A More examples of image-text pairs (no cherry picking)

Raw: *2003 Mercedes-Benz C240 sedan, Leather, MUST BE SEEN - $6199*
BLIP (finetuned): *a couple of cars parked in a parking lot with trees and cars*
BLIP2: *2002 mercedes-benz c-class for sale*
BLIP2 (finetuned): *a blue mercedes benz car parked in a parking lot next to yellow cars*
OpenCLIP-CoCa: *find used 2 0 0 1 mercedes benz c 2 4 0 base sedan 4 door 2 5 l for 2 0 0 1 mercedes benz c 2*
OpenCLIP-CoCa (finetuned): *a blue mercedes parked on the side of a road .*

Raw: *Gaziburma Ünal is one of Gespeicherte Orte von Can.*
BLIP (finetuned): *dozens of trays of different types of treats at a food stand*
BLIP2: *some trays of pastries and a knife*
BLIP2 (finetuned): *there are many trays filled with food items from the store*
OpenCLIP-CoCa: *baklava , sweets , pastries*
OpenCLIP-CoCa (finetuned): *there are trays full of different types of food .*

Raw: *Open Virgin of Kazan, Wooden Egg with Stand, Blue*
BLIP (finetuned): *a gray and white logo with the words more info in a rectangular shape*
BLIP2: *a white button with the word more info*
BLIP2 (finetuned): *more information is shown on a white button with an orange background*
OpenCLIP-CoCa: *home - page - button . png*
OpenCLIP-CoCa (finetuned): *a picture of a close up of a text message*

Raw: *2016.07.01 Nametags with Pronouns - Avery 5392_non-branded*
BLIP (finetuned): *there are no pictures here to provide a caption for*
BLIP2: *hello, my name is name, my pronouns are pronouns*
BLIP2 (finetuned): *a blue and white label with a blue and white text*
OpenCLIP-CoCa: *1 5 + hello my name is names pronunciations and meanings*
OpenCLIP-CoCa (finetuned): *hello my name is , my pronouns are .*

Raw: *Italien - Ligurien*
BLIP (finetuned): *beige colored building with tan accents and palm trees on both sides of walkway*
BLIP2: *house in villa marina, a villa with many rooms and palm trees*
BLIP2 (finetuned): *a park with lots of trees and benches in front of a large building*
OpenCLIP-CoCa: *residence - villa - maria - di - san - giovanni - near - the - sea - in - taormina*
OpenCLIP-CoCa (finetuned): *a picture of a large building with a bunch of palm trees .*

Raw: *3 formas de pedir la mano de tu novia - wikiHow*
BLIP (finetuned): *crates stacked up in a pile on top of each other*
BLIP2: *the building contains five floors of wooden planks*
BLIP2 (finetuned): *a big pile of wooden planks stacked together*
OpenCLIP-CoCa: *the cost of wood pallets*
OpenCLIP-CoCa (finetuned): *a large pile of wooden pallets mounted to a wall .*

Raw: *lutz*
BLIP (finetuned): *blond haired man in black suit looking at camera*
BLIP2: *a man sitting on a chair with a blank background*
BLIP2 (finetuned): *a man sitting in a chair with a lapel button in front*
OpenCLIP-CoCa: *actor tilda swinton is pictured during a press conference for the film ' a dangerous method ' at the 2 0 1 1 toronto film festival*
OpenCLIP-CoCa (finetuned): *a person sitting on a chair wearing a suit and tie .*

Raw: *Women Personality Creative Christmas Hat Face Expression Gold Earring Funny Cartoon Ear Stud Jewelry Accessories Gift Hot*
BLIP (finetuned): *red and gold tone emojt earring*
BLIP2: *kawaii santa emoticuos en la cabeza*
BLIP2 (finetuned): *a pair of emoji earrings with faces and hats*
OpenCLIP-CoCa: *best christmas gift for her new arrivals emoji earrings christmas emoji earrings*
OpenCLIP-CoCa (finetuned): *a pair of gold earrings with a smiley face and a christmas hat .*

