submitted each prompt to the DALL-E 2 rendering engine, and obtained the first 18 images that resulted. In a small number of cases, the prompt was rejected as a policy violation (e.g. 'a man kicking a man'). In such cases, the second entity was replaced at random until no policy violation was encountered. Our final stimulus set consisted of 1350 images (75 prompts x 18 images).

These stimuli were evaluated by 169 online participants. (See Appendix A.1 for demographic details, and human subjects protocol). Participants were informed that they would be assessing a 'picture-drawing AI', by examining grids of images that an AI drew in response to a given sentence. In each trial, participants were shown 18 images, organized into a 3x6 grid, with the target prompt at the top. Participants were instructed to select all images in the grid that matched the prompt. (See Figure A.1 for a screenshot of a typical trial, and trial sampling procedures.)

3 RESULTS

Unless otherwise noted, results are reported with the following convention: arithmetic mean [lower 95% confidence interval, upper 95% confidence interval] across participants or trials.

Participants on average reported a low amount of agreement between DALL-E 2's images and the prompts used to generate them, with a mean of 22.2% [18.3, 26.6] across the 75 distinct prompts. Agentic prompts, with a mean of 28.4% [22.8, 34.2] across 35 prompts, generated higher agreement than physical prompts, with a mean of 16.9% [11.9, 23.0] across 40 prompts ($t_{Welch}(71.82) = -2.81, p < 8.41e^{-3}, \hat{g}_{Hedges} = -0.62\ [-1.08, -0.16]$). (See also Figure A.2).

Decomposing the broad categories of physical and agentic into constituent relations, we observe a range of human agreement scores, as shown in Figure 1. While it is difficult to say what criterion establishes whether DALL-E 2 'understands' a given relation, here we report comparisons to 3 thresholds: 0%, 25%, and 50% perceived agreement, averaged across participants. Holm-corrected, one-sample significance tests for each relation suggest all 15 relations have participant agreement significantly above 0% at $\alpha = 0.95$ ($p_{Holm} < 0.05$). However, only 3 relations entail agreement significantly above 25% (touching, helping, and kicking), and no relations entail agreement above 50%. This remains true even without correction for multiple comparisons.

Considering the results qualitatively, we note that even a (relatively) high average agreement may not indicate relational understanding, but rather an influence of the training set. For example, the

