Huazhong University of Science & Technology

MSTAR:Box-free Muti-query Scene Text Retrieval

Liang Yin Xudong Xie Zhang Li Xiang Bai Yuliang Liu*

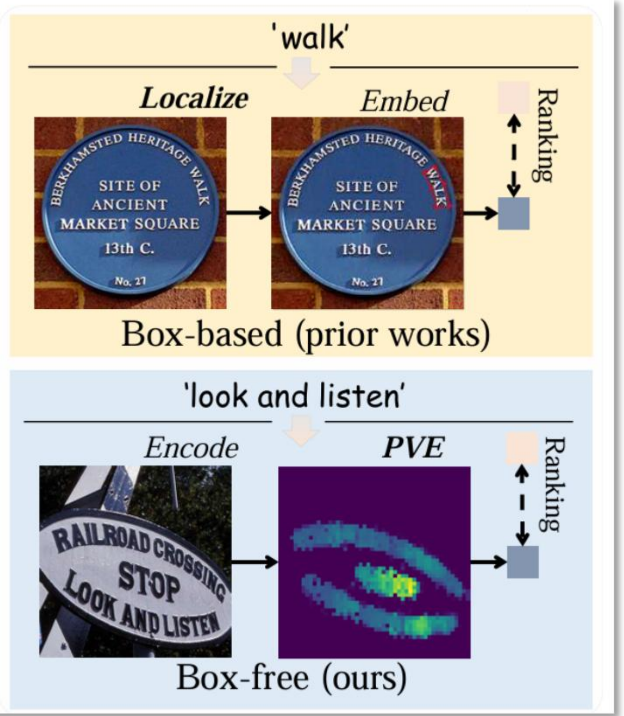
Huazhong University of Science and Technology

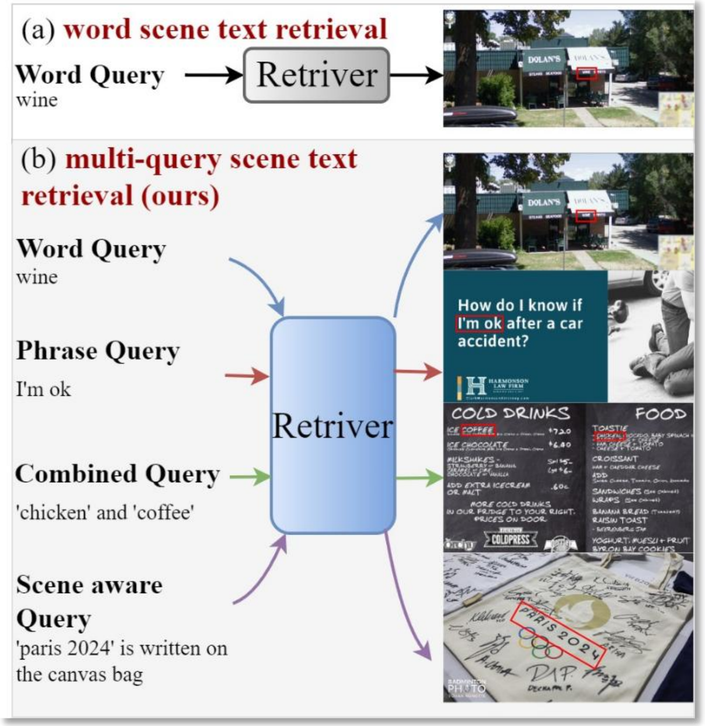
{liangyin, xdxie, zhangli, xbai, ylliu}@hust.edu.cn

Background

Scene text retrieval: given a text query and an image gallery, the aim is to search for images that writes the query from the gallery. Applications: visual document retrieval/key frame extraction.

Major contributions of this work:





Contribution1: box-free scene text retrieval that eliminates the heavy cost of spatial annotations.

