A Picture is Worth More Than 77 Text Tokens: Evaluating CLIP-Style Models on Dense Captions

Jack Urbanek[*†] Florian Bordes[1,2,3†] Pietro Astolfi[1] Mary Williamson[1] Vasu Sharma[1]
Adriana Romero-Soriano[1,3,4,5]

[1] FAIR, Meta [2] Mila [3]Universite de Montreal, [4] McGill University [5] Canada CIFAR AI chair

Abstract

Curation methods for massive vision-language datasets trade off between dataset size and quality. However, even the highest quality of available curated captions are far too short to capture the rich visual detail in an image. To show the value of dense and highly-aligned image-text pairs, we collect the Densely Captioned Images (DCI) dataset, containing 7805 natural images human-annotated with mask-aligned descriptions averaging above 1000 words each. With precise and reliable captions associated with specific parts of an image, we can evaluate vision-language models' (VLMs) understanding of image content with a novel task that matches each caption with its corresponding subcrop. As current models are often limited to 77 text tokens, we also introduce a summarized version (sDCI) in which each caption length is limited. We show that modern techniques that make progress on standard benchmarks do not correspond with significant improvement on our sDCI based benchmark. Lastly, we finetune CLIP using sDCI and show significant improvements over the baseline despite a small training set. By releasing the first human annotated dense image captioning dataset, we hope to enable the development of new benchmarks or fine-tuning recipes for the next generation of VLMs to come.

1. Introduction

State-of-the-art vision-language models (VLMs) are often trained on large scale datasets such as LAION-400M [28], YFCC100M [34], or other undisclosed datasets crawled from the web. These datasets are formed by collecting images from the web and using alt-text (or other local text on the webpage) to create loose image-text pairs. These can then be filtered down trading off on quantity for quality [26, 30]. Still, recent work has demonstrated that throwing these loose captions out entirely in favor of generated captions, with enhanced quality and density, can produce improved results [10]. Other works [1, 20, 21, 38] have demonstrated that it is possible to get CLIP-level performance using a vastly reduced compute, often by throwing away portions of the data resulting in more balance between image and text modalities. However, those approaches rely on automatic pipelines which do not generate reliable and long captions that can capture rich visual details in an image. From this it appears no existing dataset has high-quality image descriptions that are tightly-coupled enough with the image to train for or evaluate a deep alignment between the two domains.

In the absence of high quality captions to evaluate VLMs, benchmarks such as ARO [42] and VL-Checklist [45] often complement image-caption pairs with hard negatives that are generated by slightly altering the initial (positive) description. Progress on these benchmarks has been rooted in training VLMs with negatives of similar construction to the tests [42] rendering the methodologies ineffective on datasets such as Winoground [35]. Recent works [22] have called the evaluation capacity of many of these benchmarks into question, given how effective language-prior-based methods perform. More specifically, given the unlikeliness of the hard negative captions in these benchmarks, a good text encoder can achieve close to 100% accuracy without looking at the images. Moreover, Bordes et al. [3] have shown that most improvements observed on ARO or VL-Checklist do not translate on simple synthetic benchmarks for which the negative caption is as likely as the positive one. Since the use of VLMs is significantly increasing, it is crucial to make sure that we have a diverse suite of reliable benchmarks to asses their abilities.

In this paper, we introduce the Densely Captioned Images dataset, a collection of 7805 images with dense and mask-aligned descriptions averaging above 1000 words each. One such example is provided in Figure 1, displaying just a subset of the collected text paired with their aligned masks. We demonstrate how to leverage this dataset to evaluate VLMs in two ways after summarizing captions

