Evaluating Semantic Relations in Predicting Textual Labels for Images of Abstract and Concrete Concepts

Tarun Tater[1], **Sabine Schulte im Walde**[1], **Diego Frassinelli**[2,3]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Department of Linguistics, University of Konstanz, Germany
[3]Center for Information and Language Processing, LMU Munich, Germany
{tarun.tater, schulte}@ims.uni-stuttgart.de
frassinelli@cis.lmu.de

Abstract

This study investigates the performance of SigLIP, a state-of-the-art Vision-Language Model (VLM), in predicting labels for images depicting $1,278$ concepts. Our analysis across 300 images per concept shows that the model frequently predicts the exact user-tagged labels, but similarly, it often predicts labels that are semantically related to the exact labels in various ways: synonyms, hypernyms, co-hyponyms, and associated words, particularly for abstract concepts. We then zoom into the diversity of the user tags of images and word associations for abstract versus concrete concepts. Surprisingly, not only abstract but also concrete concepts exhibit significant variability, thus challenging the traditional view that representations of concrete concepts are less diverse.

1 Introduction

Concrete concepts, such as *apple* and *dog*, are easily perceivable through our senses, whereas abstract concepts such as *happiness* and *justice* lack physical referents and are not directly linked to our sensory experiences (Paivio et al., 1968; Brysbaert et al., 2014). These differences have played a crucial role in various applications involving both textual and multi-modal inputs (Turney et al., 2011; Tsvetkov et al., 2013; Köper and Schulte im Walde, 2016; Köper and Schulte im Walde, 2017; Cangelosi and Stramandinoli, 2018; Su et al., 2021; Ahn et al., 2022). Recent advances in multi-modal learning with Vision-Language Models (VLMs) like CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and SigLIP (Zhai et al., 2023), have improved the alignment of textual and visual data to generate context-aware representations. However, the ability of VLMs to capture semantic relationships between concepts and their visual representations remains underexplored. For example, the concept *idea* is semantically related to the synonym *thought* and the hyponym *belief*, and associated with *invention*. This example highlights the range of possible labels for the visual representation of a concept and the importance of including not only human-assigned labels but also their semantically related counterparts to enhance applications like image retrieval and visual question answering. Inspired by this, we evaluate how well SigLIP, a state-of-the-art VL model, predicts image labels that are generated by users, as well as their synonyms, hypernyms, co-hyponyms, and associative words. Assessing the impact on model performance, we aim to determine if integrating these relations into VLM training can potentially improve the representation of abstract and concrete concepts.

People use a variety of cues, including visual and linguistic, to perceive and understand concepts (Lynott et al., 2020). Traditionally, concrete concepts, which are directly related to sensory experiences, are considered less diverse in their visual representations compared to abstract concepts (Hessel et al., 2018; Kastner et al., 2019), and are expected to have more consistent tags and associations. In contrast, abstract concepts, being inherently diverse, are expected to have varied word associations and user tags, reflecting the complexity of understanding these concepts across modalities. To evaluate these differences, we pose the following questions:

RQ1: How do different semantic relations of user tags affect the prediction of image labels for abstract and concrete concepts?

RQ2 How do user tags given a visual cue (image), and word associations given a linguistic cue, differ in characterizing abstract versus concrete concepts?

Our findings show that SigLIP often predicts semantically related labels (such as hypernyms) instead of the original user tags. Our analysis of association data and user tags reveals that concrete concepts, like abstract ones, invoke a diverse range of descriptors, challenging the traditional view of less diversity.

