Concept-pedia: a Wide-coverage Semantically-annotated Multimodal Dataset

Karim Ghonim[1] **Andrei Stefan Bejgu**[2] **Alberte Fernández-Castro**[1,2]
Roberto Navigli[1,2]

Sapienza University of Rome
[1]surname@diag.uniroma1.it

Babelscape, Italy
[2]surname@babelscape.com

Abstract

Vision-language Models (VLMs), such as CLIP and SigLIP, have become the *de facto* standard for multimodal tasks, serving as essential building blocks for recent Multimodal Large Language Models, including LLaVA and PaliGemma. However, current evaluations for VLMs remain heavily anchored to ImageNet. In this paper, we question whether ImageNet's coverage is still sufficiently challenging for modern VLMs, and investigate the impact of adding novel and varied concept categories, i.e., semantically grouped fine-grained synsets. To this end, we introduce Concept-pedia, a novel, large-scale, semantically-annotated multimodal resource covering more than 165,000 concepts. Leveraging a language-agnostic, automatic annotation pipeline grounded in Wikipedia, Concept-pedia expands the range of visual concepts, including diverse abstract categories. Building on Concept-pedia, we also present a manually-curated Visual Concept Recognition evaluation benchmark, Concept-10k, that spans thousands of concepts across a wide range of categories. Our experiments show that current models, although excelling on ImageNet, struggle with Concept-10k. Not only do these findings highlight a persistent bias toward ImageNet-centric concepts, but they also underscore the urgent need for more representative benchmarks. By offering a broader and semantically richer testbed, Concept-10k aims to support the development of multimodal systems that better generalize to the complexities of real-world visual concepts.

1 Introduction

Vision-language models have rapidly advanced in recent years, achieving remarkable performance on zero-shot Visual Concept Recognition benchmarks such as ImageNet-1k (Deng et al., 2009) and ObjectNet (Barbu et al., 2019). Prominent models like CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023), which aim to learn a shared representation space for text and images simultaneously, have become standard approaches in multimodal research.

The success of vision-text models can largely be attributed to their training on large-scale image-text datasets, which allow them to recognize a broad spectrum of concepts represented in standard benchmarks. This has been thoroughly studied by Schuhmann et al. (2022) and Gadre et al. (2023), among others, who analyze how the quantity of data, combined with the data curation techniques used during model pre-training, impacts zero-shot concept recognition capabilities. However, despite advancements in multimodal models, concept representation capabilities are predominantly evaluated using datasets like ImageNet-1k, iNaturalist (Horn et al., 2018) and ObjectNet. iNaturalist covers more than 8,000 concepts, specializing in plants and animals, while ObjectNet includes 313 concepts and explicitly focuses on objects. Although ImageNet-1k covers just 1,000 classes, it encompasses a broader and more diverse set of categories spanning objects, animals, food, and more, establishing itself as the de facto benchmark for general-purpose Visual Concept Recognition. This extensive coverage makes ImageNet-1k a more general and versatile evaluation benchmark for multimodal models.

ImageNet has been fundamental in shaping advancements in both Computer Vision and vision-text research. However, the field's heavy reliance on ImageNet comes with some well-documented issues. First, multiple studies have raised concerns that the reported performance gains might stem from models effectively "overfitting" to ImageNet's specific collection of images and labels, rather than genuinely improving generalization abilities (Recht et al., 2019; Beyer et al., 2020). Second, ImageNet is not exempt from data-quality issues, including ambiguous or outright incorrect annotations, and its single-label assignment for each image is often in-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34406–34427

sufficient given the multi-object nature of many images (Beyer et al., 2020). These limitations can impact model accuracy without adequately reflecting real-world robustness. Moreover, with the increasingly expanding range of multimodal applications today, relying on ImageNet as a general-purpose benchmark is becoming increasingly restrictive due to its constrained coverage of concepts. Indeed, ImageNet currently includes entries for approximately 21,841 concepts, which represent 26% of all WordNet nominal synsets.[1] However, its only manually validated benchmark, ImageNet-1k, provides images for just 1,000 concepts, which the research community continues to rely on for evaluation in the absence of broader alternatives.

In this work, we investigate whether the performance gains achieved on ImageNet truly generalize to a broader and more varied range of concepts. To enable this assessment, we present Concept-pedia, a wide-coverage multimodal resource that provides semantically-annotated images and texts. Concept-pedia covers a significantly larger and more varied set of concepts compared to existing datasets. Using Concept-pedia, we show that current models exhibit a bias toward ImageNet images and concepts.

To summarize, our contributions are as follows:

- Concept-pedia, a wide-coverage semantically-annotated multimodal resource that provides multi-label annotations, covering 165,000 concepts across 28 categories;
- A novel manually-curated evaluation benchmark, Concept-10k, that covers $\approx$ 10k concepts;
- A study on the ability of models to generalize to images and concepts beyond existing benchmarks such as ImageNet.

In the hope of fostering research in Multimodal Concept Recognition, we release the code and data at https://github.com/SapienzaNLP/concept-pedia.

2 Related Work

ImageNet has acted as an anchor for tracking progress in multimodal research, particularly in assessing the proficiency of multimodal models in detecting concepts within images. This is largely because ImageNet was designed to cover a wide range of concepts, i.e., WordNet synsets, ensuring representation across diverse categories, thus serving as a pre-training resource (ImageNet-21k) and a standard evaluation benchmark (ImageNet-1k). However, the ImageNet-21k dataset (Ridnik et al., 2021) is limited to approximately 22,000 concepts, predominantly skewed toward objects. To address this limitation, we base our resource on Wikipedia, leveraging its manually added hyperlinks to ensure broader concept coverage. As a result, Concept-pedia encompasses 165,000 concepts, spanning a wider range of categories and exhibiting a more balanced distribution. Wikipedia has previously been utilized to create large-scale training corpora for various objectives. For instance, Srinivasan et al. (2021) introduced Wikipedia-based Image Text (WIT), a multilingual multimodal dataset extracted from Wikipedia. However, WIT contains image-text data only and lacks concept annotations, which Concept-pedia provides. On the other hand, Wikipedia hyperlinks have already been successfully used to automatically create datasets for several NLP tasks, such as Named Entity Recognition (Nothman et al., 2013; Al-Rfou et al., 2015; Tsai et al., 2016; Pan et al., 2017; Tedeschi et al., 2021; Martinelli et al., 2024), Word Sense Disambiguation, Semantic Role Labeling, Semantic Parsing, Relation Extraction (all introduced in Conia et al. (2024)), and Entity Linking (Wu et al., 2020). However, to the best of our knowledge, Concept-pedia is the first multimodal resource that leverages Wikipedia hyperlinks to annotate images with concepts.

Moreover, ImageNet-21k lacks an official human-validated evaluation benchmark. The only widely used ImageNet benchmark, ImageNet-1k, provides samples for just 1,000 concepts, primarily consisting of objects and animals. To evaluate models' robustness on out-of-distribution images, several variations of ImageNet have been introduced, such as ImageNet-v2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021a), and ImageNet-A (Hendrycks et al., 2021b), incorporating new images for ImageNet concepts. Another variation of ImageNet-1k is ImageNet-ReaL (Beyer et al., 2020), which addresses annotation inaccuracies and the single-label limitation in ImageNet-1k by relabeling the images and assigning multiple labels where applicable. However, all these variations remain limited to the same original set of concepts used in ImageNet. This limitation is even

