

# CLIP is Almost All You Need: Towards Parameter-Efficient Scene Text Retrieval without OCR

## Supplementary Material

### 1. Methodology Explanation and Supplement

#### 1.1. Zero-Shot Performance of ViT-based EXCLIP with the Split Strategy

As shown in Fig. 1, ViT-based EXCLIP models generalize better with the split strategy when dealing with high-resolution input. We divide input images into  $2 \times 2$  splits as the default setting for ViT-based models.

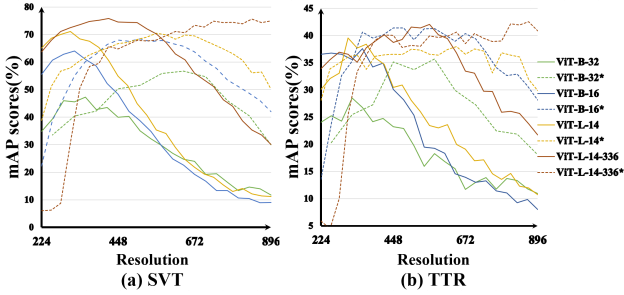


Figure 1. Zero-shot performance of EXCLIP models at different input resolutions on the SVT and TTR datasets. \* denote the ViT-based EXCLIP models with  $2 \times 2$  splits.

#### 1.2. More Explanation for Visual Context Dropout

Visual context dropout (VCD) dropouts the visual context in the last attention layer when computing the local visual features  $\mathbf{v}^l \in \mathbb{R}^{N \times N_v \times E}$ . From the perspective of scaled dot-product attention, the global visual features  $\mathbf{v}^g \in \mathbb{R}^{N \times E}$  can be obtained by:

$$\mathbf{v}_i^g = \sum_j \text{softmax}\left(\frac{f_q(\mathbf{v}_i^g) \cdot f_k(\mathbf{v}_{i,j}^l)}{C}\right) f_v(\mathbf{v}_{i,j}^l), \quad (1)$$

where  $f_q, f_k, f_v, C$  denote the linear transform for  $\mathbf{v}^g, \mathbf{v}^l, \mathbf{v}^l$ , and the learned temperature respectively. In CLIP, the query projection  $f_q$  is only optimized for the global representation  $\mathbf{v}^g$ . When we use the last attention layer to compute  $\mathbf{v}^l$ ,  $\mathbf{v}^g$  is replaced with  $\mathbf{v}^l$ , making the jointly optimized representation space biased. From the other perspective, CLIP encourages the model to learn discriminative features, resulting in focused and Gaussian-like attention, which can be approximated to one hot distribution. Based on the hypothesis, the global presentation  $\mathbf{v}^g$  is within an approximate representation space with some transformed local features in  $f_v(\mathbf{v}^l)$ . Thus we directly use a one-hot-like probability distribution to generate an approximately aligned rep-

resentation  $\mathbf{v}^l$ , which can be simplified as:

$$\mathbf{v}_{i,k}^l = \sum_j \mathbb{I}(j = k) f_v(\mathbf{v}_{i,j}^l) = f_v(\mathbf{v}_{i,k}^l). \quad (2)$$

### 2. Supplementary Experiments

#### 2.1. Datasets

Table 1. Statistics of the training and test datasets.

Dataset	MLT5k	SVT	STR	TTR	PSTR	OCTTR
#Images	5161	249	10000	300	1080	300
#Queries	17365	427	50	60	36	60

**Multi-lingual Scene Text 5k (MLT5k)** dataset [12] is a subset of MLT2017 [6]. We follow [12] to construct the dataset, resulting in 5161 images containing Latin texts.

**Street View Text (SVT)** dataset [13] has 349 images collected from Google Street View. This dataset contains 100 and 249 images in the training and the test sets, respectively. 427 annotated words in the test set are taken as the queries.

**IIIT Scene Text Retrieval (IIIT-STR)** dataset [5] consists of 50 query words and 10000 images. It is challenging due to the variation of fonts, styles, and viewpoints, as well as the large proportion of background images.

**Total-Text Retrieval (TTR)** dataset [15] originates Total-Text [1] which consists of 1255 training and 300 test images. Following [12, 15], the query is selected from the words in the test set by the frequency of instances. Specifically, words with an appearance frequency of less than 4 are filtered, resulting in 60 queries.

