Supplementary Material - WikiDO: A New Benchmark Evaluating Cross-Modal Retrieval for Vision-Language Models

A Datasheet for WikiDO dataset

A.1 Motivation

Q1 **For what purpose was the dataset created?** *Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

(a) VLMs are trained on very large amounts of diverse image and text data, thus making them robust in reasoning. Evaluating how well VLMs generalize to out-of-distribution (OOD) instances is one way to measure the robustness. This has been addressed in prior work [4, 3] by finetuning VLMs on a given corpus for a given task [5] and conducting zero-shot evaluations on a new corpus [7]. However, the mere use of an unseen corpus for evaluation does not imply it is OOD. For a more accurate characterization of generalization, the OOD nature of the evaluation data should be carefully established. Therefore, we introduce WikiDO, which serves as a testbed for VLMs to measure how well they generalize to OOD instances.

Q2 **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

(a) The dataset is created by students and the main PI of CSALT Lab, Department of CSE, IIT Bombay, in collaboration with two researchers from Google DeepMind.

Q3 **Who funded the creation of the dataset?** *If there is an associated grant, please provide the name of the grantor and the grant name and number*

(a) There is no associated grant.

Q4 **Any other comments?**

(a) No.

A.2 Composition

Q1 **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

(a) We provide 384k image-text pairs. Each instance of the data has the following information: image path, image ID (Wikipedia image ID), original caption (reference text from Wikipedia), image (unique image ID given in the dataset), page ID (Wikipedia page ID), page title (Wikipedia page title), topic (domain label obtained from Wikipedia Diversity Observatory) and caption (enhanced the original caption by passi through LLava [6].

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

Table 1: Topic-wise number of instances in the dataset

Topic	Full dataset	Train set (100 k)	OOD test set (3000)
monuments and buildings	206460	12561	0
earth	46599	12557	0
books	21189	12557	0
music creations and organizations	19743	12557	0
industry	27401	12557	0
religion	4682	4606	0
sport and teams	14254	12557	0
clothing and fashion	5939	5839	0
folk	5129	5044	0
glam	9335	9165	0
medicine	9324	0	1294
food	10428	0	1675
paintings	4265	0	31

Q2 **How many instances are there in total (of each type, if appropriate)?**

(a) There are a total of 384K pairs of image-texts. The breakdown of these instances across topics is shown in Table 2.

Q3 **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

(a) All the images and the captions were scraped from Wikipedia pages, specifically articles referred to by the Wikipedia Diversity Observatory. From each page that has a topic label, we scraped all the images and the corresponding metadata.

Q4 **What data does each instance consist of?** *"Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.*

(a) The composition of each instance is described in A.2 Q1. All the metadata corresponding to each instance are text fields. The image itself is a .jpeg file of 256x256 resolution.

Q5 **Is there a label or target associated with each instance?** *If so, please provide a description.*

(a) There is no hard label, but the text/caption associated with each image is typically considered to be a label for tasks like image captioning. We additionally provide metadata for each instance like topic name and Wikipedia page title that can be used as a classification label.

Q6 **Is any information missing from individual instances?** *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

(a) No.

Q7 **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** *If so, please describe how these relationships are made explicit.*

(a) No.

Q8 **Are there recommended data splits (e.g., training, development/validation, testing)?** *If so, please provide a description of these splits, explaining the rationale behind them.*

(a) Yes, since the dataset is mainly for the purpose of evaluating OOD generalization, we created train, test and val splits accordingly. A details explanation and rationale behind these splits is given in main paper (§3.2).

Q9 **Are there any errors, sources of noise, or redundancies in the dataset?** *If so, please provide a description.*

(a) We extract images and text pairs from Wikipedia pages. Each page is given a topic label from the Wikipedia Diversity Observatory. This topic is directly propagated from the Wikipedia page to all images on that page. This may lead to some noise in the topic labels associated with image-text pairs.

Q10 **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

(a) Yes, the dataset is self-contained. The metadata also has links to the Wikipedia pages from which the images are scraped. We provide the downloaded images and captions.

