7. LLM Summarization and Negatives

In this section we discuss the prompting strategies we used on the LLaMA 70B-chat model to produce our summarized captions and negatives.

7.1. Summarization

In order to generate multiple summaries for each image and subsection of the image, we used the following method. First, we attempted to create a first-pass summary of either the full image or of masks with the following prompts:

Full Image: `You are given a full-text description of an image. You should summarize it into about 65 words, being sure to include as much salient visual information as possible given the 65 word constraint, especially information from the start of the original description. The new description should apply for the original image. Respond with only the summary, in one line.`

Submask: `You are given a description of part of an image. You should summarize it into a single line no longer than 65 words, being sure to include as much salient visual information as possible given the 65. Don't include any details not present in the provided description. Respond with only the summary, in one line.`

This provided us with once caption per image and submask, which is used universally as the *first* caption for each. We, however, wanted to have more options to capture potential details that may have been missed by the first summarization. To that end we repeated the following prompt until we had at least 6 captions per.

Multi-caption Prompt: `You are providing descriptions of an image. The goal is to create a list of summarized descriptions that all accurately describe the same image, but may pay attention to slightly different details. A complete description will be provided. Complete 5 entries in this list, each a few sentences long. Provide in the format:\n1. <description>\n2. ...`

7.2. Negatives

To generate negatives, we came up with three distinct prompts to produce negatives when given a summary. All of the negatives are generated from the *first* summarized caption. Each attempted to raise different kinds of reasoning, either for simple basic edits, changes to structural information, or reconstructing a new sentence from the same bag of words.

basic: `You are given the description for an image. You should provide a mostly similar description, changing the original one slightly, but introducing enough significant differences such that the two descriptions could not possibly be for the same image. Keep the description length the same. Provide just the updated description.`

layout: `You are given the description for an image. You should provide a mostly similar description, with one part of it changed in a way that visibly alters the structure, layout, or content of the image. The change should introduce enough difference such that the two descriptions could not be for the same image. Keep the description length the same. Provide just the updated description.`

swaps: `You are given a description. Selecting ONLY from the same words, construct a random response with the same words in a completely new order. The new description should not remain accurate to the first. Try to use all of the words from the original description, and keep the length the same, but use the words nearly randomly such that the scene makes less sense. Try to pair nouns and adjectives differently than in the original.`

Ultimately, in initial ablations we found that the **swaps** prompt produced the best model performance, so we use these negatives as the default.

7.3. Common Issues

One main issue with LLM-based summarization is that LLaMA2 was not particularly effective at producing answers within the specified word boundary. Thus, we additionally use prompts that attempt to get the model to shorten the output response so long as it is somewhat close in length. If after enough tries this didn't work, we'd return to the original prompt.

Reduce: `You are given the description for a scene. Summarize and rephrase the description such that the new description is a few words shorter,`

`but retains key information.`

Another issue is that the chat fine-tuning on the LLaMA2 model made it frequently preface the response with additional affirmative text, like "sure I can do that...". Often this could be removed by simply removing the everything before the last line break.

7.4. Output Text Examples

Below we provide an example of a summarization for a base image (which includes details from all of the subimages during the prompting stage) as well as a summarization for one submask. We also share one example of each type of type of negative for both.

7.4.1 Base Image

Description: The upper floor of an ornate design building with the words "Dem Wahren, Schonen" visible on its top. There's a light source with five light bulbs decorating each pillar, at least two marble statues and multiple arches between each pillar. An ornate building with at least two floors. There's four small pillars with three black guard rails and four light stands on the pillars. There's two marble statues on either end of the upper floor. The upper floor also has five arches with decorative symbols above them. There's six visible tall supporting pillars on either side of the arches. On the top is the visible words: Dem Wahren Schoenon. On the bottom floor is three arches with a black light hanging from each of them. The inner arches and interiors are somewhat visible. In front of the building is a black streetlight. ...(submask text omitted)...

Selected Summaries (3/8): *(1)* An ornate building with four pillars, three arches, and a decorative facade. The building has a upper floor with five light bulbs, two marble statues, and multiple arches. The lower floor has three black guard rails and a streetlight. The image includes various decorative elements such as swirling branches, leaves, and faces.
(2) The image shows the upper floor of an ornate building with five pillars, each with a light source and decorative elements. The building has a grand entrance with multiple arches and marble statues. The words "Dem Wahren, Schonen" are visible on the top of the building.
(3) The image highlights the right side of the upper floor, focusing on the decorative elements on the pillars and the arches. The right most pillar has a thinner middle and wider bottom, and the top pillar decoration has swirly branches and a spread out decoration like a opened flower. The image also shows the inner pillars and the black guard rails.

`basic` negative: A grand building with four pillars, three arches, and a detailed facade showcasing mythical creatures and a clock. The building's upper floor boasts five light bulbs, two marble statues, and several arches, while the lower floor features three black guard rails and a streetlight that illuminates the vibrant flowers and trees outside.

`layout` negative: A grand building with intricate details, including four pillars, three arches, and a decorative facade now covered in vines and moss, giving it a weathered and natural appearance. The upper floor features five light bulbs, two marble statues, and multiple arches, while the lower floor has three black guard rails and a streetlight.

`swaps` negative: A facade with five arches, three pillars, and four light bulbs. The building has a decorative upper floor with two marble statues, multiple black guard rails, and a streetlight. The lower floor has three arches and various decorative elements such as swirling branches, leaves, and faces.

