Multimodal Data Curation via Object Detection and Filter Ensembles

Tzu−Heng Huang*, Changho Shin*, Sui Jiet Tay, Dyah Adila, Frederic Sala
Department of Computer Sciences
University of Wisconsin-Madison
{cshin23, thuang273}@wisc.edu, jiet9000@gmail.com, adila@wisc.edu, fredsala@cs.wisc.edu

Abstract

We propose an approach for curating multimodal data that we used for our entry in the 2023 DataComp competition filtering track. Our technique combines object detection and weak supervision-based ensembling. In the first of two steps in our approach, we employ an out-of-the-box zero-shot object detection model to extract granular information and produce a variety of filter designs. In the second step, we employ weak supervision to ensemble filtering rules. This approach results in a 4% performance improvement when compared to the best-performing baseline, producing the top-ranking position in the small scale track at the time of writing. Furthermore, in the medium scale track, we achieve a noteworthy 4.2% improvement over the baseline by simply ensembling existing baselines with weak supervision.

1 Introduction

Multimodal models, such as CLIP [16], DALL-E [17], Stable Diffusion [21], Flamingo [1], and FLAVA [26] have shown unprecedented performance in many vision-language tasks. Massive datasets collected from the web play a crucial role in these successes. As a result, there is renewed interest in *data-centric* approaches [28, 10] to machine learning, focusing on data rather than models, architectures, training approaches, etc. An important part of this data-centric approach involves community efforts to curate enormous open-source vision-language datasets via large crawls of the web [23, 22].

While offering impressive scale, raw web-crawled data can be noisy and lack appropriate selection. Data curation is therefore crucial. However, there are many questions on what might be the right approach for curation in the context of training large-scale models. In order to shed light on these, the *DataComp* competition invites users to propose a variety of data curation approaches while fixing model architectures, training procedures, and a raw data pool [7].

In this work, we document our data curation framework and report performance results for the DataComp filtering track at small and medium scale. Our approach is predominantly based on object detection and filter ensembles. In the small scale case, we include various additional rules generated from higher-order granular information via object detection, which yields 4.0% improvement over the best-performing existing baseline (CLIP score (L/14 30%)). Additionally, in the medium scale case, we ensemble baseline filters, which provides 4.2% improvement over the top-performing baseline (Image-based ∩ CLIP score (L/14 30%)).

2 Data Curation Framework

Our curation approach mainly focuses on filtering and ensembling. We provide an overall workflow of the proposed framework in Figure 1 and discuss each component in-depth. Broadly, our framework involves two steps. First, we design individual filters by considering multi-

*Equally contributed to this work

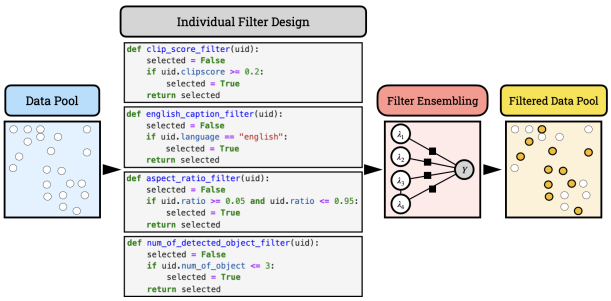

Figure 1: Overall workflow of the data curation framework. Each data point in the raw data pool is passed to individual filters, which can be designed by human heuristics, pre-computed CLIP scores, or inference results from other off-the-shelf models. We employ Grounding DINO, a zero-shot object detection model, to identify objects mentioned in the image caption. After each designed filter processes, we ensemble filtering results and curate the final refined data pool.

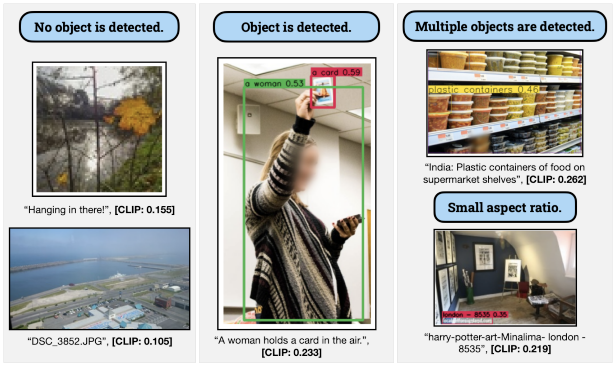

Figure 2: We showcase various image samples with their caption and CLIP score in the small scale dataset and annotate recognized objects through Grounding DINO. Nearly 38% of the images do not have any identified objects, while 18% of them have multiple detected objects. Additionally, around 3% of the images have tiny detected objects. These results offer rich information, which can be used and combined with heuristics to design additional filtering rules.

