Supplementary Material of Mixture-of-Scores: Robust Image-Text Data Valuation via Three Lines of Code

Sitong Wu[1,†] Haoru Tan[2,†] Yukang Chen[3] Shaofeng Zhang[5] Jingyao Li[1]
Bei Yu[1‡] Xiaojuan Qi[2,‡] Jiaya Jia[4,‡]

[1]The Chinese University of Hong Kong [2]The University of Hong Kong [3]NVIDIA
[4]The Hong Kong University of Science and Technology [5]Shanghai Jiao Tong University

https://github.com/dvlab-research/MoS

This material includes the following parts:

- Sec. 1 demonstrates the necessity and significance of this work.
- Sec. 2 derives the proof for the standard deviation upper-bound mentioned in the main paper.
- Sec. 3 discusses the related works.
- Sec. 4 describes the details of three data processing strategies used in our experiments.
- Sec. 5 provides the detailed experimental settings, including the datasets and settings used during training and evaluation, and the baseline methods for comparison.
- Sec. 6 shows the robustness of our method on different data filtering ratios.
- Sec. 7 provides more experiments beyond the main paper, including results on BLIP and LLaVA model, and the comparison on larger-scale dataset LAION.
- Sec. 8 and 9 present the impact and limitation of this paper.

1. Necessity & Significance of This Work

We believe our work has significance in problem formulation, method design, and outcomes.

(1) Problem significance: Although it may be obvious that the alignment score for each image-text pair will be different when measured with different models (score disparity), the degree of score disparity and its serious impact on subsequent data processing and model learning have not been investigated. Thus, most existing works still use a single quality score to process image-text pair data, which may lose valuable data and retain noisy data due to the bias of single scoring model. To our best knowledge, our paper presents the first comprehensive study on score disparity and its impacts. We show that no single score excels across all evaluation tasks and different scores have complementary effects. We believe this will inspire further research to scale up high-quality image-text data collection especially considering our work has achieved notable improvements.

(2) Method design significance: Motivated by investigation, we present the method to heuristically calculate a more reliable score from the distribution of multiple scores based on density and deviation, which is the first to extract the essence and eliminate biases of multiple scores. Our method is scalable due to its high efficiency (refer to Sec.4.2 in the main paper). We shall release codes to contribute community to amplify its impacts.

(3) Outcomes significance: We show that our method not only enhances the accuracy of image-text pair quality evaluation (refer to Sec.4.1.1 in the main paper) but also significantly improves the performance on downstream tasks, for example, +4.3 and +3.2 on Flickr30K image-text retrieval.

2. Derivation for Upper Bound of the Standard Deviation

Considering a set of quality scores $\{S^1, ..., S^M\}$, each of which is normalized within [0, 1], their standard deviation can be formulated as $\sigma = \sqrt{\frac{1}{M}\sum_{k=1}^{M}(S^k - \overline{S})^2}$, where $\overline{S} = \frac{1}{M}\sum_{k=1}^{M} S^k$ denotes the mean value. Now, we target deriving the upper bound of σ.

Proof: Note that $a \leqslant S^k \leqslant b$, where $a = 0$ and $b = 1$, is the same as

$$-\frac{b-a}{2} \leqslant S^k - \frac{a+b}{2} \leqslant \frac{b-a}{2} \tag{1}$$

Also, for every random variable that satisfying $-\frac{b-a}{2} \leqslant Z \leqslant \frac{b-a}{2}$ or equivalently $|Z| \leqslant \frac{b-a}{2}$, we have

$$Z^2 \leqslant \left(\frac{b-a}{2}\right)^2 = \frac{(b-a)^2}{4} \tag{2}$$

[†]Equal contribution
[‡]Corresponding author

Thus, we can know that

$$\mathbb{E}(Z^2) \leqslant \frac{(b-a)^2}{4} \tag{3}$$

Therefore, since variance remains invariant under the addition of a constant, we can write that

$$\begin{aligned} \sigma^2 &= \mathrm{Var}\Big(S^k\Big) \\ &= \mathrm{Var}\Big(S^k - \frac{a+b}{2}\Big) \\ &\leqslant \mathbb{E}\Big[\Big(S^k - \frac{a+b}{2}\Big)^2\Big] \\ &\leqslant \frac{(b-a)^2}{4} \end{aligned} \tag{4}$$

