

Sieve: Multimodal Dataset Pruning Using Image Captioning Models

Anas Mahmoud^{1,2*} Mostafa Elhoushi¹ Amro Abbas[†] Yu Yang³

Newsha Ardalani¹ Hugh Leather¹ Ari S. Morcos⁴

¹FAIR at Meta ²University of Toronto ³UC Los Angeles ⁴DatologyAI

Abstract

Vision-Language Models (VLMs) are pretrained on large, diverse, and noisy web-crawled datasets. This underscores the critical need for dataset pruning, as the quality of these datasets is strongly correlated with the performance of VLMs on downstream tasks. Using CLIPScore from a pretrained model to only train models using highly-aligned samples is one of the most successful methods for pruning. We argue that this approach suffers from multiple limitations including: false positives and negatives due to CLIP’s pretraining on noisy labels. We propose a pruning signal, Sieve, that employs synthetic captions generated by image-captioning models pretrained on small, diverse, and well-aligned image-text pairs to evaluate the alignment of noisy image-text pairs. To bridge the gap between the limited diversity of generated captions and the high diversity of alternative text (alt-text), we estimate the semantic textual similarity in the embedding space of a language model pretrained on unlabeled text corpus. Using DataComp, a multimodal dataset filtering benchmark, when evaluating on 38 downstream tasks, our pruning approach, surpasses CLIPScore by 2.6% and 1.7% on medium and large scale respectively. In addition, on retrieval tasks, Sieve leads to a significant improvement of 2.7% and 4.5% on medium and large scale respectively.

1. Introduction

Contrastive Language-Image Pre-training (CLIP) [32] models have shown great success in solving zero-shot image classification and multimodal retrieval tasks. In addition, many foundational Vision-Language Models (VLMs) use pretrained CLIP encoders to condition image generation on CLIP text embeddings [33] in retrieval augmented vision-language models [13, 42], and to align modalities including audio, depth, and thermal with language through CLIP image embeddings [10]. Therefore, the quality of CLIP representations can influence the performance of many VLMs.

To pretrain CLIP, billions of image-text pairs are collected using common crawl. The raw data is highly diverse but contains many noisy image-text pairs, including low quality images, low quality alternative text (alt-text), and misaligned image-text pairs. Pretraining CLIP models on noisy data can have adverse effects on the learned representations, thus leading to poor performance on downstream tasks [1].

To address this challenge, researchers have developed data pruning methods to remove low quality image-text pairs. Heuristics that filter out image-text pairs based on image dimensions, aspect ratio, alt-text length, and complexity are commonly used [9, 36] to reduce noise, but can also limit the diversity of the dataset [26]. Methods that use images or class names from datasets, like ImageNet, to sample semantically similar image-text pairs can lead to higher accuracy on downstream tasks [40], but limit the diversity of the selected samples as they sample image-text pairs close to a specific dataset.

One of the most effective pruning methods, CLIPScore [11, 36], computes the cosine similarity between image and text embeddings using a pretrained CLIP model. This score is then used to rank the alignment of image-text pairs. However, as shown in Figure 1, using CLIPScore can lead to false positives – samples that are poorly aligned but have high CLIPScore (i.e., bad samples) due to spurious correlations learned by the pretrained CLIP model [41]. In addition, using CLIPScore can lead to false negatives – samples that are aligned but have low CLIPScore (i.e., hard samples) due to the poor discrimination between hard and bad samples. Excluding hard samples and including bad samples can negatively affect the generalization of CLIP image and text encoders.

The goal of this work is to reduce both false negatives and positives induced by CLIPScore ranking by relying on an image-captioning model pretrained on small, diverse, and well-aligned image-text pairs. As shown in Figure 1, Sieve can reduce false positives or false negatives in different scenarios, such as samples where CLIPScore focuses on geographic or chronological context of an image rather than the content of an image. As depicted in Figure 2, we evalu-