[1] https://www.image-net.org/

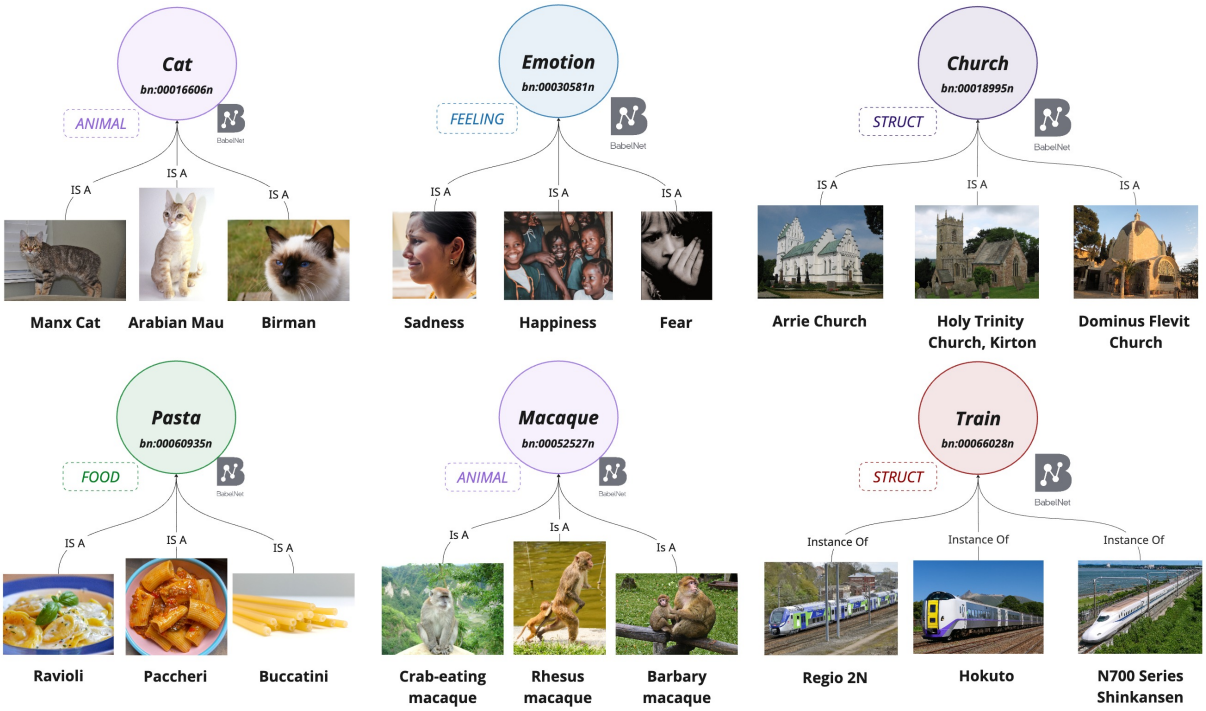


Figure 1: Examples for taxonomical concept population in Concept-pedia across different CNER categories.

more evident in ObjectNet and iNaturalist, which include images for 313 and 8,142 concepts, respectively. Similarly, ImageNet-CoG (Sariyildiz et al., 2021), which was explicitly designed to include concepts beyond those in ImageNet, remains heavily skewed toward the same categories as ImageNet-1k, while BabelPic (Calabrese et al., 2020) focuses explicitly on non-concrete (NC) concepts, such as emotions. These constraints underscore the need for a broader benchmark to complement the existing, domain-specific datasets. To address this issue, we leverage Wikipedia, in combination with semantic categories, with the novel goal of creating a more diverse and comprehensive benchmark, i.e., Concept-10k.

More specifically, Concept-10k contains images for approximately 10,000 concepts, making it substantially larger than currently used general-purpose benchmarks. Furthermore, it ensures a more balanced distribution across categories, avoiding the skew toward specific categories such as objects or animals present in benchmarks like ImageNet-1k or iNaturalist.

3 Concept-pedia

In this section, we outline the process of creating Concept-pedia, detailing how images from Wikipedia are associated with their respective concepts (see Figure 11). We use hyperlinks in image captions as the basis for mapping images to concepts with the help of BabelNet[2] (Navigli and Ponzetto, 2010; Navigli et al., 2021), a multilingual knowledge graph. However, since Wikipedia is manually edited, many captions lack direct hyperlinks. To address this limitation, we automatically annotate the captions with the missing links. Next, we generalize entities to the concepts they are instances of (e.g. Colosseum is an instance of amphitheater), and ensure that each concept in our resource is associated with an adequate number of images by exploiting BabelNet's taxonomical information. Striking this balance is crucial to maintaining the dataset's quality and representativeness.

3.1 Link propagation

Wikipedia pages provide manually annotated hyperlinks, including those found in image captions. However, many mentions of entities or concepts within captions lack corresponding hyperlinks, resulting in missed annotations, as illustrated in Figure 2. To overcome this issue, we employ a link propagation algorithm inspired by Tedeschi et al. (2021), which matches unannotated text chunks in a caption with links present in the Wikipedia page containing the caption.

The algorithm operates on the assumption that captions frequently reference mentions already

[2] https://babelnet.org

linked within the body of the corresponding Wikipedia page. Building on this assumption, the propagation algorithm attempts to match unlinked text chunks in captions with manually annotated links in the page. This matching process relies on a sequence of cascaded heuristics to expand the annotations. First, **exact string matching** identifies exact surface form overlaps between text chunks and linked mentions, ensuring high precision when captions use identical terminology. Then, **lemma matching** extends this by normalizing text to its lemmatized form, which we perform using SpaCy[3], enabling matches that account for inflectional variations such as pluralization or verb conjugation (e.g., *cathedrals* being matched with *cathedral*).

To further link more unannotated text chunks, we propose two heuristics that utilize the inter-resource mapping provided by BabelNet to match text chunks with synonyms or alternative lexicalizations. The first heuristic, referred to as **synonym matching**, obtains for each existing link in a Wikipedia page the corresponding BabelNet synset, and identifies unannotated text chunks that match WordNet synonyms or Wikidata aliases contained within that synset. For instance, *natatorium* can be linked to the *swimming pool* page since both terms are synonyms in the corresponding WordNet synset, and *gas* linked to *gasoline* due to the former being a Wikidata alias. This technique is particularly useful for handling variations in naming conventions, such as acronyms or synonyms. The second heuristic, i.e. **redirection matching**, performs an analogous matching using Wikipedia page redirections, i.e. pages that redirect to other pages. This allows us to capture alternative spellings, common misspellings, and other variations, e.g. *Red Panda* and *Lesser Panda*.

While the link propagation algorithm significantly enhances coverage, the precision of its heuristics varies. Hyperlinks manually added to captions naturally exhibit the highest precision, as they were explicitly added by Wikipedia editors. In contrast, heuristics like synonym or redirection matching introduce a degree of uncertainty, e.g. potentially introducing concept shifts (e.g. *Driver (person)* redirecting to *Driving*). To maximize coverage while maintaining transparency, Conceptpedia includes annotations derived from all heuristics and records the link propagation information for each link. This enables downstream applications to tailor their use of the annotations based on their needs – prioritizing precision (e.g., by using only existing or exact matches) or maximizing recall (e.g., by incorporating all annotations). We discuss this in more detail in Section 5.

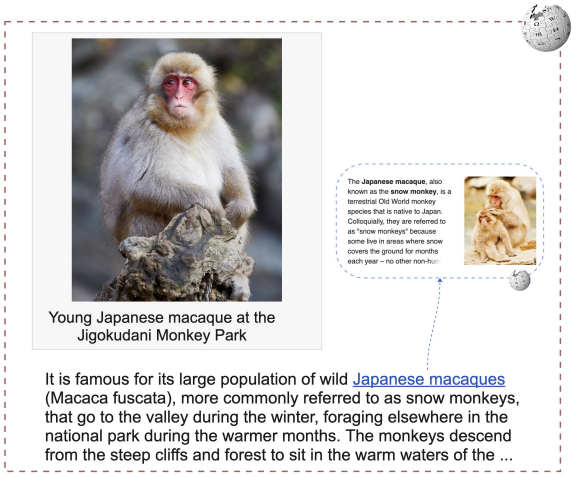


Figure 2: An example of a Wikipedia page containing an image-caption pair in which the caption lacks a hyperlink to the mentioned concept, while the page body contains the corresponding link.

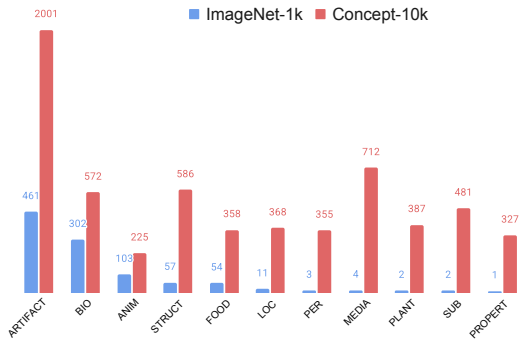
This link propagation approach to creating multimodal datasets achieves two key objectives. First, creating manually annotated datasets is notoriously expensive; this method repurposes high-quality manual annotations, thereby significantly reducing costs. Second, as highlighted by Beyer et al. (2020), images inherently contain multiple concepts, which our link propagation algorithm helps associate. By leveraging this approach, we produce what is, to the best of our knowledge, the first semantically-annotated multimodal training resource of this scale with multi-label concept annotations.