**Phrase-level Scene Text Retrieval (PSTR)** dataset [16] is a phrase-level STR dataset, in which a query contains 2 to 4 words. The text queries and images are collected from the TextOCR dataset [9] and the Google image search engine. 36 query words and 1080 images are included in PSTR.

**OCTT Retrieval (OCTTR)** dataset originates from the OCTT dataset [8] which focus on occluded text spotting. OCTT comes from the test set of the Total-Text dataset, which contains 300 images that at least one character is weakly occluded following the same procedure described in VisionLAN [14]. Despite the occluded text instances in images, the annotations of the OCTTR dataset are the same as those of the TTR dataset.

Together with the datasets mentioned in the main text, the overall statistics of training and test datasets are summarized in Tab. 1.

Table 2. Zero-shot performance of more contrastive models.

Method		params (M)	SVT	IIIT-STR	TTR	Avg.
ZS	EVA01-g-14-plus-224 [10]	1366.62	66.99	62.75	38.3	56.01
	EVA02-L-14-336 [10]	428.08	<u>79.82</u>	71.35	<b>48.02</b>	66.40
	ViT-SO400M-14-SigLIP-384 [17]	877.96	79.24	<u>76.86</u>	<u>47.97</u>	<u>68.02</u>
	ViT-L-16-SigLIP-384 [17]	652.48	79.16	74.91	46.56	66.88
	ViT-B-16-SigLIP-512 [17]	203.79	<b>80.43</b>	<b>78.80</b>	46.97	<b>68.73</b>
	BLIP-ViT-L-384 [4]	892.26	50.85	52.13	19.64	40.87
CAYN	ViT-B-16-SigLIP-512	204.02	84.89	88.81	76.08	83.26

## 2.2. Implement Details

Following existing works [12, 15], we first filter the special marks and keep characters and numbers in the text query. Then all the queries are transformed to lowercase before input into the model.

## 2.3. Experimental Results

### 2.3.1 Zero-Shot Performance of More CLIP Models

We evaluate representative *eva-clip*, *siglip*, and *blip* models in Tab. 2. SigLIP-512 achieves the best performance and can be further improved with CAYN, demonstrating resolution and CAYN matter.

### 2.3.2 Ablation Study on Incorrect Negatives Calibration and Different Sampling Strategies in ITM

The training data is constructed by generating image-text pairs, in which the text is the word that occurs in the image. In CLIP, texts or images from different matching pairs are taken as negatives. However, although the training data is randomly shuffled and the batch size is not large, a negative pair generated in this way may be a matched pair and depress the optimization process. To calibrate the effect, we introduce the use of the accurate matching label. Specifically, negative label calibration (NLC) is achieved by excluding incorrect labeling examples from the sampling process of the negatives.

Table 3. Comparison of variants of CAYN with different model configurations on the SVT and TTR datasets. “NLC” is short for negative label calibration. “US” and “HEM” denote uniform sampling and hard example mining.

#	Retrieval	Reranking		SVT	TTR
	NLC	NLC	Sampling		
0	×	-	-	81.98	66.54
1	✓	-	-	81.86	67.54
2	×	✓	US	<b>85.29</b>	<u>73.01</u>
3	×	✓	HEM	<u>85.05</u>	<b>74.04</b>
4	×	×	US	84.91	72.28
5	×	×	HEM	66.95	58.65
6	✓	✓	HEM	84.70	<u>73.01</u>

As shown in Tab. 3, using NLC in the retrieval stage overall outperforms the one without NLC (comparing #0

with #1), bringing a 1.00% increase on TTR and a 0.12% decrease on SVT. Without NLC, decreases of 0.32%/0.73% (comparing #2 with #4) with uniform sampling (US) and 18.10%/15.39% (comparing #3 with #5) with hard example mining (HEM) are observed on the SVT/TTR datasets. The results illustrate that NLC is important in the optimization of reranking, especially for the HEM strategy. In general, the utilization of the HEM strategy in the reranking stage outperforms the one with the US (comparing #2 with #3), bringing a 1.03% increase on the TTR dataset and a 0.24% decrease on SVT. However, decreases of 0.35%/1.03% (comparing #3 with #6) are found on SVT/TTR when further combining NLC in the retrieval stage. We find that removing NLC in the retrieval stage brings better hard negative examples, benefiting the optimization of the reranking stage. Finally, we adopt the configuration #3 as the default optimization setting.