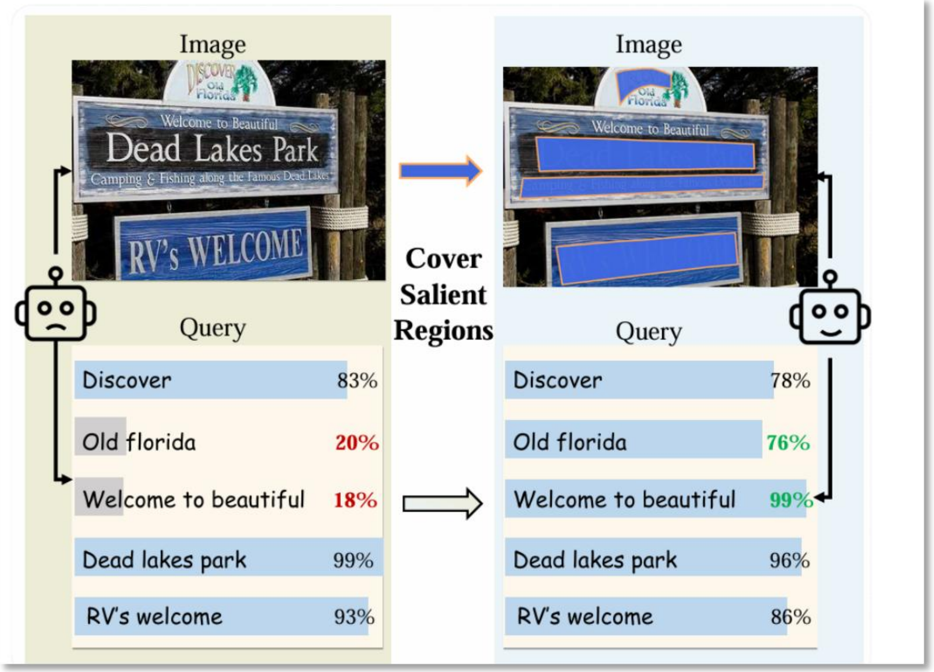
Contribution2: multi-query scene text retrieval that unifies queries of four styles.

Motivation

A possible solution to the questions before is vision-language models. We unveil the challenges of direct application of VLM on this task.

Model	Parameters	Pretraining Data	MAP%
CLIP-RN50	97M	400M images	6.6
CLIP-ViT-Base	143M	400M Images	6.8
CLIP-ViT-Large	408M	400M Images	8.1
BLIP-ViT-Large	426M	129M Images	6.9
BLIP2-ViT-Large	452M	129M images	13.3
SigLIP-ViT-Base-512	194M	9B Samples	12.8
SigLIP-ViT-Large-384	622M	9B Samples	11.7
MSTAR-ViT-Base	270M	-	60.13

Tab1. Performance on the CoCo-Text dataset.

