

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

Yucheng Suo Fan Ma Linchao Zhu[†] Yi Yang
 ReLER, CCAI, Zhejiang University, China

[†] Corresponding author

Abstract

We study the zero-shot Composed Image Retrieval (ZS-CIR) task, which is to retrieve the target image given a reference image and a description without training on the triplet datasets. Previous works generate pseudo-word tokens by projecting the reference image features to the text embedding space. However, they focus on the global visual representation, ignoring the representation of detailed attributes, e.g., color, object number and layout. To address this challenge, we propose a Knowledge-Enhanced Dual-stream zero-shot composed image retrieval framework (KEDs). KEDs implicitly models the attributes of the reference images by incorporating a database. The database enriches the pseudo-word tokens by providing relevant images and captions, emphasizing shared attribute information in various aspects. In this way, KEDs recognizes the reference image from diverse perspectives. Moreover, KEDs adopts an extra stream that aligns pseudo-word tokens with textual concepts, leveraging pseudo-triplets mined from image-text pairs. The pseudo-word tokens generated in this stream are explicitly aligned with fine-grained semantics in the text embedding space. Extensive experiments on widely used benchmarks, i.e. ImageNet-R, COCO object, Fashion-IQ and CIRR, show that KEDs outperforms previous zero-shot composed image retrieval methods. Code is available at <https://github.com/suoych/KEDs>.

1. Introduction

Composed Image Retrieval (CIR) is a task first introduced by Vo *et al.* [71], which involves retrieving the target image given a reference image and a modification description. Different from traditional image-based retrieval [10, 19, 59, 64] or text-based retrieval [63, 72] tasks, composed image retrieval requires the model to interpret both visual and text modality information. In practical scenarios, CIR allows users to specify fine-grained styles and content details in the queries, enabling flexibility and customization.

With the rise of image-text pre-training models like CLIP [61], significant improvements are achieved in the com-

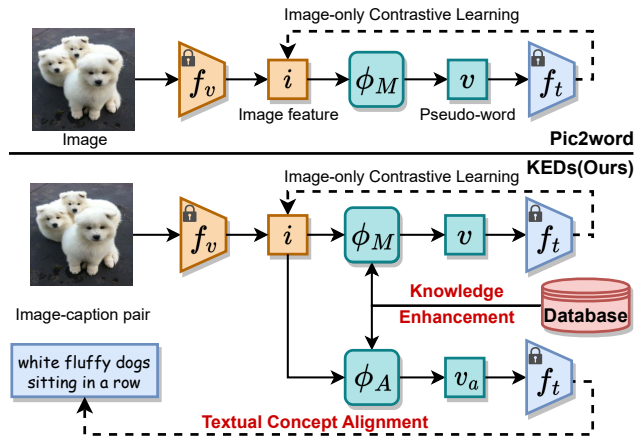


Figure 1. **Comparison between existing methods and KEDs.** Pic2word [65] learns the mapping network using image-only contrastive learning and generates pseudo word ϕ_M token v . We propose to augment the pseudo-word token with external knowledge. In addition, we introduce an extra branch ϕ_A sharing architecture with ϕ_M for textual concept alignment. Note that f_v and f_t indicate frozen CLIP visual encoder and text encoder respectively.

posed image retrieval field [2, 11, 50]. Previous supervised methods work on designing networks to generate compositional features based on the reference image and text. Training and evaluation are conducted on the triplet datasets in various domains, e.g. Fashion-IQ [73] on the fashion clothes domain, CIRR [50] on real-world images. Despite the impressive performance achieved by these supervised methods, two major limitations remain. Firstly, these approaches require extensive triplet data annotations, which are labor-intensive and time-consuming. Secondly, supervised methods train a tailored model for each dataset, thereby reducing the flexibility and generalization ability.

Recent studies [3, 65] introduce zero-shot approaches for composed image retrieval to address the above limitations. Pic2word is the pioneering work that learns a mapping layer that projects the image features into the text embedding space. This is achieved by self-supervised learning using a contrastive loss between the mapped features and original image features as the training objective. A frozen CLIP text encoder then generates the hybrid feature for inference.

This image-only training paradigm learns a single model capable of datasets in various domains without triplet data. SEARLE [3] is another work that uses a large language model to generate additional descriptions based on the object class, enhancing the alignment between mapped image features and class semantics. However, these approaches share a common limitation: they directly map the overall image feature to the text embedding space, overlooking the detailed attribute information. This reduces the retrieval accuracy since the text in composed retrieval triplets only describes the object differences between target images and reference images.

