BIOMEDICA: An Open Biomedical Image-Caption Archive, Dataset, and Vision-Language Models Derived from Scientific Literature

Supplementary Material

Supplementary Material Table of Content

10. Acknowledgments

This research was supported by NIH grants (NIH#P30AG066515 to JJN), the Chan Zuckerberg Initiative Neurodegeneration Challenge Pairs Pilot Project to SYL (2020-221724, 5022), the Wu Tsai Knight Initiative Innovation grant (#KIG 102) to SYL, Hoffman-Yee Research Grant to SYL, the Arc Institute Graduate Fellowship to AL, the Stanford Data Science Graduate Research Fellowship MW, and the Quad Fellowship to JB. SYL is a Chan Zuckerberg Biohub – San Francisco Investigator. We thank Daniel van Strien, Matthew Carrigan, and Omar Sanseviero from Hugging Face for their invaluable assistance with data upload and design planning on the Hugging Face platform.

11. Dataset Description

BIOMEDICA dataset contains a total of 24,076,288 image-caption pairs from 5,050,473 scientific articles (full article-text for the 6M is additionally provided). Each image-caption pairs is assigned article metadata and additional annotations. Table S1 shows descriptive statistics for BIOMEDICA dataset. Table S2 shows all the data and metadata fields provided in the dataset. Table S15 shows example image-caption pairs. Figure S1 shows a cohort diagram, summarizing the materialization of BIOMEDICA dataset.

Table S1. BIOMEDICA dataset statistics: Counts of articles, image-caption pairs, metadata, human annotators and specialties.

Aspect	Count
Articles	6,042,494
Articles with Images	5,050,473
Images	24,076,288
Captions	24,076,288
Figure Reference	30,711,542
Metadata	22
Global Concepts	13
Local Concepts	170
Clinical Annotators	2
Scientist Annotators	6

12. Dataset Statistics

We provide additional statistics for our dataset. The reference count, representing the number of articles citing a given article, ranges from 0 to 3346, with a median of 37 (IQR: 35). Articles with no references account for 120,870 entries. A total of 263,836,608 references is found across the entire dataset.

For MeSH terms, the dataset includes 29,859 unique terms. MeSH term counts per article range from 0 to 44, with a median of 0 (IQR: 10) and a total of 32,896,861 terms. Notably, 2,991,141 articles are annotated with MeSH terms, emphasizing the dataset's depth in biomedical categorization.

Table S10 summarizes the top 50 most frequent MeSH terms in the dataset.

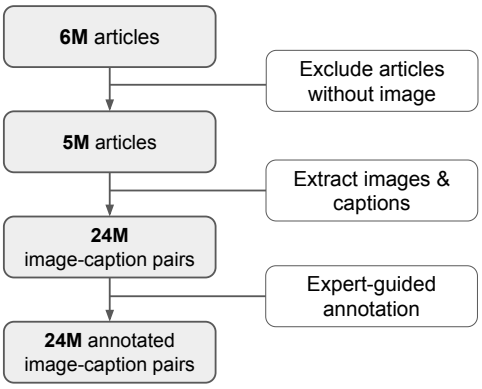


Figure S1. BIOMEDICA cohort diagram: selection criteria for the construction of relevant image-caption pairs.

Concept	Provenance
Image Data	
Image Key	Generated
Image File	File List
Caption	nXML File
Image Metadata	
Image Cluster ID	Generated
Image Hash	Generated
Image File Name	nXML File
Image Set	nXML File
Image Context	nXML File
Image Annotations	
Image Panel Type	Generated
Image Panel Subtype	Generated
Image Content Primary Label	Generated
Image Content Secondary Label	Generated
Article Metadata	
Article Keywords	nXML File
Article Category	nXML File
Article Title	nXML File
Article Abstract	nXML File
Article full text	nXML File
Article Publication Date	Entrez API
Article MeSH Terms	Entrez API
Article Journal	File List
Article PMID	File List
Article Citation	File List
Article License	File List
List of PMIDs citing article	Entrez API
Count of PMIDs citing article	Entrez API
Image Embeddings	
Image DINO-V2 Features	Generated
Image PMC-CLIP Features	Generated

Table S2. List of image data (n=3), image metadata (n=5), image annotations (n=4), article metadata (n=13), and image embeddings (n=2) provided in BIOMEDICA dataset, alongside source.

13. Data Curation Process

Besides releasing the code to make BIOMEDICA dataset fully reproducible, we provide additional descriptions and design choices for the dataset curation process. To increase efficiency and enable scaling, everything step in the data curation process is parallelized, unless specified otherwise.

13.1. PMC OA Dataset Description

The PubMed Central (PMC) Open Access (OA) Subset is a publicly accessible collection of full-text scholarly articles hosted by the National Center for Biotechnology Information (NCBI). This subset contains articles that have been made available under various open-access licenses. It covers a wide range of disciplines within biomedical and life sciences, providing rich content that includes research articles, reviews, case reports, and clinical trials. As of 2024, over six million articles are available, with tens of thousands of new articles added annually, reflecting the continuous contributions of researchers worldwide.

13.2. Data Extraction

The remote paths to the compressed files (tar.gz) containing each article's media files are listed in a CSV file, referred to as the file list. The file list provides the structure for locating and accessing the content, ensuring that all relevant files can be traced and downloaded accurately. The file list includes the following columns: *File*, *Citation*, *Accession_ID*, *Date*, *PMID*, and *License*. The server stores files with randomized paths to optimize storage; therefore, the absolute file path provided in the file list is required to retrieve the media files for a specific article. We connect to the server using the Python package `ftplib`:

```
ftp = FTP("ftp.ncbi.nlm.nih.gov")
ftp.retrbinary("RETR <remote_file_path>",
open("<local_file_path>", "wb").write)
```

Bulk downloads are available only for text files, necessitating individual retrieval of associated media files.

The server enforces a rate limit of three requests per second per IP address, with varying download speeds requiring precise scheduling to prevent disruptions. FTP connections may disconnect intermittently, requiring a robust retry mechanism with short delays to maintain data integrity. To conserve storage, only necessary files (nxml and jpg) are kept, while other files (e.g., pdf, docx, xlsx, mp4) are discarded if present.

After downloading and uncompressing all files, we create JSON files to store data extracted from the raw nxml and image files. These JSON files contain a list of dictionaries, where each dictionary holds the data for a single unique article. The `figure_set` is a list of dictionaries, where each dictionary contains the figure's `PMID`, volume number, image file, caption, and context. This structure is shown in Figure S2.