*Work done during internship at Meta

†Work done during an AI residency at Meta

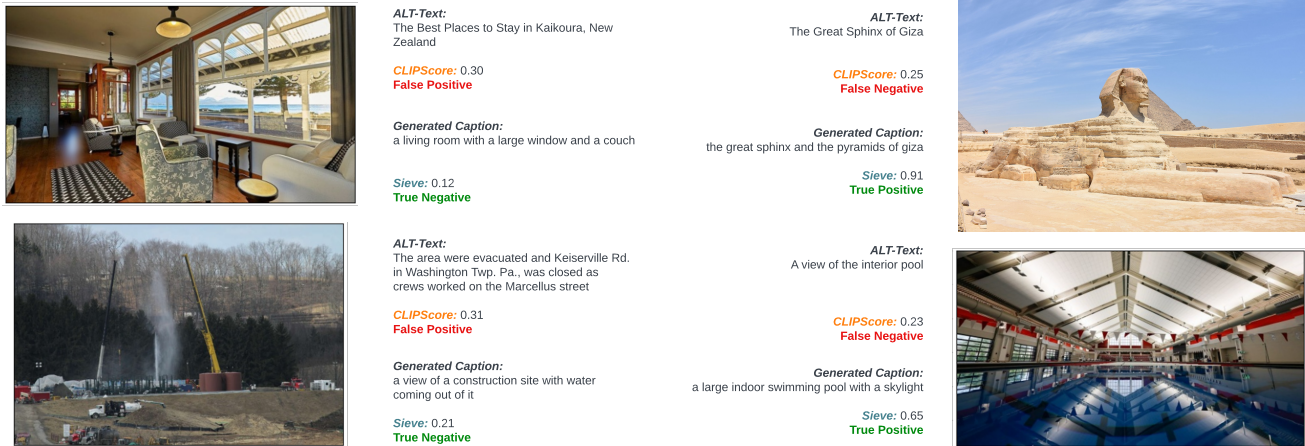


Figure 1. Examples of image-text pairs in which the scores of pretrained CLIP models, a commonly used image-text data filtering approach, fail to measure their alignment. Our proposed approach, Sieve, provides an accurate alignment score using a caption generator and sentence transformer. **Top left** and **bottom left**: Examples of false positives where alt-text describes concepts that are not found or unrelated to the image. CLIP is trained on similar noisy image-text pairs, thus, it assigns a relatively high score. Sieve can detect that such image-text pairs are misaligned. **Top right** and **bottom right**: Examples of false negatives where images are aligned with the alt-text but are assigned low CLIP scores, either due to the low likelihood of these pairs in the pretraining data, or because CLIP may have seen similar images aligned with other noisy labels. Sieve can detect that such image-text pairs are well-aligned and selects them for pretraining.

ate the alignment of web-crawled image-text pairs by first, generating multiple captions for each image using nucleus sampling [12], followed by removing phrases that describe the medium (e.g., “an image of”, “a photo of”) rather than visual concepts. Finally, to evaluate semantic similarity between the limited diversity of generated captions and the high diversity of alt-text, we utilize the embedding space of a lightweight sentence transformer pretrained on unlabeled text corpus. The alignment between the generated captions and the alt-text is then used as a proxy for image-text alignment. Our goal is to fuse this alignment signal with CLIP-Score to minimize false positives and negatives, leading to a more aligned pretraining dataset. To evaluate the effectiveness of our proposed pruning method, we utilize the DataComp [9] benchmark, which fixes the pretraining hyperparameters of CLIP and provides multiple candidate pools of noisy image-text data for pretraining CLIP models. The goal is to select a subset of noisy image-text data that leads to the best performance on 38 downstream tasks. Using image-captioning model alignment scores fused with CLIP-Score, we surpass CLIPScore filtering by 2.6% and 1.7% on the average of 38 downstream tasks on *medium* and *large* scale pool respectively. In addition, on multimodal retrieval tasks, our approach leads to an improvement of 2.7% and 4.5% on *medium* and *large* scale respectively.

2. Related Work

Heuristics are basic filtering methods including: filtering non-English alt-text using fastText [15], filtering alt-text

with a few words [9, 36], filtering alt-text with low text complexity [30], and filtering images by size or aspect ratio [9]. A combination of these unimodal filtering approaches has been explored by DataComp [9]. An example of a multimodal filtering approach is text spotting: detecting and recognizing text in images and filtering image-text pairs with high overlap between spotted text (text detected in image) and alt-text (associated label of image) [30].

Datasets as Priors was proposed in DataComp [9], relying on sampling image-text pairs that are semantically similar to diverse and curated datasets like ImageNet [5]. Text-based sampling selects image-text pairs with alt-text overlapping one of the ImageNet classes. CiT [40] uses cosine similarity to filter alt-text that are similar to ImageNet classes. Image-based sampling approaches encode images from the unfiltered candidate pool using the OpenAI CLIP’s ViT-L/14 vision encoder, and clusters the images into 100,000 groups using FAISS [14]. Then, embeddings of ImageNet training samples are used to keep the closest cluster to each sample. The main limitation of such approaches is that they bias the CLIP model and may not generalize well to new downstream tasks. We argue that selecting samples that match the distribution of downstream tasks encourages overfitting to the evaluation set and, thus limits generalization to other downstream tasks. Our approach, Sieve, does not use any dataset as a prior.

Pretrained VLMs One of the most successful methods for evaluating image-text alignment is CLIPScore [11]. LAION filtering [36] uses an OpenAI CLIP model [32] pre-

trained on 400 million image-text pairs to evaluate image-text alignment of large webscale datasets, and filter out samples with the lowest CLIPScore. Filtering using CLIPScore can suffer from false negatives, which leads to filtering out hard informative samples, and false positives, which leads to including misaligned samples. Another approach proposes a non-filtering approach that utilizes pre-trained VLMs [26], using large image-captioning models like BLIP2 [21] to replace alt-text labels with descriptive synthetic captions. The synthetic captions are then used to train CLIP models. The authors [26] demonstrate that at scale, the improvement of synthetic captions is capped by the limited diversity of generated captions compared to the high diversity of noisy text labels. Compared to [26], we do not alter the original alt-text and thus our focus is on the dataset pruning challenge.

