DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models

Royi Rassin[*1], **Shauli Ravfogel**[*1,2] **Yoav Goldberg**[1,2]
[1]Bar-Ilan University [2]Allen Institute for Artificial Intelligence
{rassinroyi, shauli.ravfogel, yoav.goldberg}@gmail.com

Abstract

We study the way DALLE-2 maps symbols (words) in the prompt to their references (entities or properties of entities in the generated image). We show that in stark contrast to the way human process language, DALLE-2 does not follow the constraint that each word has a single role in the interpretation, and sometimes re-uses the same symbol for different purposes. We collect a set of stimuli that reflect the phenomenon: we show that DALLE-2 depicts both senses of nouns with multiple senses at once; and that a given word can modify the properties of two distinct entities in the image, or can be depicted as one object and also modify the properties of another object, creating a semantic leakage of properties between entities. Taken together, our study highlights the differences between DALLE-2 and human language processing and opens an avenue for future study on the inductive biases of text-to-image models.

1 Introduction

Large diffusion-based text-to-image models, such as DALLE-2 (Ramesh et al., 2022), generate visually compelling images that condition on input texts. Yet, the extent to which such models capture properties of human language, such as its compositional structure, has been doubted (Conwell and Ullman, 2022).

A very basic property in the interpretation of natural language utterances is that each word has a specific role in the interpretation, and there is a one-to-one mapping between symbols and roles. While symbols—as well as sentence structures—may be ambiguous, after an interpretation is constructed this ambiguity is already resolved. For example, while the symbol *bat* in *a flying bat* can be interpreted as either a wooden stick or an animal, our possible interpretations of the sentence are either of

Figure 1: The word *bat* is realized as two entities given the prompt *a bat is flying over a baseball stadium*.

a *flying wooden stick* or *a flying animal*, but never both at the same time. Once the word *bat* has been used in the interpretation to denote an object (for example a wooden stick), it cannot be re-used to denote another object (an animal) in the same interpretation. Similarly, in *a fish and a gold ingot*, the word *gold* is used as a modifier of *ingot*.[1] Once it is used, it cannot be re-used to modify another object in the same interpretation, and also cannot be used as a standalone object.[2] Some linguists refer to this property as *resource sensitivity* (Salvucci et al., 2009).

We show evidence that—in stark contrast

[*]Equal contribution.

[1]We use the word "modifier" to refer to any word that influences the way another word is depicted in the output.

[2]Note that in some cases a single word can be used to modify several objects, for example, *a gold fish and ingot* can be interpreted as *a gold fish and a gold ingot*. These cases manifest in very specific syntactic configurations, and are well documented in linguistics. Indeed, the linguistic analysis of such cases are either based on mechanisms such as "duplication" of the word, or allude to a deeper version of the sentence in which the word appeared twice, and was dropped before production. While the reality behind these linguistic theories cannot be proven, the appeal to use such mechanisms as "duplication" or "deletion" (before realization) as explanations to the phenomena, highlights the naturalness of the "single use per symbol" principle.

Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 335 - 345

to humans—the text-to-image model DALLE-2 (Ramesh et al., 2022) does not respect this constraint. Indeed, we show that DALLE-2 exhibits the following behaviors that are inconsistent with the single-role principle:

- **A word or phrase is interpreted as two distinct (concurrently-incompatible) entities** and is consequently realized as multiple objects in the same scene. For instance, the prompt *a bat is flying over a stadium* is mapped to an image showing a baseball bat alongside the flying mammal (fig. 1).
- **A word is interpreted as a modifier of two different entities**,[3] and is consequently realized as a property of multiple objects in the scene. For instance, the prompt *a fish and a gold ingot* is mapped to an image showing a gold ingot as well as a goldfish (section 4.2).
- **A word is interpreted simultaneously as an entity and as a modifier of a different entity** and is consequently realized as an object in the scene as well as as a property of another object in the same scene. For instance, the prompt *a seal is opening a letter* is mapped to an image showing a seal holding a *sealed* letter (fig. 2).

Visually, the cases we highlight map to one of two failure modes from the perspective of the image designer / user: (a) sense-ambiguous words are hard to isolate, and the resulting images often exhibit the unintended sense together with the intended one (*homonym duplication*). (b) visual properties of one object in the image "leak" into other objects in the image (*concept leakage*).[4]

We also observe some *second order effects*, where a hypernym of the modifier word, or an implicit description of it, also triggers the duplication effect.