Raw: *10840 SW 126th St photo067*
BLIP (finetuned): *overview of a large backyard with a swimming pool and patio*
BLIP2: *3344 sw 7th st, palm beach*
BLIP2 (finetuned): *a house with a pool from above, with a yard*
OpenCLIP-CoCa: *home for sale in country club shores west palm beach florida*
OpenCLIP-CoCa (finetuned): *aerial image of a pool that has a little bit of shade by the side .*

Raw: *image8.JPG*
BLIP (finetuned): *members of a school play soccer in a gymnasium with a crowd*
BLIP2: *a large crowd of kids perform during a dance show*
BLIP2 (finetuned): *a group of young children standing on the basketball court*
OpenCLIP-CoCa: *kid dressed in white standing in a gym area .*
OpenCLIP-CoCa (finetuned): *a group of kids on the gym floor with fans on the floor .*

Raw: *hair oil*
BLIP (finetuned): *smiling blonde woman blow drying hair in a salon while getting a mani*
BLIP2: *hair stylist using hair spray in beauty salon*
BLIP2 (finetuned): *a person is using a hairdryer to blow dry a long blonde hair*
OpenCLIP-CoCa: *female hairdresser styling a long blond hair with hairspray in a beauty salon . concept : hair care , hair straightening , hair color correction .*
OpenCLIP-CoCa (finetuned): *a person is spraying a hair dryer on a long blonde hair .*

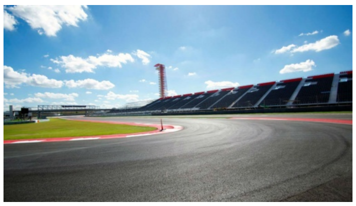
Raw: *Women long sleeve t shirt 2015 Fashion shirts woman Full Comfortable leisure fashion womens long sleeve tops*
BLIP (finetuned): *the qaoo loading logo is shown above the qaoo loading logo*
BLIP2: *qoo10 loading logo on white*
BLIP2 (finetuned): *a picture of an image of a phone screen showing a loading sign*
OpenCLIP-CoCa: *loading _ 1 1 1 1 2 0 _ 0 1 . png*
OpenCLIP-CoCa (finetuned): *a light grey font and a dark grey font with a large white background*

Raw: *Nautica NAPTYR005*
BLIP (finetuned): *navitta mens stainless steel bracelet watch with blue dial*
BLIP2: *nautica men's chronograph watch*
BLIP2 (finetuned): *nautica men's men's chronograph black dial stainless steel bracelet watch*
OpenCLIP-CoCa: *nautica newport chronograph n 2 2 0 0 3 g*
OpenCLIP-CoCa (finetuned): *a mans black watch is shown with red and blue accents*

Raw: *Greenberg Weathered Marble Plush Ivory Area Rug*
BLIP (finetuned): *grey rug with a text home on it by a table*
BLIP2: *a grey area rug on a wooden floor*
BLIP2 (finetuned): *a white coffee table with a sign saying home on it. it is sitting on a cream colored rug*
OpenCLIP-CoCa: *rugs and carpets in hyderabad : buy online at best price in ...*
OpenCLIP-CoCa (finetuned): *a rug is shown in a living room with a chair .*

Raw: *productivity, productivity, productivity*
BLIP (finetuned): *drivers guide to the truck industry*
BLIP2: *buy and sell truck parts*
BLIP2 (finetuned): *a white truck with a cover on it drives along a highway*
OpenCLIP-CoCa: *how the trucking industry is changing*
OpenCLIP-CoCa (finetuned): *there are some trucks on the road .*

Raw: *Amigas*
BLIP (finetuned): *crowd of people outside a wedding ceremony near several trees*
BLIP2: *a wedding ceremony in the middle of the street*
BLIP2 (finetuned): *a black and white photograph of a number of women in prom dresses*
OpenCLIP-CoCa: *2 0 1 3 0 8 0 5 _ wedding _ carlenan _ 0 0 3*
OpenCLIP-CoCa (finetuned): *a group of people hugging and talking in a group*



Raw: *Autozone*
BLIP (finetuned): *racing track with a line of seats and a sky background*
BLIP2: *a photo of a grand prix race track, under a blue sky*
BLIP2 (finetuned): *the circuit track is empty, but the sun beams into the sky*
OpenCLIP-CoCa: *circuit of the americas*
OpenCLIP-CoCa (finetuned): *a red and white pole next to a racing track*