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 214–220

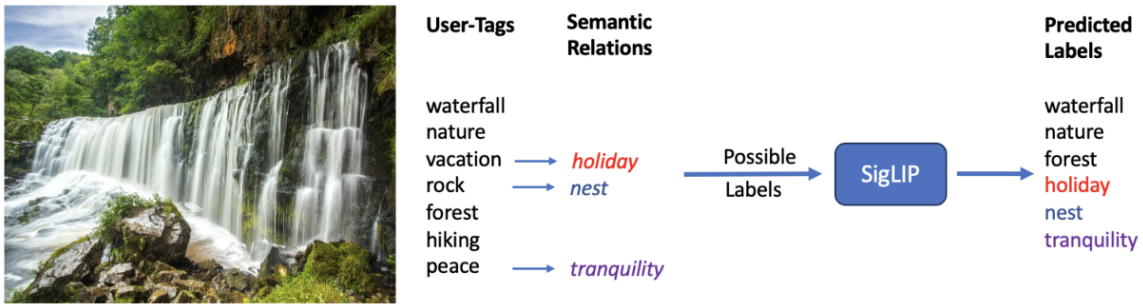

Figure 1: Example of an image and the corresponding user tags. Here, *holiday* is a synonym of *vacation*, *nest* is a co-hyponym of *rock*, and *tranquility* is a hypernym of *peace*. For this image, the SigLIP model might predict *waterfall*, *nature*, *forest*, *holiday*, *nest*, *tranquility*.

2 Experimental Design

2.1 Materials

Target Concepts & Concreteness Norms: We selected concrete and abstract nouns using the concreteness ratings from Brysbaert et al. (2014). The ratings range from 1 (abstract) to 5 (concrete) and were collected via crowdsourcing. We utilized the filtered dataset of $5,438$ nouns from Schulte im Walde and Frassinelli (2022) to reduce ambiguity through frequency thresholds and POS tagging. To better understand the differences between the two extremes of the concreteness spectrum, we focused on the most concrete and most abstract nouns. At the same time, we wanted to ensure a sufficient number of nouns from both extremes with at least 300 images available for each concept. However, acquiring a sufficient number of images (300) was challenging for many abstract nouns. Therefore, we selected concrete nouns rated from $4.5 - 5$, and used a broader range for abstract nouns from $1 - 2.5$. From these, we excluded all nouns which occur in the $1,000$ classes of the ILSVRC-2012 ImageNet dataset (Russakovsky et al., 2015), since many vision models are trained or evaluated on these classes. We also filtered out nouns that could lead to images depicting explicit content, as well as the nouns *camera*, *picture* and *photo* which were very common user tags because of the nature of the dataset.

Image Dataset and user tags: We used images from the YFCC100M Multimedia Commons Dataset (*YFCC*; Thomee et al. (2016)), the largest publicly available user-tagged dataset containing ≈ 100 million media objects from Flickr. Each image has tags provided by users (user tags) when uploading the image. For example, Figure 1 is an image with the corresponding possible user tags: *waterfall*, *nature*, *vacation*, *rock*, *forest*, *hiking*, *peace*. We only retained user tags that consisted solely of English characters. We randomly selected 300 images where the target concept appeared among the user tags and consider them as relevant images of that concept. This resulted in $1,278$ nouns (371 abstract and 907 concrete) with 300 images each.

Semantic Relations and Associations: We used WordNet (Miller, 1995) to extract synonyms, hypernyms and co-hyponyms for each user tag, and utilized association norms from De Deyne et al. (2019), which were gathered by prompting annotators to provide three words that came to mind for a given word. For example, the concept *idea* might have associations like *thought*, *bulb*, and *invention*. We restricted our analysis to nouns that were assessed by at least 100 annotators to ensure enough annotations, filtering our set to 682 nouns (527 concrete and 155 abstract) for RQ2.

2.2 Models and Evaluation

In this study, we perform multi-label classification, where each image can have multiple relevant labels. For instance, for Figure 1, the SigLIP model might predict *waterfall*, *nature* and *forest* as labels of the image[1]. Our goal is to evaluate how well the SigLIP model predicts either these user tags or their semantically related words as labels. We utilize the SigLIP model (Zhai et al., 2023), the only publicly available pre-trained multi-label classification VLM specifically trained with a contrastive sigmoid loss designed to align text and images. For each image, we evaluate various semantic relations as labels, including synonyms, hypernyms, co-hyponyms and association words in separate experiments. SigLIP assigns a score to each label,

[1]Please note this is only a walk-through example and the actual results may vary.

