

# Comparison of Image Generation Models for Abstract and Concrete Event Descriptions

Mohammed Abdul Khaliq<sup>1</sup> and Diego Frassinelli<sup>2,3</sup> and Sabine Schulte im Walde<sup>1</sup>

<sup>1</sup>Institute for Natural Language Processing, University of Stuttgart, Germany

<sup>2</sup>Department of Linguistics, University of Konstanz, Germany

<sup>3</sup>Center for Information and Language Processing, LMU Munich, Germany

{mohammed.abdul-khaliq,schulte}@ims.uni-stuttgart.de,  
diego.frassinelli@uni-konstanz.de

## Abstract

With the advent of diffusion-based image generation models such as DALL-E, Midjourney and Stable Diffusion, high-quality images can be easily generated using textual inputs. It is unclear, however, to what extent the generated images resemble human mental representations, especially regarding abstract event knowledge, in contrast to concrete event knowledge. We analyse the capabilities of four state-of-the-art models in generating images of verb-object event pairs when we systematically manipulate the degrees of abstractness of both the verbs and the object nouns. Human judgements assess the generated images and indicate that DALL-E is strongest for event pairs with concrete nouns (e.g., *pour water*; *believe person*), while Midjourney is preferred for event pairs with abstract nouns (e.g., *remain mystery*; *raise awareness*), in both cases irrespective of the concreteness of the verb. Across models, humans were most unsatisfied with images of events pairs that combined concrete verbs with abstract direct-object nouns (e.g., *speak truth*; *steal idea*). We hypothesised that this is due to the tendency of these combinations to express figurative language, which was confirmed by post-hoc collected human judgements.

## 1 Introduction

Nowadays tools for automatic image generation are accessible to laypeople as much as to experts. But do the generated images capture human mental representations? And which images are generated for abstract concepts and events that are not easily depictable, such as the concept *patience* and the event *speak the truth*, given that what we really see in the images depicting abstract knowledge are concrete objects?

The current study assesses four image generation models on how well they depict abstract vs. concrete event descriptions: we compare DALL-E 2 (Ramesh et al., 2022), Stable Diffusion (Rombach

et al., 2022), Stable Diffusion XL (Podell et al., 2023) and Midjourney<sup>1</sup>, as well as images retrieved by the search engine Bing<sup>2</sup>. Following Frassinelli and Schulte im Walde (2019), the prompts for the models are represented by 40 phrase-level events consisting of a verb and a direct object noun, where we systematically vary the words’ degrees of abstractness by relying on the ratings in Brysbaert et al. (2014), cf. *build a perspective* vs. *carry a box*. We evaluate the generated images through human ratings (i) in a standard large-scale crowd-sourcing task, and (ii) in a two-step small-scale setup where we prime our participants on their expectations by asking them to first describe what they would expect to see in an image of a specific event, before asking them to judge the quality of the automatically generated images. Our hypothesis is that humans will be less satisfied with the depiction of abstract in comparison to concrete event knowledge, while it is unclear how and to what extent the abstractness of verbs vs. nouns influences the human judgements with regard to the four-way combinations of abstract/concrete verb-noun events.

We thus propose an exploration of the capabilities of image generation models regarding abstract vs. concrete event descriptions, while previous work primarily focused on concrete events such as scenes with concrete objects and relations (Johnson et al., 2018), person appearance and shape (Tang et al., 2020), and transformer-based text-to-image generation across different styles (Ding et al., 2021), or on investigating prompts variants for optimising the generation of abstract and figurative concepts (Chakrabarty et al., 2023; Liao et al., 2023). Examples of research that not only targeted concrete but also abstract knowledge in images, are studies by McRae et al. (2018) who performed priming experiments for abstract words in images,

<sup>1</sup><https://www.midjourney.com>

<sup>2</sup><https://www.bing.com/>

Akula et al. (2023) who proposed standard vision detection and retrieval tasks to distinguish between concrete and abstract concepts in visual metaphors, and Shahmohammadi et al. (2023) who trained image generation models to illustrate any kind of textual input, including figurative language.