ple filtering sources such as existing Datacomp baselines with human heuristics, provided CLIP scores, and an additional component, object detection filters. Next, we tune the thresholds in filters to establish refining rules by evaluating the performance with downstream tasks in the Datacomp benchmark. At last, we select established filters and apply a weak supervision algorithm [18, 25] to ensemble filters and aggregate each filtered result to curate the final dataset.

2.1 Filtering Method Design

Our created filters mainly rely on two data pruning approaches and their intersections. The first approach leverages provided CLIP scores (L/14) to select "top x%" of the images that have high similarity score, while the second approach uses the inference results from a zero-shot object detection model to design rules for refining images.

In this work, we employ Grounding DINO [12][*] to identify objects mentioned in the image caption and anchor their locations. There are three types of inference outcomes provided by Grounding Dino: bounding boxes of detected objects, predicted logit scores for each object (scaled from 0 to 1), and the phrases for each detected object.

Such zero-shot object detection models have several advantages. First, the inference results offer a certain level of certification that the mentioned objects exist in the image. In addition to the pre-computed CLIP scores, identification results are more granular, providing information that can be included in filter designs based on interpretable heuristics. Furthermore, zero-shot object detection models possess the capability to detect objects that are unseen in the predefined set of classes used for training. This feature makes our filtering framework ideal for recognizing new objects and addressing diverse scenarios in image-text datasets obtained from web crawling on a large scale. Finally, Grounding DINO is simple and out-of-the-box, which does not require human efforts to customize prompts for candidate objects to query whether exist or not. Grounding DINO automatically detects phrases from the given caption and links corresponding visual elements to locate, allowing our filtering method to be easily

[*]https://github.com/IDEA-Research/GroundingDINO

applied to a massive amount of images.

Several image examples are categorized and displayed in Figure 2. With the inference results from Grounding DINO, we convert these into *filtering conditions* to refine images in accordance with three types of human heuristics. The conditions are the following:

1. Predicted logit scores: the logit score acts as a measure of how certain Grounding DINO is in identifying specific objects. To eliminate images with low logit scores, we take the average and the maximum scores of detected objects within an image and establish a threshold to refine the data pool,

2. Number of detected objects: Grounding DINO is capable of detecting multiple objects. Since the image captions are brief and short, the objects mentioned in them are expected to fall within a certain range. Therefore, we determine the total number of objects detected in an image and discard any images that have an out-of-range number of objects,

3. Aspect ratio of detected objects: In order to eliminate object localization that lacks significance, we compute the aspect ratio of every object that is detected and calculate the average ratio present in the image. Afterward, we use a threshold to eliminate images that are either too small or too large in the frame.

Obtaining these filters is cost-effective. They serve the basic purpose of checking the presence of particular objects mentioned in the caption within an image. Additionally, they are designed to be easily integrated with other contributed filters for greater adaptability and enhancement. To illustrate this notion, by combining object detection filters with CLIP score filters and analyzing their intersection, we can discard images devoid of any objects but with a high CLIP score.

2.2 Ensembling Filtering Methods

When multiple results from different filters are available, the most effective approach is to combine them using an ensembling method. Our ensembling strategy borrows from the rich literature on weak supervision [2, 19, 4, 25].

Let the raw dataset be $\mathcal{D} = \{x_i\}_{i=1}^n$, where x_i is an image-text pair, and denote $y \in \{0,1\}^n$ be the inclusion labels processed by a given filter — keep data point x_i if $y_i = 1$ — such that $D_y = \{x_i | y_i = 1, i = 1, \ldots, n\}$ be the refined dataset. Define the downstream task loss $\mathcal{R}(x, y; \mathcal{A})$, given a training model and algorithm $\mathcal{A}$, and let $y^* = \arg\min \mathcal{R}(x, y; \mathcal{A})$. While y^* can be obtained by brute force search over $O(2^n)$ combinations, this is not practically possible even for the small scale case in DataComp, where $n = 12.8$ millions. Instead, we use a variety of filtering methods λ^j $(j = 1, \ldots, m)$ such that $\lambda^j(x) \in \{0, 1\}$. These designed filters are assumed to have some level of accuracy with respect to y^*. Hence, given m filters, the most basic ensemble approach is majority voting, i.e.