Concept class	Avg. number of user tags	Avg. number of noun user tags	Avg. % of tags pred. as labels	Avg. % images where no label was pred.	Avg. % of images where target concept not pred.
Abstract	8.69	5.40	54.41	8.06	48.87
Concrete	6.85	4.58	62.17	4.11	26.93

Table 1: User-tag prediction (pred.) results for SigLIP model with 300 images of $1,278$ concepts.

Semantic Relation	Concept class	Avg. number of user tags with semantic relations	Avg. % of labels pred.	Avg. % images where ≥ 1 tag not pred. but their relation pred.	Avg. % user tag not pred. but their semantic relation pred.	Avg. % of images where no label was pred.
Hypernym	Abstract	80.67	27.96	76.40	40.06	3.39
	Concrete	71.48	28.66	66.94	32.02	1.39
Co-hyponym	Abstract	684.04	26.51	70.25	33.23	1.56
	Concrete	608.55	27.44	55.28	22.97	1.00
Synonym	Abstract	41.15	38.00	53.24	18.60	7.14
	Concrete	33.73	44.26	38.61	12.00	4.26

Table 2: SigLIP prediction (pred.) results when considering semantic relations of user tags as labels.

and those with a score ≥ 0.0001, we consider as predicted labels. This threshold is chosen to ensure that only the most relevant tags are considered. In our experiments, we compare synonyms, hypernyms, and co-hyponyms for the subset of noun tags. The primary evaluation metrics include the average percentage of labels predicted, the average percentage of images where no label was predicted, the average percentage of images where at least one user tag was not predicted but its semantic relation was, and the percentage of user tags not predicted as labels, but whose semantically related tags were predicted as labels.

3 RQ1 - Impact of Semantic Relations on Model Predictions

We first analyze the number of user tags associated with each image and the model's performance in predicting these tags as labels. Then, we evaluate whether the model also predicts synonyms, hypernyms, and co-hyponyms of user tags as possible labels. We hypothesize that synonyms will provide alternative labels that capture variations in naming, potentially improving prediction accuracy for both abstract and concrete concepts. Hypernyms are expected to offer more general category labels. Co-hyponyms will highlight sibling relationships between concepts, improving label prediction by capturing related yet distinct categories.

Table 1 presents the results of user-tag predictions comparing abstract and concrete concepts, when using 300 images.[2] We found that images associated with abstract concepts tend to have more user tags on average (8.69, with 5.40 being nouns) compared to those associated with concrete (6.85, with 4.58 being nouns). The model successfully identified a higher percentage of labels for images associated with concrete concepts (62.17%) than for abstract concepts (54.41%). There was a low ratio of images where no user tag was predicted as a label: 8.06% for abstract and 4.11% for concrete concepts. Notably, a high percentage of images did not have the target concept predicted (which we associated the image with): 48.87% for abstract and 26.93% for concrete concepts. These findings suggest that SigLIP struggles to consistently label images with the same tags used by humans. This is especially true for images of abstract concepts, highlighting the difficulty in aligning model predictions with human annotations for these concepts. The findings also point out the diversity involved in tagging abstract concepts, thus emphasizing the importance of accepting a wider selection of relevant labels for multimodal representations.

[2]We also analyzed concepts with 400 available images to check for sampling bias. We did not use them in the main study as it resulted in losing many abstract nouns and causing class imbalance. Overall, we find similar results for 400 images and include them in the Appendix.

Class	Avg. number of unique user tags	Avg. number of unique associations	Association not in user tags	Association predicted	Association predicted for at least one image
Abstract	751	36.00	64.96%	27.89%	99.67%
Concrete	747	33.75	46.79%	38.68%	99.66%

Table 3: Overlap between user tags and word associations.

Table 2 presents the model's performance in predicting synonyms, hypernyms and co-hyponyms of user tags as potential labels. The number of possible labels increases considerably when considering semantically related words of user tags, especially regarding co-hyponyms (684.04 for abstract and 608.55 for concrete concepts). Synonyms had the highest percentage of labels predicted, especially for concrete concepts (44.26%), indicating that the model better captures meaning variations for concrete nouns. For abstract concepts, a majority of images had at least one user tag not predicted, but their hypernym was (76.40%). Abstract concepts also had a higher percentage of labels (40.06%) where hypernyms were predicted but original user tags were not. Similarly, co-hyponyms also showed that abstract concepts (70.25%) had a higher percentage of images where at least one user tag was not predicted but a co-hyponym was, compared to concrete concepts (55.28%). This suggests that concepts, especially abstract, could benefit considerably from broader categorical information provided by hypernyms and co-hyponyms. There are very few images where no label was found, ranging from 1.00% to 7.14% when considering different semantically related words. Overall, our results highlight the importance of considering semantic relations as a possible means to improve the robustness and accuracy of multi-modal models for both abstract and concrete concepts.

