Evaluating Textual and Visual Semantic Neighborhoods of Abstract and Concrete Concepts

Sven Naber[1], Diego Frassinelli[2], and Sabine Schulte im Walde[1]

[1]Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart
[2]MaiNLP, Center for Information and Language Processing, LMU Munich
{sven.naber, schulte}@ims.uni-stuttgart.de
frassinelli@cis.lmu.de

Abstract

This paper presents a systematic evaluation of nearest neighbors across semantic representation spaces in both textual and visual modalities. We focus on nominal concepts with varying concreteness levels, and apply a neighborhood overlap measure to compare these target concepts differing in their linguistic and perceptual nature. We find that alignment is primarily determined by modality, and additionally by level of concreteness: Models from the same modality show stronger alignment than cross-modal models, and spaces of concrete concepts show stronger alignment than those of abstract ones. Overall, larger neighborhood size strengthens the alignment between spaces.

1 Introduction

Humans possess an intuitive understanding of concepts and their relative positions within a shared semantic space. For instance, we perceive *cat* as being more similar to *dog* than to *table*. This sense of conceptual proximity is grounded in real-world knowledge and similarity across multiple attribute dimensions—such as form (e.g., four-legged animals), category membership (e.g., pets, furniture), and function (Rosch, 1973; Talmy, 1983).

In natural language processing, prototypical attributes and similarity between concepts are generally captured by vector representations and vector distance measures. At the textual level, these representations are learned either through count-based methods such as co-occurrence matrices (Turney and Pantel, 2010), or predictive models such as shallow neural networks trained on target or context prediction objectives (Mikolov et al., 2013; Pennington et al., 2014a). In these vector spaces, relative position and structure encode semantic relatedness. Similarly, vision models such as convolutional neural networks (Krizhevsky et al., 2017) and vision transformers (Dosovitskiy et al., 2020) map images into vector spaces where proximity reflects not only semantic but also perceptual similarity (Battleday et al., 2020). While text embeddings are based on distributional patterns in language, image embeddings are grounded in visual features—such as shape, color, and scene (Krizhevsky et al., 2017).

Level	Range	Concepts
abstract	1.0 – 2.0	*idea*, *justice*
mid-scale	3.0 – 4.0	*story*, *election*
concrete	4.8 – 5.0	*apple*, *car*

Table 1: Concreteness ranges and example concepts based on Brysbaert et al. (2014).

In the current study, we build and compare representations in textual and visual modalities regarding concrete vs. abstract concepts, which differ in their perceptual nature (vision, sound, smell, taste, touch): Concrete concepts such as *apple* are more easily grounded in perceptual features, in contrast to abstract concepts such as *idea*, which lack stable visual referents and are stronger connected to linguistic context (Paivio, 1971; Andrews et al., 2009; Pecher et al., 2011; Frassinelli and Lenci, 2012; Brysbaert et al., 2014; Frassinelli et al., 2017; Naumann et al., 2018; Tater et al., 2024, 2025).

We conduct a systematic analysis of semantic attributes across modalities and across the abstractness-concreteness continuum, by relying on a simple and interpretable nearest-neighbor overlap to capture embedding space alignment between concepts. We demonstrate that modality is indeed the primary factor shaping semantic neighborhoods, with stronger alignment within modalities for both concrete and abstract concepts, while cross-modal agreement is stronger for concrete than for abstract concepts.

2 Data and Methods

In this section, we describe the target concepts, the variants of textual and visual representations, and how we identify nearest neighbors.

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 142–151

Target Concepts This study uses a curated set of 1,500 nouns drawn from the Brysbaert concreteness ratings (Brysbaert et al., 2014). To guarantee clear distinctions between sets of concepts, this set consists of 500 nouns each from three distinct levels of concreteness: extremely abstract, intermediate mid-scale, and extremely concrete noun concepts (see Table 1 for concreteness ranges and example concepts). As nearest neighbor candidates for the target concepts we include a larger set of 5,448 nouns from the Brysbaert ratings, as constructed in our previous work (Schulte im Walde and Frassinelli, 2022).