Raw: *Automne hiver enfants manteau et pantalon ensemble capuche veste de Ski et pantalon garçon fille coupe-vent imperméable en plein air camping randonnée*
BLIP (finetuned): *a man wearing a red and blue jacket and a pair of pants and a pair of sneakers*
BLIP2: *the arctic light hooded jacket and pants set*
BLIP2 (finetuned): *the colors of the jacket match the pant color of the jacket*
OpenCLIP-CoCa: *the arctic light 2 0 1 7 children 's clothing sets winter kids ski suits sets windproof waterproof warm jackets coats pants boys set*
OpenCLIP-CoCa (finetuned): *a child standing in their ski wear and a jacket and pants*

Raw: *1173x1500 Awesome Adult Coloring Pages Printable Zentangle Design*
BLIP (finetuned): *chinese dragon coloring pages dragon coloring pages for adults to print coloring pages*
BLIP2: *dragon coloring pages with large and large dragon*
BLIP2 (finetuned): *a circle with a dragon on it in the center*
OpenCLIP-CoCa: *the 2 5 best chinese dragon drawing ideas on pinterest chinese*
OpenCLIP-CoCa (finetuned): *a chinese dragon looks like a dragon from the movie the karate kid*

Raw: *Der Lieferumfang*
BLIP (finetuned): *there are several electronics laid out on the table ready to be used*
BLIP2: *samsung galaxy s10e review | a quick tour of the samsung galaxy s10e*
BLIP2 (finetuned): *wireless charging case and remote control, both packaged in the box*
OpenCLIP-CoCa: *best - wireless - chargers - for - samsung - galaxy - note - 8 - s 8 - and - iphone - 8*
OpenCLIP-CoCa (finetuned): *a set of various electronic items sitting on a table .*

B Experiment details

Refer to Appendices M and N of the DataComp benchmark [18] for training and evaluation details. To summarize, both `small` and `medium` scales use ViT-B/32 as the image encoder for CLIP, in addition to fixing the hyperparameters used for training: learning rate 5e-4, 500 warmup steps, batch size 4096, AdamW optimizer $\beta_2 = 0.98$. `Large` scale training uses the same hyperparameters, but with batch size 8192 and ViT-B/16 as the image encoder.

Using DataComp infrastructure and the AWS EC2 cloud, a `small` model takes 4 A100 hours to train, while `medium` requires 40 A100 hours and `large` utilizes 960 A100 hours. We additionally report CLIP ViT-L/14 and BLIP2 (OPT 2.7B backbone) inference costs. Recall that we run both of these models on the DataComp's `large` pool to curate the datasets used in this paper. For the CLIP model, we measure throughput at 490 samples per second on a single A100. For BLIP2, we get 75 samples per second on the same hardware. Hence, for the `large` pool of 1.28B samples, we spend 725 A100 hours computing CLIP features and 4,740 A100 hours generating BLIP2 captions.

While the annotation cost (i.e., BLIP2 caption and CLIP score generation) is 6× larger than a single training run proposed by the DataComp benchmark (which is equivalent to going through the entire candidate pool for 1 epoch), this additional cost can be easily amortized with more training epochs over the final training set, as well as with training different downstream models on the improved dataset. For reference, OpenAI trained various CLIP models on the same set of 400M curated image-text pairs; the best performing model was trained on 256 GPUs for 2 weeks, totalling about 86,000 GPU hours [2]. This scale of training is common among existing large vision models. Future work could explore the option of adaptively allocating compute to CLIP training and synthetic caption annotation given a fixed compute budget.

