Parrot Captions Teach CLIP to Spot Text

Yiqi Lin[1,2*], Conghui He[1*✉], Alex Jinpeng Wang[2*], Bin Wang[1*], Weijia Li[3], and Mike Zheng Shou[2]

[1] Shanghai Artificial Intelligence Laboratory
[2] Show Lab, National University of Singapore
[3] Sun Yat-Sen University

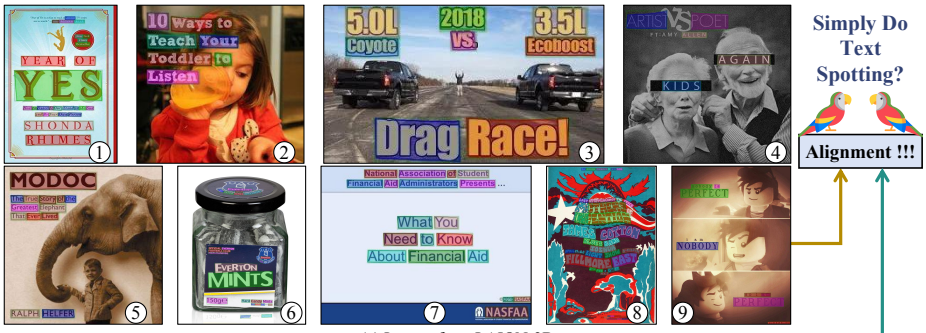
Abstract. Despite CLIP [29] being the foundation model in numerous vision-language applications, CLIP suffers from a severe text spotting bias. Such bias causes CLIP models to 'Parrot' the visual text embedded within images while disregarding the authentic visual semantics. We uncover that in the most popular image-text dataset LAION-2B [31], the captions also densely parrot (spell) the text embedded in images. Our analysis shows that around 50% of images are embedded with visual text content and around 30% of captions words are concurrently embedded in the visual content. Based on such observation, we thoroughly inspect the different released versions of CLIP models and verify that the visual text is a dominant factor in measuring the LAION-style image-text similarity for these models. To examine whether these parrot captions shape the text spotting bias, we train a series of CLIP models with LAION subsets curated by different parrot-caption-oriented criteria. We show that training with parrot captions easily shapes such bias but harms the expected visual-language representation learning in CLIP models across various vision-language downstream tasks. This suggests that it is urgent to revisit either the design of CLIP-like models or the existing image-text dataset curation pipeline built on CLIP score filtering. Project page: `https://linyq17.github.io/CLIP-Parrot-Bias/`

Keywords: Image-Text Dataset · Text Spotting Bias

1 Introduction

Recently, contrastive learning models [17, 29, 31] pre-trained with large-scale image-text pair data has led to numerous vision-language modeling task breakthroughs. Due to its efficiency and simplicity, the pioneering work CLIP [29] now serves as a foundation model in various applications [20, 26, 30, 49]. However, several works [4,13] have shown that CLIP models have perpetuating biases towards visual text [19,25], color [34,44], gender [41], etc. In this paper, we focus on probing the visual text bias in CLIP, i.e., the capacity of spotting text in images. Most of the previous cues [25, 29, 34] attribute the sources of biases to the noisy pre-training data. Therefore, we begin by taking a close look at the most popular dataset, LAION-2B [31].

* Equal contribution. ✉ Corresponding author (Email: heconghui@pjlab.org.cn).



(a) Images from LAION-2B

(b) Image Paired Captions

Fig. 1: In LAION-2B [31], image-text pairs with the Top-5% highest similarity score are most dominant by visual text! These samples have dense concurrent text appearing in captions and images (text form in pixels). We refer to their captions as **Parrot Captions** as they raise a question: *Dose CLIP Simply Parroting Text in Images for Vision-Language Alignment?* The concurrent text is spotted by the OCR model and highlighted with color in image-text pairs. (Best view in color)

Considering the massive scale of the image-text data, it is non-trivial to assess the bias simply with a rough estimation. To this end, we first do image clustering on the whole dataset and rank each cluster by CLIP scores to analyze the most preferred types of image-text pairs under CLIP score measurement. As shown in Fig. 1, we surprisingly observe that a decent number of samples with top CLIP scores have dense concurrent text appearing in the captions and the images in the form of pixels. These samples break the assumption that CLIP models leverage text supervision to align the visual and language concepts. We refer to these captions as **Parrot Captions** as they provide another shortcut to achieve the same goal by teaching CLIP to do text spotting even without perceiving the actual visual concepts. To understand the underlying impact, we analyze the parrot captions from three perspectives: dataset, widely used released models, and model training. The results lead to three key findings:

Firstly, captions in LAION-2B have a significant bias towards describing visual text content embedded in the images. We provide thorough profiling using off-the-self text spotting models on the LAION-2B dataset and show that over 50% of the images are embedded with visual text content. Moreover, by examining the spotted text content and the paired caption in the subset that images embedded with text, we find that over 60% of the captions at least have one concurrent word and reach at least around 30% words overlap

between the caption and spotted text from images. *This finding suggests that the basic assumption of image-text semantic alignment in CLIP does not fully stand its ground when training with LAION-style data.*

Secondly, released CLIP models have strong text spotting bias almost in every style of web images, resulting in CLIP-filtering datasets inherently biased towards visual text-dominant data. We investigate OpenAI released CLIP model's behaviors in the LAION-2B dataset by examining the difference between alignment scores before and after text removal. The results show that CLIP model predictions densely correlate the visual text embedded in images with their parrot captions. Next, we further study the preference of the text spotting capacity on text content in CLIP and OpenCLIP models. Note that CLIP is trained on WIT-400M, while OpenCLIP uses the LAION-2B dataset. Therefore, we use synthetic images embedded with specific rendered text to avoid overfitting in OpenCLIP models. Our analysis shows that OpenCLIP is more biased toward text spotting than CLIP. *We believe that the parrot caption plays a lurking role in training these released CLIP models and is the source of text spotting capacity instead of emergence behavior* [42].