Q11 **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** *If so, please provide a description.*

(a) No, all the data is scraped from Wikipedia. All the text has a license of CC-BY-SA. All the images are public domain or CC-BY-SA.

Q12 **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** *If so, please describe why.*

(a) Since the images are scraped from Wikipedia, which already performs checks for any offensive content, this dataset has very little probability for such content. However, some images from the topic of medicine are related to diseases and surgery, which might make certain users anxious.

Q13 **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

(a) People may be present in the images or textual descriptions, but people are not the sole focus of the dataset.

Q14 **Does the dataset identify any subpopulations (e.g., by age, gender)?**

(a) We do not provide any labels of subpopulation as attributes of the image-text pairs but it is possible to deduce them from the Wikipedia diversity Observatory.

Q15 **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** *If so, please describe how.*

(a) Yes, it may be possible to identify people using facial recognition. We do not offer or attempt to provide such means, but institutions that possess large amounts of facial recognition features can identify specific individuals in the data set. People can also be identified via the associated text.

Q16 **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** *If so, please provide a description.*

(a) Since the data was extracted from Wikipedia pages that have been carefully moderated and checked for sensitive content, we do not believe there is any sensitive or personally identifiable information in WIKIDO. Our source, Wikipedia Diversity Observatory, does contain data from diversity axes such as gender, ethnicity, sexual orientation. However, we do not use any data from these axes and restrict WIKIDO to the "topics" axis that covers broad topics such as buildings, food, paintings, etc.

Q17 **Any other comments?**

(a) We urge the user to exercise discretion and demand responsible use of the dataset exclusively for research purposes.

A.3 Collection Process

Q1 **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

(a) We first get a list of Wikipedia page URLs and the corresponding topic label from Wikipedia Diversity Observatory. After crawling all image-text pairs from these articles, we filter images and their associated alt-text. We only take English articles. All metadata associated with the image-text pairs were extracted.

Q2 **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?**

(a) We ran a crawling script to extract all image URLs and texts from Wikipedia dump [2]. For each image URL, we saved and reshaped the images using the img2dataset repository[1]. We ran a preprocessing and filtering script in Python. The final set of image-text pairs was obtained after filtering. These were then passed through LLava [6] to get the enhanced (and more descriptive) captions. After splitting the dataset into train, test and validation sets, the in-domain (ID), out-of-domain (OOD) test sets and validation sets' captions are manually verified for any hallucinations. For each image, the evaluator was specifically asked, ”Is there any made-up/ hallucinated content in the caption that is not supported by the image/reference text?” with an option to answer with a ”Yes” or ”No”. If ”Yes”, then the evaluator was asked to correct the reference text by mainly removing the hallucinations in the enhanced captions.

Q3 **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

(a) All English articles listed on Wikipedia Diversity Observatory were selected.

Q4 **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

(a) Human evaluators were used to verify the test and val sets. They were employed via a data annotation company located in India. The entire annotation exercise involving 9.1K captions cost Rs 75,000.

Q5 **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?** *If not, please describe the timeframe in which the data associated with the instances was created.*

(a) The data was filtered from July to August 2023, but the editors of the Wikipedia page might have included content from before then. The exact date the content was updated will be known from the metadata of the corresponding Wikipedia page.

Q6 **Were any ethical review processes conducted (e.g., by an institutional review board)?** *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

(a) No.

Q7 **Does the dataset relate to people?** *If not, you may skip the remaining questions in this section.*

(a) People are not focus of this dataset, although they may appear in the images and descriptions.

Q8 **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

(a) We extract the data from Wikipedia.

Q9 **Were the individuals in question notified about the data collection?** *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

(a) Individuals were not notified about the data collection.

Q10 **Did the individuals in question consent to the collection and use of their data?** *If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

(a) We follow Wikipedia's images and text licenses. Thus, editors consent to their pages being crawled. However, those depicted in the photograph might not have given their consent to its upload.

