Representing Abstract Concepts with Images: An Investigation with Large Language Models

Ludovica Cerini[1], **Alessandro Bondielli**[1,2], **Alessandro Lenci**[1]
[1]Department of Philology, Literature and Linguistics, University of Pisa
[2]Department of Computer Science, University of Pisa
ludovica.cerini@phd.unipi.it
{alessandro.bondielli, alessandro.lenci}@unipi.it

Abstract

Multimodal metaphorical interpretation of abstract concepts has always been a debated problem in many research fields, including cognitive linguistics and NLP. With the dramatic improvements of Large Language Models (LLMs) and the increasing attention toward multimodal Vision-Language Models (VLMs), there has been pronounced attention on the conceptualization of abstracts. Nevertheless, a systematic scientific investigation is still lacking. This work introduces a framework designed to shed light on the indirect grounding mechanisms that anchor the meaning of abstract concepts to concrete situations (*e.g. ability - a person skating*), following the idea that abstracts acquire meaning from embodied and situated simulation. We assessed human and LLMs performances by a situation generation task. Moreover, we assess the figurative richness of images depicting concrete scenarios, via a text-to-image retrieval task performed on LAION-400M.

Keywords: LLMs, figurative language, multimodality, indirect grounding

1. Introduction

A naive view of abstract words regards them as expressing concepts that are not perceptually grounded, because they do not refer to entities with physical existence and perceivable by our senses. However, the relationship between abstract concepts and perceptual data is much more complex than it appears *prima facie*. Firstly, even if they lack direct grounding, abstract concepts are **indirectly grounded** (Louwerse, 2011; Dove, 2014; Utsumi, 2022), by being closely related to our embodied experience of the world (Paivio, 1990; Barsalou, 2008; Borghi et al., 2017). In this sense, abstract concepts succeed in acquiring perceptual representations via their association with concrete situations (e.g., the concept of *love* is grounded in the event of a mother hugging her child). Secondly, images can represent abstract concepts too. For instance, Figure 1 can be said to represent not only a mother with her baby, but also the concept of love. Actually, it is the association with abstract concepts that gives rise to metaphorical and figurative interpretations of images. Our hypothesis is that the ability of images to represent abstract concepts is determined by the indirect grounding of the latter: *An image represents an abstract concept A, if I depict a situation associated with A*. In this work, we use this hypothesis to address two main questions:

Q1 Do Large Language Models (LLMs) have human-analogue indirect grounding abilities to produce concrete situations that are strongly associated with abstract concepts?

Q2 Can the situations generated by the LLMs be used to retrieve images that represent abstract concepts? In addressing these questions, our main goal is to establish a framework aimed at enriching linguistic and multimodal resources for the study of metaphorical grounding of abstract concepts, with a focus on the Italian context. To this end, we selected a set of Italian abstract nouns and we set-up a **situation generation task** to compare the human-generated and LLM-generated situations produced in response to the target abstract word prompts. Then, we used the LLM-generated data in a **text-to-image retrieval task** from the LAION-400M dataset (Schuhmann et al., 2021a). Finally, the retrieved images were evaluated via crowdsourcing with respect to their ability to represent the target abstract concepts. The results of our experiments show that the ability of LLMs to ground abstract concepts on situations is very similar to the human one. Moreover, images retrieved through these situations strongly represent the target concepts used to generate them, suggesting that this method might be used to develop datasets of images annotated with their figurative meanings and to enhance the competencies of multimodal models to cope with metaphorical interpretations.[1]

2. Related Work

In linguistics, a number of approaches have been proposed to investigate the pragmatic abilities of

[1]We release the data collected across our experiments at `https://github.com/lcerini/SituaMet` (In preparation).

The Workshop on Cognitive Aspects of the Lexicon, pages 107–113
20 May, 2024.

Figure 1: Example of an image depicting a concrete situation evoking an abstract concept.