7.4.2 Submask

Description: *Front black street light* - The black street light in front of the ornate building. It has a thin support beam holding it up with little lines going down it. There's curved branches on the beams supporting the lights. Each light has the same design, with little pointy spires or points on its top tip. There's about 6 visible lights on this street light. ...(submask text omitted)...

Selected Summaries (3/8): *(1)* A black street light stands in front of an ornate building, supported by a thin beam with curved branches. Six lights with pointy spires are visible, with the closest one featuring clear glass panes and a lantern-like shape.
(2) The front black street light is decorated with curved branches and has a lantern-like shape. The light closest to the image has a clear glass pane with ridges on the edges and a small decorative circle above it.
(3) This image highlights the unique design of a black street light, with a lantern-like shape and pointed spires on top. The light has a clear glass panel with ridges and a decorative circle, and is supported by a thin beam and curved branches. The image also shows the ornate building in the background.

`basic` negative: A white street light stands in front of a modern skyscraper, supported by a thick concrete pillar. Five lights with sleek, angular designs are visible, with the closest one featuring frosted glass panes and a cylindrical shape.

`layout` negative: A black street light stands in front of a modern skyscraper, supported by a thick concrete pillar. Six LED lights with sleek, angular designs are visible, with the closest one featuring a frosted glass cover and a cylindrical shape.

`swaps` negative: A lantern-like shape stands in front of an ornate building, supported by a clear glass panes with curved branches. Six pointy spires are visible, with the closest one featuring a thin beam and black street light.

8. Crowdsourcing Methodology

Extending from the overall description for the main task in Section 3.2, we here provide an overview of the precursor qualification tasks, as well as some additional details of the preparation and packaging process for the dataset. The complete set of instructions, as well as code to be able to directly reproduce the collection for this dataset, is available on the project GitHub[5].

Note, we do not include the code for our initial version of the collection task, wherein workers were asked to both select the regions of the image that were worth annotating *and* annotate them in the same pass, as these proved both incredibly time-consuming and often lower quality than using model-based mask generation and allowing workers to filter low-quality masks.

8.1. Quality Assurance

High-quality datasets rely on getting solidly-performing crowdworkers, which nowadays can prove to be an adversarial task initially. To remedy this situation, we set up a multi-stage process wherein workers could complete precursor tasks (for which they were compensated) that we could use to determine eligibility in the main task pool.

Stage one of the task asked a few questions about a preset image, wherein we asked workers to provide a few sentences describing the image, then note a few things in the image they might include descriptions of if they were to need to write 1000 meaningful words about the image. After filtering out answers from bots, we manually reviewed answers for quality on the provided description, as well as having included any of a few details in the image that we felt were hard to notice on first pass in either the description or list of additional things they might describe.

Workers with solid fluency in English and solid attention to detail were moved into stage two, wherein they had access to the full task and were eligible to complete it three times. In this stage we evaluated responses for a general understanding of the more complex task interface, and allowed workers who completed this stage or only made minor mistakes to the full tasks. Workers who did make minor mistakes were given feedback to help understanding.

In the last stage, work was audited regularly from each worker to provide feedback about description quality, proper use of disqualifying bad masks, and overall tradeoff between time spent and work completed. We used this information to limit over-contribution from single individual, filter out a few workers that provided decreasing quality over time, and bonus workers who were slower than our target pace but providing exceptional quality.

8.2. Instruction Details

Before entering the task, the worker was walked through an example task to familiarize them with the interface and the goal of the task. We additionally provided a few specific scenarios to anticipate common questions, like what to do when two masks are the same, or how to deal with writing for a mask that contains a mask that was already done, or what to do on images that were much simpler than the norm. A complete list of these instructions is available with our released code.

8.3. Worker Metrics

Over the course of a month, we made the first qualification round eligible to a large cohort of workers on Mechanical Turk, and fielded over 800 submissions. Of these, roughly 250 workers made it through to the second task.

In the second stage, we hand-reviewed around 600 tasks (up to 3 from each worker from the first stage), and ultimately moved a group of 120 workers to the full task. Of these, around 80 were regular contributors over the course of our collection.

During the main task, we enacted controls to ensure that no worker provided more than 10% of the currently collected data at any time. We used automated metrics around words per image, words per minute, and unique words per image to act as an overview of worker quality, and hand-reviewed examples that were far outside of our expected bounds for these metrics.

9. Extended Ablations

In order to evaluate ideal use techniques for the Densely Captioned Images dataset for the purpose of fine tuning, we run some ablation experiments and compare to performance on fine tuning on COCO and Localized Narratives.

9.1. DCI Fine-tuning ablation methods

9.1.1 PickN Caption Training

In order to make use of all of the available captions for each image, we pick some number of captions from those available. For $N = 1$, this just results in selecting one of the captions for each image randomly in each epoch. For $N > 1$,

[5] https://github.com/facebookresearch/DCI/tree/main/reproduction/crowdsourcing

we instead provide multiple captions during *both* CLIP loss and negatives loss calculations. In these circumstances, loss is calculated between the *worst* positive and the *best* negative. As each image is supposed to have unique and high-quality captions across the whole set, the expectation here is that any caption for one image should score better than any caption for another.

9.1.2 Image-based Batching

When training with $sDCI$ submasks, we have the opportunity to provide the model with exceptionally difficult negatives during the CLIP loss calculation, namely other submasks from the same image. We call this ablation $ImgGroup$ and run it for each $sDCI$ model to determine the impact that hard negatives like these have on CLIP training.