C Temperature ablations

Captioning model	Metric	T=0.5	T=0.75	T=1.0	T=1.5
BLIP (finetuned)	ImageNet accuracy	-	0.207	**0.212**	-
	Average accuracy	-	0.303	**0.312**	-
BLIP2	ImageNet accuracy	0.212	**0.281**	0.280	0.251
	Average accuracy	0.300	**0.357**	0.353	0.332
BLIP2 (finetuned)	ImageNet accuracy	-	0.227	**0.234**	0.221
	Average accuracy	-	0.325	**0.326**	0.311
OpenCLIP-CoCa	ImageNet accuracy	0.306	**0.321**	0.314	-
	Average accuracy	0.366	**0.371**	0.370	-
OpenCLIP-CoCa (finetuned)	ImageNet accuracy	-	0.252	**0.264**	0.262
	Average accuracy	-	0.364	**0.374**	0.364

Table 2: Performance on ImageNet and averaged across 38 tasks when training on the captions generated by captioning models in Table 1, with different softmax temperatures. We find that $T = 0.75$ and $T = 1.0$ generally lead to good performance for CLIP training.

D More filtering baselines

[2] `https://openai.com/research/clip`

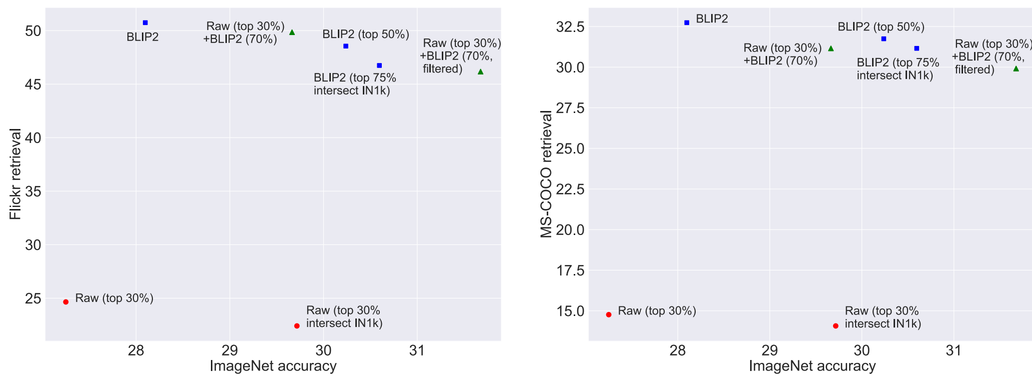


Figure 9: Retrieval performance on Flickr (left) and MS-COCO (right) versus ImageNet accuracy for select baselines. Similar to the findings in Figure 2, we find that using BLIP2 captions or including them in the training data with raw captions significantly boosts performance.

Baseline	Training set size	ImageNet accuracy	Average accuracy
`small` scale (12.8M candidate pool, 12.8M training steps)			
Raw captions (no filtering)	12.8M*	0.025*	0.132*
BLIP2 captions (no filtering)	12.8M	0.076	0.200
Raw captions (top 30%)	3.8M*	0.051*	0.173*
BLIP2 captions (top 50%)	6.4M	**0.080**	**0.203**
Raw captions (intersect IN1k and top 30%)	1.4M*	0.039*	0.144*
BLIP2 captions (intersect IN1k and top 75%)	2.4M	0.073	0.192
Raw captions (top 30%) + BLIP2 captions (70%, filtered), intersect IN1k	2.2M	0.045	0.153
Raw captions (top 30%) + BLIP2 captions (70%, filtered)	8.4M	0.076	0.197
`medium` scale (128M candidate pool, 128M training steps)			
Raw captions (no filtering)	128M*	0.176*	0.258*
BLIP2 captions (no filtering)	128M	0.281	0.357
Top BLIP2 captions across all temperatures (no filtering)	128M	0.293	0.368
Raw captions (top 30%)	38M*	0.273*	0.328*
BLIP2 captions (top 50%)	64.1M	0.302	0.370
Raw captions (intersect IN1k and top 30%)	14.0M*	0.297*	0.328*
BLIP2 captions (intersect IN1k and top 75%)	23.6M	0.306	0.360
Raw captions (top 30%) + BLIP2 captions (70%, filtered), intersect IN1k	22.2M	0.281	0.314
Raw captions (top 30%) + BLIP2 captions (70%, filtered)	83.6M	**0.317**	0.368
BLIP2 captions (top 50%) + Raw captions (50%, filtered)	75.3M	0.310	**0.376**
`large` scale (1.28B candidate pool, 1.28B training steps)			
Raw captions (no filtering)	1.28B*	0.459*	0.437*
BLIP2 captions (no filtering)	1.28B	0.487	0.505
Raw captions (top 30%)	384M*	0.578*	0.529*
BLIP2 captions (top 50%)	641M	0.526	0.522
Raw captions (intersect IN1k and top 30%)	140M*	0.631*	0.537*
BLIP2 captions (intersect IN1k and top 75%)	237M	0.533	0.527
Raw captions (top 30%) + BLIP2 captions (70%, filtered), intersect IN1k	222M	**0.643**	0.549
Raw captions (top 30%) + BLIP2 captions (70%, filtered)	834M	0.598	**0.551**

