Unveiling the mystery of visual attributes of concrete and abstract concepts: Variability, nearest neighbors, and challenging categories

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

Abstract

The visual representation of a concept varies significantly depending on its meaning and the context where it occurs; this poses multiple challenges both for vision and multimodal models. Our study focuses on concreteness, a well-researched lexical-semantic variable, using it as a case study to examine the variability in visual representations. We rely on images associated with approximately 1,000 abstract and concrete concepts extracted from two different datasets: Bing and YFCC. Our goals are: (i) evaluate whether visual diversity in the depiction of concepts can reliably distinguish between concrete and abstract concepts; (ii) analyze the variability of visual features across multiple images of the same concept through a nearest neighbor analysis; and (iii) identify challenging factors contributing to this variability by categorizing and annotating images. Our findings indicate that for classifying images of abstract versus concrete concepts, a combination of basic visual features such as color and texture is more effective than features extracted by more complex models like Vision Transformer (ViT). However, ViTs show better performances in the nearest neighbor analysis, emphasizing the need for a careful selection of visual features when analyzing conceptual variables through modalities other than text.

1 Introduction

Language and vision play a crucial role for the understanding of the world surrounding us. Among the five senses, vision is considered the primary source of perceptual information for our mental representations when experiencing the real world (Brysbaert et al., 2014; Lynott et al., 2020). Based on these premises, computational studies have leveraged the strong interaction between visual and textual information to uncover the latent relationships between these two modalities and to build richer and more precise representations. In most cases, the contribution of these two very different modalities is asymmetric, with the textual modality having a stronger influence on model performance; for example, when investigating the concreteness of a concept, its compositionality, or its semantic representation (Bhaskar et al., 2017; Köper and Schulte im Walde, 2017; Hewitt et al., 2018). The exact reasons behind such asymmetry are still unclear, and especially the role of the visual elements has been explored significantly less. Therefore, this paper focuses explicitly on the nature and contribution of the visual component. To this end, we analyze the different characteristics of concrete and abstract concepts to determine whether and how visual information can help distinguish between them. Our analysis is particularly important when addressing the complex task of modeling abstract concepts, which often lack a distinctive visual component, unlike their concrete counterpart. For example, concrete concepts like `banana` and `chariot` evoke vivid mental images anchored to objects that are easy to visualize. In contrast, abstract concepts like `accountability` and `allegiance` are more challenging and subjective to visualize (Paivio et al., 1968; Kastner et al., 2020).

Various studies have successfully attempted to predict the concreteness score of a concept by exploiting the visual information extracted from multiple images associated with it in combination with more traditional textual representations (Kiela et al., 2014; Hessel et al., 2018; Charbonnier and Wartena, 2019). A building assumption of these visual models is a certain degree of visual coherence that facilitates the construction of stable visual representations. While images of concrete concepts are generally expected to show greater consistency, a notable variability is still present in *both* concrete and abstract concepts, i.e., the properties of these images, including color, shape, size, and other visual details, may vary significantly, thus reflecting the diversity of the intrinsic nature of the concept.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21581–21597
November 12-16, 2024 ©2024 Association for Computational Linguistics

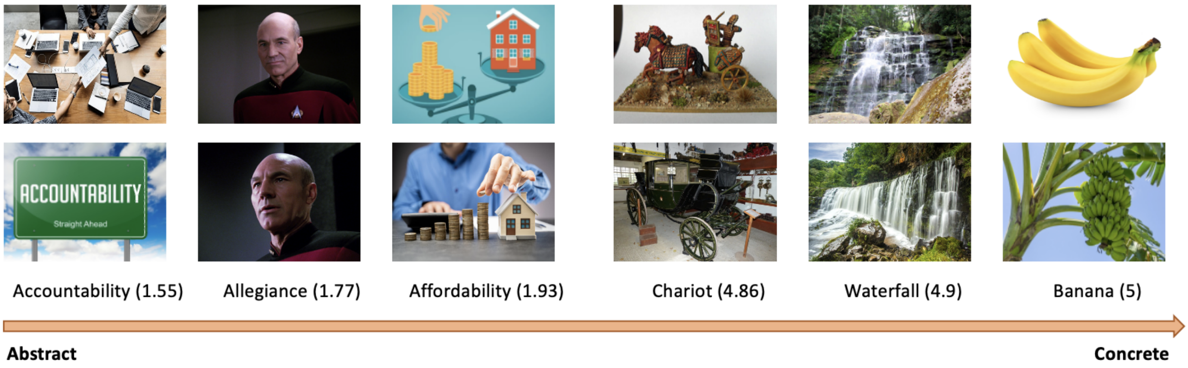


Figure 1: Images of concrete and abstract concepts with varying concreteness ratings on a scale from 1 (clearly abstract) to 5 (clearly concrete), and two plausible visual representations each. The examples are extracted from the *Bing* dataset described in Section 3.2.

Figure 1 illustrates the variability in the images associated with abstract and concrete concepts. Images can be highly representative of a concept and be visually similar (e.g., `affordability`, `waterfall`) or be rather different from one another (e.g., `chariot`). Conversely, images of a concept can be very similar but not informative representations of the concept (e.g., `allegiance`). Finally, they can be highly different yet individually all strongly associated with the same target concept (e.g., `banana, accountability`). These degrees of variation highlight some of the inherent challenges computational methods face in constructing a comprehensive visual representation of a concept and mapping it to its labels. These challenges are orthogonal to previously raised issues regarding depictions of (mostly concrete) semantic concepts such as variability of prototypicality (Gualdoni et al., 2023; Harrison et al., 2023; Tagliaferri et al., 2023), and will be explored further in the course of this study (see RQ3 below).

Our research targets the challenges of precisely quantifying the contribution of visual information in describing concrete versus abstract concepts, using interpretable representations to explore the following three research questions:

RQ1: Can visual diversity differentiate between concrete and abstract concepts?
RQ2: How consistent are visual attributes across multiple images of the same concept?
RQ3: What are inherent yet plausible failure categories for unimodal visual representations?

In Study 1, we address **RQ1** by classifying approximately 500 concrete and 500 abstract concepts based on the diversity in visual features extracted from images associated to each concept. This approach helps us identify the most salient visual features that distinguish between concepts based on their concreteness. In Study 2, we address **RQ2** and analyse the consistency of these features across multiple concept images, by performing a nearest-neighbor analysis of image representations. Finally, Study 3 targets **RQ3** by qualitatively analyzing the failures in Study 2 and manually determining categories of problematic issues.

To our knowledge, this is the first large-scale study conducting a detailed quantitative and qualitative investigation into how visual features contribute to representing abstract and concrete concepts. By focusing exclusively on the visual component, we can systematically identify the strengths and weaknesses of using such extremely rich source of information. Additionally, compared to previous studies, our methodology highlights cases that are particularly challenging because they are equally plausible rather than erroneous.

2 Related Work

The distinction between abstract and concrete words is highly relevant for natural language processing and has been exploited for metaphor detection (Turney et al., 2011; Tsvetkov et al., 2013; Köper and Schulte im Walde, 2016; Maudslay et al., 2020; Su et al., 2021; Piccirilli and Schulte im Walde, 2022), lexicography (Kwong, 2011), and embodied agents and robots (Cangelosi and Stramandinoli, 2018; Rasheed et al., 2018; Ichter et al., 2023), among others. Most studies addressing this distinction have primarily focused on the textual modality alone (Frassinelli et al., 2017; Ljubešić et al., 2018; Naumann et al., 2018; Charbonnier and Wartena, 2019; Frassinelli and Schulte im Walde, 2019; Schulte im Walde and Frassinelli, 2022; Tater et al., 2022). First extensions to further modalities explored free associations and imageability (Hill et al., 2014; Kiela et al., 2014; Köper and Schulte im Walde, 2016). Since the primary distinction between degrees of abstractness is influenced by the strength of sense perception, with vision being considered the main source of perceptual information (Brysbaert et al., 2014; Lynott et al., 2020), later studies began to explore bimodal approaches that combine text and images (Bhaskar et al., 2017; Hessel et al., 2018). Compared to text, the visual component has provided less definitive insights, and it is unclear whether this is due to architectural choices or to the inherent challenge triggered by depicting abstract concepts. Cerini et al. (2022) analyzed the mechanism behind this indirect grounding of abstract concepts by collecting word association data and pairs of images and abstract words. Kastner et al. (2019) discuss the visual variety of a dataset using mean shift clustering, where the dataset is designed to contain images in the same ratio of sub-concepts as in real life. Kastner et al. (2020) performed a regression study to predict the imageability of concepts using the YFCC100M dataset; our feature selection builds on their results. Kiela et al. (2014) and Hessel et al. (2018) postulated that concreteness in images varies across datasets and is not directly connected to the underlying linguistic concept. Pezzelle et al. (2021) evaluated the alignment of semantic representations learned by multimodal transformers with human semantic intuitions, finding that multimodal representations have advantages with concrete word pairs but not with abstract ones. Vaze et al. (2023) argue that there are multiple notions of "image similarity" and that models should adapt dynamically. For example, models trained on ImageNet tend to prioritize object categories, while a user might want the model to focus on colors, textures, or specific elements in the scene. They introduce the GeneCIS benchmark, assessing models' adaptability to various similarity conditions in a zero-shot evaluation setting. They observe that even robust CLIP models struggle to perform well, and performance is only loosely connected to ImageNet accuracy. Most recently, Tater et al. (2024) examined to which degree SigLIP, a state-of-the-art Vision-Language model (VLM), predicts labels for images of abstract and concrete concepts that are semantically related to the original labels in various ways: synonyms, hypernyms, co-hyponyms, and associated words. The results show that not only abstract but also concrete concepts exhibited significant variability in semantically appropriate label variants.