Thirdly, CLIP models can easily learn text spotting capacity from parrot captions while failing to connect the vision-language semantics, just like a text spotting parrot. We sample different LAION-2B subsets curated by text-orientated criteria, including the embedded text ratio, the concurrent word ratios, and the relative CLIP score from text removal to train CLIP models under the same setting. The results show that using parrot captions data, CLIP model can learn strong text spotting capacity but lose most of the zero-shot generalization ability on image-text downstream tasks. Moreover, we also observe similar behaviors on various downstream tasks such VQA, captioning, and retrieval on BLIP [21] models. *Lastly, we argue that the existing data curation pipeline built on CLIP score and the contrastive fashion urgently needs to be re-examined by considering such hidden parrot captions.*

2 Related Work

2.1 Contrastive Vision-Language Pre-training

Modeling vision and language by aligning the embedding similarity between paired image-text data [17, 29, 31] has shown great potential for transferable to downstream vision-language tasks. The pre-training techniques mainly contain the vision encoder [10, 15] for image embedding encoding, text encoder [9] for text embedding modeling, and cross-modal contrastive learning [17, 22, 29, 47] for learning a joint embedding space of vision and language. The pioneering work CLIP [29] leverages 400 million noisy image-text pairs to learn transferable visual representation from text supervision and show impressive zero-shot performance for various vision-language tasks. Following CLIP, several vision-language models such as ALIGN [17], BASIC [27], and Open-CLIP [31] are proposed, and CLIP models have been replicated on various datasets including WIT [29], LAION [31], COYO [6], and DataComp [11]. We mainly profile

the LAION-2B [31] dataset due to its large scale and wide usage [26, 30] and two versions of pre-trained models, CLIP and OpenCLIP. Note that the 2 billion image-text pairs in the LAION-2B dataset are filtered by OpenAI released CLIP models, making OpenCLIP connect to CLIP closely.

2.2 Studying of CLIP Behaviors

Despite the strong zero-shot and transferable performance of CLIP, the perpetuating biases [1, 3, 13, 19, 41, 46] in CLIP are still not well investigated due to its large-scale noisy training data. Much research [2, 25, 34, 37, 40, 43, 44] focuses on revealing or enhancing the downstream performance with discovered bias in CLIP. For example, colorful masks [44] or red circles [34] applied to images can improve the zero-shot performance on visual localization tasks. In studying visual text content bias, [13] shows the multimodal neurons of CLIP not only respond to visual content and the visual text embedded in the image. Another work [19] shows that image recognition in CLIP can be strongly dominated by the visual text embedded in the image. To disentangle such bias, [25] attempts to separate the text spotting representation in pre-trained CLIP by training representation projection. Meanwhile, LoGoPrompt [33] enhances the classification performance by utilizing the visual text content as auxiliary prompts as input. Also, CLIPPO [39] shows that directly aligning the image and synthetic images print with the captions can perform similarly to CLIP without a text encoder.

2.3 Data Curation with Text Removal

Due to the successful practice of data curation in LAION datasets [31, 32] on scaling up the image-text datasets, searching advanced selection strategy to improve the data quality from common crawl data pool gains a growing interest [11]. Recently, several works [7, 24, 28] suggest that introducing text-related filtering methods improves the pre-training dataset quality. In DiHT [28], the data curation steps include filtering out the image-text pairs with high OCR confidence and matching text ratio. Moreover, [7, 24] mainly focus on studying the importance of filtering out the text-dominate images utilizing OCR models to improve pre-training dataset quality. Maini et al. [24] also draw the observation that 40% of LAION's image text is highly correlated with the caption, but only performing a small pilot study on 500 samples with manual judgment. *Differently, this paper makes the first attempt to systemically reveal the source of text spotting capacity in CLIP is the data bias and the consequences of such bias in existing widely used datasets and pre-trained models.*

3 Terminology

The data processing on images in the following sections mainly covers clustering, text spotting (OCR), and text inpainting. Firstly, we cluster all images based on feature similarity. For each image-text pair, we then use the pre-trained text

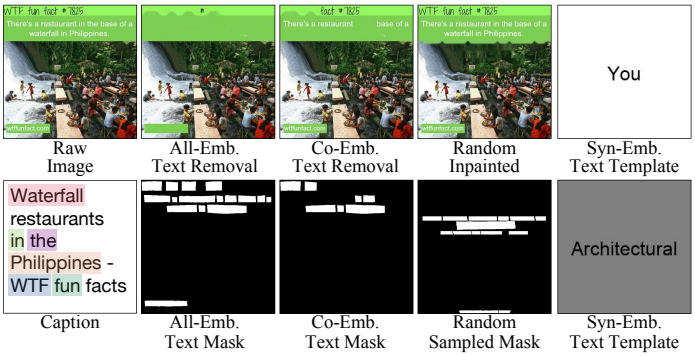


Fig. 2: Visualization of defined terminologies. Co-Emb. Text is marked in the caption.

Algorithm 1 Pseudocode of Detecting Co-Emb. Text (Rate)

```
# caption: captions from LAION-2B dataset.
# ocr_text: text spotted by OCR model.
cap_words, ocr_words = set(caption.split()), set(ocr_text.split())
co_emb_text = intersection(cap_words, ocr_words)
co_emb_text_rate = len(co_emb_text) / len(cap_words)
```

spotting model to detect and recognize the text print in image pixels. The mask images in Fig. 2 are the spotted text area. Next, we match the spotted text with the caption using Algorithm 1 to obtain the concurrent words and their ratio in captions. Lastly, we use inpainting to remove the text from the image for CLIPs' pattern ablation. To avoid confusion, we define these concepts as follows,