4 RQ2 - Relationship between Association Norms and User Tags

We analyzed the overlap between user tags for images and word associations. We expected abstract concepts to show higher diversity in both visual associations (user tags of images) and word associations due to their inherently diverse nature, while we expected concrete concepts to show a larger overlap between user tags and word associations. We analyzed 682 concepts (527 concrete and 155 abstract) for which we had 300 images and 100 annotators. From De Deyne et al. (2019), we selected association words with a frequency of ≥ 2. For each image where the target concept was one of the user tags, we evaluated how well SigLIP predicts the associated words of the target concept as possible labels.

Table 3 presents the number of unique associations and user tags, and the performance of SigLIP for multi-label classification of images considering association words as possible labels. Contrary to our hypothesis, the average number of unique user tags for both abstract (751) and concrete (747) concepts across 300 images is similar. This indicates that people associate diverse words with both abstract and concrete concepts when tagging images, suggesting high diversity in visual interpretation even for concrete concepts, which are traditionally considered less vague and less diverse. However, abstract concepts have a slightly higher average number of unique associations (36.00) compared to concrete concepts (33.75), indicating a slightly greater associative diversity for abstract concepts. This shows that when users are presented with images of a concept, they produce a greater variety of descriptive tags than when producing associations, highlighting the impact of visual context on the descriptive process. It is important to note that this comparison is a bit skewed because our dataset averages 7 tags per image for 300 images of a concept, while word associations are gathered from 100 annotators with 3 associations each per concept. Another surprising finding is the proportion of associations not present as user tags: 64.96% for abstract concepts and 46.79% for concrete concepts. This discrepancy highlights that, despite the directly perceivable nature of concrete concepts, they evoke different personal or contextual mental associations that may not directly translate into visual depictions and vice-versa, similar to abstract concepts. This suggests that concrete concepts possess more semantic diversity than what is visually observable. The SigLIP model on average predicted 27.89% associations as labels for images associated with abstract concepts vs. 38.68% for concrete concepts. However, almost all the asso-

ciations (99.67% for abstract and 99.66% for concrete) were predicted for at least one of the 300 images associated with each concept, indicating that the model could recognize the majority of associations as labels across different images.

These findings point towards the need to further explore whether and how current models trained on user-generated tags might fail to capture the full range of human conceptual associations. Models might benefit from integrating these broader associative data to enhance their understanding and representation of concepts.

5 Conclusion

Our study highlights the potential of integrating diverse semantic relationships in improving the representations in Vision-Language Models (VLMs), particularly SigLIP, for abstract and concrete concepts. The results demonstrate that for images associated with both abstract and concrete concepts, SigLIP often predicts semantically related words such as synonyms, hypernyms, and co-hyponyms of a user tag even when the user tag itself is not predicted as a label. Furthermore, the distinction between visual and linguistic associations revealed differences in how these concepts are perceived and described. Our findings suggest that leveraging semantic relationships and associations should be further explored to enhance representations of abstract and concrete concepts in VLMs, aligning them more closely with human cognitive variation.

Limitations

Our findings are based on the SigLIP model, other VLMs may yield different results. Additionally, we have considered all labels from SigLIP with probability scores ≥ 0.0001, and the results may vary if a different threshold is considered. Also, another selection of images may introduce some variability into the results.

Ethics Statement

We anticipate no ethical concerns with this work. All modeling experiments utilized open-source libraries, which have been appropriately cited.

Acknowledgements

This research is supported by the DFG Research Grant SCHU 2580/4-1 *Multimodal Dimensions and Computational Applications of Abstractness*.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022. Do as I can, not as I say: Grounding language in robotic affordances. arXiv:2204.01691v2.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 64:904–911.