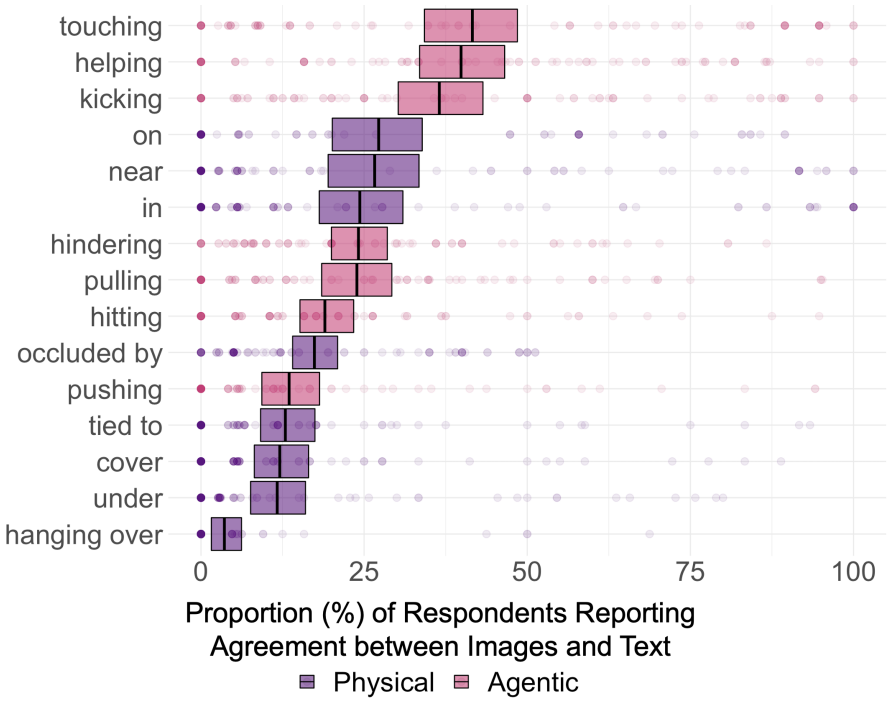

Figure 1: The proportion of participants reporting agreement between image and prompt, by the specific relation being tested. Points are the means of individual images, averaged across participants. There is a large range of reported agreement between image and text, though no relation entails average agreement significantly greater than 40%.

'touching' relation generated maximal average agreement (at a mean of 42% [34.3, 49.6] across 90 images), but with varied, bimodal success at the level of individual prompts. For example, the prompt 'child touching a bowl' generated 87% [80.1, 93] agreement on average, while 'a monkey touching an iguana' generated 11% [5.3, 19.7] agreement on average (see Figure A.3). It may be then that the combination of 'child' and 'bowl' is likely to generate images of a child touching a bowl simply given the training data. We consider this point further in the discussion.

While there are many factors that influence the quality of DALL-E 2's generated outputs, one particular parameter of interest is the CLIP score of the generated images: That is, the similarity (as determined by CLIP) between the generated image, and the text prompt used to generate that image (36). Intuitively, this is one of the parameters most responsible for the match between the target linguistic concept (in this case, a relation) and its depiction, but it's not necessarily a given that CLIP accounts for relations specifically. To examine the relationship between CLIP similarity and human perception, we used OpenAI's open-source ViT-L/14 model to calculate the similarity score between each image in our image set and their associated prompts. We then averaged the CLIP scores across the 18 images generated from each prompt, and correlated this average with the average perceived agreement provided by the human respondents. We found a moderate relationship between the two: $\hat{\rho}_{Spearman} = 0.39\ [0.17, 0.57], p = 5.5e^{-4}$ (and see Figure 2), suggesting CLIP is at least partially sensitive to the kinds of relations we've tested. (See Appendix A.3 for a demonstration of how CLIP similarity interacts with relation-type.)

4 Discussion

Relational understanding is a fundamental component of human intelligence, which manifests early in development (43), and is computed quickly and automatically in perception (15). DALL-E 2's difficulty with even basic spatial relations (such as *in, on, under*) suggests that whatever it has learned, it has not yet learned the kinds of representations that allow humans to so flexibly and robustly structure the world. A direct interpretation of this difficulty is that systems like DALL-E 2 do not yet

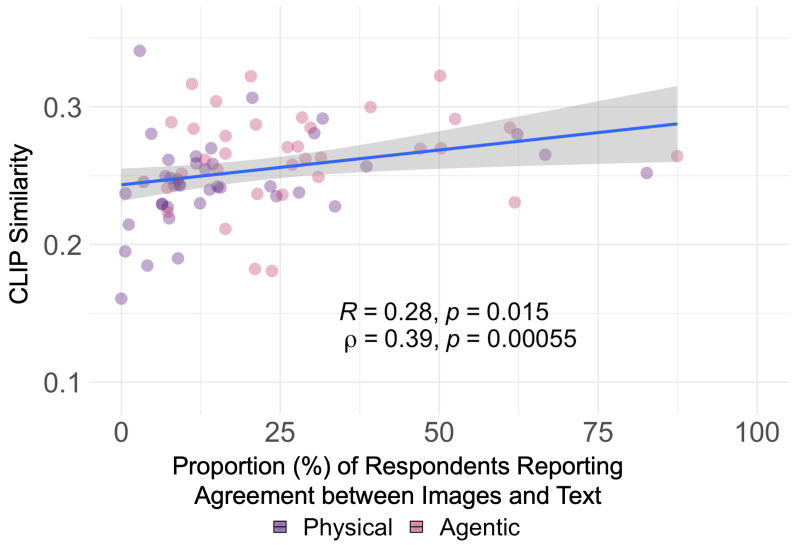

Figure 2: Relationship between CLIP (ViT-L/14) similarity scores and human agreement scores, averaged over images and participants. Each point is 1 / 75 prompts.

have relational compositionality. This is a point that has now been addressed by a variety of similiar works (27; 51), so we won't belabor it too intently here. (Though see A.4 for coverage of key points.)