Taken together, the phenomena we examine provides evidence for limitations in the linguistic ability of DALLE-2 and opens avenues for future research that would uncover whether those stem from issues with the text encoding, the generative model, or both. More generally, the proposed approach can be extended to other scenarios where the decoding process is used to uncover the inductive bias and the shortcomings of text-to-image models.

[3]While not under a syntactic configuration that allows such duplication.

[4]Importantly, we note that both phenomena rarely occur when explicit specification is provided, e.g, the prompt *a fruit bat is flying over a stadium* generates only the flying mammal. See Hutchinson et al. (2022) for a discussion on under-specification in text-to-image generation.

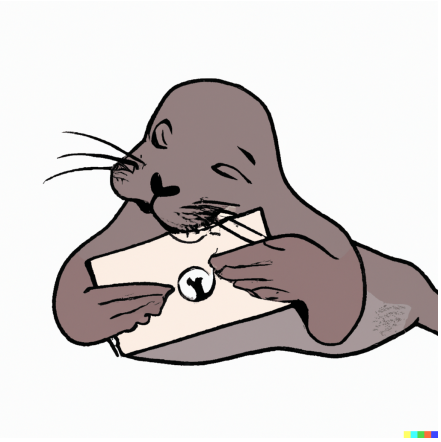


Figure 2: The word *seal* is realized both as an object and as a modifier of another word (*letter*) in the prompt *a seal is opening a letter*, resulting in *concept leakage*. The prompt tends to generate *sealed* letters, while minimally-different prompts that do not contain the noun *seal* do not.

2 Previous Work

Promptly after the introduction of DALLE-2 (Ramesh et al., 2022), analyses of the system began to surface (Marcus et al., 2022; Conwell and Ullman, 2022). Conwell and Ullman (2022) showed that DALLE-2 does not understand spatial relations. Marcus et al. (2022) reported that if an object in the prompt is said to have some property, then the image will likely show that property somewhere, however, not necessarily on the correct object. We take this analysis a step further, showing that the same property sometimes simultaneously modifies several unrelated objects. Ramesh et al. (2022) allude to the phenomena we analyze: they mention the issues with binding separate attributes to separate objects, and hypothesize that it is because CLIP (Radford et al., 2021) embeddings do not explicitly bind attributes to objects. Accordingly, they observe that that reconstructions from the decoder often mix up attributes and objects. Concurrent to this work, Hutchinson et al. (2022) discuss underspecification in text-to-image generation, a condition which we find to be associated with some of the behaviors we identify; and Anonymous (2023) discuss several cases of “leakage", and propose an intervention in the decoding process that is aimed to mitigate them.

3 Methdos

Stimuli We construct linguistic stimuli (prompts) which evoke a behavior that is inconsistent with the single-use-per-symbol axiom. For the first case (words interpreted as two entities), we use *homonyms*, words that have two distinct senses. DALLE-2 often generates two objects, one corresponding to each of the senses. For the modification cases, we construct the following prompts. For the case of a word that is interpreted as a modifier of two entities, we include prompts with two entities e_1 and e_2 and a modifier word m that is semantically compatible with both of them, but is only syntactically modifying the second one (For instance, "A *fish*$_{e_1}$ and a *gold*$_m$ *ingot*$_{e_2}$"). For the case of a word that is simultaneously interpreted as an entity as well as a modifier, we construct stimuli that contain two entities e_1 and e_2, such as that one of them is semantically associated with the other noun—for instance, "classic *butter*$_{e_1}$ and *peanuts*$_{e_2/m}$".

The complete list of stimuli is presented in tables 1, 2 and 3 in appendix A.1, together with samples of the generated images.

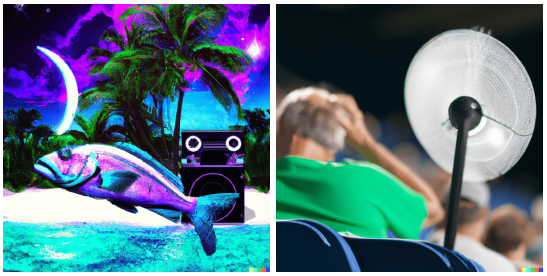
Control Stimuli In order to argue for the existence of a leakage of property P between two entities e_1 and e_2, we need to assess whether DALLE-2 does not tend to depict e_2 with that property, regardless of e_1. For instance, if the default *groceries* for DALLE-2 in different contexts is a groceries basket, it is hard to argue that it is the presence of the word *basket* in "a *basketball* near *groceries*" that elicited DALLE-2 to generate a basket. With this in mind, we generate a set of control prompts, which are minimally-different variations over the challenge prompts, that were built so as to prevent the possibility of a leakage by changing either n_1 or n_2 alone, depending on which change can prevent the leakage in a given prompt. The full set of controls are also detailed in tables 2 and 3.