LLMs (Seals and Shalin, 2023; Hu et al., 2023). Barattieri di San Pietro et al. (2023) show that liguistic competence could be encoded distributionally in LLMs, thus allowing us to leverage such models to extract cognitive pattern of linguistic phenomena. This is particularly interesting for abstract conceptualization. Existing resources such as WordNet do not always reflect human representation of hierarchical relations among concepts (Bolognesi and Caselli, 2022; Liao et al., 2023). However, pragmatic inference involving meta-representional information still are not fully achieved in LLMs (Barattieri di San Pietro et al., 2023). Studies on non-literal understanding are still lacking, and the indirect grounding of abstract terms is understudied. Meta-representational, embodied, situated and multimodal aspects of conceptualization are the key to investigate metaphorical realization in a more complete and complex fashion. Major approaches from the NLP community in this context have been focused on text only (Shutova et al., 2010; Mohler et al., 2013; Shutova et al., 2016; Pramanick et al., 2018; Liu et al., 2020). Data scarcity and the cost of creating multimodal datasets impede research in this sense. Large datasets of image-text pairs, built by querying search engines, have been built to effectively train Large Vision-Language Models (VLMs) (Desai et al., 2021). However, due to their end goal, these dataset are less appropriate to study the more abstract and figurative aspects of images. Emerging approaches to multimodal figurative language have been proposed, and new effort has been made in the realization of multimodal metaphors datasets (Zhang et al., 2021; Akula et al., 2023). None, however, take fully into account the indirect grounding view of abstract concepts and images. Metaphorical images obtained by juxtaposition, resemblance, or fusion mechanisms instead lie outside the scope of the present work.

3. Situation generation task

The mental representation of abstract concepts makes use of associative relations to acquire meaning (Crutch and Warrington, 2005). To examine humans' abilities in grounding abstract concepts onto situations, and thus the capacity of LLMs to associate abstract concepts with real-world knowledge, we proceeded as follows. We first designed an elicitation task to investigate humans representation of abstract-related situations. Then, we translated it in a few-shot prompting task fed to a GPT-3 `davinci-003` model (Brown et al., 2020), which is the last non chat-oriented OpenAI GPT model and has shown state-of-the-art abilities in this regard.

Eliciting situations from humans To collect human data, we designed a test to elicit situations associated with 107 abstract stimuli. We divided the stimuli into three levels of concreteness (low, medium, and high) using norms from Brysbaert et al. (2013). *Low* abstract are concepts perceived as more anchored to concrete ideas or entities, while *high* abstract concepts are perceived as less anchored to concrete ideas or entities. *Victory* for example is seen as more concrete due to its association to experiences, unlike *justice*, which aligns with moral and social aspects.

Participants in a crowdsourcing experiment were asked to describe a situation that came to their mind given an abstract concept, with instructions including examples of situation formats to guide their responses. We submitted the test to 60 participants and obtained 539 situations in total, with an average of 5 situations per stimulus. We used Prolific[2] to crowdsource participants. An example of the abstract stimulus and the resulting situation is shown in Table 1.

Abstract Stimulus	Situation
Ability	Athletes performing acrobatic feats
Speed	A lion running

Table 1: Example situations generated by humans (translated from Italian).

Generating situations with an LLM To generate linguistic situations from abstract concepts with an LLM, we used the pre-trained `Davinci-003` GPT-3 model, following a structured output generation design. We exploited a few-shot prompting method, to obtain a specific format of generated situations, with $temperature = 0.5$ and $toppenalty = 1$. The few-shot prompt was constructed by using the same abstract stimuli used in the human elicitation task, followed by an arrow operator and two examples taken from human situations (See Fig. 2), ensuring consistency with the protocol used for human elicitation. We generated 10 situations per abstract concept, totalling 1070

[2] `prolific.co`

situations.

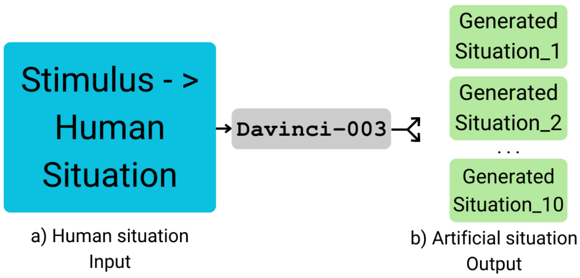

Figure 2: Few-shot prompt schema. Example of human input: *ability -> a person jumping*

Evaluation and Analysis Human-Elicited (HES) and Artificially-Generated (AGS) situations were then evaluated based on their similarity density and the associative strength to the abstract stimuli.