- **Embedded Text**: text spotted by OCR models from the images. To study the correlation of embedded text with captions, we define different kinds of embedded text as, **1) All-Emb. Text**: all the text is spotted from an image; **2) Co-Emb. Text**: spotted text concurrently appears in the image's corresponding captions; **3) Syn-Emb. Text**: synthetic text rendered in an image with a fixed font and a blank background. Fig. 2 shows examples of spotted embedded text by binary mask and the rendering synthetic text.
- **Co-Emb. Text Rate (CoTR)**: the word set IoU of Co-Emb. text and captions (See Algorithm 1).
- **Image w/ or w/o Embedded Text**: spotted text results of a given image are none-empty or empty.
- **Text Removal Image**: do inpainting in the specific spotted text area (All-Emb., Co-Emb., or Random). Random is implemented by sampling other image's text areas. For the different inpainting results, see Fig. 2.
- **Relative Scores (RSA/RSC)**: the difference in CLIP score between images modified by different inpainting operations while keeping the same captions. **RSA** and **RSC** are the short for the relative scores before and after removing All-Emb. text and Co-Emb. text.
- **Image Clusters**: image partitions based on K-Means.

Table 1: Overall parrot captions statistic. More than 50% of images are embedded with text, and 30% of caption words are printed in images!

Number of Total Images	1,985,284,122
Number of Images w/ Emb. Text	1,083,896,427
Number of Images w/ Co-Emb. Text	663,600,432
Co-Emb. Text Rate (in Total)	15.42%
Co-Emb. Text Rate (in Images w/ Emb. Text)	28.24%
Fuzzy Co-Emb. Text Rate (in Total)	30.46%
Fuzzy Co-Emb. Text Rate (in Images w/ Emb. Text)	55.79%

- **CLIP and OpenCLIP**: CLIP models are trained on WIT-400M [29] and LAION-2B [31] dataset.
- **N-gram Vocabulary (Vocab)**: the set of all contiguous N word sequences extracted from a text corpus, such as the collection of all captions.

4 Profiling LAION-2B Data

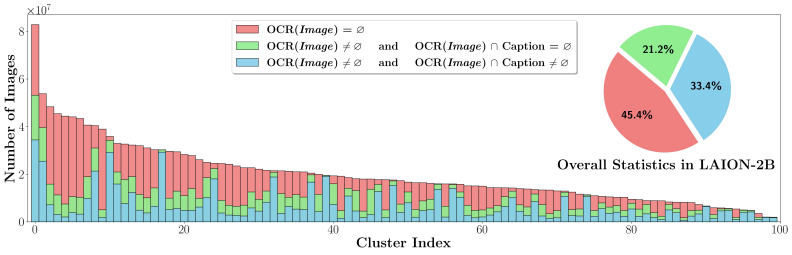
To better profile the image-text pair data on a billion scale, we first cluster all the images based on CLIP features into 4,000 clusters and sort each cluster with CLIP scores. After obtaining all the cluster labels, we use the SOTA text spotting model [45] to get the visual text content on all the collected images. Finally, we aggregate all the model-predicted results and compare them with their corresponding captions to bring out our observations.

4.1 Implementation Details

Clustering with CLIP Features: We first train K-Means (implemented by Faiss [18]) on the LAION-400M [32] subset using ViT-B-32 [10] CLIP features to speed up the clustering process. Due to the large memory consumption, we reduce the feature dimensions from 512 to 256 using PCA. Then, we partition the whole dataset using trained K-Means with the same feature extraction pipeline.
Text Spotting and Matching: To detect and recognize text across various scenes, we adopt DeepSolo [45] as our text spotting model and use the pre-trained checkpoints with the ViTAEv2-S [48] backbone in default setting. The output format of the text spotting model is a sequence of polygons of text location and their recognized characters. Despite its strong performance, we empirically find that DeepSolo can not handle the crowd scenes well (with more than 100 separate words) which is only a small proportion of the dataset ($\sim$2%). To identify the correlation between the spotted text and captions, we use Algorithm 1 to calculate the Co-Emb. text rate in each image-text pair. Considering the imperfect text spotting predictions might miss or misspell words, we also use Levenshtein distance to calculate the fuzzing similarity and reported in Tab. 1.

4.2 Statistic and Observations from LAION-2B

The overall statistics of the 2 billion image-text pairs are reported in Tab. 1. In summary, the images embedded with visual text content reach a surprisingly



(a) The ratio of different OCR-oriented data types in LAION-2B clusters.

(b) Top CLIP score samples visualization from 50 clusters with ▢ ratio over 80%.

Fig. 3: (a): Based on the OCR prediction results, the image-text pairs are divided into three types: ▢ image without visual embedded text content; ▢ the spotted text from the image has no concurrent text with the caption; ▢ the spotted text at least share one concurrent word with the caption. The clusters are merged from 4000 into 100 for a better view. **(b)**: In the clusters with high ▢ ratio, the top CLIP score samples contain various text sources, such as posters, book covers, advertisements, and even slides.

high proportion of 54.60% in the investigated data. Around 15% of words in the dataset captions are Co-Emb. text, and the proportion of Co-Emb. text can further reach 30% when considering the fuzzy matching results of the spotted text and captions. This suggests that CLIP models trained on these data might lead to a high bias toward text spotting. To better visualize the data distribution, we provide cluster-specific statics results and top CLIP score samples of text-dominated clusters in Fig. 3. We divide all images into 100 clusters based on visual similarity and visualize them according to the OCR results. Every cluster contains more or less images embedded with text. Combined with sample visualization, we observe that in the LAION collected data, the parrot captions cover various scenes. In the subsets of images embedded with text, around 60% of captions at least precisely parrot one concurrent word (Co-Emb. Text Rate > 0) appearing in the image. It suggests that the data collection pipeline of LAION [31] has a strong bias to introduce parrot captions from web data.