Table 3: Performance for select baselines at `small`, `medium` and `large` scales of DataComp. * indicates numbers obtained from the original paper [18]. Underlined numbers are best-performing baselines from the DataComp benchmark, trained on only raw web-crawled captions. Bolded numbers are the updated state-of-the-art figures after comparing with baselines involving synthetic captions. In general, given a fixed training budget, it is helpful to include more samples in the training pool by carefully replacing noisy raw captions with synthetic captions (i.e., Raw (top 30%) + BLIP2 (70%, filtered) versus Raw (top 30%)). We experiment with many more filtering and mixing methods at the `medium` scale and report how the performance varies with CLIP score filtering threshold, see Table 4.

CLIP score filtering	10%	20%	30%	50%	75%	90%
Cosine similarity threshold						
Raw captions	0.295	0.266	0.243	0.203	0.160	0.129
BLIP2 captions	0.315	0.292	0.277	0.251	0.217	0.187
Only raw captions						
Training set size	12.8M*	25.7M*	38.4M*	64.1M*	96.1M*	115M*
ImageNet accuracy	0.198*	0.260*	0.273*	0.254*	0.212*	0.188*
Average accuracy	0.277*	0.322*	0.328*	0.315*	0.285*	0.266*
Only BLIP2 captions						
Training set size	12.8M	25.6M	38.5M	64.1M	96.0M	115M
ImageNet accuracy	0.146	0.249	0.275	0.302	0.300	0.293
Average accuracy	0.254	0.333	0.356	0.370	0.365	0.366
Only BLIP2 captions, for top % based on cosine similarity of image and *raw* text						
Training set size	12.8M	25.7M	38.4M	64.1M	96.1M	115M
ImageNet accuracy	0.192	0.245	0.261	0.266	0.267	0.276
Average accuracy	0.280	0.330	0.346	0.342	0.349	0.356
Raw captions for top % + BLIP2 captions for the remaining examples						
Training set size	128M	128M	128M	128M	128M	128M
ImageNet accuracy	0.286	0.296	0.297	0.286	0.250	0.215
Average accuracy	0.360	0.357	0.365	0.349	0.323	0.293
Raw captions for top % + BLIP2 captions for the remaining examples, subject to the same cosine similarity threshold						
Training set size	30.5M	59.5M	83.6M	114M	127M	128M
ImageNet accuracy	0.267	0.310	**0.317**	0.296	0.251	0.212
Average accuracy	0.343	0.372	0.368	0.352	0.313	0.285
BLIP2 captions for top % + raw captions for the remaining examples, subject to the same cosine similarity threshold						
Training set size	17.1M	32.8M	47.7M	75.3M	105M	121M
ImageNet accuracy	0.212	0.272	0.298	0.310	0.298	0.285
Average accuracy	0.305	0.353	0.367	**0.376**	0.375	0.355
Concatenate raw & BLIP2 captions for top % + BLIP2 captions for the remaining examples, subject to the same cosine similarity threshold						
Training set size	30.5M	59.5M	83.6M	114M	127M	128M
ImageNet accuracy	0.250	0.287	0.299	0.286	0.269	0.262
Average accuracy	0.336	0.368	0.367	0.359	0.340	0.337
Top % raw captions + top % BLIP2 captions						
Training set size	25.6M	51.3M	76.9M	128M	-	-
ImageNet accuracy	0.238	0.285	0.297	0.300	-	-
Average accuracy	0.318	0.358	0.366	0.356	-	-
BLIP2 captions - top % intersect with examples from IN1k clustering						
Training set size	-	-	10.0M	16.4M	23.6M	27.1M
ImageNet accuracy	-	-	0.243	0.289	0.306	0.301
Average accuracy	-	-	0.310	0.343	0.360	0.344