The qualitative analysis of HES revealed a certain degree of prototypical but diverse associations, i.e. similar situations among the different abstract concept groups. This is in line with literature indicating that people tend to anchor the representation of categories, simulating them in a typical perceptual situation (W.Yeh and Barsalou, 2006).

Starting from this, we explored how the two groups behaved in terms of typicality/diversity. We used `bert-base-italian-cased`, an Italian BERT model (Devlin et al., 2019), to obtain vector representations of the generated situations. For both HES and AGS groups we computed the the average cosine similarity between situations for each given abstract stimulus. A key aspect of the methodology involved constructing a dictionary that mapped each concept to its associated phrases. First, we iteratively processed the dataset and organized phrases according to their respective concepts. Then, we quantified the semantic similarity of phrases within each concept by computing the cosine similarity between each pair of phrases belonging to the concept. Note that we used the BERT tokenizer and right-padded all phrases to a standardized length (i.e., the maximum input size of the model), as it is customary when modelling texts with encoder-only Transformer architectures. We obtained embeddings for each phrase with a forward pass on the model, and then computed their pairwise similarity scores. Finally, we averaged the similarity scores to provide a quantitative measure of the semantic density of the phrases associated with the target concepts.

Figure 3 shows the average cosine similarities taking into account also the concreteness of the stimulus. Concreteness does not seem to affect similarity distributions. The median similarity values for the HES group was 0.365, while for the AGS group it was 0.375, indicating a slight central tendency difference between the two groups. As the two groups of items have been independently generated, we asses their statistical differences through a Wilcoxon test. The test showed a statistically significant difference between them ($W =$ 1203.5, $p <$0.05). However, the small difference in median values suggests that the magnitude of these differences might not be large in practical terms. The qualitative analysis of the results allowed us to identify several outliers that offer some insights into the nuances of human versus artificial sentence generation. In some concepts, AGS situations depict scenarios that, despite being distinct from each other, follow a similar structure in terms of the entities represented and the type of action involved. For example, the artificially generated sentences for the concept *power* predominantly feature situations involving entities associated with positions of power, such as *presidents, judges, and military figures*. This contrasts with the human-produced sentences, where more metaphoric situations also emerge, such as *a CEO's desk* versus an artificially generated scenario of *a CEO making decisions*. Similarly, in the case of the concept *peace*, human-generated sentences depict scenarios that do not necessarily involve animate entities (e.g., *a tranquil place*), offering a more abstract or symbolic representation of peace. Conversely, all artificially generated instances involve animate entities, such as *a group of people meditating together*. Despite the observed differences in similarity density and the varied entity types these outliers represent, they nonetheless appear to be associated with the abstract stimulus. To quantitatively evaluate this aspect and further understand how effectively HES and AGS capture and reflect the abstract concepts they are meant to represent, we proceeded to measure the associative strength. This measure aims to quantify the extent to which the generated phrases, regardless of their surface differences, retain a strong conceptual linkage to the original ideas they express.

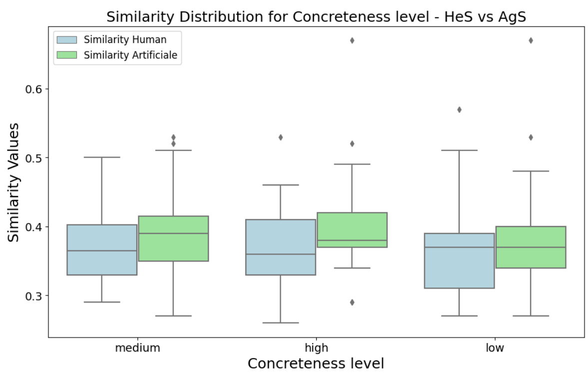

Figure 3: Similarities distribution in HES and AGS divided for concreteness level of abstract stimuli