Entrez is a search and retrieval system from NCBI that we use to collect additional metadata, including publication details and MeSH terms, which are not available in the file list or raw nxml files. The Entrez API supports batch queries with up to 200 PMIDs per request:

Text Statistics	Min	Max	Median	IQR	Total
Caption Token Length	1	12389	64	134	2.84×10^9
Caption Character Length	1	25539	246	498	1.05×10^{10}
Figure Reference Token Length	17	699117	338	299	1.28×10^{10}
Figure Reference Character Length	39	1395195	1323	1162	4.82×10^{10}
Full Text Token Length	20	2449711	10306	8831	6.65×10^{10}
Full Text Character Length	70	8701762	39072	32080	2.47×10^{11}
Image Statistics					
Image Width (pixels)	1	52490	709	150	-
Image Height (pixels)	1	65081	476	406	-
Image Area (pixels²)	1	1.95×10^9	334400	307272	-

Table S3. Overview of dataset statistics, detailing text token and character lengths, and image dimensions. For text statistics, tokens are generated using the BPE tokenizer from the `tiktoken` library

```
from Bio import Entrez
Entrez.email = "<your_email>"
handle = Entrez.efetch(
    db = "pubmed",
    id = "<comma_separated_PMIDs>",
    retmode = "xml")
```

Each JSON file is limited to a maximum of 200 articles to comply with the Entrez API batch limit and to keep file sizes manageable for processing.

13.3. Dataset Serialization

As shown in Figure S2 A, the retrieved data is a collection of articles with full metadata and a figure set (containing multiple images and captions). This structure can natively be serialized by article; however, it requires an extra iteration within the figure set. Instead, we decide to serialize the dataset by figure, such that a row (figure within a figure set) becomes a unique image-caption pair rather than an article. This implies that each image-caption includes all the corresponding metadata. Note that this comes with the disadvantage of repeated entries for images belonging to the same figure set (e.g. within the same publication). In other words, if two different images come from the same article figure set, then all the metadata and nxml will be repeated twice. Figure S2 shows the data structure before (A) and after (B) serialization.

PMC-OA Subsets were serialized in parallel. Table S4 shows the total serialization run time. Notably, we can serialize the entirety of PMC-OA within a single day.

13.4. Tokenized Caption Distribution

Note that the histogram in Figure S5 shows a long right tail in the distribution of caption token lengths, with many captions exceeding the CLIP model's context length limit. **Need to provide numbers, mean min, max, median etc**

This issue is even more pronounced for figure reference.

Subset	Serialization Time (Hrs)
Commercial	23:50:57
NonCommercial	7:36:35
Other	1:36:42
Total	33:04:14

Table S4. Total Serialization time by PMC-OA subset

14. Concept Labeling: Additional Details

We developed an AI-assisted pipeline to categorize similar concepts within PMC-OA, as described in section Section 3.2. In summary, this process involves four stages: First, we define and organize similar images in clusters applying unsupervised clustering on image content, second, a group of 2 clinicians and 1 scientist use these clusters and taxonomies to create a hierarchical taxonomy for PMC-OA (see Figure S6), then a group of 2 clinicians and 6 scientists use this taxonomy to annotate each cluster. Lastly, metadata is propagated to each cluster instance.

In this section, we provide additional details and experiments for each design choice.

14.1. Dimensionality Reduction

We used DINOv2 (ViT-L/14 distilled) to generate a 1024-dimensional vector for each image in PMC-OA . However, directly clustering such high-dimensional data can lead to poor performance due to the "curse of dimensionality", where the increased sparsity in high-dimensional spaces reduces the reliability of distance-based measures like those used in K-means clustering.

To address this, we applied PCA (Principal Component Analysis) to reduce the dimensionality of the embeddings. A scree plot analysis was performed to determine the minimum number of principal components required to retain 99% of the data variance. As shown in Table S3, 25 principal components were sufficient to achieve this threshold.

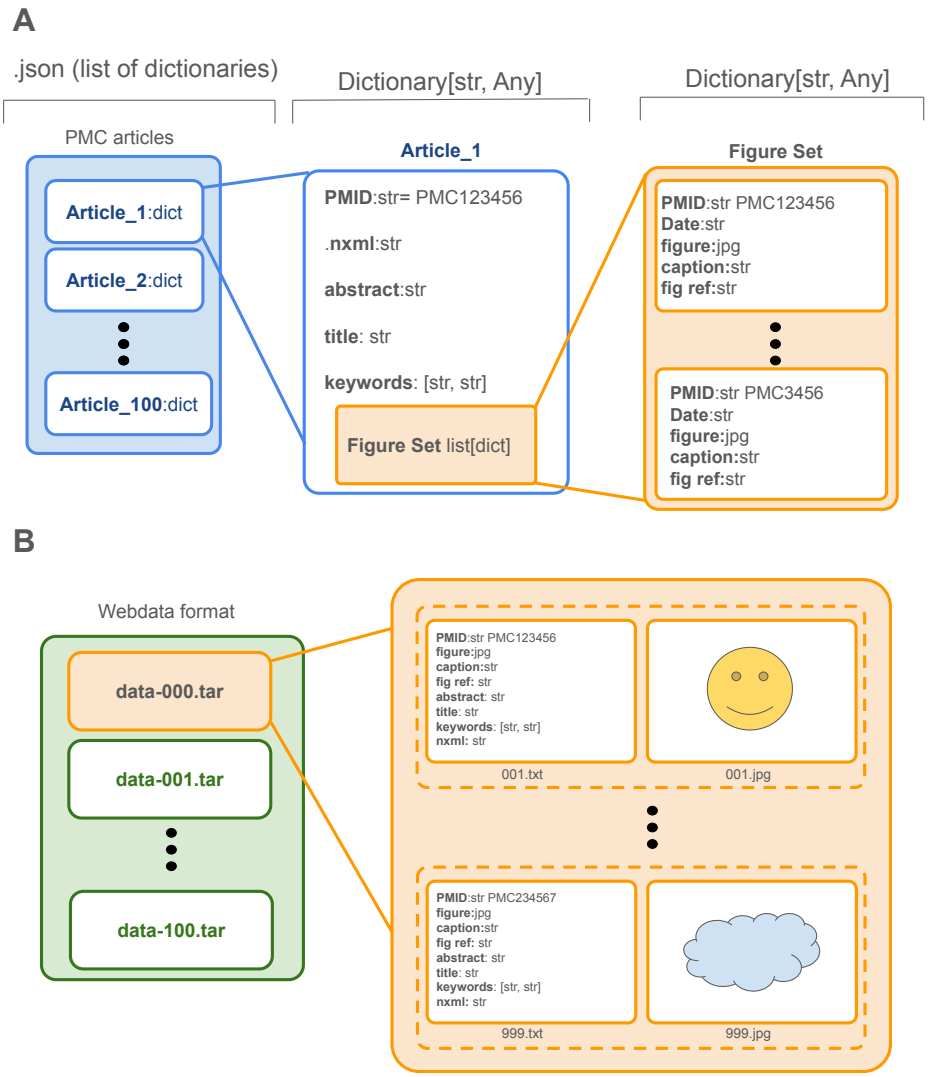


Figure S2. A) Diagram illustrating the structure of a JSON file containing a list of dictionaries representing PMC articles. Each article dictionary includes metadata fields such as PMID, nXML path, abstract, title, keywords, and a nested figure set. The figure set is a list of dictionaries, where each dictionary contains the figure's PMID, volume number, image file, caption, and context. B) Diagram illustrating the WebDataset format, where data is stored across multiple .tar archives (e.g., data-000.tar, data-001.tar). Each archive contains paired text and image files representing individual records.