9.1.3 Negatives Loss

In order to evaluate the impact that over-weighing negatives during train time has on model performance, we launch one job with 9 times the negatives loss used in our standard experiments. This attempts to evaluate how easily the negative construction techniques in these standard benchmarks could be 'gamed'.

9.1.4 Negatives Selection

While we have LLM-generated negatives readily available for the $sDCI$ dataset (as described in Section 7, we also wanted to compare to negatives for the LN and $COCO$ datasets. For this we used a `spacy`-based swapping technique similar to NegCLIP [42] or DAC [10], wherein noun phrases, verbs, and adjectives were randomly swapped inside of a given caption to create a negative. We used these `spacy`-swaps for all $COCO$ and LN runs, but we also include `spacy`-swap ablations for DCI-trained models to compare with the LLM-generated negatives.

9.2. Datasets and Training methods

Overall we use the same training parameters outlined in Section 5. We use the following five datasets in our ablations:

1. $sDCI_{sub}$: All of the complete images and subimages with LLM generated captions (referred to above as $DCI_{sub<77}$).
2. LN: All images and captions from the COCO subset of Localized Narrations with a CLIP token count under 77 (referred to above as $LN_{COCO<77}$).
3. LN_{7805}: The first 7805 images of LN, used to be a same-size comparison to DCI-trained models.
4. $COCO$: All images and captions from the COCO 2017 set.
5. $COCO_{7805}$:The first 7805 images of $COCO$, used to be a same-size comparison to DCI-trained models.

For each sweep we select the best model as determined by highest score on validation metrics from the same dataset used for training (for instance using $COCO$ valid for a model trained on $COCO$ train).

For $sDCI$, we split the test set into a training set of 7599, a valid set of 98, and a test set of 108. All metrics below are reported on the 108 image test set.

9.3. Results

9.3.1 Aggregate DCI Ablations

A complete table of our sDCI ablations can be seen in Table 4, however it is more dense information than is likely useful in seeing overall trends. Instead, we aggregate over the different ablation methods.

In Table 5, for ARO and VL-C we observe that generally constructing batches from the entire dataset rather than from masks in the same image is better for performance.

We also see that using *Pick1* is the most effective method for making use out of the LLM-summarizations of captions we've created, as it allows the model to see more of the related text (better than *first*) without potentially penalizing it for situations where two related overlapping captions are provided at the same time (which may be an issue with *Pick5*).

We do observe that using high negatives loss is very effective at gaining additional performance on most of these metrics, however this does not include VL-Object and VL-Attribute, likely due to our method of constructing negatives not correlating very well with the types of negatives created in these tests. While this provides great scores, it mostly just provides evidence towards the aforementioned issues with these types of evaluations [22].

In Table 6, we observe a slightly different story. Here $ImgGroup$ appears to be the best method for selecting images, only performing worse on the Base Neg test. This is expected though, as the test set is constructed with sequential examples in a manner similar to the $ImgGroup$ ablation, which doesn't have any effect on Base Neg given there are no submasks to deal with in that test.

No individual setting for negatives loss performs best on all metrics, however it is unsurprising that negatives loss 0 results in the highest performance on Subcrop-Caption Matching tasks (given their similarity to the CLIP-style learning objective), and not as well as negatives-trained models on negatives tasks.

The PickN ablations are somewhat surprising, as *first* captions generally performed the best overall, and Pick1 outperformed training on Pick5 when testing on Pick5.

Conditions			ARO				VL-Checklist			sDCI All		All Pick5		Base	All
Capts	NL	Batch	VG-R	VG-A	COCO	FLICKR	Object	Attribute	Relation	SCM	Neg	SCM	Neg	Neg	H-Neg
First	9	Rand	75.34%	62.72%	84.33%	87.86%	76.99%	67.77%	68.27%	51.71%	**97.20%**	8.69%	79.41%	**98.21%**	88.30%
First	9	Group	74.82%	58.26%	85.05%	88.70%	69.44%	68.65%	68.93%	54.38%	96.79%	9.44%	79.69%	94.64%	**88.65%**
Pick1	9	Rand	76.04%	65.51%	86.69%	89.72%	73.35%	65.04%	69.48%	44.12%	95.55%	12.59%	93.57%	93.75%	82.97%
Pick1	9	Group	76.41%	63.71%	82.89%	88.04%	72.68%	65.44%	63.33%	50.14%	95.62%	19.36%	94.05%	92.86%	84.06%
Pick5	9	Rand	**78.61%**	65.65%	86.83%	**91.40%**	68.61%	66.10%	**71.82%**	39.33%	96.10%	7.46%	**94.94%**	**98.21%**	85.64%
Pick5	9	Group	75.65%	**68.92%**	85.45%	90.26%	70.17%	65.09%	63.80%	43.23%	96.17%	11.56%	94.32%	96.43%	83.79%
First	1	Rand	64.00%	60.79%	83.93%	87.46%	76.73%	68.75%	63.53%	54.17%	96.17%	14.64%	75.92%	96.43%	85.91%
First	1	Group	73.09%	62.60%	79.54%	84.84%	72.64%	69.37%	61.68%	61.76%	95.76%	19.63%	76.81%	97.32%	85.43%
Pick1	1	Rand	76.23%	67.56%	**88.58%**	91.30%	80.71%	68.69%	70.12%	50.48%	94.39%	19.02%	88.85%	95.54%	81.40%
Pick1	1	Group	64.97%	65.05%	80.32%	87.62%	75.01%	66.70%	60.75%	58.14%	94.60%	26.81%	89.88%	96.43%	83.58%
Pick5	1	Rand	72.56%	61.32%	78.00%	83.68%	77.64%	66.74%	66.65%	49.93%	94.19%	18.60%	86.87%	95.54%	80.78%
Pick5	1	Group	57.96%	61.49%	78.27%	84.06%	64.80%	63.05%	61.73%	51.44%	94.66%	19.97%	90.63%	93.75%	81.81%
First	0	Rand	55.61%	55.83%	29.60%	39.22%	80.70%	67.71%	62.10%	57.25%	74.21%	22.23%	30.30%	80.36%	67.58%
First	0	Group	49.91%	59.04%	41.68%	50.90%	73.11%	67.35%	60.03%	63.13%	75.44%	27.84%	33.11%	81.25%	69.49%
Pick1	0	Rand	57.34%	61.98%	39.36%	44.62%	**88.37%**	**70.42%**	61.28%	56.77%	74.08%	23.73%	34.54%	83.93%	64.43%
Pick1	0	Group	50.88%	56.67%	46.62%	51.82%	76.83%	67.41%	58.98%	**64.02%**	71.55%	**31.60%**	35.16%	78.57%	66.28%
Pick5	0	Rand	53.64%	62.39%	36.70%	43.78%	81.33%	69.79%	61.40%	56.77%	75.99%	22.16%	30.78%	82.14%	68.60%
Pick5	0	Group	46.89%	55.67%	32.05%	37.06%	75.24%	68.98%	63.35%	61.97%	77.84%	26.88%	32.35%	79.46%	73.05%
CLIP Baseline			59.98%	63.18%	47.9%	60.2%	81.17%	67.67%	61.95%	37.82%	60.12%	10.94%	23.19%	67.86%	39.95%