To address the aforementioned issues, we propose a Knowledge-Enhanced Dual Stream zero-shot composed image retrieval framework (KEDs). First of all, KEDs incorporates a Bi-modality Knowledge-guided Projection network (BKP) that generates pseudo-word tokens based on external knowledge. Specifically, BKP incorporates a database to provide relevant images and captions to generate comprehensive pseudo-word tokens. The advantage is that the retrieved captions and images enrich the pseudo-word tokens with shared attribute information, *e.g.*, object number and layout. This approach is akin to an “open-book exam”, where the database serves as a reference sheet to better identify the reference image. We simply construct the database by random sampling a portion of the image-text pairs from the training dataset.

Image-only contrastive training does not align the pseudo-word tokens with real text concepts, bringing difficulty in composing reference images with text during inference. Therefore, we introduce an extra stream to generate pseudo-word tokens aligned with textual semantics. This stream is trained on pseudo-triplets mined from image-text pairs. This approach facilitates the interaction between pseudo-word tokens and diverse text during inference while maintaining the object semantics. Note that the pseudo-triplet mining process does not require generating extra data by external models. During inference, KEDs combines the output of the two streams, allowing for controlled alignment with specific modalities.

To evaluate the effectiveness of KEDs, we conduct experiments across four datasets, *i.e.* ImageNet-R [12, 24], COCO [46], Fashion-IQ [73], CIRR [50]. The four datasets test KEDs on different aspects of compositional ability, showcasing the generalization ability. The result indicates that KEDs surpasses all previous methods, especially in the ImageNet-R domain conversion task, achieving a remarkable boost of 7.9% in Recall@10 and 12.2% in Recall@50 on average. Additionally, ablation studies are also provided to inspect the details of the method design. Overall, our contributions can be concluded as follows:

- We propose a Knowledge-Enhanced Dual Stream framework (KEDs) for zero-shot composed image retrieval,

where an external database enriches the mapping network with knowledge, enhancing the generalization ability.

- A new textual concept alignment training paradigm utilizing pseudo-triplets mined from the image-text pairs, ensuring semantic alignment between the mapped visual features and rich semantics.
- Extensive experiments on four datasets demonstrate the effectiveness of the proposed framework.

2. Related Work

2.1. Composed Image Retrieval

Composed Image Retrieval (CIR) is a compositional task first introduced by Vo *et al.* [71], which aims to retrieve a target image given a reference image and a modification description [78]. A critical aspect of the task is the extraction and combination of information from both reference images and text [6, 74]. Current supervised methods train a cross-modal network using a fusion paradigm, which learns a joint embedding combining image and text. Representative methods include CoSMo [37], DCNet [36], ARTEMIS [11], CLIP4Cir [2], *etc.* These supervised methods are trained on various labeled triplet data benchmarks including Fashion-IQ [73], CIRR [50], *etc.*

Considering the labor-intensive process of obtaining triplet data, Saito *et al.* [65] introduce a new setting to train a network without triplet data under a zero-shot setting. Their pioneer work Pic2word learns a mapping network that projects the image feature to the text embedding space, achieving impressive performance over datasets like CIRR and Fashion-IQ. Baldrati *et al.* design a text inversion network distilling the mapping network for better alignment with nouns [3]. Another line of work focuses on generating extra triplet data using generative models [21, 49]. In this paper, we propose a method trained on image-text pairs since image-text data pairs are easy to acquire and contain pair-matching prior information.

2.2. Knowledge Enhanced Methods

Knowledge-enhanced methods are widely used in the natural language processing community [22, 38]. The fundamental concept is to improve the performance by incorporating external knowledge [75]. In essence, knowledge-enhanced methods require the model to generate predictions considering prior knowledge [30]. The advantage lies in their ability to inject interpretable world knowledge, particularly beneficial for knowledge-intensive tasks like open-domain question answering. With the rise of Large Language Models, knowledge-enhanced methods are used for comprehending long context and improving generation quality [25, 31, 47, 60, 62].