More important than the failures themselves are the reasons for the failures, and how we might address them. There are many potential reasons for Dall-E 2's current lack of relational understanding, and they range from the minutia of technical implementation, to larger disjuncts between the computational principles underlying human intelligence and those underlying many current artificial intelligence systems (including almost all foundation models). One such disjunct is the way in which 'place' is explicitly coded for in both the generative image and text models that constitute text-guided image generation algorithms. Perhaps the only explicit encoding of relational order in such models is to be found in the positional embeddings of the text transformer in CLIP – effectively an auxiliary input that might easily be outweighed by the dozen or so nonlinear attention heads between them and the model's final outputs. This design choice (which in some cases produce models that function more or less as highly nonlinear bags-of-words (56)) is a marked difference from earlier iterations of natural language processing algorithms that provide syntactic parse trees in conjunction with the tokens corresponding to individual morphemes and words (48). At the level of images, there is an incompatibility between many modern machine vision algorithms – often designed *explicitly* to mimic the primate ventral visual stream – and the explicit representation of relations (spatial and otherwise) in the primate dorsal stream (46). Text-guided image generation algorithms might well benefit from mimicking algorithms in robotics (e.g. CLIPort 41), which combine CLIP's semantic flexibility with spatial transformers to model object identities and affordances simultaneously.

Another plausible upgrade that may boost model performance on relations are architectural adjustments that allow for multiplicative effects in a single layer of computation (44). These kinds of adjustments are inspired by biological perceptual systems, including the dorsal stream, that contain mixed selectivity neurons and lateral sub-circuits that facilitate the representation of interactions at multiple levels of the information-processing hierarchy (42; 10).

DALL-E 2 and other current image generation models are things of wonder, but they also leave us wondering what exactly they have learned, and how they fit into the larger search for artificial intelligence. DALL-E 2 has seemingly done what many models before it have failed to do, and bound the abstractions of natural language to clear points of perceptual reference. But that binding so far remains far more tenuous than the binding that defines the clear referents of standard human communication. The case of relational understanding provides a clear target for bringing an already meaningful advancement in artificial intelligence even closer to human meaning. While this analysis focuses exclusively on Dall-E, it can be replicated quickly and at low cost (see Appendix A.5) for similiar models (e.g. Stable Diffusion, Imagen, Parti, et cetera). *Any* foundation model that makes

a claim to meaningfully model the myriad, structured outputs of human perception and language should be able to reason about relations, and arguably *must* do so to instantiate further progress.

5 Related Work

The intense interest in the further development of DALL-E and other similiar text-to-image models (e.g. Parti (55), MUSE (2), EDIFFI (1), GLIDE (40), GALIP (50), and Stable Diffusion (38)) has produced a commensurately intense interest in formalizing and tracking the progress of these models in generating images that satisfy the semantic and syntactic specifications of their prompts (31; 39; 51; 4; 13; 30; 23; 32; 25). The work we have presented here is thus just one in an increasingly diverse mosaic of such works, taking as its singular focus the various kinds of relationships considered significant in developmental cognitive science, but sharing a number of commonalities with other benchmarks and probes of reasoning in text-to-image models. Below, we cover a sample of these works in a bit more detail.