Table 4. Full $sDCI$ ablation analysis against all benchmarks. Cells are colored in comparison to the CLIP Baseline.

	ARO				VL-Checklist		
Ablation	VG-R	VG-A	COCO	FLICKR	Object	Attribute	Relation
Rand	**67.71%**	**62.64%**	**68.22%**	73.23%	**77.71%**	**67.89%**	**66.07%**
ImgGroup	63.40%	61.27%	67.99%	**73.70%**	72.21%	66.89%	62.51%
Neg Loss 9	**76.15%**	**64.13%**	**85.21%**	**89.33%**	71.87%	66.35%	**67.61%**
Neg Loss 1	68.14%	63.14%	81.44%	86.49%	74.59%	67.22%	64.08%
Neg Loss 0	52.38%	58.60%	37.67%	44.57%	**78.43%**	**68.61%**	61.19%
First	65.46%	59.87%	67.36%	73.16%	74.94%	**68.27%**	64.09%
Pick1	**66.98%**	**63.41%**	**70.74%**	**75.52%**	**76.99%**	67.28%	63.99%
Pick5	64.22%	62.57%	66.22%	71.71%	72.97%	66.63%	**64.79%**

Table 5. $sDCI$ Grouped ablation against standard benchmarks. Results from Table 4 are averaged across different ablations.

9.3.2 DCI fine-tuning performance

In Table 7 we report our different ablations performance compared to the LN and COCO baselines, as well as CLIP and DAC as comparison points. We expect sDCI models to outperform all other baselines, given the test set is out-of-distribution for the other models. Still, there are some interesting observations available in this table.

First, for sDCI, Localized Narrations, and COCO, only training on the smallest subset of images (7800) and without using masks, DCI ends up performing only as well as COCO for Subcrop-Caption matching, but both outperform Localized Narratives by a noticeable margin. This may point to data in Localized Narratives generally being less sample-efficient than baseline COCO captions.

Second, moving to LLM-based captions increases performance on all negatives at the expense of performance on Subcrop-Caption masking. This would imply that the captions generated by LLMs may actually be overfitting to the test task to the detriment of performance on other metrics.

Third, training directly on negatives as a method of improving models' vision-language understanding universally decreases performance on Subcrop-Caption masking, a task that also definitely takes strong vision-language understanding. This is seen regardless of the training dataset used, or of any other sDCI ablation involved.

9.3.3 Linear Transferability

We evaluate a subset of all of our models, trained on each of our candidate datasets, on the Elevater [17] benchmark to determine linear transferability for our models. We expect some degradation given the relatively small size of our training datasets. We also evaluate them on zero-shot ImageNet specifically.

Overall, in table 9 we observe a slight degradation in linear probe performance on their included datasets across all shots. In our ImageNet zero-shot evaluation reported in Table 8, we note that sDCI trained without negatives suffers much less degradation compared to with negatives, ending

Ablation	All SCM	All Neg	All Pick5 SCM	All Pick5 Neg	Base Neg	All Hard Negs
Rand	51.17%	88.65%	16.57%	68.35%	**91.57%**	78.40%
ImgGroup	**56.47%**	**88.71%**	**21.45%**	**69.56%**	90.08%	**79.57%**
NL 9	47.15%	96.24%	11.52%	**89.33%**	95.68%	**85.57%**
NL 1	54.32%	**94.96%**	19.78%	84.83%	**95.84%**	83.15%
NL 0	**59.99%**	74.85%	**25.74%**	32.71%	80.95%	68.24%
First	**57.07%**	**89.26%**	17.08%	62.54%	**91.37%**	**80.89%**
Pick1	53.95%	87.63%	**22.19%**	**72.68%**	90.18%	77.12%
Pick5	50.45%	89.16%	17.77%	71.65%	90.92%	78.95%

Table 6. $sDCI$ Grouped ablation against the 112 heldout sDCI test images. Results from Table 4 are averaged across different ablations.