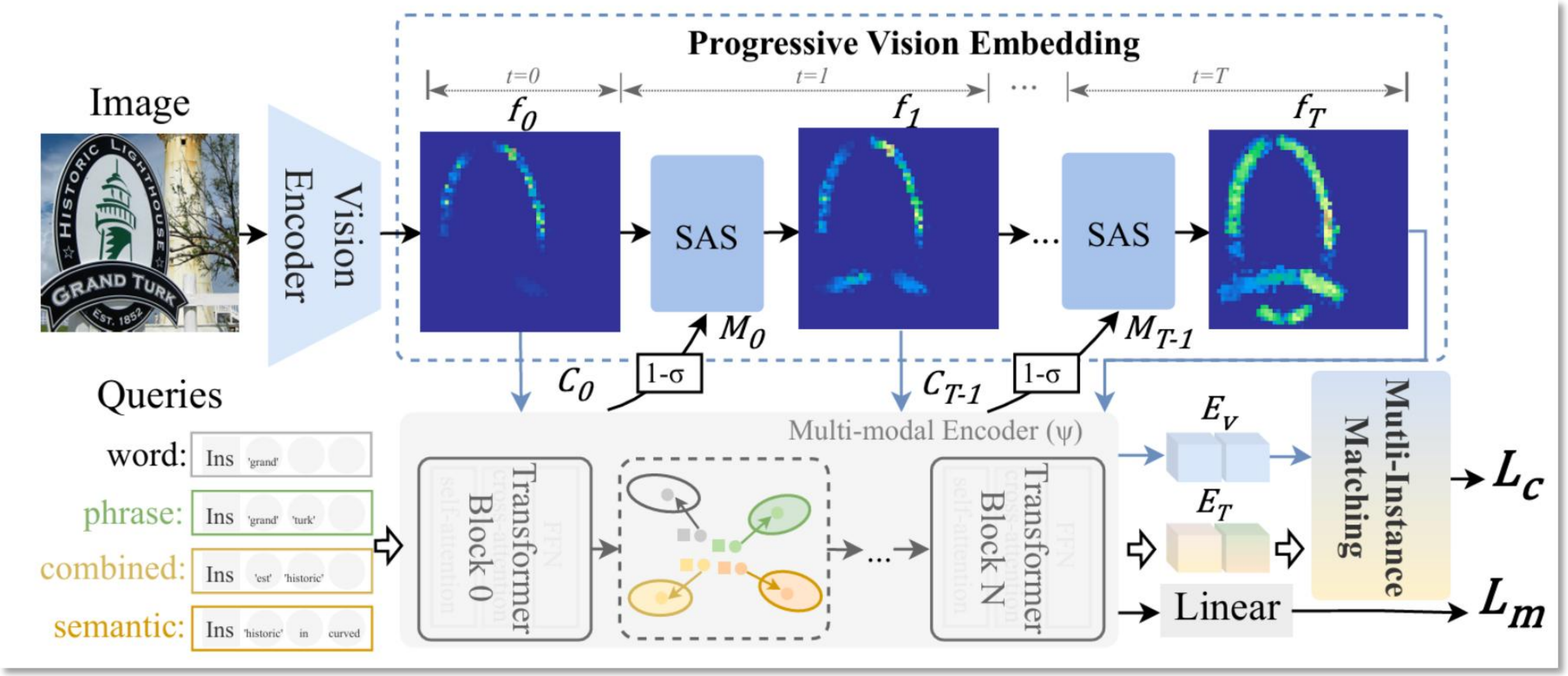


Observation1: Traditional CLIP-style models performs pool on the CoCo-Text dataset which features dense and small text instances.

Observation2: VLMs can transfer their vision attention to the smaller text after the larger text is masked.

Method

The MSTAR model consists of three major components: vision encoder, progressive vision embedding, multi-model encoder, multi-instance matching modules.

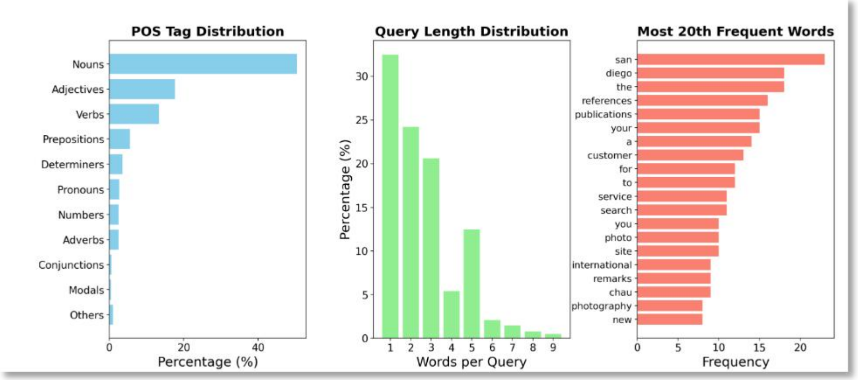


- Progressive Vision Embedding: shift the vision attention to the regions less attented to aviod the missing of small text instances.
- Multi-instance matching: explicitly assing the matching relations of vision embeddings (Ev) and text embeddings (ET).

MQTR dataset

dataset	Query Type				Query	Images
	word	phrase	combined	semantic		
SVT [42]	✓	✗	✗	✗	427	249
Total-text [42]	✓	✗	✗	✗	60	300
CTW [42]	✓	✗	✗	✗	100	500
ICDAR15 [42]	✓	✗	✗	✗	100	500
CTR [39]	✓	✗	✗	✗	500	7196
STR [11]	✓	✗	✗	✗	50	10000
CSVTR [40]	✗	✓	✗	✗	23	1667
PSTR [50]	✗	✓	✗	✗	36	1080
MQSTR	✓	✓	✓	✓	686	16000

Tab2. Comparisons of MQTR and prios works.