Distributional Word Representations We built textual representations for our target concepts and nearest neighbor candidates by training both count-based and predictive word embedding models on the ENCOW16AX corpus (Schäfer and Bildhauer, 2012; Schäfer, 2015). **Count-based embeddings** were created with a symmetric window of ±20 nouns, verbs and adjectives occurring at least 50 times in the corpus; each word is represented by a 46,716-dimensional vector. **GloVe embeddings** were created using a symmetric window of ±15 and a minimum frequency of 5, relying on the original implementation in Pennington et al. (2014b); each word is represented by a 50-dimensional embedding. **Word2Vec embeddings** were created with a context window of ±5 and a minimum frequency of 5 using the skip-gram objective over 10 epochs with the Gensim package (Řehůřek and Sojka, 2010); each word is represented by a 300-dimensional embedding. **FastText embeddings** were created with a context window of ±5 and a minimum frequency of 2 over 5 epochs using the FAIR FastText implementation (Bojanowski et al., 2017); each word is represented by a 100-dimensional embedding.

Sentence Representations We built a second set of transformer-based textual representations for our target concepts, as a more direct point of comparison with representations based on vision-transformers (see below), by mean-aggregating embeddings of 35 sentences of each noun retrieved from the ENCOW16AX corpus. Embeddings for each sentence were generated from the models with the SentenceTransformer package (Reimers and Gurevych, 2019). **Mpnet embeddings** uses a pre-trained mpnet-base (Song et al., 2020) finetuned on 1B sentence pairs. We use `all-mpnet-base-v2` (Hugging Face, 2021); the resulting embeddings have a dimensionality of 768. **Gemma embeddings** is a 300M parameter embedding model derived from Gemma 3 (Gemma Team et al., 2025). We use `google/embeddinggemma-300m` (Schechter Vera et al., 2025); the resulting embeddings have a dimensionality of 768. **Qwen3 embeddings** use a 0.6B parameter embedding model from the Qwen model family (Yang et al., 2025). We rely on `Qwen/Qwen3-Embedding-0.6B` (Zhang et al., 2025); the resulting embeddings have a dimensionality of 1024.

Visual Representations We built visual representations relying on the top 35 images for each noun using Bing Image Search (Microsoft Corporation, 2025). Images identified as corrupted or irrelevant were automatically replaced. Image quality was probed as described in Appendix A.1. The visual embedding of each concept was created by mean aggregating the `[CLS]` token embeddings, by passing through the respective models the top-n images (with $1 \leq n \leq 35$). **Vision Transformer (ViT)** embeddings use the model `google/vit-base-patch16-224-in21k` (Google, 2021). The resulting embeddings have a dimensionality of 768, serving as a baseline for concept representation from a pure vision transformer trained in a supervised manner. **DINOv2 embeddings** represent concepts using a vision transformer trained in a self-supervised manner. We rely on the `facebook/dinov2-base` model (Oquab et al., 2024; FacebookAIResearch, 2023); the resulting embeddings have a dimensionality of 768. **Hiera embeddings** represent concepts using a vision transformer with a hierarchical architecture trained in a supervised manner. We rely on the `facebook/hiera-large-224-hf` model (FacebookAIResearch, 2024); the resulting embeddings have a dimensionality of 768. **CLIP embeddings** represent concepts using a vision transformer trained to align images and captions with contrastive loss. We rely on the image encoder of the `openai/clip-vit-base-patch32` (OpenAI, 2021; Radford et al., 2021) model; these embeddings have a dimensionality of 512.

Nearest Neighbor Identification We identify the nearest neighbors of a specific concept within a specific semantic embedding space by calculating the cosine between vector representations of the concept and each nearest neighbor candidate in that space, and then sorting the neighbors by decreasing cosine score.

