	Size (image-text pairs)	Quality of each Metadata	How was it attained	Where has it been used
Laion 5B	5.85 billion	quality of the captions is generated by 40 (or more) candidate captions for each image and then ranking them using OpenAI's CLIP ViT-L/14 & CLIP-Resnet50x64. First we ranked all candidates with ViT-L/14 and then we ranked the top-5 results again using Resnet50x64 (thus high quality)	Highly-acclaimed CLIP filtering images. detection scores for watermark and NSFW	1) Generative models: training image/text generative models, e.g autoregressive models like DALL-E or diffusion models like GLIDE 2)Models with contrastive losses: self-supervised training on image/text pairs using contrastive losses, e.g CLIP 3)Classification models: e.g,

				performing zero-shot classification by extracting pseudo labels from queries on the dataset Increased accuracy of CLIP (https://twitter.com/wightmanr/status/1527752308267724800)
CC12M https://arxiv.org/pdf/2102.08981.pdf	12 million	reliance on text during pre-training, which hurts the quality of its image representations two annotators to rate how well the given alt-text fits the image on a 1–5 scale: 1 (no fit), 2 (barely	CC12M, we keep the image-text filtering intact from CC3M, and relax the unimodal filters only	See CC3M

		fit), 3 (somewhat), 4 (good fit, but disfluent language), 5 (perfect)		
CC3M	3.3 million	Similar to CC12M, using human judgment. To evaluate the precision of our pipeline, we consider a random sample of 4K examples extracted from the test split of the Conceptual Captions dataset. We perform a human evaluation on this sample, using the same methodology described in Section	The construction of CC3M used three main filtering types: image-based, textbased, and image-text–based.	Practical applications of automatic image description systems include leveraging descriptions for image indexing or retrieval, and helping those with visual impairments by transforming visual signals into information that can be communicated via text-to-speech technology. The scientific challenge is seen as aligning, exploiting, and pushing further the latest improvements at the intersection of Computer Vision and Natural Language Processing.
Laion 400	400 million	We can use the CLIP filter tool along with this index to produce subsets using search terms efficiently. We	Same as Laion 5B	Same as Laion 5B

		also provide two 16GB knn indices of higher quality.		
Mscoco (Microsoft Common Objects in Context)	600 000	Mscoco provides a high quality dataset due to the following reasons: Object segmentation with detailed instance annotations Superpixel stuff segmentation 80 object categories, the “COCO classes”, which include “things” for which individual instances may be easily labeled 5 captions per image		
Coco	330K images (>200K labeled)	High Quality due to	This pipeline processes	object detection — model should

		Object segmentation *- Recognition in context* *- Superpixel stuff segmentation* *- 330K images (>200K labeled)* *- 1.5 million object instances* *- 80 object categories*	billions of Internet webpages in parallel. From these webpages, it extracts, filters, and processes candidate himage, captioni pairs.	get bounding boxes for objects, i. e. return list of object classes and coordinates of rectangles around them; objects (also called “things”) are discrete, separate objects, often with parts, like humans and cars; the official dataset for this task also contains additional data for object segmentation object/instance segmentation model should get not only bounding boxes for objects (instances/“things”), but also segmentation masks, i. e. coordinates of polygon closely around the

				object stuff segmentation — model should do object segmentation, but not on separate objects (“things”), but on background continuous patterns like grass or sky In computer vision, those tasks have tremendous usage, e. g. for self-driving vehicles (detection of people and other vehicles), AI-based

				security (human detection and/or segmentation) and object re-identification (object segmentation or removing background with stuff segmentation helps with checking object identity).
YFCC100M https://arxiv.org/abs/1503.01817	100 million media objects, of which approximately 99.2 million are photos and 0.8 million are videos	the dataset is represented by its metadata in the form its Flickr identifier, the user that created it, the camera that took it, the time at which it was taken and when it was uploaded, the location	All data attained by Flickr (image and video uploading site)	