Also, we observe that if we define $S^k = a$ with probability of 0.5 and $S^k = b$ also with probability of 0.5, then,

$$\begin{aligned} \mathrm{Var}\Big(S^k\Big) &= \frac{1}{2}\Big(a - \frac{a+b}{2}\Big)^2 + \frac{1}{2}\Big(b - \frac{a+b}{2}\Big)^2 \\ &= \Big(\frac{b-a}{2}\Big)^2 \\ &= \frac{(b-a)^2}{4} \end{aligned} \tag{5}$$

The above shows that the upper-bound $\sigma^2 \leqslant \frac{(b-a)^2}{4} = \frac{1}{4}$ can be achieved by a valid distribution. Therefore, we have $\sigma \leqslant \frac{1}{2}$, that is, the upper-bound of σ is 0.5.

3. Related Work

3.1. Vision-Language Pre-training

Vision-language pre-training (VLP) aims to derive common knowledge from large amounts of image-text data and develop a vision-language foundation model that generalizes well on various vision and linguistic tasks.

The pioneering works CLIP [1] and ALIGN [2] employed a dual-encoder architecture with a contrastive objective to learn aligned cross-modal representations from large-scale noisy image-text pairs crawled from the web, and achieved remarkable performance on zero-shot transferability for various downstream tasks. Subsequent works further enhanced the contrastive paradigm through various training objectives [3–7] and additional text decoder enabling the text generation ability [7, 8].

Recently, with the rise of large language models (LLM) [9–12], researchers began to incorporate the extensive knowledge of LLM with the vision-language model. These models are generally pre-trained on massive image-text pairs to learn an intermediate network that bridges the gap between the embedding space of the pre-trained vision encoder and LLM [13–15].

Our MoS can be applied on both classic models (such as CLIP [1] and BLIP [7]) and recent advanced models (such as LLaVA [15]).

3.2. Image-text Data Quality Metric

With the increasing prevalence of data collection from the Internet, the evaluation of data quality has become increasingly important.

The quality of image-text pair data is commonly assessed by examining the consistency between the visual content of the image and its text description. In the early stages, researchers relied on manual metrics for evaluation, including human judgement [16], fixed rules and heuristics [17, 18]. Since the advent of dual-encoder vision-language models (VLM) [1, 6, 7, 13] with aligned visual and linguistic embedding space, model-based metrics have gained significant popularity, where the quality of image-text data is measured by the similarity between the image embedding and text embedding extracted from the VLM.

Among these metrics, CLIP-Score [19] stands out as the most widely used. It has been applied in the data processing for various well-known datasets [20–22] and vision-language models [23, 24]. Furthermore, the follow-up works of CLIP [1], such as EVACLIP [6], BLIP [7], BLIP2 [13], etc., can also be employed to obtain this quality metric based on feature similarity, following the same principle as CLIP-Score.

In contrast, our paper is the first to reveal a phenomenon: existing model-based quality scores are significantly different from each other, which further leads to the serious disparity in the performance of model trained with dataset processed by distinct scores due to their inherent bias.

4. Image-Text Data Processing Strategies

Data filtering. Data filtering is an intuitive and popular way to avoid the detrimental impact of noise data, which concentrates on removing a set of low-quality data from the training set. In our experiments, given multiple quality scores from different scoring models, we select $\rho\%$ lowest quality data as the filtered data individually based on each quality score.