### 2.3.3 Does Textual Context Dropout Work?

To validate this, we perform a context drop in the last attention layer in textual encoders as in VCD and compare it with the vision-only query in ITM (in Eq. (5) of the main text). The performance drops from 72.80%/61.03% (refer to Tab. 8 in the main text) to 66.88%/54.37% on the SVT/TTR datasets. We find the reason lies in the casual attention and strong sequential dependency within the CLIP textual encoder. The decrease demonstrates that the alignment between the visual and textual modalities is further depressed. So we keep the original attention for computing the local textual feature  $t^l \in \mathbb{R}^{N \times N_T \times E}$ .

### 2.3.4 Choice of Hyper-Parameter $\lambda$

In Tab. 4, with the increase of  $\lambda$ , optimization focuses more on image-text matching. It works well when  $\lambda$  is set to 1 or 2. We choose  $\lambda = 1$  as the default value.

Table 4. Performance of different  $\lambda$  with CAYN-RN50.

$\lambda$	0.5	1.0	2.0	4.0
SVT	85.02	<u>85.05</u>	<b>85.41</b>	84.91
TTR	72.74	74.04	<u>74.20</u>	<b>74.38</b>
IIIT-STR	82.84	<b>82.88</b>	<u>82.73</u>	82.33

Table 5. Replacing VPA with other adapters in CAYN.

CAYN-RN50	Params	SVT	STR	TTR	Avg.	CAYN-ViT-B-16	Params	SVT	STR	TTR	Avg.
CLIP-adapter [2]	0.52M	82.53	76.18	63.72	74.14	LST [11]	2.67M	71.57	66.96	60.07	66.20
VPA (Ours)	0.20M	85.05	81.79	74.04	80.29	VPA (Ours)	0.22M	77.44	79.73	68.45	75.21

Table 6. Comparison with subdivision-enhanced FDP-L.

Method	Params	Supv.	SVT	IIIT-STR	TTR	Avg.	FPS
FDP-L (MM'24)	33.45M	L+T	89.63	89.46	79.18	86.09	11.82
FDP-L* (MM'24)	33.45M	L+T	91.18	<b>91.49</b>	82.02	88.23	3.04
CAYN-RN50x16	0.45M	T	<u>92.46</u>	89.49	<u>85.98</u>	<u>89.31</u>	<b>38.79</b>
CAYN-RN50x16-768	0.45M	T	<b>92.72</b>	<u>90.91</u>	<b>86.90</b>	<b>90.18</b>	<u>28.78</u>

Table 7. Statistics on Factors Considering the Model’s Efficiency.

Method	Parameters	Time Consumption		GPU Memory Usage
	Tuned/Total	Training	Inference	Training/Inference
CAYN-RN50	0.20/102.21M	44min	80.13FPS	9.37/1.13GB
CAYN-RN50x4	0.31/178.62M	48min	53.22FPS	12.49/1.83GB
CAYN-RN50x16	0.45/291.43M	88min	38.79FPS	18.29/2.83GB
FDP-L*(reproduced)	33.45/324.43M	571min	4.33FPS	33.12/6.39GB

### 2.3.5 Choice of Reranking Image Number $K$

As shown in Tab. 8, reranking brings consistent improvement across different  $K$  values. Though carefully tuning  $K$  may achieve better performance, we set  $K$  to 32, proving to work well on both datasets.

Table 8. Comparison of different choices of  $K$  in reranking on the SVT and TTR datasets regarding mAP (%).

$K$	2	4	8	16	32	64	128
SVT	83.41	85.36	<u>85.47</u>	<b>85.75</b>	85.05	84.89	84.60
TTR	67.45	68.66	69.72	71.88	<u>74.04</u>	<b>74.54</b>	74.09

### 2.3.6 Replacing VPA with Other Adapters

Unlike general PET methods, VPA aims to compensate for visual position information in high-resolution representation (placed after “Img2Tokens” rather than other positions in Fig.5 and evaluated in Tab.5 of the main text), which matters for STR. Tab.5 shows VPA outperforms *CLIP-adapter* by 4.10%/10.53% on SVT/TTR in the retrieval-only setting. In Tab. 5, replacing VPA with *CLIP-adapter* or *LST* in CAYN results in significant decreases of 6.15% or 9.01% in average mAP (SVT/STR/TTR) and increased parameters, demonstrating the superiority of VPA.