[*] Work done while at Meta
[†] Equal contribution

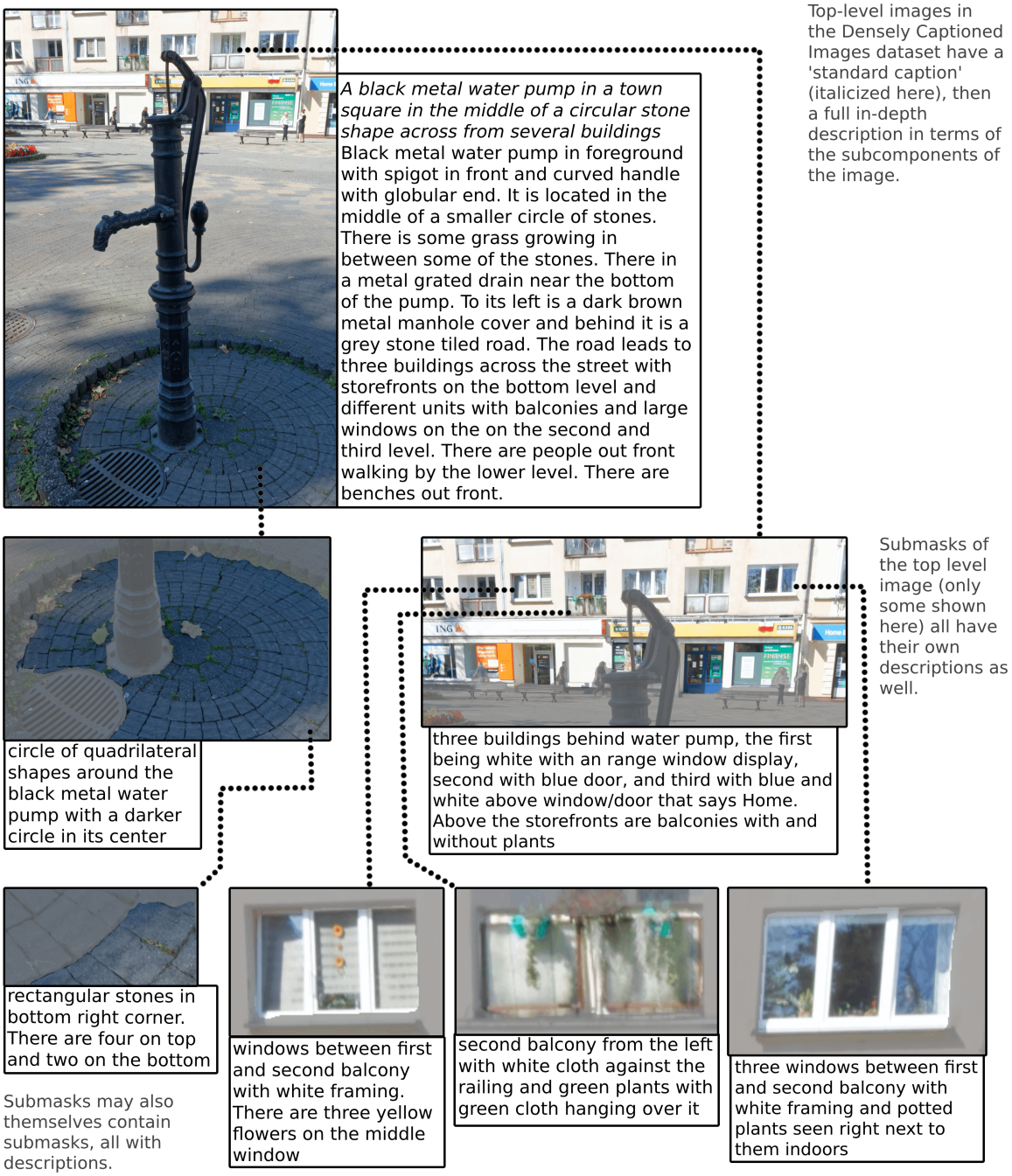

Figure 1. One example from the Densely Captioned Images dataset. Only part of the submask hierarchy is shown.

to fit into CLIP's 77 token limit, both with a negatives-based test as well as a novel matching task, referred to as *subcrop-caption matching*, that requires selecting appropriate captions for different regions of the same image. We evaluate existing baselines, and observe that no models perform well at both concurrently, and improved performance via negatives-based training comes at the cost of decreased performance on subcrop-caption matching. We also run some experiments using the summarized DCI as a fine-tuning dataset to evaluate the effectiveness of these captions for improving a model's performance on other benchmarks, and compare the efficiency per-example

to that from the automated annotation setup in DAC [10].

To summarize, our contributions are:

- We release the Densely Captioned Images (DCI) dataset, which contains dense and mask-aligned captions, alongside an LLM-summarized version (sDCI) containing captions under 77 tokens for use with current VLMs.
- We provide a new benchmark for VLMs based on sDCI to evaluate fine-grained vision-language understanding, and show that no existing model can perform well at matching captions from within one image to corresponding subsections of that image.
- We show that fine-tuning with high quality image-caption pairs is as good on ARO and VL-Checklist as fine-tuning on at least $10\times$ the automatically annotated data, and that even without utilizing explicit negatives these pairs can improve performance on VL-C-Object from 81.17% to 88.37% .

2. Related Works

The massive, loosely-labeled dataset approach that has enabled VLMs like CLIP [27] and powerful successors like BLIP2 [19], Flamingo [2], CM3leon [41], and many others, has been a clear forward step in vision-language modeling. Still recent benchmarks show that models trained in this manner display clear drawbacks in reasoning skills. Additional techniques have been proposed and adopted recently to close this gap, discussed below.