To better understand Co-Emb. Text, we provide a more thorough analysis of the word counting and text size of parrot captions. As shown in Fig. 4a, the results show that a large proportion of the Co-Emb. Text only takes a few words. However, we also find a large number of captions that are almost full parrot captions (see areas around the heatmap diagonal). Next, in Fig. 4b and Fig. 4c, we investigate the correlation between the size of concurrent words box in the image and CLIP score. The results show that the large text size does not usually lead to a higher score; meanwhile, the small text size can also dominate the score as the score can be significantly different after removing them. (Details

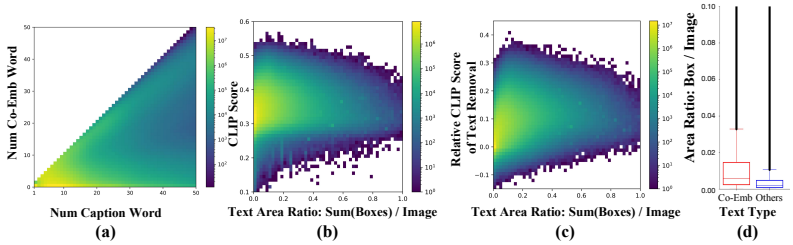


Fig. 4: (a): The number of caption words and associated spotted concurrent words based on precise word matching. **(b):** Distribution of total box area of concurrent words in the image and its CLIP score. **(c):** Distribution of total box area of concurrent words and its relative CLIP score before and after removing them from the image. **(d):** Distribution of text size of the single concurrent word and other spotted word.

of text removal described in Sec. 5.1). One possible reason is the text content and input resolution may matter more for CLIP. Moreover, we discover that the larger text is more likely to be parroted in the captions, as shown in Fig. 4d.

5 Inspecting Pre-Trained CLIP Models

It is important to note that the LAION-2B dataset collection pipeline uses CLIP score from OpenAI's model to filter out the image-text pair below **0.28**. Therefore, we inspect these two released CLIP models [29, 31] to answer better why LAION data contains such a high proportion of parrot captions. Specifically, OpenAI's CLIP model is trained on the WIT dataset (out-of-domain model), and OpenCLIP is trained on LAION-2B (in-domain model). We first study whether the embedded text is the key factor in CLIP filtering by ablating the embedded text using text inpainting. Moreover, we further investigate whether the text spotting capacity prefers specific text content by examining synthetic images with Syn-Emb. text.

5.1 Ablation of Embedded Text Removal

Text Removal via Inpainting: Given the OCR predicted results, we use the fast marching method [38] to inpaint the area of the spotted text polygons. We generate two versions of text removal results for each image with embedded text, i.e., All-Emb. text removal and Co-Emb. text removal, as the parrot caption prediction is imperfect due to the limitation of OCR models. We also generate random inpainting images with randomly sampled spotted text polygons from other images to ablate the distribution shift caused by image inpainting. Examples of the spotted text masks and inpainting results are shown in Fig. 2.
Results: Based on the OCR predicted results and text inpainting operations, we can obtain six types of LAION images, including ●): images without any embedded text (OCR results are empty); ●): images with any embedded text

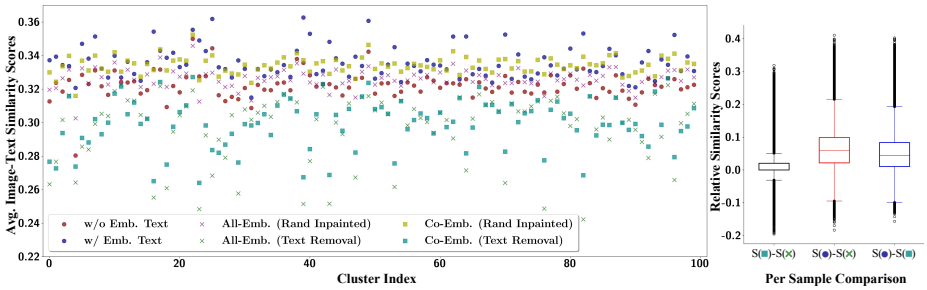


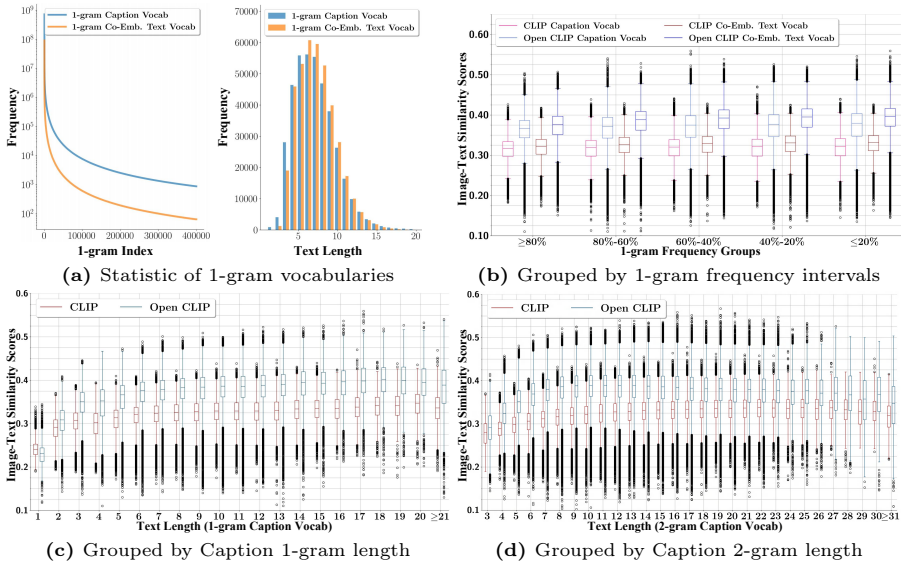
Fig. 5: Left: Mean CLIP scores of image-text pairs with different text removal operations depicted in Sec. 5.1, and grouped by cluster the same as Fig. 3. **Right:** Overall relative CLIP score distribution by comparing different text removal operations.

Table 2: Mean CLIP score of different setups of text removal.