3.2 Taxonomical Concept Propagation

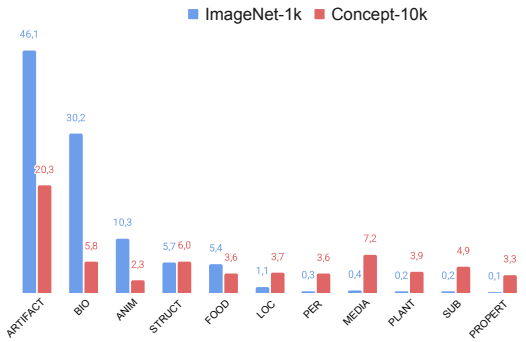
After the link propagation step, 74.5% of the 4.6 million English Wikipedia captions (Nov, 2023 dump) are annotated with at least one link. We leverage the correspondence established in BabelNet between its synsets and Wikipedia links to map these annotations. This enables us to exploit BabelNet's synset classification into concepts (e.g. *City*) and named entities (e.g. *Paris*), revealing that approximately 85% of the synsets involved are classified as named entities.

To maximize the size of our dataset and maintain a focus on concepts rather than specific entities, we transform each entity annotation into its cor-

[3] https://spacy.io/



(a) Number of concepts covered per category.



(b) Distribution (%) of categories for each benchmark.

Figure 3: Comparison of concept and category distributions between Concept-10k and ImageNet-1k.

Propagation	Capt.	Links	Concept	
			Unique	Freq.
w/o	1.1M	0.6M	114K	46K
+ Link	3.4M	1.2M	165K	48K
+ Taxonomical	3.4M	1.2M	101K	101K

Table 1: Statistics of Concept-pedia at each stage of the propagation algorithm, showing the number of captions with at least one link (Capt.), unique Wikipedia link annotations (Links), unique concepts (Unique), and frequent concepts that appear more than 5 times (Freq.) at each step.

responding generalized concept using BabelNet's taxonomic information, which is derived from the integration of WordNet and Wikidata hypernymy and instance-of relations. For example, the entity *Arrie Church* is generalized to the concept *Church* (see Figure 1). Additionally, we observe that many concepts are associated with only a few images (see Table 1, last two columns). To ensure that each concept is adequately represented, we iteratively refine the annotations for underrepresented synsets, i.e. those synsets with fewer images than a predetermined threshold. To do so, we map each annotation to its direct parent concept (e.g., mapping *Hokuto* to *Train*). This process is repeated until every concept meets the minimum image requirement, ensuring a balanced distribution across the training, development, and test splits.

An additional benefit of grounding our dataset in BabelNet is that its taxonomy preserves a hierarchical structure, similarly to ImageNet in this respect, while covering a broader range of synsets beyond those available in WordNet.

3.3 Dataset Statistics

Table 1 presents Concept-pedia statistics at each step, including the number of captions, links, and concept annotations. Additionally, we report the number of concepts that occur at least five times, referred to as frequent concepts.

We observe that, while the Link Propagation step (Section 3.1) doubles the number of links (row 2 of Table 1) and adds approximately 50,000 new concepts, it has little effect on frequent concepts, i.e., concepts appearing more than 5 times. In contrast, the Taxonomical Propagation step (Section 3.2) flattens the set of unique concepts to 101,354 and ensures that each of them is associated with at least 5 images in the dataset (row 3 of Table 1).

4 Concept-10k

Having created a large silver dataset with Concept-pedia, we can proceed to develop a gold-standard benchmark. Based on the results of our manual validation, presented in Section 5, we observed that larger captions or those containing multiple concepts are more prone to propagating concepts that are not visually represented in the images. To address this, starting with concepts that have at least 10 images, we apply three heuristics to Concept-pedia to select the most representative image-concept pairs: 1) we filter out all image-caption pairs with multiple concepts, leaving their inclusion to future work; 2) for each concept, we rank images by caption length and retain the k shortest-caption images (we set $k = 5$); 3) we remove any concepts that are parents of other concepts within the same set according to the BabelNet taxonomy, e.g., if we have *Car* and *Race car*, we remove *Car* to avoid false negatives at test time.

	Concepts	Categories
ImageNet-1k	1,000	11
ObjectNet	312	1
iNaturalist	8,142	3
Concept-10k	**9,837**	**28**

Table 2: Number of unique concepts and semantic categories covered in each dataset.

Propagation Type	P
Hyperlink in caption	100
Exact string matching	100
Lemma matching	91
Synonym matching	96
Redirection matching	84
Taxonomical propagation	96

Table 3: Precision (%) scores for the link and taxonomical concept propagation heuristics.

The final step in creating our Concept-10k gold-standard benchmark involves manually validating each image-concept pair. In this task, two expert linguists[4] with extensive experience in annotating lexical-semantic tasks were presented with image-concept pairs and instructed to perform a binary classification (True or False) based on whether the image was labeled with the correct concept. Additionally, the annotators were tasked with filtering out any Not Safe For Work (NSFW) or corrupt images. The inter-annotator agreement for this task was measured, yielding a Cohen's kappa coefficient of $\kappa = 89.3$, indicating almost perfect agreement. Further details regarding this annotation process can be found in Appendix A.3.

4.1 Benchmark Statistics and Analysis

The above procedure results in Concept-10k, a manually validated benchmark of 49,185 image-concept pairs, spanning 9,837 unique concepts, each represented by five images. To analyze the domain coverage of our final benchmark, we utilize the Concept and Named Entity Recognition (CNER) category set introduced by Martinelli et al. (2024), which maps BabelNet synsets to a predefined set of 29 semantic categories shared across concepts and entities. This mapping enables a systematic comparison of the diversity of our resource with that of ImageNet-1k, enabling an in-depth analysis of the category coverage of the two resources. We present Concept-pedia instances of concepts from various CNER categories in Appendix K.

ImageNet-1k spans only 11 out of 29 CNER categories, as shown in Table 2, with just 6 containing more than four distinct concepts. In contrast, Concept-10k covers 28 categories, excluding only *DateTime*, and includes a significantly higher number of concepts for the categories it shares with ImageNet-1k, as shown in Figure 3a. This demonstrates that Concept-10k enables a broader evaluation, encompassing not only concrete concepts but also abstract ones, such as *Feelings* or *Events*, which are often underrepresented in existing benchmarks. A comparison with other benchmarks, including those from Radford et al. (2021), is provided in Appendix I, showing that Concept-10k introduces more categories (28) than all other benchmarks combined (18). Furthermore, as we show in Figure 3b, not only does Concept-10k cover more concepts per category than ImageNet-1k but it also exhibits a less biased distribution toward categories such as *Artifacts*, i.e., objects, as seen in ImageNet-1k. Additional details on the distribution of the remaining categories in Concept-10k are provided in Appendix J.

5 Dataset Evaluation

To evaluate the quality of Concept-pedia and Concept-10k, we conduct a series of manual annotation experiments.

5.1 Annotation Quality Assessment

We first evaluate the effectiveness of the Link Propagation step in our Concept-pedia creation methodology (Section 3.1). To do so, we randomly sample 200 image-caption–link instances for each annotation type (i.e., source links and the heuristics described) and manually evaluate them. As shown in Table 3, precision is generally high, with all heuristics exceeding 90%, except for redirection matching, which, as expected, yields slightly lower precision. For the Taxonomical Concept Propagation step, we evaluate 300 randomly sampled links generalized to concepts, and observe a precision of 96%.

[4]Annotators were compensated according to the standard salary for their geographical location.

	ImageNet-1k	Concept-1k
High	88.4%	95.7%
Low	83.7%	76.8%

Table 4: Label accuracy of manually annotated subsets of ImageNet-1k and Concept-1k.

5.2 Comparison with ImageNet-1k

To further assess the quality of our silver dataset, we focus on two key aspects: the accuracy of the images in representing the given concepts and the precision of our automatic annotations compared to existing benchmarks, such as ImageNet. To this end, we conduct an experiment using an equal number of images from ImageNet and a restricted version of Concept-pedia that contains only ImageNet-1k concepts, referred to as Concept-1k. We evaluate 2,000 examples by randomly sampling two images per concept from both Concept-1k and ImageNet-1k, ensuring equal representation across concepts. This setup enables a fair comparison of quality between our dataset and a well-established benchmark.

Two expert linguists, following the guidelines detailed in Appendix A.2, are provided with 10 possible candidate labels, i.e., Wikipedia titles, for each image, retrieved from different vision-text models[5], and instructed to select all the prominent concepts depicted in the image. To ensure unbiased evaluation, instances are shuffled across resources and are shown without any source information. We accept *None* as a possible label to identify instances where the concept is correctly tagged given the context in the caption but not visually represented in the image, as shown in Figure 4. Given the fine-grained nature of ImageNet-1K classes (e.g., specific dog breeds), annotators are also asked to rate their confidence in their labels as High or Low. This process results in 84% of items in Concept-1K and 89% in ImageNet-1K being tagged with high confidence by both annotators.

