A Pedestrian is Worth One Prompt: Towards Language Guidance Person Re-Identification

Supplementary Material

In this supplementary material, we provide more analyses of our proposed PromptSG, which are difficult to elaborate in the main paper due to space limitations. A quick preview of the additional material is shown below:

We give more analyses that may be of interest to the reader, including 1) the impact of various multimodal fusion techniques, and 2) more qualitative results.

1. Ablation Study

Analysis of different multimodal fusion. We delve into the inherent benefits of our interaction module by comparing it with two different multimodal fusion modules. As illustrated in Fig. 1, we can simply consider the combination operation for multimodal fusion. In another widely-used fusion module referred to as merged attention, multi-head attention is applied to the concatenated image and language modalities. As observed in Tab. 1, our method outperforms the alternative methods across two datasets. Tab. 2 further ablate over the design choice for the cross-attention. We evaluate the variants that drop the CLS token, i.e., the global visual embedding $\tilde{v}$ in patch embeddings. The results show that such variants lead to a slight degradation in performance.

Method	Market-1501		MSMT17	
	mAP	R-1	mAP	R-1
Combination	87.5	94.2	68.9	86.6
Merged attention	92.9	95.1	75.2	89.5
Ours	94.6	97.0	87.2	92.6

Table 1. Comparisons between different multimodal fusion modules on Market-1501 and MSMT17.

Method	Market-1501		MSMT17	
	mAP	R-1	mAP	R-1
Cross-attn w/ CLS token	94.6	97.0	87.2	92.6
Cross-attn w/o CLS token	94.2	97.1	86.9	92.5

Table 2. Ablation of the design choice for the cross-attention on Market-1501 and MSMT17.

2. Qualitative Results

Top-8 retrieval results. Fig. 2 exhibits two examples of top-8 retrieval results, where the first row and the second row present the results from CLIP-ReID and PromptSG, respectively. It can be seen that both methods tend to match easy positives with similar lighting and angles. However, we notice a significant improvement in our method as it excels in subsequently recalling more challenging samples.

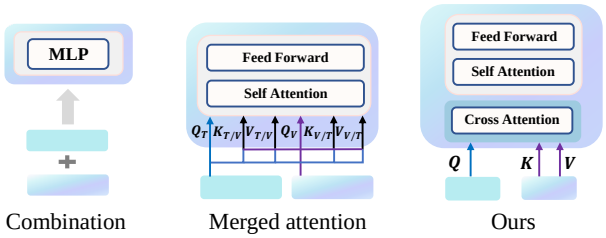

Figure 1. Illustration of three types of multimodal fusion modules.

Interpretation of the pseudo tokens. We endeavor to provide interpretation for the learned pseudo tokens by conducting a search in the vocabulary for words that are closest to the vectors using Euclidean distance. However, as noted in prior research [2, 3], since the vectors are optimized in a continuous space, it is difficult to be mapped into discrete meaningful codes of words. Therefore, our initial step involves assembling a curated list of attributes representing various characteristics of a person. These attributes, such as 'teenager', 'elderly', 'dress', 'pants', 'eyeglasses', 'bag', 'handbag', 'black', 'white', and 'brown', are thoughtfully selected from the Market-1501-attributes [1]. Then we evaluate the distance of the pseudo token from each of the words in the attribute set. The results are shown as word clouds for simplicity, with the size of each word corresponding to its distance from the pseudo token. As observed in Fig. 2, the model has learned to associate the term 'young' or 'elderly' with age-related attributes. Furthermore, the model has successfully identified some characteristics of clothing, such as 'handbag', 'dress', and 'pants'.

More visualization results. We present additional visualization results of attention maps for the ViT-based method TransReID and Vanilla CLIP. The results demonstrate that TransReID optimized without explicit semantic guidance can only focus on certain non-contiguous local regions. Furthermore, directly using the Vanilla CLIP visual model as a visual encoder fails to effectively capture the discriminative clues. This motivates us to explore an alternative language-guided approach, seeking to enhance the model's ability to extract informative information.