### 2.3.7 Comparison with Subdivision-Enhanced FDP-L (FDP-L\*)

In Tab. 6, besides the fewer tuned parameters and weaker required labels, CAYN can outperform FDP-L variants on SVT/TTR and is comparable to them on IIIT-STR. With an increasing resolution of 768 and a reranking number of 96, CAYN can be further improved to an average mAP of

90.18%. Direct subdivisions do not work well on CAYN due to severe scale mismatch in train/test (Fig.4 of main text) and semantic destruction from text cutting.

### 2.3.8 Discussing the Proposed PFCA with that in Text-Video Retrieval (TVR) Methods

Unlike the attention in TVR which directly extends the contrastive relation from images to frames, PFCA explores the transferable contrastive relation from inter-image to intra-image together with VCD for STR, which is crucial for perceiving visual text and is absent in TVR.

### 2.3.9 More in-Depth Analysis on Efficiency

Besides the *training* and *inference time*, we supplement the *total parameters* and *memory usage* for comprehensive analysis in Tab. 7 on a NVIDIA GeForce RTX 4090 GPU. Compared with FDP-L\*, CAYN models require low sources for training (<2H & <20GB) due to a relatively short path for gradient propagation and achieve a low proportion of tuned parameters and fast inference speed, demonstrating that CAYN can be both parameter- and memory-efficient.

## 2.4. Flexibility on Phrase-Level Text Retrieval

To demonstrate the flexibility of the proposed method, we evaluate CAYN-RN50 on the PSTR dataset. As shown in Tab. 9, the proposed CAYN achieves 92.42% mAP and outperforms TDSL [12] and FDP-S [16] under the same RN50 backbone, demonstrating the flexibility and potential of the CAYN to deal with queries with more complex semantics. Besides, CAYN-R50 runs the fastest at 79.62 FPS among these methods. When scaling the model size, CAYN

RN50x4 achieves 95.60% at 52.88 FPS, and the performance tends towards saturation with RN50x16 (95.90%).

Table 9. Comparison with existing methods on the PSTR dataset.

Method	mAP	FPS
Gomez <i>et al.</i> [3]	68.01	42.45
TDSL [12]	89.40	11.34
FDP-S [16]	92.28	45.11
CAYN-RN50 (Ours)	92.42	<b>79.62</b>
CAYN-RN50x4 (Ours)	<b>95.60</b>	<b>52.88</b>
CAYN-RN50x16 (Ours)	<b>95.90</b>	38.54

## 2.5. Robustness on Occluded Text Retrieval

We construct OCTTR from the OCTT dataset [7], which includes occluded and deformed text and originates from Total-Text. OCTTR shares the same GT with TTR except for images. In Tab. 10, CAYN outperforms existing methods significantly and shows good robustness against occlusion, demonstrating explicit OCR process can be omitted by unleashing cross-modal semantic knowledge from CLIP.

Table 10. Occluded and deformed text retrieval on OCTTR.

Method	TTR	OCTTR	$\Delta$
Gomez <i>et al.</i> (ECCV’18)	66.02	63.93	2.09
TDSL (CVPR’21)	76.38	72.78	3.60
CAYN-RN50 (Ours)	74.04	69.32	4.72
CAYN-RN50x4 (Ours)	81.90	79.53	2.37
CAYN-RN50x16 (Ours)	85.98	84.44	1.54

## 2.6. More Visualization Results

### 2.6.1 Visualization of Retrieval Results

As shown in Fig. 2, CAYN can accurately retrieve scene text with different scales and shapes as well as complex semantics, demonstrating the effectiveness and robustness of the proposed method.

### 2.6.2 Visualization of Attention Maps

As shown in Fig. 3, PFCA enables the CLIP models to focus on the region in document images corresponding to the query text. In this way, an implicit localization is performed with PFCA elegantly, making the STR free of an explicit localization process. In addition to the effectiveness and efficiency brought to reranking, PFCA provides a direct reference for the interpretability of scene text retrieval.