Dataset	Training Parameters Captions	Negatives	Batching	All SCM	All Neg	All Pick5 SCM	All Pick5 Neg	Base Neg	All Hard Negs
$sDCI_{7805}$	First	LLM	Rand	38.30%	84.61%	9.10%	69.22%	92.86%	76.54%
$sDCI$	Pick1	LLM	ImgGroup	58.14%	**94.60%**	26.81%	**89.88%**	96.43%	83.58%
$sDCI$	Pick1	LLM	Rand	50.48%	94.39%	19.02%	88.85%	95.54%	81.40%
$sDCI$	First	LLM	ImgGroup	61.76%	95.76%	19.63%	76.81%	**97.32%**	85.43%
$sDCI$	First	LLM	Rand	54.17%	96.17%	14.64%	75.92%	96.43%	**85.91%**
$sDCI$	First	Spacy	Rand	55.13%	87.35%	19.08%	59.30%	89.29%	75.85%
$sDCI$	Pick1	None	ImgGroup	**64.02%**	71.55%	**31.60%**	35.15%	78.57%	66.28%
$sDCI$	Pick1	None	Rand	56.77%	74.08%	23.73%	34.54%	83.93%	64.43%
$sDCI$	First	None	ImgGroup	63.13%	75.44%	27.84%	33.11%	81.25%	69.49%
$sDCI$	First	None	Rand	57.25%	74.21%	22.23%	30.30%	80.36%	67.58%
LN	First	Spacy	Rand	37.82%	76.95%	9.37%	46.31%	86.61%	63.20%
LN_{7805}	First	Spacy	Rand	34.27%	75.58%	7.73%	37.82%	83.04%	61.63%
LN	First	None	Rand	41.45%	58.82%	12.72%	21.75%	80.36%	53.42%
$COCO$	First	Spacy	Rand	40.97%	79.21%	12.65%	52.74%	91.07%	64.71%
$COCO_{7805}$	First	Spacy	Rand	38.51%	79.75%	11.70%	55.27%	86.61%	64.16%
$COCO$	First	None	Rand	42.00%	61.35%	13.95%	21.41%	82.14%	52.60%
CLIP Baseline				37.82%	60.12%	10.94%	23.19%	67.86%	39.95%
DAC_{LLM}				36.87%	81.12%	8.00%	35.91%	86.61%	70.66%
DAC_{SAM}				36.46%	84.40%	6.91%	40.83%	89.29%	73.94%

Table 7. Dense Captions test results. We compare DCI-trained models to models trained on Localized Narratives and COCO datasets, as well as to baselines.

ImageNet 0-Shot	
Model	Valid. Accuracy
CLIP (baseline)	60.96%
$sDCI_{P1}$	42.51%
$sDCI_{P1NL0}$	51.44%
DAC-LLM	52.65%
DAC-SAM	53.43%

Table 8. ImageNet zero-shot.

	Elevater N-Shot				
Model	0	5	20	50	Full
$sDCI_{P1}$	45.10%	60.78%	69.54%	72.91%	77.49%
$sDCI_{P1NL0}$	51.29%	61.59%	70.80%	73.39%	77.92%
CLIP	55.59%	64.85%	71.90%	74.38%	78.96%

Table 9. Elevater scores for linear probe across 20 benchmark datasets

up with comparable performance to the DAC models.

10. Additional Selected Examples

We include a few additional examples from the DCI dataset, selected from a random subset of 20 instances to highlight certain elements of the dataset. In each we share a subset of

the masks available per image.

Figure 4 shows the level in-depth that the descriptions go to. Despite only being an image of shoes on some grass, DCI contains descriptions down to the details of the crossing pattern on the toecap or the specks of light colored materials in a clump of dirt on the ground.

Figure 5 contains a fairly complex scene of various tiled stone buildings and light fixtures, however the description is able to identify a tree that is mostly obscured by a foreground building, as well as be in-depth enough to describe the shape at the top of a lamppost in the image as a "tiny urn".

Figure 6 stands as another example of retaining useful and aligned information even when there's a high amount of potential complexity in the image. While "An antique blue car in front of a row of trees" may be a standard caption for this image, we instead have details of the orientation of the car, details of what is visible in frame, and the resolution of the text goes all the way down to the small circular sticker on the rear passenger side door, or the screws on the reflector on the front bumper.

Figure 7 displays a more active scene of two women cooking bread, however it still captures in-depth descriptions of everything contained in the image including flour spread on a table, ornamental details on a tablecloth, and additionally the placement of a bowl and a knife that were not captured in their own masks, but still were successfully annotated.

11. Evaluating additional baselines on sDCI

We evaluate a large quantity of available VLMs from RN50, ViT, roberta, convnext, and coca architectures on the DCI test set. Results can be seen in Table 10.

Generally we observe that larger models perform better, but no model excels at all tasks. In contrast to what is observed for models trained with negatives, performance from these pretraining objectives is positively correlated between Subcrop-Caption Matching and detecting negatives.