Vision-Language Datasets. Over the last decade, there have been significant dataset collection efforts connecting images and text. Earlier works focused on curating datasets by leveraging human annotations, see *e.g.*, **COCO** [8], **Visual Genome** [16], and **Flickr30k** [40]. The process resulted in high quality annotations, which were however oftentimes limited by the caption content – *i.e.*, relatively short phrases (5.1 to 10.3 words on average) grounded at image level or region level – and the data annotation scale (30k to 130k images). To increase scale, researchers gathered web-crawled data and introduced large scale datasets such as **YFCC100M** [34], which contains 100M media objects. Yet, crawling the web oftentimes results in little correspondence between image and text pairs. To reduce noise between image and text pairs, efforts such as **SBU** [24] queried Flickr and filtered the noisy results, obtaining a $\sim$1M images. Moreover, **Conceptual Captions** (CC) [30] crawled a dataset of $\sim$12M images and alt-text pairs, and included a protocol to filter noisy text-image pairs, resulting in 3M data points. Relaxing the filtering protocol allows to trade data quality for scale. Crawling alt-text also resulted in relatively short text descriptions with 10.3 words on average, which are most often grounded at image level. **Localized Narratives** [25] was introduced as a dense visual grounding dataset leveraging a multimodal annotation procedure, collecting $\sim$850k text-image pairs with 36.5 words/caption on average. **RedCaps** [9] constituted another effort yielding large scale ($\sim$12M) web-curated data by exploring alternate data sources of high quality data instead of devising complex filtering strategies. **Wikipedia-based image-text dataset** (WIT) [32] extended dataset creation efforts by gathering a multilingual dataset of text-image-pairs consisting of $\sim$11.5M images. **LAION-5B** [29] further increased the web-crawling efforts by gathering a multilingual dataset of text-image pairs, and filtered the collected data with a pre-trained CLIP [27] model. Following, **LAION-CAT** [26] reduced noisy examples from LAION-5B by filtering for caption complexity, *i.e.*, captions that do not contain any action, and for text spotting, *i.e.*, images that contain rendered text. **MetaCLIP** [39] has also been released as an open dataset for reproducing CLIP. These very large scale datasets have been successfully used to advance the state-of-the-art of VLMs.

Vision-Language Evaluation Benchmarks. Several recent advances in visual-language learning have focused on creating comprehensive benchmarks to evaluate model performance in more holistic ways. These benchmarks are instrumental in pushing the envelope of what VLM can understand and process, ensuring they move beyond superficial image-text matching towards genuine understanding of intricate relationships between visual and linguistic elements. In particular, **VL-CheckList** [45] and **ARO** [42] assess the VLM capabilities beyond average downstream task accuracy, by focusing on a model's ability to understand objects, attributes, order or relations. ARO's extensive scope, uncovers limitations in VLMs such as poor relational understanding and lack of order sensitivity. **Winoground** [35] tests models for visio-linguistic compositional reasoning by asking VLM to match two images with two captions containing the same set of words but in different orders. This task requires models to discern the meaning conveyed by the order of words, reflecting different visual scenes. Current VLMs perform only marginally better than chance, highlighting a significant gap in compositional reasoning. **CREPE** (Compositional REPresentation Evaluation) [23] evaluates two aspects of compositionality: systematicity and productivity. Systematicity is measured by the model's ability to represent seen versus unseen atoms and their compositions, while productivity gauges the model's capacity to understand an unbounded set of increasingly complex expressions. Finally, **PUG** (Photorealistic Unreal Graphics) [3] uses synthetic data to asses the compositional reasoning abilities of VLMs by progressively increasing the complexity of a given generated scene. One issue with these evaluation datasets is their frequent reliance on COCO, either directly as in ARO, or through Visual Genome as in

VL-Checklist or CREPE. It is difficult to find an evaluation dataset of sufficient scale without COCO.

Vision-Language models. Recent VLM advancements have built upon the foundational work of **CLIP** [27], which leveraged large-scale image-text pairs to jointly pre-train an image encoder and a text encoder to predict which images are paired with which texts in a contrastive learning paradigm. **NegCLIP** build upon CLIP by leveraging negative captions when training. **BLIP** (Bootstrapping Language-Image Pre-training) [18] uses a new framework that bootstraps the captions from noisy web data for both understanding and generation tasks. Its successor **BLIP-2** [19] further streamlines the process by utilizing off-the-shelf frozen pre-trained image encoders and language models, bridging the modality gap with a lightweight querying mechanism. **Clip-rocket** [12] improves VLM baselines by showing that applying image and text augmentations makes up for most of the improvement attained by prior VLMs. **Flava** [31] proposes a foundation VLM model by combining existing VLMs objectives together with auxiliary in-modality losses for the text and vision encoders. **X-VLM** [43] achieves success with a pretraining method matching sub-portions of the text to regions of the image at multiple granularities. These models introduces improvements over CLIP, focusing on efficiency, adaptability, and reducing the need for extensive labeled datasets, thereby pushing the boundaries of vision-language pre-training. The closest work to our approach is **DAC** (Densely Aligned Captions) [10], which improves with an automated LLM based pipeline the *caption quality* and *density*. By showing that DAC-enhanced CLIP models exhibit substantial gains on some benchmarks, this work underscores the critical role that caption quality and density play in the efficacy of VLMs. We build on this insight and explore how to further increase the caption quality and density by relying on human annotators, and analyze how that impacts downstream model performance.

3. Dataset Construction

The Densely Captioned Images dataset, or **DCI**, consists of 7805 images from SA-1B [15], each with a complete description aiming to capture the full visual detail of what is present in the image. Much of the description is directly aligned to submasks of the image. An example is shown in Figure 1. In the top left we see the full image of a water pump, with an associated description. The italicized section is collected as a *standard caption*, aiming to summarize the full image in about a sentence, similar to existing caption datasets. The remainder of that first description contains details about the relationship between visible entities in the image, as well as in-depth descriptions of regions that are not described as part of the submasks. All other text describing the image is associated with submasks of the image. Each submask has its own free-text label (not pictured) and description, and may also contain further submasks. Here for instance we see submasks for windows and balconies as being contained in the submask capturing three buildings in the background.