4 Results

We generate 12 images per stimulus, and report the average over all stimuli and images.

4.1 A single word realized as multiple entities

We have 17 stimuli that elicit DALLE-2 to assign a single word two roles. Homonym duplication occurred in 80.3% of the 216 images. We found that, out of context, the vast majority of words are biased towards a particular sense, specifically, out of our 17 homonyms, only *bass* and *date* reveal homonym duplication without added context. Using *bass* as a prompt will[5] return an image of a *seabass* and a *guitar bass* or a *speaker bass*, and for *date*, DALLE-2 will return a romantic couple, holding together a date (the fruit). Most of the homonym duplication prompts were created by including context related to the non-default role of the homonym. For instance, for the homonym *fan*, DALLE-2 is biased towards the cooling device, but if we are to add context that is related to the other possible role a *fan* can have (a sports fan), and that still applies towards the cooling device, we will get both: *A fan at a hot sports event* (top of section 4.1).



left: *A bass lounging in a tropical resort in space, vaporwave*.
right: *a fan at a hot sport event*.

4.2 A single word realized as a modifier of multiple entities

Main (left): *a **fish** and a **gold ingot***.
Control (right): *a fish and an ingot*.

Given a pair of entities (e_1, e_2), we observe that the underspecified entity, e_2 is depicted by DALLE-2 with properties from e_1: in the prompt *A fish and a gold ingot*, *gold* modifies *fish*, and the image consistently contains a golden fish.[6] This is

[5] Using just a homonym as a prompt does not consistently elicit the phenomenon as a prompt that includes context with the homonym.

[6] Interestingly, The golden fish also resemble goldfish, suggesting that priming effect also extends to compounds, even though the word *gold* in *gold ingot* is a part of a phrase with a compositional structure, while "goldfish" is not compositional.

likely because *fish* is an underspecified noun and *gold* is an amicable property to fish. In the control, *a fish and an ingot*, golden fish do not appear at all in the output.

For the two-properties case, we have 6 stimuli-control pairs. The output of the stimuli prompts display the shared property in 97.2% of the cases, while the control prompts show that property in only 15.2% of the cases.

4.3 A single word is realized as an entity and as a modifier of another entity

Main (left): *A **zebra** and a **street**.*
Control (right): *A zebra and a gravel street.*

For the entity-to-property case, we have 10 stimuli-control pairs. On average, the stimuli prompts exhibit the shared property in 92.5% of the cases while the control prompt shows it in only 6.6% of the cases.

In similar fashion to the two-properties case, n_2 is depicted by DALLE-2 with properties from n_1, however, this time, n_1 is not a property, but an entity. To demonstrate, consider *a zebra and a street*, here, *zebra* is an entity, but it modifies *street*, and DALLE-2 constantly generates crosswalks, possibly because of the zebra-stripes' likeness to a crosswalk. And in line with our conjecture, the control *a zebra and a gravel street* specifies a type of street that typically does not have crosswalks, and indeed, all of our control samples for this prompt do not contain a crosswalk. When n_1 is a homonym, the entity-to-property case is more nuanced, for example, in *food and a cone hat, zoomed out photo*, the word *cone* modifies *food* by generating ice-cream cones, while its control: *food and a birthday hat, zoomed out photo*, a semantically and physically equivalent replacement of *cone*, does not lead to the generation of ice-cream cone.

4.4 Second-order stimuli

Main (left): *a tall, long-legged, long-necked bird and a construction site.*
Control (right): *a bird and a construction site.*

Concept Leakage can be taken a step further, by masking the affecting noun and receiving similar results: *a tall, long-legged, long-necked bird and a construction site* is describing *a crane and a construction site*, but nowhere in the prompt a *crane* is mentioned. Yet, it will cause DALLE-2 to generate two types of cranes: a bird crane, and a construction crane. We also observe that the bird-crane's head and legs share physical properties with the crane. From the same challenge prompt, the border between a bird-crane and a construction-crane is especially blurred. This behavior is not repeated when the description of a bird-crane is removed from the prompt: *a bird and a construction site*. This suggests that the model first maps entities into a concept space, and only then renders.