Concurrent Work Inspired by text spotting [30], T-MARS [24] is concurrent work that detects and masks text regions in images before computing CLIPScore, resulting in improved visual representations. We reason that T-MARS is orthogonal to our approach as we can apply the same masking before calculating Sieve. In addition, Devil is In the Details (DID) [44], utilizes a combination of multiple filters and sampling approaches, which is orthogonal to our approach as Sieve can be combined to other filters. Moreover, most of DID’s accuracy improvements is from aligning the selected data distribution with the DataComp [9] evaluation set (downstream tasks), which overfits downstream tasks as discussed in previous sections. More importantly, T-MARS [24] and DID [9] only perform pruning experiments on the *medium* scale of DataComp [9], while Sieve shows results on both *medium* and *large* scales. Finally, DFN [8], uses CLIPScore from a CLIP model pretrained on a private 357 million human verified image-text labels, which is $25\times$ bigger than the dataset our selected captioning models were pretrained on, and is bigger than DataComp’s *medium* scale pool.

3. Methodology

Let $\mathcal{D} = \{(I_i, T_i)\}_{i=1}^N$ denote an uncurated dataset consisting of N image-text pairs crawled from the web. Our goal is to curate a dataset, $\mathcal{D}' = \{(I_{i'}, T_{i'})\}_{i'=1}^{N'}$, that is a subset of the pool, $\mathcal{D}' \subseteq \mathcal{D}$, $N' \leq N$, to train a new CLIP model from uninitialized weights, Θ_0 , to new weights, Θ' :

$$\Theta' = \text{train}(\Theta_0, \mathcal{D}') \quad (1)$$

For a given scoring function, f , that maps an image-text pair to a scalar value, $s = f(I_i, T_i)$, we express a pruning function that selects a fraction, k , of dataset, \mathcal{D} , using function, f :

$$\text{prune}_f(\mathcal{D}, k) \quad \text{s.t. } 0 \leq k \leq 1, f : (I_i, T_i) \rightarrow \mathbb{R} \quad (2)$$

where $\text{prune}_f(\mathcal{D}, k)$ applies function, f , on each image-text sample in \mathcal{D} to obtain a score for each sample, ranks the scores in descending order, and returns a set of the top k portion of the samples.

One common approach for pruning is CLIPScore [9, 36]. Let E be a CLIP model consisting of an image encoder, E_{image} , that maps an image, I , to an embedding vector, $E_{\text{image}}(I) \in \mathbb{R}^d$, and a text encoder, E_{text} , that maps a text sample, T , to an embedding vector, $E_{\text{text}}(T) \in \mathbb{R}^d$. CLIPScore is a measure of alignment between I and T , and is defined as:

$$f_{\text{CLIP}}(I, T) = \langle E_{\text{image}}(I), E_{\text{text}}(T) \rangle \quad (3)$$

where $\langle \mathbf{x}, \mathbf{y} \rangle$ is the cosine similarity between two vectors, \mathbf{x} and \mathbf{y} , which is defined as the dot product of the l_2 normalized vectors. The most common CLIP model used for pruning is pretrained on 400 million noisy image-text pairs [36]. Our proposed pruning method, Sieve attempts to minimize the false positives and negatives induced by CLIPScore filtering. Sieve consists of two main components: Image-Captioning and Sentence Transformer.

Image-Captioning Let G be a captioning model that generates text, T_i^G , describing the content of image, I_i :

$$T_i^G = G(I_i) \quad (4)$$

Given a captioning model pre-trained on a small, representative and well-aligned dataset of image-text pairs, we are interested in estimating the alignment between image-text pairs sampled from a very large, diverse but noisy dataset. The alignment score can then be used as a ranking metric for dataset filtering. We hypothesize that:

- The probability of generating a caption that is semantically similar to the alt-text from an aligned pair is much higher than that from a misaligned pair.
- The probability of generating a caption that is semantically similar to a hard alt-text is higher than generating a caption that is semantically similar to a misaligned alt-text. Here, a hard alt-text is a text label with low likelihood with respect to the captioning model, but is aligned with the image content.

As images can contain multiple objects with complex attributes and relationships, there exist multiple ways to describe their content. Given the inherent many-to-many relationship between images and text labels, our goal is to increase the probability of generating a caption that matches an aligned alt-text. To achieve this, we utilize nucleus sampling [12], a decoding strategy used to sample multiple captions, r , per-image:

$$G(I, r) = \{T_0^G, T_1^G, \dots, T_{r-1}^G\} \quad (5)$$

Sentence Transformer Given an image, its alt-text, and a set of generated captions, our goal is to estimate the align-