Statistics of MQTR

Samples from MQTR

We built a large-scale scene text retrieval datasets called MQTR. It includes 4 styles of queries (word, phrase, combined and semantic) and 16k images. This is the first benchmark for multi-query scene text retrieval evaluation.

Experiments: multi-query retrieval

Method	Venue	AVG.	Word	Phrase	Combined	Semantic
Box Based						
ABCNet [22]	TPAMI'21	24.13	26.14	15.15	36.47	18.74
MaskTextSpotter [20]	ECCV'20	32.43	46.72	27.53	29.08	26.37
TDSL [37]	CVPR'21	58.25	69.11	40.83	72.71	50.36
Deepsolo [46]	CVPR'23	52.04	67.54	25.68	72.14	42.79
TG-Bridge [11]	CVPR'24	54.09	69.89	30.21	**75.53**	40.73
Box Free						
SPTSv2 [23]	TPAMI'23	35.18	33.56	21.24	50.76	35.16
BLIP2 [19]	PMLR'23	36.13	17.31	32.76	25.80	68.63
SigLIP [49]	CVPR'23	36.06	17.81	32.88	21.81	72.23
BLIP2 (FT) [19]	PMLR'23	58.11	58.09	42.23	60.84	71.24
MSTAR	-	**66.78**	**73.27**	**44.22**	74.48	**75.14**

Tab 3. Results on the MQTR dataset. FT denotes fine-tuned.

BLIP2 [19]	TDSL [37]	SigLIP [49]	FDP [48]	MSTAR
85.49	89.40	89.56	92.28	**95.71**

Tab 4. Results on the PSTR dataset.

1.Box-based methods struggle with phrase/semantic retrieval tasks that require semantic understanding.

2.Box-free methods underperform in fine-grained perception tasks such as word and combined retrieval.

3.MSTAR demonstrates significant improvements over previous methods on MQTR.

- It outperforms the previous highest performance across all four subsets: word, phrase, combined, and semantic retrieval.
- Compared to the baseline BLIP2, it achieves an average performance improvement of 11.99%.
- On the public dataset PSTR, it reaches a MAP of 95.71%.

Experiments: word-level retrieval

Tab 5. Word retrieval performance surpasses previous fully-supervised SOTA retrieval models and achieves comparable results.

Method	Venue	SVT	STR	CTR	Total-Text	CTW	IC15	Avg.	FPS
Box Based									
Mishra *et al.* [27]	ICCV'13	42.70	56.24	-	-	-	-	-	0.1
Jaderberg *et al.* [12]	IJCV'16	86.30	66.50	-	-	-	-	-	0.3
Gomez *et al.* [9]	ECCV'18	83.74	69.83	41.05	-	-	-	-	43.5
Mafla *et al.* [26]	PR'21	85.74	71.67	-	-	-	-	-	42.2
TDSL [37]	CVPR'21	89.38	77.09	66.45	74.75	59.34	77.67	74.16	12.0
Wang *et al.* [38]	TPAMI'24	-	81.02	**72.95**	-	-	-	-	9.3
Wen *et al.* [40]	WSDM'23	90.95	77.40	-	80.09	-	-	-	11.0
FDP-RN50×16 [48]	ACM MM'24	89.63	**89.46**	-	79.18	-	-	-	11.8
Box Free									
BLIP2 (FT) [19]	PMLR'23	88.73	85.40	45.75	77.20	82.33	55.13	72.42	37.2
MSTAR	-	**91.31**	86.25	60.13	85.55	90.87	81.21	82.56	14.2
MSTAR (+rerank)	-	91.11	86.14	65.25	**86.96**	**92.95**	**82.69**	**84.18**	6.9

Tab 6. Word retrieval performance achieves comparable results with fully-supervised SOTA text-spotting models and