## 2 Target and Data Collections

As the basis for our experiments we create verb-noun event pairs of varying degrees of concreteness (Section 2.1). These event pairs are used as prompts for the image generation models (Section 2.2).

Verb	Score	Noun	Score	Category V + N
eat	4.44	meal	4.66	C + C
know	1.68	man	4.79	A + C
raise	3.80	awareness	1.84	C + A
assume	1.75	responsibility	1.40	A + A

Table 1: Examples of verb-noun event pairs, together with the individual verb/noun mean concreteness rating scores from Brysbaert et al. (2014) on a scale from 1 (abstract) to 5 (concrete), and the event category type.

### 2.1 Verb-Noun Event Pairs

We rely on the concreteness ratings by Brysbaert et al. (2014) to systematically create a total of 40 pairs combining 10 strongly concrete verbs and strongly concrete nouns (ConcV+ConcN), 10 strongly abstract verbs and strongly concrete nouns (AbstV+ConcN), 10 strongly concrete verbs and strongly abstract nouns (ConcV+AbstN), and 10 strongly abstract verbs and strongly abstract nouns (AbstV+AbstN). Table 1 presents one example per verb-noun event category and the corresponding individual word concreteness ratings. The full table is provided in Appendix A.

### 2.2 Image Generation

We employ four image generation models. In addition to these models we also use Bing images.

**DALL-E 2** is a text-to-image image generation model from OpenAI released in April, 2022. DALL-E 2 can be accessed through the OpenAI’s API at a fixed cost per image basis. It is able to create an image in 1:1 aspect ratio with a maximum resolution of 1024x1024, which is what we use.

**Midjourney (MJ) v5.1** is a text-to-image model developed by Midjourney Inc. Unlike the other models, Midjourney is not accessible through an API, and it requires manual prompting in a Discord interface. It also has a fixed subscription-based

payment to generate images. Midjourney v5.1 generates images at 1024x1024 resolution which can be altered for different aspect ratios. We use the default 1024x1024 resolution of v5.1.

**Stable Diffusion (SD) v2.1** is a text-to-image model developed by Stability AI which makes use of the latent diffusion model architecture to generate images. It is open-source and can be run locally or accessed via API through DreamStudio<sup>3</sup>. It is able to create images of varying aspect ratios and resolutions at the cost of degrading quality the further you go away from the 768x768 native resolution. We use the 768x768 resolution for all our generations setting the inference steps to 75.

**Stable Diffusion XL (SDXL) v1.0** is the latest Stable Diffusion model from Stability AI. It improves over Stable Diffusion v2.1 by requiring shorter and less detailed prompts and being able to generate text within the images. Additionally, its three times larger UNet Backbone (used for image segmentation) and architectural improvements enable it to create more prompt-consistent and high-quality images with a native resolution of 1024x1024. It is open-source and can be run locally or accessed via API through DreamStudio. We set the resolution to 1024x1024 with the number of inference steps set to 50 (default).

**Bing** is a search engine that we use for image search as an upper bound to evaluate the image generation models. We feed our prompts via the Bing API to retrieve images that are not restricted by resolution or aspect ratio.

For all four models as well as Bing, we use as prompts the verb-noun event pairs introduced above. The image generation models were prompted using their default parameters. We collect four images from each of the four models’ outputs as well as from Bing, for each of the 40 verb-noun pairs, a total of 800 images. Figure 1 presents one example image for each model and for two event pairs, *serve food* (ConcV+ConcN) and *remain mystery* (AbstV+AbstN).

## 3 Model Evaluation

We evaluate the generated images through human ratings in two studies. The images, the full annotation instructions and all collections are publicly available at <https://www.ims.uni-stuttgart.de/data/image-generation>.