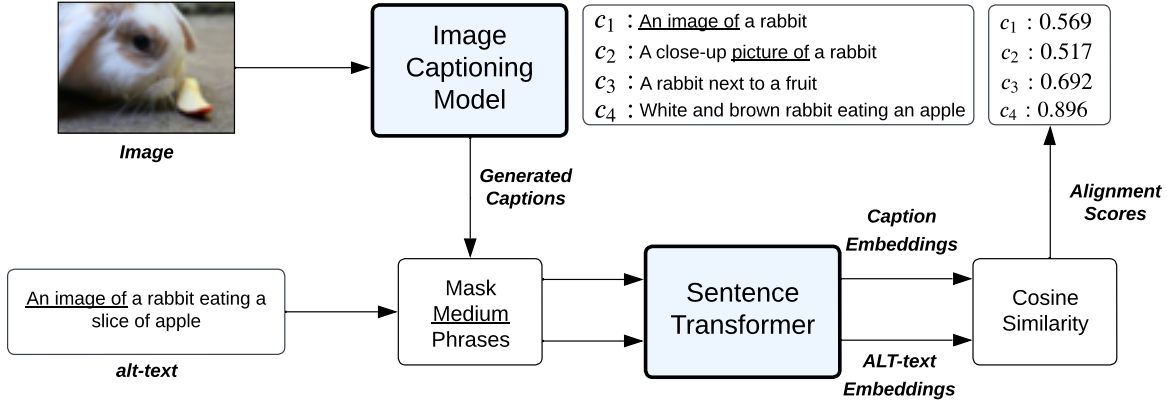


Figure 2. Our proposed framework enables dataset pruning using image-captioning models. To evaluate the alignment of a noisy image-text pair, we generate multiple captions per image using nucleus sampling. Then medium phrases, like “an image of” or “a photo of”, are masked from alt-text and generated captions. Finally, a lightweight sentence encoder is used to semantically compare the generated captions with alt-text.

ment between the generated captions and the alt-text. However, there is a very large diversity gap between the generated captions and the highly diverse alt-text as measured by the number of unique nouns and trigrams [26]. On the other hand, constructing a large, diverse and curated image-text dataset is expensive, which limits the diversity of the generated captions. We propose to bridge this gap by utilizing a light-weight sentence similarity model to encode the alt-text and the generated captions. We expect the semantically similar alt-text and generated caption pairs to be closely clustered in the embedding space compared to semantically distinct pairs. We reason that the rich semantic textual embedding space of the sentence similarity model enables pretraining the captioning model only on a small but curated image-text dataset. Thus, we rely on the semantic understanding of the sentence similarity model to bridge the gap between the limited diversity in the generated captions and the highly diverse alt-text labels.

To estimate the alignment score, we compute the cosine similarity between embeddings of each generated caption and text label. Let S be a language model that encodes a text sample, T , to a vector, $S(T) \in \mathbb{R}^d$. We define the alignment between two text samples, T_a and T_b , as the cosine similarity between their language model encodings:

$$\langle S(T_a), S(T_b) \rangle \quad (6)$$

This estimate can then be used as a proxy for the image-text alignment of an image, I , and text, T :

$$\langle S(G(I)), S(T) \rangle$$

If the image captioning model generates r different caption candidates for an image, I , we can use the maximum alignment between each of the generated captions,

$G(I, r) = \{T_0^G, T_1^G, \dots, T_{r-1}^G\}$, and a text sample, T :

$$\max_{T_j^G \in G(I, r)} \langle S(T_j^G), S(T) \rangle \quad (7)$$

In literature, there are different models and approaches to obtain text embeddings. [16] uses the average of N-gram features of each word in a text sample to obtain an embedding. A more common option is to use the logits of the last token generated by a decoder-only language model, which is the approach taken with CLIP’s text encoder [32], in GPT-1 [31], as well as in [1]. Encoder-only models, such as BERT [6] or RoBERTa [23], can also be used, where the embedding vector may be either the logits of the classification token, or the average pool of the logits of all tokens. Although such language models may have strong generation or classification capabilities, they were not optimized for sentence similarity tasks, but either for next word prediction (i.e., causal language modeling) or masked word prediction (i.e., masked language modeling) tasks. Therefore, their embeddings may not be ideal to measure alignment between sentences. More importantly, such models are large in size and hence slow to infer on large datasets. A language model finetuned on a sentence similarity task, such as SNLI [3], aligns with the goal of estimating semantic textual similarity between alt-text and captions. We find that sentence similarity models [34] pretrained using a self-supervised instance discrimination task on unlabeled corpus of sentences perform well in estimating the alignment between text pairs, and are lightweight in size and latency (e.g., tens of millions of parameters in contrast to billions of parameters of performant decoder-only large language models).

Masking Medium Words Phrases such as “image of”, “picture of”, or “photo of” can appear in either alt-text or

generated captions. We refer to such phrases as “medium phrases”, as they describe the medium rather than the contents of an image. We notice that the existence of such medium phrases adds noise to the sentence similarity score, as shown in Figure 3. A pair of sentences that each have a medium phrase are assigned a misleadingly high sentence similarity score by a sentence transformer, as they have been trained on a wide and diverse corpus of text, rather than on image captions. Hence, the existence of medium phrases may increase their attention to the topic of images or media, rather than the topic of the content of such images. Therefore, we neutralize the effect of medium phrases by removing them from both alt-text and generated captions. We express the operation of masking medium words in a text sample, $M(T)$, on text, T , as masking all possible contiguous subsequences of the text, where masking on a phrase, t , removes it if it is in the pre-determined list of medium phrases, $\mathcal{M} = \{\text{“image of”, “picture of”, “photo of”, ...}\}$, or keeps it otherwise.