To compute the associative strength between the stimuli abstract concepts the situations in HES and AGS we performed a rating assessment test using crowdsourcing via Prolific. Given an abstract concept and a situation obtained for it, we asked participants to rate how much the situation represents the paired abstract concept on a Likert scale from 1 (not at all) to 7 (very much). We collected ratings for all HES and AGS situations. Each concept- situation pair was rated by 10 participants, and we took the average rating. Figure 4 shows the distribution for each group. Both HES and AGS have a similar distribution, with median values 5.08 for HES and 5.07 for AGS ($\rho = 0.45$, p-value 0.00). This responds to **Q1** by suggestsing that i.) the LLM was able to generate human-analogue, coherent, and correct situations associated with the abstract stimuli, and that ii.) the generated situations could be considered as a proxy for the indirect grounding mechanism, providing evidence that diverse scenarios could suggest a figurative link between abstract concepts and real-world concrete events.

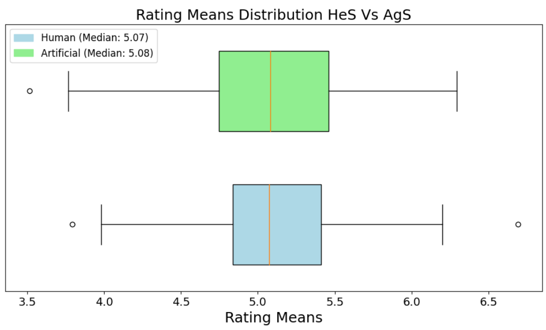

Figure 4: Rating avg. distribution in HES vs. AGS

4. Text-to-image retrieval task

Given the results of the previous experiment, we investigate on Q2 to understand whether i.) LLM-generated situations could be used to retrieve images able to represent an abstract concept and ii.) the same associative strength could be confirmed in the abstract-visual situation pair associations. To do so, we retrieved images based on the AGS, which were then evaluated via crowdsourcing.

Image Retrieval For our experiments we used the LAION-400M dataset (Schuhmann et al., 2021b). It contains 400 million of CLIP-filtered image-text pairs. We used AGS as queries for retrieving similar images from LAION-400M. To do so, we used CLIP (Radford et al., 2021) within the clip-retrieval library.[3]

[3]`https://github.com/rom1504/clip-retrieval`

For each situation in AGS, we collected the 10 most similar images. Thus, we select a total of 100 images per abstract concept and a grand total of 10,700 images. We set `aesthetic_score = 5` and `aesthetic_weight = 0.5` in clip-retrieval, to obtain images depicting real scenarios and limit the possibility of retrieving artistic illustrations.

Evaluation and Analysis To test whether querying LAION-400M with situations could retrieve images strongly associated with the target abstract concepts, we performed a rating assessment test. Further, we evaluated whether we can leverage the proposed retrieval methodology to analyze the connection between real-world scenarios and the figurative interpretation of images. We chose to test situations from AGS with an average rating above 5 (see Sec.3) to ensure that they were highly representative of the concepts for humans: 54 situations met this criterion. We selected the top-2 images per situation, for a total of 878 images. We employed two distinct multiple-choice tests for evaluation, gathering 60 participants via Prolific for each test (120 participants in total). The stimuli were divided into 12 sub-tests, with 5 annotators dedicated to each.