The most performant model on SCM averaged between the two tasks is coca_ViT-L-14 mscoco_finetuned_laion2b_s13b_b90k. The most performant model in average across negatives tasks is ViT-g-14 laion2b_s34b_b88k

Figure 4. One example from the Densely Captioned Images dataset, highlighting how in-depth descriptions are provided even for relatively simple scenes.

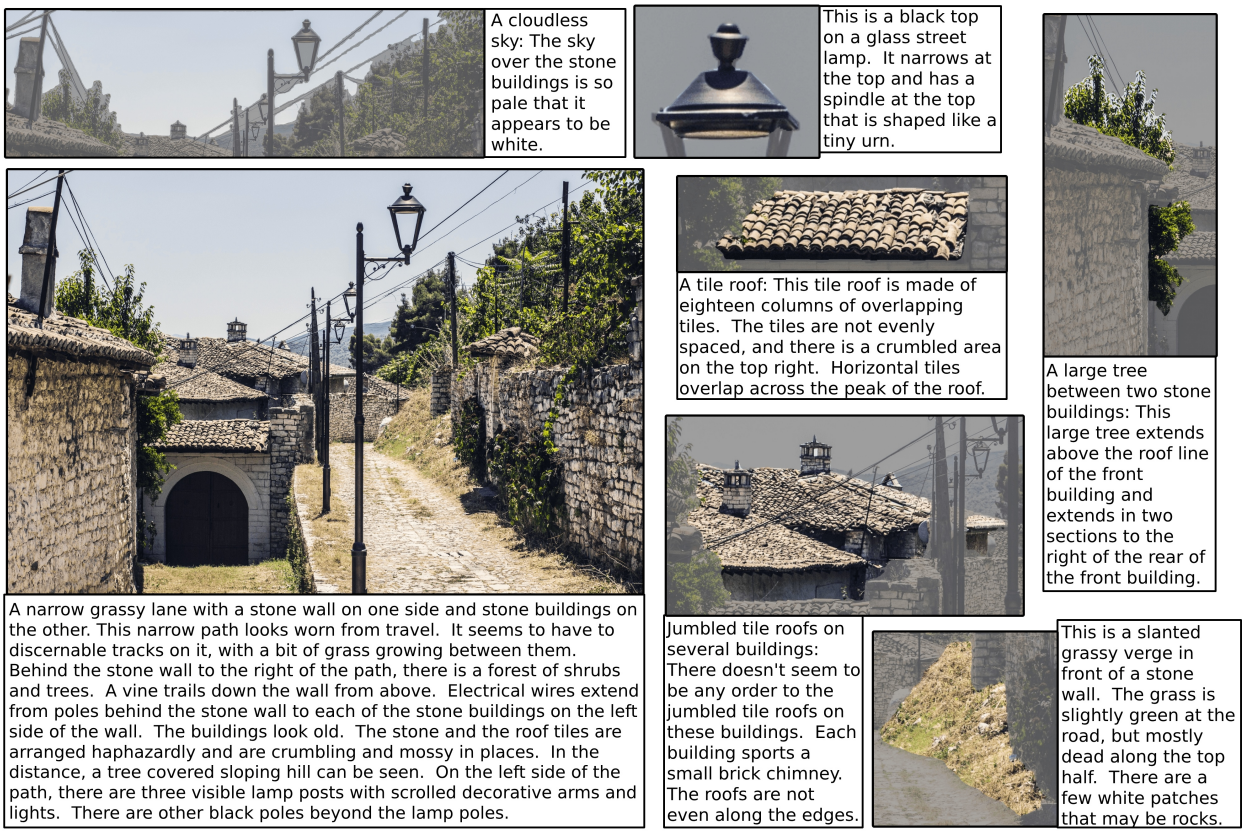

Figure 5. One example from the Densely Captioned Images dataset, highlighting how text is still highly aligned even with complex masks.

Figure 6. One example from the Densely Captioned Images dataset, highlighting the high resolution of the text in alignment with the image, down to details of the stickers on car windows or screws on the reflectors.

Figure 7. One example from the Densely Captioned Images dataset, displaying a scene with a complex interaction. The aligned description captures the action itself, alongside in-depth details like the clasp on the box on the wall, or the hand of one of the women pictured.