<sup>3</sup><https://dreamstudio.ai/generate>



Figure 1: Example images for the event pairs *serve food* and *remain mystery*, as generated by the four models.

<b>serve food</b>	<b>reduce noise</b>	<b>steal idea</b>	<b>remain mystery</b>
a waiter bringing a platter filled with food to a table at a dimly lit diner, three people sitting at the table; a man stands behind a counter and dishes up a variety of foods to a customer	a slider with a speaker symbol next to it and an arrow over the slider pointing away from the speaker symbol; a grainy picture followed by an arrow and a very soft looking version of the picture	a person in a lab coat leafing through a notebook, the body language shows unease; person with thought bubble above their head, the thought bubble is being snatched away by another person	a woman burns a letter from an ex without reading it; an archaeologist tries to decipher a text from an unknown language

Table 2: Examples of human descriptions for four verb-noun event pairs in Task 1 of the expectation-based study.

**Study 1: Crowdsourcing Ratings** We gather ratings of the generated images for our verb-noun events from Amazon Mechanical Turk (AMT)<sup>4</sup> workers based in either USA or UK, and with more than 10,000 prior submissions and a  $\geq 99\%$  approval rate. The workers are asked to rate on a scale from 1 to 6 how well each of the 800 generated images depicts the associated verb-noun pair event. We also add 80 images as sanity checks; these include an obviously wrong image for additional verb-noun pairs, e.g., an image of a car for *play football*.

**Study 2: Expectation-based Ratings** This evaluation is conducted in two consecutive tasks.

In Task 1 we aim at collecting precise descriptions of what our participants expect to see in an image of a particular verb-noun event, by asking them to provide one or more phrases describing the mental image they created of the given event. In this way, participants can reflect on the given event and the mental representations they are generating.

In Task 2, the same participants are presented with the same verb-noun pairs, their own descriptions for the pairs, and four images from each of the four models and Bing. They are asked to select all images that depict the event well, without providing any ranking. The annotators can also select images that do not directly match their own descriptions, as long as they judge the image good.

The annotators are university students highly proficient in English (B2 level or higher). We collect 19 responses from our annotators describing their image expectations for the verb-noun event in Task 1 (see examples in Table 2). 12 out of the 19 annotators also completed Task 2.

## 4 Results

**Study 1: Crowdsourcing Ratings** We collected a total of 7,200 ratings for our 800 images, with nine unique annotators rating each image on a scale from 1 to 6. After removing all ratings by annotators that failed the sanity check, and using only those images that received  $\geq 4$  approved ratings, our final set contains 4,212 ratings.

<sup>4</sup><https://www.mturk.com/>

These ratings distribute over our event categories as follows.

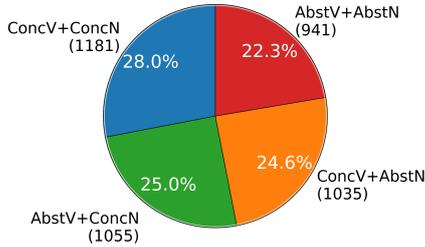


Figure 2: Final set of ratings across categories.

Figure 3 presents the proportions of how often a model received an extremely low (bad) rating of 1 or 2 (left plot) or an extremely high (good) rating of 5 or 6 (right plot), out of the total number of ratings for that model and a specific verb-noun event category. For example, SD received a very low rating for 62 (48%) of the generated images in the AbstV+AbstN category and a very high rating for only 13 (10%) generated images in this category.

Overall, we can clearly see that SD (orange bars) received most low ratings and fewest high ratings across event categories; Bing (blue bars) serves as an upper bound (i.e., receiving few low and many high ratings across most event categories); and DALL-E, SDXL and MJ show more variable results across event categories. More specifically, the right plot in Figure 3 displays closer competitions across the image generation models: Our best performing model for AbstV+ConcN and ConcV+ConcN is DALL-E, while MJ is best regarding the other two categories. Therefore, DALL-E performs best when the direct-object noun is concrete, while MJ performs best when the direct-object noun is abstract, irrespective of the concreteness of the verb. MJ also exhibits a rather uniform success rate across categories. SDXL (green bars) is the second best generation model in three out of four categories.