Sample weighting. Sample weighting aims to prevent the model from over-fitting noisy data during learning by adjusting the contribution of clean and noisy samples to the loss. Particularly, it assigns an individual weight to each data according to the quality of the data. Lower-quality data has a smaller loss weight. In our experiments, for each image-text pair, we use the normalized quality score (ranging from 0 to 1) as its loss weight.

Image re-captioning. Image re-captioning is a recently popular data processing strategy to cope with noisy image-text pair data. Image re-captioning is a recently popular data processing strategy to cope with noisy image-text pair data. It utilizes a powerful image captioning model to synthesize a new caption for the image, and then replace the original web text with this synthetic caption. Generally, compared

with the web text, the synthetic caption is more consistent with the visual content of image. In our experiments, given multiple quality scores from different scoring models, we select $\rho\%$ lowest quality data for re-captioning individually based on each quality score.

5. Experimental Settings

5.1. Training

Vision-language models. Our experiments are mainly conducted on the well-known and foundational CLIP [1] model with ViT-B/32 [25] as the image encoder. In addition, we also experiment on some more recent models, including BLIP [7] and LLaVA [15] model.

Datasets. We conduct experiments on two image-text pair datasets with different sizes (3M and 100M), respectively.

- **CC3M [17].** It was collected from 5 billion web pages, and public by Google in 2018. It contains 3,318,333 image-text pairs, where the image descriptions are obtained from the HTML alt-text attribute. Unfortunately, around 0.5M images are inaccessible due to the broken image links, so we finally collected around 2.8M image-text pairs for our experiments in vision-language pre-training.
- **LAION [22].** LAION-400M is an open-source dataset containing 400 million image-text pairs, scraped from the internet. We randomly selected 100M data from LAION-400M for experiments due to the limitations of our GPU resources. This 100M subset is referred to as LAION-100M in the following text.

Training settings. All the experiments are conducted on 16 NVIDIA V100 GPUs. The settings vary from different vision-language models. We follow most of the settings provided in their original paper.

- The CLIP model [1] is trained for 32 epochs with AdamW [26] optimizer, weight decay 0.2, and a batch size of 2048. After one warmup epoch, the learning rate gradually decreases from 1e-4 following the cosine strategy.
- For BLIP model [7], we train it for 20 epochs with a batch size of 1260 and AdamW optimizer. The weight decay is set to 0.05. The initial learning rate is 3e-4 and linearly decreases with a rate of 0.85.
- For LLaVA model [15], we follow the original settings. The model is pre-trained for 1 epoch with AdamW optimizer and a batch size of 256. The learning rate decreases from 1e-3 via the cosine strategy. The weight decay is set to zero.

5.2. Evaluation

We perform evaluation across a wide range of vision-language benchmarks under zero-shot setting, including classification, retrieval, generation, and grounding tasks.

For the zero-shot classification, we evaluate the model on four well-known benchmarks, including *e.g.,* ImageNet-1K [16], ImageNet-R [27], CIFAR100 [28] and VOC2007 [29]. We report the top-1 accuracy (denoted as "Acc@1").

For the image-text retrieval task, it contains two sub-tasks, *e.g.,* image-to-text (I2T) retrieval and text-to-image (T2I) retrieval. We report the Recall@1 (denoted as "R@1") on two widely-used benchmarks, *e.g.,* Flickr30K [30] and MSCOCO [31].

For the generation task, we consider three sub-tasks including image captioning, visual reasoning and visual question answering. For image captioning, we evaluate the models on two mainstream benchmarks (NoCaps [32] and COCO Caption [33]) and report the CIDEr [34] metric. For visual reasoning, models are evaluated on NLVR dataset [35] with accuracy metric. For visual question answering, we report the accuracy on the classical VQA dataset [36].

For the grounding task, we evaluate the models on the commonly-used RefCOCO+ dataset [37] and report the accuracy.