Setup	Average CLIP Score
● Raw Images w/o Emb. Text	0.3223 ± 0.0078
● Raw Images w/ Emb. Text	0.3358 ± 0.0094
× All-Emb. Random Inpainted	0.3260 ± 0.0057
× All-Emb. Text Removal	0.2974 ± 0.0197
■ Co-Emb. Random Inpainted	0.3341 ± 0.0051
■ Co-Emb. Text Removal	0.2993 ± 0.0146

(OCR results are none-empty); ×): images with random inpainting by other image's All-Emb. text area; ×): images removed All-Emb. text (Inpaint all the areas of OCR predicted text); ■): images randomly inpainted by other image's Co-Emb. text area, and ■): images removed Co-Emb. text (Inpaint the areas of concurrent text in OCR predicted text and captions). Then, we calculate CLIP scores of all the groups of images and their paired captions using OpenAI released CLIP model (ViT-B-32). Fig. 5 reports the mean scores of different types of images in each cluster and raises four observations as follows: **I).** The images embedded with text achieve higher CLIP scores in most clusters than those without embedded text; **II).** CLIP scores significantly drop once we remove the text from the images compared to its random inpainting baseline. It indicates that the parrot captions correlate highly with CLIP score measurement; **III).** Not all the samples are dominated by the embedded text, as some samples achieve higher scores after removing text, indicating the embedded text can also be a distractor; **IV).** Most of the relative CLIP scores (S(■) - S(×)) between images removed Co-Emb. text and All-Emb. text are positive, as shown in the right of Fig. 5. The straightforward reason is the images lose more visual information due to the larger in-painting area. Another possible reason is the imperfect text spotting prediction or the corner cases in the matching algorithm leaking parts of the concurrent text in images.

Discussion: Due to the text removal, the image distribution may shift from the CLIP training set. Therefore, we provide two random inpainting baselines to examine the effect of distribution shift. In Tab. 2, we report the mean scores



(a) Statistic of 1-gram vocabularies

(b) Grouped by 1-gram frequency intervals

(c) Grouped by Caption 1-gram length

(d) Grouped by Caption 2-gram length

Fig. 6: OpenCLIP more bias towards text spotting than CLIP model. Grouped score distributions of prompting CLIP and OpenCLIP models with N-gram Syn-Emb. text and synthetic images for model preference investigation.

of different setups. Results show that the random baselines are very close to the raw image baseline, indicating that CLIP model is robust to the distribution shift caused by information loss in inpainted regions.

5.2 Prompting with Syn-Emb. Text

Generating Synthetic Images from N-gram Vocabulary: To investigate CLIP models' text spotting preference, we adopt a similar strategy in [25] to use synthetic images to embed specific text content by rendering text in a blank background. For each text, we use four fore-background style rendering templates (black-white, black-grey, white-grey, and white-black), as shown in Fig. 2. Different from the uniformly sampling letters in [25], we generate the text content from the N-gram vocabulary built from captions and Co-Emb. text to study the text spotting pattern. We only select the top frequent 400,000 grams for each vocabulary. The statistics of 1-gram vocabulary are shown in Fig. 6a, which is a long-tail distribution. Next, we calculate the synthetic images and the rendered text similarity on released ViT-B-32 CLIP and OpenCLIP models.

Results: Firstly, we examine whether CLIP models prefer recognizing more commonly seen words (with high frequency in vocabulary). Therefore, we group the 1-gram results based on their frequency interval in the whole vocabulary, as shown in Fig. 6b. The results show that OpenCLIP model clearly has a stronger text spotting capacity than CLIP, i.e., more biased towards text spotting. We also observe that all CLIP models are more sensitive to the vocabulary built

from the concurrent words. Interestingly, both CLIP and OpenCLIP models have slightly higher scores on the less frequent grams. Secondly, considering the long-tail grams might contain more characters, we further group the 1-gram and 2-gram results based on their text length in Fig. 6c and Fig. 6d. Note that the Co-Emb. text is not regularly arranged in the images, making it hard to extract continuous word sequences. Results show that all the models are better at spotting the longer words, possibly due to the tokenizer used in the text encoder, making them more discriminative. Meanwhile, in the groups of 2-gram samples, the scores gradually drop when spotting the highly long text, indicating that the spotting capacity of CLIP models is possibly built on word-by-word.

6 Training on Emb. Text Curated Data

Next, we dive deeper into training CLIP and BLIP [21] models on different Emb. text curated subsets and studying various downstream task behaviors.

Implementation Details: For CLIP, we use the open-source software OpenCLIP [16] for all CLIP model training. Our experiments are conducted on both ViT-B [10] and RN50 [15]. We use 4,096 batch size for 3M and 8,192 for 12M scale subsets. Other settings remain the same as [31]. For BLIP, we mainly conduct on 3M scale subsets with ViT-B [10]. The BLIP models are pre-trained for 10 epochs with an AdamW [23] optimizer. For downstream tasks, we finetune 10 epochs for VQA and 5 epochs for captioning and retrieval.

Evaluation: For CLIP, we follow the DataComp benchmark [11] using 38 zero-shot classification and retrieval tasks as evaluation. We report the average performance (Avg.) of the DataComp benchmark and two subset track performances, ImageNet (IN) and Retrieval (Ret.). To evaluate the text spotting capacity, we use the synthetic benchmark illustrated in Sec. 5.2 and a real-world benchmark sampled from LAION-2B as the validation set. In the synthetic benchmark, we calculate the similarity of all the 1-gram synthetic image-text pairs from caption vocabulary and report all the trained model results in Fig. 7. For the real-world benchmark, we sample 1M image-text pairs without any embedded text and 1M samples dominated by the parrot caption (RSC $\geq$ 0.2). Fig. 8 aggregates the mean scores of the 2M evaluation set and also reports the mean scores of applying text removal on the 2M evaluation set results. For BLIP, inspired by [12], we further evaluate the model behavior on downstream tasks requiring reading text, including Visual Question Answering (VQA), Image Captioning, and Text-Image Retrieval. Specifically, for the text-oriented tasks, we use Text VQA [36] and ST-VQA [5] for VQA, and TextCaps [35] for captioning and retrieval. Moreover, we also provide the same tasks on the datasets that only require the model to see, i.e., the natural image dataset. Similarly, we use VQAv2 [14] for VQA and COCO [8] for captioning and retrieval.

