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

T Pavan Kalyan* IIT Bombay

Piyush Singh Pasi* [*] Amazon

Sahil Nilesh Dharod IIT Bombay

Azeem Azaz Motiwala IIT Bombay

Preethi Jyothi IIT Bombay

Aditi Chaudhary Google, Deepmind

Krishna Srinivasan Google, Deepmind

Abstract

Cross-modal (image-to-text and text-to-image) retrieval is an established task used in evaluation benchmarks to test the performance of vision-language models (VLMs). Several state-of-the-art VLMs (e.g. CLIP, BLIP-2) have achieved near-perfect performance on widely-used image-text retrieval benchmarks such as MSCOCO-Test-5K and Flickr30K-Test-1K. As a measure of out-of-distribution (OOD) generalization, prior works rely on zero-shot performance evaluated on one dataset (Flickr) using a VLM finetuned on another one (MSCOCO). We argue that such comparisons are insufficient to assess the OOD generalization capability of models due to high visual and linguistic similarity between the evaluation and finetuning datasets. To address this gap, we introduce WIKIDO (drawn from **Wiki**pedia **D**iversity **O**bservatory), a new cross-modal retrieval benchmark to assess the OOD generalization capabilities of pretrained VLMs. This consists of 384K image-text pairs from Wikipedia with domain labels, along with carefully curated, human-verified in-distribution (ID) and OOD test sets of size 3K each. The image-text pairs are very diverse in topics. We evaluate different VLMs of varying capacity on the WIKIDO benchmark; BLIP-2 achieves zero-shot performance of R@1$\approx$ 66% on the OOD test set, compared to $\approx$ 81% on MSCOCO and $\approx$ 95% on Flickr. When fine-tuned on WIKIDO, the R@1 improvement is at most $\approx$ 5% on OOD instances compared to $\approx$ 12% on ID instances. WIKIDO offers a strong cross-modal retrieval benchmark for current VLMs, especially for evaluating OOD generalization. Our benchmark is hosted as a competition at `https://kaggle.com/competitions/wikido24` with public access to dataset and code.

1 Introduction

Vision-language models (VLMs) are multimodal models that jointly reason on image and text. VLMs are pretrained on very large amounts of diverse image and text data, thus making them capable of robust reasoning. A true measure of this capability is to evaluate how well VLMs generalize to out-of-distribution (OOD) instances. This has been addressed in prior work [1, 2] by finetuning VLMs on a given corpus for a given task and conducting zero-shot evaluations on a new corpus. 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.

[*] Work done as a student at IIT Bombay.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

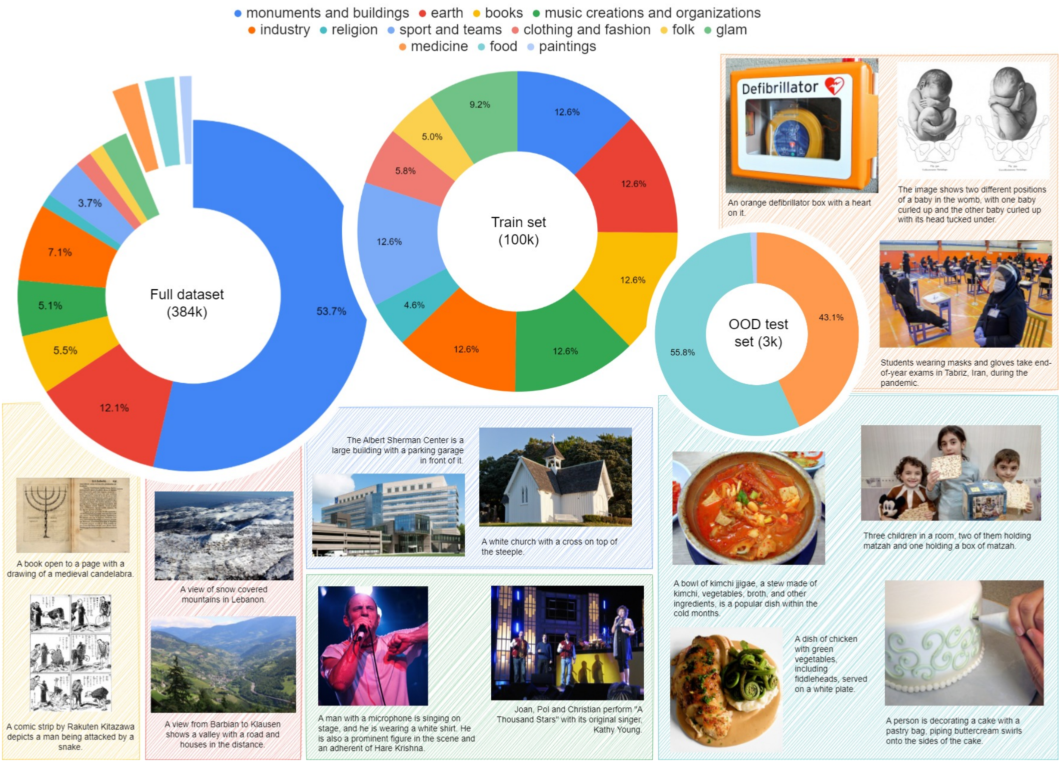

Figure 1: Distribution of topics in the final filtered dataset

In this work, we present a new benchmark WIKIDO that serves as a testbed for VLMs to measure how well they generalize to OOD instances. WIKIDO consists of image-text data derived from *Wikipedia Diversity Observatory*, a diverse source of Wikipedia articles spanning several diversity axes including geography, gender, ethnicity and domains/topics. We focus on the "domains" axis that is most diverse in terms of coverage and spans different topics (as determined via topic labels assigned to each article) such as food, books, fashion and sports. We curate a dataset consisting of 1) 354K training images with corresponding text and 2) two evaluation sets – an in-domain (ID) set and an out-of-domain (OOD) set[2] drawn from domains that are seen and unseen during training, respectively. Our OOD evaluation set is carefully constructed to be used as a reliable testbed for VLMs.