For these items, as shown in Table 4, Concept-1k labels are confirmed as correct 95.7% of the time, compared to 88.4% for ImageNet-1k, validating the quality of our automatic labels and highlighting our dataset's competitiveness with a well-established benchmark. The inter-annotator agreement, measured on a shared subset of 200 instances, yields a Cohen's kappa coefficient of $\kappa = 88.1$, indicating almost perfect agreement.

[5]We use SigLIP-base, SigLIP-large, and SigLIP-(SO)-400M, CLIP-base and CLIP-large and select the 10 most predicted labels.

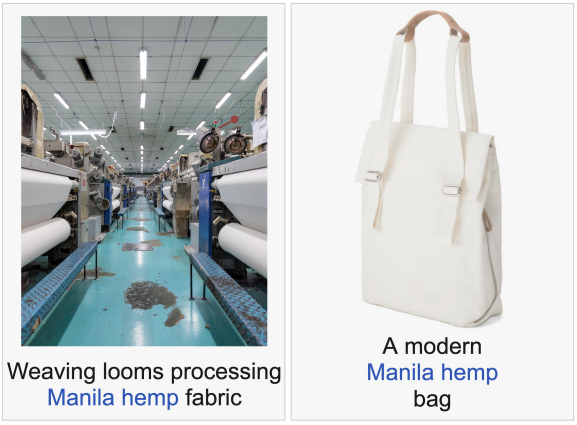


Figure 4: Examples of correct annotations excluded from Concept-10k. This is due to the concept not being visually represented in the image.

6 Experiments and Results

6.1 Experimental setup

We assess whether current models can generalize to new concepts beyond those included in ImageNet. To achieve this, we evaluate a diverse set of open-source vision-text transformer-based models, using them as baselines in our experiments and analyzing their accuracy on standard variations of ImageNet-1k (i.e., ImageNet-1k Val, ImageNet-v2, and ImageNet-ReaL) as well as on Concept-10k. Our evaluation includes the SigLIP family of models (Zhai et al., 2023), featuring three configurations: SigLIP-base, SigLIP-large, and SigLIP with a Shape-Optimized (SO) 400M parameter Vision Transformer (ViT) architecture. Additionally, we evaluate two models of the CLIP family (Radford et al., 2021), namely CLIP-base and CLIP-large. We select these models based on their strong performance (see Appendix G).

To test the models' ability to recognize concepts, we embed candidate concepts and input images into the same latent space and predict the class with the highest cosine similarity to the image embedding, and thus measure top-1 accuracy. This unified evaluation framework standardizes predictions across models, ensuring a fair and consistent comparison. Although SigLIP and CLIP are trained with different optimization objectives, as SigLIP employs a sigmoid-based loss function and CLIP utilizes a softmax loss, we apply this methodology to evaluate both models consistently. We carry out our

Model	ViT size	Params	Fine-Tuning	ImageNet-1k			Concept-10k
				Validation	v2	ReaL	
CLIP	B	150M	✗	68.3	61.9	-	22.0
SigLIP	B	203M	✗	**76.2**	**69.6**	**82.8**	31.1
SigLIP	B	203M	✓	75.9	69.1	81.5	**36.1**
CLIP	L	428M	✗	75.5	69.0	-	26.7
SigLIP	L	652M	✗	**80.5**	**74.2**	**85.9**	36.1
SigLIP	L	652M	✓	79.3	73.5	85.3	**41.9**
SigLIP	SO (400M)	878M	✗	**83.2**	**77.2**	**87.5**	40.3
SigLIP	SO (400M)	878M	✓	83.0	76.7	87.1	**45.0**

Table 5: We report the zero-shot and fine-tuned performance of different models. Performance is measured using top-1 accuracy (acc@1) for all datasets, except for ImageNet ReaL, where we report ReaL accuracy instead. We also report the performance of SigLIP fine-tuned on Concept-pedia on ImageNet test sets as well as on Concept-10k.

	ImageNet-1k	Concept-1k
SigLIP-base	86.3	79.1
SigLIP-large	87.6	83.0

Table 6: Model prediction accuracy (%) on manually annotated subsets of ImageNet-1k and Concept-1k, evaluated against high-confidence manual annotation labels.

evaluation in two scenarios, namely, zero-shot and fine-tuned, presented hereafter.

6.2 Zero-shot Visual Concept Recognition

In this section, we present the zero-shot performance of the models evaluated on both ImageNet-1k variants and our novel benchmark, Concept-10k. The latter aims to provide insight into the ability of pre-trained models to scale on categories not included in ImageNet-1k. As shown in Table 5, models perform relatively well on ImageNet-1k concepts, but struggle to accurately recognize the wide range of concepts present in the Concept-10k test set across the board. Even SigLIP-SO400m, the best-performing model on all variations of ImageNet-1k, fails to exceed 40.3% accuracy on Concept-10k.

To determine whether the introduced categories are responsible for the observed drop in performance, we evaluate model accuracy on Concept-1k, which includes only the concepts shared with ImageNet-1k Validation, as discussed in Section 5.2. As shown in Table 6, performance on Concept-1k decreases compared to ImageNet-1k. A similar performance drop is observed with ImageNet-v2 (cf. Table 5), which, like Concept-1k, introduces new images for ImageNet-1k concepts. These results suggest that the performance drop is primarily driven by the increased diversity of concepts rather than by the quality of the images or annotations.

6.3 Concept-pedia Fine-tuning

To determine the quality and usefulness of Concept-pedia in improving models' ability to detect new concepts, we first filter out all Concept-10k images. We then fine-tune SigLIP on the resulting dataset using a contrastive learning approach. SigLIP is selected in our experiments for two main reasons: first, it is the best-performing model for its size, as shown in Table 5, and second, it is less sensitive to batch size compared to CLIP, making it a more practical choice given our academic budget constraints. Further details about the hyperparameters used can be found in Appendix B. As shown in Table 5, SigLIP models of different sizes show consistent accuracy improvements from fine-tuning on the Concept-pedia training data, demonstrating an enhanced ability to detect a broader range of concepts, with an average accuracy increase of 5.2 percentage points. For instance, SigLIP-SO400m achieves 45.0% accuracy, up from 40.3%, highlighting that Concept-10k remains highly challenging even after fine-tuning. Moreover, we observe consistent performance improvements across semantic categories when models are fine-tuned using Concept-pedia, particularly in categories where models struggle in the zero-shot setting. As shown in Table 7, SigLIP-large shows a notable increase of 7.7 percentage points in the LAW category, while also improving on more common categories such as ANIM and FOOD. Interestingly, model perfor-

Semantic Type	SigLIP -base	SigLIP -base-ft	SigLIP -large	SigLIP -large-ft	SigLIP -SO400m	SigLIP -SO400m-ft
ANIM	59.6	**60.4**	66.9	**67.1**	69.2	**71.9**
BIO	51.1	**55.1**	58.4	**62.0**	60.1	**63.0**
FOOD	50.9	**54.4**	56.3	**62.4**	59.7	**65.3**
CEL	44.1	**52.9**	44.2	**55.9**	**61.8**	55.9
PLANT	40.1	**46.0**	45.3	**53.7**	49.1	**55.3**
MON	40.0	**48.7**	49.3	**61.3**	45.3	**55.3**
GROUP	34.1	**37.8**	40.6	**43.7**	44.3	**48.2**
ARTIFACT	32.4	**39.8**	36.5	**45.2**	41.3	**48.2**
DISEASE	**30.7**	**30.7**	35.3	**36.6**	42.4	**46.0**
PART	28.5	**33.8**	35.4	**38.6**	36.5	**40.8**
STRUCT	27.4	**31.8**	30.5	**37.0**	39.3	**43.7**
ASSET	25.9	**35.8**	27.2	**44.4**	33.3	**46.9**
EVE	24.1	**28.7**	28.2	**34.4**	33.4	**36.1**
LOC	23.7	**26.5**	28.9	**31.7**	32.6	**33.4**
DISCIPLINE	23.2	**28.1**	28.8	**35.0**	32.0	**36.1**
SUPER	23.1	**31.3**	31.3	**34.3**	31.3	**45.5**
SUB	22.3	**28.3**	28.0	**34.5**	33.3	**38.4**
MEDIA	20.6	**25.9**	25.5	**32.0**	28.7	**34.3**
PER	22.3	**23.8**	**29.2**	**29.2**	32.2	**37.6**
PROPERTY	22.1	**26.6**	27.2	**30.9**	**30.3**	**30.3**
LANGUAGE ITEM	17.9	**22.9**	22.9	**27.1**	27.9	**30.0**
CULT	20.0	**21.0**	22.1	**26.7**	26.7	**30.3**
ORG	18.8	**19.8**	22.7	**27.1**	30.5	**35.9**
RELATION	18.7	**21.5**	**25.8**	21.5	**26.3**	23.4
MEASURE	18.6	**21.9**	25.4	**25.6**	24.5	**26.0**
LAW	14.6	**18.5**	16.9	**24.6**	19.2	**23.1**
PSY	13.3	**18.5**	19.2	**20.4**	23.2	**24.1**

Table 7: We report the zero-shot and fine-tuned (ft) SigLIP models across all semantic categories presented in Concept-10k. Results in **bold** indicate the best performance per model size.

mance on ImageNet-1k variations is minimally affected, despite differences in category coverage and distribution. Moreover, as shown in Appendix H, fine-tuning on Concept-pedia for Visual Concept Recognition does not degrade performance on cross-modal retrieval. Finally, in Appendix C, we analyze the impact of incorporating multi-label annotations in Concept-pedia on overall performance.