6. Robustness on Different Data Filtering Ratios

To study whether our MoS is robust enough to the data filtering ratio, we compare the performance of models trained on the filtered LAION-100M dataset under three kinds of data filtering ratios (10%, 20% and 30%). As shown in Table 5 and Table 2, our MoS still outperforms any single baseline quality score and the naive average ensemble strategy under all these filtering ratios (10%, 20% and 30%).

Our comparisons were restricted to filtering ratios of 30% or less, since models trained on LAION-100M with higher filtering ratios showed degraded performance compared to models trained on unfiltered data. This may be because the negative impact of reducing the amount of data has outweighed the positive impact of removing noisy data.

Quality Score		NoCaps	COCO Caption
Type	Scoring Model	CIDEr	CIDEr
CLIP Score	ViT-B/32	105.8	133.5
	ViT-B/16	104.2	134.8
	ViT-L/14	107.7	131.4
EVACLIP Score	ViT-L/14	107.0	132.6
	ViT-G/14	107.4	134.2
BLIP Score	ViT-B/16 pt	108.0	137.4
	ViT-B/16 ft	107.9	137.7
	ViT-L/16 pt	107.5	137.5
	ViT-L/16 ft	108.9	138.0
BLIP2 Score	ViT-G/14 pt	106.5	137.0
	ViT-G/14 ft	108.3	138.2
MoS (Ours)	All of above	110.8	139.4

Table 1. Comparison on the performance of LLaVA model trained on the CC3M dataset [17] via different quality scores. The model is pre-trained using sample weighting strategy.

Data Filtering Ratio	Quality Score Type	Quality Score Scoring Model	Flickr30K I2T R@1	Flickr30K T2I R@1	MSCOCO I2T R@1	MSCOCO T2I R@1	ImageNet-1K Acc@1	ImageNet-R Acc@1	CIFAR100 Acc@1	VOC2007 Acc@1
0%	-	-	70.2	53.7	35.5	31.3	61.1	62.0	60.3	65.9
20%	CLIP Score	ViT-B/32 ViT-L/14	75.9 75.6	55.4 55.9	39.5 40.1	34.0 33.9	64.0 63.9	65.7 65.2	62.0 62.2	68.2 68.0
	EVACLIP Score	ViT-L/14 ViT-G/14	75.7 76.6	55.6 56.4	39.8 39.3	34.6 34.0	63.7 64.1	65.4 64.6	61.8 61.4	68.6 69.3
	BLIP Score	ViT-B/16-pt ViT-L/16-ft	76.4 77.9	56.8 57.0	40.4 40.0	34.5 34.6	63.8 63.5	64.9 64.3	61.6 61.2	66.0 66.4
	BLIP2 Score	ViT-G/14-pt ViT-G/14-ft	78.5 78.1	57.5 57.1	39.8 40.6	35.3 35.2	63.3 63.0	65.1 64.3	61.8 60.5	67.5 66.6
	MoS (Ours)	All of above	79.8	58.4	41.5	35.9	65.0	66.2	63.0	70.8
30%	CLIP Score	ViT-B/32 ViT-L/14	76.4 75.9	56.0 56.5	39.7 40.0	34.3 34.0	63.8 63.5	66.0 65.3	62.4 62.5	68.7 68.4
	EVACLIP Score	ViT-L/14 ViT-G/14	76.5 76.8	56.2 57.0	40.2 39.6	34.4 34.5	64.0 64.6	65.6 64.5	61.8 61.3	68.5 69.5
	BLIP Score	ViT-B/16-pt ViT-L/16-ft	77.0 78.4	57.3 56.9	40.5 40.3	34.7 34.6	63.8 63.6	65.2 64.8	61.8 61.5	66.3 66.8
	BLIP2 Score	ViT-G/14-pt ViT-G/14-ft	79.2 78.8	57.9 56.9	39.9 40.4	35.5 35.4	63.5 63.1	65.1 64.6	62.0 60.5	67.9 67.0
	MoS (Ours)	All of above	81.0	59.2	42.0	36.4	65.3	66.8	63.3	71.4

Table 2. Comparison under different data filtering ratios (20% and 30%). All the experiments are conducted on CLIP (ViT-B/32) model, trained on filtered LAION-100M dataset using different quality scores. Note that the first line in this table denotes the performance without data filtering. For the comparison under filtering 10% data, please see Table 5.