Having demonstrated success in NLP, the effectiveness of knowledge-enhanced methods also shows in other fields

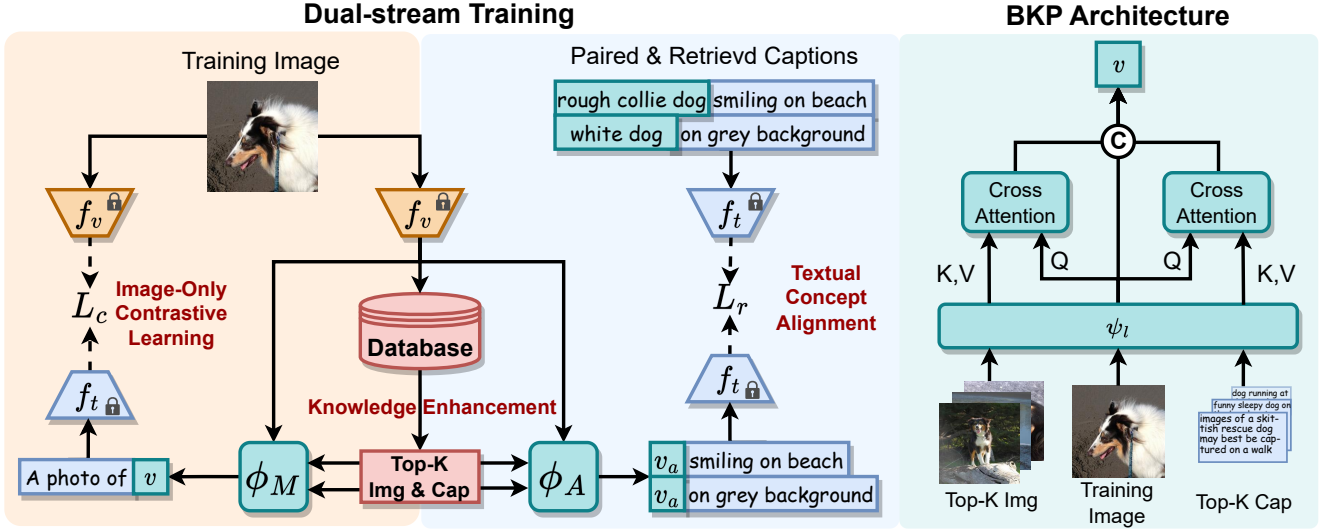


Figure 2. **Overall framework of KEDs.** The left part of the figure represents the dual-stream training of KEDs, consisting of the image-only contrastive training (orange) and textual concept alignment branch (blue). The right part represents the architecture of the proposed Bi-modality Knowledge-guided projection.

[4, 5, 20]. Recent works related to zero-shot recognition [7, 35, 42, 51, 58, 67]. For instance, K-LITE [67] trains a vision-text model with expanded entities via retrieving words from Wordnet [56] or Wikitionary[55]. RA-CLIP [75] utilizes an extra feature database to enhance cross-model knowledge, achieving the inference process in an “open-book exam” style since the feature database can be considered as a “cheating sheet”. RECO [30] explores the retrieving modalities and fusion methods in a similar setting. In this paper, we retrieve relevant images and captions to enrich the pseudo-word token semantics.

2.3. Vision-language Pretraining

Vision-language pre-training has been a long-standing research topic with wide real-world applications [13, 18, 27, 33, 53, 54, 81]. CLIP [61] is a representative work using image-text pairs to train the visual encoder and text encoder in a contrastive manner, achieving remarkable zero-shot performance on downstream tasks, including classification and image-text retrieval. CLIP lays the foundation of later vision-text pertaining models such as CoCa [77], BLIP [40, 41], ALBEF [39], Flamingo [1], etc.

Researchers have explored the strong zero-shot ability of CLIP in open-domain tasks like open-set recognition [9, 16, 23, 43, 80], open-vocabulary detection [15, 29, 48, 57, 68, 79] and segmentation [28, 32, 45, 76]. Cohen *et al.* propose PALAVRA [8], a learning paradigm for personalized concepts. PALAVRA learns mapping networks to project image embeddings to the text embedding space or project text embeddings to the image embedding space. This self-supervised method also appears in recent zero-shot papers [3, 17, 44, 65, 70]. In this paper, our goal is also to learn a mapping network but with an extra objective.