Figure 1 highlights the main aspects of WIKIDO including the domains spanned by the articles, their distribution and a few illustrative image-text pairs. In this work, we focus on cross-modal (image-to-text and text-to-image) retrieval tasks. We show retrieval performance of well-known VLMs, namely CLIP [3], BLIP [1] and BLIP-2 [2], on WIKIDO test sets before and after finetuning on the WIKIDO training instances. The best-scoring VLM, CLIP, achieves a modest zero-shot R@1 of 68% on the OOD test set. Finetuning on the WIKIDO train set improves zero-shot R@1 on the OOD set by only 5%, while zero-shot R@1 on the ID set improves substantially more by 12% further highlighting the difference between the two evaluation sets. We have hosted our code, WIKIDO datasets and a leaderboard with our current VLMs at `https://kaggle.com/competitions/wikido24`.

2 Related Work

Image-text datasets. The rapid progress of vision-language models (VLMs) in recent years can be largely attributed to the emergence of high-quality multimodal datasets. These include large, automatically-filtered datasets that are crawled from the web and smaller datasets that are human-annotated. The larger datasets comprising millions of instances include SBU [4], CC3M [5], CC12M [6], YFCC-100M [7], WIT [8], LAION-400M [9], and LAION-5B [10]. These large datasets have

[2] OOD might be more appropriately expanded as out-of-domain rather than out-of-distribution in WIKIDO, given that both ID and OOD images are extracted from the same source (i.e., Wikipedia).

been primarily used for pretraining VLMs to achieve good zero-shot performance on downstream tasks. The smaller datasets are typically created by first crawling images from the internet and then manually annotating labels, regions and textual descriptions for the images. These include Flickr30K [11], MSCOCO [12] and Visual Genome [13]. Flickr30K and MSCOCO consist of images of everyday activities, most commonly used as evaluation benchmarks for cross-modal retrieval.

Domain generalization datasets. Existing domain generalization benchmarks such as Office-Home [14], PACS [15], and VLCS [16] are predominantly focused on image classification and are not multimodal. A recent effort to extend the task of domain generalization to image-text tasks is Domain Generalization for Image Captioning (DGIC) [17]. DGIC collates popular existing datasets from five domains: common domain sourced from MSCOCO, assistive domain sourced from Vizwiz [18], social domain sourced from Flickr30k, avian domain sourced from CUB-200 [19], and floral domain sourced from Oxford102 [20]. While datasets like MSCOCO, Vizwiz, and Flickr30k represent common objects from daily life and are easier domains for VLMs, the avian and floral domains are significantly more challenging. Prior work [17, 21, 22] using any of Flickr30k, MSCOCO, and Vizwiz as test domains show good generalization performance since these datasets contain images of generic objects that appear commonly across datasets. We aim for WIKIDO to serve as a more challenging benchmark to evaluate the generalization abilities of VLMs.

Vision-language models. VLMs are broadly focused on tasks related to cross-modal understanding and cross-modal generation. A critical component of understanding is aligning the visual and textual features. Models like CLIP [3] and ALIGN [23] use a dual-encoder model to individually extract features and align them through a global contrastive loss. UNITER [24] utilizes a multimodal encoder to extract visual and textual characteristics jointly. ALBEF [25] introduced image-text matching and masked language modelling to align the image-text representations. FILIP [26] works at the granularity of image patches and textual words to further refine the alignment. BLIP [1] introduces a new vision-language pretraining framework with both vision-language understanding (image-text contrastive loss and image-text matching loss) and generation objectives (language modelling loss). Similar to BLIP, BLIP-2 [2] also uses both kinds of objectives but bootstraps vision-language pre-training from off-the-shelf pre-trained image encoders and large language models as textual encoders. BLIP-2 introduced a lightweight Querying Transformer, which is trained in two stages to bridge the modality gap. The first stage uses a frozen vision encoder for vision-language representation learning. The second stage bootstraps vision-to-language generative learning from a frozen language model. We evaluate all these three VLMs, CLIP, BLIP and BLIP-2, on WIKIDO.

3 WIKIDO: A New Evaluation Benchmark

We present WIKIDO, a new image-text retrieval dataset for the improved evaluation of VLMs for OOD generalization. We will first describe the source of the dataset (§3.1), followed by details about the data curation process and how the final data splits were obtained (§3.2).

3.1 Source of WIKIDO

WIKIDO is derived from the *Wikipedia Diversity Observatory*[3] (WDO). WDO consists of data, visualizations and tools to analyze and bridge the gap in content in Wikipedia, based on the current state of diversity across Wikipedia articles. This diversity is assessed based on a few specific categories: geographical location, gender, sexual orientation, ethnic groups, religious groups and topical coverage. We chose English articles from the topical coverage category to create the WIKIDO dataset, since this was most extensive in terms of coverage across topics. The Wikipedia articles in this category are labelled with one of the following topics: Earth, Monuments and Buildings, GLAM (Galleries, Libraries, Archives and Museums), Folk, Food, Books, Paintings, Clothing and Fashion, Sports and Teams, Music Creations and Organizations, and People. This categorization of articles into topics also aids the construction of ID and OOD test sets in WIKIDO. Other diversity axes also offer potential for creating multimodal benchmarks to evaluate visual language models. By tagging and categorizing data across dimensions such as geography, gender, sexual orientation, ethnicity, religion, and topics, researchers can assess cultural biases and generalization capabilities in

[3] `https://wdo.wmcloud.org/`

multimodal contexts for both monolingual and multilingual settings.

3.2 Details of Data Curation

The data from WDO contains meta-information about an article and the corresponding topic label. We find all Wikipedia pages from Wikipedia dumps [27] and extract the URLs of the images from the articles. The topic label associated with the page is assumed to be the topic label for all the images in the page. For each image, we extracted metadata like the page URL, page title, height, width, and three different types of text associated with the image. These are: 1. Reference description: The caption that is visible on the Wikipedia page just below the image. 2. Attribution: Text appearing on the Wikimedia page of the image. 3. Alt-text description: Text used by accessibility/screen readers when the image is not visible. This crawled dataset consists of 2.7M image-text pairs out of which 1.2M are unique images. Based on the data pipeline adopted by WIT [8], we used the following filtering steps:

Table 1: Dataset schema

Key	Description
image_path	path of the image
image_id	Wiki ID of the image
orig_cap	Reference text from Wikipedia
image	Unique image ID given in the dataset
page_id	Wiki ID of the page from which the image was extracted
page_title	Title of the wikipedia article from which the image was extracted
topic	Topic label from Wikipedia Diversity Observatory
caption	Caption obtained by passing orig_cap through LLava (for test, val sets also human verified)

1. We only retained images that have a research-permissive license such as Creative Commons; the text of Wikipedia is licensed under a CC-BY-SA license.
2. We only retained images that have reference descriptions. These textual descriptions were used as image captions. Reference texts are contextual, and therefore, instead of describing the image, they provide additional information about the context of the image. This property makes this dataset more appropriate for the task of cross-modal retrieval, rather than caption generation. Only those texts were retained that were at least of length three.
3. Only jpg and png images with a height and width of more than 150 pixels were retained.
4. Certain image-text pairs were repeated frequently. These were de-duplicated to retain only single instances. We also removed generic image-text pairs such as flags, maps, logos, etc.