In early analyses probing the limitations of text-to-image models, Marcus, Davis, and Aaronson gave Dall-E 2 several informal prompts that elicited failures in common sense, anaphora, relations, negation, and number (27). AI blogger 'Swimmer963' (47) reported informal tests along similar lines, and concluded DALL-E 2 has weaknesses with multiple characters, text, novel words, and foreground-background. Farid (8) pointed out the implausibility of cast shadows and reflections in DALL-E 2. Other limitations of text-to-image models were recognized by the developers of the models themselves. For example, Ramesh et al. described difficulties with binding, relative size, text, and other issues (Section 7 in 37). Saharia et al. proposed the 'DrawBench' benchmark, which includes a head-to-head comparison of the Imagen model to DALL-E 2, GLIDE (29), VQ-GAN-CLIP (5), and Latent Diffusion (38), on images from prior work that demonstrated failures with multiple counts, unorthodox color, positional arguments, rare words, and text generation.

Perhaps most closely related to the current work is Dall-Eval (4), a benchmark of text-to-image generative models that considers both 'physical reasoning and social biases'. Dall-Eval is comprehensive, probing a suite of inferential 'skills' ranging from object recognition and counting to '2D spatial relations' (some of which overlap with our own spatial relation probes), as well as common social biases (e.g. whether the otherwise unelaborated prompt 'a picture of a nurse' produces images that recapitulate the frequently gendered stereotypes of many professions). (These biases received even further attention in the 'Winoground' benchmark proposed by Thrush et al. (51)).

With our more intensive focus only on relations, the benchmark we have proposed here should be seen as complementary to many of these previous works, expounding especially on certain core agentic relations (e.g. helping or hurting) that other benchmarks consider only tangentially in the form of social bias. For those that are interested in building models that recapitulate many of the core compositional competences observed in human children, ours is the most directly linked (as far as we are currently aware) to developmental cognitive science.

Already we have seen how the focus on compositionality, in particular, has borne some fruit in the machine learning literature. Soon after Dall-E 2 was published, Liu et al. (24) proposed a composable diffusion model, and show that it outperforms other text-to-image models in the generation of structured images, by using basic conjunction (AND) and negation (NOT). Other models (e.g. Composer (19) and Ti2-Adapter (28)) have followed suit, and shown even further gains. More broadly, works like those of Li et al. (22) have suggested that human feedback (of the same variety used to condition LLM foundation models like ChatGPT) may be useful in guiding text-to-image models towards more reasonably human-like outputs.

Acknowledgments

We thank OpenAI for providing access to DALL-E-2 engine, and Jiayi Wang for her help in creating the stimuli. This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF1231216.

REFERENCES

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[2] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[3] Lin Chen. Topological structure in visual perception. *Science*, 218(4573):699–700, 1982.

[4] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022.

[5] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. *arXiv preprint arXiv:2204.08583*, 2022.

[6] Christian Dobel, Heidi Gumnior, Jens Bölte, and Pienie Zwitserlood. Describing scenes hardly seen. *Acta psychologica*, 125(2):129–143, 2007.

[7] Sébastien Ehrhardt, Oliver Groth, Aron Monszpart, Martin Engelcke, Ingmar Posner, Niloy Mitra, and Andrea Vedaldi. Relate: Physically plausible multi-object scene synthesis using structured latent spaces. *Advances in Neural Information Processing Systems*, 33:11202–11213, 2020.

[8] Hany Farid. Perspective (in) consistency of paint by text. *arXiv preprint arXiv:2206.14617*, 2022.

[9] Chaz Firestone and Brian Scholl. Seeing physics in the blink of an eye. *Journal of Vision*, 17 (10):203–203, 2017.

[10] Stefano Fusi, Earl K Miller, and Mattia Rigotti. Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology*, 37:66–74, 2016.

[11] Tao Gao, George E Newman, and Brian J Scholl. The psychophysics of chasing: A case study in the perception of animacy. *Cognitive psychology*, 59(2):154–179, 2009.

[12] Reinhild Glanemann, Pienie Zwitserlood, Jens Bölte, and Christian Dobel. Rapid apprehension of the coherence of action scenes. *Psychonomic bulletin & review*, 23(5):1566–1575, 2016.

[13] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022.