Consequently, we selected PCA (n=25) to transform the data before applying K-means clustering.

14.2. Over-Clustering

We opt to over-cluster using KMeans with (K=2000), as this approach allows us to thoroughly explore the PMC-OA dataset. By focusing on fine-grained patterns and subgroups, this strategy ensures a detailed analysis, particularly when the optimal number of clusters remains uncertain. This number was also selected to achieve an effective annotation time of at most **48 hours per annotator**.

To do this at scale, we extract DINO-v2 features in parallel with NVIDIA Triton Inference. PCA and K-means are not parallelized. However, due to the simplicity of these approaches, they scaled to 24M pairs. Finally, cluster labels are stored along with image-uid. This is later used to propagate metadata to individual instances.

14.3. Online Cluster Annotations Form

Via a form hosted on Google Forms we asked two practicing clinicians and 5 scientists to annotate each image cluster by answering the following questions:

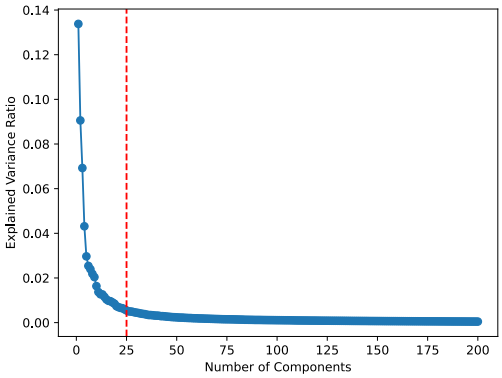


Figure S3. DinoV2 features Scree plot

1. *Are the majority images within this cluster single panel or multiple panel?* Options:
   Single Panels
   Multiple panels with non-biomedical imaging
   Multiple panels with biomedical imaging and plots
   Multiple panels with biomedical imaging and assays
2. *Think of the main characteristics the MAJORITY of images in this cluster share in common. Given these characteristics, what is the most likely global class for this group?* Options:
   Clinical Imaging
   Microscopy
   Immuno Assays
   Illustrative Diagrams
   Hand Drawn and Screen Based Visuals
   Tables
   Plots and Charts
   PCR
   Graph and Network
   Scientific Formulae and Equations
   Chemical Structures
   Maps
   Tools and Materials
3. *Given these characteristics, what is the most likely local class for this group? (remember to use the hierarchical taxonomy provided)*

Due to the quantity of high-resolution images, is not possible to compile all clusters together, thus 20 Google Forms, each containing 100 clusters were automatically created using Google Apps Script. Within each form, we provide the cluster image on top of each question to facilitate annotations. Annotators are calibrated by doing a practice run on 20 examples (these annotations are not added to the analysis). Annotators were given two weeks **336 hours**) to fill all assigned forms. Annotators could only be assigned up to 1000/2000 clusters (10 forms). Lastly, forms were overlapped with a maximum of three annotators, meaning that annotations could have at most three labelers.

14.4. Taxonomy Curators Statistics

Table S11 shows statistics for taxonomy curators (denoted with *). Curators have a minimum of 3 years of experience, with a median of 11 years and a maximum of 14 years. Both clinical curators have an additional PhD. All curators have wet-lab experience.

14.5. Cluster Annotator Statistics

Table S11 shows statistics for cluster annotators. Annotators have a minimum of 3 years of experience, with a median of 4 years and a maximum of 14 years. On average, annotators spent 17.66 hours to finish all assigned forms. Collectively, annotators spent 103 hours annotating all 2000 clusters.

For our statistics, experience in the biomedical domain is measured from the time a person starts their graduate studies (such as a master's or Ph.D.) and continues to accumulate as they work in the field. For this work, undergraduate studies do not contribute to the experience count, meaning that time spent pursuing a Bachelor's degree or any other undergraduate-level studies is excluded from the total experience calculation, even when it's biomedical related.

14.6. Dataset Taxonomy

The full delivered taxonomy is shown in Figure S6. Examples of topics with they corresponding image cluster are shown in Table S14. In total the taxonomy spans 13 global topics enumerated below:

- Clinical Imaging
- Microscopy
- Plots and Charts
- Immuno Assays
- Illustrative Diagrams
- Scientific Formulae and Equations
- Tables
- Hand Drawn and Screen Based Visuals
- Graph and Network
- Chemical Structures
- Maps
- Tools and Materials
- PCR (Polymerase Chain Reaction)

Notice that in contrast to the derived taxonomy, "Clinical and Scientific Image Data" is not included as a topic. Instead we itemize this concept by its children (clinical imaging and microscopy). This design choice facilitates labeling. For the rest of this work, clinical imaging and microscopy do not share any parent.

14.7. Label Assignment and Propagation

After labeling, each cluster has at most three expert-level annotations pertaining to panel type, global taxonomy, and local taxonomy. To resolve annotations, we follow a three-step pipeline:

1. Text Prepocessing Since panel-type and global taxonomy are multiple-choice questions, no further reprocessing is necessary. Local taxonomy, however, consists of open questions (users are asked to follow the taxonomy as shown in Figure S6). To this end, answers in this section are preprocessed by being lowercased, and white spaces and dashes are deleted.
2. Majorty vote After preprocessing the final label for each field is resolved by taking the consensus of annotations by using the majority vote.
3. Label Propagation The metadata for the labeled cluster is then propagated to each image within that cluster. This is accomplished by retrieving the cached cluster label and image-uid (see Over-Clustering section) along with the corresponding resolved clustered metadata and adding this information to each item in within the serliaized data.

Concept	Min	Max	Mean	Median	IQR
Panel	0.0	66.66	13.54	0.0	33.33
Global	0.0	66.66	8.39	0.0	0.0
Local	0.0	75.0	19.91	0.0	50.00

Table S5. Statistical overview of inter-annotator disagreement (lower is better).

14.8. Inter-annotator Disagreement

Table S5 shows a statistical overview of inter-annotator disagreement. The median disagreement for all concepts is 0.0, while the maximum disagreement (present in local taxonomy) is 19.91%, highlighting a high inter-annotator agreement greater than 80% for all concepts. 'Easier' concepts, such as global taxonomy, show a low mean disagreement of 8.39%. Figure S4 shows a histogram summarizing these findings.

15. Data Upload

Data is uploaded to HuggingFace using "upload_large_folder" with 40 workers. This upload mechanism enables parallelism but relies on an already structured dataset.

```
from huggingface_hub import HfApi

api = HfApi(token=HF_TOKEN)
api.upload_large_folder(
    repo_type    = "dataset",
    repo_id      = REPO_ID,
    folder_path  = LOCAL_PATH,
    num_workers  = 40)
```

16. Model Training

16.1. Base Model Selection

In biomedical model development, it is common to skip ablations when selecting a strong base model for continual pretraining, often defaulting to models that yield state-of-the-art results in general domains. To this end, we use the validation sets of four random datasets to identify a robust base model for further fine-tuning (see Table S6). Our analysis reveals that, generally, ViT models trained on datasets such as CommonPool, Liabo2B, and WebLi achieve the strongest performance in the selected biomedical domains, while previously favored models like CoCa show the weakest results. Based on these findings, we select CLIP, ViT-L-14 Base as our model for subsequent experiments.