6.1 Ablation Study of Data Curation on CLIP

Curation I: Embedded Text in Images. To study the impact of embedded text on overall pre-train data quality, we sample three subsets: random baseline,

Table 3: Ablation of images embedded with or without text.

Data	Model	IN	Ret.	Avg.
3M Random	RN50	0.204	0.222	0.256
3M w/o Emb. Text	RN50	**0.228**	**0.232**	**0.282**
3M w/ Emb. Text Only	RN50	0.071	0.139	0.164
3M Random	ViT-B	0.131	0.148	0.210
3M w/o Emb. Text	ViT-B	**0.162**	**0.164**	**0.234**
3M w/ Emb. Text Only	ViT-B	0.052	0.111	0.153
12M Random	RN50	0.360	0.354	0.354
12M w/o Emb. Text	RN50	**0.409**	**0.361**	**0.372**
12M w/ Emb. Text Only	RN50	0.129	0.192	0.218
12M Random	ViT-B	0.314	0.299	0.351
12M w/o Emb. Text	ViT-B	**0.370**	**0.318**	**0.364**
12M w/ Emb. Text Only	ViT-B	0.129	0.172	0.225

Table 4: Ablation of different Co-Emb. Text Rate (CoTR).

Data (3M)	Model	IN	Ret.	Avg.
CoTR = 0.0	RN50	**0.193**	**0.229**	**0.247**
CoTR ≥ 0.3	RN50	0.031	0.110	0.137
CoTR ≥ 0.5	RN50	0.021	0.099	0.124
CoTR ≥ 0.8	RN50	0.012	0.082	0.096
CoTR = 1.0	RN50	0.012	0.074	0.102
CoTR = 0.0	ViT-B	**0.132**	**0.164**	**0.206**
CoTR ≥ 0.3	ViT-B	0.029	0.084	0.130
CoTR ≥ 0.5	ViT-B	0.021	0.082	0.119
CoTR ≥ 0.8	ViT-B	0.012	0.076	0.104
CoTR = 1.0	ViT-B	0.013	0.076	0.103

Table 5: Ablation of models trained on subsets sampled by different RSA.

Data (3M)	Model	Avg.S(●)	IN	Ret.	Avg.
RSA < 0.0	RN50	0.319	**0.181**	**0.220**	**0.239**
RSA ≥ 0.0	RN50	0.339	0.126	0.180	0.215
RSA ≥ 0.1	RN50	0.351	0.041	0.123	0.148
RSA ≥ 0.2	RN50	0.360	0.017	0.094	0.109
RSA ≥ 0.3	RN50	0.376	0.009	0.075	0.097
RSA < 0.0	ViT-B	0.319	**0.123**	**0.159**	**0.198**
RSA ≥ 0.0	ViT-B	0.339	0.079	0.129	0.185
RSA ≥ 0.1	ViT-B	0.351	0.031	0.103	0.134
RSA ≥ 0.2	ViT-B	0.360	0.012	0.080	0.103
RSA ≥ 0.3	ViT-B	0.376	0.006	0.070	0.096

Table 6: Ablation of models trained on subsets sampled by different RSC.

Data (3M)	Model	Avg.S(●)	IN	Ret.	Avg.
RSC < 0.0	RN50	0.326	**0.125**	**0.171**	**0.209**
RSC ≥ 0.0	RN50	0.345	0.062	0.129	0.168
RSC ≥ 0.1	RN50	0.354	0.014	0.091	0.106
RSC ≥ 0.2	RN50	0.364	0.008	0.084	0.104
RSC ≥ 0.3	RN50	0.380	0.005	0.058	0.084
RSC < 0.0	ViT-B	0.326	**0.079**	**0.129**	**0.174**
RSC ≥ 0.0	ViT-B	0.345	0.045	0.119	0.149
RSC ≥ 0.1	ViT-B	0.354	0.018	0.091	0.116
RSC ≥ 0.2	ViT-B	0.364	0.008	0.076	0.106
RSC ≥ 0.3	ViT-B	0.380	0.004	0.059	0.091

images without any embedded text, and images all embedded with text from LAION-2B. The subsets include 3M and 12M scales. The results in Tab. 3 show that images embedded with text generally reduce the pre-training dataset quality as all performance tracks significantly decrease. Meanwhile, in Fig. 7, the model trained with the images embedded with text achieves the strongest text spotting capacity compared to the random and images without text baselines.

Curation II: Co-Emb. Text Rate (CoTR). Tab. 3 reports CLIP models trained on parrot captions with different CoTR. We first select all the images with embedded text and then sample images based on the CoTR depicted at Algorithm 1 with different thresholds. With increasing CoTR, all the zero-shot benchmark performance drops significantly. Despite the images in the subset (CoTR = 0) all embedded with text, the pre-trained model performs similarly to the random baseline in Tab. 3. It indicates that the parrot caption is more crucial than embedded text in reducing the pre-trained data quality. For the text spotting capacity, Fig. 7 and Fig. 8 show that the increasing CoTR does not lead to stronger text spotting capacity, possibly due to the average length of captions decreasing in higher CoTR data.