7 Conclusion

In this work, we present an automatic annotation methodology for creating multimodal datasets grounded in Wikipedia, resulting in Concept-pedia, a semantically-annotated, wide-coverage dataset that we exploit for Visual Concept Recognition. Additionally, we introduce Concept-10k, a manually validated benchmark that covers a considerably broader range of concept categories than existing benchmarks, offering a more challenging evaluation for VLMs. Our experiments show that, in the zero-shot setting, even VLMs that perform well on established benchmarks, such as ImageNet (e.g., SigLIP-SO400m achieving 83.2% acc@1), struggle to generalize to the diverse and challenging concepts included in Concept-10k, with the best-performing model achieving just 40.3% accuracy. Our work highlights that Concept-pedia can improve multimodal model performance on new concept categories, yielding an average accuracy gain of 5.2 percentage points, while maintaining performance on ImageNet. For future work, we aim to expand Concept-pedia to cover a wider range of languages, ensuring greater linguistic diversity, and to incorporate multi-label annotations into our benchmark.

Limitations

Concept-pedia is currently based solely on the English version of Wikipedia, potentially introducing a cultural bias. Expanding to other languages could help mitigate this issue, enrich existing concepts with additional images, and introduce entirely new concepts. Additionally, the final set of concepts included in our benchmark is selected using a frequency-based algorithm, which may result in suboptimal class diversity within categories. Addressing these limitations is left for future work.

Beyond Visual Concept Recognition, Concept-pedia offers a rich set of annotations that include captions, entity annotations, and concept annotations of varying levels of granularity. This could support a range of tasks such as Visual Question Answering (focused on both entities and concepts) and Multimodal Entity Linking. It may also enable the development of semantically-grounded multimodal retrieval approaches that go beyond the current image–caption-centric methods. We leave these directions for future work.

Acknowledgements

We gratefully acknowledge the support of the PNRR MUR project PE0000013-FAIR.

We also gratefully acknowledge the CREATIVE project (CRoss-modal understanding and gEnerATIon of Visual and tExtual content), which is funded by the MUR Progetti di Ricerca di Rilevante Interesse Nazionale programme (PRIN 2020). Karim Ghonim and Alberte Fernández-Castro conducted this work during their enrollment in the Italian National Doctorate in Artificial Intelligence at Sapienza University of Rome. The work of Alberte Fernández-Castro is co-funded by Babelscape.

References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. POLYGLOT-NER: massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, April 30 - May 2, 2015*, pages 586–594. SIAM.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9448–9458.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. 2020. Are we done with ImageNet? *CoRR*, abs/2006.07159.

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020. Fatality killed the cat or: BabelPic, a multimodal dataset for non-concrete concepts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4680–4686. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Simone Conia, Edoardo Barba, Abelardo Carlos Martinez Lorenzo, Pere-Lluís Huguet Cabot, Riccardo Orlando, Luigi Procopio, and Roberto Navigli. 2024. MOSAICo: a multilingual open-text semantically annotated interlinked corpus. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7990–8004. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M. Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. 2023. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems 36: Annual*

Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8320–8329. IEEE.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021b. Natural adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15262–15271. Computer Vision Foundation / IEEE.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7514–7528. Association for Computational Linguistics.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. 2018. The inaturalist species classification and detection dataset. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8769–8778. Computer Vision Foundation / IEEE Computer Society.

Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Giuliano Martinelli, Francesco Molfese, Simone Tedeschi, Alberte Fernández-Castro, and Roberto Navigli. 2024. CNER: Concept and named entity recognition. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8336–8351, Mexico City, Mexico. Association for Computational Linguistics.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of BabelNet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4559–4567. ijcai.org.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artif. Intell.*, 194:151–175.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1946–1958. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR.

Tal Ridnik, Emanuel Ben Baruch, Asaf Noy, and Lihi Zelnik. 2021. ImageNet-21K pretraining for the masses. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Mert Bülent Sariyildiz, Yannis Kalantidis, Diane Larlus, and Karteek Alahari. 2021. Concept generalization in visual representation learning. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9609–9619. IEEE.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems 35:*

Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2443–2449. ACM.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2521–2533. Association for Computational Linguistics.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 219–228. ACL.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE.

A Annotation Guidelines

In this section, we outline the annotation guidelines employed in our study. We assign each annotation task to two expert linguists. Annotators were paid according to the standard salaries for their geographical location. Annotators were tasked with performing three distinct annotation tasks to evaluate the automatic annotations and the associated images. Due to the different requirements of these tasks, we developed a custom application for each in order to facilitate the process.

Annotators received detailed, task-specific guidelines, which we describe below. As a general guideline for all tasks, annotators were instructed to discard corrupt or unclear images, as well as NSFW images and concepts.
Furthermore, for every concept utilized across tasks, we provided the annotators with the definition retrieved from different resources through BabelNet, e.g. Wikidata, WordNet. In case of persistent ambiguity occurring with a given concept, if the human annotator could not determine the correct annotation, with a high degree of confidence, they were required to flag the given sample.

A.1 Task 1: Silver Labels Evaluation

Task Description In this task, annotators are presented with image-caption pairs where hyperlinks are visible and clickable, as they appear in Wikipedia. We supplement the existing hyperlinks in the captions with additional links provided by our link propagation approach, as described in 3.1. Annotators are required to click on the provided hyperlink in the caption and verify whether it corresponds it is associated correctly. Some examples are provided in Figure 5.

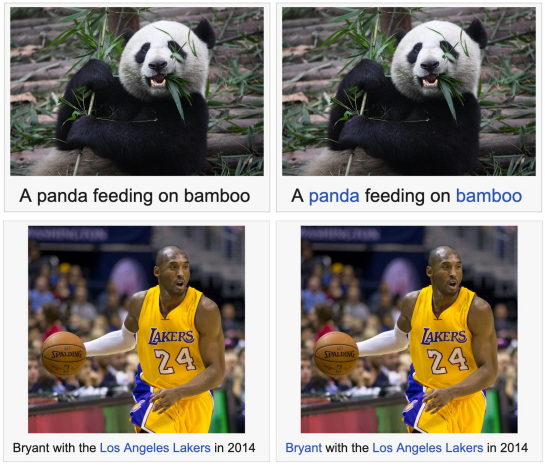


Figure 5: Examples of link propagation annotations in Concept-pedia.

A.2 Task 2: Concept-1k vs. ImageNet-1k

Task Description

1. Check what the main concept *could* be in the image. It needs to have visually clear features.

2. The image can be inside or outside of a building as long as it is visually clear. If an image clearly depicts a specific location or structure, such as the interior of a church with recognizable features like stained glass windows and pews, it should be tagged accordingly.

3. If the concept does not cover the *majority* of the image, then it is not considered the main concept.

4. The image needs to depict most of the concept. For example, an image focused only on an "altar", which is peculiar to a church but not representing a church, can not be tagged as a "church". On the contrary, an image with the inside of a church should be tagged as church because it is the biggest inner part of the church.

5. When an object is the image's primary focus, and a person is present but not the main subject, prioritize tagging the object. For example, if an image shows a person kayaking on a lake, the appropriate tag would be "kayak" rather than "person," since the kayak is the specific object of interest.

6. In the case of two competing concepts, tag the image with both concepts.

7. The maximum number of tags should be two. If there are more than two, tag the image as **uncertain** and provide an explanation in the notes tab.

8. If none of the concepts provided correctly describe the image, leave the sample untagged, and mention it in the notes with the probable main concept in your opinion.