[14] Chenxiao Guan and Chaz Firestone. Seeing what's possible: Disconnected visual parts are confused for their potential wholes. *Journal of experimental psychology: general*, 149(3):590, 2020.

[15] Alon Hafri and Chaz Firestone. The perception of relations. *Trends in Cognitive Sciences*, 25 (6):475–492, 2021.

[16] Alon Hafri, Michael F Bonner, Barbara Landau, and Chaz Firestone. A phone in a basket looks like a knife in a cup: The perception of abstract relations. *PsyArXiv*, 2020.

[17] J Kiley Hamlin, Karen Wynn, and Paul Bloom. Social evaluation by preverbal infants. *Nature*, 450(7169):557–559, 2007.

[18] Susan J Hespos and Elizabeth S Spelke. Conceptual precursors to language. *Nature*, 430(6998): 453–456, 2004.

[19] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023.

[20] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 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.

[21] Philip J Kellman and Elizabeth S Spelke. Perception of partly occluded objects in infancy. *Cognitive psychology*, 15(4):483–524, 1983.

[22] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

[23] Evelina Leivada, Elliot Murphy, and Gary Marcus. Dall-e 2 fails to reliably capture common syntactic processes. *arXiv preprint arXiv:2210.12889*, 2022.

[24] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. *arXiv preprint arXiv:2206.01714*, 2022.

[25] Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M Dai. Mind's eye: Grounded language model reasoning through simulation. *arXiv preprint arXiv:2210.05359*, 2022.

[26] Andrew Lovett and Steven L Franconeri. Topological relations between objects are categorically coded. *Psychological science*, 28(10):1408–1418, 2017.

[27] Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*, 2022.

[28] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[29] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[30] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. *arXiv preprint arXiv:2304.01816*, 2023.

[31] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[32] Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*, 2022.

[33] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.

[34] Francis Jeffry Pelletier. The principle of semantic compositionality. *Topoi*, 13(1):11–24, 1994.

[35] Francis Jeffry Pelletier. Semantic compositionality. In *Oxford research encyclopedia of linguistics*. 2016.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

[39] Irene Russo. Creative text-to-image generation: Suggestions for a benchmark. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pp. 145–154, 2022.

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[41] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2021.

[42] R Angus Silver. Neuronal arithmetic. *Nature Reviews Neuroscience*, 11(7):474–489, 2010.

[43] Elizabeth Spelke. Initial knowledge: Six suggestions. *Cognition*, 50(1-3):431–445, 1994.

[44] Julia Steinberg and Haim Sompolinsky. Associative memory of structured knowledge. *bioRxiv*, 2022.

[45] Brent Strickland and Frank Keil. Event completion: Event based inferences distort memory in a matter of seconds. *Cognition*, 121(3):409–415, 2011.

[46] Christopher Summerfield, Fabrice Luyckx, and Hannah Sheahan. Structure learning and the posterior parietal cortex. *Progress in neurobiology*, 184:101717, 2020.

[47] Swimmer963. What dall-e 2 can and cannot do, 2022. URL `https://www.lesswrong.com/posts/uKp6tBFStnsvrot5t/what-dall-e-2-can-and-cannot-do#DALLE_s_weaknesses`.

[48] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

[49] Leonard Talmy. Lexicalization patterns: Semantic structure in lexical forms. *Language typology and syntactic description*, 3(99):36–149, 1985.

[50] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. Galip: Generative adversarial clips for text-to-image synthesis. *arXiv preprint arXiv:2301.12959*, 2023.

[51] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 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.

[52] Tomer Ullman, Chris Baker, Owen Macindoe, Owain Evans, Noah Goodman, and Joshua Tenenbaum. Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, 22, 2009.

[53] Benjamin van Buren, Stefan Uddenberg, and Brian J Scholl. The automaticity of perceiving animacy: Goal-directed motion in simple shapes influences visuomotor behavior even when task-irrelevant. *Psychonomic bulletin & review*, 23(3):797–802, 2016.