Curation III: Relative Score from Text Removal (RSA & RSC). Given the observations in Sec. 5.1, we further select a series of subsets based on the relative score of images before and after text removal. The higher relative scores mean the masked text is more dominant in CLIP score measurement. In Tab. 5 and Tab. 6, we report the zero-shot performance of models trained on subsets with different relative score thresholds. CLIP models pre-trained with higher

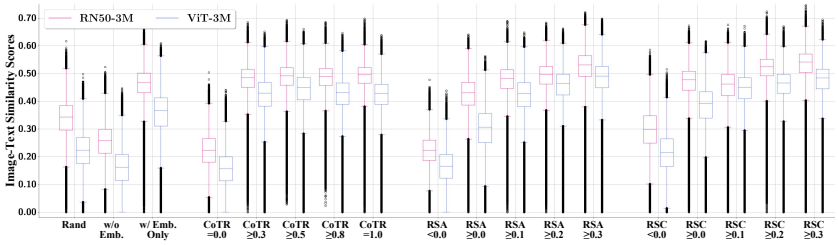


Fig. 7: CLIP models learn text spotting well from parrot captions. Benchmarking text spotting capacity of CLIP models with 1-gram caption vocabulary synthetic images dataset as the same as Sec. 5.2.

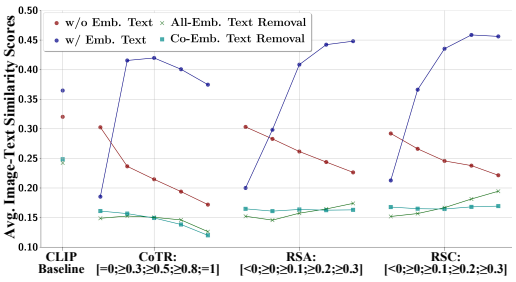


Fig. 8: Text spotting capacity validation on real images. Models trained with more parrot captions are better at aligning the image with parrot captions but perform worse at aligning images without embedded text.

RSA or RSC both get worse downstream performance. Notably, the average raw CLIP scores S(●) of these subsets have a positive correlation with RSA or RSC, indicating using CLIP scores from a biased pre-trained model as the data filtering strategy can be unreliable. When comparing the RSA and RSC, the results show that the samples dominated by the latter, i.e., parrot captions, are less informative for CLIP training. Meanwhile, Fig. 7 and Fig. 8 show that the text spotting capacity of CLIP can be further improved by training on the samples using relative scores as data curation criteria against CoTR.

6.2 More Investigation on Text-Oriented Tasks

Inspired by [12], we further train BLIP [21] models on the curated subsets to better understand the impact of parrot caption on various vision-language tasks, especially the tasks that require models to read the text. We chose BLIP for the ablation study instead of CLIP as it can be directly applied to all these tasks. As shown in Tab. 7, training BLIPs to spot text can boost their performance on the downstream tasks requiring the model to read but impede the performance of downstream tasks only requiring the model to see, which are consistent with the observation on classification tasks. Nevertheless, when BLIPs mainly focus

Table 7: BLIP downstream tasks performance of pre-training on different curated 3M subsets. The gray color represents tasks requiring the model to read the text from images, i.e., spotting text from images.

BLIP Data (3M)	Visual Question Answering (Acc)			Image Captioning (CIDEr)		Text-to-Image Retrieval (R@1)		Image-to-Text Retrieval (R@1)	
	VQAv2	TextVQA	ST-VQA	COCO	TextCaps	COCO	TextCaps	COCO	TextCaps
Rand	71.07	15.36	10.48	115.6	53.7	48.91	56.34	65.46	72.45
w/ Emb. Text	68.94	**19.05**	**12.65**	108.9	**92.1**	42.89	**70.1**	58.5	**81.42**
w/o Emb. Text	**71.22**	13.65	9.29	**116.2**	41.5	**49.96**	31.83	**66.5**	48.7
CoTR = 0.0	**71.11**	13.97	9.75	**116.3**	44.6	**49.55**	38.05	**66.08**	54.57
CoTR ≥ 0.3	67.4	19.28	11.81	104.9	**96.9**	37.78	**67.28**	51.98	**78.2**
CoTR ≥ 0.5	67.02	**19.64**	12.38	102.7	94.1	35.94	65.24	50.32	76.94
CoTR ≥ 0.8	66.38	18.50	12.00	100.9	91.6	34.13	62.65	46.9	73.56
CoTR = 1.0	66.18	18.47	**12.80**	101.2	91.3	33.55	61.83	46.62	73.05
RSA < 0.0	**70.79**	14.16	9.64	**115.7**	44.9	**48.25**	36.85	**64.72**	54.7
RSA ≥ 0.0	70.03	18.76	11.81	111.9	84.5	46.25	**68.61**	62.92	**81.23**
RSA ≥ 0.1	68.14	19.48	**13.33**	105.6	**96.1**	39.96	68.13	54.64	79.37
RSA ≥ 0.2	66.01	**21.06**	11.85	98.7	94.4	33.03	64.17	47.12	75.33
RSA ≥ 0.3	64.20	18.44	12.04	95.26	91.1	26.64	60.11	37.3	70.24
RSC < 0.0	**70.13**	15.19	10.74	**112.2**	46.7	**46.8**	41.95	**63.24**	58.05
RSC ≥ 0.0	68.86	20.12	**13.75**	107.8	93.5	42.0	**69.78**	57.42	**80.92**
RSC ≥ 0.1	67.35	**20.54**	12.84	103.4	**96.9**	36.4	66.69	51.02	77.79
RSC ≥ 0.2	62.62	20.32	13.14	98.7	92.8	30.08	61.38	42.96	71.98
RSC ≥ 0.3	63.75	18.94	13.03	92.9	88.7	24.23	58.35	34.72	68.95

on reading, e.g. (RSA ≥ 0.3), their text-oriented and natural downstream performance also decreases. In other words, the parrot captions can benefit the text-orient downstream tasks while requiring careful data mixing trade-off. We believe that understanding parrot captions is essential for revealing the underlying mechanisms of existing large vision-language systems.

7 Conclusion and Discussion

The popularity of vision-language contrastive loss stems from its efficiency and simplicity. However, the analysis and experiments we presented show that the embedded text in images and their parrot captions plant significant text spotting bias due to such contrastive fashions. Firstly, almost 30% of the captions in the widely used LAION-2B dataset are biased towards parroting the embedded text in images. Secondly, the pre-trained CLIP models have strong preferences for the image-text pair with parrot captions, which achieve higher similarity scores than those without. Finally, using data biasing to parrot captions, we can easily train a CLIP model with a strong text spotting bias. Our work demonstrates the emergency of reviewing the impact of parrot captions in the entire ecosystem of CLIP models. Our future endeavor involves building a bias-aware data curation pipeline and a robust training function to mitigate such issues.