3 Experimental Design

In the following sections, we present the resources used in our analyses. We introduce the target concepts under investigation, their abstractness scores, and the associated images. Subsequently, we describe the algorithms employed to extract the visual attributes from the images.

3.1 Target Concepts & Concreteness Norms

To select a balanced amount of concrete and abstract targets, we use the concreteness ratings from Brysbaert et al. (2014) (henceforth, *Brysbaert norms*) that were collected via crowd-sourcing, and range from 1 (clearly abstract) to 5 (clearly concrete). Our analyses focus on 500 highly abstract (concreteness range: $1.07 - 1.96$) and 500 highly concrete ($4.85 - 5.00$) nouns. We excluded nouns with mid-range concreteness scores as they are typically more challenging for humans and thus lead to noisier distributional representations (Reilly and Desai, 2017; Pollock, 2018; Knupleš et al., 2023).

3.2 Image Datasets

We extracted images for each target noun – both concrete and abstract – from two distinct datasets: (i) the YFCC100M Multimedia Commons Dataset (*YFCC*; Thomee et al. (2016)); and (ii) *Bing*[1].

For the YFCC dataset, we randomly selected 500 images tagged with each target concept. The YFCC dataset is the largest publicly available user-tagged dataset containing 100 million media objects extracted from the online platform Flickr. Its images exhibit diversity in quality, content, visual coherence, and annotation consistency. Thus, we use them to test the robustness of the methods adopted and support the ecological validity of our studies despite introducing a significant level of noise from variable image quality and annotation inaccuracies.

For the Bing dataset, the images were selected by directly querying the target word. To avoid duplicates, we automatically excluded images where all the pixel values were exactly the same as another image and downloaded new ones if necessary (continuing recursively). Subsequently, we manually inspected the remaining images for inappropriate content (e.g., sexual content) and removed them.

[1] https://www.bing.com/images/

We kept a maximum of 25 images for each target concept as this was the highest number consistently available across all target concepts. Given that Bing was our control condition, maintaining a balanced dataset was important. Finally, for both YFCC and Bing, we only included images with a size of 256×256 pixels or higher and resized them to a uniform size as required for each feature analysis.

Despite the huge size of the YFCC dataset, we were unable to extract the desired number of 500 images across all our $1,000$ targets (500 concrete and 500 abstract). Table 1 shows for how many concrete and abstract target nouns we were able to retrieve $25 \ldots 500$ images. For example, we could only retrieve 500 images for subsets of 463 concrete and 151 abstract nouns. For the following analyses, it is therefore important to remember that abstract targets are more affected than concrete targets regarding the available numbers of images.

# Images	**25**	**100**	**200**	**300**	**400**	**500**
Concrete	498	494	481	475	472	463
Abstract	420	304	237	197	172	151

Table 1: Number of abstract and concrete target nouns for different number of images per target (YFCC).

3.3 Extraction of Visual Attributes

When evaluating an image, it is crucial to consider the visual properties that help us capture its most prominent characteristics. We extracted a series of independent visual features (attributes) for each image associated with our target words. Furthermore, we utilized two SOTA visual models to generate comprehensive image representations and use them as benchmarks for our analyses.

We start with low-level features, including colors, shapes, and textures. Colors are described as distributions in the *HSV* space: hue, saturation, value (Joblove and Greenberg, 1978). Shapes and structures in an image are quantified using the *Histogram of Oriented Gradients* (*HOG*; Dalal and Triggs (2005)): this feature descriptor captures the occurrences of gradient orientation in localized image segments. We capture texture information using two methods: the *Gray-Level Co-occurrence Matrix* (*GLCM*; Haralick et al. (1973)) and the *Local Binary Patterns Histograms* (*LBPH*; Ojala et al. (2002)). GLCM is a statistical measure that considers the spatial relationship of pixels represented as a co-occurrence matrix. This approach quantifies how often pairs of pixel values appear together at a specified spatial orientation. LBPH, on the other hand, calculates a local representation of texture by comparing each pixel with its neighbors.

We also include more complex features representing objects and their relationships in a scene. Low-dimensional abstract representations of a scene are computed using *GIST* (Oliva and Torralba, 2001). To identify similar sub-regions and patches across images, we use the Speeded-Up Robust-Features feature descriptor (*SURF*; Bay et al. (2008)) combined with a Bag-of-Words model (*BOW*; Csurka et al. (2004)) using k-means clustering. The objects occurring in an image are detected using the YOLO9000 model (*YOLO*; Redmon and Farhadi (2017)) pre-trained on $9,418$ object classes. We then extract hypernymy relationships from WordNet (Miller, 1995) to reduce the number of object types detected from the original $9,418$ to $1,401$ classes of hypernyms. With this approach, we substantially alleviate sparsity while retaining most of the information captured by the model since the hypernyms contain information specific enough to qualify the objects in an image. We then determine the location of the objects detected in the image and quantify their spacial relationship by using an overlapping 10×10 grid and counting the number of objects co-occurring in each cell. On average, only 10% of the images associated with each target noun contain an object detected by the YOLO model, even though 330 of our 500 concrete concepts are also in the $9,000$ object classes in YOLO (for more details, see Table 3 in the Appendix).

Finally, we generate comprehensive visual representations with two pre-trained models for feature extraction: *SimClr* (Chen et al., 2020) and *Vision Transformer* (*ViT*; Dosovitskiy et al. (2021)). We use these models as a benchmark against basic features since they are more advanced models and are the backbone of most currently used multi-modal models (e.g., CLIP uses a ViT encoder). SimClr builds image representations using contrastive learning trained on images only. It maximizes the agreement between differently augmented views of the same image using a contrastive loss. ViT is a supervised model for image classification trained by splitting an image into patches, which are then combined and converted into linear embeddings using a transformer network. ViT uses attention maps to deduce an image's most informative parts. It is pre-trained on the ILSVRC-2012 ImageNet dataset with $1,000$ classes. Only 36 of our target concepts completely overlap with these $1,000$ classes, indicating that our results are generalizable and not the consequence of the overlap between the classes from ImageNet and our target concepts.

3.3.1 Feature Combination

As traditionally done in the literature (e.g., Kiela et al. (2014); Bhaskar et al. (2017)), we create one single visual representation for each concept combining the information from the different images. To achieve this, we compare the feature vectors of all images of the same concept. This results in nine square similarity matrices (one per visual attribute) of size $N \times N$ (the number of images), which are symmetrical. These matrices capture the characteristics of a concept and, at the same time, highlight the variability across its different visual representations. Given that the similarity matrix's values depend on the order of the images, we calculate the N eigenvalues of each similarity matrix to provide an invariant representation that is order-independent. This also helps us reduce the dimensionality of features and make them consistent, while still encoding the core characteristics of each feature.

4 Study 1: Classifying Concepts using Visual Information

This first study aims to identify the visual features that are most useful for discriminating between images of concrete vs. abstract nouns. We utilize three different classifiers: Support Vector Machine (SVM) with *rbf* kernel, Random Forest (RF), and Logistic Regression (LR) with hyper-parameter tuning, while using the eigenvalues of the combined visual features described above as predictors.[2] In the main text, we report the performance of the RF model as, overall, it outperforms the other two classifiers (the results for LR and SVM are reported in Figures 5 and 6 in the Appendix). We evaluate the predictive power of our features independently and by concatenating them. To account for data skewness between classes, we apply 5-fold cross-validation.