3.1. Preparation

In order to collect the data, we first select images from a random privacy-mitigated subset of SA-1B. We then procedurally extract subregions of each image to annotate, as we found in initial trials that crowdsourcing both regions and descriptions concurrently overcomplicated the task and successful annotation rate. For this process, we turn to the Segment Anything Model (SAM) [15] and adapt their standard method to extract all masks from an image.

For the extraction process, SAM usually relies on a grid of points across the entire image. In order to increase the possibility of selecting interesting regions worth annotating, we additionally apply a canny filter [4] and select random points within a radius from discovered edges. We then run SAM to detect all masks using both the grid and the near-edge points. Once the masks are returned, we establish a hierarchy of submasks by thresholding the number of overlapping pixels between two masks to determine if one is a submask of the other, or if the two masks should be joined. This helps reduce some of the noise introduced by the automatic masking process, and leaves us with a tree-like structure for the masks. Lastly, we remove any masks that are too small. We note that undergoing this process does not result in every detail of each image being selected as a candidate for annotation, and as such instances in the DCI dataset are not expected to have *complete* submask-aligned coverage of all elements one could recognize in or discuss about an image.

3.2. Collection Process

We use Mephisto [37] to host our task, pay crowdworkers to provide annotations on the dataset, and additionally run qualification steps. Workers that pass our qualifications are eligible to work on the main task which contains 3 stages:

1. Workers are provided with the whole image, and asked to provide a short description of it. This is considered the *standard caption*.
2. Workers are provided with submasks of the image, one at a time starting with the leaves of the mask tree, displaying a SAM-selected region of the image as well as an indicator for where that region comes from. They are generally asked to provide a label and complete description for the pictured region, though are allowed to mark the region as 'uninteresting' and only provide a label, or 'bad' and provide nothing. These options allow us to focus worker time on useful annotations and help capture some of the noise of the automatic selection pro-

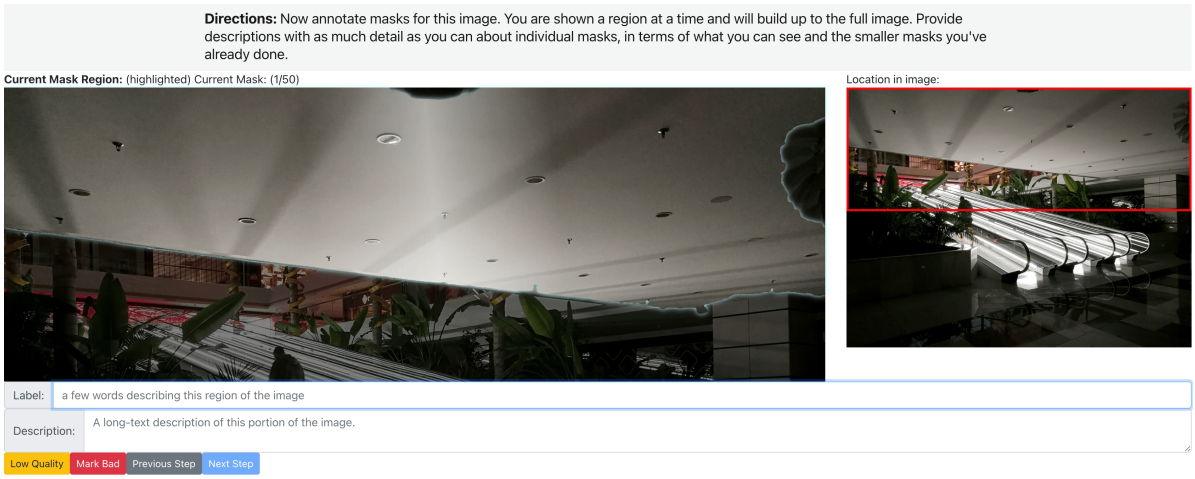

Figure 2. Annotation view for writing description for masks of the image. The masked region appears highlighted for clarity.