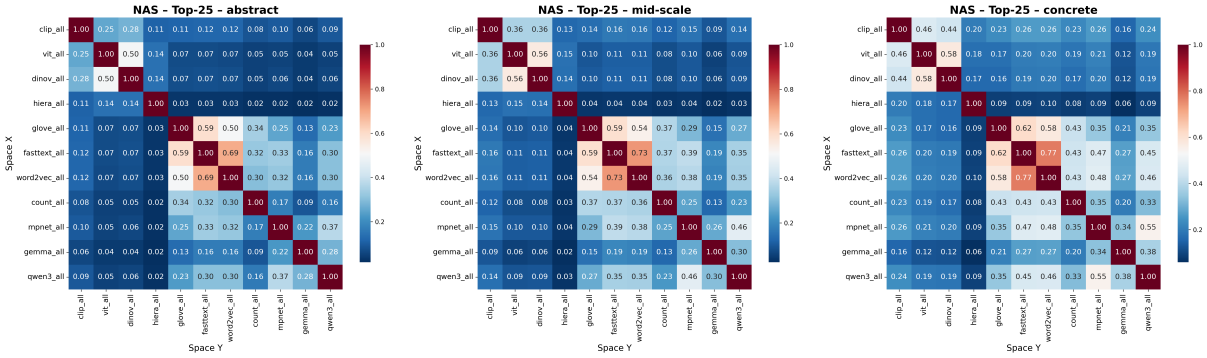

Figure 1: Neighborhood overlap (top-25) across representations for abstract, mid-scale, and concrete concepts.

We vary both the number and start rank of the nearest neighbors we sample, by including the nth to kth nearest neighbors, $1 \leq n \leq k \leq 100$; for example, with $n = 1$ and $k = 5$ we would include the 5 nearest neighbors, and with $n = 5$ and $k = 10$ we would include the 5th to 10th nearest neighbors.

At the core of comparing nearest neighbors across textual and visual semantic representations, we apply a measure of neighborhood overlap in the following way. First, we calculate the overlap $O_c^{n,k}$ of nearest neighbors for a concept c, where $N_{space_i}(c)[n:k]$ denotes the set of neighbors of c ranked from n to k (inclusive) in a given space i:

$$O_c^{n,k} = \frac{|N_{space_1}(c)[n:k] \cap N_{space_2}(c)[n:k]|}{k - n + 1} \quad (1)$$

The concept-specific overlap scores $O_c^{n,k}$ are then aggregated by averaging across all target concepts c in the set C (e.g., the set of all abstract concepts):

$$O_{obs} = \frac{1}{|C|} \sum_{c \in C} O_c^{n,k} \quad (2)$$

In order to interpret these observed overlap scores O_{obs} relative to random chance, we define a normalized alignment score (NAS) by taking into account the expected overlap O_{rand} between two random sets of x neighbors from a candidate set of N concepts. This reflects the expected proportion of shared neighbors for two spaces with random concept positions:

$$O_{rand} = \frac{x}{N - 1} \quad (3)$$

Our NAS score rescales the observed overlap O_{obs} against the expected overlap O_{rand}. A score of 0 indicates overlap by chance; 1 indicates perfect alignment in the observed neighborhood band.

$$\text{NAS} = \frac{O_{obs} - O_{rand}}{1 - O_{rand}} \quad (4)$$

Significance Testing We evaluate whether a single alignment score is significantly greater than chance using a non-parametric sign-flip permutation test which assumes the null hypothesis of symmetric scores around. Given a score vector $\boldsymbol{x} = (x_1, \ldots, x_N)$ across N concepts we compute the observed mean and compare it to a null distribution obtained by randomly flipping signs of x_i across $B = 10{,}000$ permutation. The two-sided p-value is the proportion of permuted means at least as extreme as the observed.

Because our concept sets are large (N=5,448 for the full set of neighbor candidates from Brysbaert, and N=500 for each concreteness band), and because of a standard error of mean scaling with $1/\sqrt{N}$ (Good, 2004), statistically significant alignment differences are at the level of $\overline{\text{NAS}} \approx 0.01$.

3 Experiments and Results

Given that we are interested in overlapping vs. complementary nearest neighbors of abstract and concrete concepts in various modality spaces, we report as results and main insights variants of pairwise comparisons relying on NAS.

Effect of Models and Modalities Figure 1 shows NAS scores for all model pairings across the three concept sets (abstract, mid-scale and concrete, from left to right) at a fixed neighborhood size of 25. The top and left-most four models in the matrices refer to the visual representations, the bottom and right-most seven (four distributional word and three sentence representations) refer to the textual ones.