4.1 Results

Figure 2 reports the F1-scores obtained by the RF classifier. We compare the performance of low-level visual features used individually and in combination (Combined Basic), as well as advanced features derived from ViT and SimClr, along with their combinations. The different columns reflect the number of images available for each target. Notably, the model trained on a mix of only basic features consistently obtains the highest F1-scores (darker color) across all datasets and image counts. Incorporating more sophisticated visual features, such as SimClr or ViT, offers limited advantages and only when merged with the basic feature set. When comparing the performance for Bing and YFCC, images extracted from Bing consistently outperform those from YFCC across all feature types and number of images. Furthermore, a trend emerges with YFCC images: increasing the number of images from 25 to 500 leads to a steady improvement in performance.

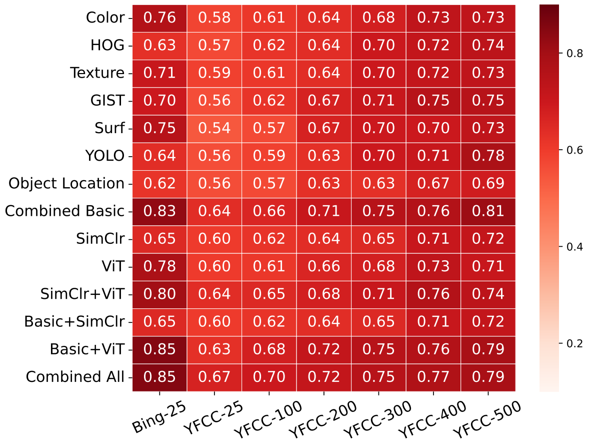


Figure 2: Weighted F1-scores for different features and different dataset sizes for Bing and YFCC using RF.

In Figure 3 we report the same results but separately for abstract vs. concrete concepts. It is striking to see that, on average, the RF model classifies more effectively concrete than abstract concepts, simply based on their visual diversity. We also see that while adding visual information is beneficial for classifying concrete nouns, it is detrimental for abstract nouns. This is strongly influenced by the marked reduction in the abstract target nouns when increasing the number of images (see Table 1).

4.2 Discussion

This study tested how reliable are visual attributes in capturing the diversity of images to distinguish between concrete vs. abstract concepts. Overall, low-level features like color and patch similarity (SURF) play a more vital role in predicting ab-

[2] We also conduct a regression analysis and present the results in the Appendix.

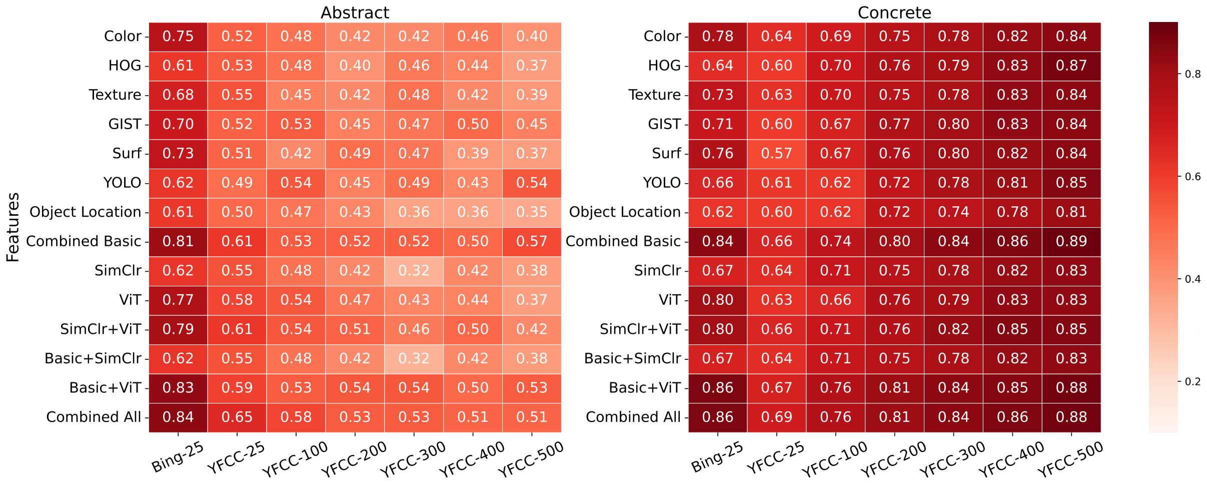


Figure 3: Class-wise F1-scores of abstract and concrete concepts for RF, across features and dataset sizes.

stractness than more complex feature types like object location and detection. This suggests that while high-level object information may vary considerably, low-level features remain more consistent across different depictions of the same concept, which is crucial in classifying concepts based on their abstractness. This observation extends to more sophisticated feature representations such as ViT and SimClr as well.

The results in Figure 3 show that for coherent and less noisy images of concepts in the Bing dataset, the model shows comparable performance for both concrete and abstract nouns, mirroring the general patterns discussed above. However, when increasing the number of images for the YFCC dataset, the performance of the model progressively increases for concrete nouns with the addition of more images while the performance for abstract nouns decreases. Particularly when evaluating the performance of the model with 500 images per concept, it becomes evident that basic features are all very good predictors (all above 0.84) of concreteness. Notably, also more complex features, such as YOLO and object location, show a steady improvement and achieve a level of performance that closely aligns with that of the simpler, low-level features regardless of the low number of objects detected. Once again, the use of more sophisticated representations does not show any substantial improvement in the performance of the models. When examining abstract nouns, the drastic reduction in the number of target nouns with the addition of more images inevitably impacts the performance of the model in a negative way. This reduction renders any subsequent analysis of this particular subset less informative.

5 Study 2: Inspecting Visual Nearest Neighbors

In our second study, we directly build on the evidence from Study 1 and perform a nearest neighbors analysis to inspect the consistency of visual attributes across multiple images of the same concept. We compute the cosine similarity of each image of a concept with all other images of all concepts in the same dataset, represented by using the same features as before. We then inspect the top N (where $N = [25, \dots, 500]$) most similar images and compute the percentage of neighbors associated with the same concept; e.g., how many nearest neighbor images of an image of banana are also images of banana.

5.1 Results

Table 2 presents the average percentage of visual neighbors associated with the same concept across different features for the Bing, YFCC-25 and YFCC-500 datasets (see Table 5 in the Appendix for YFCC-100, 200, 300 and 400). Overall, the results across features and datasets are very low. On average, less than 1% of the images closest to a specific target are associated with it, both for concrete and abstract targets, but interestingly exhibiting divergent patterns. With Bing, even though not as strongly as we initially expected, the nearest neighbors of concrete concepts show a higher similarity than those of abstract concepts. Among simple features, object detection (YOLO) marginally

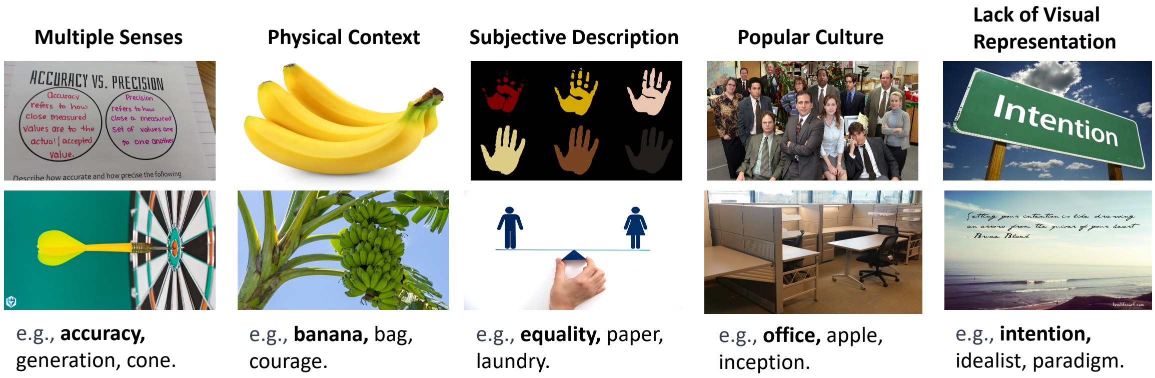

Multiple Senses
Physical Context
Subjective Description
Popular Culture
Lack of Visual Representation
e.g., **accuracy,** generation, cone.
e.g., **banana,** bag, courage.
e.g., **equality,** paper, laundry.
e.g., **office,** apple, inception.
e.g., **intention,** idealist, paradigm.
Intention

Figure 4: Five most frequent reasons (top row) of visual diversity among images associated with the same concept (indicated by the bold font in the example list below each image).

outperforms the rest. However, for the YFCC-25 dataset, all basic features except object location produce better results for abstract concepts. When we include more images (YFCC-500), the percentage of correct neighbors drop even more. Unlike the results of the classification study discussed in Section 4, employing more sophisticated representations, such as Vision Transformer, yields the best outcomes, although the performance levels remain low. Moreover, abstract concepts in the YFCC-25 dataset perform similarly to, or even better than, their counterparts in the Bing dataset, despite still showing overall poor performance.