16.2. Modeling Hyperparameters

We detail the key configurations for training our model as follows:

- **Batch Size and Accumulation**: We use a batch size of 1024 per GPU on 4 GPUs with a batch accumulation frequency of 2, yielding an effective batch size of 8192.
- **Learning Rate**: We perform a sweep of learning rates from $1e-6$ to $1e-8$. We select the biggest learning rate with stable training curve. To this end , We use a learning rate of 1e-6 and a warmup phase of 1000 steps.
- **Optimizer**: We use the Adam optimizer with parameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.95$, and a weight decay of 0.2.
- **Floating-Point Precision**: All computations are performed with 32-bit floating-point precision.

Table S7 summarizes all parameters used during training.

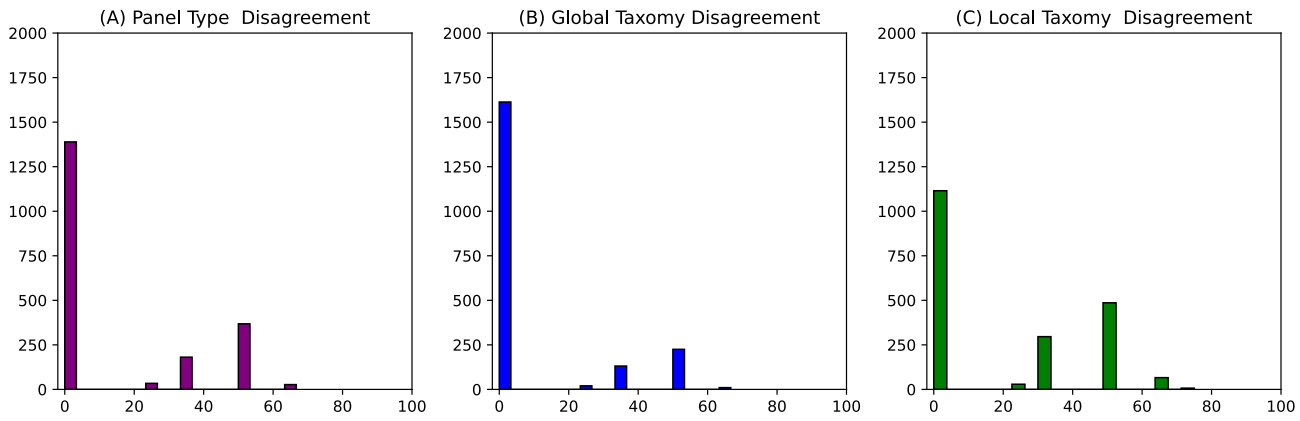


Figure S4. Inter-annotator disagreement (lower is better).

Model	Pretraining Dataset	Mean
ViT-L-14	commonpool	38.465
ViT-B-32	laion2b	37.892
ViT-SO400M-14-SigLIP	webli	35.010
ViT-L-14	laion2b	33.951
convnext-larged-320	laion2b	33.773
ViT-L-14	datacomp	33.148
EVA02-B-16	merged2b s8b b131k	32.995
RN50-quickgelu	openai	32.736
ViT-B-16-SigLIP-384	webli	32.482
ViT-B-16-SigLIP	webli	32.399
ViT-B-16-SigLIP-256	webli	31.762
EVA02-L-14	merged2b s4b b131k	31.680
ViT-B-16-SigLIP-512	webli	31.319
ViT-B-32	commonpool	30.614
coca-ViT-B-32	laion2b	29.133
ViT-B-32	datacomp	29.077
convnext-B	laion2b	28.701
ViT-B-16	laion2b	25.786
coca-ViT-L-14	laion2b	25.212

Table S6. Average Accuracy of Base Models (with Corresponding Pretraining Datasets) on 5 biomedical classification tasks

17. Evaluation

17.1. Closed VQA Benchmark

17.1.1. Closed VQA Formulation

A total of 39 existing classification tasks are formulated as multiple-choice visual question answering. The following subsection provides additional details for evaluation reformulation.

We first collect the test set of each dataset, yielding image-label pairs.

$$D_i = \{(x_i^1, l_i^1), \ldots, (x_i^{N_i}, l_i^{N_i})\}$$

where N_i corresponds to the total number of samples in the test subset of the i-th dataset.

For each image-label pair, the label l_i^j corresponds to one of M_i possible classes defined for the i-th dataset..

$$l_i^j \in C_i,$$

where $C_i = \{c_1, c_2, \ldots, c_{M_i}\}$, $|C_i| = M_i$

To convert each classification task into a closed VQA task, we define a mapping function f_i for each dataset. This function maps a given label l_i^j to a human written textual description:

$$f_i : l_i^j \rightarrow t_i^j,$$

where t_i^j is the textual descriptor of label l_i^j. This process is applied to the entire dataset:

$$D_i = \{(x_i^1, t_i^1), \ldots, (x_i^{N_i}, t_i^{N_i})\}$$

Then the reminder of (incorrect) classes textual descriptions are added to each data point to create a multiple-choice list: $A_i = \{t_i^j, a_i^1, a_i^2, \ldots, a_{M_i^{i-1}}\}$. Lastly, a random permutation is applied, storing the position of the correct label after this operation k_i^j.

These operations convert the initial dataset to a collection of image-text pairs, where each image x_i is associated with: 1. A list of multiple-choice answers 2. The correct index of the label within this list, denoted as k_i:

$$D_i = \{(x_i^1, A_i^1, k_i^1), \ldots, (x_i^{N_i}, A_i^{N_i}, k_i^{N_i})\}.$$

Where π denotes the random permutation function applied to the answers in $A_i^{\pi(j)}$.

Model	Continual Training Data	Hyperparameters	Values
CLIP: ViT-L-14 (commonpool)	Full Data (24M)	batch size (per GPU)	1024
		GPUs	4xH100
		accumulation frequency	2
		effective batch size	8192
		learning rate	$1e{-6}$
		beta1	0.9
		beta2	0.95
		warmup	1000
		epochs	9
		precision	FP32
		gradient clipping norm	1.0
		dataset type	WebDataset
	Concept Balanced Data (8M)	batch size (per GPU)	1024
		GPUs	4xH100
		accumulation frequency	2
		effective batch size	8192
		learning rate	$1e{-6}$
		beta1	0.9
		beta2	0.95
		warmup	1000
		epochs	27
		precision	FP32
		gradient clipping norm	1.0
		dataset type	WebDataset
	Concept Filtered Data (6M)	batch size (per GPU)	1024
		GPUs	4xH100
		accumulation frequency	2
		effective batch size	8192
		learning rate	$1e{-6}$
		beta1	0.9
		beta2	0.95
		warmup	1000
		epochs	36
		precision	FP32
		gradient clipping norm	1.0
		dataset type	WebDataset

Table S7. Hyper-parameters used for continual pretraining.