**Study 2: Expectation-based Ratings** Figure 4 shows how many images from each model were selected by the annotators across verb-noun categories in Task 2, after they had previously described their expectations (see examples in Table 2). Similar to our large-scale experiment, we notice the consistently poor performance of SD, while DALL-E, SDXL and MJ are more favoured, and Bing serves as the upper bound. The plot confirms that DALL-E performs best when the direct-object noun is concrete, while MJ performs best when the

direct-object noun is abstract. Finally, the annotators were much less satisfied across models with images for the ConcV+AbstN event category than with images for any of the other event categories.

Table 3 once more confirms the general trends by showing the total number of images for each model that were selected in Task 2. Again we notice that MJ, DALL-E and also SDXL are more favoured than SD, and that Bing serves as the upper bound. Table 3 also shows the mean and standard deviation scores across our four event categories, pointing out that especially DALL-E varies strongly.

	#selected	mean	stdev
Bing	760	190.00	47.10
DALL-E	513	128.25	60.75
MJ	530	132.50	33.27
SD	181	45.25	19.76
SDXL	429	107.25	35.91

Table 3: Overall selected images per model/Bing.

Overall, our human expectations evaluation confirms the general trends from the crowdsourced evaluation regarding (dis)preferences that annotators perceived when judging the generated images. In fact, Figure 4 presents a similar yet sharper picture of the human evaluation preferences in comparison to Figure 3.

**Abstract Events and Figurative Language** Our initial hypothesis was that humans would be less satisfied with the depiction of abstract in comparison to concrete event knowledge. Looking into our best model results, this hypothesis has been confirmed but in an unexpected way. We found that DALL-E performs best when the direct-object noun is concrete (however with a rather large standard deviation), while MJ performs best when the direct-object noun is abstract, irrespective of the concreteness of the verb. In particular, annotators were much less satisfied across models with images for the ConcV+AbstN event category. So overall it seems as if the abstractness of the noun plays a core role in how well the generated images depict verb-noun events.

We suspected that this is the case because ConcV+AbstN events predominantly express figurative language usage, as suggested by Frassinelli and Schulte im Walde (2019), which is inherently difficult to depict. In order to look into this follow-up hypothesis, we ran an additional annotation study by asking 12 annotators for their binary judgements

