

Focus, Distinguish, and Prompt: Unleashing CLIP for Efficient and Flexible Scene Text Retrieval (Supplementary Material)

Anonymous Authors

1 PSTR DATASET

To validate whether the STR models can be generalized to arbitrary-length query text, we introduce a new **Phrase-level Scene Text Retrieval (PSTR)** dataset. Specifically, we select 36 phrases that occur frequently in life as queries, each containing 2 to 4 words. All queries are listed in Tab.1.

Table 1: The list of queries in PSTR dataset.

Length	Queries
2	"apple store", "bud light", "coming soon", "low price", "no smoking", "one way", "school bus", "second edition", "upper canada", "brewing company", "fly emirates", "blue moon", "nutrition facts", "cabernet sauvignon", "fitting room", "ice cream", "joe boxer", "macbook air", "caps lock", "san francisco", "sony ericsson", "the original"
3	"bank of america", "do it yourself", "handle with care", "happy new year", "made in china", "india pale ale", "olive oil soap", "pedro benito urbina", "slide to unlock"
4	"in god we trust", "share a coke with", "have a nice day", "pink lady apple juice", "single malt scotch whisky"

2 ABLATION OF LOSS WEIGHTS

The hyperparameters λ_1 , λ_2 , and λ_3 are used to balance the loss items \mathcal{L}_{loc} , \mathcal{L}_{align} , and $\mathcal{L}_{distract}$ respectively in training FDP. Considering \mathcal{L}_{align} is the main loss for contrastively aligning the matched (Image, Query) pairs, we set $\lambda_2 = 1$ and conduct ablation studies of λ_1 and λ_3 on IIIT-STR dataset. The results are reported in Tab.2 and Tab.3. According to the ablation results, FDP reaches the best performance when $\lambda_1 = 1$ and $\lambda_3 = 1$.

Table 2: Ablation of the hyperparameter λ_1 when $\lambda_3 = 1$.

λ_1	0.5	1	3	5
mAP	81.33	81.77	80.98	80.46

Table 3: Ablation of the hyperparameter λ_3 when $\lambda_1 = 1$.

λ_3	0.5	1	3	5
mAP	81.60	81.77	81.73	80.20

3 MORE QUALITATIVE RESULTS

To further support our claim in the paper, we provide more qualitative examples retrieved by FDP-S in Fig.1. The proposed FDP method could deal with scene text in various scenarios, and has ability to recalling complicated cases such as multi-oriented and curve text. In particular, with the aid of our semantic-aware prompting technique, the retrieval accuracy on the function words (such as "the") is significantly strengthened compared to the original CLIP model. Nevertheless, for the challenging SVT and TotalText benchmarks, FDP still suffers from some limitations. On the one hand, when the query text to be retrieved is extremely tiny and meanwhile there are many disturbing words appearing in the image, the model has difficulty locating the target text. On the other hand, FDP still tends to return the images containing the similar words (e.g., "port" vs. "sport", "since" vs. "venice") from the image gallery, suggesting that the ability of fine-grained character discrimination needs to be further improved. We would like to go on with exploration for addressing these problems in the future.

Benchmark	Query	Retrieval Results				
IIIT-STR	<i>“free”</i>					
	<i>“department”</i>					
SVT	<i>“street”</i>					
	<i>“the”</i>					
TotalText	<i>“since”</i>					
	<i>“port”</i>					
PSTR	<i>“bank of america”</i>					
	<i>“have a nice day”</i>					

Figure 1: Visualization of the rank@1-5 retrieval results from FDP-S on IIIT-STR, SVT, TotalText and PSTR benchmarks. The correct results are highlighted in green while the incorrect ones are highlighted in red. Best viewed in zoom.