Attribute	Bing-25		YFCC-25		YFCC-500	
	A	C	A	C	A	C
Color	0.68	0.96	1.70	0.95	0.81	0.65
HOG	0.48	1.44	0.68	0.58	0.36	0.44
Texture	0.29	0.33	0.35	0.26	0.28	0.27
GIST	0.55	1.88	1.03	0.76	0.52	0.56
SURF	0.64	1.70	0.93	0.62	0.40	0.38
YOLO	2.25	3.19	1.09	1.03	1.64	1.57
Object Loc.	0.18	0.39	0.15	0.18	0.24	0.27
Combined	0.64	2.14	1.40	0.99	0.69	0.75
Simclr	0.65	1.49	1.15	0.79	0.53	0.55
ViT	2.83	26.44	3.71	6.67	2.27	6.63

Table 2: Average percentage of nearest neighbors (out of top 25 or 500, respectively) associated with the same abstract (A) or concrete (C) concept.

5.2 Discussion

This study demonstrated that images with similar labels share very little visual information. While Hessel et al. (2018) and Hewitt et al. (2018) have already discussed the lack of a univocal visual representation for abstract concepts, our results reveal a more nuanced pattern. Surprisingly, we found significant visual variability even among concrete concepts, which challenges the assumption that images of the same target share consistent visual features. More complex models (like ViT) can capture the higher agreement between concrete concepts, indicating that images of concrete concepts are generally more consistent or similar. However, basic features may encode more distinctive information related to individual abstract concepts than concrete concepts. Moreover, combined basic features, which performed better than ViT in Study 1, do not encode enough information for nearest neighbors compared to ViT.

6 Study 3: Exploring Factors Behind Visual Diversity

As discussed in Section 5 when analyzing nearest neighbors, the biggest challenge in using images of a concept comes from the diversity of the images associated with it. The same concept, whether abstract or concrete, can be depicted in many different yet plausible ways, thus relating to previously discussed issues regarding the variability of prototypical attributes in depictions of the semantic concepts (Gualdoni et al., 2023; Harrison et al., 2023; Tagliaferri et al., 2023). In our final analysis, we provide a manual classification of the critical factors influencing the nearest neighbors of our target concepts.

We identified five primary reasons for visual diversity, as exemplified in Figure 4. For concepts like `accuracy`, `generation`, and `cone`, the words used as a proxy to our concepts may be lexically ambiguous and have *multiple senses*. According to Wordnet (Miller, 1995), 650 out of the 918 concepts used in our studies have more than one sense, and 248 concepts have four or more senses. A further source of variability is *physical context*, manifesting itself as different background information, objects, etc. In our example, both images depict bananas, but they differ visually: in the bottom image, the bananas are still hanging on a banana tree, which dominates the scene. Another form of visual diversity is triggered by *subjective representations*: concepts like equality and paper show very high variability. People have different visual interpretations and realizations of these concepts, even when the underlying conceptual meaning is understood in the same way. *Popular culture* often associated with films and books represents a kind of variability that introduces visual representations often completely disjoint from the original meaning of the concept: for example, images of the concepts inception and office contain images which are from the movie “Inception” and the TV show “The Office”, respectively. Finally, primarily abstract concepts like intention and idealist *lack distinctive visual representations* and are hard to depict. These concepts are often represented by writing the associated word into an image.

Another orthogonal source of variability in visual representations comes from the selection process in the source dataset. For example, the YFCC dataset contains images from Flickr that are uploaded by users, resulting in a lot of variability and bias toward specific senses of a tagged concept.

6.1 Experimental Design and Results

Given that we are the first suggesting this categorization of "challenges" related to very diverse but still plausible images associated with specific concepts, we ask 13 participants to evaluate our five categories using a subset of target images related to abstract and concrete concepts. We selected two images each for a subset of 18 concepts, while ensuring that we included potentially “problematic" cases. The experiment was conducted on Google Forms, where the participants could choose at least one reason (and possibly more) why two images of the same concept differed.

Figure 7 in the Appendix presents the 18 concepts, the image pairs, and the results of the annotation. For most of the target images, we see high agreement between annotators on a specific “reason for visual diversity”, with Krippendorff’s $\alpha = 0.29$ (Krippendorff, 1980; Artstein and Poesio, 2008). For example, 12 out of 13 ratings assigned the visual ambiguity for the concept banana to variability in the *physical context*. And 10 out of 16 ratings for intention are linked to a *lack of visual representation*. To further inspect the variability and complexity of plausible, but yet diverse, visual representations across images of these 18 images, we set up an Amazon Mechanical Turk[3] study where nine native English speakers (from the UK and USA) had to describe in one word "what is depicted in an image". As an example of the plausible variability in the response, when evaluating the response for the image of equality showing six colorful hands (see Figure 4), 27 out of 39 participants listed words referring to the colors in the image. Even though colors provide relevant attributes of the image, they do not represent generally salient meaning components of the associated concept. See Tables 7 and 8 in the Appendix for the complete lists of words generated for the 10 images associated with concrete and abstract concepts.[4]

7 Conclusion

We performed three empirical studies to understand how abstract and concrete concepts are depicted in images. Compared to existing studies, we focused exclusively on the role of variability in the visual information. After automatically generating nine different feature representations for the images, we tested their reliability in a classification study to distinguish between concrete and abstract concepts. We showed that, overall, combining low-level features produces good results. We then investigated the consistency of the visual attributes across multiple images of the same concept by looking at the nearest neighbors of each image in the two datasets. The results across feature types, datasets, and concreteness scores were very low; overall, abstract concepts showed considerably higher cases where none of the most similar images were associated with the same concept. The results also showed that both concrete and abstract concepts lack a univocal visual representation in terms of objects depicted and, in general, basic visual properties. Finally, in an error analysis study with human participants, we highlighted the five most frequent reasons explaining visual diversity among images associated with the same concept.

[3] https://www.mturk.com

[4] The complete dataset of human-generated words (manually checked for offensive content) can be found here: https://github.com/TarunTater/AbstractConceptsInImages/tree/main/depict_image_annotations

Overall, our research significantly advances the understanding of the role of the visual component in tasks that heavily rely on the integration of multiple types of information beyond just text.

Limitations

The number, random selection, and content of the images used in this study may introduce some variability in the results. Moreover, any interpretation based on the output of the object detection systems should be made with caution, especially considering the very low number of images where an object was detected.

Ethics Statement

We see no ethical issues related to this work. All experiments involving human participants were voluntary, with fair compensation (12 Euros per hour), and participants were fully informed about data usage. We did not collect any information that can link the participants to the data. All modeling experiments were conducted using open-source libraries, which received proper citations.

Acknowledgements

This research is supported by the DFG Research Grant SCHU 2580/4-1 *Multimodal Dimensions and Computational Applications of Abstractness*. We thank Allison Keith, Amelie Wührl, Ana Baric, Christopher Jenkins, Filip Miletic, Hongyu Chen, Iman Jundi, Lucas Moeller, Maximilian Martin Maurer, Mohammed Abdul Khaliq, Neele Falk, Prisca Piccirilli, Simon Tannert, Tanise Ceron, and Yarik Menchaca Resendiz for their help in the evaluation tasks.

References

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. Speeded-up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359.

Sai Abishek Bhaskar, Maximilian Köper, Sabine Schulte im Walde, and Diego Frassinelli. 2017. Exploring Multi-Modal Text+Image Models to Distinguish between Abstract and Concrete Nouns. In *Proceedings of the IWCS Workshop on Foundations of Situated and Multimodal Communication*.

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.