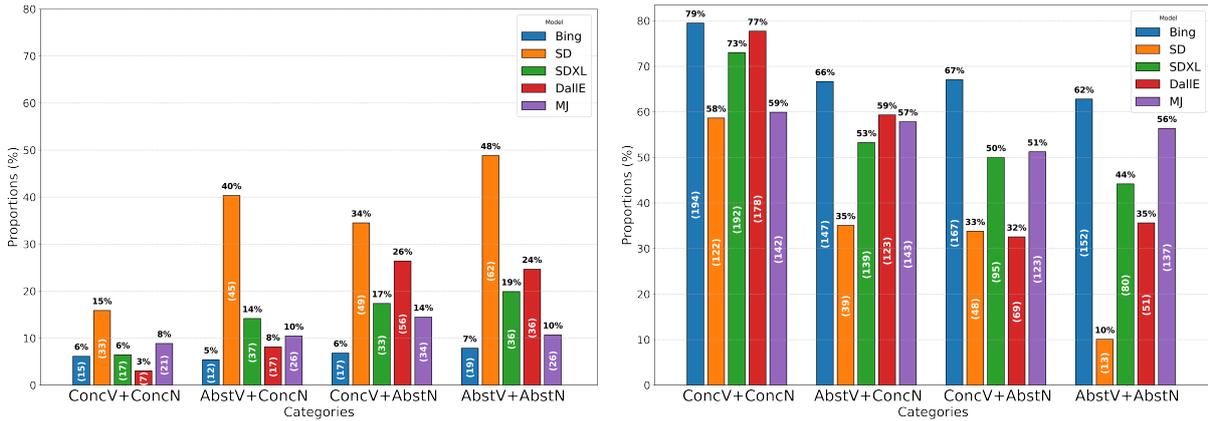


Figure 3: Proportions of how often the four models or Bing received an extremely low rating (1 or 2, *left plot*) or an extremely high rating (5 or 6, *right plot*) in the crowdsourcing evaluation, out of the total number of ratings for that model and a specific event category.

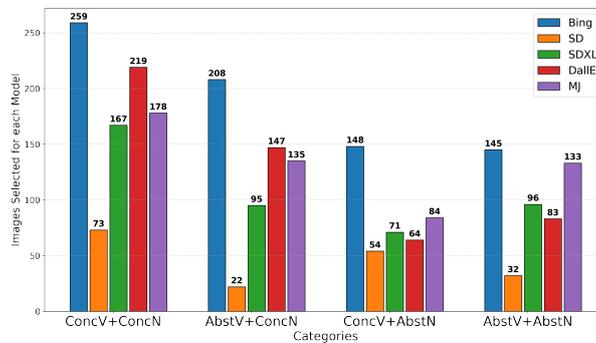


Figure 4: Number of images selected for each model in Task 2 of the human expectations setup, i.e., where the annotators judged the images as well-depicting the respective events.

on figurative vs. literal language of our 40 event pairs.<sup>5</sup> Figure 5 shows that indeed ConcV+AbstN (and to a lesser degree also the most abstract combination AbstV+AbstN) are strongly perceived as figurative language.

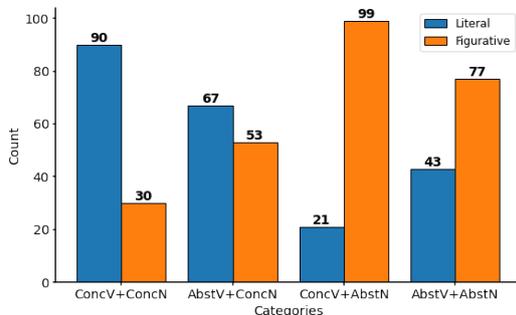


Figure 5: Number of literal vs. figurative language judgements of event pairs across event categories.

<sup>5</sup>We also asked the annotators to provide example sentences, so that we could check that they understood the task, and to obtain textual event information. The data are publicly available from the same URL as above.

## 5 Conclusion

This paper systematically assessed image generation models on their capacity to generate images for abstract vs. concrete event descriptions. We demonstrated through human evaluations that DALL-E is strongest for event pairs with concrete nouns, while MJ is strongest for event pairs with abstract nouns. Regarding images for events with a concrete verb and an abstract direct-object noun, humans were generally not satisfied with any model, which is an additional annotation attributed to a strong tendency for representing figurative language.

We cannot conclusively say why some models perform better than others, but we suspect that this is due to reasons such as MJ’s tendency to produce more creative images in contrast to DALL-E producing simplistic and to-the-point images (which humans seem to like for concrete nouns). Overall, all models were outperformed by Bing images, which we attribute to less artifacting, randomness and consistency issues in those images.

## Acknowledgements

This research was supported by the DFG Research Grant SCHU 2580/4-1 (*MUDCAT – Multimodal Dimensions and Computational Applications of Abstractness*). We also thank the reviewers and the SemRel group for useful feedback and suggestions, and our friends and colleagues who supported us in our collections of image expectations and expectation-based ratings.

## Ethics Statement

We are aware that image generation models – as all models trained on some selection of natural language data – are likely to capture biases regarding societal inequalities. With respect to our experiments involving human participants, we do not see any ethical issues related to this work: All collections were conducted on a voluntary basis with a fair compensation (12 Euros per hour), and we kept the data collection anonymous.