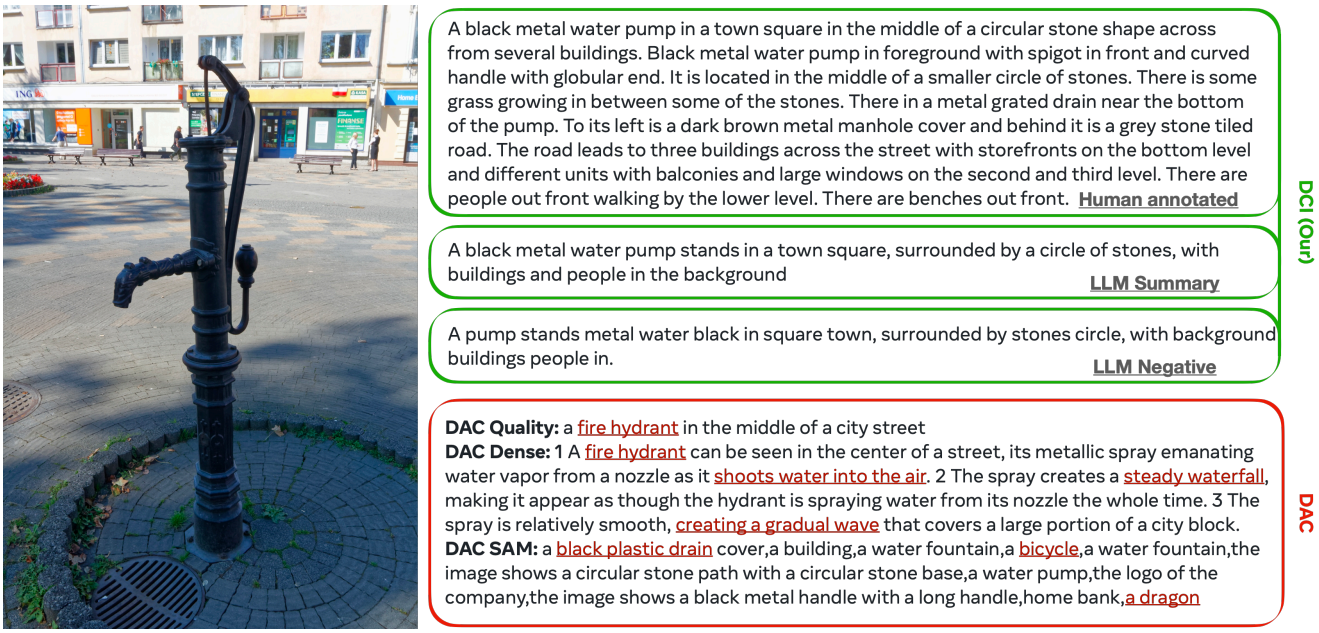

Figure 3. Example of a Llama2-generated summary and negative that comprise sDCI. Each image and submask have multiple summarizations and negatives. We also compare the caption quality between DAC [10] and DCI. In contrast to DCI that relies on human annotations, DAC used an automatic pipeline based on LLM for captioning. As we observe in this example, the DAC captions can suffer from hallucinations and miss important elements of the photo. In this work we argue that while improving automatic pipeline is an important research direction, for now the captions proposed are not reliable enough to be used to evaluate models and assess their abilities.

cess. This is shown in Figure 2. For masks that contain submasks, workers are also provided with overlays that show the regions already annotated, and are asked to annotate in terms of what has already been written.

3. After completing all the submasks, the worker is then shown the complete image again and asked to provide an overall description, paying attention to the relationship between previously annotated regions.

An in-depth description of the filtering and quality assurance process can be found in Appendix 8 while the Datasheet [13] is available in Appendix 12. Complete annotation instructions, dataset download links as well as reproducible code are avaliable on our GitHub[1]. The DCI dataset is released under the CC-BY-NC license.

[1] https://github.com/facebookresearch/DCI

3.3. Fitting DCI into 77 CLIP tokens

Ultimately, we collected an average of 1111 words (1279 CLIP tokens) per image, with a median of 941 words. This proves problematic for evaluating or fine-tuning CLIP-based VLMs given their maximum text token length of 77. Embedding pooling methods [7] to extend the effective input size for text modeling is an active research area [6, 44], and current work suggests average-pooling embeddings over these longer descriptions would be ineffective.

One possible approach would be to utilize the subsections of the image while providing the corresponding sub-caption, in a manner akin to a multi-modal multi-crop approach [5]. Still, even when considering just the 91,424 submasks, the average token length is nearly 200 per caption. We instead use the longer context capabilities of Llama2 [36] to summarize down the overall information in the image into CLIP-consumable portions. We generate multiple captions for each image and submask, using prompts that attempt to summarize down recursively until the result is in bounds. As this modification to the dataset is generated automatically, the summarizations may have introduced noise, and may not capture all of the detail in the full original captions. Summarizations also occasionally mix references or include context in a submask that isn't the main focus. Still, the summaries are fairly high quality and more dense than those found in other datasets, especially when using more than one distinct summarization per image. We also prompt the LLM to generate negatives from these summaries, achieving a set of particularly hard negatives for CLIP to evaluate. We call this version of the dataset **summarized DCI (sDCI)**. Examples of full caption, *LLM summary* and *LLM negative* are included in Figure 3 and contrasted with DAC [10] data. More detail including the prompts used can be found in Appendix 7.

Ultimately, this fitting step produces a *lower bound* on the level of vision-language understanding 'resolution' that the overall DCI dataset is capable of evaluating a model for. As newer models arise that are able to handle embedding much larger quantities of text content, it will be possible to make full use of DCI's original annotated captions.

3.4. Statistics

All-in-all the Densely Captioned Images dataset is far more dense than Localized Narratives on COCO images [25] (later referred to as LN_{COCO}) and nearly $100\times$ more dense than standard COCO captions [8]. After reducing to CLIP-bounded summaries, it still contains more text density than both. Complete details can be found in Table 1.

Here we see that the multiple-summarization method of sDCI produces fairly similar token per image values to the original dataset while keeping individual captions' token lengths in bounds for CLIP. To get Localized Narratives into the 77 token bound, we simply drop longer examples.