9. Kindly use the notes tab extensively to signal any issues regarding the process or the data.

Examples

- **Example 1: Clear Single Concept**
  - **Image Description:** A close-up photograph of a sunflower in a field, filling the entire frame with its petals and center clearly visible.
  - **Tag:** "sunflower"
  - **Motivation:** The sunflower is the dominant and sole subject of the image, with no other competing elements.
- **Example 2: Specific Object with a Person Present**
  - **Image Description:** An image of a person kayaking on a serene lake, with the kayak being prominently visible and the person facing away from the camera.
  - **Tag:** "kayak"
  - **Motivation:** The kayak is the main focus, and while a person is present, they are not the primary subject. The emphasis is on the activity and the kayak itself.
- **Example 3: Two Competing Concepts**
  - **Image Description:** A dog and a cat sitting side by side on a couch, both looking directly at the camera and occupying equal space in the frame.
  - **Tags:** "dog," "cat"
  - **Motivation:** Both animals are equally prominent and central to the image, making it appropriate to tag both.

A village covered by volcanic ash.

Volcanic ash

The Toyota SORA fuel-cell bus.

Fuel-cell

Figure 6: Filtered negative examples from Concept-10k.

A.3 Task 3: Concept-10k Validation

We followed the instructions outlined in (Krizhevsky and Hinton, 2009), with one key difference: we provided definitions for each concept, extracted from sources such as Wikidata, Wikipedia, and WordNet, to eliminate any potential ambiguity. Annotators were tasked with selecting the images to be included in the final benchmark. This step was essential because, in some cases, the label propagated by our algorithm was correct in the context of the caption but the concept was not visually represented in the image.

In Figure 6, we present two such examples. The propagated label is shown in red, while the caption is shown in blue. Although the label is accurate relative to the caption, it is not suitable as a Visual Concept Recognition label.

Model	Zero-shot	Single-label	Multi-label
SigLIP -base	31.1	33.9	**36.1**
SigLIP -large	36.1	39.4	**41.9**

Table 8: Performance comparison between models fine-tuned on Concept-pedia using single-label and multi-label annotations.

Model	Training Data	in-1k	Con-10k
SigLIP -base	Concept-pedia	75.9	36.1
	ImageNet-21k	78.5	30.2
SigLIP -large	Concept-pedia	79.3	41.9
	ImageNet-21k	81.2	32.6

Table 9: Performance comparison between models fine-tuned on ImageNet-21k and Concept-pedia.

	in-1k	Con-1k	Con-10k
High	10.4%	34.3%	51.2%
Low	9.1%	27.3%	52.3%

Table 10: Performance breakdown on SigLIP's ability to detect negatives for manually annotated subsets of ImageNet-1k and Concept-pedia.

B Training details

In this section, we detail our fine-tuning approach using the Concept-pedia dataset.

We fine-tuned open-sourced SigLIP-base [6], SigLIP-large [7], and SigLIP-SO400m [8] on the complete training split of Concept-pedia. The models are trained with a batch size of 256 for 50,000 steps using AdaFactor (Shazeer and Stern, 2018) as optimizer with a learning rate of $1 \cdot 10^{-6}$.

C Multi-Label Ablation

We ablate the impact of leveraging multi-label annotations in Concept-pedia for fine-tuning. To do so, we construct a single-label version of Concept-pedia and fine-tune both SigLIP-base and SigLIP-large on it, comparing their performance against models trained using the original multi-label annotations.

To generate the single-label version of Concept-pedia, we retain the label assigned by the propagation heuristic with the highest precision, as reported in Table 3. For samples with multiple annotations of the same type, we resolve conflicts by randomly selecting one label.

As shown in Table 8, regardless of model size, while fine-tuning on single-label annotations still improves performance, models benefit even more from multi-label annotations. This confirms the advantage of using multi-label annotations for images when suitable and available.

D Model-based Filtering

It is important to note that we opted against using model-based selection methods (e.g., CLIPScore (Hessel et al., 2021)). There are two main reasons behind this decision. First, such methods inherently introduce bias (Gadre et al., 2023). Second, we qualitatively observe that current models struggle to accurately assess the similarity between the visual and textual representations of new concepts, often resulting in the erroneous exclusion of valid samples.

E Negative Detection

In this experiment, we investigate whether models truly detect the correct concept or merely tend to predict "true" regardless of correctness. This is particularly relevant for SigLIP, which is trained to perform multi-label classification by generating a score for each label and applying a threshold to determine whether the label is predicted as true or false.

We examine this from two perspectives: first, whether the models are biased towards producing high scores for ImageNet labels, even when the predictions are incorrect, and second, whether they are biased towards predicting incorrect labels (as per our annotations) for ImageNet images.

As part of our manual annotation process, we are able to detect a significant number of negative samples, i.e., where the label is not visually represented in the image. With this, using SigLIP, we predict a probability for this concept (which we know is incorrect) being correct given an image. In this case, we want to check if the model is able to correctly determine that this concept is not visually represented in the image. We show that for ImageNet, for example, the model consistently predicts high probabilities for incorrect labels. In-

[6] google/siglip-base-patch16-256
[7] google/siglip-large-patch16-256
[8] google/siglip-so400m-patch14-384

Model	ImageNet-1k	Concept-10k
CLIP-base	68.3	22.0
SigLIP-base	76.2	31.1
CLIP-large	75.5	26.7
SigLIP-large	80.5	36.1
SigLIP-so400m	83.2	40.3
FLAVA	36.2	4.8
LLava-1.5	22.8	2.8
InstructBLIP	14.6	2.6

Table 11: Comparison of zero-shot model performance (top-1 accuracy) on ImageNet-1k and Concept-10k.

terestingly, however, for our split, the model still struggled to perform this task correctly, but not as badly as with ImageNet. This suggests that models overfit on ImageNet concepts, as well as ImageNet annotations. Moreover, this bias is less pronounced in the Concept-10k classes, as shown in Table 10, which provides yet further confirmation of the bias towards ImageNet classes.

F Comparison with ImageNet-21k

We compare ImageNet-21k, another large-scale resource, to Concept-pedia in terms of category coverage and distribution, as illustrated in Figure 8. Furthermore, we evaluate its effectiveness as a training resource for recognizing the newly introduced concepts in Concept-10k. As shown in Table 9, while fine-tuning on ImageNet-21k improves model performance on ImageNet-1k, it leads to a deterioration in performance on Concept-pedia. We attribute this to two factors: first, ImageNet-21k does not cover all CNER categories present in Concept-pedia and subsequently in Concept-10k; second, similarly to ImageNet-1k, it is heavily skewed toward a subset of categories, such as Artifact and Plant.

G MLLM and VLM Evaluation

Additionally, we evaluate the visual concept recognition capabilities of Multimodal Large Language Models (MLLMs) such as LLava-1.5 (Liu et al., 2024) and InstructBLIP (Dai et al., 2023), both of which use a 7-billion-parameter LLM. Given that Concept-10k comprises approximately 10,000 classes and these models support a maximum context length of 4K tokens, it is infeasible to include all class names directly in the prompt. To address this, inspired by Cao et al. (2021), we adopt a constrained decoding approach that restricts the model's output space to the set of candidate classes.

Model	Fine-tuning	I → T	T → I
CLIP-base	✗	56.3	36.5
SigLIP-base	✗	64.4	47.2
SigLIP-base	✓	63.7	46.7
CLIP-large	✗	57.9	37.1
SigLIP-large	✗	69.5	51.1
SigLIP-large	✓	68.3	50.2
SigLIP-SO400	✗	70.2	52.0
SigLIP-SO400	✓	69.1	50.9

Table 12: Cross-modal (Text-to-Image and Image-to-Text) retrieval performance on the MSCOCO test set. We report model performance in both the zero-shot setting and after fine-tuning on Concept-pedia.

H Cross-modal Retrieval Performance

The results presented in Table 12 show that fine-tuning on Concept-pedia does not considerably degrade the performance of VLMs, even on retrieval tasks, which are not the focus of our study or dataset. The observed drop in retrieval performance is minimal (only 0.9 points on average), and our fine-tuned SigLIP models still outperform similarly sized models, such as CLIP.

I Benchmark Category Coverage

In Table 13, we report the concept and semantic category coverage of commonly used benchmarks, including ImageNet-1k, iNaturalist, ObjectNet, and the 27 datasets used in (Radford et al., 2021), and compare them with Concept-10k. We show that Concept-10k not only includes significantly more concepts than the other benchmarks, but also covers a broader range of semantic categories. Notably, even when combining all other benchmarks, the total number of unique semantic categories amounts to 18, which Concept-10k surpasses with 28 categories.

J Comparison with ImageNet-CoG

As previously mentioned, ImageNet-CoG was created to provide concepts that are not included in ImageNet. In Figures 9 and 10 we show how it compares against Concept-10k in terms of CNER category coverage and population. Even though ImageNet-CoG includes more concept categories

than ImageNet-1k, we can clearly see that it suffers from a similarly skewed category distribution toward Artifacts.