In *Test 1*, for each visual stimulus, participants were asked to label the image choosing between 4 options to label the image: the associated (i.e., correct) abstract concept and 3 other abstract concepts not associated with the image, used as distractors. In *Test 2*, we used different distractors. Specifically, we used a concrete word representing an object or entities depicted in the image, and an abstract word and a concrete word not associated with the image. Distractors were chosen to exclude synonyms or concepts directly related to the correct option. This format allows us to assess the strength of the relationship between the abstract concept and the image, while still presenting options that could be plausibly related. We adopetd a multiple-choice framework that incorporates distractor options for its robustness against possible bias and vagueness of answers. In fact, visuals inherently carry diverse interpretations, enabling a singular image to be associated with numerous abstract concepts. Introducing a variety of abstract choices allows for a more precise evaluation of the strength of the association between the correct abstract stimulus and the image. Using a rating system that includes only the correct notions could potentially introduce bias, as it pre-defines the links between images and abstract concepts. On the contrary, an open-ended response format might elicit a wide array of answers, which could be less beneficial in analyzing the connection between specific abstract-visual scenario pairs. The lack of constraints in responses to tasks centered on abstract concepts would have resulted in extremely varied outcomes. This is due to the complex nature of abstract concepts and their associative paths, which may manifest differently across contexts and intensities. Typically, unrestricted tasks lead participants to associate abstract concepts with synonyms or to describe images in literal terms initially. Given the study's goal to delve into a specific representation of abstract concepts, specifically their metaphorical interpretation through images, we believe that the use of distractors serves to probe these kinds of associative links more effectively.

Table 2 provides the results for both tests. For Test 1, when we look at *labels distribution*, we see that participants chose the correct abstract concept as label in most cases (69,3%), suggesting a degree of association between the two. To further evaluate their *association strength*, we proceeded as follows: we binarized the results based on whether the associated abstract concept was chosen by at least 50% of participants across all images for the same stimulus. In this case, we can suppose that the images retrieved could be interpreted figuratively via indirect grounding. This evaluation shows that 90.7% of images are vehicle of the associated abstract concepts. In test 2, results indicate that, notwithstanding the presence of a concrete term corresponding to the visual representation in the image, participants selected the correct abstract labels in 29.1% of instances. In this 29.1%, 5.6% have more than 50% of correct abstract labels. These results confirm the variability of visual semantics. Nevertheless, still a fair enough percentage of correct abstract labels were assigned. We can argue that our findings suggest a positive answer for Q2.

	Test 1		Test 2	
	Corr.	Incorr.	Corr.	Incorr.
Labels distribution	**69.3%**	30.7%	29.1%	70.6%
Association strength	**90.7%**	9.3%	5.6%	94.4%

Table 2: % of correct/incorrect labels.

5. Discussion and Conclusions

In this work, our primary aim was to provide a framework for building or enriching linguistic/multimodal resources to delve into the metaphorical grounding of abstract concepts, focusing on its applicability within the Italian linguistic context. We evaluated the abilities of humans and a LLM, namely `Davinci-003` GPT-3 model to generate situations that could provide indirect grounding to abstract concepts.

Our first experiment suggests that LLMs-generated situations are comparable to those generated by humans. Similarity density analysis suggests that the AGS align with the cognitive system's ability to produce diverse conceptualizations for an abstract concept (Barsalou, 2003). The proposed text-to-image retrieval method confirmed that images depicting situations grounding abstract concepts can represent these concepts, adding a small but significant piece to the indirect grounding theories that still lack empirical evidence (Utsumi, 2022). However, the evaluation task conducted with concrete distractors, confirms the idea that visual properties of an image do not always coincide with linguistic properties (Giunchiglia et al., 2023), generating a mismatch between abstract and concrete classification. By analyzing the images that struggled to recall the associated abstract concepts, we also found that retrieved images for these situations (i.e., the AGS query) is hard to be visually represented. For example, the AGS *a child having to face the loss of a parent* generated for the concept *adversity* has retrieved the image of a sad girl, which only partially represents the scenario depicted in the generated sentence. The proposed method may prove to be beneficial for at least two reasons. First, its direct outcome is a way to automatically obtain metaphorically-rich images from dataset aimed at Computer Vision or Language-Vision problems, by exploiting LLMs generative abilities and VLM-based retrieval. Second, this kind of data may become a valuable asset in facing the lack of metaphorical multimodal dataset, to achieve a better understanding of the indirect grounding mechanisms in a multimodal setting. Moreover, it could enrich the abstract concepts understanding capabilities of VLMs by training on metaphorical- and abstract-oriented data. To these ends, we release all the data collected for the present work. Our future plans include expanding the present work by further exploring prompting techniques and the use of LLMs and VLMs. Moreover, we intend to adopt the proposed method to augment existing datasets, either linguistic, visual and multi-modal, with information concerning abstract concepts and figurative interpretations. To enhance our framework, we also plan to broaden the set of abstract concepts beyond the initial 107, and to incorporate more comprehensive measures beyond similarity density for evaluating the generated data. Additionally, we aim to test the framework across various linguistic and cultural systems. We believe that the proposed research framework can indeed be utilized to explore the phenomenon of abstract concept representation via images in other languages as well. This would allow us to ascertain whether cross-cultural differences emerge in the anchoring of abstract terms to situations, or whether similar patterns of metaphorical grounding are observed across different linguistic and cultural landscapes.