$$\hat{y}_i = \frac{1}{m} \sum_{j=1}^{m} \lambda^j(x_i)$$

However, majority voting fails to consider the accuracy and correlation of filtering methods. A standard approach in weak supervision is to encode filters' accuracy and correlations with the Ising model [2, 19, 6], i.e.,

$$p(\lambda^1, \ldots, \lambda^m, x_i, y_i) = \frac{1}{Z} \exp\Big(\sum_{j=1}^{m} \theta_j \lambda^j(x_i) y_i + \sum_{(j,k) \in E} \theta_{j,k} \lambda^j(x_i) \lambda^k(x_i) + \theta_Y y_i \Big),$$

where $\theta_j, \theta_{j,k}, \theta_Y$ are the canonical parameters encoding accuracy, correlation, and class balance respectively. E is the set of candidate filters correlations, and Z is a normalization constant to ensure the probability is valid. After learning the parameters in the Ising model, we can infer the most probable inclusion labels by computing

$$\hat{y}_i = \arg \max_{y' \in \{0,1\}} p(y_i = y' | \lambda^1(x_i), \ldots, \lambda^m(x_i))$$

Finally, $\hat{y}_i$ is the aggregated decision to include x_i or not, and $D_{\hat{y}} = \{x_i | \hat{y}_i = 1, i = 1, \ldots, n\}$ becomes the final curated dataset to be used to train the model.

3 Experiments

Implementation Details We document our implementation and report each step in detail. We used the

img2dataset package* to download small and medium scale datasets in the filtering track, succeeding in downloading 11.9M (93.32%) samples and 115.0M (89.89%) samples. In the small scale case, we double-checked that the amount of downloaded data is comparable to the DataComp team's data by reproducing "Baseline: CLIP score (L/14 30%)", which gives a slightly lower but comparable performance (ImageNet accuracy 0.045/0.051 and Average performance 0.168/0.173).

In Grounding DINO, the inference rate is not consistent and can be slow when processing with the original image size. To improve computational efficiency, we resized the images. However, this trade-off between performance and inference rate must be taken into account. After attempting to resize with various dimensions including (224, 224), (400, 400), and (800, 800), we ultimately settled on using (400, 400) as it provided similar performance. Among variants of Grounding DINO, we choose Grounding-DINO-T, which used Swin-T [13] as an image backbone and BERT [11] as a text backbone.

In the ensembling step, we employed Snorkel [2]*, a well-known framework for weak supervision. We set the class balance parameter to 0.3 and 0.2 for small and medium scale datasets respectively, based on prior works [7] that investigated the relationship between the size of the filtered dataset and downstream task performance. To ensemble multiple filter results, we trained the Snorkel label model for 1000 epochs with learning rate 0.01. All the experiments were performed on an NVIDIA A6000 GPU.

Filtering Conditions Next, we discuss considered conditions when designing the three types of object detection filters — logit score, detected number, and aspect ratio. First, if there is no object identified by Grounding DINO, the output is empty. There are about 38% of images in this setting, and we eliminated them from the raw data pool.

For the logit score filter, we selected image examples with the highest average score and highest maximum score based on the top 30%. Another filter was designed to detect numbers within a specific range of 1 to 4 and 1 to 3. The aspect ratio filter discarded images with an average aspect ratio smaller than 5% or larger than 95%. Finally, to ensure well-aligned image-text examples, we intersected each of the above filters with various CLIP score filters, including CLIP L/14 30%, 50%, and 55%. Our design of these thresholds took into account the size of the resulting dataset.

*https://github.com/rom1504/img2dataset

*https://github.com/snorkel-team/snorkel.

4 Results

Small Scale Dataset In order to evaluate the effectiveness of the proposed object detection filters and tune threshold parameters to optimize results, we examined the performance of each filter on the small scale dataset. We anticipate that Grounding DINO's outputs will provide additional granular information that can be used to create better filtering rules. Table 1 shows the performance of the curated training dataset with various designed filter combinations. As expected, when combined with the CLIP score filter, the object detection filters outperform most of DataComp's existing baselines. Additionally, by setting the threshold correctly, we find that the aspect ratio filter produces the best average performance of 0.178 among all methods, while another designed filter that counts the number of detected objects achieves the suboptimal performance at 0.174. With the optimal threshold and its resulting filters, we were able to establish filtering conditions and used them in the ensembling step.