Table 4: Summary of how various filtering and mixing strategies perform on ImageNet and on average across 38 evaluation tasks in DataComp, given a 128M candidate pool (`medium` scale). * indicates numbers obtained from Gadre et al. [18]. Note that all resulting training sets are trained for a fixed number of steps (128M samples seen) and other training variables (e.g., architecture, hyperparameters) are kept constant. Synthetic captions are generated using pre-trained BLIP2 model with top-K sampling (K = 50) and softmax temperature 0.75. We find that at this scale, approaches that yield the best ImageNet and average accuracies leverage a combination of raw and synthetic captions.

E Synthetic caption analysis

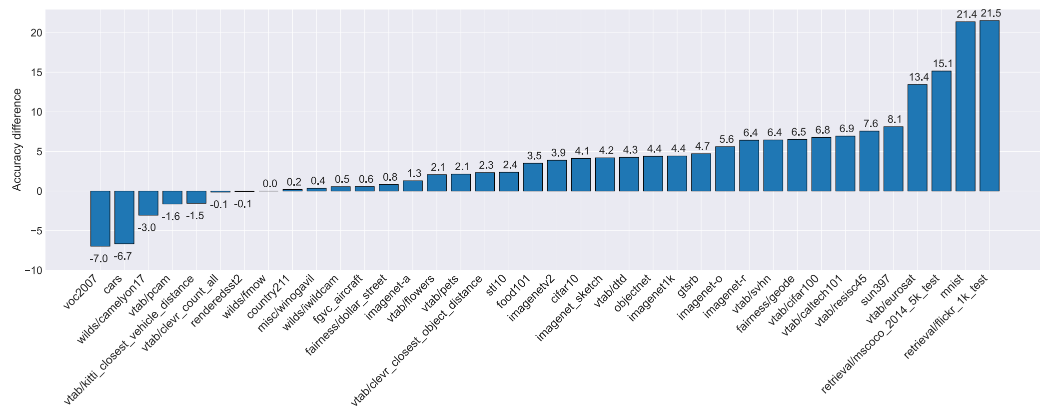


Figure 10: **We find that expanding a training set of filtered raw data by using BLIP2 captions for some of the discarded images improves performance on 30 out of 38 evaluation tasks, in addition to boosting average accuracy by 4%.** We compare performance on each task between training on top 30% of examples with raw captions (based on CLIP score) and training on the same set of examples but with the addition of BLIP2 captions for the remaining 70% images, filtered by the same CLIP score threshold. In Table 3, we have shown that adding BLIP2 captions improves ImageNet accuracy by 4.4% and average accuracy by 4%. With this breakdown, we find that the performance improvement applies to most of the tasks in the evaluation set, especially retrieval.

We investigate whether there are systematic differences in training with raw and generated text when it comes to recognizing certain object categories. To do so, we examine two CLIP models that perform similarly on ImageNet (i.e., $\pm 0.2\%$): one trained on only raw captions and one trained on only generated captions, both training sets have been filtered with CLIP score ranking to select the top 30% image-text pairs. In Figure 11, we analyze performance on each ImageNet class, categorized as either 'living' or 'non-living' thing based on where the classname synset is located in the WordNet hierarchy. We observe that class-wise classification performance is scattered evenly around the $y = x$ line, indicating that compared to web-crawled captions, synthetic captions do not exhibit a particular disadvantage on either 'living' or 'non-living' concepts.