We can observe that the attention map of accurately localized regions tends to be concentrated and Gaussian-like. The reason may be that the conservative language-image pre-training encourages the model to focus on the most discriminative features. Interestingly, in attention maps with

phrase-level queries, we find the attention map tends to focus on the word with concrete meanings, e.g., “yourself” in “do it yourself” and “original” in “the original”.

## 2.7. Failure Case Analysis

We perform failure case analysis on the four datasets and analyze the queries with the ten lowest APs. The typical failure cases can be summarized in Fig. 4: 1) Incorrect annotation. We find there exist missing labels in existing datasets, especially in the SVT dataset, as shown in Fig. 4(a). 2) Tiny scale. As shown in Fig. 4(b), tiny scale, as well as low quality, brings serious challenges, which are also shared by other scene text extraction and understanding tasks. 3) Extreme shape. Due to the lack of curved texts in training, as shown in Fig. 4(c), it’s difficult for the model to perceive nearly vertical and reversed curved text instances. 4) Semantic bias. As shown in Fig. 4(d), the query with “technology” performs much worse than “institute”, even though “institute” and “technology” always occur together within the phrase “indian institute of technology” in the IIIT-STR dataset.

To better illustrate this phenomenon, we provide a visualization of attention maps of PFCA. As shown in Fig. 5, the attention maps with the query “technology” are less accurate or concentrated than those with “institute”, which qualitatively explains the bad performance for the query “technology”.

We attribute the reason to the semantic bias in CLIP models, which may be caused by imbalanced data distribution in the training datasets. Another phenomenon we observed is that content words tend to perform better than function words, consistent with the conclusion in FDP [16]. The reason may lie in the dominant role of content words in the semantics of sentences in the CLIP textual encoder, which tends to correspond to a concrete entity. However, it’s a common issue that is shared in the CLIP-based method.

Among all datasets, the SVT is relatively sparse, where a query matching one image occurs frequently, and missing retrieving the query easily depresses the overall performance. Compared with the other three datasets, the performance of the PSTR dataset is much better. Even the lowest AP with the query “bud light” achieves an AP of 75.57%. We guess the reasons are that 1) the phrase-level provides more specific semantics and 2) a phrase-level query contains at least one content word.

## 2.8. Limitation

Due to the semantic bias in CLIP, in which Latin text dominates the pre-training data, the performance on multilingual scene text and out-of-vocabulary (OOV) words is not good enough. However, this characteristic originates from CLIP. We’d like to explore the improvement in future work.



Dataset	Query	Retrieval Results				
SVT	"street"					
	"museum"					
IIIT-STR	"accenture"					
	"nokia"					
TTR	"restaurant"					
	"company"					
PSTR	"bank of america"					
	"have a nice day"					

Figure 2. Visualization of the top-5 retrieval results from CAYN-RN50 on the SVT, IIIT-STR, TTR, and PSTR datasets. The correct results are highlighted in green, while the incorrect ones are highlighted in red.


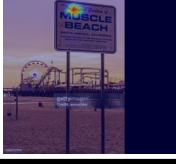

Dataset	Query	Visualized Attention Maps				
SVT	"hotel"					
	"the"					
IIIT-STR	"coffee"					
	"school"					
TTR	"market"					
	"shop"					
PSTR	"do it yourself"					
	"the original"					

Figure 3. Visualization of the attention maps in the parameter-free cross-attention (PFCA) with CAYN-RN50 on the SVT, IIIT-STR, TTR, and PSTR datasets. The input images are padded to the same shape as stated in the main body of the submission.

## References

- [1] Chee Kheng Ch'ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition.

In 2017 14th IAPR international conference on document analysis and recognition (ICDAR), pages 935–942. IEEE, 2017. 1