Subsequently, to evaluate the performance of our ensemble method, we began by aggregating the filtered datasets created by DataComp's baselines. The performance of the final curated datasets, obtained through different filter combinations, is presented in Table 2. We observe that both majority voting and weak supervision techniques yield results that are comparable or even superior to the best individual rule when the dataset is curated solely by the baseline ensemble. Additionally, weak supervision demonstrates superior performance compared to majority voting by 4.1%. These demonstrate the advantages of our ensemble technique and validate the curation framework effectively.

Afterward, we apply this setup and incorporate object detection filters and weak supervision techniques with the class balance parameter 0.3 to ensemble filters for performance enhancement. Finally, by combining various sources of filters, the combination of baselines and logit score filters achieves the best 4.0% improvement over the existing best-performing baseline (CLIP L/14 30%) on

Table 1: Performance comparison of individual filters in small scale dataset.

Method	Dataset size	ImageNet acc.	Average perf.
CLIP L/14 30% (DataComp)	3.84M	.051	.173
OD Avg. Logit 30% ∩ CLIP L/14 50%	3.84M	.054	.164
OD Avg. Logit 30% ∩ CLIP L/14 30%	2.38M	**.059**	**.172**
OD Max. Logit 30% ∩ CLIP L/14 30%	2.39M	.054	.172
OD Max. Logit 30% ∩ CLIP L/14 50%	3.84M	**.059**	**.173**
OD Num. of Objects (≤ 4) ∩ CLIP L/14 50%	4.43M	.052	.170
OD Num. of Objects (≤ 3) ∩ CLIP L/14 55%	4.61M	.050	.173
OD Num. of Objects (≤ 3) ∩ CLIP L/14 50%	4.36M	**.052**	**.174**
OD Avg. Aspect Ratio ∩ CLIP L/14 55%	4.19M	.053	.173
OD Avg. Aspect Ratio ∩ CLIP L/14 50%	3.98M	**.053**	**.178**

Table 2: Performance comparison of ensemble filters in small scale dataset.

Method	Dataset size	ImageNet acc.	Average perf.
MV (baselines)	2.39M	.060	.168
WS (baselines, class balance 0.5)	6.38M	.043	.153
WS (baselines, class balance 0.2)	2.49M	.058	.174
WS (baselines, class balance 0.3)	3.20M	**.059**	**.175**
WS (baselines + All the OD Filters)	4.10M	.055	.169
WS (baselines + OD Max. Logit + OD Avg. Aspect Ratio)	3.92M	.059	.172
WS (baselines + OD Num of Objects + OD Avg. Aspect Ratio)	4.14M	.056	.173
WS (baselines + OD Avg. Logit + OD Max. Logit)	4.11M	**.056**	**.180**

average.

Medium Scale Dataset Our team predominantly focused on the small scale track. However, the ensemble approach with the provided baseline filters can be used to curate the dataset in the medium case without significant additional costs. By setting the class balance at 0.2 and using weak supervision techniques, the final refined dataset achieves ImageNet accuracy of 0.305 and average performance of 0.342, exhibiting 4.2% improvement compared to the top-performing baseline filter. Our ensemble approach has shown its adaptability by successfully incorporating other contributed filters effectively.

5 Conclusion

Large-scale web-crawled data has been widely collected for training multimodal models, producing high-performance models. However, web-crawled data often contains noise, low-quality samples, and suffers from poor selection. To address this issue, we proposed a framework for refining data pool and improving curated datasets in the 2023 DataComp competition. Our approach involves designing various filters using off-the-shelf zero-shot object detection models and applying weak supervision-based ensembling techniques. Through empirical validation, we showed the effectiveness of the designed filters and ensemble methods in our data curation framework on both small and medium scale datasets. Our approach has resulted in a 4% performance improve-

ment compared to the best existing baselines, producing the top position in the leaderboard for the small scale track at the time of writing.