K CNER categories

In Table 14, we present Concept-pedia instances of concepts from different CNER categories.

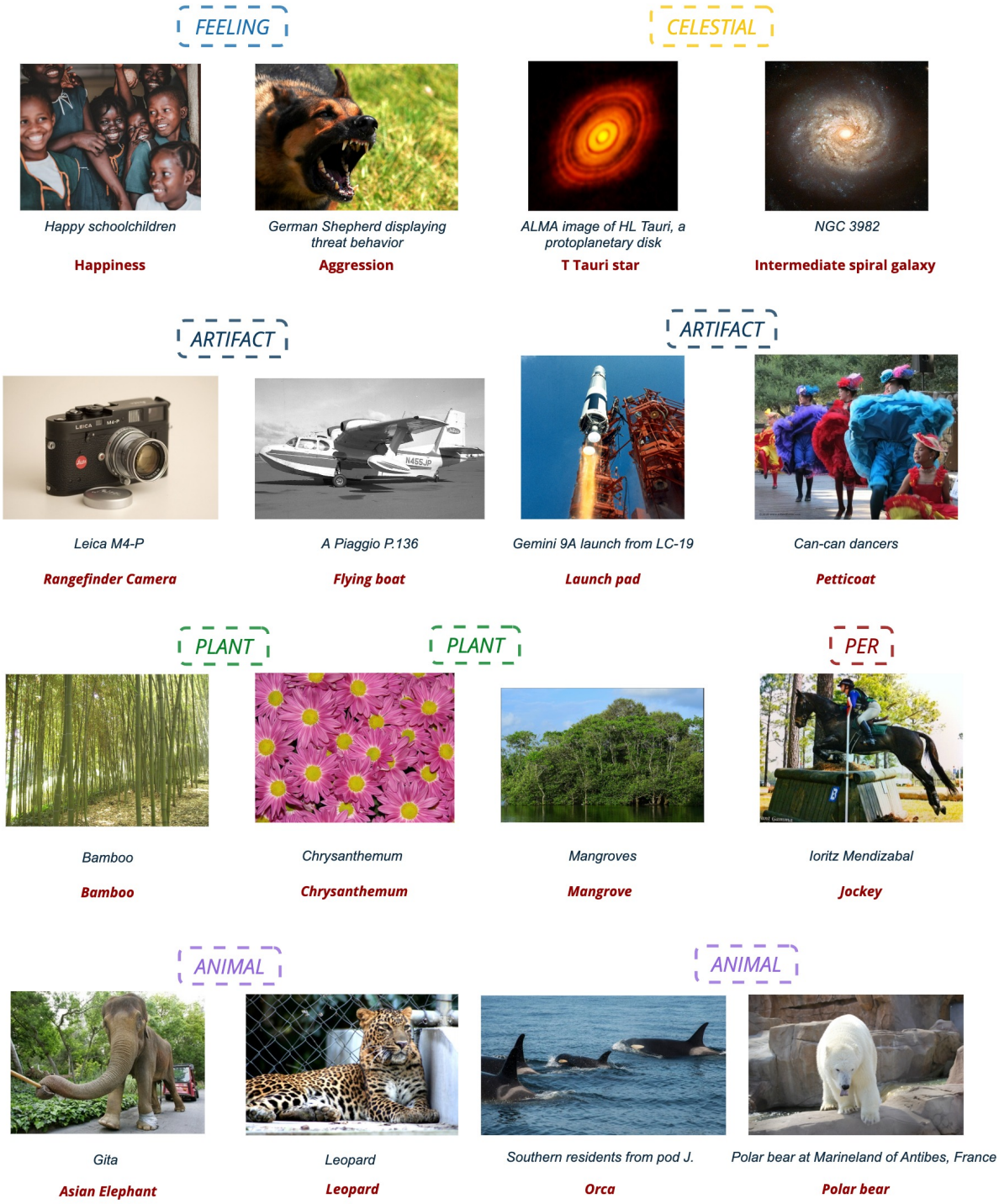


Figure 7: Examples from Concept-pedia. We provide the caption (in blue) and the concept (in red) for each image.

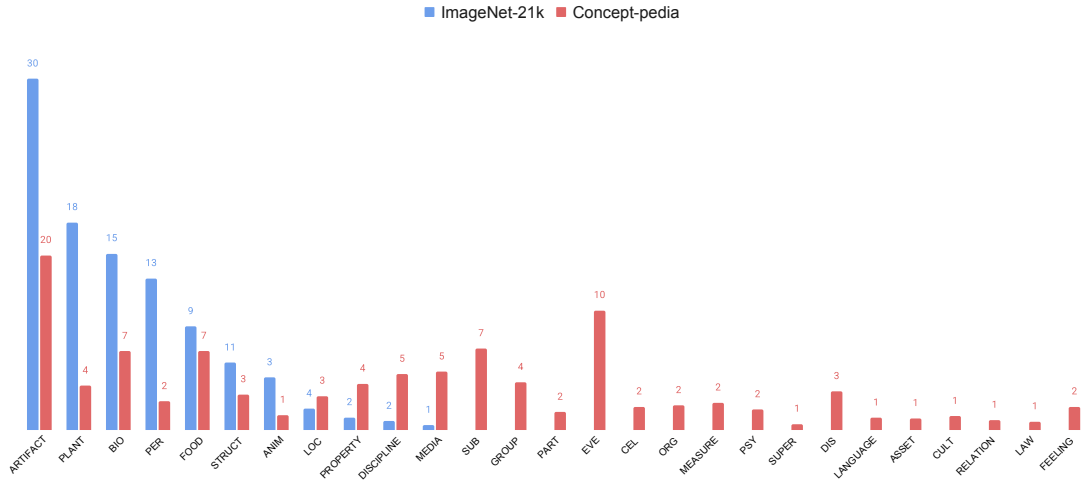


Figure 8: Comparison of CNER category population between Concept-pedia and ImageNet-21k.

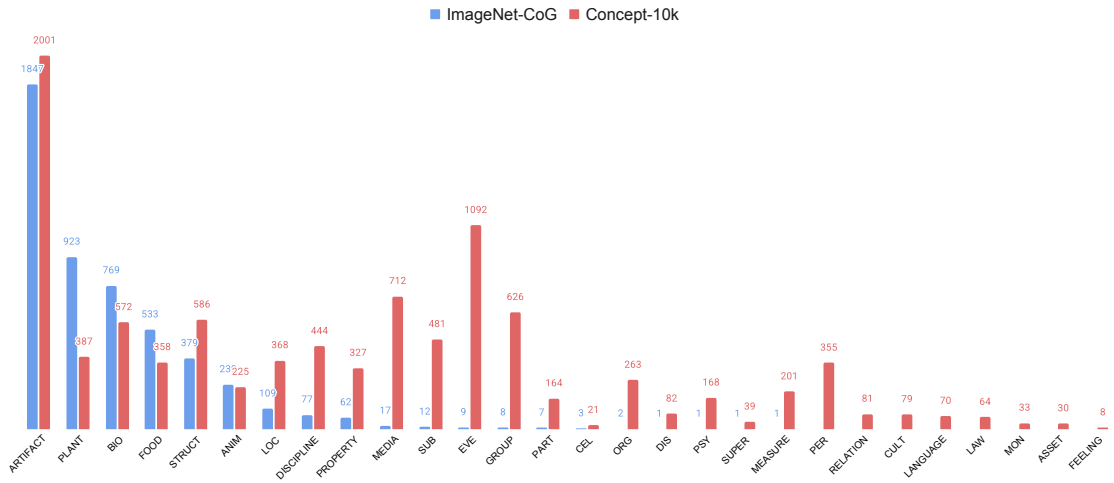


Figure 9: Comparison of number of concepts per CNER category between Concept-10k and ImageNet-CoG.