Quality Score Type	Quality Score Scoring Model	Flickr30K I2T R@1	Flickr30K T2I R@1	MSCOCO I2T R@1	MSCOCO T2I R@1	ImageNet-1K Acc@1	ImageNet-R Acc@1	VQA Acc	NLVR Acc	COCO Caption CIDEr	RefCOCO+ Acc
-	-	69.4	57.0	40.3	29.5	58.2	62.6	69.3	74.6	118.0	69.6
CLIP Score	ViT-B/32 ViT-L/14	72.4 73.5	59.0 59.2	41.2 42.1	30.5 31.5	61.0 60.9	65.8 66.2	71.0 71.5	74.8 75.1	118.1 118.7	69.8 70.3
EVACLIP Score	ViT-L/14 ViT-G/14	73.0 74.5	60.2 60.1	41.7 42.0	31.5 31.5	60.5 61.0	65.8 66.1	71.3 70.3	75.8 75.3	118.0 119.2	70.5 70.9
BLIP Score	ViT-L/16-pt ViT-L/16-ft	75.2 75.3	60.3 60.5	42.3 43.0	31.0 32.5	60.5 60.8	65.0 65.8	70.0 71.1	75.0 75.9	118.7 119.9	72.0 71.7
BLIP2 Score	ViT-G/14-pt ViT-G/14-ft	75.5 75.0	60.2 59.8	43.6 43.2	32.5 32.2	60.5 60.0	65.4 65.0	69.4 69.9	74.6 75.2	119.4 119.5	72.7 73.4
MoS (Ours)	All of above	77.0	61.3	44.5	33.7	62.0	67.5	73.8	77.0	123.5	75.2

Table 3. Comparison on the performance of BLIP-Base model trained on CC3M dataset [17] processed by data filtering strategy. We filter out 10% lowest-quality data based on each quality score. The first line in this table denotes the performance of model trained on full CC3M dataset without any data filtering.

7. More Experiments

In this section, we provide some important experiments beyond the main paper.

7.1. Results on LLaVA Model

We also conduct experiments on a more modern vision-language model, LLaVA [15], and observe similar quality score disparity and corresponding model performance disparity phenomena. We pretrain LLaVA model under sample weighting strategy where the weight of each data is determined by different quality scores. As shown in Table 1, our MoS is still better than each of single baseline quality scores.

7.2. Results on BLIP Model

We report the results on BLIP model Table 3. For zero-shot classification and retrieval tasks, our MoS shows consistent advantages when training BLIP model with around 2 points.

For VQA task, our MoS has an advantage of +3.4 compared with the average ensemble strategy. When compared with single baseline quality scores, our MoS is +3.2 better than the average performance of all the baseline scores and +2.3 better than the best one among all the baseline scores on VQA dataset. For the visual reasoning task, our MoS outperforms the best baseline quality score by +1.1 on NLVR dataset, and the average performance of all baseline quality