Putting it all together, we define the Sieve score function as:

$$f_{\text{Sieve}}(I, T) = \max_{T_j^G \in G(I, r)} \langle S(M(T_j^G)), S(M(T)) \rangle \quad (8)$$

The dataset pruned using Sieve with the top k portion of its samples can be expressed as:

$$\mathcal{D}_{\text{Sieve}, k} = \text{prune}_{f_{\text{Sieve}}}(\mathcal{D}, k) \quad (9)$$

We summarize our approach in Figure 2 and as pseudocode in Algorithm 1 in the Appendix.

4. Experiments

4.1. Training and Evaluation

We utilize the DataComp benchmark to evaluate the utility of image-captioning models for multimodal dataset pruning. Two candidate pools are considered, the *medium* and the *large* scale, consisting of 128 million and 1.28 billion image-text pairs, respectively. To train CLIP models, we use DataComp’s hyperparameters and architectures to standardize training [9]: 5×10^{-4} learning rate, 500 iterations warmup, AdamW optimizer, for *medium* scale: ViT-B/32 image encoder [7], batch size 4096, 128M training samples as a compute budget, and for *large* scale: ViT-B/16 image encoder, batch size 8192, 1.28B training samples as a compute budget. We evaluate the zero-shot performance on 38 downstream tasks, including classification and retrieval tasks [18, 32, 45].

For our captioning model, we utilize BLIP with ViT-B/16 image encoder pretrained on 14 million image-text pairs [20], including highly curated web datasets: Conceptual Captions [37], Conceptual 12M [4] and SBU captions [29]; as well as small human-annotated datasets:

COCO, and Visual Genome [17]. To compute the alignment between generated captions and alt-text, we use a lightweight distilled sentence transformer, all-MiniLM-L6-v2 [39], further finetuned using self-supervised contrastive learning on unlabeled text corpus.

4.2. Main Results

Table 1 reports multiple baselines from DataComp [9], including applying no filtering, basic filtering, and CLIPScore filtering. On the *medium* scale, Sieve with an image-captioning model pretrained on 30 times less but curated data surpasses CLIPScore by 1.8% on average. In addition, we demonstrate that Sieve’s alignment score provides a complimentary signal to CLIPScore by fusing Sieve with CLIPScore. We apply min-max normalization to Sieve alignment scores and CLIPScore independently, then take the per-sample weighted average of both scores:

$$f_{\text{Sieve+CLIP}}(I, T) = (1 - \alpha) \times \bar{f}_{\text{Sieve}}(I, T) + \alpha \times \bar{f}_{\text{CLIP}}(I, T)$$

$$\text{s.t.} \quad \bar{f}(I, T) = \frac{f(I, T) - \min_{(I_i, T_i) \in \mathcal{D}} f(I_i, T_i)}{\max_{(I_i, T_i) \in \mathcal{D}} f(I_i, T_i) - \min_{(I_i, T_i) \in \mathcal{D}} f(I_i, T_i)}$$

where the weight α used in the reported results is 0.5. Finally, we select the top 20% of samples. We observe that the fused approach improves CLIPScore average performance by 2.6% and 1.7% on *medium* and *large* scale respectively, demonstrating the complementarity of Sieve’s alignment score. Moreover, Sieve without fusion achieves the best performance on retrieval tasks on both *medium* and *large* scale experiments. Relative to CLIPScore, a significant portion of Sieve’s retrieval gain is effectively transferred when fusing Sieve with CLIPScore (i.e., +2.7% and +4.5% on *medium* and *large* scale). This is also promising as retrieval performance is critical for retrieval augmented VLMs [13, 42].

4.3. Per-Task Performance

Figure 5 shows the change in accuracy introduced by Sieve as well as Sieve+CLIPScore on each task compared to CLIPScore on *medium* scale, and Figure 6 of the Appendix shows for the *large* scale. We observe that in addition to outperforming on image retrieval tasks, Flickr [43], and MS COCO [22], Sieve’s greatest performance boost comes from WingoGAViL [2], a retrieval task which requires diverse reasoning skills, including general knowledge, common sense, and abstraction. This high performance can be attributed to Sieve’s preference towards keeping samples where the alt-text correctly describes visual concepts and

Sentence 1	Sentence 2	Sentence Similarity
A picture of a cat	A picture of a happy dog	0.520
	An animal	0.469
	A mammal	0.399
a cat	a happy dog	0.389
	An animal	0.569
	A mammal	0.502

Sentence 1	Sentence 2	Sentence Similarity
An image of a beautiful park	Image of a building	0.484
	An image of a factory	0.359
	Trees and grass	0.357
a beautiful park	a building	0.338
	a factory	0.248
	Trees and grass	0.367

Figure 3. Masking medium phrases improves the ranking of sentence similarity scores. On the **left**, sentence pairs with misleadingly high (or low) sentence similarity due to the existence (or absence) of medium phrases are highlighted in dark red (or light red). On the **right**, similarity scores that are more aligned with semantics are highlighted in dark green. The sentence similarity scores are computed using the all-MiniLM-L6-v2 sentence transformer [39].