Training Parameters		All		All Pick5		Base	All
Arch	Dataset	SCM	Neg	SCM	Neg	Neg	Hard Negs
RN50	yfcc15m	39.95%	55.85%	6.72%	19.30%	71.63%	55.35%
RN50	cc12m	41.32%	48.79%	8.65%	19.69%	65.38%	49.72%
RN50-quickgelu	openai	41.54%	62.20%	11.58%	23.51%	72.24%	54.00%
RN50-quickgelu	yfcc15m	40.08%	55.89%	6.63%	19.06%	70.51%	55.14%
RN50-quickgelu	cc12m	41.97%	47.70%	9.18%	17.98%	65.41%	48.63%
RN101	yfcc15m	39.72%	55.85%	7.02%	19.73%	72.01%	54.88%
RN101-quickgelu	openai	40.44%	62.70%	10.53%	19.48%	76.52%	56.22%
RN101-quickgelu	yfcc15m	39.88%	55.51%	7.14%	19.31%	71.07%	54.47%
ViT-B-32	openai	40.06%	60.79%	11.23%	24.12%	67.56%	41.34%
ViT-B-32	laion400m_e31	45.44%	62.96%	14.75%	22.29%	79.33%	57.27%
ViT-B-32	laion400m_e32	45.31%	62.95%	14.65%	22.07%	79.13%	57.27%
ViT-B-32	laion2b_e16	45.59%	60.13%	15.34%	20.23%	78.60%	56.26%
ViT-B-32	laion2b_s34b_b79k	45.56%	60.12%	15.30%	20.58%	78.10%	56.28%
ViT-B-32	datacomp_xl_s13b_b90k	45.62%	59.92%	15.09%	20.24%	78.17%	56.19%
ViT-B-32	datacomp_m_s128m_b4k	40.29%	58.53%	8.93%	19.93%	73.43%	55.70%
ViT-B-32	commonpool_m_clip_s128m_b4k	41.58%	58.35%	10.60%	19.63%	71.95%	54.76%
ViT-B-32	commonpool_m_laion_s128m_b4k	40.25%	59.00%	8.92%	20.45%	71.89%	55.81%
ViT-B-32	commonpool_m_image_s128m_b4k	41.14%	58.33%	9.95%	20.28%	73.66%	55.54%
ViT-B-32	commonpool_m_basic_s128m_b4k	40.86%	58.32%	9.81%	19.54%	73.41%	55.24%
ViT-B-32	datacomp_s_s13m_b4k	28.98%	55.47%	1.31%	18.44%	61.73%	54.16%
ViT-B-32	commonpool_s_clip_s13m_b4k	33.96%	59.22%	4.62%	21.46%	68.17%	56.34%
ViT-B-32	commonpool_s_laion_s13m_b4k	28.99%	55.50%	1.37%	18.45%	61.43%	54.27%
ViT-B-32	commonpool_s_image_s13m_b4k	28.98%	55.47%	1.33%	18.31%	61.73%	54.16%
ViT-B-32	commonpool_s_basic_s13m_b4k	32.31%	57.03%	3.39%	19.01%	64.84%	54.89%
ViT-B-32	commonpool_s_s13m_b4k	32.45%	58.97%	4.80%	22.63%	67.50%	55.64%
ViT-B-32-quickgelu	openai	40.06%	60.79%	11.22%	24.18%	67.56%	41.34%
ViT-B-32-quickgelu	laion400m_e31	45.19%	61.19%	14.75%	21.76%	78.27%	55.67%
ViT-B-32-quickgelu	laion400m_e32	45.08%	61.14%	14.83%	21.88%	78.45%	55.60%
ViT-B-32-quickgelu	metaclip_400m	44.27%	60.73%	12.81%	20.28%	77.68%	57.23%
ViT-B-32-quickgelu	metaclip_fullcc	44.63%	59.75%	14.05%	20.61%	79.09%	56.34%
ViT-B-16	openai	39.33%	61.40%	11.30%	26.02%	70.98%	51.96%
ViT-B-16	laion400m_e31	44.66%	61.46%	14.66%	22.60%	78.07%	56.09%
ViT-B-16	laion2b_s34b_b88k	45.28%	60.81%	15.28%	20.60%	77.72%	57.44%
ViT-B-16	datacomp_xl_s13b_b90k	45.31%	59.96%	15.39%	20.42%	76.76%	56.48%
ViT-B-16	datacomp_l_s1b_b8k	44.41%	59.45%	14.12%	20.32%	76.99%	55.91%
ViT-B-16	commonpool_l_clip_s1b_b8k	44.06%	60.71%	14.00%	20.73%	76.82%	55.88%
ViT-B-16	commonpool_l_laion_s1b_b8k	44.53%	60.45%	14.22%	20.40%	76.66%	56.24%
ViT-B-16	commonpool_l_image_s1b_b8k	44.02%	58.72%	13.86%	19.56%	76.50%	55.23%
ViT-B-16	commonpool_l_text_s1b_b8k	44.48%	60.21%	14.02%	20.67%	78.46%	56.91%
ViT-B-16	commonpool_l_basic_s1b_b8k	44.19%	59.99%	13.89%	19.79%	77.98%	56.47%
ViT-B-16	commonpool_l_s1b_b8k	43.11%	58.31%	13.38%	19.10%	75.81%	54.20%
ViT-B-16-quickgelu	metaclip_400m	43.88%	61.31%	13.28%	21.47%	79.17%	57.95%
ViT-L-14	openai	38.09%	59.76%	11.05%	24.82%	69.15%	52.42%
ViT-L-14	laion400m_e31	44.09%	63.61%	14.89%	21.92%	80.22%	57.90%
ViT-L-14	laion400m_e32	44.09%	63.43%	14.86%	21.65%	80.04%	57.78%
ViT-L-14	laion2b_s32b_b82k	45.19%	58.82%	15.25%	19.82%	76.96%	56.73%
ViT-L-14	datacomp_xl_s13b_b90k	44.37%	63.19%	15.37%	22.43%	79.99%	59.45%
ViT-L-14	commonpool_xl_clip_s13b_b90k	44.63%	61.28%	15.52%	21.37%	77.50%	58.05%
ViT-L-14	commonpool_xl_laion_s13b_b90k	45.34%	62.77%	15.77%	21.50%	79.55%	59.30%
ViT-L-14	commonpool_xl_s13b_b90k	43.50%	62.09%	14.49%	21.65%	76.41%	58.67%
ViT-L-14-quickgelu	metaclip_400m	42.91%	61.60%	12.90%	21.43%	78.62%	59.11%
ViT-L-14-quickgelu	metaclip_fullcc	43.02%	61.92%	13.70%	21.93%	81.63%	59.22%
ViT-L-14-quickgelu	dfn2b	46.37%	60.27%	16.88%	20.65%	80.08%	58.98%
ViT-H-14	laion2b_s32b_b79k	45.32%	63.44%	15.51%	21.35%	79.06%	60.24%
ViT-H-14-quickgelu	metaclip_fullcc	43.12%	62.44%	13.65%	21.72%	81.02%	59.99%
ViT-H-14-quickgelu	dfn5b	46.39%	60.99%	16.93%	21.21%	82.36%	60.19%
ViT-g-14	laion2b_s12b_b42k	45.06%	61.44%	15.83%	21.22%	78.95%	58.44%
ViT-g-14	laion2b_s34b_b88k	44.95%	63.72%	15.58%	22.15%	77.94%	60.40%
ViT-bigG-14	laion2b_s39b_b160k	45.09%	62.14%	15.67%	20.48%	79.37%	60.44%
roberta-ViT-B-32	laion2b_s12b_b32k	12.58%	47.85%	0.04%	14.40%	48.48%	47.37%
xlm-roberta-base-ViT-B-32	laion5b_s13b_b90k	12.96%	50.62%	0.03%	16.57%	49.08%	50.49%
xlm-roberta-large-ViT-H-14	frozen_laion5b_s13b_b90k	11.99%	51.06%	0.01%	17.67%	50.67%	50.59%