Modality is clearly the overall dominant factor in alignment scores, i.e., across concreteness levels, model pairs from the same modality (text-text and image-image, cf. top left and bottom right parts of matrices) show substantially stronger overlap than cross-modal

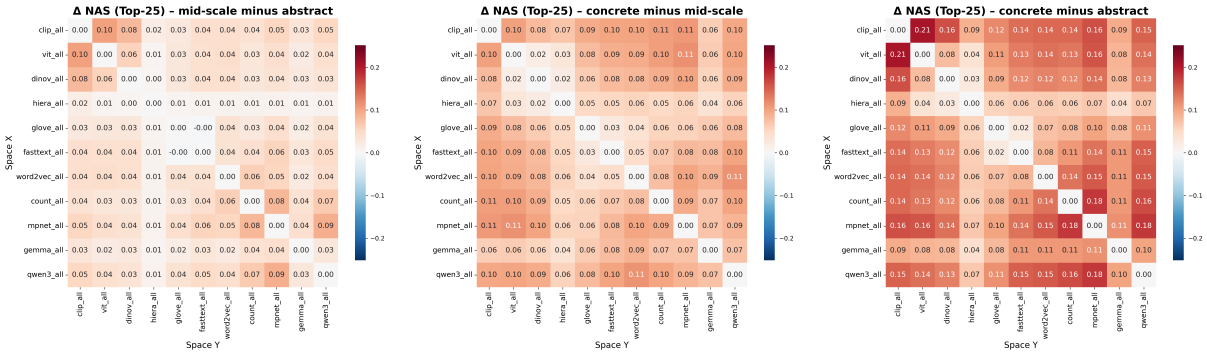

Figure 2: Neighborhood overlap (top-25) differences: $\Delta_{\text{mid}-\text{abstract}}$, $\Delta_{\text{concrete}-\text{mid}}$, and $\Delta_{\text{concrete}-\text{abstract}}$.

pairs: Across abstract/mid-scale/concrete concept sets, alignment scores of image-image spaces reach $0.50/0.56/0.58$ (ViT/DINOv2), and alignment scores of text-text spaces reach $0.69/0.73/0.77$ (Word2Vec/FastText), respectively, while across-modality alignment scores only reach $0.14/0.16/0.26$.

The architectural differences of models within a modality also influence the alignment (with strongest within-modality alignments for word-based distributional models, and weakest ones for sentence-based representations), however to a lesser degree. Overall, the presented alignment patterns thus indicate that while semantic spaces are structured similarly within the same modality, visual and textual modalities shape them in fundamentally different ways.

The modality gap is strongest for abstract concepts, which are inherently more challenging to ground in visual features. For example, the 3×3 top left parts of the matrices in Figure 1 (image-image alignment) show increases from alignment range $[0.25, 0.50]$ for abstract concepts to $[0.36, 0.56]$ for intermediate and $[0.44, 0.58]$ for concrete concepts; the 4×4 middle part of the matrices (distributional text-text alignment, which also presents the overall strongest alignments) shows increases from alignment range $[0.50, 0.69]$ for abstract concepts to $[0.54, 0.73]$ for intermediate and $[0.58, 0.77]$ for concrete concepts.

We also find – but to a more subtle degree – that both within and across modalities the alignment is consistently higher for concrete concepts in comparison to mid-scale concepts, where it is again higher than for abstract concepts. Figure 2 illustrates this pattern across concreteness levels, by showing the differences (Δ scores) between alignment scores of mid-scale vs. abstract concept sets (left), concrete vs. mid-scale sets (middle) and concrete vs. abstract sets (right). We can see that the Δ scores strongly increase from left to right, with differences of up to 0.21 for image-image alignment and up to 0.18 for text-text alignment. This confirms prior research that concrete concepts tend to have less diverse neighbors than abstract concepts (Recchia and Jones, 2012; Kiela et al., 2014; Danguecan and Buchanan, 2016; Reilly and Desai, 2017; Naumann et al., 2018; Schulte im Walde and Frassinelli, 2022; Tater et al., 2024).

Effect of Image Aggregation We now look into the effect of increasing the number of images used for creating visual representations. Figure 3 compares the top-100 overlap of ViT embeddings with an increasing number of images against textual embeddings from FastText, and visual embeddings from DINOv2 and CLIP. The plot illustrates that aggregating image embeddings via mean pooling notably increases alignment scores when adding more images. This development can be attributed to the aggregation process mitigating the multiplicity and saliency issues, which are inherent in any visual representation – challenges which are more pronounced when visual cues are more variable. Accordingly, the aggregation effect is stronger within the vision modality, as it does not affect the gap between the difference in concept representations across modalities. Appendix A.2 further explores the differences of image aggregation across concreteness levels.