17.1.2. Closed VQA Evaluation

All evaluated contrastive models have a vision encoder E_{image} and text encoder E_{text}. We first compute the image embedding, $z_{x_i} = E_{image}(\mathbf{x}_i^j)$, along with each candidate answer: $z_{a_i^\pi(j)} = E_{text}(a_i^\pi(j))$ for $l \in [1, M_i]$. Then we compute the cosine similarity score for each caption, $s_{ij} = z_{a_i^\pi(j)} \cdot z_{x_i}^T$ for $l \in [1, M_i]$. The option with the largest s_{ij} is then assigned as the final prediction. If argmax(s_{ij}) has the same index as the corresponding correct answer k_i^j the question is marked as correct, incorrect otherwise.

17.1.3. Closed VQA Conversion Prompts

If a dataset did not explicitly contain a Closed-VQA form, then a group consisting of a biomedical informatics, pathologist converted each class to unique its corresponding caption.

17.2. Retrieval Benchmark Evaluation

Given a dataset of images and captions

$$D_c = \{(x^1, c^1), \ldots, (x^{N_i}, c^{N_i})\}.$$

We evaluate retrieval performance using Recall@k, using

Table S8. Provenance of evaluation benchmark. Each row lists dataset name (with it's corresponding citation), task description, image modality, and number of classes in each corresponding tasks.

Dataset	Task	Modality	N. Classes
Cell Biology			
Cell Cycle & Stage Identification			
BBBC048 (BF) [20]	Cell cycle phase	Fluresence Microscopy	7
BBBC048 (DF) [20]	Cell cycle phase	Fluresence Microscopy	7
BBBC048 (EF) [20]	Cell cycle phase	Fluresence Microscopy	5
Cell Profiling			
PCST Contour [10]	Cell contour	Fluresence Microscopy	3
PCST-Texture [10]	Cell texture	Fluresence Microscopy	3
PCST Eccentricity [10]	Cell eccentricity	Fluresence Microscopy	3
Cell & Structure Identification			
Fluorescence Cells [40]	Organisms and structures	Fluresence Microscopy	13
EMPIAR SBF-SEM [30]	Organisms and structures	Electron Microscopy	5
ICPR2020 Pollen [6]	Pollen structures	Fluresence Microscopy	4
Pathology			
Cytology			
Acevedo et al 2020 [1]	White blood cell	Light Microscopy (Giemsa)	8
Jung et al 2022 [33]	White blood cell	Light Microscopy (Giemsa)	5
Pap Smear 2019 [26]	Pap smear grading	Light Microscopy (Pap Smear)	4
Neoplastic Histopathology			
Kather et al 2016 [34]	Colorectal tissue	Light Microscopy (H&E)	8
LC25000 (Lung) [8]	Lung tissue classification	Light Microscopy (H&E)	3
PCAM [37]	Lymph node classification	Light Microscopy (H&E)	2
LC25000 (Colon) [8]	Colon tissue classification	Light Microscopy (H&E)	2
Non-neoplastic Histopathology			
Tang et al 2019 [55]	Amyloid morphology	Light Microscopy (IHC)	4
Wong et al 2022 [60]	Amyloid morphology	Light Microscopy (IHC)	4
Nirschl et al 2018 [44]	Clinical chronic heart failure	Light Microscopy (H&E)	2
Wu et al 2023 [63]	Mitochondrial morphology	Light Microscopy (IHC)	2
General Microscopy			
Micro-Bench Submodality [40]	Microscopy Submodality	Microscopy	6
Micro-Bench Stain [40]	Microscopy Stain	Microscopy	6
Micro-Bench Domain [40]	Microscopy Domain	Microscopy	6
Micro-Bench Modality [40]	Microscopy Modality	Microscopy	3
Radiology			
Diagnostics			
Chexpert [29]	Chest X-Rays Findings	Chest X-Ray	5
RSNA 2018	Chest X-Rays Findings	Chest X-Ray	2
BreastMNIST [4]	Breast cancer diagnosis	Breast Ultrasound	2
Dermatology			
HAM1000 [57]	Common skin lesions	Dermatoscope	7
Surgery			
Dresden Anatomy Dataset [13]	Anatomy in surgery	Laparoscopic Surgery	7x2
Ophthalmology			
DeepDRiD [38]	Diabetic retinopathy	Retina Fundus Images	4

Text Statistics	Min	Max	Median	IQR	Total
Caption Token Length	4	1324	23	63	4.5×10^5
Caption Character Length	10	3287	78	149	1.10×10^6
Image Statistics					
Image Width (pixels)	70	1024	1024	1	
Image Height (pixels)	70	1024	768	87	
Image Area (pixels²)	4900	1048576	786432	109206	

Table S9. Overview of Biomedical Flickr statistics, detailing text token and character lengths, and image dimensions. Tokens are generated using the BPE tokenizer from the `tiktoken` library

the following protocol:

All evaluated contrastive models have a vision encoder E_{image} and text encoder E_{text}. We first compute the image embedding, $z_{x_i} = E_{image}(\mathbf{x}^i)$, along with each caption: $z_{c_i} = E_{text}(c^i)$ for $l \in [1, M_i]$. Then we compute the cosine similarity score for each caption, $s_{ij} = z_{c^i} \cdot z_{x_i}^T$ for $l \in [1, N_i]$. Captions are arranged from the largest to smallest similarity (s_{ij}). If the correct caption is within the first k-th arranged items, then the option is considered relevant, irrelevant otherwise. lastly, we calculate Recall@k using the following equation:

$$\text{Recall@k} = \frac{\text{Number of relevant items in the top } k \text{ results}}{\text{Total number of relevant items in the dataset}}$$

17.3. Computing Confidence Intervals

Error bars represent 95% confidence intervals (CI) computed via nonparametric bootstrapping using the SciPy $stats.bootstrap$ function with 1000 resampling and default settings.

18. Flickr Dataset Description

The Biomedical Flickr dataset consists of 7k image-caption pairs retrieved from flicker channels with permissive licenses. It mostly spans microscopy. Table S9 shows statistics for the dataset. Table S16 shows 10 random samples from the dataset.

19. Compute Environment

Experiments are performed in a local on-prem university compute environment using 24 Intel Xeon 2.70GHz CPU cores, 8 Nvidia H100 GPUs, 16 Nvidia A6000 GPUs, and 40 TB of Storage.