Q11 **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

(a) Not applicable

Q12 **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

(a) Since the data was scraped from pages listed on the Wikipedia diversity observatory, the image-text pairs are selected from diverse topics across different axes of diversity, including gender, religion and ethnic groups, etc. The use of Llava for caption enhancement might induce some biases based on its training data, but it is unlikely as the original caption and image are provided as inputs to the model. However, the authors also note that this dataset posits currently the only openly available solution for studying the generalization of multimodal models, examining their potential benefits and harms.

Q13 **Any other comments?**

(a) No

A.4 Preprocessing, Cleaning, and/or Labeling

Q1 **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** *If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

(a) An entire set of preprocessing and filtering steps was applied to all the instances to remove low-quality image-text pairs. These details are mentioned in the main draft. The preprocessing of the images is done by resizing them to 256x256 by adding white borders.

Q2 **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** *If so, please provide a link or other access point to the "raw" data.*

(a) The raw images are not saved. However, the links of all images (prior to filtering) along with the text and metadata will be provided.

Q3 **Is the software used to preprocess/clean/label the instances available?** *If so, please provide a link or other access point.*

(a) The preprocessing and filtering script is available in our Github repository: `https://kaggle.com/competitions/wikido24`.

Q4 **Any other comments?**

(a) No.

A.5 Uses

Q1 **Has the dataset been used for any tasks already?** *If so, please provide a description*

(a) This dataset is introduced in this work, and is used for image-text/text-image retrieval tasks.

Q2 **Is there a repository that links to any or all papers or systems that use the dataset?** *If so, please provide a link or other access point.*

(a) No.

Q3 **What (other) tasks could the dataset be used for?**

(a) We encourage future researchers to use WIKIDO for several tasks. Particularly, we see applications of the dataset in image and text representation learning, image-to-text generation, image captioning, and other common multimodal tasks. It presents a unique opportunity to test domain generalization for all these multimodal tasks. We also use the Wikipedia Diversity Observatory as the source, which contains many more axes of diversity (e.g., gender, ethnicity, etc.) beyond topics. These could be explored further to construct other variants of WIKIDO.

Q4 **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** *For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

(a) As this data is curated from Wikipedia, it mirrors the biases of the content available on Wikipedia. This can also be noted in multiple visualizations by Wikipedia Diversity Observatory. Therefore, this dataset should not be used directly to make decisions about people.

Q5 **Are there tasks for which the dataset should not be used?** *If so, please provide a description.*

(a) Any model trained or fine-tuned on this data as-is should not be used or deployed directly. It can exhibit biases, and hence any direct use would not be responsible.

Q6 **Any other comments?**

(a) No.

A.6 Distrbution

Q1 **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** *If so, please provide a description.*

(a) Yes, the dataset will be open-source.

Q2 **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?** *Does the dataset have a digital object identifier (DOI)?*

(a) The data will be available through the website link and will also be hosted on Huggingface.

Q3 **When will the dataset be distributed?**

(a) On (and after) 12/06/2024.

Q4 **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

(a) Dataset will be distributed under CC-BY-SA-4.0

Q5 **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** *If so, please describe these restrictions and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

(a) The images from Wikipedia have license of public domain and text have license CC-BY-SA. We do not own the copyright of the images or text.

Q6 **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

(a) No.

Q7 **Any other comments?**

(a) No.

A.7 Maintenance

Q1 **Who will be supporting/hosting/maintaining the dataset?**

(a) Hosting will be done at Huggingface and maintained by the authors.

Q2 **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

(a) Please contact the authors through email.

Q3 **Is there an erratum?** *If so, please provide a link or other access point*

(a) Errata will be documented as future releases on the dataset website.

Q4 **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

(a) WIKIDO will be updated to include more kinds of domains in future, encompassing different axes of diversity. The updates will be reported on the website for WIKIDO.

Q5 **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** *If so, please describe these limits and explain how they will be enforced.*

(a) Not applicable.