Effect of Neighborhood In our last analysis, we explore the role of neighborhood sizes, by computing NAS profiles for increasing neighborhoods. The plots in Figure 4 shows these profiles for selected model pairings within and across modalities as well as across concreteness level: abstract, mid-scale, concrete, and all concepts.

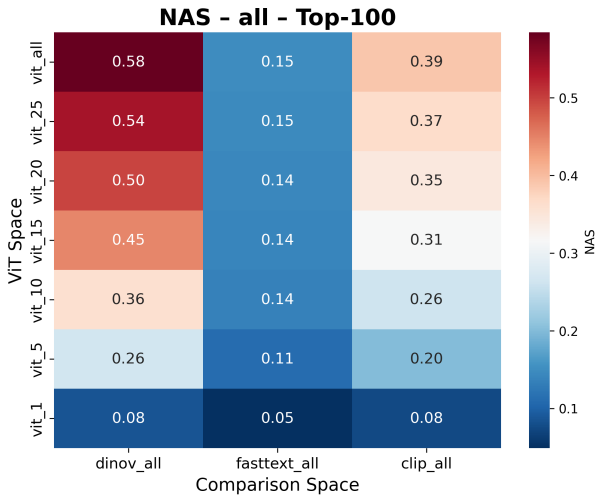

Figure 3: Neighborhood overlap (top-100) across representations and concepts using 1–35 images.

We can see that all alignments involving text models (top three) show a recurring pattern: the alignment is clearly highest for concrete concepts (green lines), and it is lowest for abstract concepts (orange lines), while alignments for mid-scale concepts and across all concepts (red/blue, respectively) are in-between. For the image-image pairing (bottom), the picture is slightly different for concrete concepts, whose alignment declines relative to the other concepts with increasing neighborhood size. Across all pairings, the alignment tends to monotonically increase with larger neighborhood sizes, except in the case of concrete concepts in image-image pairings. The above observations hold for most model pairings.

Overall, our analysis hints at (i) higher variability of immediate neighbors, while larger neighborhoods capture broader semantic similarities, and (ii) more variability in neighborhoods of more abstract in comparison to more concrete concepts. The contrast between within- and cross-modal alignment remains, thus reaffirming the impact of modality on semantic structure.

4 Conclusion

We presented a systematic evaluation of nearest neighbors of abstract and concrete concepts by applying a simple, interpretable, modality- and model-agnostic metric to comparing a variety of textual and visual semantic embedding spaces. Our results confirm that modality is the primary factor shaping semantic structure: Alignments of neighborhoods within the same modality are stronger than alignments across modalities. Specifically focusing on concepts across concreteness levels, we found that concrete concepts show higher alignment of semantic space neighbors than abstract ones, which confirms the difference in perceptual strength in the visual domain, where grounding them is especially challenging. A mean aggregation of images strengthened the alignment with diminishing returns beyond 20–25 images per concept; also, larger neighborhood sizes evoked stronger alignments. Our findings provide a foundation for further analysis of cross-model and cross-modal differences in meaning representation.