MeSH Term	Frequency
Humans	2,189,713
Female	990,873
Male	897,332
Animals	775,314
Adult	521,585
Middle Aged	483,806
None	375,532
Aged	371,170
Mice	288,320
Young Adult	205,580
Adolescent	200,700
Retrospective Studies	178,740
Child	166,187
COVID-19	149,167
Cross-Sectional Studies	135,551
Risk Factors	135,111
Treatment Outcome	131,638
Aged, 80 and over	125,351
Signal Transduction	109,685
SARS-CoV-2	100,287
Cell Line, Tumor	100,154
Surveys and Questionnaires	98,150
Prospective Studies	96,729
Prognosis	89,765
Rats	89,433
Pregnancy	88,439
Child, Preschool	79,751
Mutation	79,073
Biomarkers	77,093
Disease Models, Animal	76,046
Cell Proliferation	75,080
Time Factors	75,034
Mice, Inbred C57BL	72,500
Infant	71,594
Pandemics	70,915
China	68,963
Algorithms	64,904
Neoplasms	64,674
Cohort Studies	64,563
Reproducibility of Results	62,995
Phylogeny	62,412
Prevalence	61,952
Apoptosis	61,057
Cells, Cultured	58,811
Cell Line	57,933
Gene Expression Profiling	57,680
Brain	57,472
Case-Control Studies	57,044
Quality of Life	56,170
Infant, Newborn	55,529

Table S10. Top 50 most common MeSH Terms and their frequencies

	Years of experience	Field of study	Role
Annotator 1	3	Developmental biology	PhD Student
Annotator 2	3	Microbiology	PhD Student
Annotator 3*	3	Biomedical Informatics	PhD Student
Annotator 4	5	Biomedical Informatics	PhD Student
Annotator 5	3	Genetics	PhD Student
Annotator 6	9	Biomedical Informatics	MD-PhD Student
Annotator 7*	11	Biomedical Engineering, Molecular Biology, Surgical Data Science, ML	Industry Director, Post-doc
Annotator 8*	14	Pathology, Cell biology, Neuroscience	Attending Pathologist, Post-doc

Table S11. Description of cluster annotators. Years of experience include years of research or laboratory experience in a biology/biomedical or microscopy related discipline. * Annotator developed taxonomy.

Category	Other	Commercial	Noncommercial	Total
Scientific Formulae and Equations	2,322	20,384	6,182	28,888
PCR	2,307	25,353	7,693	35,353
Tools and Materials	10,320	210,740	51,339	272,399
Maps	18,700	264,092	41,473	324,265
Hand Drawn and Screen Based Visuals	24,690	356,047	76,404	457,141
Graph and Network	26,453	415,737	83,470	525,660
Tables	26,190	269,384	343,455	639,029
Immuno Assays	38,055	651,339	215,238	904,632
Chemical Structures	92,881	839,082	196,119	1,128,082
Clinical Imaging	82,349	1,078,901	766,908	1,928,158
Microscopy	101,617	1,818,302	597,413	2,517,332
Illustrative Diagrams	140,925	2,227,690	551,380	2,919,995
Plots and Charts	542,718	9,426,147	2,394,358	12,389,066

Table S12. Number of images by global taxonomy concepts.

License Type	Number of Articles
CC0	132592
CC BY	3795419
CC BY-SA	1287
CC BY-ND	7900
CC BY-NC	771755
CC BY-NC-SA	275638
CC BY-NC-ND	642982
Other	414857

Table S13. Number of articles by license type. As described in the PMC website: "Commercial use allowed: CC0, CC BY, CC BY-SA, CC BY-ND. Non-commercial use only: CC BY-NC, CC BY-NC-SA, CC BY-NC-ND. Other: no machine-readable Creative Commons license, no license tagged, or a custom license."

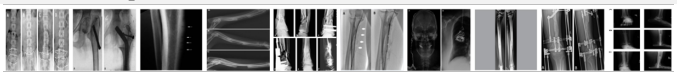
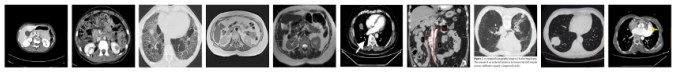
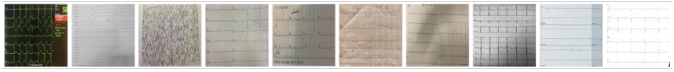
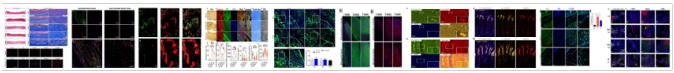
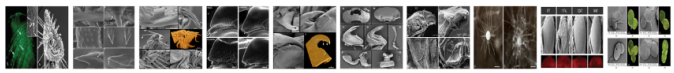
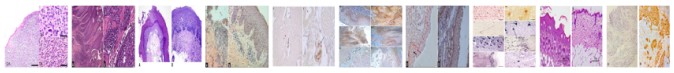
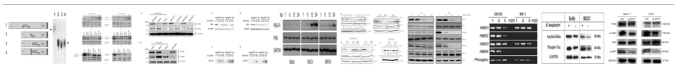
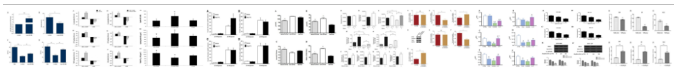
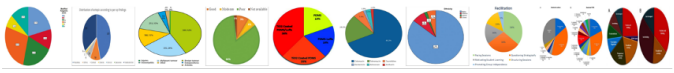
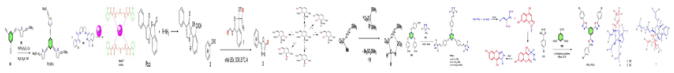
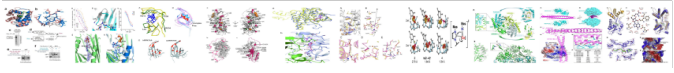
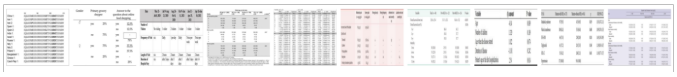
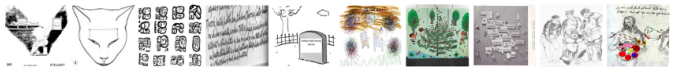
Cluster Example	Global Taxonomy	Local Taxonomy	Multi-panel
	Clinical Imaging Data	x-ray radiography	✓
	Clinical Imaging Data	computerized tomography	✓
	Clinical Imaging Data	electrocardiogram	×
	Microscopy	fluorescence microscopy	✓
	Microscopy	electron microscopy	✓
	Microscopy	light microscopy	✓
	Immuno Assays	gel electrophoresis	✓
	Plots and Charts	bar plot	✓
	Plots and Charts	pie chart	×
	Chemical Structures	2D chemical reaction	×
	Chemical Structures	3D chemical structure	×
	Illustrative Diagrams	signaling pathway	×
	Tables	table	×
	Maps	map	×
	Drawings	drawing	×
	Tools and Materials	lab equipment	✓
	Natural Images	natural image	×

Table S14. Taxonomy of clusters with example images. Images resized to a uniform width of 10cm and height of 1cm. Column widths adjusted to fit the page.