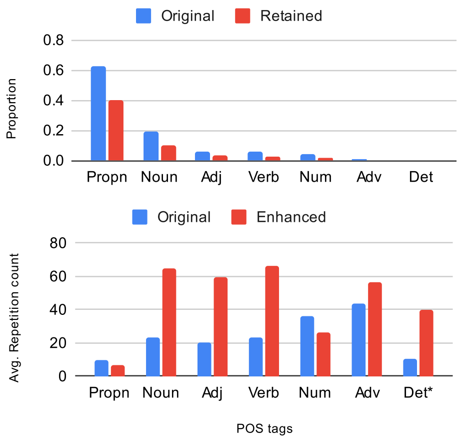

Figure 2: Top: Proportion of POS tags in original captions and those retained in the enhanced captions. Bottom: Average repetition count of retained POS-word pairs in the original and enhanced captions. $*$ denotes that the repetition count of determiners is scaled down by a factor of 1000 for visualization.

Caption enhancement. The final filtered dataset consists of 384K unique image-text pairs, each labelled with a topic label. The distribution across topics is shown in Figure 1. Reference texts in WIKIDO tend to either be descriptive with domain-specific terminology or very concise and non-descriptive. In order to maintain a balance between the original reference texts and more detailed textual descriptions of the image, we passed the original text through the visual instruction-tuned model LLaVA [28]. For each instance, we provided LLaVA with the image and the reference text and prompted it to provide a concise caption describing the image without missing any information from the reference text. The prompt template and some examples are provided in the appendix A.2. To analyze how the LLaVA-enhanced captions differ from the original reference texts, we use a Part-of-Speech (POS) tagger [29] to compute POS tags for every word. Figure 2 shows the POS tag distributions for both original and enhanced captions. While there is a loss of unique POS-word pairs after enhancement, the retained POS-word pairs tend to repeat at a much higher rate in the enhanced captions compared to the original captions. While retained proper nouns do not often repeat in enhanced and original captions, common nouns, adjectives, verbs, and determiners tend to repeat a lot in enhanced captions. This may be due to the replacement/paraphrasing of specific proper nouns and nouns with more general nouns. Qualitative examples of the most frequently occurring nouns in original and enhanced captions are given in the appendix A.2.

Measuring the domain gap. The final WIKIDO dataset we use in all our experiments uses LLaVA-enhanced captions. To create train-val-test splits, we identified a subset of topics that were semantically different from the rest, both visually and linguistically. We randomly sampled 1000 instances from each topic and passed the instances through CLIP (ViT-L) to get embeddings. Figure 3 shows the t-SNE plots for both image and text embeddings. Although most topics in the dataset overlap in the representation space, paintings, medicine and food (shown in blue) are fairly well-separated in both image and text space. We further validate this quantitatively by measuring the domain gap using Maximum Mean Discrepancy (MMD) [17]. We observe that food, medicine and paintings differ linguistically from the other topics. MMD for visual embeddings show that earth, food and medicine are the most distant topics. The appendix A.3 provides implementation details and results for MMD.

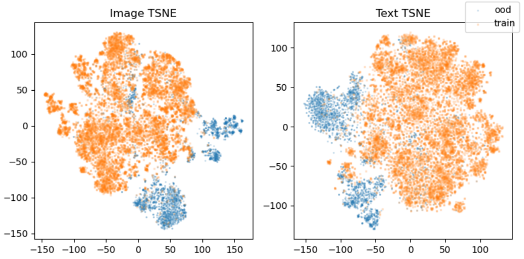

Figure 3: T-SNE plot for embeddings of 1000 random images and texts per topic with perplexity 32

Data splits. Based on the t-SNE plots and MMD analysis, we chose the topics food, paintings and medicine to appear in the OOD test set. The remaining topics are included in the train, validation and in-domain test sets. To further sample a smaller evaluation set of size 3K (comparable to existing test set sizes in MSCOCO and Flickr) from all the samples across the OOD topics, we use the following strategy. We find the image-image similarity and text-text similarity between each OOD instance and all instances of the train set. Then, we pick the top-k similarity scores for each instance of the text and image modalities. If the average of each top-k for each modality crosses a certain threshold, we discard those samples from the OOD test set. Therefore, we only retain those samples in the OOD test set that are highly dissimilar from the train set w.r.t. both image and text modalities.

To mimic the original data distribution, we randomly sample 1000 and 3000 instances from the train set instances to create validation and in-domain (ID) test splits. 354K samples remain in the train set after creating the validation and test splits. As the distribution of topics is highly skewed towards a few topics, we created three different kinds of train splits – a balanced train set consisting of almost equal number of samples from each topic amounting to a total of 100K instances. Henceforth, this set will be referred to as the train set unless mentioned otherwise. Similarly, a balanced training set is created using 200K samples and finally, the training set containing all 354K samples.

Human verification. Since the reference texts were enhanced using LLaVA and could result in hallucinations, we revised the validation, ID, OOD test set captions via a human verification pass. 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 they answer "Yes", then the evaluator was asked to correct the reference text by mainly removing the hallucinations in the enhanced captions. Figure 4 shows that the percentage of instances marked as having hallucinations is comparatively much smaller

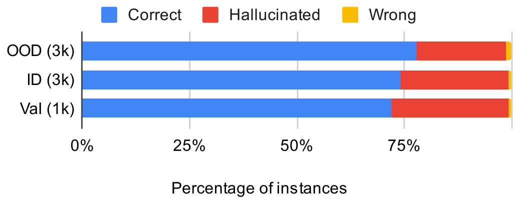

Figure 4: Percentage of instances that were marked as Correct (no hallucination), Hallucinated, and Wrong (caption and image do not match).