Q6 **Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.**

(a) Yes, older versions will continue to be hosted on the same website. If new datasets are added, they will be communicated on the WIKIDO website.

Q7 **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*

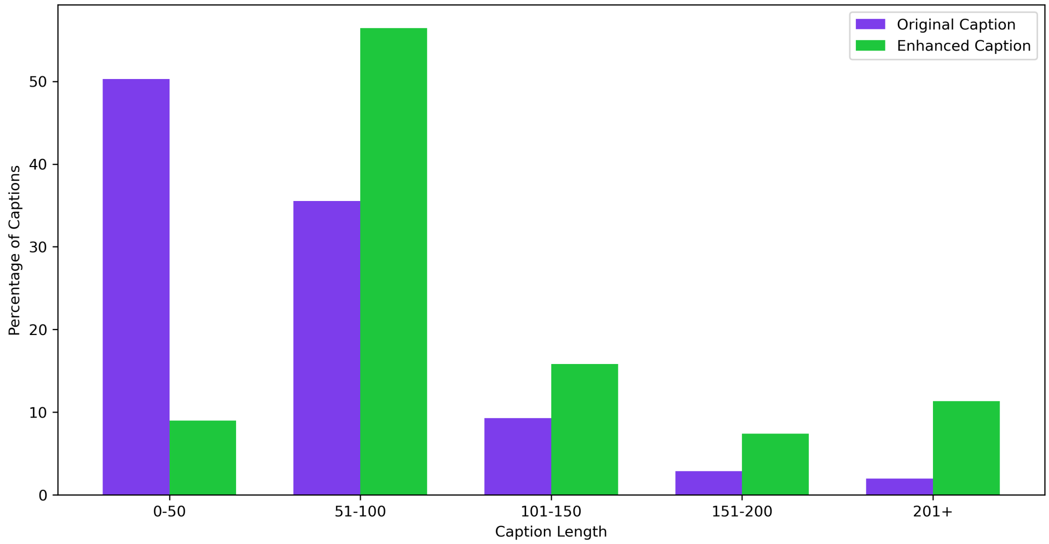
(a) All future researchers are encouraged to build upon this dataset; however, we will not be able to verify these contributions. These researchers may mail the authors who will update such contributions on WIKIDO website for further reach.

Q8 **Any other comments?**

(a) No.

B Dataset Statistics & Analysis

Figure 1: Percentage of captions that range between a given character length.



B.1 Caption length

Figure 1 shows the percentage of captions that are within a given length bucket. It is clear from the image that enhanced captions are longer than the original captions. More than 50% enhanced captions range between 51-100 characters whereas about 50% original captions range are under 50 characters. Figure 2 shows topic-wise average length of captions. Across all topics, enhanced captions have increased their length on average; captions of topics such as folk, paintings have doubled in length. This increase in length of enhanced captions (and increased common noun frequency, c.f., Figure 2 in Section 3.2 of the main paper) hints towards original captions becoming more descriptive with enhancement.

Figure 2: Average length of caption for each topic.