Data Processing	Quality Score Type	Scoring Model	Flickr30K I2T R@1	Flickr30K T2I R@1	MSCOCO I2T R@1	MSCOCO T2I R@1	ImageNet-1K Acc@1	ImageNet-R Acc@1	CIFAR100 Acc@1	VOC2007 Acc@1
Data Filtering	CLIP Score	R50	20.5	15.3	9.9	7.5	14.9	17.5	29.0	38.2
		R101	20.1	14.6	9.6	7.6	14.6	16.5	28.6	38.1
		ViT-B/32	22.7	15.3	10.8	7.8	15.0	17.7	28.8	38.5
		ViT-B/16	22.6	14.8	10.7	8.1	14.6	16.8	28.9	38.4
		ViT-L/14	22.9	14.6	10.6	7.8	14.0	16.2	28.5	38.3
		ViT-L/14-336px	22.0	15.5	10.4	7.6	14.5	17.0	28.3	38.1
	EVACLIP Score	ViT-B16	20.7	14.6	10.4	7.7	14.8	17.3	28.0	38.0
		ViT-L/14	22.2	16.2	11.2	8.2	14.6	17.0	28.9	38.3
		ViT-L/14-336px	22.5	15.3	10.4	8.0	13.8	15.3	26.1	38.1
		ViT-G/14	22.4	15.2	10.0	7.7	14.7	17.1	28.7	38.5
		ViT-G/14-plus	21.9	15.0	11.1	7.6	13.8	15.4	26.2	38.0
	BLIP Score	ViT-B/16-pt	23.5	16.7	11.5	8.2	14.3	16.5	29.5	39.0
		ViT-L/16-pt	23.1	16.1	11.2	8.3	13.3	15.0	28.9	38.7
		ViT-B/16-ft	23.0	16.3	11.7	8.6	14.1	16.0	29.9	39.2
		ViT-L/16-ft	22.6	15.9	11.4	8.5	13.5	14.8	28.5	38.4
	BLIP2 Score	ViT-L/14-pt	22.8	16.2	11.2	8.1	14.0	16.2	29.3	39.0
		ViT-G/14-pt	23.4	16.6	11.3	8.4	14.1	16.5	29.6	38.9
		ViT-G/14-ft	22.9	15.3	11.6	8.8	14.0	16.0	28.6	38.7
Sample Weighting	CLIP Score	R50	21.8	15.2	10.1	7.7	14.6	17.0	28.8	38.8
		R101	21.5	14.6	10.2	7.6	14.0	15.8	28.5	38.3
		ViT-B/32	22.6	15.2	10.8	7.9	14.5	16.6	29.0	38.7
		ViT-B/16	21.9	14.5	10.0	7.4	13.9	16.0	28.4	38.0
		ViT-L/14	22.4	14.9	10.4	7.4	14.4	16.7	28.5	38.4
		ViT-L/14-336px	21.1	14.1	10.2	7.3	14.5	16.3	28.0	37.7
	EVACLIP Score	ViT-B16	22.7	15.3	10.4	7.5	13.8	15.5	28.3	37.9
		ViT-L/14	21.7	14.9	10.8	7.8	14.2	16.0	28.6	38.2
		ViT-L/14-336px	22.3	15.6	10.6	7.3	14.4	16.4	28.2	37.8
		ViT-G/14	23.3	15.9	10.9	8.0	14.7	16.8	28.4	38.6
		ViT-G/14-plus	21.5	14.4	10.3	7.2	14.2	15.9	28.1	37.9
	BLIP Score	ViT-B/16 pt	23.9	16.1	11.0	8.4	13.2	16.0	29.8	39.0
		ViT-L/16 pt	22.2	15.6	10.7	8.5	13.7	15.7	29.5	39.1
		ViT-B/16 ft	23.4	16.0	11.3	8.9	13.0	15.1	29.7	39.4
		ViT-L/16 ft	23.0	15.9	11.1	8.6	13.3	15.3	29.4	39.2
	BLIP2 Score	ViT-L/14 pt	22.5	15.8	10.9	8.2	13.6	15.0	28.9	38.6
		ViT-G/14 pt	23.0	16.0	10.7	8.1	13.8	15.6	29.3	39.0
		ViT-G/14 ft	21.9	15.1	11.0	8.7	13.2	15.4	28.5	39.2
Image Re-captioning	CLIP Score	R50	39.5	26.5	21.2	14.8	16.2	21.0	23.0	35.8
		R101	38.5	26.0	20.8	13.5	15.9	20.5	22.1	35.2
		ViT-B/32	39.2	26.7	20.7	14.5	16.2	20.8	22.4	35.7
		ViT-B/16	38.1	26.0	20.2	13.8	15.5	20.2	22.2	35.3
		ViT-L/14	37.6	26.2	21.0	13.4	15.2	20.8	22.0	35.5
		ViT-L/14-336px	38.4	26.3	20.5	13.2	15.0	20.0	21.8	34.9
	EVACLIP Score	ViT-B16	37.3	24.8	20.6	14.0	15.0	20.2	21.6	34.6
		ViT-L/14	38.5	26.3	20.9	14.0	16.4	21.2	22.5	36.0
		ViT-L/14-336px	38.1	26.1	20.7	13.9	15.9	20.4	22.0	35.0
		ViT-G/14	39.4	26.5	21.0	13.2	15.5	20.5	22.3	35.5
		ViT-G/14-plus	37.6	25.4	20.1	13.3	16.5	20.9	22.8	35.9
	BLIP Score	ViT-B/16 pt	40.0	27.6	21.6	14.9	15.2	20.0	24.0	36.4
		ViT-L/16 pt	39.7	27.1	21.7	14.5	15.3	20.2	23.6	36.2
		ViT-B/16 ft	40.4	27.2	22.0	14.8	14.9	19.1	23.9	36.0
		ViT-L/16 ft	39.9	26.8	21.4	14.0	15.1	20.0	23.5	36.5
	BLIP2 Score	ViT-L/14 pt	39.1	26.0	21.0	14.1	15.1	19.5	23.0	35.8
		ViT-G/14 pt	40.0	26.9	21.5	14.5	15.0	20.2	23.5	36.0
		ViT-G/14 ft	39.8	26.5	21.8	14.6	15.3	19.8	23.2	36.2