Dataset	Imgs	Caps	Toks/Cap	Toks/Img
DCI	7,805	7,805	1,282.09	**1,282.09**
DCI_{sub}	96,007	96,007	199.33	199.33
sDCI	8,012	87,268	49.21	536.00
sDCI_{sub}	96,007	714,630	36.60	263.01
LN_{COCO}	142,845	142,845	49.11	**49.11**
$\text{LN}_{COCO<77}$	127,456	127,456	43.70	43.70
COCO	123,287	616,767	13.54	67.74

Table 1. Comparison of DCI dataset statistics to other datasets, focusing on average CLIP tokens per image or caption. Note the **26x** difference between DCI and the previous longest annotated dataset, Localized Narratives (LN). sub denotes including submasks and their descriptions as examples, and sDCI refers to the LLM-summarized version of DCI that fits captions to 77 tokens (Sec. 3.3), while $\text{LN}_{COCO<77}$ simply drops examples longer than 77 tokens ($\sim 10.8\%$).

4. Evaluating VLMs with summarized DCI

4.1. Methodology

Using the 7805 images in the summarized Densely Captioned Images (sDCI) dataset, we construct a few different evaluations. As noted above, the ability to select multiple submasks from the same image and include them in the same batch allows us to create a CLIP-style test, wherein the model can evaluate a full batch of images and captions and score correctly which caption belongs to which image. As we provide models with a crop around the selected masks, we call this *Subcrop-Caption Matching (SCM)*, and we use a batch size of 8. We can run against our LLM-generated negatives as well. Given that LLM-summarization has provided us with multiple captions and negatives per image and submask, we supply the first unless noted otherwise. With this in mind, we construct 6 evaluations as follows:

[All SCM]: Group each image with their subcrops, alongside one summarized caption per subcrop. Then use the model to find the most likely caption associated to each subcrop. This test measures the ability of the VLM to distinguish between the different parts that compose an image.[2]

[All Neg]: Select one LLM summarized caption and the corresponding LLM-generated negative for each image and subcrop. Score a model on its ability to distinguish between the positive and negative.

[All Pick5-SCM]: Use the same setup as **All SCM**, but rather than using only one caption per subcrop, we use 5 LLM generated captions per subcrop. We score a model as succeeding only when the worst-scoring positive caption

[2]Since we used sDCI to fit current models token length, it is possible that some of the summaries remove the information that make possible to distinguish between the captions. Ideally this test should be performed on the non-summarized version once VLMs can handle 1000+ tokens.

Model	All		All Pick5		Base	All
	SCM	Neg	SCM	Neg	Neg	Hard Negs
CLIP Baseline [27]	40.06%	60.79%	11.21%	24.06%	67.56%	41.34%
NegCLIP [42]	**43.35%**	56.00%	**13.22%**	4.82%	76.69%	50.84%
BLIP [18]	39.13%	54.02%	10.73%	5.51%	63.41%	53.23%
Flava [31]	38.08%	47.99%	8.01%	9.82%	11.6%	45.59%
X-VLM [43]	38.45%	53.46%	10.96%	5.10%	44.29%	52.42%
DAC_{LLM}[10]	37.45%	81.71%	8.13%	37.84%	**90.56%**	71.21%
DAC_{SAM} [10]	37.90%	**84.17%**	6.70%	**39.94%**	89.66%	**73.61%**

Table 2. sDCI test result: We compare existing baselines on our Subcrop-Caption Matching (SCM) and negatives tests. Additional results are available in Table 10 in the Appendix. We note our best model fine-tuned on sDCI from section 5 achieved 64.02% and 31.60% on a held-out test of **All SCM** and **All SCM Pick5** respectively, setting an upper bound for model performance.

scores higher than the best-scoring caption of any other image in the batch. This test evaluates if the representation space is structured such that captions belonging to a specific image are closest to the target image in the space.

[All Pick5-Neg]: Use the same setup as **All Neg**, but rather than using one caption, we use 5 LLM summarized captions for each image and subcrop. If any of these captions score worse than the negative, the model fails the example.

[Base Neg]: Using only the 7805 base images without subcrops, evaluate the model's ability to distinguish between an LLM generated caption and its corresponding LLM-generated negative. Note, this is a strict subset of **All Neg**, though these captions are on the longer side on average and cover a different distribution.

[All Hard-Negs]: Using the same setup as **All Neg**, but rather than using a single negative, use the negative across all LLM-generated negatives that CLIP scores highest.

4.2. Results

We compare in Table 2 the sDCI performances given by different state-of-the-art models: CLIP [27], NegCLIP [42], BLIP [18], Flava [31] and X-VLM [43]. Additional experiments on different architectures and pretraining datasets are available in Table 10 (see Appendix). The CLIP baseline starts at 40.12% on **All SCM** and 60.63% on **All Neg**. The only model to improve over CLIP on SCM tasks is NegCLIP, which follows the fact that the hard image negatives that NegCLIP is trained on provide the most similar task to what we test of any of these models. None of the models trained without an explicit CLIP-loss component outperform CLIP on SCM tasks, but DAC ultimately performs the worst.

Performance on the Pick5 variations of each task follow the trends of the standard performance. Performance on **Base Neg** for Flava point to a weakness in comparing longer text examples, given the significant drop from 47.99% to 11.6% that is not demonstrated in other models.