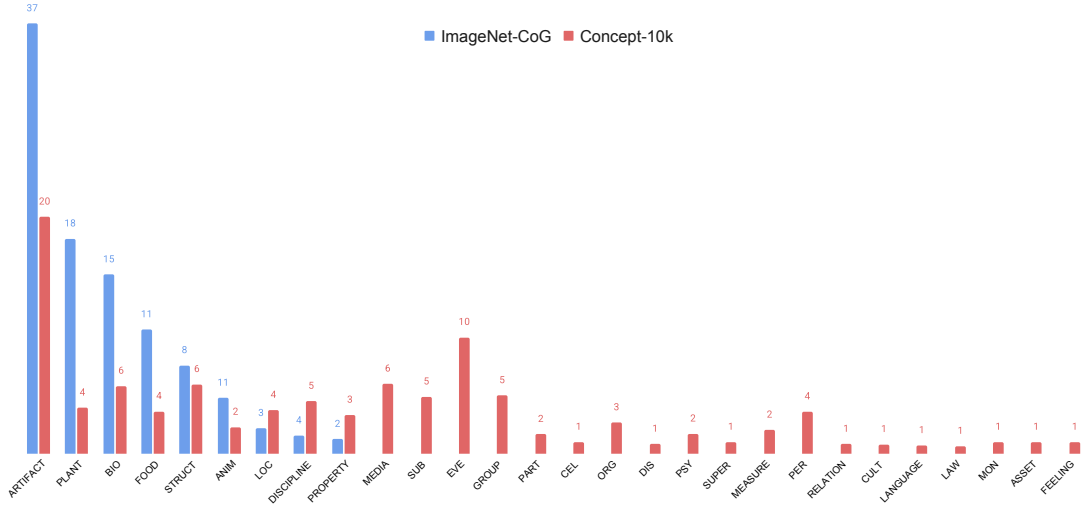


Figure 10: Comparison of CNER category distribution (%) between Concept-10k and ImageNet-CoG.

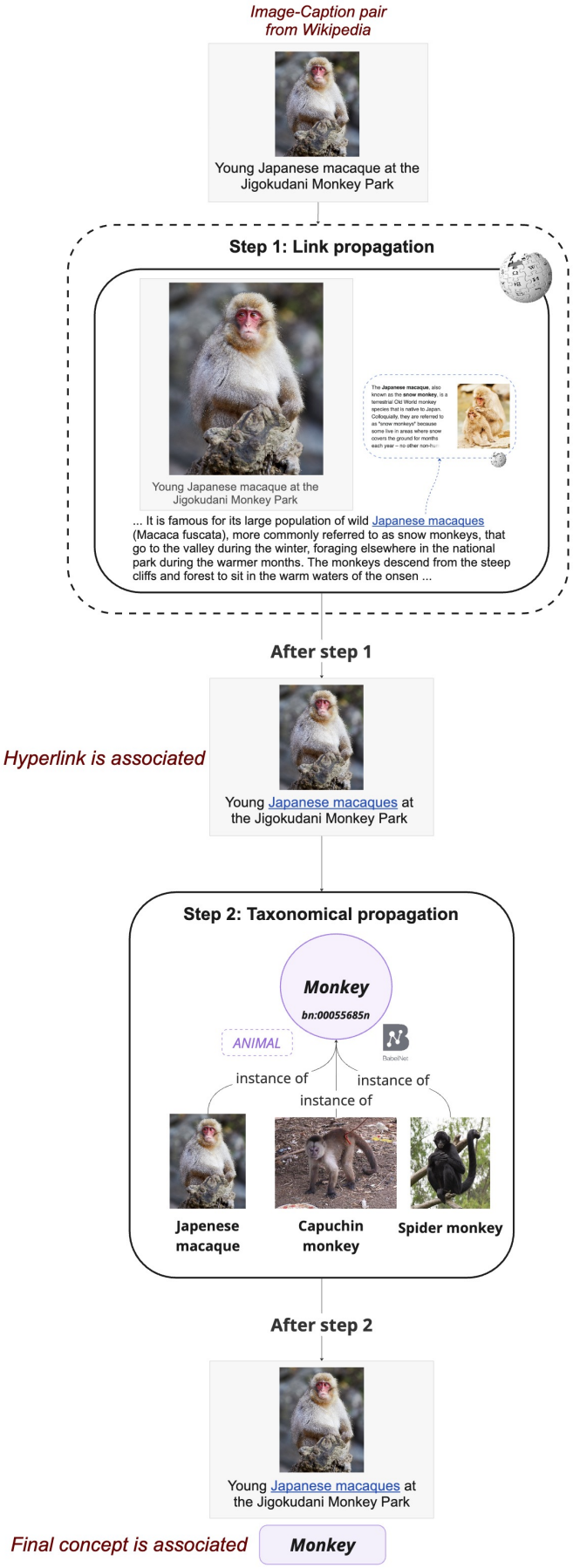


Figure 11: Concept-pedia construction process.

	Concepts (Classes)	Semantic Categories	CNER Categories
Food-101	102	4	FOOD, BIO, PLANT, SUBSTANCE
CIFAR-10	10	2	ANIMAL, ARTIFACT
Birdsnap	500	1	ANIMAL
SUN397	397	3	LOC, STRUCT, ORG
Stanford Cars	196	1	ARTIFACT
FGVC Aircraft	100	1	ARTIFACT
Pascal VOC 2007 Classification	20	4	ANIMAL, ARTIFACT, PER, PLANT
Describable Textures	47	1	PROPERTY
Oxford-IIIT Pets	37	1	ANIMAL
Caltech-101	102	4	ANIMAL, ARTIFACT, PLANT, PER
Oxford Flowers 102	102	1	PLANT
MNIST	10	1	MEASURE
Facial Emotion Recognition 2013	8	1	FEELING
STL-10	10	2	ANIMAL, ARTIFACT
EuroSAT	10	3	PLANT, STRUCT, LOC
GTSRB	43	3	ARTIFACT, MEASURE, LAW
PatchCamelyon	2	2	BIO, DISEASE
UCF101	101	3	PER, DISCIPLINE, CULTURE
Kinetics700	700	3	PER, DISCIPLINE, CULTURE
CLEVR Counts	8	2	MEASURE, ARTIFACT
Rendered SST2	2	2	PER, CULTURE
Hateful Memes	2	4	PER, CULTURE, ORG, LAW
ObjectNet	312	1	ARTIFACT
iNaturalist	8142	3	ANIMAL, PLANT, BIO
KITTI	4	5	ARTIFACT, PER, STRUCT, LOC, MEASURE
RESISC45	45	6	LOC, STRUCT, ARTIFACT, PLANT, ORG, CULTURE
Country211	211	6	LOC, STRUCT, CULTURE, PLANT, ANIMAL, PER
CIFAR-100	100	7	ANIMAL, PLANT, FOOD, ARTIFACT, STRUCT, LOC, PER
ImageNet-1k	1000	11	ANIMAL,ARTIFACT, BIO, STRUCT, FOOD, LOC, MEDIA, PER, PLANT, SUB, PROPERTY
Total (combined)	-	18	
Concept-10k	**9837**	**28**	

Table 13: Comparison of concept and semantic category coverage between Concept-10K and existing benchmarks. We report the number of unique concepts and semantic categories for each dataset, as well as the total number of distinct semantic categories covered across all the benchmarks (**Total (combined)**), including those in the CLIP Benchmark. Concept-10k covers all CNER classes except *DateTime*.

Category	Examples
ANIMAL	*Lion, American bison, Orca, Leopard, Reindeer*
ARTIFACT	*Racing car, Water tower, Monoplane, Wind turbine, Arcade video game*
ASSET	*Stock, Debit card, Mortgage loan, Equity, Health insurance*
BIOLOGY	*Proton, Protein dimer, Collagen*
CELESTIAL	*Pulsar, Star, Ring galaxy*
CULTURE	*Humanism, Sufism, Islam, Pacifism*
DISEASE	*Covid-19, Tuberculosis, Malaria, Smallpox, Ebola*
DISCIPLINE	*Entrepreneurship, College football, Mixed martial arts, Professional boxing, Contemporary art*
EVENT	*Freestyle swimming, Climate change, Deforestation*
FEELING	*Love, Nirvana, Hatred, Pride, Sympathy*
FOOD	*Chocolate bar, Honey, Mussels, Sushi, Ramen*
GROUP	*Brigade, Basketball team, Opera company*
LANGUAGE	*Ambigram, Synonym, Vowel*
LAW	*International law, Regulation, Independence*
LOC	*Vineyards, Playground, Bus stop*
MEASURE	*Kilometre, Latitude, Mile*
MEDIA	*Music video, Self-portrait, Visual novel*
MONEY	*Euro, Dirham, Dollar,*
ORG	*Music industry, Democratic Party, Commercial enterprise*
PART	*Liver, Stomach, Wrist, Retina*
PER	*Gymnast, Jockey, Ski jumper, Cyclist*
PLANT	*Bamboo, Sugarcane, Bonsai, Wheat*
PROPERTY	*Linear, Dimension, Shape*
PSYCH	*Sweetness, Cognition, Attention, Necessity*
RELATION	*approximation, Comparison, Relatedness, Distance (graph theory), Function*
STRUCT	*Supermarket, Marina, Hospital*
SUBSTANCE	*Aluminium, Silver, Sodium*
SUPER	*Unicorn, Zombie, Goblin*

Table 14: Frequent concepts from `Concept-pedia` categorized by CNER groups.