Ludovica Cerini, Eliana Di Palma, and Alessandro Lenci. 2022. From Speed to Car and Back. An Exploratory Study About Associations Between Abstract Nouns and Images. In *Proceedings of the 1st (Dis)embodiment Workshop*, pages 80–88.

Jean Charbonnier and Christian Wartena. 2019. Predicting Word Concreteness and Imagery. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 176–187.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607.

Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. 2004. Visual Categorization with Bags of Keypoints. In *Proceedings of the Workshop on Statistical Learning in Computer Vision*, volume 1, pages 1–2.

Navneet Dalal and Bill Triggs. 2005. Histograms of Oriented Gradients for Human Detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893.

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. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*.

Diego Frassinelli, Daniela Naumann, Jason Utt, and Sabine Schulte im Walde. 2017. Contextual Characteristics of Concrete and Abstract Words. In *Proceedings of the 12th International Conference on Computational Semantics*.

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.

Eleonora Gualdoni, Thomas Brochhagen, Andreas Mädebach, and Gemma Boleda. 2023. What's in a Name? A Large-Scale Computational Study on How Competition Between Names Affects Naming Variation. *Memory and Language*, 133.

Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. 1973. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, (6):610–621.

Sophia Harrison, Eleonora Gualdoni, and Gemma Boleda. 2023. Run Like a Girl! Sport-Related Gender Bias in Language and Vision. In *Findings of the Association for Computational Linguistics*, pages 14093–14103.

Jack Hessel, David Mimno, and Lillian Lee. 2018. Quantifying the Visual Concreteness of Words and Topics in Multimodal Datasets. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2194–2205.

John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. 2018. Learning Translations via Images with a Massively Multilingual Image Dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2566–2576.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Multi-Modal Models for Concrete and Abstract Concept Meaning. *Transactions of the Association for Computational Linguistics*, 2(1):285–296.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. 2023. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 287–318.

George H Joblove and Donald Greenberg. 1978. Color Spaces for Computer Graphics. In *Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques*, pages 20–25.

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.

Marc A Kastner, Ichiro Ide, Frank Nack, Yasutomo Kawanishi, Takatsugu Hirayama, Daisuke Deguchi, and Hiroshi Murase. 2020. Estimating the Imageability of Words by Mining Visual Characteristics from Crawled Image Data. *Multimedia Tools and Applications*, 79:18167–18199.

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*, pages 835–841.

Urban Knupleš, Diego Frassinelli, and Sabine Schulte im Walde. 2023. Investigating the Nature of Disagreements on Mid-Scale Ratings: A Case Study on the Abstractness-Concreteness Continuum. In *Proceedings of the SiGNLL Conference on Computational Natural Language Learning*, pages 70–86.

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.

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.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*.

Oi Yee Kwong. 2011. Measuring Concept Concreteness from the Lexicographic Perspective. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 60–69.

Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. 2018. Predicting Concreteness and Imageability of Words Within and Across Languages via Word Embeddings. In *Proceedings of The 3rd Workshop on Representation Learning for NLP*, pages 217–222.

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.

Rowan Hall Maudslay, Tiago Pimentel, Ryan Cotterell, and Simone Teufel. 2020. Metaphor Detection Using Context and Concreteness. In *Proceedings of the 2nd Workshop on Figurative Language Processing*, pages 221–226.

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

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 7th Joint Conference on Lexical and Computational Semantics*, pages 76–85.

Timo Ojala, Matti Pietikainen, and Topi Maenpaa. 2002. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 24(7):971–987.

Aude Oliva and Antonio Torralba. 2001. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42:145–175.

Allan Paivio, John C Yuille, and Stephen A Madigan. 1968. Concreteness, Imagery, and Meaningfulness Values for 925 Nouns. *Journal of Experimental Psychology*, 76(1p2):1.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. 2021. Word Representation Learning in Multimodal Pre-trained Transformers: An Intrinsic Evaluation. *Transactions of the Association for Computational Linguistics*, 9:1563–1579.

Prisca Piccirilli and Sabine Schulte im Walde. 2022. Features of Perceived Metaphoricity on the Discourse Level: Abstractness and Emotionality. In *Proceedings of the 13th International Conference on Language Resources and Evaluation*.

Lewis Pollock. 2018. Statistical and Methodological Problems with Concreteness and Other Semantic Variables: A List Memory Experiment Case Study. *Behavior Research Methods*, 50:1198–1216.

Nadia Rasheed, Shamsudin H.M. Amin, Umbrin Sultana, Abdul Rauf Bhatti, and Mamoona N. Asghar. 2018. Extension of Grounding Mechanism for Abstract Words: Computational Methods Insights. *Artificial Intelligence Review*, 50(3):467–494.

Joseph Redmon and Ali Farhadi. 2017. YOLO9000: Better, Faster, Stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7263–7271.

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

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 Pado (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.

Claudia Tagliaferri, Sofia Axioti, Albert Gatt, and Dennis Paperno. 2023. The Scenario Refiner: Grounding subjects in images at the morphological level. In *Proceedings of LIMO@KONVENS: Linguistic Insights from and for Multimodal Language Processing*.

Tarun Tater, Diego Frassinelli, and Sabine Schulte im Walde. 2022. Concreteness vs. Abstractness: A Selectional Preference Perspective. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 92–98.

Tarun Tater, Sabine Schulte Im Walde, and Diego Frassinelli. 2024. Evaluating Semantic Relations in Predicting Textual Labels for Images of Abstract and Concrete Concepts. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 214–220.

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 1st Workshop on Metaphor in NLP*, pages 45–51.

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.

Sagar Vaze, Nicolas Carion, and Ishan Misra. 2023. GeneCIS: A Benchmark for General Conditional Image Similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6862–6872.

8 Appendix

8.1 Feature Availability - YOLO

As mentioned in Section 3.3, there are instances where the YOLO9000 model does not detect any objects in an image. In Table 3 we examine the percentage of images per concept where at least one object was detected. On average, around 10% of the images associated to an abstract concept have at least one object detected. For concrete concepts, this value is slightly lower, ranging between 8.5% to 9.5%. We hypothesize that the surprisingly low number of images where objects are detected (only 10% of the images) is very likely due to the following reasons. Firstly, the YFCC dataset exhibits high visual variability in terms of informativeness and quality of the user-tags used on Flickr. For example, a tag like 'dessert' might be attributed to vastly different types of images, ranging from cakes to fruit platters or ice creams. In such cases, the user tag may describe concepts or objects that fall under the same broad category but differ from the specific items the object detection model is trained to recognize. Some tags may also refer to objects that are not very salient in the visual scene, making them difficult for the model to detect. This mismatch between the user tags and the model's ability to identify objects likely contributes to the low detection rate observed. Moreover, we used YOLO9000 (released in 2016–17) because it is the only available model with $9,000$ classes, even though there are more powerful object detection models (YOLOv9) available. For our task, this was one of the crucial reasons for selecting the model. We wanted to detect as many object classes as possible since we can not know which of these object classes may be present within images of concepts, especially for abstract concepts.

Dataset	Number of Images (%) A	C
YFCC - 500	10.02	9.48
YFCC - 400	10.01	9.37
YFCC - 300	10.07	9.28
YFCC - 200	10.09	9.06
YFCC - 100	10.00	8.87
YFCC - 25	10.08	8.64
Bing - 25	14.28	15.28

Table 3: Average number (percentage) of images for abstract (A) and concrete (C) concepts containing at least one object detected by the YOLO9000 model.

8.2 Classification Results for Different Classifiers

In the classification study in Section 4, we experimented with three different classifiers: Support Vector Machines (SVM) with *rbf* kernel, Random Forests (RF), and Logistic Regression (LR). The results for the RF model are reported in the main text (Figures 2 and 3). The results, combined and by class, for Logistic Regression can be found in Figure 5. The results for SVM are presented in Figure 6.

8.3 Eigenvalues and How to Infer Them?

We use eigenvalues to extract characteristics of the similarity matrix in Study 1. The top eigenvalues capture the most information about the similarity matrix as they represent the variance of principal components. Hence, they are expected to have the most information on the similarity/variance of images. So, all high eigenvalues would indicate very diverse images for a feature, whereas all low eigenvalues would suggest high similarity.

8.4 Nearest Neighbor Results

Table 5 supplements the results shown in Table 2 by incorporating the nearest neighbor analysis with varying quantities of images per concept (ranging from 100 to 400) extracted from the YFCC dataset.