Interestingly, models trained absent of CLIP (BLIP, Flava, X-VLM) experience a far less noticeable drop in performance between **All Neg** and **All Hard Negs**. This validates that sDCI's CLIP-hard negatives are not simply a higher proportion of 'impossible' negatives, but rather capture some underlying trait about the negatives that CLIP models and their descendants all struggle with.

None of the models presented perform well across all of the sDCI test set. Given each of the CLIP-style models have some kind of advantage on this test set due to being trained on some objective that sDCI directly evaluates, we expect that the BLIP, Flava, and X-VLM scores are somewhat representative for existing state-of-the-art models' true performance on this test set.

5. Using summarized DCI as fine-tuning dataset

To evaluate the use and difficulty of the sDCI dataset for training, we fine-tune state-of-the-art models with it. In particular, we use a ViT/32B CLIP model in all of our experiments, requiring use of the CLIP-bounded version of our dataset. We split sDCI into 7800 train, 100 validation, 112 test samples for this purpose. We use a training batch size of 32 and a learning rate of `5e-5` for all experiments, and run for up to 10 epochs. We train using both standard CLIP loss as well as an additional Negatives loss component, which follows the 'text negative' of NegCLIP [42]. Given the tiny size of our finetuning sets relative to the 400M pretraining images, we use LoRA [14] to reduce the trainable parameters. We train a model with and without negatives loss.

In order to make good use of the multiple summarized captions we have per image and submask, we randomly select one to be used in each individual epoch. We call this method *Pick1*. We describe this method and other ablations we attempted in more detail in Appendix 9.

We follow the experimental setup of DAC [10] by evaluating our sDCI fine-tuned CLIP on the ARO and VL-Checklist benchmarks. We compare to DAC directly as it is the most similar work to ours in attempting to increase

Model	ARO				VL-Checklist		
	VG-R	VG-A	COCO	FLICKR	Object	Attribute	Relation
$sDCI_{P1}$	76.23%	67.56%	88.58%	91.30%	80.71%	68.69%	70.12%
$sDCI_{P1NL0}$	57.34%	61.98%	39.36%	44.62%	*88.37%*	70.42%	61.28%
$DAC_{LLM_{10,000}}$	61.53%	63.89%	46.28±1.5%	59.41±1.9%	66.90%	57.4%	56.96%
$DAC_{LLM_{100,000}}$	61.0%	63.6%	48.2%	61.42%	66.87%	57.22%	57.18%
$DAC_{LLM_{500,000}}$	60.1%	63.8%	50.2%	61.6%	66.54%	57.39%	56.77%
$DAC_{LLM_{3,000,000}}$	**81.28%**	**73.91%**	**94.47%**	**95.68%**	87.30%	**77.27%**	86.41%
$DAC_{SAM_{3,000,000}}$	77.16%	70.5%	91.22%	93.88%	**88.50%**	75.83%	**89.75%**
CLIP Baseline [27]	59.98%	63.18%	47.9%	60.2%	81.17%	67.67%	61.95%
BLIP2 [19]	41.16%	71.25%	13.57%	13.72%	84.14%	80.12%	70.72%
NegCLIP [42]	81%	71%	86%	91%	81.35%	72.24%	63.53%
SVLC [11]	80.61%	73.03%	84.73%	91.7%	85%	71.97%	68.95%

Table 3. sDCI fine-tuned CLIP performance against the ARO and VL-Checklist benchmark. We compare CLIP fine-tuned with sDCI against models fine-tuned using DAC captions. Since the DAC dataset contains 3M images whereas sDCI contains only 7805 images, we performed an ablation of the number of training images used in the DAC dataset. In this instance, $DAC_{LLM_{10000}}$ refer to fine-tuning CLIP using only 10,000 images from DAC. We plot the mean across 5 different seeds and display the standard deviation when it is above 1% accuracy. We observe that training on sDCI lead to significant improvement in comparison to DAC for a comparable number of examples.

caption density. As noted in Figure 3, these automatically generated captions are generally noisy. As DAC is using 3M images for fine-tuning, we performed a small ablation on the number of DAC images to use for fine-tuning to be similar to our base image count (10,000 compared to our 8,012), or to our full mask count (100,000 compared to our 99,445).

5.1. Results

In Table 3, we show that, while the DCI Pick1 model trained with negatives loss ($sDCI_{P1}$) does not reach the performance of DAC models trained on 3M images, it does improve over the CLIP baseline on most metrics[3], and outperforms some baselines trained on more data. $sDCI_{P1}$ does however outperform both sample-limited ablations of DAC, suggesting that a small number of highly aligned image to dense text pairs are more effective for training models than larger quantities of more loosely aligned or sparse data. Unsurprisingly, the version trained without negatives loss, $sDCI_{P1NL0}$, does not improve across most benchmarks, and even somewhat degrades when compared to the CLIP baseline.[4] Of note however is the significant bump in VL-Object, alongside some improvement to VL-Attribute. Improvements here suggest that the sDCI dataset successfully includes more object, and to a lesser degree attribute, information than the captions in the source dataset for CLIP. It does, however, point to a limitation of using the LLM summarizations and not incorporating mask information, as relational information is sometimes lost.