References

[1] 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. 1

[2] Stephen H Bach, Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cassandra Xia, Souvik Sen, Alex Ratner, Braden Hancock, Houman Alborzi, et al. Snorkel drybell: A case study in deploying weak supervision at industrial scale. In *Proceedings of the 2019 International Conference on Management of Data*, pages 362–375, 2019. 3, 4, 8

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 7

[4] Mayee F Chen, Daniel Y Fu, Dyah Adila, Michael Zhang, Frederic Sala, Kayvon Fatahalian, and Christopher Ré. Shoring up the foundations: Fusing model embeddings and weak supervision. In *Uncertainty in Artificial Intelligence*, pages 357–367. PMLR, 2022. 3, 9

[5] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021. 10

[6] Daniel Y. Fu, Mayee F. Chen, Frederic Sala, Sarah M. Hooper, Kayvon Fatahalian, and Christopher Ré. Fast and three-rious: Speeding up weak supervision with triplet methods. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, 2020. 3, 8

[7] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 1, 4

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 7

[9] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 7

[10] Mohammad Hossein Jarrahi, Ali Memariani, and Shion Guha. The principles of data-centric ai (dcai). *arXiv preprint arXiv:2211.14611*, 2022. 1

[11] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 4

[12] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 7

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[14] 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. 10

[15] M Minderer, A Gritsenko, A Stone, M Neumann, D Weissenborn, A Dosovitskiy, A Mahendran, A Arnab, M Dehghani, Z Shen, et al. Simple open-vocabulary object detection with vision transformers. arxiv 2022. *arXiv preprint arXiv:2205.06230*. 7

[16] 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, 7

[17] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[18] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the 44th International Conference on Very Large Data Bases (VLDB)*, Rio de Janeiro, Brazil, 2018. 2

[19] A. J. Ratner, B. Hancock, J. Dunnmon, F. Sala, S. Pandey, and C. Ré. Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, 2019. 3, 8

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object

detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 7

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[22] 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. 1

[23] 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

[24] Changho Shin, Sonia Cromp, Dyah Adila, and Frederic Sala. Mitigating source bias for fairer weak supervision. *arXiv preprint arXiv:2303.17713*, 2023. 9

[25] Changho Shin, Winfred Li, Harit Vishwakarma, Nicholas Carl Roberts, and Frederic Sala. Universalizing weak supervision. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3

[26] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 1

[27] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 7

[28] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023. 1

[29] Jieyu Zhang, Linxin Song, and Alex Ratner. Leveraging instance features for label aggregation in programmatic weak supervision. In *International Conference on Artificial Intelligence and Statistics*, pages 157–171. PMLR, 2023. 9

The appendix is organized as follows. In the Appendix A, we provide an analysis of correlation across all the individual filters we designed and their estimated accuracies. Next, we discuss related works in our framework in the Appendix B. Last but not least, we provide more insights about the properties of conservative loss as part of a future work discussion in Appendix C.

A Correlation and Estimated Accuracy Across Designed Filters

The correlation between baselines and designed filters listed in Table 1 is presented in Figure 3. As expected, the correlation is low for baseline filters and moderate to high for CLIP and Grounding DINO filters. This observation could explain the results in Table 2, where the baseline ensemble performs well, while including Grounding DINO does not produce the best results due to the high correlation. To resolve this issue, the dependency graph can be taken into account, which was avoided for simplicity.

Figure 4 shows the estimated accuracy of each filter in the Ising model (i.e. $P(y = \lambda^j(x))$). Though the estimated accuracy can be affected by the violation of conditional independence, we notice that the estimated accuracy is correlated to the performance that we display in Table 1.

B Related Works

Zero-Shot Object Detection Zero-shot object detection differs from traditional object detection methods [3, 20, 8]. Unlike the latter, it is not limited to detecting only pre-defined object classes in the training data. This technique does not require fine-tuning of model parameters to introduce novel object classes. Instead, it uses multi-modal representations and language generalization to perform detection for such objects. Most of the current zero-shot object detection models [15, 12, 27, 9] use CLIP [16] as their query module to align textual embeddings and visual components. In our work, we use Grounding DINO [12] as a detector to check that the mentioned objects exist in the image. We then use the returned information to design additional filters.