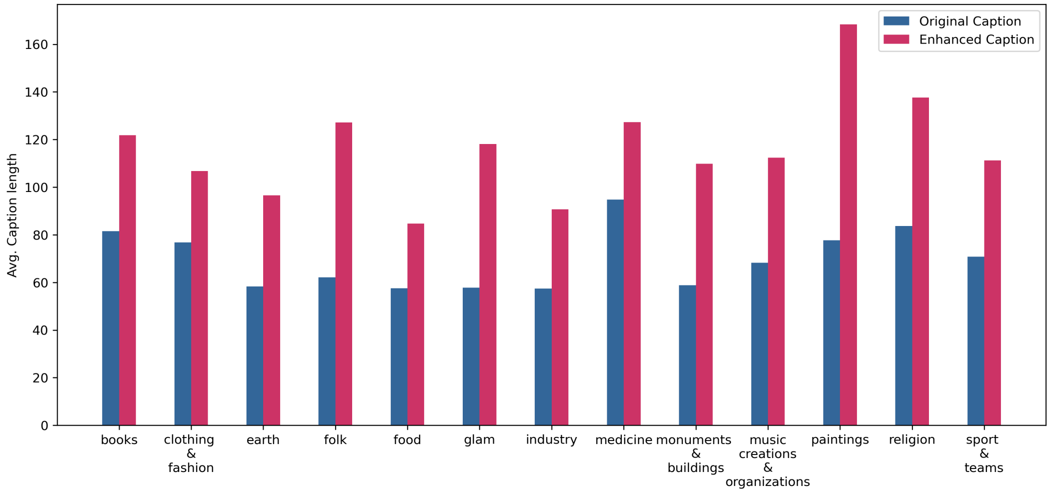


Table 2: Topic-wise number of image objects found in Grounding analysis

Topic	No. of Objects	
	Enhanced Captions	Original captions
monuments and buildings	547427	302466
earth	108612	66474
industry	67693	44113
music creations and organizations	64236	37116
books	54814	32086
sport and teams	40660	23684
food	31385	21548
glam	26339	15991
medicine	22446	15447
clothing and fashion	17789	13351
folk	14045	8417
religion	13626	8305
paintings	12961	7668
Total	1022033	596666

B.2 Grounding

We ran Grounding-DINO on 338K instances of WikiDO. There are 1M (noun-phrase, image bounding-box) pairs for enhanced captions and 596K (noun-phrase, image bounding-box) pairs for original captions corresponding to these 338K instances. Before running Grounding-Dino, 1.9M noun phrases were extracted using the spaCy parser for enhanced captions and 1M noun phrases were extracted for the original captions. Meaningful grounding retained by Grounding-DINO for enhanced captions is more than that for original captions by a fraction of 0.67.

C Additional Experiments

To empirically establish superior quality of enhanced captions over original captions we finetune CLIP (best performing model for WikiDO) on original captions. As shown in Table 3, $\approx 15\%$ reduction in R@1 for WikiDO ID and $\approx 10\%$ reduction in R@1 for WikiDO OOD is observed for both text-to-image and image-to-text retrieval. This in conjunction with grounding object analysis,

Table 3: Performance of CLIP model on original captions. Here Z denotes zero-shot and W denotes finetuned on WikiDO 100K train split.

Model		WikiDO ID Test set (3K) (N=128)						WikiDO OOD Test set (3K) (N=128)					
		Image → Text			Text → Image			Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP (ViT-L)-428M	Z	61.6	79.7	85.6	60.7	78.0	83.7	61.1	80.9	86.1	61.0	80.9	85.6
	W	66.9	84.3	89.3	66.4	83.5	88.6	63.6	82.2	86.8	63.2	82.7	87.1

Table 4: Performance of CLIP model trained on WikiDO 100K split with varying random seeds

Model	seed	WikiDO ID Test set (3K) (N=128)						WikiDO OOD Test set (3K) (N=128)					
		Image → Text			Text → Image			Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP (ViT-L)-428M	2024	82.3	94.9	97.5	81.3	93.6	97.0	73.6	88.0	92.0	73.1	88.2	91.6
	42	82.8	95.0	97.4	81.5	94.4	96.7	73.4	87.7	91.8	72.9	88.3	91.9

and less correction of enhanced captions by human annotators gives a strong evidence that enhanced captions are superior to original captions.

To verify the improvement is not due to a randomly good combination of mini-batches, we train CLIP (best performing model on WikiDO) with a different random seed. Table 4 shows that the performance of CLIP does not differ by much. We were limited by compute constraints, and hence evaluated using the best-performing model across 2 different seeds.

References

[1] Romain Beaumont. img2dataset: Easily turn large sets of image urls to an image dataset. `https://github.com/rom1504/img2dataset`, 2021.

[2] Wikimedia Foundation. English wikipedia dump. `https://dumps.wikimedia.org/`, June 2024.

[3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[4] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[5] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

[6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[7] Peter Young, Alice Lai, Micah Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.