[3]The decreased performance on VL-Object may be explained by our LLM-generated negatives not closely covering the test set negatives.

[4]The degradation is likely due to the distribution shift and small sample size, given the training objective is the same as CLIP.

6. Conclusion and Future Work

We introduce the Densely Captioned Images dataset, and display clear use for it as a evaluation benchmark. We also show initial potential for using the dataset for fine-tuning. Given that in order to evaluate today's models on DCI we had to reduce the size of the text to only 77 tokens, DCI should prove to be useful for a longer period of time as models that are able to consume and utilize larger amounts of text context become the norm. We envision that in those cases the full human annotated captions without length reduction would be provided. Today's context size limitation also prevented us from fine-tuning existing models on the highly aligned text-image data within DCI, as existing models don't have enough context size to handle the full text, but the dataset isn't nearly large enough to pre-train a new set of models that could use the full text. It could be relevant to treat developing highly aligned text-image datasets in a similar manner to that used in machine translation for low-resource languages, which run into a similar issue with cost and difficulty to collect. This area of work has relied on automated methods such as bitext mining [33] to bootstrap up from an initial set of expertly collected examples, which DCI may already provide the foundation for. Further, we haven't attempted to incorporate the pixel-level masks that the dataset has in any of our experiments, instead opting to use crops around the masks to retain parity with our test set. This dataset is unique for both the extreme density and high degree of alignment present, and in this introductory work we've only scratched the surface of using this information to its fullest extent.

References

[1] Amro Kamal Mohamed Abbas, Kushal Tirumala, Daniel Simig, Surya Ganguli, and Ari S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023. 1

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. 3

[3] Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari S. Morcos. Pug: Photorealistic and semantically controllable synthetic data for representation learning. In *Advances in Neural Information Processing Systems*, 2023. 1, 3

[4] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. 4

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021. 6

[6] Ilias Chalkidis, Xiang Dai, Manos Fergadiotis, Prodromos Malakasiotis, and Desmond Elliott. An exploration of hierarchical attention transformers for efficient long document classification, 2022. 6

[7] Qian Chen, Zhen-Hua Ling, and Xiaodan Zhu. Enhancing sentence embedding with generalized pooling, 2018. 6

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. 3, 6

[9] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: web-curated image-text data created by the people, for the people, 2021. 3

[10] Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Dense and aligned captions (dac) promote compositional reasoning in vl models, 2023. 1, 3, 4, 5, 6, 7

[11] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision&language concepts to vision&language models, 2023. 8

[12] Enrico Fini, Pietro Astolfi, Adriana Romero-Soriano, Jakob Verbeek, and Michal Drozdzal. Improved baselines for vision-language pre-training. *Transactions on Machine Learning Research (TMLR)*, 2023. 4

[13] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. Datasheets for datasets, 2021. 5, 13

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 7

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 4, 13

[16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. 3

[17] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. Elevater: A benchmark and toolkit for evaluating language-augmented visual models, 2022. 5

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. *CoRR*, abs/2201.12086, 2022. 4, 7

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 3, 4, 8

[20] Xianhang Li, Zeyu Wang, and Cihang Xie. An inverse scaling law for clip training, 2023. 1

[21] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking, 2023. 1

[22] Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Visualgptscore: Visio-linguistic reasoning with multimodal generative pre-training scores, 2023. 1, 4

[23] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally?, 2023. 3

[24] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2011. 3

[25] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives, 2020. 3, 6

[26] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6967–6977, 2023. 1, 3

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 3, 4, 7, 8

[28] 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. 1

[29] 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, 2022. 3

[30] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics. 1, 3

[31] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *CVPR*, 2022. 4, 7

[32] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Mike 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)*, 2021. 3

[33] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022. 8

[34] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100m. *Communications of the ACM*, 59(2): 64–73, 2016. 1, 3

[35] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022. 1, 3

[36] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 6

[37] Jack Urbanek and Pratik Ringshia. Mephisto: A framework for portable, reproducible, and iterative crowdsourcing, 2023. 4

[38] Hu Xu, Saining Xie, Po-Yao Huang, Licheng Yu, Russell Howes, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Cit: Curation in training for effective vision-language data. *arXiv preprint arXiv:2301.02241*, 2023. 1

[39] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data, 2023. 3

[40] Peter Young, Alice Lai, Micah Hodosh, and Julia 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. 3

[41] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, Candace Ross, Adam Polyak, Russell Howes, Vasu Sharma, Puxin Xu, Hovhannes Tamoyan, Oron Ashual, Uriel Singer, Shang-Wen Li, Susan Zhang, Richard James, Gargi Ghosh, Yaniv Taigman, Maryam Fazel-Zarandi, Asli Celikyilmaz, Luke Zettlemoyer, and Armen Aghajanyan. Scaling autoregressive multi-modal models: Pretraining and instruction tuning, 2023. 3

[42] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. 1, 3, 7, 8, 4

[43] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 4, 7

[44] Hang Zhang, Yeyun Gong, Yelong Shen, Weisheng Li, Jiancheng Lv, Nan Duan, and Weizhu Chen. Poolingformer: Long document modeling with pooling attention, 2022. 6

[45] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations, 2023. 1, 3