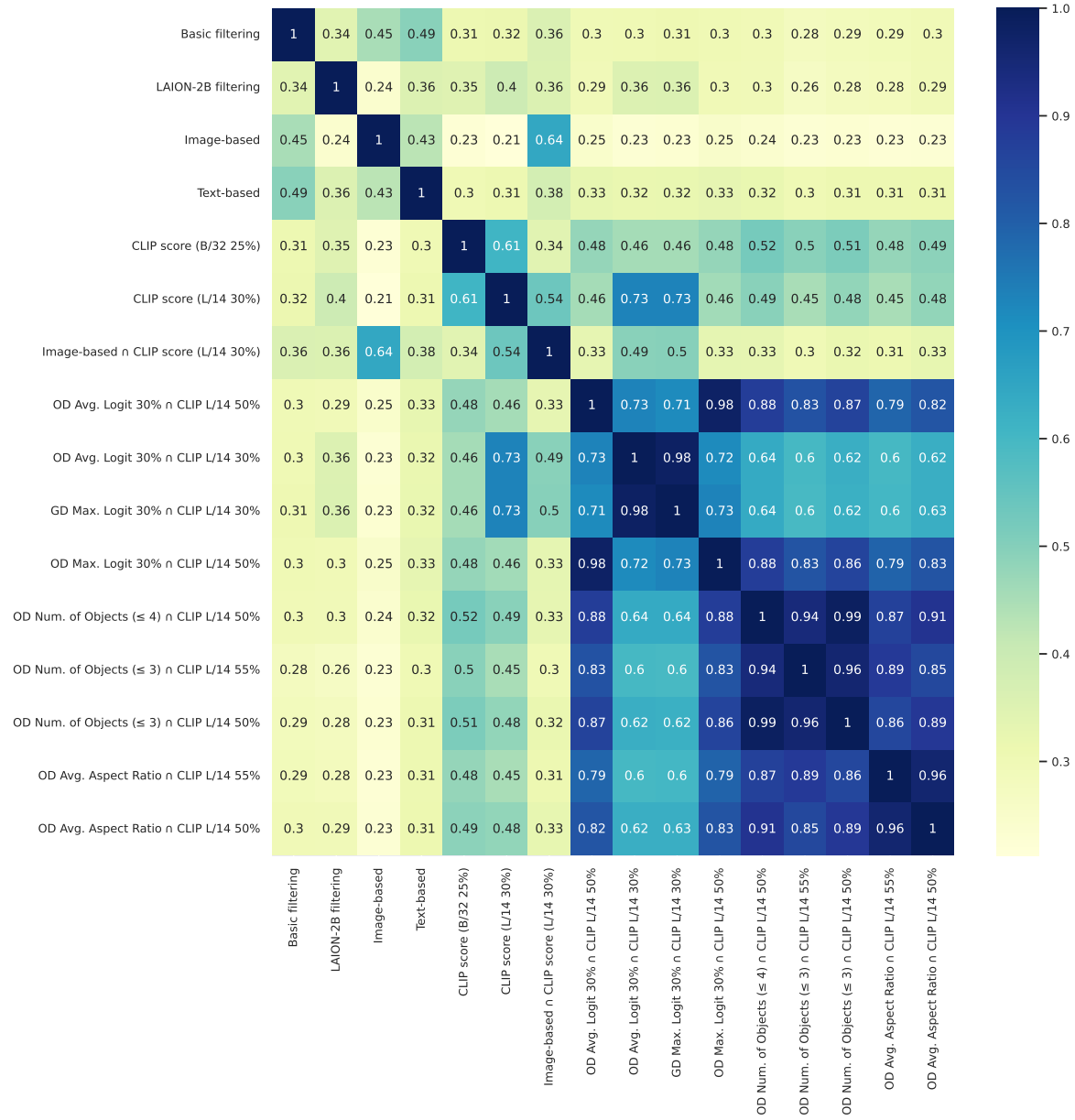

Figure 3: Correlation across designed filters.

Weak Supervision In our ensemble step, we used the most standard label model built on the Ising model [2, 19, 6]. While simple, the main drawback of such models is that they do not exploit input geometry, assuming globally uniform accuracy for each filter rule. To overcome such limitations, several existing works incorporate the