Table 3: Comparison of state-of-the-art VLMs. Z denotes zero-shot and W denotes model fine-tuned on 100K split of WikiDO dataset. Number of parameters are listed alongside model names.

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
BLIP (ViT-B)-223M	Z	58.8	81.0	87.7	63.8	82.9	88.7	55.1	73.2	79.4	58.7	76.1	81.6
	W	73.2	90.8	94.6	73.4	89.7	93.9	62.3	79.6	84.3	62.8	80.0	85.2
BLIP (ViT-L)-446M	Z	61.6	83.5	89.7	65.8	85.4	91.1	58.9	76.4	82.6	62.1	79.0	83.9
	W	72.6	90.8	94.6	73.7	90.3	94.2	63.6	80.8	86.1	65.9	81.8	86.9
CLIP (ViT-L)-428M	Z	72.9	88.8	93.1	69.5	87.2	91.6	68.2	85.8	90.3	66.3	84.3	88.9
	W	**82.8**	**95.0**	**97.4**	81.5	**94.4**	**96.7**	**73.4**	**87.7**	**91.8**	**72.9**	**88.3**	**91.9**
BLIP-2 (ViT-L)-473M	Z	70.3	87.8	91.8	74.1	89.3	93.2	66.4	82.2	86.9	70.4	84.9	88.6
	W	82.1	94.0	96.4	**82.5**	94.3	**96.7**	72.1	85.9	90.3	73.6	87.1	90.3
BLIP-2 (ViT-G)-1172M	Z	70.8	89.1	93.1	75.3	90.6	94.1	66.1	81.4	86.2	69.3	84.7	88.3
	W	79.4	93.3	96.2	80.0	93.3	96.1	70.5	84.3	88.2	72.0	85.9	89.2

than correct captions across all splits. Almost all edits done by human raters to the hallucinated captions are "deletion" edits to remove hallucinations. Please refer to the appendix A.4 for more details on edits made by the human raters.

4 Experiments and Results

We benchmark the performance of pretrained CLIP, BLIP, and BLIP-2 models on WikiDO, MSCOCO and Flickr. We show zero-shot performance and the effect of finetuning with different objectives using these pretrained models on all three datasets.

4.1 Experimental Setup

We use the standard train, validation and test sets introduced in MSCOCO [12] and Flickr [11]. For WikiDO, we use the splits introduced in Section 3.2. To finetune BLIP and BLIP-2, we follow the official code published by the authors. To finetune CLIP, we use LAVIS codebase[4]. We use two variants of BLIP (ViT-L, ViT-B) and BLIP-2 (ViT-L, ViT-G) as well as CLIP (ViT-L/14@336px). All models are trained for 6 epochs on 4 A100 80GB Nvidia GPUs. We used a cosine learning rate scheduler. Hyperparameter settings are given in Table 2. A description of the model variants and their pretraining objectives can be found in Appendix B.1. Unlike CLIP, both BLIP and BLIP-2 use a re-ranking strategy for evaluation. For instance, in the re-ranking strategy, we first select the top N captions ($N = 128$ for all experiments) for a given image using ITC (image-text contrastive) scores, i.e., cosine similarity scores. Then, we compute ITM (image-text matching) scores between the image and each of these N texts. The final scores used for ranking are obtained by adding both ITC and ITM scores. For CLIP, we only use cosine similarity (ITC scores) between the image and text for ranking. Conversely, the same applies to text-to-image retrieval. All evaluations use the Recall@k (R@k, k= $1, 5, 10$) metric.

Table 2: Hyperparameter settings

	BLIP ViT-L (ViT-B)	**BLIP-2** ViT-L (ViT-G)	**CLIP**
Batch size	256	224	256
Queue size	57600	57600	-
Pixel Res.	256	256	336
Optimizer	AdamW	AdamW	Adam
lr	$5e^{-6}(1e^{-5})$	$5e^{-6}(1e^{-5})$	$1e^{-6}$
Decay	0.05	0.05	$1e^{-3}$
β_1, β_2	0.9, 0.999	0.9, 0.98 (0.9, 0.999)	0.9, 0.98

[4] https://github.com/salesforce/LAVIS

Table 4: Overview of results for the current way of showing OOD generalization. Z denotes zero-shot, C denotes model fine-tuned on MSCOCO, and F denotes model fine-tuned on Flickr.

Model		COCO Test set (5K) (N=128)						Flickr Test set (5K) (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	57.5	80.7	87.8	36.6	60.9	71.0	86.6	98.0	99.1	67.2	88.9	93.4
	C	75.4	92.8	96.2	58.6	82.2	89.3	94.5	99.7	99.7	83.1	96.9	98.5
	F	68.9	87.6	92.7	51.8	75.8	84.0	95.5	99.5	99.9	85.0	97.7	98.9
BLIP-2 (ViT-L)-473M	Z	78.9	93.9	96.9	62.4	84.1	90.2	95.3	99.7	100.0	85.2	96.9	98.3
	C	83.2	95.9	98.0	66.1	86.6	91.8	97.1	100.0	100.0	88.3	98.0	98.9
	F	80.7	94.7	97.5	64.4	85.4	91.1	97.0	100.0	100.0	89.9	98.4	99.2

4.2 Results and Analysis

Zero-shot. Models BLIP and BLIP-2 perform better on ID than OOD by 3-7% across all R@K. CLIP, on the other hand, performs almost similarly on both ID and OOD (≈1% gap), suggesting better domain coverage during pretraining. Zero-shot performance of CLIP and BLIP-2 is *significantly higher* than that of BLIP. This could be attributed to the larger training data of CLIP compared to BLIP (>3x). Despite BLIP and BLIP-2 being trained using the same dataset, BLIP-2 utilizes a frozen CLIP image encoder, potentially helping to gain further improvement on CLIP, as seen in Table 3. It is interesting that BLIP-2 performs much better on text-to-image retrieval (≈5% for R@1 and ≈2% for R@10 improvement) compared to CLIP. This could be due to the caption-generation-based pretraining objective used in BLIP-2.