[54] Ilker Yildirim, Max H Siegel, and Joshua B Tenenbaum. Perceiving fully occluded objects via physical simulation. In *Proceedings of the 38th annual conference of the cognitive science society*, 2016.

[55] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[56] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bag-of-words models, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

A Appendix

A.1 Human Subjects Protocol + Demographics

We recruited 180 participants online (33) via the Prolific platform (`https://www.prolific.co`). Participants were restricted to those located in the USA, having completed at least 100 prior studies on Prolific, with an acceptance rate of at least 90%. The mean age of the participants was 33.8; 59% of participants identified as female, 40% identified as male, and one did not identify as either. Of this sample, 11 participants failed to pass two attention checks, and were removed from analysis, leaving 169 participants in the final sample.

A.2 Trial Sampling

The 10 prompts any given participant rated were randomly drawn from the full set of 75 prompts. This resulted in variability in the number of participants that evaluated any given image. The number of participants that rated a given image ranged from 15 to 43, with an average of 23. After participants finished evaluating 10 prompts, they were given another attention check, thanked for their time, and given an opportunity to provide feedback.

A.3 Further Details on CLIP Similarity + Relation Type

To assess more finely the combined influence of broad relation type ('agentic' or 'physical') and CLIP scores on the human-perceived match between text and image, we used two Bayesian multilevel (mixed-effects) models: a zero-inflated binomial model calculated directly over the participant-level choice data (with additive effects for relation type and CLIP score, plus random intercepts for subject and the order of image presentation [0-18]), and a zero-one-inflated beta model calculated over the average scores per image (again with additive effects for relation type and CLIP score, but with a random intercept for the order of image presentation alone). We use zero-inflated models in both cases, given the outsize quantity of images that participants labeled as not matching the target prompt. Controlling in both cases for variance injected by factors outside the study design (i.e. random effects), these models suggest small-to-midsize significant effects of both relation type and CLIP score on the probability of human respondents designating a target image as matching its prompt. Results from these regressions are summarized in Table 1.

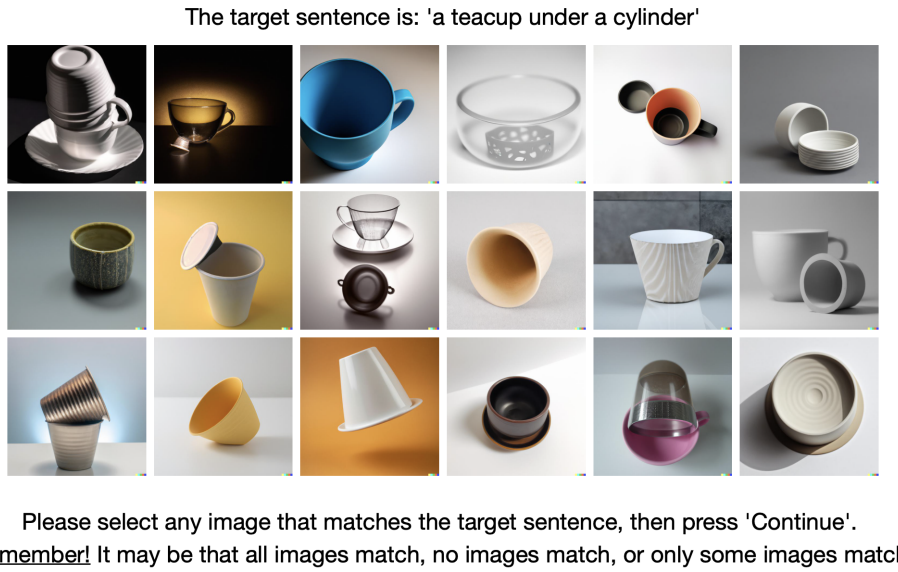

Figure A.1: Screenshot from a trial in our Experiment. Participants were presented with grids of images, and a sentence prompt. Participants selected images that matched the target sentence.