8.5 Cosine Similarity Comparison between Abstract and Concrete Concepts

Attribute	Bing-25		YFCC-25	
	A	C	A	C
Color	0.91	0.92	0.92	0.92
HOG	0.78	0.80	0.80	0.81
Texture	0.99	0.99	0.99	0.99
GIST	0.91	0.91	0.93	0.93
SURF	0.61	0.64	0.42	0.42
YOLO	0.95	0.89	0.91	0.86
Object Loc.	0.85	0.84	0.81	0.80
Combined	0.98	0.98	0.98	0.98
Simclr	0.98	0.98	0.99	0.99
ViT	0.58	0.56	0.56	0.52

Table 4: Average cosine similarities for abstract (A) and concrete (C) concepts for the Bing-25 and YFCC-25 datasets.

Table 4 shows a comparison of cosine similarity scores for the top 25 nearest neighbors of an image, evaluated across different visual features. The similarity scores are generally consistent across feature type both for concrete and abstract targets, and across different datasets. Vision Transformer (ViT) stand out for having lower scores compared to the other features.

8.6 Crowd-sourcing Collections

As discussed in Section 6, we collected data using crowd-sourcing methods. The classification of 18 concepts (8 concrete and 10 abstract) in five “reasons for visual diversity" is reported in Figure 7. Tables 7 and 8 provide examples of words describing the images of five concrete and five abstract concepts.

8.7 Model Details

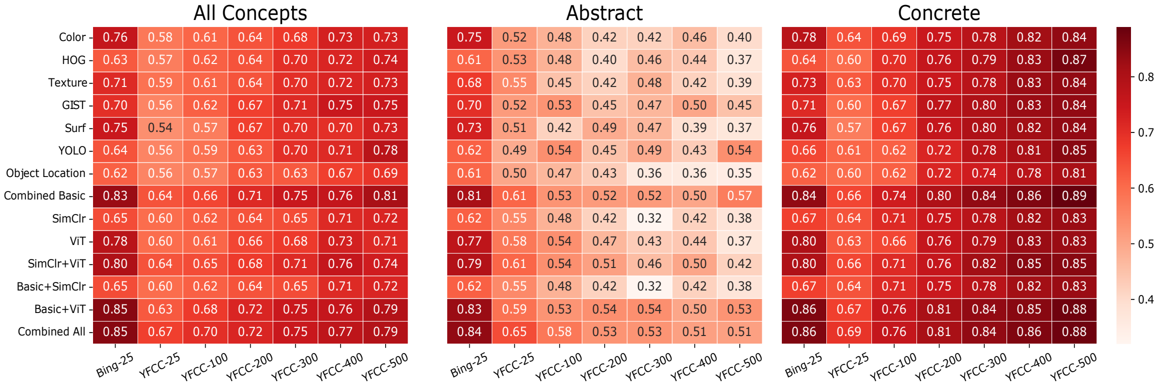
In Study 1, we used three classifiers: Random forest (RF), SVM and Logistic Regression from the scikit-learn library (Pedregosa et al., 2011), and performed an extensive hyper-parameter search with 5-fold cross-validation. For RF, the hyper-parameters included *number of estimators* (trees), *max_depth* (maximum depth of the tree), *min_samples_split* (minimum number of samples required to split an internal node), *min_samples_leaf* (minimum number of samples required at a leaf node) and *max_features* (number of features to be considered for determining the best split).

For feature extraction for the YOLO model, we used an NVIDIA RTX A6000 GPU. It takes around 8 hours of GPU processing to extract YOLO features. The computation of nearest neighbors takes multiple weeks.

8.8 Regression Analysis

We use a *Gradient Boosted trees* model to predict the concreteness of each target concept using the eigenvalues of the combined visual features described in Section 3.3 as predictors. The predicted concreteness scores are compared against the *Brysbaert* norms using Spearman’s rank-order correlation coefficient ρ. We use an 80:20 data split between train and test sets with Monte Carlo cross-validation.

As shown in Table 6, the combination of all the low-level features (Combined) achieves the highest results for both datasets and outperforms both ViT and SimClr more complex representations. This is in line with the classification results. Similar to classification, we also further investigate the sampling bias of images, we conduct similar analysis for concepts with $100, 200, 300, 400$ and 500 images. We see similar results as depicted in Figure 3 in the main text. As expected, Spearman correlations generally improve with the inclusion of more images, as increased data helps to average out noise.



All Concepts

	Bing-25	YFCC-25	YFCC-100	YFCC-200	YFCC-300	YFCC-400	YFCC-500
Color	0.76	0.58	0.61	0.64	0.68	0.73	0.73
HOG	0.63	0.57	0.62	0.64	0.70	0.72	0.74
Texture	0.71	0.59	0.61	0.64	0.70	0.72	0.73
GIST	0.70	0.56	0.62	0.67	0.71	0.75	0.75
Surf	0.75	0.54	0.57	0.67	0.70	0.70	0.73
YOLO	0.64	0.56	0.59	0.63	0.70	0.71	0.78
Object Location	0.62	0.56	0.57	0.63	0.63	0.67	0.69
Combined Basic	0.83	0.64	0.66	0.71	0.75	0.76	0.81
SimClr	0.65	0.60	0.62	0.64	0.65	0.71	0.72
ViT	0.78	0.60	0.61	0.66	0.68	0.73	0.71
SimClr+ViT	0.80	0.64	0.65	0.68	0.71	0.76	0.74
Basic+SimClr	0.65	0.60	0.62	0.64	0.65	0.71	0.72
Basic+ViT	0.85	0.63	0.68	0.72	0.75	0.76	0.79
Combined All	0.85	0.67	0.70	0.72	0.75	0.77	0.79

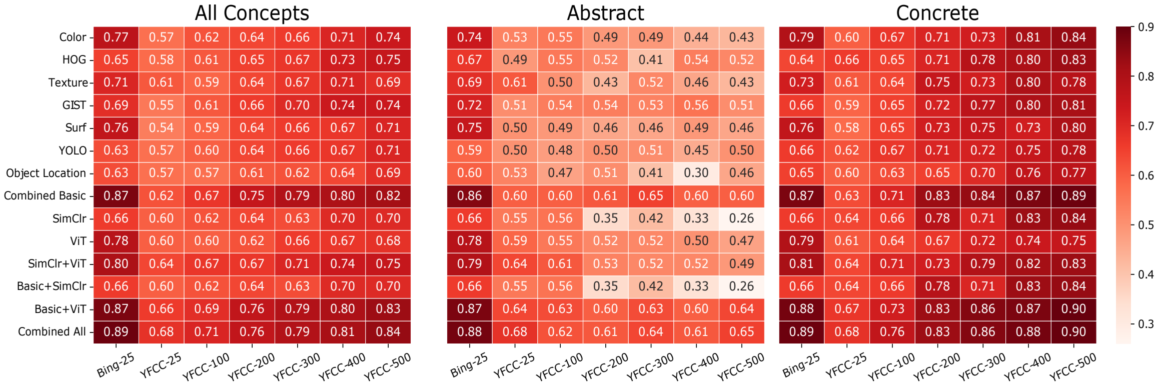
Abstract

	Bing-25	YFCC-25	YFCC-100	YFCC-200	YFCC-300	YFCC-400	YFCC-500
Color	0.75	0.52	0.48	0.42	0.42	0.46	0.40
HOG	0.61	0.53	0.48	0.40	0.46	0.44	0.37
Texture	0.68	0.55	0.45	0.42	0.48	0.42	0.39
GIST	0.70	0.52	0.53	0.45	0.47	0.50	0.45
Surf	0.73	0.51	0.42	0.49	0.47	0.39	0.37
YOLO	0.62	0.49	0.54	0.45	0.49	0.43	0.54
Object Location	0.61	0.50	0.47	0.43	0.36	0.36	0.35
Combined Basic	0.81	0.61	0.53	0.52	0.52	0.50	0.57
SimClr	0.62	0.55	0.48	0.42	0.32	0.42	0.38
ViT	0.77	0.58	0.54	0.47	0.43	0.44	0.37
SimClr+ViT	0.79	0.61	0.54	0.51	0.46	0.50	0.42
Basic+SimClr	0.62	0.55	0.48	0.42	0.32	0.42	0.38
Basic+ViT	0.83	0.59	0.53	0.54	0.54	0.50	0.53
Combined All	0.84	0.65	0.58	0.53	0.53	0.51	0.51