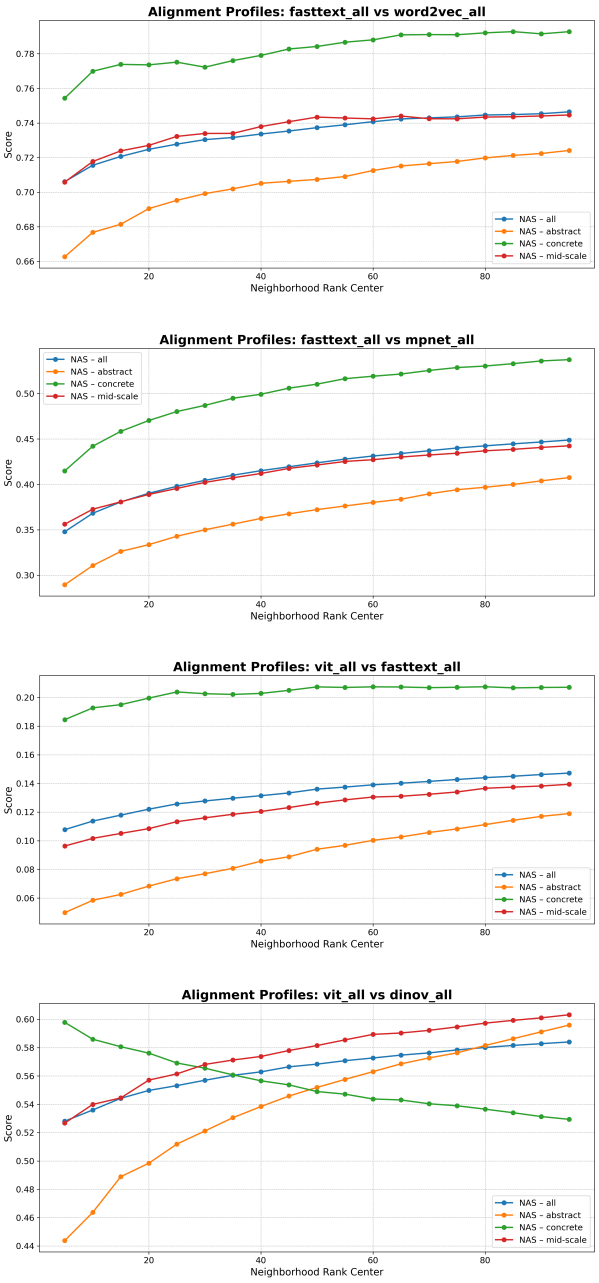

Figure 4: NAS profiles for increasing neighborhood sizes and across concreteness levels.

Limitations

While our method provides interpretable, model- and modality-agnostic comparisons of semantic spaces, several limitations should be acknowledged. Hubness in high-dimensional semantic spaces may inflate alignment scores for some central positioned concepts, which can distort the overlap based comparison for this subset of concepts. Bing Image Search introduces both variation and bias based on cultural and temporal factors (time of retrieval) in its search engine ranking. The retrieved images may therefore not be a comprehensive or completely representative sample of the possible depictions for a target concept. Our approach primarily focuses on local neighborhood structure by comparing ranked orderings. It does not account for global structural properties of these embedding spaces and also overlooks differences in distances between concepts.

Ethical Statement

We do not see any ethical issue related to this work. All our modeling experiments were conducted with open-source libraries, which received due citations.

Use of AI Assistants. The authors acknowledge the use of AI assistants solely for correcting grammatical errors, formatting table boundaries, and providing assistance with coding.

Code and Data Availability

All code and data used in this paper as well as additional plots are openly available at: https://github.com/SNaber/StarSem2025_SemanticNeighborhoods

Acknowledgements

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

References

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating Experiential and Distributional Data to Learn Semantic Representations. *Psychological Review*, 116(3):463–498.

Ruairidh Battleday, Joshua Peterson, and Thomas Griffiths. 2020. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, 11:5418.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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.

Ashley N. Danguecan and Lori Buchanan. 2016. Semantic Neighborhood Effects for Abstract versus Concrete Words. *Frontiers in Psychology*, 7(1034).

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

FacebookAIResearch. 2023. facebook/dinov2-base. https://huggingface.co/facebook/dinov2-base. Accessed: 2025-04-28.

FacebookAIResearch. 2024. Hiera: Hierarchical vision transformer models. https://huggingface.co/facebook/hiera-large-224-hf. Accessed: 2025-04-28.

Diego Frassinelli and Alessandro Lenci. 2012. Concepts in Context: Evidence from a Feature-Norming Study. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, Sapporo, Japan.

Diego Frassinelli, Daniela Naumann, Jason Utt, and Sabine Schulte m Walde. 2017. Contextual characteristics of concrete and abstract words. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Short papers*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Phillip I. Good. 2004. *Permutation, Parametric, and Bootstrap Tests of Hypotheses (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Google. 2021. google/vit-base-patch16-224-in21k. https://huggingface.co/google/vit-base-patch16-224-in21k. Accessed: 2025-04-28.

Hugging Face. 2021. sentence-transformers/all-mpnet-base-v2. https://huggingface.co/sentence-transformers/all-mpnet-base-v2. Accessed: 2025-09-29.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 835–841, Baltimore, Maryland. Association for Computational Linguistics.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90.

Microsoft Corporation. 2025. Bing image search api. https://learn.microsoft.com/en-us/bing/search-apis/bing-image-search/overview.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.

Daniela Naumann, Diego Frassinelli, and Sabine Schulte im Walde. 2018. Quantitative semantic variation in the contexts of concrete and abstract words. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 76–85, New Orleans, Louisiana. Association for Computational Linguistics.