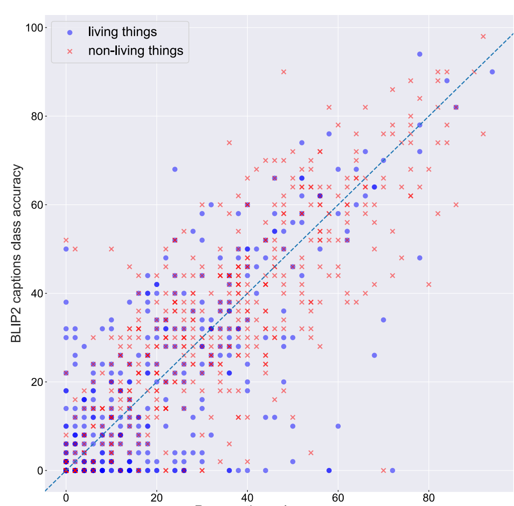


Figure 11: We break down per-class performance on ImageNet, between a CLIP model trained on only raw captions and one trained on only synthetic captions with similar overall ImageNet accuracy. We find no systematic trends in the performance of either model when it comes to classifying 'living' or 'non-living' things.

F Performance at Scale

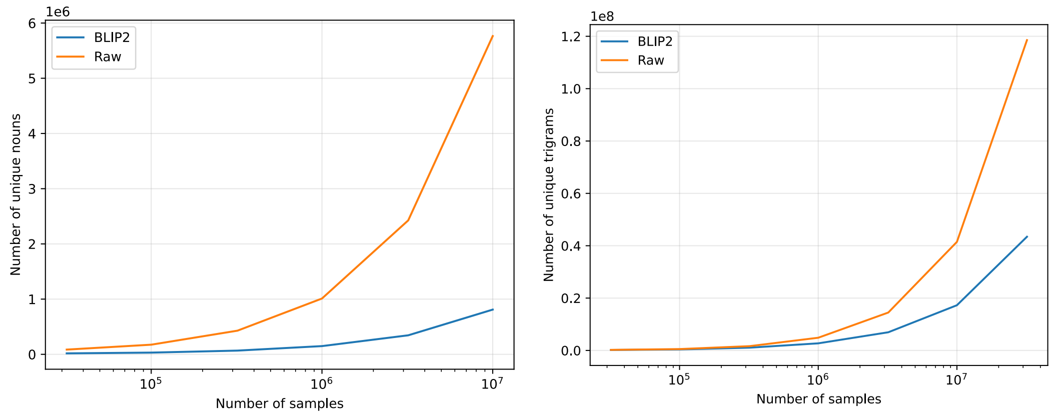
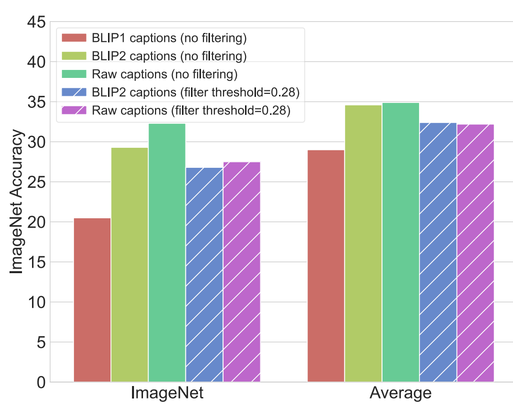


Figure 12: **Our simple analyses of text properties suggest that the text diversity provided by synthetic captions may not scale as well as that of raw captions scraped from the Internet.** We measure the number of unique nouns and unique trigrams for random subsets of BLIP2 and raw captions of various sizes. We observe that on both metrics, the scaling trend for synthetic captions is worse than that of raw captions. This increasing gap in data diversity may impact the performance benefits we can expect to obtain from using synthetic captions, when dealing with a larger scale of training data.

G Experiments with LAION-COCO

Our experiments with synthetic captions are partly inspired by the release of LAION-COCO dataset [45], which used BLIP [29] with various hyperparameter settings to caption LAION-5B data [46], and then selected the top synthetic caption for each image based on the cosine similarity output by OpenAI's CLIPs [40]. We pick a random set of 100M samples from LAION-COCO and train on this set using DataComp's `medium` scale configuration (i.e., 128M steps), with either only the raw captions or only the top BLIP captions that come with the dataset. We find that training on BLIP captions significantly lags behind training on raw captions, measured by both ImageNet and average accuracies (Figure 13). Consequently, a natural question is how much of this gap can be overcome with progress in image captioning models, e.g. the release of BLIP2.