Angelo Cangelosi and Francesca Stramandinoli. 2018. A review of abstract concept learning in embodied agents and robots. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1752):20170131.

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51:987–1006.

Jack Hessel, David Mimno, and Lillian Lee. 2018. Quantifying the visual concreteness of words and topics in multimodal datasets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2194–2205, New Orleans, Louisiana. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916.

Marc A Kastner, Ichiro Ide, Yasutomo Kawanishi, Takatsugu Hirayama, Daisuke Deguchi, and Hiroshi Murase. 2019. Estimating the visual variety of concepts by referring to web popularity. *Multimedia Tools and Applications*, 78:9463–9488.

Maximilian Köper and Sabine Schulte im Walde. 2016. Distinguishing literal and non-literal usage of German particle verbs. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362, San Diego, CA, USA.

Maximilian Köper and Sabine Schulte im Walde. 2017. Complex verbs are different: Exploring the visual modality in multi-modal models to predict compositionality. In *Proceedings of the 13th Workshop on Multiword Expressions*, pages 200–206, Valencia, Spain.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The Lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52:1–21.

George A Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Allan Paivio, John C. Yuille, and Stephen A. Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology (Monograph Supplement)*, 76(1/2):1–25.

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.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

Sabine Schulte im Walde and Diego Frassinelli. 2022. Distributional measures of abstraction. *Frontiers in Artificial Intelligence: Language and Computation 4:796756. Alessandro Lenci and Sebastian Padó (topic editors): "Perspectives for Natural Language Processing between AI, Linguistics and Cognitive Science".*

Chang Su, Weijie Chen, Ze Fu, and Yijiang Chen. 2021. Multimodal metaphor detection based on distinguishing concreteness. *Neurocomputing*, 429:166–173.

Bart Thomee, Benjamin Elizalde, David Shamma, Karl Ni, Gerald Friedland, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59:64–73.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51, Atlanta, Georgia.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, UK.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

6 Appendix

6.1 Extending association norms analysis

In our main study, we consider association words for a noun if they have a frequency of ≥ 2, meaning the association is provided by at least 2 annotators. Here, we validate our findings by incorporating all possible association words, without a frequency threshold. The results, presented in Table 4, are similar to those in Table 3. The number of unique associations for abstract concepts (130) remains higher than for concrete concepts (105), and are also high for concrete concepts. A majority of association words (80% for abstract and 69% for concrete concepts) do not appear among the user tags associated with the target concepts, thus demonstrating the gap between associations given a linguistic cue and user tags given an image (visual cue). Similar to Table 3, almost all associations were predicted for at least one image, out of the 400 images associated with the target concept.

Class	Avg. unique assoc.	Assoc. not in user tags	Assoc. pred.	Assoc. pred. for any image
Abstract	130	80%	25%	99.28%
Concrete	105	69%	31%	98.94%

Table 4: Overlap between user tags and word associations.

6.2 Multi-label prediction with 400 images

To ensure that our findings were not influenced by sampling bias, we also experimented with the subset of concepts where 400 images were available. These are 1,191 concepts with 400 images (864 concrete and 327 abstract). We present the results in Table 5. The results are similar to when considering 300 images for each concept.

Semantic Relation	Concept class	Avg. number of user tags with semantic relations	Avg. % of labels pred.	Avg. % of images where no label was pred.	Avg. % images where ≥ 1 tag not pred. but their relation pred.	Avg. % user tag not pred. but their semantic relation pred.
Hypernym	Abstract	80.98	27.68	3.40	76.35	39.97
	Concrete	71.50	28.59	1.39	67.00	32.04
Co-hyponym	Abstract	692.67	26.41	1.62	70.23	33.23
	Concrete	606.32	27.40	0.98	55.22	22.97
Synonym	Abstract	41.57	37.81	7.26	52.97	18.47
	Concrete	33.79	44.21	4.30	38.68	12.03

Table 5: SigLIP prediction (pred.) results when considering semantic relations of user tags as labels for 400 images per concept.