Concrete

	Bing-25	YFCC-25	YFCC-100	YFCC-200	YFCC-300	YFCC-400	YFCC-500
Color	0.78	0.64	0.69	0.75	0.78	0.82	0.84
HOG	0.64	0.60	0.70	0.76	0.79	0.83	0.87
Texture	0.73	0.63	0.70	0.75	0.78	0.83	0.84
GIST	0.71	0.60	0.67	0.77	0.80	0.83	0.84
Surf	0.76	0.57	0.67	0.76	0.80	0.82	0.84
YOLO	0.66	0.61	0.62	0.72	0.78	0.81	0.85
Object Location	0.62	0.60	0.62	0.72	0.74	0.78	0.81
Combined Basic	0.84	0.66	0.74	0.80	0.84	0.86	0.89
SimClr	0.67	0.64	0.71	0.75	0.78	0.82	0.83
ViT	0.80	0.63	0.66	0.76	0.79	0.83	0.83
SimClr+ViT	0.80	0.66	0.71	0.76	0.82	0.85	0.85
Basic+SimClr	0.67	0.64	0.71	0.75	0.78	0.82	0.83
Basic+ViT	0.86	0.67	0.76	0.81	0.84	0.85	0.88
Combined All	0.86	0.69	0.76	0.81	0.84	0.86	0.88

Figure 5: Weighted F1-scores (overall and by class) for different features and different dataset sizes for Bing and YFCC using **Logistic Regression**.



All Concepts

	Bing-25	YFCC-25	YFCC-100	YFCC-200	YFCC-300	YFCC-400	YFCC-500
Color	0.77	0.57	0.62	0.64	0.66	0.71	0.74
HOG	0.65	0.58	0.61	0.65	0.67	0.73	0.75
Texture	0.71	0.61	0.59	0.64	0.67	0.71	0.69
GIST	0.69	0.55	0.61	0.66	0.70	0.74	0.74
Surf	0.76	0.54	0.59	0.64	0.66	0.67	0.71
YOLO	0.63	0.57	0.60	0.64	0.66	0.67	0.71
Object Location	0.63	0.57	0.57	0.61	0.62	0.64	0.69
Combined Basic	0.87	0.62	0.67	0.75	0.79	0.80	0.82
SimClr	0.66	0.60	0.62	0.64	0.63	0.70	0.70
ViT	0.78	0.60	0.60	0.62	0.66	0.67	0.68
SimClr+ViT	0.80	0.64	0.67	0.67	0.71	0.74	0.75
Basic+SimClr	0.66	0.60	0.62	0.64	0.63	0.70	0.70
Basic+ViT	0.87	0.66	0.69	0.76	0.79	0.80	0.83
Combined All	0.89	0.68	0.71	0.76	0.79	0.81	0.84

Abstract

	Bing-25	YFCC-25	YFCC-100	YFCC-200	YFCC-300	YFCC-400	YFCC-500
Color	0.74	0.53	0.55	0.49	0.49	0.44	0.43
HOG	0.67	0.49	0.55	0.52	0.41	0.54	0.52
Texture	0.69	0.61	0.50	0.43	0.52	0.46	0.43
GIST	0.72	0.51	0.54	0.54	0.53	0.56	0.51
Surf	0.75	0.50	0.49	0.46	0.46	0.49	0.46
YOLO	0.59	0.50	0.48	0.50	0.51	0.45	0.50
Object Location	0.60	0.53	0.47	0.51	0.41	0.30	0.46
Combined Basic	0.86	0.60	0.60	0.61	0.65	0.60	0.60
SimClr	0.66	0.55	0.56	0.35	0.42	0.33	0.26
ViT	0.78	0.59	0.55	0.52	0.52	0.50	0.47
SimClr+ViT	0.79	0.64	0.61	0.53	0.52	0.52	0.49
Basic+SimClr	0.66	0.55	0.56	0.35	0.42	0.33	0.26
Basic+ViT	0.87	0.64	0.63	0.60	0.63	0.60	0.64
Combined All	0.88	0.68	0.62	0.61	0.64	0.61	0.65

Concrete

	Bing-25	YFCC-25	YFCC-100	YFCC-200	YFCC-300	YFCC-400	YFCC-500
Color	0.79	0.60	0.67	0.71	0.73	0.81	0.84
HOG	0.64	0.66	0.65	0.71	0.78	0.80	0.83
Texture	0.73	0.61	0.64	0.75	0.73	0.80	0.78
GIST	0.66	0.59	0.65	0.72	0.77	0.80	0.81
Surf	0.76	0.58	0.65	0.73	0.75	0.73	0.80
YOLO	0.66	0.62	0.67	0.71	0.72	0.75	0.78
Object Location	0.65	0.60	0.63	0.65	0.70	0.76	0.77
Combined Basic	0.87	0.63	0.71	0.83	0.84	0.87	0.89
SimClr	0.66	0.64	0.66	0.78	0.71	0.83	0.84
ViT	0.79	0.61	0.64	0.67	0.72	0.74	0.75
SimClr+ViT	0.81	0.64	0.71	0.73	0.79	0.82	0.83
Basic+SimClr	0.66	0.64	0.66	0.78	0.71	0.83	0.84
Basic+ViT	0.88	0.67	0.73	0.83	0.86	0.87	0.90
Combined All	0.89	0.68	0.76	0.83	0.86	0.88	0.90

Figure 6: Weighted F1-scores (overall and by class) for different features and different dataset sizes for Bing and YFCC using **Support Vector Machines**.

Attribute	**YFCC-100**		**YFCC-200**		**YFCC-300**		**YFCC-400**	
	A	**C**	**A**	**C**	**A**	**C**	**A**	**C**
Color	1.28	0.79	0.99	0.72	0.88	0.68	0.86	0.66
HOG	0.47	0.48	0.34	0.46	0.32	0.45	0.34	0.44
Texture	0.30	0.24	0.26	0.25	0.26	0.26	0.27	0.26
GIST	0.69	0.61	0.53	0.58	0.50	0.57	0.51	0.56
SURF	0.65	0.55	0.44	0.53	0.40	0.53	0.52	0.52
YOLO	1.58	1.38	1.70	1.46	1.65	1.50	1.66	1.54
Object Loc.	0.20	0.23	0.19	0.24	0.21	0.25	0.23	0.26
Combined	1.03	0.85	0.78	0.80	0.71	0.76	0.70	0.76
Simclr	0.80	0.65	1.67	1.67	1.40	1.45	0.53	0.56
ViT	2.79	6.71	2.30	6.67	4.55	11.99	2.26	6.55

Table 5: Average percentage of visual nearest neighbors (out of 100, 200, 300 or 400, respectively) associated with the same abstract (A) or concrete (C) concept.

Visual Attribute	**Bing**		**YFCC**	
	ρ	**RMSE**	ρ	**RMSE**
Color	0.52	1.34	0.16	1.58
HOG	0.24	1.53	0.12	1.60
Texture	0.42	1.41	0.17	1.57
GIST	0.38	1.43	0.07	1.61
SURF	0.49	1.34	0.07	1.61
YOLO	0.26	1.54	0.07	1.61
Object Location	0.21	1.67	0.01	1.62
Combined	**0.63**	**1.12**	**0.30**	**1.51**
SimClr	0.28	1.87	0.17	1.90
ViT	0.56	1.27	0.20	1.85

Table 6: Spearman correlation scores (ρ) and Root-mean-squared-error (RMSE) comparing the predicted concreteness scores using different visual attributes to the *Brysbaert* norms. Results for the Bing and the YFCC datasets. In bold-font we highlight the highest scores for each dataset.