Acknowledgements

This project was supported by the National Key R&D Program of China (No. 2022ZD0160101) and the Shanghai Artificial Intelligence Laboratory. Mike Shou does not receive any funding for this work.

References

1. Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J.W., Brundage, M.: Evaluating clip: towards characterization of broader capabilities and downstream implications. arXiv preprint arXiv:2108.02818 (2021)
2. Alabdulmohsin, I., Wang, X., Steiner, A.P., Goyal, P., D'Amour, A., Zhai, X.: Clip the bias: How useful is balancing data in multimodal learning? In: ICLR (2024)
3. Ali, J., Kleindessner, M., Wenzel, F., Budhathoki, K., Cevher, V., Russell, C.: Evaluating the fairness of discriminative foundation models in computer vision. In: AIES. pp. 809–833 (2023)
4. Berg, H., Hall, S.M., Bhalgat, Y., Yang, W., Kirk, H.R., Shtedritski, A., Bain, M.: A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. arXiv preprint arXiv:2203.11933 (2022)
5. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: ICCV. pp. 4291–4301 (2019)
6. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset` (2022)
7. Cao, L., Zhang, B., Chen, C., Yang, Y., Du, X., Zhang, W., Lu, Z., Zheng, Y.: Less is more: Removing text-regions improves clip training efficiency and robustness. arXiv preprint arXiv:2305.05095 (2023)
8. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. arXiv preprint arXiv:2304.14108 (2023)
12. Ganz, R., Nuriel, O., Aberdam, A., Kittenplon, Y., Mazor, S., Litman, R.: Towards models that can see and read. arXiv preprint arXiv:2301.07389 (2023)
13. Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., Olah, C.: Multimodal neurons in artificial neural networks. Distill **6**(3), e30 (2021)
14. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR. pp. 6904–6913 (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
16. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021), `https://doi.org/10.5281/zenodo.5143773`
17. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. pp. 4904–4916. PMLR (2021)

18. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2019)
19. Lemesle, Y., Sawayama, M., Valle-Perez, G., Adolphe, M., Sauzéon, H., Oudeyer, P.Y.: Language-biased image classification: evaluation based on semantic representations. arXiv preprint arXiv:2201.11014 (2022)
20. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: ICLR (2022)
21. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML. pp. 12888–12900. PMLR (2022)
22. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. NeurIPS **34**, 9694–9705 (2021)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
24. Maini, P., Goyal, S., Lipton, Z.C., Kolter, J.Z., Raghunathan, A.: T-mars: Improving visual representations by circumventing text feature learning. arXiv preprint arXiv:2307.03132 (2023)
25. Materzyńska, J., Torralba, A., Bau, D.: Disentangling visual and written concepts in clip. In: CVPR. pp. 16410–16419 (2022)
26. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
27. Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A.W., Yu, J., Chen, Y.T., Luong, M.T., Wu, Y., et al.: Combined scaling for zero-shot transfer learning. Neurocomputing **555**, 126658 (2023)
28. Radenovic, F., Dubey, A., Kadian, A., Mihaylov, T., Vandenhende, S., Patel, Y., Wen, Y., Ramanathan, V., Mahajan, D.: Filtering, distillation, and hard negatives for vision-language pre-training. In: CVPR. pp. 6967–6977 (2023)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
31. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. NeurIPS **35**, 25278–25294 (2022)
32. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
33. Shi, C., Yang, S.: Logoprompt: Synthetic text images can be good visual prompts for vision-language models. In: ICCV. pp. 2932–2941 (2023)
34. Shtedritski, A., Rupprecht, C., Vedaldi, A.: What does clip know about a red circle? visual prompt engineering for vlms. arXiv preprint arXiv:2304.06712 (2023)
35. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: ECCV. pp. 742–758. Springer (2020)
36. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: CVPR. pp. 8317–8326 (2019)
37. Tanjim, M.M., Singh, K.K., Kafle, K., Sinha, R., Cottrell, G.W.: Discovering and mitigating biases in clip-based image editing. In: WACV. pp. 2984–2993 (2024)

38. Telea, A.: An image inpainting technique based on the fast marching method. Journal of graphics tools **9**(1), 23–34 (2004)
39. Tschannen, M., Mustafa, B., Houlsby, N.: Clippo: Image-and-language understanding from pixels only. In: CVPR. pp. 11006–11017 (2023)
40. Wang, H., Zhan, Y., Liu, L., Ding, L., Yu, J.: Balanced similarity with auxiliary prompts: Towards alleviating text-to-image retrieval bias for clip in zero-shot learning. arXiv preprint arXiv:2402.18400 (2024)
41. Wang, J., Liu, Y., Wang, X.E.: Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. arXiv preprint arXiv:2109.05433 (2021)
42. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022)
43. Xu, Y., Xu, Z., Chai, W., Zhao, Z., Song, E., Wang, G.: Devil in the number: Towards robust multi-modality data filter. arXiv preprint arXiv:2309.13770 (2023)
44. Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.S., Sun, M.: Cpt: Colorful prompt tuning for pre-trained vision-language models. arXiv preprint arXiv:2109.11797 (2021)
45. Ye, M., Zhang, J., Zhao, S., Liu, J., Liu, T., Du, B., Tao, D.: Deepsolo: Let transformer decoder with explicit points solo for text spotting. In: CVPR. pp. 19348–19357 (2023)
46. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: ICLR (2022)
47. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: CVPR. pp. 18123–18133 (2022)
48. Zhang, Q., Xu, Y., Zhang, J., Tao, D.: Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. IJCV pp. 1–22 (2023)
49. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV **130**(9), 2337–2348 (2022)