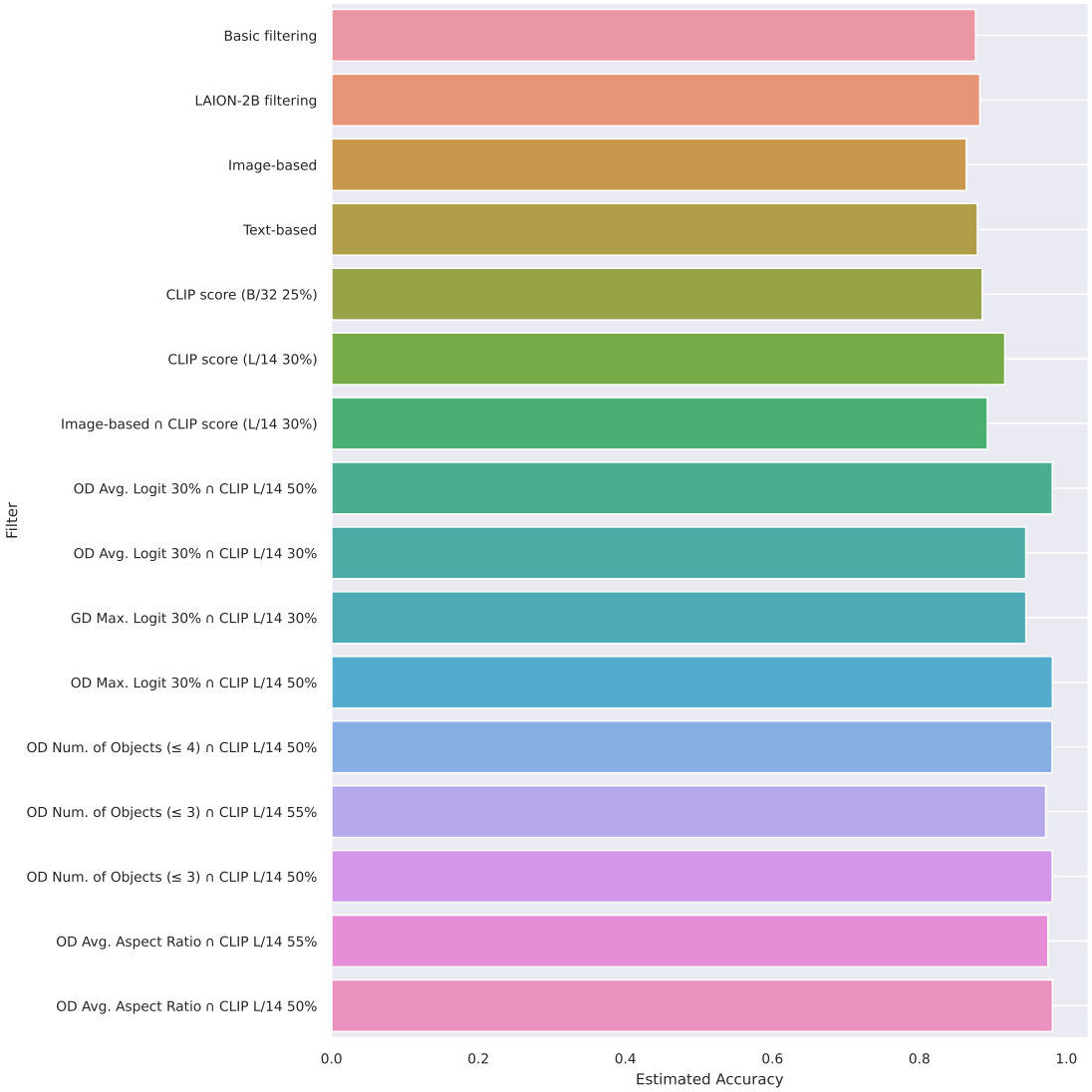

Figure 4: Estimated accuracy across designed filters.

input space into the label model. [4] suggested a partition-based label model, which separated parameters in each input partition. [29] applied a Gaussian process and a Bayesian label model to leverage input features. [24] provided a label model based on accuracy center and slope. While we mainly used Ising model-based techniques to ensemble filters, applying an embedding-based weak supervision approach may be more useful to aggregate filtered results, enabling each filter's strengths in specific input space to be better exploited.

C Discussion

Property of Contrastive Loss The contrastive loss underpins the training approach of CLIP. Though it is natural to filter the data via CLIP scores, some properties of the contrastive loss can provide further insights to curate a dataset. For example, [5] found 1) the contrastive loss is feasible with multiple objects while too many objects may undermine model learning, 2) the presence of dominant objects may suppress the learning of feature of small objects, 3) easy-to-learn features may suppress the learning of other features.

The first and the second points support our motivation when we were considering additional components — object detection filters, especially designing filters considering the number of objects and object relative size. The third point is related to another work, T-MARS [14], which enhances the filtering scheme by masking recognized text in the image and then re-scoring as text features are typically easy-to-learn, undermining other features. As such, exploring the properties of the contrastive loss may yield more insights and heuristics to design and craft filtering methods.