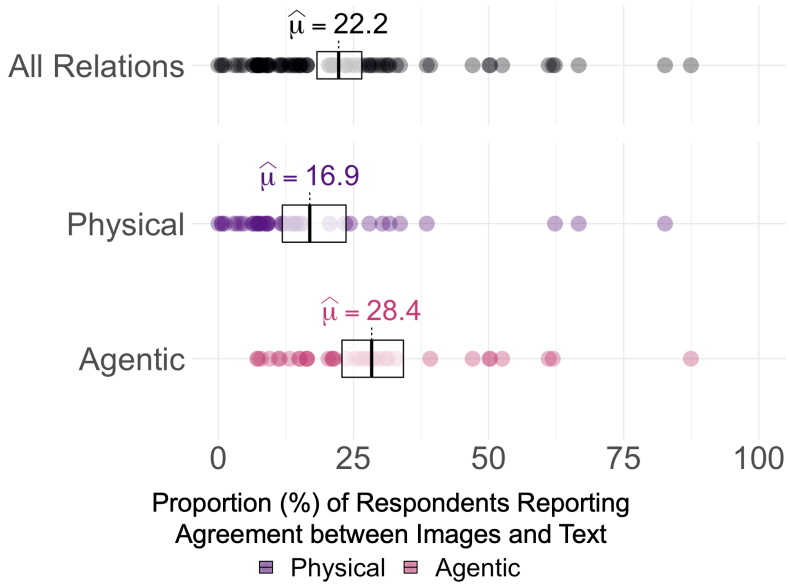

Figure A.2: Experiment results, participant agreement that images matched a prompt. Each point is an individual prompt. Points in black show all prompts. Points in color break down the prompts by whether the subject of the prompt was an object (physical) or agent (agentic).

	Zero-Inflated Binomial		**Zero-One-Inflated Beta**	
Predictors	*Odds Ratios*	*CI (95%)*	*Estimates*	*CI (95%)*
Intercept	0.29	0.22 – 0.38	0.40	0.36 – 0.44
RelationType: Agentic	2.13	1.93 – 2.36	1.41	1.22 – 1.62
ClipScore (Scaled)	1.59	1.50 – 1.71	1.15	1.07 – 1.24
Random Effects				
σ^2	3.29		1.00	
τ_{00}	0.04 $_{\text{ImageOrder}}$		0.00 $_{\text{ImageOrder}}$	
	0.47 $_{\text{SubjectID}}$			
ICC	0.13		0.00	
N	169 $_{\text{SubjectID}}$		18 $_{\text{ImageOrder}}$	
	18 $_{\text{ImageOrder}}$			
Observations	30398		1350	
Marginal R^2 / Conditional R^2	0.050 / 0.093		0.017 / 0.017	

Table 1: Results of two mixed effects regressions of relation type and CLIP score on human agreement, either at the individual subject level (zero-inflated binomial) or the image level (zero-one-inflated-beta).

A.4 Further Discussion: Failure of Compositionality in Image-Text Models

The notion that systems like DALL-E 2 do not have compositionality may come as a surprise to anyone that has seen DALL-E 2's strikingly reasonable responses to prompts like 'a cartoon of a baby daikon radish in a tutu walking a poodle'. Prompts such as these often generate a sensible approximation of a compositional concept, with all parts of the prompts present, and present in the

Figure A.3: Grids for two example prompts that probed the *touching* relation. While the average agreement was 42%, the underyling distribution of prompt responses was effectively bimodal, with e.g. the prompt ‘a child touching a bowl’ generating high agreement (87%), and ‘a monkey touching an iguana’ generating low agreement (11%).

right places. Compositionality, however, is not only the ability to glue things together – even things you may never have observed together before. Compositionality requires an understanding of the *rules* that bind things together. Relations are such rules.