OpenAI. 2021. openai/clip-vit-base-patch32. https://huggingface.co/openai/clip-vit-base-patch32. Accessed: 2025-04-28.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. Dinov2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

Allan Paivio. 1971. Imagery and Language. In Sydney Joelson Segal, editor, *Imagery: Current Cognitive Approaches*, pages 7–32. Academic Press, New York and London.

Diane Pecher, Inge Boot, and Saskia Van Dantzig. 2011. Abstract Concepts. Sensory-Motor Grounding, Metaphors, and Beyond. *Psychology of Learning and Motivation – Advances in Research and Theory*, 54:217–248.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014a. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014b. Glove: Global vectors for word representation. https://github.com/stanfordnlp/GloVe. Accessed: 2025-04-28.

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. In *International Conference on Machine Learning*.

Gabriel Recchia and Michael N. Jones. 2012. The Semantic Richness of Abstract Concepts. *Frontiers in Human Neuroscience*, 6(315).

Megan Reilly and Rutvik H. Desai. 2017. Effects of Semantic Neighborhood Density in Abstract and Concrete Words. *Cognition*, 169:46–53.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Eleanor Rosch. 1973. Natural categories. *Cognitive Psychology*, 4(3):328–350.

Roland Schäfer. 2015. Processing and Querying Large Web Corpora with the COW14 Architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28–34, Mannheim, Germany.

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493, Istanbul, Turkey.

Henrique* Schechter Vera, Sahil* Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 69 others. 2025. Embeddinggemma: Powerful and lightweight text representations.

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".*

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *ArXiv*, abs/2004.09297.

Leonard Talmy. 1983. How Language structures Space. In Herbert L. Pick, Jr. and Linda P. Acredolo, editors, *Spatial Orientation: Theory, Research, and Application*, pages 225–282. Plenum Press, New York and London.

Tarun Tater, Diego Frassinelli, and Sabine Schulte im Walde. 2025. AbsVis – benchmarking how humans and vision-language models "see" abstract concepts in images. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China.

Tarun Tater, Sabine Schulte Im Walde, and Diego Frassinelli. 2024. Unveiling the mystery of visual attributes of concrete and abstract concepts: Variability, nearest neighbors, and challenging categories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21581–21597, Miami, Florida, USA. Association for Computational Linguistics.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. pages 45–50.

A Appendix

A.1 Image Quality Assessment

To assess image quality, we manually rated images for 30 randomly sampled concepts (10 per category) on a Likert-scale based on how well the image visually represented the concept. Images that contained the concept word in written form were counted separately. As shown in Table 2, concrete concepts had the highest visual quality and lowest proportion of textual depictions, while abstract concepts were harder to depict and often appeared as symbolic or textual representations.

Level	Mean Rating	Std Rating	Text
abstract	1.99	0.98	0.39
mid-range	3.59	0.94	0.11
concrete	4.67	0.21	0.01

Table 2: Mean image quality ratings and proportion of images with textual depictions, by concreteness level.

A.2 Effect of Aggregation by Concept Set

In Figure 5 we can see that the effect of aggregation is more pronounced for more abstract concepts, which often do not have direct visual referents and are therefore more variable in their visual representation. In the text-vision comparison the effect is also visible if one accounts for the stronger effect of concreteness levels on final alignment.

A.3 Qualitative Analysis of Nearest Neighbors

We inspected the top-5 nearest neighbors for four concepts of varying concreteness: *eye* (concrete 4.9), *goal* (mid-scale 3.06), *probability* (abstract, symbolically depictable, 1.65) and *ethos* (abstract, symbolically non-depictable, 1.58). Both vision and text models produce plausible neighbors for all concepts, with one exception: lacking clear visual referents, *ethos* poses a challenge for the image models – only CLIP finds meaningful neighbors, as seen in Table 3.