		where it was taken (if available), and the CC license it was published under.		
Align dataset https://ai.googleblog.com/2021/05/align-scaling-up-visual-and-vision.html	1.8B image-text pairs		A surprising property of word vectors is that word analogies can often be solved with vector arithmetic. A common example, "king − man + woman = queen". Such linear relationships between image and text embeddings also emerge in ALIGN. Specifically, given a query image and a text string, we add their ALIGN embeddings together and use it to retrieve relevant images using cosine similarity, as shown below. These examples not only demonstrate the compositionalit	

			y of ALIGN embeddings across vision and language domains, but also show the feasibility of searching with a multi-modal query. For instance, one could now look for the "Australia" or "Madagascar" equivalence of pandas, or turn a pair of black shoes into identically-looki ng beige shoes. Also, it is possible to remove objects/attribut es from a scene by performing subtraction in the embedding space, shown below.	
Visual Genome (relationships) https://arxiv.org/pdf/1602.07332.pdf	100K images where each image has an average of 21 objects, 18 attributes, and 18 pairwise relationships between objects.		These are raw texts without any restrictions on length or vocabulary. Next, we extract objects, attributes and relationships from our descriptions. Together, objects,	

			attributes and relationships fabricate our scene graphs that represent a formal representation of an image. VQA: annotated with one or more question answers	
Redcaps https://arxiv.org/pdf/2111.11431.pdf	12M image-text pairs		Uses images and captions in manually curated subreddits	

Align VS CLIP

ALIGN is also a strong image representation model. Shown below, with frozen features, ALIGN slightly outperforms CLIP and achieves a SotA result of 85.5% top-1 accuracy on ImageNet. With fine-tuning, ALIGN achieves higher accuracy than most generalist models, such as BiT and ViT, and is only worse than Meta Pseudo Labels, which requires deeper interaction between ImageNet training and large-scale unlabeled data.

Model (backbone)	Acc@1 w/ frozen features	Acc@1	Acc@5
WSL (ResNeXt-101 32x48d)	83.6	85.4	97.6
CLIP (ViT-L/14)	85.4	-	-
BiT (ResNet152 x 4)	-	87.54	98.46
NoisyStudent (EfficientNet-L2)	-	88.4	98.7
ViT (ViT-H/14)	-	88.55	-
Meta-Pseudo-Labels (EfficientNet-L2)	-	***90.2***	***98.8***
ALIGN (EfficientNet-L2)	***85.5***	88.64	98.67

ImageNet classification results comparison with supervised training (fine-tuning).

I also think we should cite the YFCC100M at the start because it is one of the first multimodal datasets.

The Align dataset was created with alt text of the web images; sometimes the alt-text is extremely inaccurate so, like the CC3M, there is a heavy filtering process and post-processing. Instead, the Align dataset only applies minimal frequency-based filtering. The result is a much larger but noisier dataset of 1.8B image-text pairs.

This highlights the efficiency of Laion 5B as it takes the best of both worlds by supplying a decent set of image-text pairs (quality) and 5.85 image-text pairs (quantity). Using CLIP as a filtering process, we ensure its quality.

“We hypothesize that the generated captions match (& sometimes even surpass) the average quality of the human captions of MS COCO (which are sometimes also far from perfect) in most cases, but sometimes (in less than <10%) contain obvious mistakes, that humans would not make, because deeper kind of world knowledge & „common sense“ would be necessary in those cases.”
(something from Laion 5B I think worth mentioning with regards to mscoco vs Laion 5B)

Also I think we could mention how the accuracy of our image-text pairs is bolstered by BLIP

Redcaps is an interesting dataset that doesn’t rely on alt-texts (it tries to minimize the filtering process involved in many datasets), but I think there’s a lot of limitations to it, namely how Reddit posts can be opinionated so the labels are subjective. This makes the labels inaccurate so again a complex filtering process is still needed to filter out the good image-text pairs.