Acknowledgments

This research was partly funded by PNRR—M4C2—Investimento 1.3, Partenariato Esteso PE00000013—“FAIR—Future Artificial Intelligence Research”—Spoke 1 “Human-centered AI,” funded by the European Commission under the NextGeneration EU programme.

Bibliographical 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.

Chiara Barattieri di San Pietro, Federico Frau, Veronica Mangiaterra, and Valentina Bambini. 2023. The pragmatic profile of chatgpt: Assessing the communicative skills of a conversational agent. *Sistemi intelligenti, Rivista quadrimestrale di scienze cognitive e di intelligenza artificiale*, (2/2023):379–400.

Lawrence Barsalou. 2003. Situated simulation in the human conceptual system. *Language and Cognitive Processes*, 18(5-6):513–562.

Lawrence Barsalou. 2008. Grounded cognition. *Annual review of psychology*, 59:617–45.

Marianna Bolognesi and Tommaso Caselli. 2022. Specificity ratings for italian data. *Behavior Research Methods*, pages 1–18.

Anna Borghi, Ferdinand Binkofski, Cristiano Castelfranchi, Felice Cimatti, Claudia Scorolli, and Luca Tummolini. 2017. The challenge of abstract concepts. *Psychological Bulletin*, 143.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Marc Brysbaert, Amy Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46.

Sebastian Crutch and Elizabeth Warrington. 2005. Abstract and concrete concepts have structurally different representational frameworks. *Brain : a journal of neurology*, 128:615–27.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. Redcaps: web-curated image-text data created by the people, for the people. *CoRR*, abs/2111.11431.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Guy Dove. 2014. Thinking in words: Language as an embodied medium of thought. *Topics in cognitive science*, 6.

Fausto Giunchiglia, Mayukh Bagchi, and Xiaolei Diao. 2023. Aligning visual and lexical semantics. In *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity*, pages 294–302, Cham. Springer Nature Switzerland.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models.

Jiayi Liao, Xu Chen, and Lun Du. 2023. Concept understanding in large language models: An empirical study.

Jerry Liu, Nathan O'Hara, Alexander Rubin, Rachel Draelos, and Cynthia Rudin. 2020. Metaphor detection using contextual word embeddings from transformers. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 250–255, Online. Association for Computational Linguistics.

Max M. Louwerse. 2011. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2):273–302.

M. Mohler, D. Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. pages 27–35.

A. Paivio. 1990. *Dual Coding Theory. In Mental Representations: A Dual Coding Approach*. Oxford: Oxford University Press.

Malay Pramanick, Ashim Gupta, and Pabitra Mitra. 2018. An LSTM-CRF based approach to token-level metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 67–75, New Orleans, Louisiana. Association for Computational Linguistics.

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

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021a. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021b. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

S. M. Seals and Valerie L. Shalin. 2023. Long-form analogies generated by chatgpt lack human-like psycholinguistic properties.

Ekaterina Shutova, Lin Sun, Elkin Gutiérrez, Patricia Lichtenstein, and Srini Narayanan. 2016. Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. *Computational Linguistics*, 43:1–88.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. volume 2, pages 1002–1010.

Akira Utsumi. 2022. A test of indirect grounding of abstract concepts using multimodal distributional semantics. *Frontiers in psychology, 13, 906181*.

W.Yeh and LW. Barsalou. 2006. The situated nature of concepts. *J Psychol.*, pages 119(3):349–84.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. MultiMET: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3214–3225, Online. Association for Computational Linguistics.