Scale	Filtering	Dataset Size	ImageNet	ImageNet dist. shifts	VTAB	Retrieval	Average over 38 datasets
Medium (128 Million)	No Filtering	128M	17.6	15.2	25.9	21.9	25.8
	Basic Filtering	30M	22.6	19.3	28.4	25.1	28.5
	LAION Filtering	13M	23.0	19.8	30.7	23.3	29.2
	CLIPScore	38M	27.3	23.0	33.8	25.1	32.8
	Sieve	24M	29.4	25.0	35.2	28.9	34.6
	Sieve+CLIPScore	24M	30.3	25.4	36.2	27.8	35.4
Large (1.28 Billion)	No Filtering	1.28B	45.9	37.8	42.6	41.9	43.7
	Basic Filtering	298M	51.6	42.3	44.6	48.0	45.8
	LAION Filtering	130M	55.3	45.3	51.0	49.5	50.1
	CLIPScore	384M	57.8	47.4	53.8	46.6	52.9
	Sieve	235M	57.3	47.8	52.0	52.0	52.3
	Sieve+CLIPScore	235M	59.7	49.1	54.8	51.1	54.6

Table 1. Zero-shot performance of CLIP models pretrained using various filtering strategies on *medium* and *large* scale pools of the DataComp benchmark. Sieve fused with CLIPScore beats CLIPScore by 2.6% and 1.7% on *medium* and *large* scale respectively. In addition, on retrieval tasks, Sieve achieves best performance on both scales and after fusing with CLIPScore, leads an improvement on retrieval over CLIPScore of 2.7% and 4.5% on *medium* and *large* scale respectively.

their attributes and relations. Sieve, especially when combined with CLIPScore, significantly outperforms on medical diagnosis tasks, Camelyon17 and PatchCamelyon.

Sieve mainly underperforms in tasks requiring parsing text from images, such as MNIST [19], SVHN [25], and Rendered SST-2 [28], concluding that Sieve is less likely to select image-text pairs that are useful for OCR tasks. In addition, Sieve underperforms CLIPScore on context-based tasks like Country211 [27], a task assessing the geolocation capability of visual representations, demonstrating Sieve’s preference towards selecting samples based on the alignment of alt-text with visual concepts rather than context. Interestingly, when fusing with CLIPScore, we improve the performance of all these tasks while retaining the advantage

of Sieve especially at *large* scale (see Figure 6).

4.4. Ablation Studies

We conduct studies on the *medium* scale pool and report the average of three runs per experiment.

Pretraining data-distribution We study the effect of the pretraining data distribution used to train the captioning model on the quality of the alignment score. This is measured based on the downstream performance of the CLIP model trained using the selected image-text pairs. Two pretraining data distributions proposed by BLIP [20] are investigated. The first uses 14 million curated image-text pairs, while the second uses an additional 115 million web images with noisy alt-text [36]. Although the original BLIP

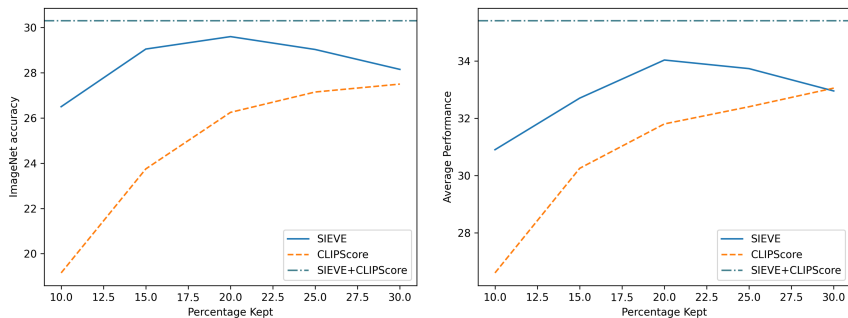


Figure 4. Evaluating CLIP models pretrained on different fractions of the top ranked samples based on our proposed approach (Sieve), CLIPScore, and fusing Sieve with CLIPScore (Sieve+CLIPScore), on *medium* scale.

Percentage Kept	Caption Generator Pretraining Data	ImageNet	ImageNet dist. shift	VTAB	Retrieval	Average over 38 datasets
10	BLIP-129M	23.00	20.60	30.20	21.40	30.00
	BLIP-14M	26.50	22.50	32.10	23.75	30.90
15	BLIP-129M	25.95	22.80	32.80	24.40	32.40
	BLIP-14M	29.05	24.60	33.35	26.95	32.70
20	BLIP-129M	27.85	23.65	33.45	26.35	33.05
	BLIP-14M	29.60	24.93	35.07	28.57	34.03

Table 2. Effect of a caption generator’s pretraining data-distribution on Sieve. The 14M pretraining dataset consists of curated image-text pairs, while the 129M dataset includes an additional 115M noisy image-text pairs from LAION [36].

work reports higher captioning performance when pretraining on 115 million samples, our results in Table 2 indicate that for the purpose of dataset pruning, using curated image-text pairs results in a better alignment score than using a much larger noisy dataset. This highlights the importance of using a captioning model pretrained on higher quality data for large-scale dataset pruning.