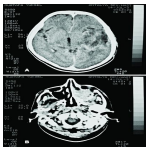
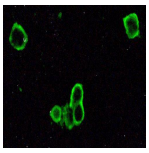
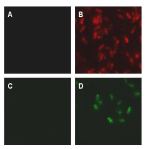
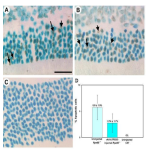
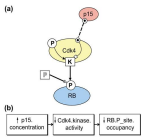
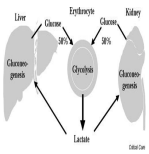
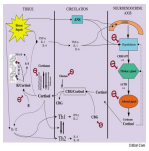
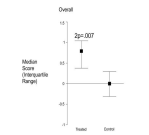
Image	Caption
	Four weeks after the accident: the radiograph shows good alignment (lateral view).
	CT-images belong to brain and maxillary sinus. Ct-images taken before any procedure applied illustrating two separate tumors in the brain (2a) and left-maxillary sinus (2b). Note the enhancing heterogeneous tumor at the left-temporoparietal lobe shifting the midline to the right (2a); and invasion of the tumor (T4) in the left-maxillary sinus into the adjacent tissues (2b).
	Detection of HSV-1 antigen. An impression cytology smear obtained from a patient with HSK showing the presence of rounded up corneal epithelial cells positive for viral antigen. Infected cells show brilliant apple green fluorescence. Note the absence of background staining. Indirect immunofluorescence assay, × 500.
	Detection of presence of CotC-LTB and CotC-TTFC by immunofluorescence microscopy. Sporulation of B. subtilis strains was induced by the resuspension method, and samples were taken 6 h after the onset of sporulation and analysed by immunofluorescence microscopy as described previously [46]. Samples were labelled with mouse anti-LTB antibody followed by anti-mouse IgG-TRITC conjugate (red fluorescein, Panels A & B), or rabbit anti-TTFC antibody followed by anti-rabbit IgG-FITC conjugate (green fluorescein, Panels C & D). Panels A & C, wild type spores; Panel B, isogenic spores expressing CotC-LTB); Panel D, isogenic spores expressing CotC-TTFC.
	Clinical photograph of the abdomen (close-up view): The surgical scar of implantation of baclofen pump is seen. The pigtail catheter emerges close to the scar. The skin around the pigtail catheter is red and angry-looking. Approximate position of baclofen pump is marked on the skin with a pen.
	Comparison of apoptotic cells numbers Photomicrographs of the outer nuclear layer of 8 mo uninjected Rpe65-/- (A), rAAV.RPE65 injected Rpe65-/- (B) and C57BL/6J mice (C) stained for apoptotic nuclei (arrows). (D) Graphical presentation of the percentage of photoreceptor nuclei that are apoptotic in uninjected Rpe65-/-, rAAV.RPE65-injected Rpe65-/- and uninjected C57BL/6J mice. Apoptotic and total photoreceptor nuclei were counted along 60 μm lengths of the outer nuclear layer of mice at 7 mo post-injection (8 mo of age). Average total photoreceptor counts: uninjected Rpe65-/- = 106.8 ± 22.9, rAAV.RPE65 injected Rpe65-/- = 134 ± 30.3, uninjected C57 = 213.5 ± 3.3. All data are mean ± S.D.
	Flow diagram showing randomisation and response rates of the survey.
	(a) A simplified model of a Cdk4 kinase molecule illustrates how basic BioD icons and action arrows can concisely represent intra- and inter-molecular actions. The Cdk4 molecule includes a kinase site (K) that, when active, phosphorylates a phosphorylation site (P) on the RB protein. The kinase site on Cdk4 is activated (filled arrow) by occupancy of its phosphorylation site and inhibited (open-squared arrow) by occupancy of the binding site (dotted circle) that binds the Cdk4 inhibitor p15. (b) An 'event model' derived from the model above. Events are defined as changes of state of one or more functional properties of icons in a state model. Here, for instance, the event model displays a chain of events triggered by an increase of p15 concentration (see text).
	Renal contribution to endogenous glucose release from lactate during the postabsorptive phase. Data from [1].
	Strategy for detection and treatment of adrenal failure during sepsis. ACTH, adrenocorticotrophic hormone.
	Mammographic changes in treated and control individuals. Values are expressed as median change from baseline and interquartile range.

Table S15. Bomedica dataset Image-caption examples Benchmark Example: 10 examples from Biomedica dataset grouped by concept

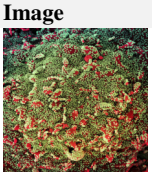
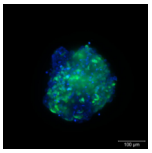
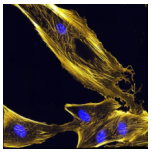
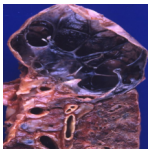
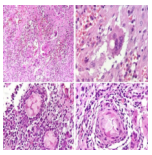
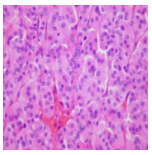
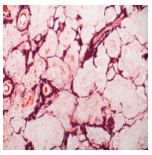
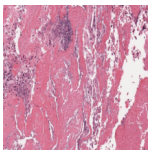
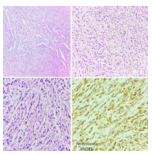
Image	Caption
	Colorized scanning electron micrograph of a cell (red) heavily infected with SARS-CoV-2 virus particles (green), isolated from a patient sample.
	HBE human derived cell lines cultured as micro-tissues (DAPI staining in blue; fixed with PFA). Infection with human Adenovirus type 2 expressing GFP.
	Immunofluorescence image of actin bundles in muscle precursor cells called myoblasts. The actin is labeled with fluorescently-tagged phalloidin, which is a toxin from the Amanita phalloides mushroom. Nuclei are shown in blue.
	Cut surface of a large apical bulla. Involved hilar lymph nodes also present.
	In this pleural biopsy, a chronic inflammatory reaction with giant cells is seen, reacting to the presence of food material originating from an esophageal fistula. In the two bottom photographs, the structure of this material, although deteriorated, is still preserved. These are particles of vegetable material which, due to their size and shape, surely are seed -derived storage cells.
	Typical carcinoid tumor with organoid/insular growth and oncocytic tumor cells. There are many morphologic variants of carcinoid tumors.
	Metastatic calcifiication of alveolar walls and blood vessels in an area of acute pneumonitis.
	Dicrofilarium
	Pleomorphic lung carcinoma is composed of 10% or more of spindle cells and/or giant cells admixed with variable amounts of adenocarcinoma, squamous carcinoma, or large cell carcinoma. Some are composed solely of spindle cells and/or tumor giant cells. The diagnosis can only be made in a surgical specimen, not in a biopsy specimen. The type(s) of non-spindle cell carcinoma that are present should be mentioned in the pathology report. The term “sarcomatoid carcinoma” should be avoided because it is an umbrella term encompassing pleomorphic carcinoma, carcinosarcoma, and pulmonary blastoma.These images of pleomorphic carcinoma show malignant spindle cells admixed with adenocarcinoma
	Normal PA chest x-ray