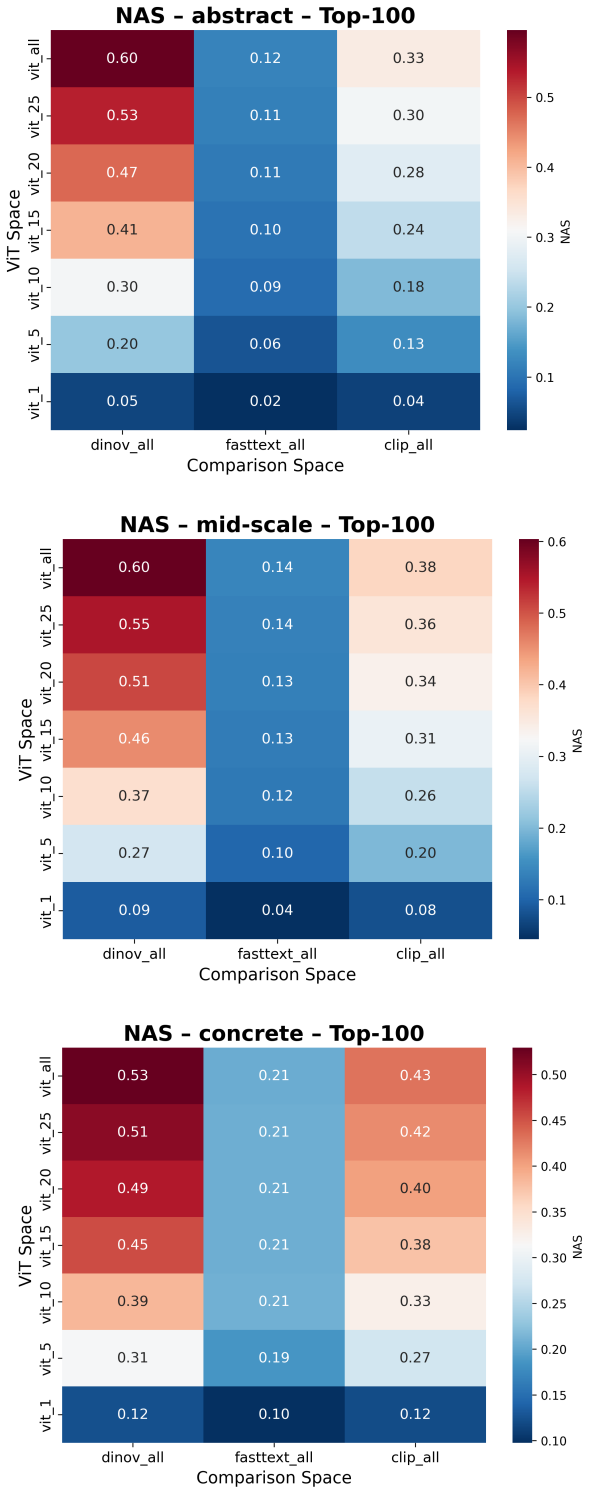

Figure 5: Neighborhood overlap (top-100) across representations for abstract, mid-scale and concrete concepts (top to bottom) using increasingly more images (1–35).

Concept: *eye*

ViT	Hiera	CLIP	Count	GloVe	FastText
pupil	cataract	pupil	nose	hair	eyelid
eyelid	pupil	macro	moment	nose	nose
macro	eyelid	eyelid	spectacle	body	eyesight
cataract	macro	blindness	mouth	ear	ear
eyesight	cavity	cataract	amazement	skin	forehead

Concept: *goal*

ViT	Hiera	CLIP	Count	GloVe	FastText
competition	spontaneity	effort	ambition	effort	effort
greatness	effort	life	team	chance	scorer
optimism	choice	achievement	effort	success	goalkeeper
determination	inaction	teammate	motivation	advantage	striker
confidence	enthusiasm	victory	ability	momentum	ball

Concept: *probability*

ViT	Hiera	CLIP	Count	GloVe	FastText
interpolation	interpolation	interpolation	likelihood	correlation	likelihood
denominator	prevalence	subset	sensitivity	likelihood	variance
differentiation	subroutine	permutation	estimation	variance	estimation
combination	vertex	combination	occurrence	prediction	prediction
fraction	hypothesis	approximation	propensity	estimation	approximation

Concept: *ethos*

ViT	Hiera	CLIP	Count	GloVe	FastText
competence	iteration	pathos	academy	ethic	ethic
outcome	proposition	personality	school	mindset	attitude
competency	validity	ethic	community	commitment	commitment
analysis	percentage	trait	pupil	professionalism	tradition
epidemiology	tradeoff	empathy	aspiration	culture	individuality

Table 3: Five nearest neighbors per concept in vision (ViT, Hiera, CLIP) and text (Count, GloVe, FastText) models.