4.5 Other text-to-image-models

The empirical focus of this work is DALLE-2. In preliminary experiments, we found that DALLE-mini (Dayma et al., 2021)—a much smaller model—did not replicate the observed phenomena. Particularly, only the "common" word sense is depicted for homonyms. As for Stable-diffusion (Rombach et al., 2021), the phenomena seem to occur less frequently, particularly because in many times the model does not follow the prompt at all. Other models, such as (Saharia et al., 2022; Yu et al., 2022), are not publicly available. We hypothesize that—paradoxically—it is the lower capacity of DALLE-mini and Stable-diffusion and the fact they do not robustly follow the prompts, that make them appear "better" with respect to the flaws we examine. A thorough evaluation of the relation between scale, model architecture, and concept leakage is left to future work.

5 Conclusions

We demonstrate that across a set of stimuli, DALLE-2 does not follow basic principles of symbol-to-entity mapping in language. This flaw is especially pressing given that DALLE-2 is directed to the general population, where the majority of the users are probably not aware of the phenomenon.[7] Future work should trace back the issues we diagnose to specific components in the model, such as the text encoding or the generative decoder; and study its dependence on scale and architecture. Additionally, the development of post-hoc mitigation techniques is especially important, given that most users lack the resources needed to train the models from scratch.

6 Acknowledgements

We thank Carlo Meloni for his valuable feedback. This project received funding from the Europoean Research Council (ERC) under the Europoean Union's Horizon 2020 research and innovation programme, grant agreement No. 802774 (iEXTRACT).

7 Limitations

One limitation of this work is the focus on DALLE-2 a commercial product, whose source code is not publicly available. As discussed, capacity issues prevented a thorough examination of the phenomena we focus on in other available models. We hope that additional models of similar scale will be released to the public in the future, enabling a more thorough examination. Additionally, our analysis is focused on a manually constructed set of stimuli, that were designed so as to surface the issues we focus on in the clearest way. Scaling the analysis would rely on automatic or semi-automatic generation of stimuli of the kind we use, which is a challenging task. Yet, we believe that such large-scale analysis is important in order to robustly quantify the issues we observe.

References

Anonymous. 2023. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *Submitted to The Eleventh International Conference on Learning Representations*. Under review.

Colin Conwell and Tomer Ullman. 2022. Testing relational understanding in text-guided image generation.

Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khac, Luke Melas, and Ritobrata Ghosh. 2021. Dall·e mini.

Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. 2022. Underspecification in scene description-to-depiction tasks. *arXiv preprint arXiv:2210.05815*.

Gary Marcus, Ernest Davis, and Scott Aaronson. 2022. A very preliminary analysis of dall-e 2.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models.

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, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding.

Dario D Salvucci, Niels A Taatgen, and Jelmer P Borst. 2009. Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 25, item 90.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling autoregressive models for content-rich text-to-image generation.

[7] While the issues we focus on can easily go unnoticed when examining a randomly-selected set of outputs–which mostly look very good—our examination of carefully-selected stimuli clearly reveals them. We believe that such analysis is important, as human language is characterized by many long-tail phenomena.

A Appendix

A.1 Data

The full image-dataset can be accessed at https://github.com/RoyiRa/DALLE2-Flaws-in-word-to-concept-mapping.

The following tables contain the stimuli we use to test the reference system of DALLE-2 .

Prompt	Duplication Ratio
A person wearing a **cone** hat is eating, a full body photo	7/12
A **bat** is flying over a baseball stadium	11/12
A **bass** lounging in a tropical resort in space, vaporwave	10/12
A **fan** at a hot sports event	12/12
A **crane** carrying a heavy material, fish in the background	12/12
A banker collecting **dough**	9/12
Two men having a loud **beef**, restaurant in the background	10/12
A model with an **hourglass** figure	12/12
A gentleman with a **bow** in the forest	11/12
A woman is pouring water into her **glasses**	9/12
A man stuck in a **jam**, eating	7/12
A great **ruler**	8/12
Apple commercial	12/12
A person with **metal nails**	6/12
date	7/12
a person with **chicken legs**, full body image	12/12
a **board** meeting	9/12

Table 1: List of stimuli for the **multiple entities** case (Section 4.1). The noun that is mapped to two entities is denoted with **bold**