To the extent that DALL-E 2 is only able to generate relations some of the time is the extent to which DALL-E 2 is actively *not* compositional. These failure cases are important, because they tell us something about the way DALL-E 2 is getting things *right*. The fact that DALL-E 2 seems able to easily generate ‘a spoon in a cup’, but not ‘a cup on a spoon’ (see Figure A.4), means that even when it is getting ‘a spoon in a cup’ right this is likely due to a great deal of prior exposure to images of spoons in cups, rather than an understanding of ‘in’ or ‘on’ – precisely the kinds of syntactic rules that define compositionality. Real compositionality should be invariant at the level of the relation, which is to say that ambiguity in meaning should come from the semantic elements involved in the relation, and not from the relation itself (34; 35).

In addition to effects of training data on apparent successes, it is possible that DALL-E 2’s slightly better performance with more abstract relations like ’helping’ is due to visual ambiguities, and the interpretive steps that people take on top of a given image. That is, when seeing an image of a robot touching another robot and the prompt ‘a robot helping a robot’, people may be thinking ‘Well, I guess this *could* be helping, if...’. This is a tentative suggestion, but it could be tested empirically by showing people images generated through prompts like ‘helping’ but without labeling, and having them either freely describe the image, or giving people a forced choice among several relations.

Figure A.4: Illustrative example, images generated given ‘a spoon in a cup’ and ‘a cup on a spoon’. Examining just the left images may lead to the conclusion that Dall-E 2 captures the *in* relation, but the right images suggest this is simply an effect of training images that involve *spoon* and cup.

Even with the occasional ambiguity, the current quantitative gap between what DALL-E 2 produces and what people accept as a reasonable depiction of very simple relations is enough to suggest a *qualitative* gap between what DALL-E 2 has learned, and what even infants seem already to know. This gap is especially striking given DALL-E 2’s staggering diet of image content.

A.5 Replicating these Benchmarks for Other Image-Text Models

While human labeling of images may seem a suboptimal method for benchmarking, we take our work as a demonstration that this is not necessarily the case. Our full sample of human participants (N = 180) was collected in approximately 36 hours, at a cost of $2.00USD per participant.

The prompts we use in this experiment are already freely available at `https://osf.io/sm68h`. Full Javascript code for the behavioral experiment will be made available via GitHub on publication.

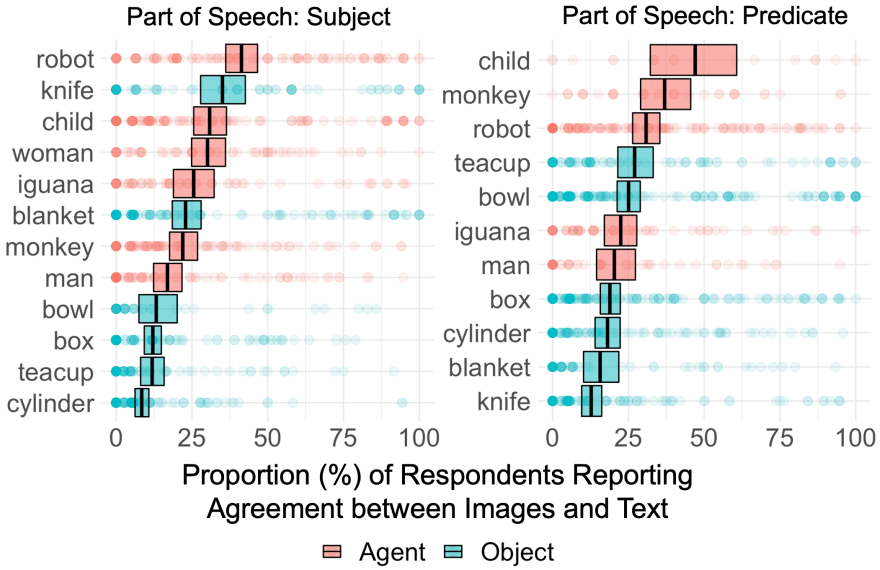

Figure A.5: The proportion of respondents reporting agreement between image and prompt, broken down by each entity’s part of speech.