Method	Venue	SVT	STR	CTR	Total-Text	CTW	IC15	Avg.	FPS
Box Based									
ABCNet [22]	TPAMI'21	82.43	67.25	41.25	73.23	74.82	69.28	68.04	17.5
MaskTextspotterV3 [20]	ECCV'20	83.14	74.48	55.54	83.29	80.03	77.00	75.58	2.4
Deepsolo [46]	CVPR'23	87.15	76.58	67.22	83.19*	87.67*	82.80*	80.77	10.0
TG-Bridge [11]	CVPR'24	87.23	81.30	**70.08**	**87.11***	88.39*	**83.55***	82.94	6.7
Box Free									
SPTSv2 [23]	TPAMI'23	78.08	62.11	48.39	73.61*	83.30*	66.27*	68.63	7.6
MSTAR	-	**91.31**	**86.25**	60.13	85.55	90.87	81.21	82.56	14.2
MSTAR (+rerank)	-	91.11	86.14	65.25	86.96	**92.95**	82.69	**84.18**	6.9

Experiments: ablation studies

Ins	MIM	PVE	CTR	SVT	STR	Total-Text	CTW	IC15	MQTR
✗	✗	✗	52.87	90.07	81.57	82.32	87.28	76.71	65.79
✓	✗	✗	54.65	90.70	82.81	83.19	88.96	77.15	66.15
✓	✓	✗	55.77	91.02	85.00	84.01	90.31	79.23	65.69
✓	✓	✓	60.13	91.31	86.25	85.55	90.87	81.21	66.78

Tab7. Ablation Study — Instruction (Ins), Multi-Query Matching (MQM), Progressive Vision Embedding (PVE)

σ	CTR	Total-Text	IC15
No PVE	55.76	84.01	79.23
Zero Pad	56.67	83.79	79.27
TH+CC	59.66	85.17	80.17
TH+WS+CC	60.13	85.55	81.21

Tab 8. Ablation Study — different choice of binary algorithm.

T	CTR	Total-Text	CTW	FPS
0	55.76	84.01	90.31	16.5
1	60.13	85.55	90.87	14.2
2	60.47	86.68	90.95	12.9
3	60.87	87.66	91.24	11.2

Tab 9. Ablation Study — impact of recurrent steps T.

Visualization

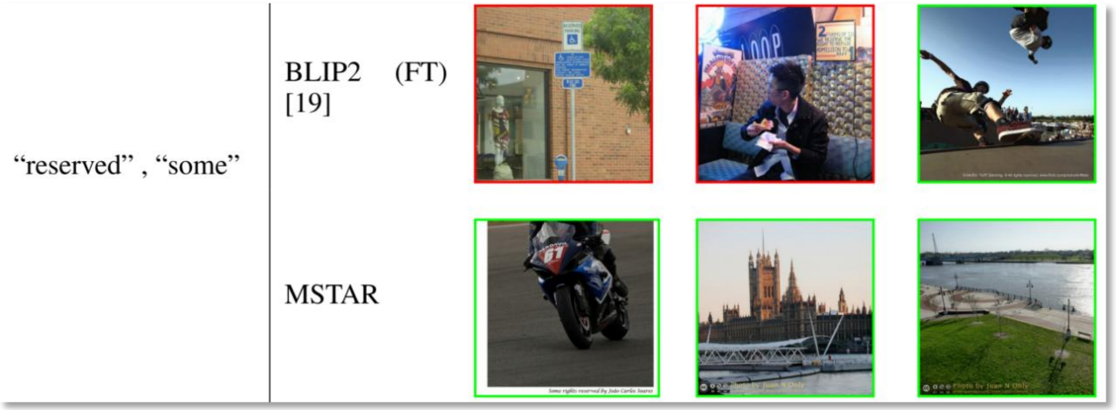
(a) 'dream big' tshirt (b) 'lighthouse' (c) '1888', 'celtic' (d) 'restaurant' (e) 'copyright'

Method	Mean IoU	High-Quality Masks (IoU $\geq$ 0.5)
BLIP-2	21.78	129 (25.8%)
MSTAR	**50.82**	**304 (60.8%)**

Table 12: Quantitative comparison of text region localization on the CTW dataset. MSTAR produces substantially more accurate and higher-quality text masks than BLIP-2.

Analysis of localization of text regions

Advantages on lingusitc semantics



Advantages on fine-grained features

Analysis of retrieval results

Thanks