Table 4. Performance of each baseline quality score under three data processing strategies. We train the CLIP ViT-B/32 model on CC3M dataset [17] processed using different baseline quality scores.

scores by +1.8. For the image captioning task, our MoS has an advantage of +4.6 on COCO Caption compared to the average performance of all baseline quality scores, and is +3.6 better than the best baseline quality score. For visual grounding task, our MoS is +1.8 better on RefCOCO+ dataset than the best baseline quality score, and +3.8 better than the average performance of all baseline scores. Compared with the naive average ensembling all baseline scores, our MoS shows an advantage of +3.4.

7.3. Results on CLIP Model

Train on CC3M dataset. Due to the space limitations of the main paper, we supplement the performance of each baseline score in Table 1 (main paper) for CLIP ViT-B/32 model in Table 4.

Train on LAION Dataset. Similar to the experimental setting of CC3M, we first calculate the quality scores for each data in LAION-100M using eight commonly-used off-the-shelf scoring models, and finally obtain eight baseline quality scores. These eight scoring models include CLIP (ViT-B/32 and ViT-L/14), EVACLIP (ViT-L/14 and ViT-G/14), BLIP (ViT-L/16-pt and ViT-L/16-ft) and BLIP2 (ViT-G/14-pt and ViT-G/14-ft). Then, we calculate our MoS metric based on these eight baseline quality scores.

In particular, the distribution of standard deviations between these eight baseline quality scores shows a similar pattern to Figure 1(a) of the main paper, but with a higher mean value of 0.23 (*vs.* 0.14) and a higher maximum value of 0.39 (*vs.* 0.26). It means that the quality score disparity