Challenge and Control	Challenge Ratio	Control Ratio
a person **jay**walking and a *bird* A person crossing the street unlawfully and a bird	11/12	0/12
food and a **cone** hat, zoomed out photo food and a birthday hat, zoomed out photo	12/12	0/12
A *fish* and a **gold** ingot A fish and an ingot	12/12	0/12
A **basket**ball near *groceries* A football near groceries	12/12	5/12
A **lady**bug and a *person*, full body image A bug and a person, full body image	11/12	6/12
An *animal* alongside a **piggy bank** An animal alongside a table	12/12	0/12

Table 2: Challenge and Control Prompts in **word-to-multiple-modifiers** case (section 4.2). The modifier nouns in the Challenge prompts are denoted with **bold**, and modified nouns with *italic*

Challenge and Control	Challenge Ratio	Control Ratio
A **zebra** and a *street* A zebra and a gravel street	10/12	0/12
A **seal** is opening a *letter* A large water mammal is opening a letter	10/12	1/12
Two people play **squash** with *food* Two people play squash	12/12	0/12
a **pool** and a *table* a pool and a bar	10/12	0/12
A person is washing hair with **Dove** *shampoo* A person is washing hair with shampoo	12/12	0/12
an **armadillo** and an *elephant* a rhino and an elephant	12/12	0/12
a fish and an **elephant** a rhino and an elephant	12/12	0/12
A **cross** and a *sidewalk* A crucifix and a sidewalk	10/12	2/12
classic butter and **peanuts** classic butter and cucumbers	12/12	0/12
A **leopard** and a *piece of cloth* a leopard and a black towel	11/12	5/12

Table 3: Challenge and Control Prompts in the **word-to-entity-and-modifier** case (Section 4.3). The modifier nouns in the Challenge prompts are denoted with **bold**, and modified nouns with *italic*

Challenge and Control	Challenge Ratio	Control Ratio
An **armadillo** on a *sea* shore A dog on a sea shore	12/12	0/12
A **pinniped** is opening a *letter* A large water mammal is opening a letter	7/12	1/12
a **tall, long-legged, long-necked bird** and a *construction site* a bird and a construction site	12/12	0/12
A *person* and a **bug with red spots**, full body image A bug and a person, full body image	11/12	6/12

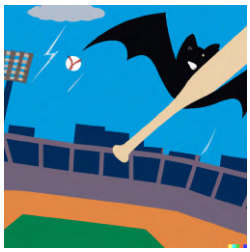
Table 4: Challenge and Control Prompts in **Second-Order Concept Leakage** (Section 4.4). The modifier nouns in the Challenge prompts are denoted with **bold**, and modified nouns with *italic*

Single word realised as multiple entities

*A person wearing a **cone** hat is eating, full body photo*

*A **crane** carrying a heavy material, fish in the background*

*A **fan** at a hot sports event*



*A **bat** is flying over a baseball stadium*

*A banker collecting **dough***

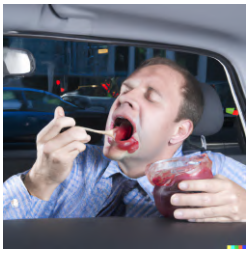
*A model with an **hourglass** figure*

*A **bass** lounging in a tropical resort in space, vaporwave*

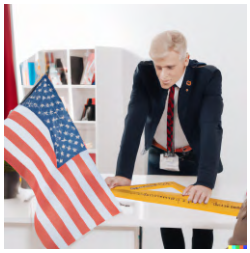
*Two men having a loud **beef**, restaurant in the background*

*A gentleman with a **bow** in the forest*

*A woman is pouring water into her **glasses***

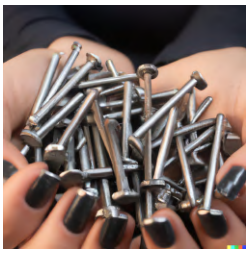


*A man stuck in a **jam**, eating*



*A great **ruler***

***Apple** commercial*



*A person with **metal nails***

date

*A **board** meeting*

*A person with **chicken legs,** full body image*

A word acting as a modifier of two different entities

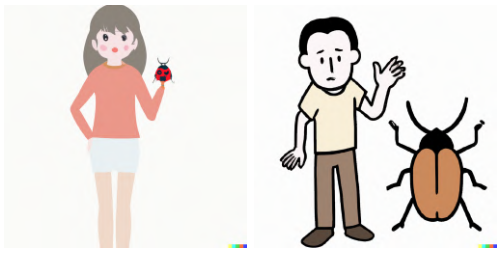
Bold items are **entities**, and bold+underline is the **modifier**.