Finetuning. We use the 100K balanced train set to finetune all models and evaluate on WikiDO ID and OOD test sets; these numbers are denoted as W in Table 3. The gap between ID and OOD sets have significantly grown from 3-7% in the zero-shot setting to 9-11% across all R@K. All models show >8% improvement for R@1 on the ID set, and the majority of models have >=5% improvement for most R@K (BLIP-2 being slightly behind, possibly due to stronger zero-shot). For OOD, R@K for the majority of models are under 5%, and none of them are above 10%. Vision backbones of BLIP (ViT-L), BLIP-2 (ViT-L), and BLIP-2 (ViT-G) have the same architecture as CLIP but differ in pretraining data. BLIP-2 (ViT-L) benefits from pretrained CLIP ViT-L as a robust vision encoder as compared to BLIP-2 (ViT-G), which uses Eva-CLIP [30] as the backbone.

MSCOCO & Flickr. Here, we establish that fine-tuning VLMs on MSCOCO and testing on Flickr is not a reliable test for OOD generalization. Table 4 shows finetuning on MSCOCO significantly improves Flickr performance. CLIP's zero-shot performance is lower than that of the BLIP model as the latter has already seen image-caption pairs similar to those of MSCOCO during pretraining. For BLIP, finetuning on MSCOCO and testing on Flickr is almost the same as finetuning on Flickr. Similarly, finetuning on Flickr significantly boosts zero-shot MSCOCO. Such improvements suggest that both datasets significantly overlap, making MSCOCO-Flickr a not-so-strong pair for testing generalization. A full comparison of all models is provided in Appendix B.2.

Table 5: Effect of scaling the ID data on OOD generalization using BLIP (ViT-L).

# samples	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
100K	72.6	90.8	94.6	73.7	90.3	94.2	63.6	80.8	86.1	65.9	81.8	86.9
200K	74.1	91.4	95.4	75.4	91.1	94.9	64.4	81.1	86.1	66.5	82.1	86.7
354K	76.2	92.2	96.0	76.5	92.2	95.6	64.3	80.8	86.1	66.6	82.1	86.7

Effect of scaling ID data. To find out whether the performance gap between the ID and OOD test sets can be abridged by adding more data from the ID, we train BLIP with 200K and 354K image-text

Table 6: Performance of models trained on different finetuning objectives trained on 100K split. ViT-L backbone is used for both BLIP and BLIP-2.

		WikiDO ID Test set (3K) (N=128)						WikiDO OOD Test set (3K) (N=128)					
Loss	Model	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
ITC	CLIP	82.8	95.0	97.4	81.5	94.4	96.7	73.4	87.7	91.8	72.9	88.3	91.9
	BLIP	68.9	88.3	93.0	69.0	88.4	93.1	59.6	77.5	83.4	60.4	78.1	83.7
	BLIP-2	74.4	92.9	95.7	75.6	92.3	95.7	61.2	80.1	85.6	62.5	82.0	87.6
ITC+ITM	BLIP	72.6	90.8	94.6	73.7	90.3	94.2	63.6	80.8	86.1	65.9	81.8	86.9
	BLIP-2	80.4	93.2	96.3	80.6	93.3	95.6	70.9	85.4	89.5	73.3	86.7	90.2
ITC+ITM+ITG	BLIP-2	82.1	94.0	96.4	82.5	94.3	96.7	72.1	85.9	90.3	73.6	87.1	90.3

pairs. The results in Table 5 show minimal improvements in OOD, suggesting that scaling ID data is insufficient to close the performance gap. In addition, the distribution of the largest train set is heavily biased towards only a few domains, indicating the need for more diverse data during training.

Ablations on finetuning objectives. All three models were trained with different pre-training objectives (described in Section 2). Table 6 shows the results of using different losses during fine-tuning. All models are of comparable size. Even without the use of additional objectives, CLIP proves to be very robust. In BLIP-2, the addition of ITM as an additional fine-tuning objective results in the largest R@1 improvement of $\approx$ 5-6%, and ITG slightly improves performance. The performance improvement for BLIP by adding ITM is limited to approximately $\approx$ 3-4%.

5 Discussion and Limitations

To understand why there is any improvement on the OOD test set when fine-tuned on the ID data, we first use a parser [29] to extract noun phrases from all sentences. Next, we use Grounding-DINO [31] to detect object boxes from the corresponding image and label each box with the corresponding noun chunk if they semantically represent the same thing. We recognize roughly 1M image boxes with the corresponding noun chunks in the text. We pass these image boxes to DINOv2 [32] to extract the image features. After applying K-means clustering to these embeddings with K=100, we obtain 100 meaningful clusters. To visualize this, we select 1000 boxes per cluster that are closest to the centroid. Figure 5 shows these object clusters with the difference between the objects present in OOD compared to ID. While there are a few clearly separated clusters for OOD objects, there are clusters that contain both objects in OOD and ID instances. This object overlap explains the gains in R@K for OOD after fine-tuning. While our work presents a carefully constructed test bed for OOD evaluation of VLMs, it is important to acknowledge several limitations:

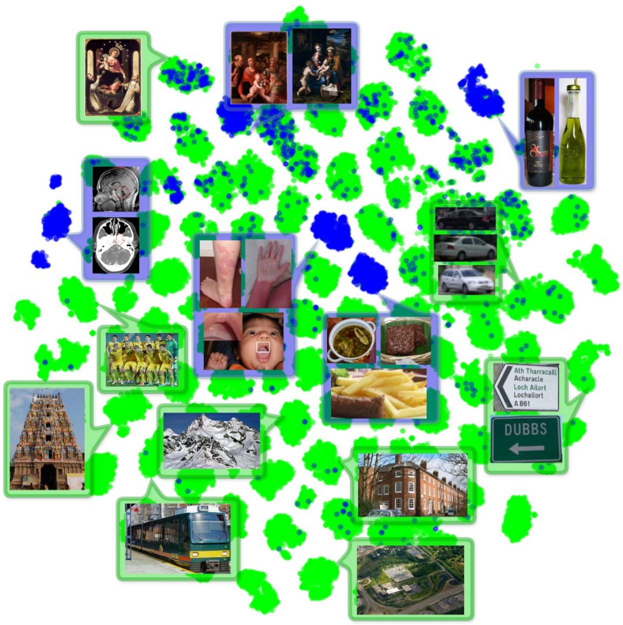

Figure 5: TSNE of 100 object clusters. Blue shows OOD objects, and green denotes objects from ID images.