Table S16. BiomedFlickr Benchmark Example: 10 Random examples from biomedflcikr

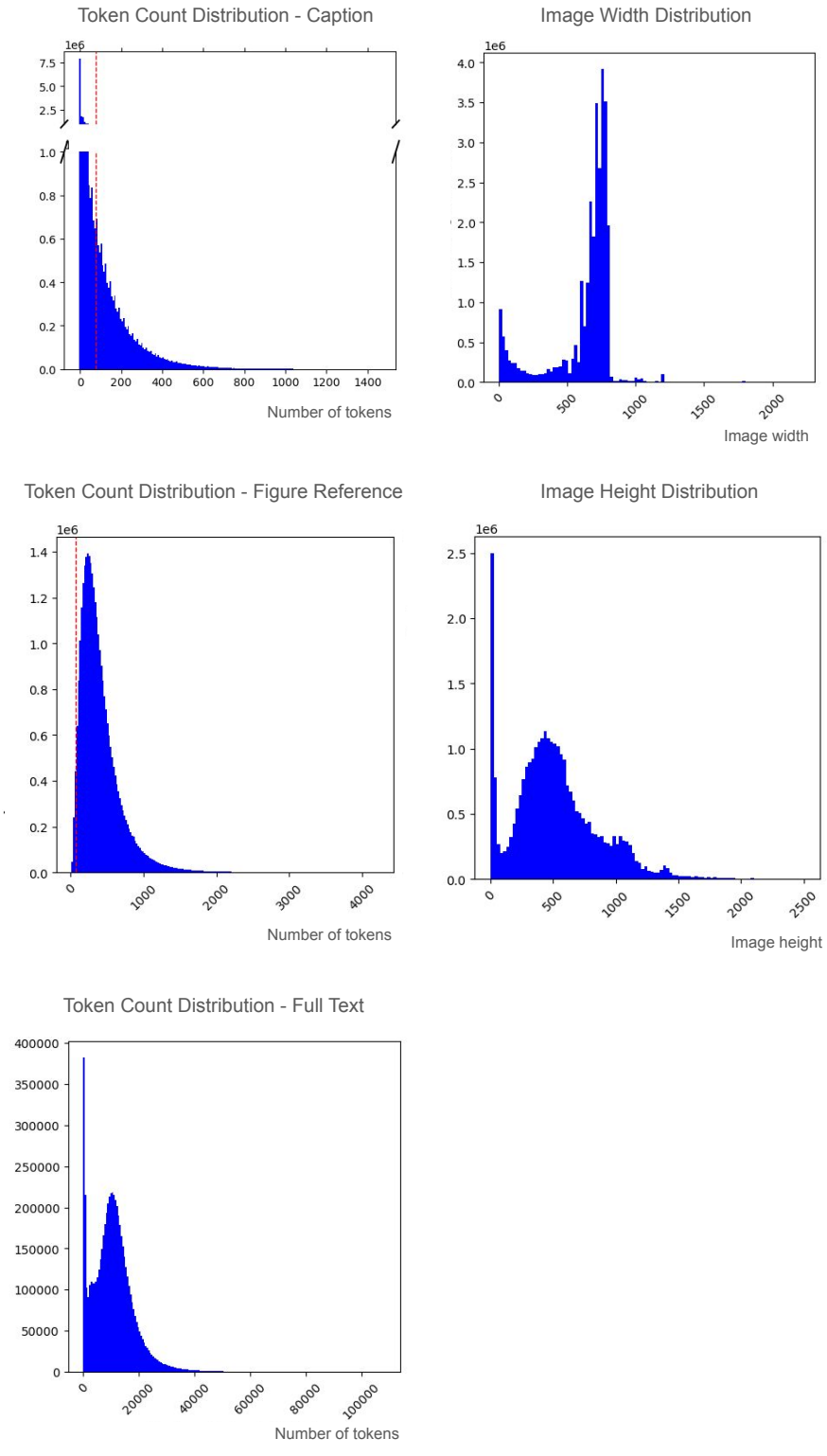


Figure S5. Distributions of token counts and image dimensions in the dataset. Histograms are shown for token counts in captions, figure references, and full text, as well as for image widths and heights. Outliers have been excluded to highlight the central tendencies and areas of higher data density.

```
TAXONOMY = {'Ambiguous': ['ambiguous'],

    'Chemical Structures': ['2D chemical reaction','3D chemical reaction','2D chemical structure',
                            '3D protein structure','3D chemical structure'],

    'Clinical Imaging': ['x-ray radiography','optical coherence tomography','endoscopy',
                         'intraoral imaging','angiography','procedural image','skull','patient photo',
                         'functional magnetic resonance', 'magnetic resonance','eye','mammography',
                         'electrocardiography', 'clinical imaging', 'skin lesion','ultrasound',
                         'specimen','computerized tomography','laryngoscopy','teeth',
                         'intraoperative image','surgical procedure','brain'],


    'Graphs and Networks': ['graph','neural network','network'],

    'Illustrative Diagrams': ['sankey diagram','metabolic pathway','scientific illustration','diagram',
                              'signaling pathway','illustrative diagram','flow diagram',
                              'cohort selection flowchart','illustration','drawing','system diagram',
                              'flowchart'],

    'Immuno Assays': ['immunocytochemistry','karyotype','gel electrophoresis','immunoassay',
                      'immunoblot','assay','immunohistochemistry'],

    'Laboratory Specimens and Cultures': ['reagents','laboratory specimen','bacterial culture'],

    'Maps': ['map'],

    'Microscopy': ['scanning electron microscopy','electron microscopy','flowcytometry',
                   'transmission electron microscopy','light microscopy','fluorescence microscopy',
                   'phase contrast microscopy','confocal microscopy','epifluorescence microscopy',
                   'microscopy'],

    'Natural Images': ['face','aerial photography','natural image','human head','humans and devices',
                       'human','insects','nature'],

    'PCR': ['qPCR','RT PCR'],


    'Plots and Charts': ['violin plot','bar plot','roc curve','sequence plot','radial plot','plot',
                         'matrix plot','phylogenetic tree','process chart','dot plot','pyramid chart',
                         'forest plot','box plot','survival curve','circos plot','venn diagram',
                         'heatmap plot','circular plot','scatter plot','word cloud','list','tree',
                         'density plot','funnel plot','plot and chart','2D mesh','3D plot',
                         'radial diagram','pie chart','manuscript','histogram',
                         'differential gene expression matrix','line plot','signal plot'],


    'Screen Based Visuals': ['screenshot','user interface'],

    'Scientific Formulae and Equations': ['algorithm'],

    'Tables': ['table','checklist table'],

    'Tools and Materials': ['medical equipment','microscope','electronic circuit','lab equipment',
                            'tool']}
```

Figure S6. Hierarchical Taxonomy. Filtered hierarchical taxonomy with topics included in the BIOMEDICA dataset.