	ImageNet accuracy	Average accuracy	Dataset size
BLIP	20.5	29.0	104M
BLIP2	29.3	34.6	104M
Raw	32.3	34.9	104M
BLIP2 (CLIP score ≥ 0.28)	26.8	32.4	41M
Raw (CLIP score ≥ 0.28)	27.5	32.2	41M

Figure 13: **BLIP2 significantly closes the performance gap between BLIP captions and raw captions on LAION-COCO; when controlled for noise level, the performance difference between using BLIP2 and using raw captions is actually negligible.** We use BLIP2 [30] to generate captions for 100M random samples from the LAION-COCO dataset [45], which already come with corresponding BLIP [29] captions. We find that advances in the BLIP model family help synthetic captions close the gap with raw captions, measured by the zero-shot performance of CLIP trained on the captions. After applying a cosine similarity threshold of 0.28 to the BLIP2 training pool, just like how LAION data was originally selected, we find that using either raw captions or synthetic captions for the resulting set of examples makes little difference (hatched columns).

We proceed to generating BLIP2 captions for the same set of 100M images, using only one configuration from the original hyperparameter grid in [45] due to compute constraints. Despite the lack of tuning, the new BLIP2 captions manage to close the previous ImageNet performance gap by 75% and come close to the average accuracy obtained from training on raw captions (see table in Figure 13). Since raw data in LAION was already filtered with a CLIP score threshold of 0.28 during the dataset construction, we next experiment with applying the same filtering to BLIP2 captions, in order to control for noise quality in the caption data. On the resulting 41M images, using BLIP2 captions is about as effective as using raw captions (-0.7% ImageNet accuracy and +0.2% average accuracy).

We note that LAION is considered a curated web dataset, with heavy cosine similarity filtering being one of the preprocessing steps. This in turn leads to approximately 90% of the raw data from Common Crawl to be discarded, according to Schuhmann et al. [46]. Since LAION only retains about 10% of the original candidate pool, similar experiments in DataComp [18] have shown that further CLIP score filtering on these top examples will only hurt performance. In addition, given that the selected raw captions are already relatively clean (measured via image-text cosine similarity), and there is no record of datapoints that were filtered out for further experimentation, we find LAION-COCO to be an unsuitable benchmark for studying the utility of synthetic captions. Our experiments here mainly seek to demonstrate that progress in image captioning models (e.g., the BLIP model family) can translate to better text supervision for CLIP training that rivals the effectiveness of using raw captions.

H Fairness implications of using synthetic captions

We examine zero-shot classification accuracy of predicting race and gender from face images in the Fairface dataset [26], for a model trained on only filtered raw captions, one trained on only filtered synthetic captions, and one trained on both. We acknowledge that there are limitations to these evaluations as race and gender should not be considered fixed categories.

With Fairface, we find that using synthetic captions improves the classification performance on the disadvantaged group (e.g. female) significantly, and reduces the performance gap between male and female groups while still boosting the overall performance on all race categories. We leave more extensive study of the fairness implications of using synthetic data (including and beyond gender biases) to future work.

Gender	Model	Race						
		Black	White	Indian	Latino/ Hispanic	Middle Eastern	South East Asian	East Asian
Male	Raw (top 30%)	93.0	88.8	91.2	90.8	92.3	85.3	81.3
	BLIP2 (top 30%)	87.2	73.7	77.2	74.9	78.6	72.0	64.0
	Raw (top 30%) + BLIP2 (70%, filtered)	90.5	75.0	79.7	79.4	81.1	72.4	65.3
Female	Raw (top 30%)	20.3	47.1	35.1	42.0	40.9	44.9	56.8
	BLIP2 (top 30%)	36.9	70.8	57.9	67.5	67.4	64.1	78.4
	Raw (top 30%) + BLIP2 (70%, filtered)	32.9	74.8	56.5	66.3	67.9	67.8	81.9
Overall	Raw (top 30%)	56.7	68.0	63.2	66.4	66.6	65.1	69.1
	BLIP2 (top 30%)	62.1	72.3	67.6	71.2	73.0	68.1	71.2
	Raw (top 30%) + BLIP2 (70%, filtered)	61.7	74.9	68.1	72.9	74.5	70.1	73.6

Table 5: Using synthetic captions in the training mix improves classification performance on Fairface for the minority group (i.e. female) across all race categories.