Limited Scope of Image-Text Retrieval. Our primary focus has been on image-text retrieval. Although this approach can be extended to other tasks, such as generation and contextual understanding, our current evaluation framework does not cover these tasks. Since the data is extracted from Wikipedia along with the meta-data like page ID and title, it can be used for tasks like contextual image-captioning [33], image-suggestion and image-promotion [34].

Use of Topics Axis Only. In WIKIDO, we have primarily explored diversity only through the lens of topical content. There are numerous other diversity axes, such as cultural context, ethnicity, gender and religion, etc. that could provide a more robust and diverse evaluation framework. Additionally, our data is currently limited to English despite the availability of similar data in multiple languages. Expanding our evaluation to include multilingual datasets would help evaluate multilingual VLMs.

Lack of Manual Verification for Enhanced Training Set. The enhanced training set captions, due to their large size, have not been manually verified. While our test and validation sets indicate that the quality of the enhanced captions is high, the absence of manual verification could mean that some errors remain in the training data.

6 Conclusion

In this work, we introduced WIKIDO, a novel benchmark specifically designed to evaluate the out-of-distribution (OOD) generalization capabilities of vision-language models (VLMs) in the context of cross-modal retrieval. Unlike existing benchmarks, WIKIDO draws from the diverse content of Wikipedia, providing a robust dataset that includes 384K image-text pairs categorized by domain. Our benchmark includes both in-distribution (ID) and OOD test sets, each comprising 3K carefully curated and human-verified pairs, allowing for a comprehensive assessment of model performance across a wide range of topics. Our evaluations of various state-of-the-art VLMs, such as CLIP and BLIP-2, on the WIKIDO benchmark revealed insights into their OOD generalization capabilities. While BLIP-2 demonstrated superior zero-shot performance with R@1$\approx$ 66% on the OOD test set, this was notably lower compared to its performance on traditional benchmarks like MSCOCO and Flickr. Moreover, fine-tuning on WIKIDO yielded a relatively modest improvement of approximately $\approx$5% on OOD instances, suggesting inherent challenges in achieving robust OOD generalization. These findings underscore the limitations of current VLMs in handling truly OOD data. WIKIDO thus serves as an effective testbed to help develop, evaluate and guide future VLMs towards superior generalization capabilities.

References

[1] 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.

[2] 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.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[4] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[5] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018.

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *CoRR*, abs/2102.08981, 2021.

[7] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[8] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA, 2021. Association for Computing Machinery.

[9] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021.

[10] 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. Laion-5b: An open large-scale dataset for training next generation image-text models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc., 2022.

[11] 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.

[12] 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.

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[14] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

[15] D. Li, Y. Yang, Y. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5543–5551, Los Alamitos, CA, USA, oct 2017. IEEE Computer Society.

[16] Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *2013 IEEE International Conference on Computer Vision*, pages 1657–1664, 2013.

[17] Yuchen Ren, Zhendong Mao, Shancheng Fang, Yan Lu, Tong He, Hao Du, Yongdong Zhang, and Wanli Ouyang. Crossing the gap: Domain generalization for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2871–2880, 2023.

[18] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[19] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*. Jul 2011.

[20] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

[21] Hongchen Wei and Zhenzhong Chen. Improving generalization of image captioning with unsupervised prompt learning. *ArXiv*, abs/2308.02862, 2023.

[22] Roberto Dessì, Michele Bevilacqua, Eleonora Gualdoni, Nathanaël Carraz Rakotonirina, Francesca Franzon, and Marco Baroni. Cross-domain image captioning with discriminative finetuning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6935–6944, 2023.

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[24] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[25] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[26] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *International Conference on Learning Representations*, 2022.

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

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

[29] Matthew Honnibal and Ines Montani. spacy: Industrial-strength natural language processing in python, 2020. Version 2.0.

[30] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[31] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499, 2023.

[32] Maxime Oquab, Timoth'ee Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023.

[33] Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A. Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. Wikiweb2m: A page-level multimodal wikipedia dataset. *ArXiv*, abs/2305.05432, 2023.

[34] Jheng-Hong Yang, Carlos Lassance, Rafael Sampaio de Rezende, Krishna Srinivasan, Miriam Redi, Stéphane Clinchant, and Jimmy J. Lin. Atomic: An image/text retrieval test collection to support multimedia content creation. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

[35] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 3.2 and Section 4 to validate the dataset contribution and experimental claims.
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We present data scraped from Wikipedia with open licenses to evaluate current vision-language models. Our data is objectively limited to domains and does not reflect any societal elements.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] We do not present any theoretical results in the paper
   (b) Did you include complete proofs of all theoretical results? [N/A] We do not present any theoretical results in the paper
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Section 4.1 has details of codebases and datasets used. Section 1 contains the URL for proposed dataset and code.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Section 4.1 has detailed explanation of training setting and Table 2 has hyperparameter details.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Due to compute limitations, we have ran experiment with multiple seed only for the best model. Results can be found in supplementary.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Section 4.1 has details of the amount of compute used.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Section 2 contains the details of all models and data used.
   (b) Did you mention the license of the assets? [Yes] Section 3.2 contains the license of the assets used.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Section 1 has the URL for all the assets created.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We are using publicly available data at `https://wdo.wmcloud.org/`. Details are provided in the supplementary material.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Details are provided in the supplementary material.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] Used human annotators to help with caption enhancement. Annotator guidelines are mentioned in the supplementary material.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] No potential participant risks, as research is not based around human subjects.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] Yes, human annotators were paid. Details of compensation is mentioned in the supplementary material.

A Dataset Details

A.1 Sub topic distribution

Each topic comprises of many subtopics. Figure 6 shows a distribution of these subtopics.

