Knowledge-Enhanced Dual-stream Zero-shot Composed Image Retrieval

Supplementary Material

1. Experimental Details

In this section, we provide further details of the proposed KEDs in several aspects.

Network Architecture. The Bi-modality Knowledge-guided Projection network consists of a shared ψ_l and two identical cross-attention blocks. The detailed structures are illustrated in Figure 1. Note that the cross-attention block contains three cross-attention layers illustrated in the figure. The overall parameter of KEDS is 10.5M.

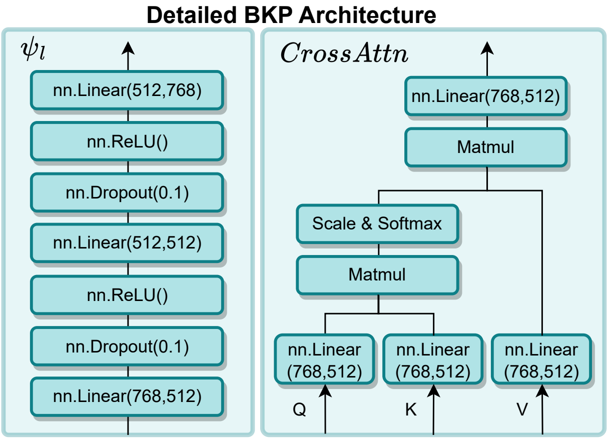

Figure 1. Visualization of the detailed BKP architecture. Left lane represent the linear block while the right lane illustrates the Cross-attention layer.

Efficiency Analysis. Efficiency is vital for real-world practice. To this end, we compare the inference speed with the previous method pic2word. The speed test is conducted using a single RTX 4090 on the CIRR validation set. Results show that the average inference time per batch (batch size of 64) is 1.04s using pic2word and 1.17s using KEDs respectively.

Subject Parsing Details. In the dual-stream training section, we introduce mining pseudo-triplets from the image-caption pairs. The subject phrases of the captions are substituted by the pseudo-word tokens. Specifically, we conduct dependency parsing for each caption, identify all noun phrases in the caption then select the one that contains the syntactic subject noun. Examples are shown in Figure 2.

2. Addtional Ablations

In this section, we provide the results of additional ablation studies.

Database Feature Extractor. In our implementation, we simply acquire the visual feature in the database through the default CLIP transformer backbone. To further explore the influence of the visual backbone, we conduct experiments using DINOv2 ViT-B/14 to encode visual features for the database. This experiment aims to test the robustness of KEDs toward different database feature backbones. Results show that KEDs is robust to different feature extractors. Using CLIP achieves higher recall. A potential reason could be that using the same backbone helps the KEDs to capture differences between the training image and retrieved images.

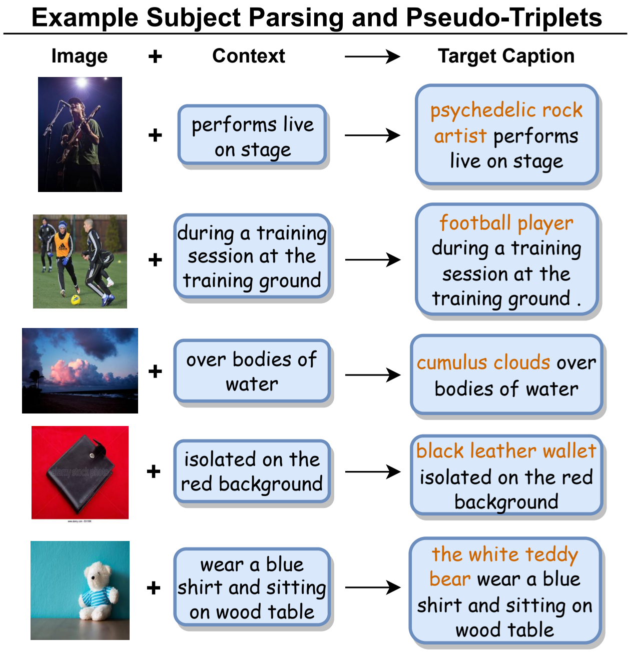
Example Subject Parsing and Pseudo-Triplets
Image
+
Context
Target Caption
performs live on stage
psychedelic rock artist performs live on stage
during a training session at the training ground
football player during a training session at the training ground .
over bodies of water
cumulus clouds over bodies of water
isolated on the red background
black leather wallet isolated on the red background
wear a blue shirt and sitting on wood table
the white teddy bear wear a blue shirt and sitting on wood table

Figure 2. Example subject parsing and pseudo-triplets. Note that text in orange represents the subject phrase.

Method	Database	R1	R5	R10
Pic2word	-	22.6	52.6	66.6
KEDs	**CLIP**	**27.3**	**56.4**	**69.2**
	DINO	26.7	55.5	68.5
	DINO†	26.9	56.2	68.8

Table 1. **Database Feature Extractor Ablation.** † indicates trained on CLIP-based database and inference on DINO-based database.

Test-time Robustness. In the Ablation study in the original manuscript, we conduct training time experiments to test whether KEDs is robust in different settings. In this section, we visualize the test-time ablation results on the CIRR validation set in Figure 3 (a) and (b). We find that KEDs

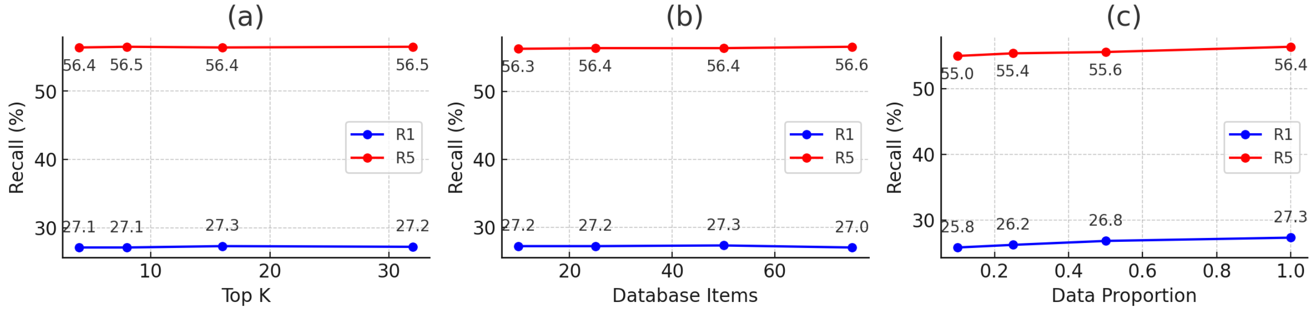

Figure 3. **Training Samples and Test-time Ablation.** Visualization of how (a) the item amount of the database during inference, (b) the number of retrieved neighbors during inference, (c) the proportion of training samples influence the performance respectively on the CIRR validation set.

is flexible with the number of items in the database and retrieved neighbors.

Training Sample requirements. To test whether KEDs is data-hungry, we train ϕ_M with different proportions of the training set. Results are shown in Figure 3 (C). KEDS is also effective with limited training data.