Captioning models We conduct experiments to investigate the generalizability of Sieve to other captioning models. In Table 3, we compare GIT [38], which utilizes an image encoder, and a text decoder captioning model, pretrained on 10 million image-text pairs to BLIP [20], which utilizes an image encoder, text encoder and a text decoder and is pretrained on 14 million image-text pairs. We observe that fusing Sieve with CLIPScore based on GIT and BLIP improves retrieval performance by 3.13% and 3.83% and improves average performance on 38 downstream tasks by 1.07% and 2.52% respectively.

Text embedding space In Table 4, we ablate over embeddings from different text models and show that embeddings from our selected sentence transformer perform better than embeddings from CLIP and BLIP text encoders. The CLIP text encoder was pretrained along with the CLIP vision encoder to map text and images to the same embedding space, and is used in diffusion models to condition image genera-

Filtering	ImageNet	VTAB	Retrieval	Avg
CLIPScore	27.13	33.90	24.17	32.63
Sieve-GIT	27.47	33.20	27.27	32.40
+CLIPScore	28.90	34.83	27.30	33.70
Sieve-BLIP	29.60	35.07	28.57	34.03
+CLIPScore	30.35	35.90	28.00	35.15

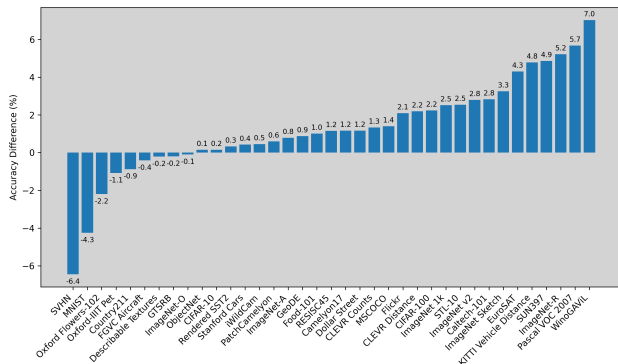
Table 3. Effect of captioning models, GIT [38] pretrained on 10M image-text pairs, and BLIP [20] pretrained on 14M pairs

tion [33]. However, we observe that the CLIP text encoder suffers from poor semantic textual understanding, leading to a large drop in accuracy when used as a caption similarity measure. BLIP’s text encoder performs better than that of CLIP, but the lightweight sentence transformer specifically pretrained on aligning semantically similar texts performs significantly better with $\geq 2\%$ improvements across various task types. In Figure 7 in the Appendix we show how cosine similarities of sentence similarity models result in better semantic textual clustering compared to CLIP and BLIP text encoders.

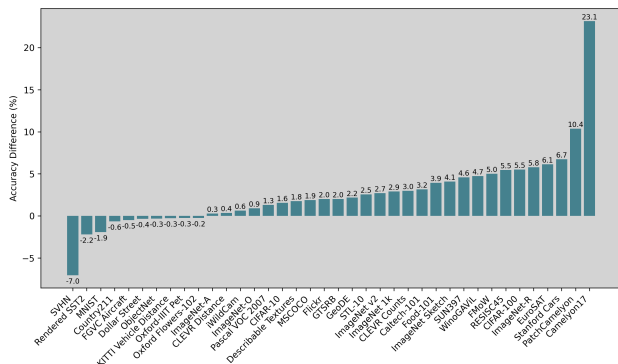
Pruning percentage We study the effect of the fraction of samples selected for pretraining. For each experiment, we compute the Sieve alignment score and CLIPScore for each

Text Encoder	ImageNet	ImageNet dist shift	VTAB	Retrieval	Average over 38 datasets
CLIP	18.00	15.65	26.90	20.95	25.90
BLIP	27.20	22.10	32.70	25.45	31.85
Sentence Transformer	29.60	24.93	35.07	28.57	34.03

Table 4. Effect of the sentence encoder on the performance of Sieve. CLIP uses the text encoder pretrained on 400M samples [36], BLIP uses the text encoder pretrained on the curated 14M samples defined in [20], and Sentence Transformer uses a language model pretrained on unlabeled text corpus [39]. Each encoder encodes generated captions and the alt-text where the textual semantic alignment is computed.



(a) Sieve gain over CLIPScore on *medium* scale pool.



(b) Sieve+CLIPScore gain over CLIPScore on *medium* scale pool.