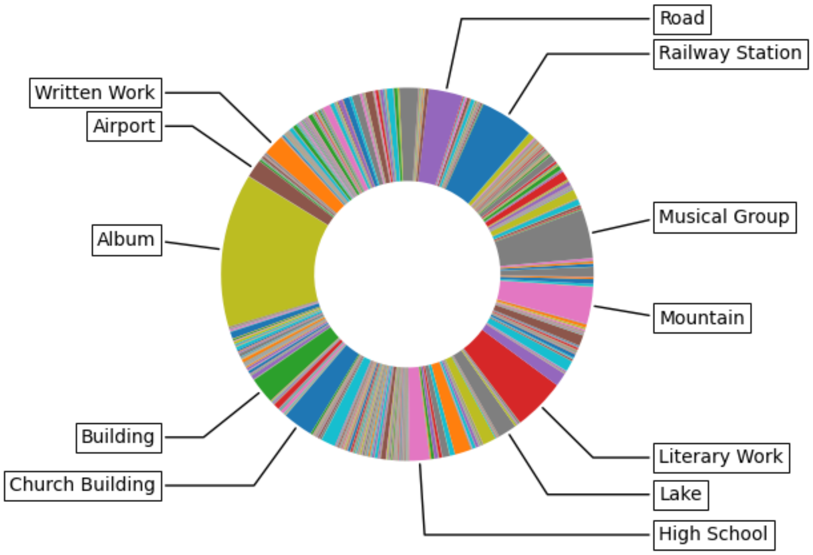

Figure 6: Pie-plot of subtopics in the dataset. Only a few subtopics are labelled to avoid clutter. Each color in the plot denotes a different subtopic.

A.2 Caption enhancement

Here is the prompt template for LLava.

> <Image> Wikipedia caption: <Caption>
> Given the image and the wikipedia caption above, give an exact and concise caption.
> Do not miss any information from the wikipedia caption.

Some examples of original and enhanced captions are given in Table 8.

Most frequently occurring noun phrases in both original and enhanced captions are given in Table 7.

A.3 MMD

The MMD distance between domains $\mathcal{D}^S$ and $\mathcal{D}^T$ can be measured according to the following equation:

$$\mathrm{MMD}(\mathcal{D}^S, \hat{\mathcal{D}}^T) = \left\|\mathbb{E}_{X\sim\mathcal{D}^S}[\varphi(X)] - \mathbb{E}_{Y\sim\hat{\mathcal{D}}^T}[\varphi(Y)]\right\|_{\mathcal{H}} \tag{1}$$

$$= \frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s} k(x_i, x_j) + \frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} k(y_i, y_j) - \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} k(x_i, y_j) \tag{2}$$

where k represents the RBF kernel and n_s, n_t represent the sample sizes in the source and target domains. For visual representations, we use pretrained ResNet-101 [35] to extract final 2048-D

Table 7: Most frequently occurring noun phrases in original and enhanced captions.

Original		Enhanced		Original		Enhanced	
Proper noun	freq.	Proper noun	freq.	Common noun	freq.	Common noun	freq.
st	7390	st	5253	view	16582	building	83017
street	6687	park	4279	station	12932	front	38729
park	6070	museum	4268	building	7086	people	35632
house	5907	new	4196	entrance	7008	man	35472
museum	5835	school	4069	church	4057	train	26638
church	5528	house	4024	century	3772	view	26415
new	4666	street	3891	side	3520	group	24159
hall	4641	church	3747	c	3511	sign	23713
school	4473	hall	3386	right	3398	image	23262
us	4357	national	3039	bridge	3349	background	23221
de	4330	de	3008	train	3339	photo	20967
bridge	4070	island	2763	part	3297	street	19211
lake	3939	city	2655	line	3149	station	17741
national	3879	lake	2601	tower	3062	painting	17497
station	3790	river	2453	background	2994	stone	16995
island	3779	john	2434	construction	2987	brick	16812
road	3651	station	2416	image	2978	tower	16508
river	3267	us	2397	site	2903	woman	16065
castle	3197	bridge	2385	end	2739	side	15880
city	3176	north	2330	railway	2635	scene	15390

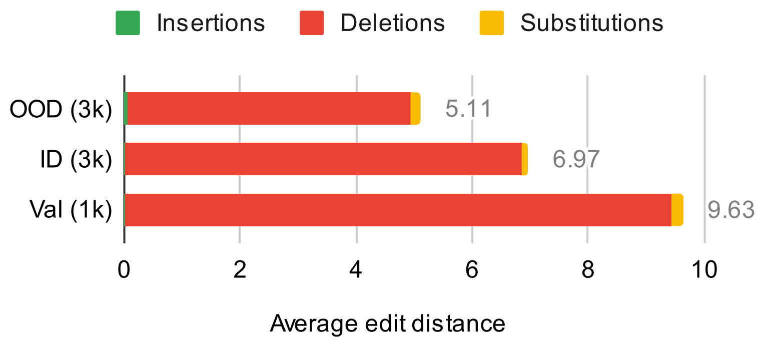

Figure 7: Average edit distance between human-corrected and hallucinated captions is shown. The contribution of various edit types is shown.

embeddings $\{v_i, v_j\}$ for images in each pair of datasets $\{\mathcal{D}^S, \hat{\mathcal{D}}^T\}$. For the semantic representation of the captions, we choose pre-trained BERT [36] to encode captions $\{q_i, q_j\}$ from pairwise datasets $\{\mathcal{D}^S, \hat{\mathcal{D}}^T\}$.

A.4 Human Verification

Since the task of human verification is to label is caption as hallucinated or not and remove any hallucinations from the incorrect captions, we expect most edits done by humans are deletions. This is validated by finding the Edit distance between the enhanced caption and the human-verified caption. Figure 7 summarizes different types of edits and shows the edit distance.

Table 8: Examples of original and enhanced version of captions

Original Caption	**Enhanced Caption**	**Image**
Rakuten Kitazawa created the first modern Japanese comic strip. (Tagosaku to Mokube no T0̆14dky0̆14d Kenbutsu,[f] 1902)	A comic strip by Rakuten Kitazawa depicts a man being attacked by a snake.	
Highest peak, Mogielica (top centre)	A mountain range with the highest peak being Mogielica.	
Ceremonial bag of the Fr0̆0edas culture	A ceremonial bag from the Frías culture is displayed in a glass case.	
The Model A Ford Museum	The Model A Ford Museum is a large brick building with a blue flag on top. The building is surrounded by a parking lot and has a sign out front.	