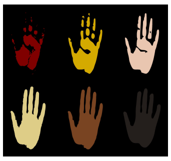
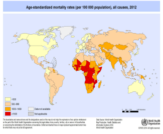
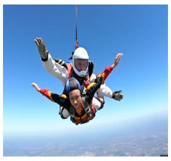
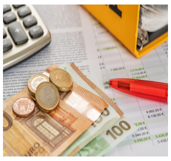
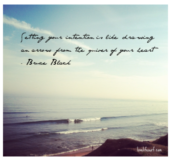
	equality (1.41)	**mortality** (1.46)	**courage** (1.52)	**accountancy** (1.68)	**intention** (1.70)
					
number of distinct annotations	17	18	24	10	18
Annotations	red: 7 yellow: 7 brown: 5 hand: 3 grey: 2 pink: 2 black: 2 sandal: 2 hi: 1 ash: 1 orange: 1 white: 1 hand print: 1 color: 1 fingers: 1 six hand: 1 six colors: 1	map: 6 world map: 4 sea: 4 country: 3 continent: 3 ocean: 2 yellow: 2 earth: 2 orange: 1 india: 1 world: 1 desert: 1 articles: 1 red: 1 letters: 1 lands: 1 mortality rate:1 population: 1	sky: 6 fly: 6 adventures: 3 diving: 2 exciting: 2 air: 2 helmet: 2 person: 2 man: 1 women: 1 skydive: 1 nature: 1 coat: 1 rope: 1 hand: 1 focus: 1 two: 1 male: 1 female: 1 advancer: 1 skydress: 1 flying: 1 hanging: 1 helpmate: 1	coin: 9 pen: 8 calculator: 8 money: 5 file: 5 calculate: 1 rupees: 1 pencil: 1 paper: 1 euro notes: 1	sky: 6 sea: 5 beach: 3 waves: 3 sand: 3 quotes: 2 water: 2 stone: 1 white: 1 motivation: 1 happy life: 1 good intention:1 peaceful: 1 set goal: 1 blue: 1 post card: 1 ocean: 1 words: 1

Table 7: Words generated by nine participants when answering to the question “What is depicted in each image?". Examples for five images of *abstract concepts* (and their concreteness score).

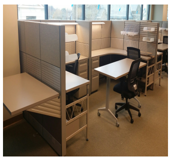
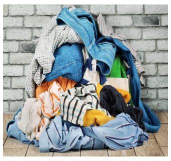
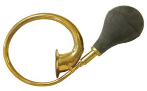
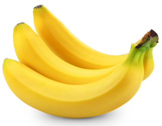
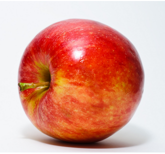
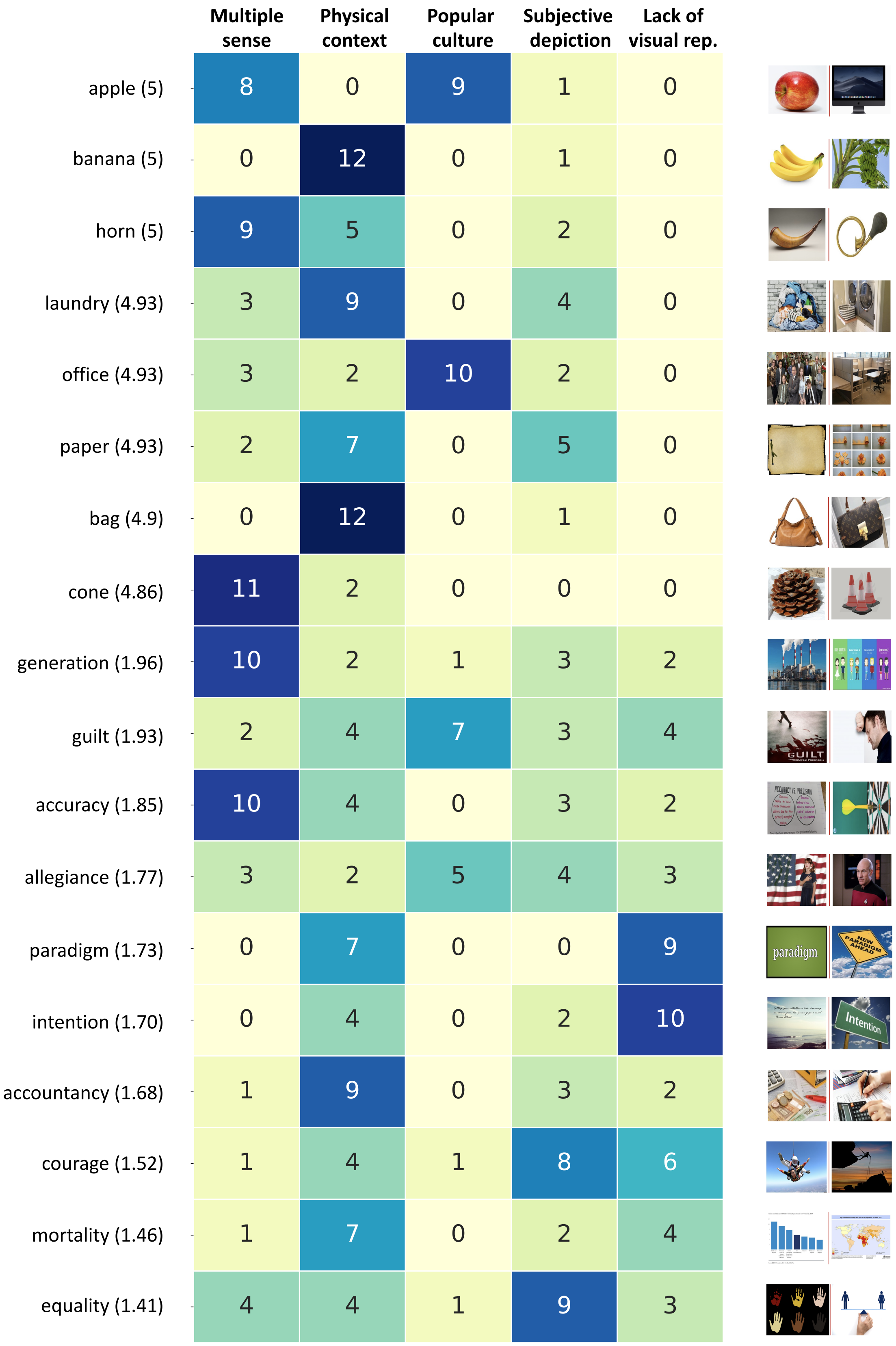
	office (4.93)	**laundry** (4.93)	**horn** (5.00)	**banana** (5.00)	**apple** (5.00)
					
number of distinct annotations	14	16	29	14	16
Annotations	chair: 9 table: 8 window: 7 desk: 3 room: 2 glass: 2 light: 1 furniture: 1 office furniture: 1 building: 1 cotton: 1 floor: 1 office: 1 drawer: 1	clothes: 8 wall: 6 pant: 4 jeans: 3 laundry: 2 dress: 2 shirt: 2 floor: 2 bricks: 2 color: 2 garments: 1 trousers: 1 stone: 1 tiles: 1 tshirt: 1 blue: 1	horn: 5 brass: 3 retro old-timer: 1 brass bulb: 1 motor horn: 1 rubber horn: 1 steel: 1 rubber: 1 circle: 1 oval: 1 honking sound : 1 brass honking instrument: 1 sound: 1 metal: 1 mike: 1 black: 1 rubber bulb: 1 musical instrument: 1 sound instrument: 1 signal horn: 1 military bugle: 1 brass instrument: 1 hunting horn:1 conical horn: 1 honk: 1 rubber top: 1 metal instrument: 1 bulb horn: 1 large circular: 1	banana: 13 yellow: 8 three: 3 fruit: 2 three banana: 1 fresh fruit: 1 very sweet fruit: 1 white: 1 green: 1 fresh: 1 sweet: 1 healthy: 1 curved: 1 ripened: 1	fruit: 7 apple: 6 fresh: 4 red apple: 3 red: 3 stem: 2 health: 2 eating: 1 good for health: 1 organic: 1 one: 1 fruits: 1 fresh fruits: 1 paradise apple: 1 shadow: 1 healthy: 1

Table 8: Words generated by nine participants when answering to the question "What is depicted in each image?". Examples for five images of *concrete concepts* (and their concreteness score).



	Multiple sense	Physical context	Popular culture	Subjective depiction	Lack of visual rep.
apple (5)	8	0	9	1	0
banana (5)	0	12	0	1	0
horn (5)	9	5	0	2	0
laundry (4.93)	3	9	0	4	0
office (4.93)	3	2	10	2	0
paper (4.93)	2	7	0	5	0
bag (4.9)	0	12	0	1	0
cone (4.86)	11	2	0	0	0
generation (1.96)	10	2	1	3	2
guilt (1.93)	2	4	7	3	4
accuracy (1.85)	10	4	0	3	2
allegiance (1.77)	3	2	5	4	3
paradigm (1.73)	0	7	0	0	9
intention (1.70)	0	4	0	2	10
accountancy (1.68)	1	9	0	3	2
courage (1.52)	1	4	1	8	6
mortality (1.46)	1	7	0	2	4
equality (1.41)	4	4	1	9	3

Figure 7: Main reasons of visual diversity between two images of 18 concepts (8 concrete and 10 abstract) according to 13 participants. At least one reason had to be selected for each concept.