Table 10. Summarized Dense Captions test results on OpenCLIP models. We compare various baseline models on our Subcrop-Caption Matching (SCM) and negatives tests.

12. DCI Dataset Datasheet

The following are our answers to the Datasheets for Datasets [13] question list.

12.1. Motivation

For what purpose was the dataset created? To create an initial dataset of highly aligned text and image pairs that were not yet available, primarily for evaluating how well existing models can make use of all of the data.
Who created the dataset and on behalf of which entity? Researchers on Meta's FAIR research team created it on their own behalf.
Who funded the creation of the dataset? Meta

12.2. Composition

What do the instances that comprise the dataset represent? Images with text annotations
How many instances are there in total? 7805 images with complete mask-aligned annotations.
Does the dataset contain all possible instances or is it a sample of instances from a larger set? Images were sampled from a random subset of SA-1B's [15] underlying image dataset.
What data does each instance consist of? One image, a top-level caption and description, and then a list of submask-subcaption pairings covering a significant portion of the image's content
Is there a label or target associated with each instance? Just the text descriptions, no categorization is done.
Is any information missing from individual instances? Not all of the image is covered in the submask-aligned captions, so the descriptions may still be considered incomplete.
Are relationships between individual instances made explicit? There is no clear relationships between instances in the dataset.
Are there recommended data splits? The dataset is intended **primarily** as a test set, however we also provide a finetuneing train/valid/test split for those wanting to use it for experiments.
Are there any errors, sources of noise, or redundancies in the dataset? Though attempts were made to keep the dataset high-quality, annotator error can be present through the dataset, modeling errors may cause some masks to have been omitted, and the LLM-based augmentation for scaling captions to CLIP length may introduce noise as well.
Is the dataset self-contained, or does it link to or otherwise rely on external resources? Self-contained
Does the dataset contain data that might be considered confidential? Not to the authors' knowledge
Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? Not to the author's knowledge

12.3. Collection Process

How was the data associated with each instance acquired? The images were selected from the SA-1B image dataset, and underwent a combination of automated and manual annotation
What mechanisms or procedures were used to collect the data? We use the Mephisto framework, as well as custom annotation interfaces, to collect the data. Complete details are available on the project's GitHub.
If the dataset is a sample from a larger set, what was the sampling strategy? Random selection from a single subset
Who was involved in the data collection process and how were they compensated? Crowdworkers were paid well above minimum wage for their time spent.
Over what timeframe was the data collected? Spring through Fall of 2023.
Were any ethical review processes conducted? This collection process underwent internal review.

12.4. Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done? The dataset was preprocessed using the Segment Anything Model in order to identify the regions of the image to be annotated.
Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data? The raw data is unmodified by the extraction process.
Is the software that was used to preprocess/clean/label the data available? All of the software used to construct the dataset will be made available alongside the dataset release on the project's GitHub.

12.5. Uses

Has the dataset been used for any tasks already? The dataset is used for the Densely Captioned Images test set.
Are there tasks for which the dataset should not be used? The dataset is intended as a test set. Any use outside of this is unplanned by the authors.

12.6. Distribution

Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created? The dataset will be made broadly available
How will the dataset will be distributed? A download script will be made available on the project GitHub, alongside a copy of the code used to collect and prepare the original dataset.
When will the dataset be distributed? Upon release of the associated publication.
Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under

applicable terms of use (ToU)? The dataset will be released under CC-By-NC.
Have any third parties imposed IP-based or other restrictions on the data associated with the instances? Not to the authors' knowledge.
Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? Not to the authors' knowledge.

12.7. Maintenance

Who will be supporting/hosting/maintaining the dataset? Meta's FAIR team will host this dataset.
How can the owner/curator/manager of the dataset be contacted? On the project's GitHub page.
Is there an erratum? Changes will be noted on the project's github.
Will the dataset be updated? The authors have no clear schedule to update or alter the dataset.
If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances? All known instances of people in the dataset have been face-blurred as per the SA-1B release, and no retention policy is known.
If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? They may do so from the project GitHub, following the terms included therein.