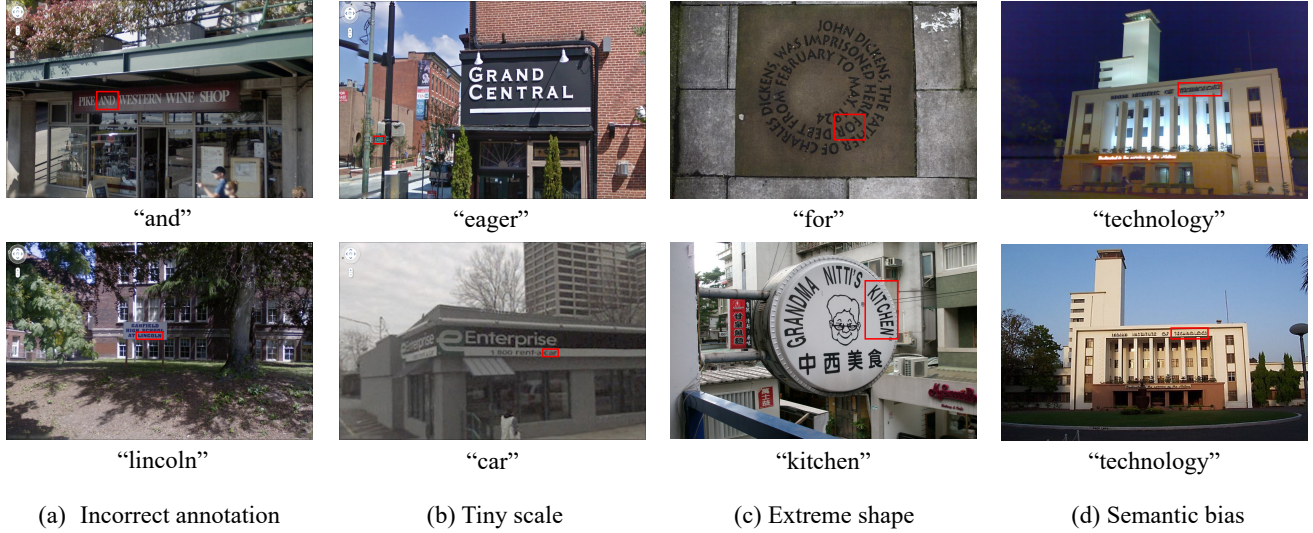


Figure 4. Failure case analysis. (a) Incorrect annotation; (b) Tiny scale; (c) Extreme shape; (d) Semantic bias. We highlight the position of the queried scene text with red boxes for better visualization.



Figure 5. Visualization of the attention maps in the parameter-free cross-attention (PFCA) with CAYN-RN50 on the STR dataset. (a) original input; (b) attention maps queried by “institute”; (c) attention maps queried by “technology”.

- [2] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 3
- [3] Lluís Gómez, Andrés Mafla, Marçal Rusinol, and Dimosthenis Karatzas. Single shot scene text retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pages 700–715, 2018. 4
- [4] 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
- [5] Anand Mishra, Kartek Alahari, and CV Jawahar. Image retrieval using textual cues. In *Proceedings of the IEEE inter-*

*national conference on computer vision*, pages 3040–3047, 2013. [1](#)

- [6] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, pages 1454–1459. IEEE, 2017. [1](#)
- [7] Zobeir Raisi and John Zelek. Occluded text detection and recognition in the wild. In *2022 19th Conference on Robots and Vision (CRV)*, pages 140–150. IEEE, 2022. [4](#)
- [8] Zobeir Raisi and John Zelek. Occluded text detection and recognition in the wild. In *2022 19th Conference on Robots and Vision (CRV)*, pages 140–150. IEEE, 2022. [1](#)
- [9] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8802–8812, 2021. [1](#)
- [10] 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](#)
- [11] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *Advances in Neural Information Processing Systems*, 35:12991–13005, 2022. [3](#)
- [12] Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, and Wenyu Liu. Scene text retrieval via joint text detection and similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4558–4567, 2021. [1](#), [2](#), [3](#), [4](#)
- [13] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *2011 International conference on computer vision*, pages 1457–1464. IEEE, 2011. [1](#)
- [14] Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. From two to one: A new scene text recognizer with visual language modeling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14194–14203, 2021. [1](#)
- [15] Lilong Wen, Yingrong Wang, Dongxiang Zhang, and Gang Chen. Visual matching is enough for scene text retrieval. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 447–455, 2023. [1](#), [2](#)
- [16] Gangyan Zeng, Yuan Zhang, Jin Wei, Dongbao Yang, Peng Zhang, Yiwen Gao, Xugong Qin, and Yu Zhou. Focus, distinguish, and prompt: Unleashing clip for efficient and flexible scene text retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2525–2534, 2024. [1](#), [3](#), [4](#)
- [17] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. [2](#)