## References

- Arjun R. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T. Freeman, Yuanzhen Li, and Varun Jampani. 2023. MetaCLUE: Towards comprehensive visual metaphors research. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 23201–23211, Vancouver, Canada.
- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. [Concreteness ratings for 40 thousand generally known English word lemmas](#). *Behavior Research Methods*, 46(3):904–911.
- Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. [I spy a metaphor: Large language models and diffusion models co-create visual metaphors](#). *arXiv preprint arXiv:2305.14724*.
- Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. CogView: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835.
- Diego Frassinelli and Sabine Schulte im Walde. 2019. Distributional interaction of concreteness and abstractness in verb–noun subcategorisation. In *Proceedings of the 13th International Conference on Computational Semantics*, pages 38–43, Gothenburg, Sweden.
- Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1219–1228, Salt Lake City, Utah.
- Jiayi Liao, Xu Chen, Qiang Fu, Lun Du, Xiangnan He, Xiang Wang, Shi Han, and Dongmei Zhang. 2023. Text-to-image generation for abstract concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 38(4), 3360–3368, Vancouver, Canada.
- Ken McRae, Daniel Nedjadrassul, Raymond Pau, Bethany Pui-Hei Lo, and Lisa King. 2018. Abstract concepts and pictures of real-world situations activate one another. *Topics in Cognitive Science*, 10:518–532.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. [SDXL: Improving latent diffusion models for high-resolution image synthesis](#). *arXiv preprint arXiv:2307.01952*.
- Anirudh Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. [High-resolution image synthesis with latent diffusion models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.
- Hassan Shahmohammadi, Adhiraj Ghosh, and Hendrik Lensch. 2023. ViPE: Visualise pretty-much everything. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5477–5494, Singapore.
- Hao Tang, Song Bai, Li Zhang, Philip H. S. Torr, and Nicu Sebe. 2020. XingGAN for person image generation. In *Proceedings of the European Conference on Computer Vision*, pages 717–734. Springer.

## A All 40 Verb-Noun Pairs and their Event Categories

Verb	Score	Noun	Score	Category V + N
eat	4.44	meal	4.66	ConcV+ConcN
write	4.22	song	4.66	
pour	4.14	water	5.00	
throw	4.04	money	4.54	
carry	4.04	weight	3.94	
raise	3.80	family	4.23	
serve	3.78	food	4.80	
build	3.71	company	4.11	
hold	3.68	pillow	5.00	
read	3.56	paper	4.93	
put	2.50	weight	3.94	AbstV+ConcN
keep	2.37	money	4.54	
investigate	2.27	case	3.93	
generate	2.23	electricity	3.90	
sustain	2.17	injury	4.00	
educate	2.12	child	4.78	
reduce	2.00	noise	3.52	
develop	1.87	company	4.11	
know	1.68	man	4.79	
believe	1.55	person	4.72	
pave	4.03	way	2.34	ConcV+AbstN
seize	3.97	moment	1.61	
steal	3.84	identity	2.00	
steal	3.84	idea	1.61	
raise	3.80	awareness	1.84	
raise	3.80	expectation	1.62	
build	3.71	perspective	2.38	
speak	3.70	truth	1.96	
hold	3.68	responsibility	1.40	
unfold	3.55	drama	2.34	
understand	2.28	reason	1.93	AbstV+AbstN
understand	2.28	meaning	1.85	
learn	2.20	language	2.35	
reduce	2.00	loss	2.19	
remain	1.96	mystery	2.33	
develop	1.87	idea	1.61	
improve	1.82	safety	2.37	
improve	1.82	health	2.28	
fulfill	1.78	obligation	2.04	
assume	1.75	responsibility	1.40	

Table 4: All our 40 verb-noun event pairs, together with the individual verb/noun mean concreteness rating scores from Brysbaert et al. (2014) on a scale from 1 (abstract) to 5 (concrete), and the event category type.