Quality Score		Flickr30K		MSCOCO		ImageNet-1K	ImageNet-R	CIFAR100	VOC2007
Type	Scoring Model	I2T R@1	T2I R@1	I2T R@1	T2I R@1	Acc@1	Acc@1	Acc@1	Acc@1
-	-	70.2	53.7	35.5	31.3	61.1	62.0	60.3	65.9
CLIP Score	ViT-B/32	75.5	55.3	38.2	33.9	63.5	64.4	61.9	68.1
	ViT-L/14	75.3	55.6	38.1	33.6	63.3	65.0	61.7	68.5
EVACLIP Score	ViT-L/14	75.9	55.5	38.6	34.0	63.0	65.0	61.5	68.4
	ViT-G/14	76.2	56.0	38.0	33.4	63.4	64.3	61.8	68.9
BLIP Score	ViT-B/16-pt	76.7	56.3	38.5	34.0	62.9	64.5	61.0	65.6
	ViT-L/16-ft	77.5	57.1	38.9	34.2	62.8	64.0	61.5	66.3
BLIP2 Score	ViT-G/14-pt	78.0	56.9	39.1	34.8	62.8	64.7	61.7	67.0
	ViT-G/14-ft	77.8	57.0	39.5	34.7	62.4	64.1	60.6	66.4
MoS (Ours)	All of above	79.0	57.7	40.6	35.3	64.1	65.5	62.4	70.2

Table 5. Comparison on the performance of CLIP (ViT-B/32) model trained on the filtered LAION dataset [22] using different quality scores. We filter out 10% lowest-quality data based on each quality score. Note: the first line in this table denotes the performance without data filtering.

phenomenon becomes more obvious on the LAION dataset than CC3M dataset. This may be because the LAION dataset is more noisy, and we found that medium-quality data are more likely to have larger quality score disparity (discussed in Sec. 2.3 of the main paper).

We also compare the performance of model (CLIP ViT-B/32) trained with the filtered datasets using different quality scores. As shown in Table 5, it can be found that the performance is also sensitive to the scoring model. Our MoS still obviously outperforms these baseline quality scores and the naive average ensemble strategy on all evaluation tasks.

8. Broader Impact

This paper proposes a simple but powerful method to obtain good data quality metrics that can generalize well. There are some potential positive societal effects, such as helping people better understand the role of data to develop more robust deep learning systems and possibly even be used to eliminate data bias. As for the potential negative impacts, we believe that this technology, and even the entire field of artificial intelligence, may be applied to inhumane social surveillance, which should be taken seriously by legislative bodies worldwide.

9. Limitations

The proposed approach in this paper has only been studied in the field of vision-language so far, without being validated in broader areas of multi-modal learning, such as speech-vision or speech-text. In the future, we plan to further update the research content of this paper and validate it in a wider range of multi-modal learning domains.

References

[1] 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. 2, 3

[2] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International conference on machine learning, pages 4904–4916. PMLR, 2021. 2

[3] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208, 2021. 2

[4] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10995–11005, 2023.

[5] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[6] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389, 2023. 2

[7] 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. 2, 3

[8] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Ja-

son J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. arXiv preprint arXiv:1909.11059, 2019. 2

[9] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023. 2

[10] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[11] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068, 2022.

[12] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. Journal of Machine Learning Research, 25(70):1–53, 2024. 2

[13] 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. 2

[14] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716–23736, 2022.

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024. 2, 3, 4

[16] 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. Ieee, 2009. 2, 3

[17] 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, 2018. 2, 3, 4, 5

[18] 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. 2

[19] 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. 2

[20] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems, 35:25278–25294, 2022. 2

[21] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

[22] 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. 2, 3, 6

[23] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. arXiv preprint arXiv:2211.01335, 2022. 2

[24] Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. Urbanvlp: A multi-granularity vision-language pre-trained foundation model for urban indicator prediction. arXiv preprint arXiv:2403.16831, 2024. 2

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 3

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017. 3

[27] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8340–8349, 2021. 3

[28] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[29] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 3

[30] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision, pages 2641–2649, 2015. 3

[31] 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: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014. 3

[32] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Ste-

fan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In Proceedings of the IEEE/CVF international conference on computer vision, pages 8948–8957, 2019. 3

[33] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325, 2015. 3

[34] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4566–4575, 2015. 3

[35] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. arXiv preprint arXiv:1811.00491, 2018. 3

[36] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433, 2015. 3

[37] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pages 69–85. Springer, 2016. 3