Figure 5. The relative performance gain of Sieve and Sieve+CLIPScore relative to CLIPScore on 38 downstream tasks on the *medium* scale pool.

sample. The top- $k\%$ and pretraining CLIP models are then selected. Here, $k\%$ is set to 10%, 15%, 20%, 25% and 30%. Finally, we report the zero-shot performance on ImageNet and the average on 38 tasks in Figure 4. We observe that Sieve achieves the best performance using 20% of the data, while CLIPScore peaks at 30% (similar to results reported in 9). Hence, pruning using Sieve achieves better performance with less data, compared to CLIPScore.

Number of generated captions and fusion with CLIP-Score We study the effect of using multiple captions per image to maximize the alignment of the generated captions with the alt-text. For nucleus sampling [12], we set the

Generated Captions	CLIP weight	ImageNet	Average over 38 datasets
1	0.0	28.60	32.50
2	0.0	29.00	33.40
4	0.0	29.53	33.70
8	0.0	29.60	34.03
8	0.3	30.10	34.40
8	0.5	30.35	35.15
8	0.7	30.25	34.35

Table 5. Effect of number of generated captions and weight of CLIPScore on zero-shot performance of pretrained CLIP models

cumulative probability of the smallest set of words to 0.9, and the minimum and maximum sequence lengths to 5 and 20, respectively. We study the effect of sampling 1, 2, 4, and 8 captions. For each input image-text pair, we assign the maximum alignment score between the alt-text and the generated captions. We observe in Table 5 that increasing the number of generated captions improves the performance on downstream tasks. We reason that due to the many-to-many relationship between images and captions, generating more captions increases the probability of matching a hard aligned alt-text. We also investigate the effect of fusing the Sieve alignment score with CLIPScore in Table 5. Each score is independently normalized, and a weighted average is applied between the two scores. Finally, the top 20% of samples ranked by Sieve+CLIPScore are selected. We observe that a weight of 0.5 achieves the best performance.

5. Conclusion

We introduce a novel method, Sieve, that enables pruning large-scale noisy web-crawled image-text datasets. We propose utilizing synthetic captions from image-captioning models pretrained on small, diverse, and curated datasets to evaluate the alignment of noisy image-text pairs. Using the embedding space of a lightweight sentence transformer, we compute an alignment score between generated captions and alt-text. We demonstrate that Sieve provides a complementary pruning signal to CLIPScore, effectively minimizing false positives and negatives, leading to improved zero-shot classification and retrieval performance.

References

- [1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication, 2023. [1](#), [4](#)
- [2] Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, and Roy Schwartz. WinoGAVil: Gamified association benchmark to challenge vision-and-language models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. [5](#)
- [3] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015. [4](#)
- [4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. [5](#)
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. [2](#)
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. [4](#)
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. [5](#)
- [8] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. [3](#)
- [9] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. [1](#), [2](#), [3](#), [5](#), [8](#), [12](#)
- [10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. [1](#)
- [11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. [1](#), [2](#)
- [12] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*. OpenReview.net, 2020. [2](#), [3](#), [8](#)
- [13] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23369–23379, 2023. [1](#), [5](#)
- [14] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. [2](#)
- [15] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016. [2](#)
- [16] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, 2017. [4](#)
- [17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. [5](#)
- [18] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022. [5](#)
- [19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. [6](#), [11](#)
- [20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. [5](#), [6](#), [7](#), [8](#)
- [21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. [3](#)
- [22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. [5](#)
- [23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. [4](#)
- [24] Pratyush Maini, Sachin Goyal, Zachary C Lipton, J Zico Kolter, and Aditi Raghunathan. T-mars: Improving visual representations by circumventing text feature learning. *arXiv preprint arXiv:2307.03132*, 2023. [3](#), [11](#)
- [25] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bisacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. [6](#), [11](#)

- [26] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Se-woong Oh, and Ludwig Schmidt. Improving multi-modal datasets with image captioning. *arXiv preprint arXiv:2307.10350*, 2023. [1](#), [3](#), [4](#)
- [27] OpenAI. The country211 dataset, . [6](#)
- [28] OpenAI. The rendered sst2 dataset, . [6](#)
- [29] 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. [5](#)
- [30] 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. [2](#), [3](#), [11](#)
- [31] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. [4](#)
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [1](#), [2](#), [4](#), [5](#)
- [33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. [1](#), [7](#)
- [34] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. [4](#)
- [35] William A Gavia Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. [11](#)
- [36] 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. *arXiv preprint arXiv:2111.02114*, 2021. [1](#), [2](#), [3](#), [6](#), [7](#), [8](#)
- [37] 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. [5](#)
- [38] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. [7](#)
- [39] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. [5](#), [6](#), [8](#)
- [40] 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. 2023. [1](#), [2](#)
- [41] Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. Mitigating spurious correlations in multi-modal models during fine-tuning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 39365–39379. PMLR, 2023. [1](#)
- [42] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. 2023. [1](#), [5](#)
- [43] 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. [5](#)
- [44] Haichao Yu, Yu Tian, Sateesh Kumar, Linjie Yang, and Heng Wang. The devil is in the details: A deep dive into the rabbit hole of data filtering. *arXiv preprint arXiv:2309.15954*, 2023. [3](#)
- [45] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruysen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. [5](#)