B Experiments

B.1 Models

CLIP We use CLIP with Vision Transformer ViT/14 with pixel resolution 336px (ViT/14@336px). This is the best model reported by the authors of CLIP. CLIP is trained on 400 million image-text pairs collected from publicly available sources on the Internet. CLIP is pretrained with contrastive learning objective (ITC) in a shared image-text space with large data making it robust to unseen domains.

BLIP introduces a unified vision-language pretraining method which jointly optimizes three objectives: image-text contrastive learning (ITC), image-text matching (ITM), and image-conditioned language modeling. BLIP is trained with 129M images, including MSCOCO, Visual Genome, CC3M, CC12M, SBU, and 115M images from the LAION-400M dataset. We perform experiment on different image encoders i.e BLIP with ViT-B and ViT-L.

Table 9: Measuring the topic gaps with MMD. Red is linguistic domain gaps over 2048-D ResNet embeddings, and green is visual domain gaps over 768-D BERT embeddings. Following is the order of topics: food (1), medicine (2), industry (3), sport and teams (4), paintings (5), religion (6), folk (7), books (8), glam (9), music creations and organizations (10), clothing and fashion (11), monuments and buildings (12), earth (13).

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	-	0.02	0.05	0.06	0.07	0.05	0.05	0.04	0.05	0.08	0.04	0.06	0.07
2	0.03	-	0.04	0.04	0.04	0.03	0.03	0.01	0.04	0.04	0.02	0.05	0.06
3	0.02	0.03	-	0.03	0.06	0.03	0.02	0.04	0.01	0.05	0.04	0.01	0.05
4	0.03	0.03	0.01	-	0.04	0.03	0.02	0.04	0.03	0.03	0.02	0.04	0.06
5	0.05	0.04	0.04	0.04	-	0.02	0.02	0.03	0.04	0.03	0.03	0.05	0.07
6	0.03	0.03	0.02	0.02	0.02	-	0.01	0.02	0.01	0.04	0.02	0.02	0.05
7	0.03	0.02	0.01	0.01	0.03	0.01	-	0.03	0.02	0.02	0.02	0.02	0.05
8	0.02	0.02	0.01	0.02	0.02	0.01	0.01	-	0.03	0.04	0.02	0.04	0.06
9	0.03	0.03	0.01	0.02	0.03	0.01	0.01	0.01	-	0.05	0.04	0.01	0.04
10	0.03	0.03	0.02	0.01	0.04	0.02	0.01	0.02	0.02	-	0.03	0.06	0.09
11	0.02	0.02	0.01	0.01	0.03	0.01	0.01	0.01	0.02	0.01	-	0.05	0.08
12	0.02	0.03	0.01	0.02	0.03	0.01	0.01	0.01	0.0	0.02	0.02	-	0.03
13	0.02	0.03	0.01	0.02	0.04	0.02	0.02	0.01	0.01	0.02	0.02	0.01	-

Table 10: Another table to empirically show coco flickr are not great for OOD generalization: zhsot, coco finetuned, flickr finetuned for MSCOCO and Flickr splits

		COCO Test set (5k) (K=128)						Flickr Test set (5k) (K=128)					
Model		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
BLIP (ViT-B)-223M	Z	70.6	90.2	94.4	56.4	80.4	87.4	87.2	98.0	99.1	78.2	94.1	96.9
	C	81.9	95.2	97.6	64.3	85.7	91.5	96.0	99.9	100.0	85.0	96.8	98.6
	F	77.9	93.3	96.6	61.3	83.7	89.9	97.2	99.9	100.0	87.3	97.6	99.0
BLIP (ViT-L)-446M	Z	73.7	91.6	95.6	58.2	81.7	88.7	89.9	98.8	99.7	80.4	94.9	97.1
	C	82.3	95.3	97.9	65.1	86.3	91.9	96.7	100.0	100.0	86.7	97.3	98.7
	F	78.9	93.7	97.1	62.7	84.7	90.6	97.4	99.8	99.9	87.6	97.7	99.0
CLIP (ViT-L)-428M	Z	57.5	80.7	87.8	36.6	60.9	71.0	86.6	98.0	99.1	67.2	88.9	93.4
	C	75.4	92.8	96.2	58.6	82.2	89.3	94.5	99.7	99.7	83.1	96.9	98.5
	F	68.9	87.6	92.7	51.8	75.8	84.0	95.5	99.5	99.9	85.0	97.7	98.9
BLIP-2 (ViT-L)-473M	Z	78.9	93.9	96.9	62.4	84.1	90.2	95.3	99.7	100.0	85.2	96.9	98.3
	C	83.2	95.9	98.0	66.1	86.6	91.8	97.1	100.0	100.0	88.3	98.0	98.9
	F	80.7	94.7	97.5	64.4	85.4	91.1	97.0	100.0	100.0	89.9	98.4	99.2
BLIP-2 (ViT-G)-1172M	Z	81.1	95.1	97.6	64.5	85.1	90.7	94.8	99.8	99.9	86.4	97.1	98.5
	C	83.9	96.5	98.2	67.0	86.8	92.0	96.7	99.9	100.0	87.2	97.3	98.7
	F	82.5	95.8	98.0	65.9	85.9	91.5	97.7	100.0	100.0	89.5	98.2	99.2

BLIP-2 bridges the modality gap of existing pretrained frozen image and text encoders using a lightweight Querying Transformer (Q-Former) which uses learnt prompt queries and a BERT-based text encoder. BLIP-2 is trained with same data and same training objectives as BLIP. Similar to BLIP, we experiment with two BLIP-2 model variants: ViT-L and ViT-G. ViT-L is pretrained CLIP ViT-L and ViT-G is pretrained Eva-CLIP ViT-G. ViT-L and ViT-G are trained on different but similar sized (400M image-text pairs) datasets (CLIP-400M vs LAION-400M).

B.2 Flickr-COCO

We finetune all models MSCOCO and test on Flickr to see the impact of OOD generalization. Conversely, we also finetune with Flickr and test on MSCOCO. It is empirically evident MSCOCO and Flickr benefit from each other. For all models, finetuning dataset helps the test dataset significantly. In fact, either finetuning with MSCOCO or Flickr gives almost equal gains in Flickr Test set performance, suggesting overlap between MSCOCO and Flickr.