Main (left): *A **person** **jay**walking and a **bird***
Control (right): *A person crossing the street unlawfully and a bird*

Main (left): ***food** and a **cone** **hat**, zoomed out photo*
Control (right): *food and a birthday hat, zoomed out photo*

Main (left): *A **fish** and a **gold** **ingot***
Control (right): *A fish and an ingot*

Main (left): *A **basketball** near **groceries***
Control (right): *A football near groceries*



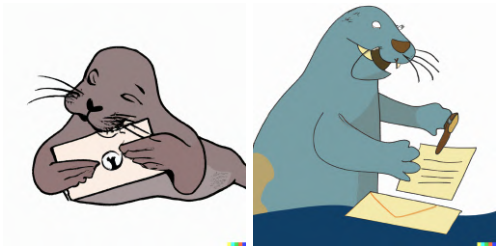
Main (left): *A **ladybug** and a **person**, full body image*
Control (right): *A bug and a person, full body image*

Main (left): *An **animal** alongside a **piggy** **bank***
Control (right): *An animal alongside a table*

Same word realised as an entity as well as a modifier of another entity

Bold item is the **entity**, bold+underline is acting as both **modifier and entity**.

Main (left): *A **zebra** and a **street***
Control (right): *A zebra and a gravel street*



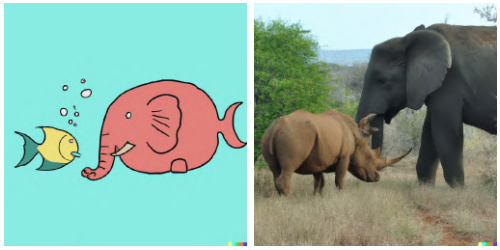
Main (left): *A **seal** is opening a **letter***
Control (right): *A large water mammal is opening a letter*

Main (left): *Two people play **squash** with **food***
Control (right): *Two people play squash*

Main (left): *a **pool** and a **table***
Control (right): *a pool and a bar*

Main (left): *A person is washing hair with **Dove shampoo***
Control (right): *A person is washing hair with shampoo*

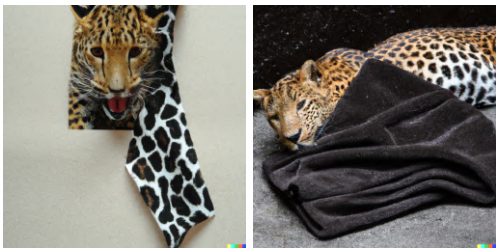
Main (left): *an **armadillo** and an **elephant***
Control (right): *a rhino and an elephant*



Main (left): *a **fish** and an **elephant***
Control (right): *a rhino and an elephant*

Main (left): *A **cross** and a **sidewalk***
Control (right): a crucifix and a sidewalk

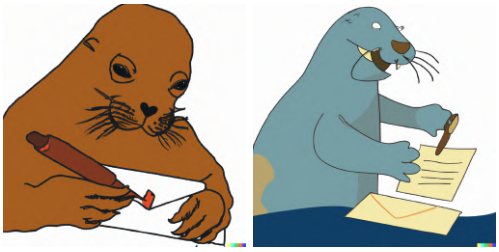
Main (left): *classic **butter** and **peanuts***
Control (right): *classic butter and cucumbers*



Main (left): *A **leopard** and a piece of **cloth***
Control (right): *a leopard and a black towel*

Second order concept leakage

Main (left): *A tall, long-legged, long-necked bird and a construction site*
Control (right): *a bird and a construction site*
Explanation: the *tall, long-legged, long-necked bird* is identified as a **crane** and realised as two kinds of cranes.



Main (left): *A pinniped is opening a letter*
Control (right): *A large water mammal is opening a letter*
Explanation: the *pinniped* interpreted as a **seal,** which is realised both as the actor and as a seal on the letter.

Main (left): *A person and a bug with red spots, full body image*
Control (right): *A bug and a person, full body image*
Explanation: The *bug with red spots* is identified as a **ladybug**, and the person is female (*lady*).

Main (left): *An Armadillo on a sea shore*
Control (right): *A dog on a sea shore*
Explanation: the *armadillo* has a **shell**, and there are shells on the beach.