3. Method

In this section, we describe the proposed method in detail. The overall pipeline of KEDs is illustrated in Figure 2. We first introduce the preliminaries in Section 3.1. Then in Section 3.2, we introduce the Bi-modality knowledge-enhanced projection. The dual-stream alignment training paradigm is discussed in section 3.3. Additionally, We introduce the inference process of KEDs in Section 3.4.

3.1. Preliminaries

In this work, all retrieval processes are accomplished via Contrastive Language Image Pre-training (CLIP). CLIP consists of a visual encoder f_v and a text encoder f_t trained on image-text pairs with a contrastive objective. Specifically, given an image I and corresponding caption T , the visual encoder extracts visual feature $i = f_v(I) \in \mathbb{R}^d$, and the text encoder extracts the overall caption feature $t = f_t(T) \in \mathbb{R}^d$. To align the image-text pair in a contrastive manner, CLIP calculates a symmetric cross-entropy loss \mathcal{L}_c [69] which can be formulated as:

$$\mathcal{L}_{i2t} = -\frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \log \frac{\exp(\tau i_n^T t_n)}{\sum_{m \in \mathcal{B}} \exp(\tau i_n^T t_m)}, \quad (1)$$

$$\mathcal{L}_{t2i} = -\frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \log \frac{\exp(\tau t_n^T i_n)}{\sum_{m \in \mathcal{B}} \exp(\tau t_n^T i_m)}, \quad (2)$$

$$\mathcal{L}_c = \mathcal{L}_{i2t} + \mathcal{L}_{t2i}. \quad (3)$$

Note that the image features and the text features are normalized before calculating the contrastive loss.

3.2. Bi-modality Knowledge-guided Projection

To train a model for zero-shot composed image retrieval without triplet datasets, we follow previous work [65] to learn a mapping network that projects the image feature into the text embedding space, forming a pseudo-word token. As shown in Figure 2, we prompt the pseudo-word token v with the text "A photo of" and encode with a frozen CLIP model, the generated feature is used for calculating the contrastive loss \mathcal{L}_c with the image feature i . However, this image-only training process focuses on the global image representation, neglecting detailed attribute information.

To overcome this issue, we propose a Bi-modality Knowledge-guided Projection network (BKP) ϕ_M to generate the pseudo-word token v extracting information from relevant images and captions.

Top-K images and captions retrieval. As illustrated in Figure 2, we provide bi-modality knowledge for each training image by retrieving Top-K image features $\{i_k^r\}_{k=1}^K$ and caption features $\{t_k^r\}_{k=1}^K$ from a database. The Top-K image and caption features provide context for the projection, augmenting the vanilla image feature mapping network with shared attribute information. The database is simply constructed by random sampling 0.5M image-caption pairs from the training set and encoded by a pre-trained CLIP model. BKP retrieves items from the database using the faiss library [34], ensuring training efficiency. For a thorough context comprehension, we set K to 16.

Bi-modality context Fusion. After obtaining bi-modality knowledge from the database, we fuse the reference image feature with the knowledge with a simple network. The training image feature i and the retrieved features $\{i_k^r\}_{k=1}^K$ and $\{t_k^r\}_{k=1}^K$ are first projected into a common feature space through a linear block ψ_l to bridge the modality gap. Then the training image feature i is used as the query to interact with the retrieved image features and caption features via two cross-attention blocks respectively to facilitate comprehension of context. The final output $v \in \mathbb{R}^{3 \times d}$ is the concatenation of three tokens, *i.e.*, a mapped image feature token $\hat{i} = \psi_l(i)$ and two context-aware mapped tokens $v_i \in \mathbb{R}^d, v_c \in \mathbb{R}^d$. The procedure can be formulated as:

$$v_i = \text{CrossAttn}(\hat{i}, \psi_l(\{i_k^r\}_{k=1}^K), \psi_l(\{i_k^r\}_{k=1}^K)), \quad (4)$$

$$v_c = \text{CrossAttn}(\hat{i}, \psi_l(\{t_k^r\}_{k=1}^K), \psi_l(\{t_k^r\}_{k=1}^K)), \quad (5)$$

$$v = \text{concat}(\hat{i}, v_i, v_c). \quad (6)$$

BKP encodes the bi-modality context together with the reference images, thereby generating pseudo-word tokens with comprehensive attribute information.

3.3. Dual-stream Semantic Alignment

Previous work [65] trains the mapping network through image-only contrastive training as illustrated in Figure 2.

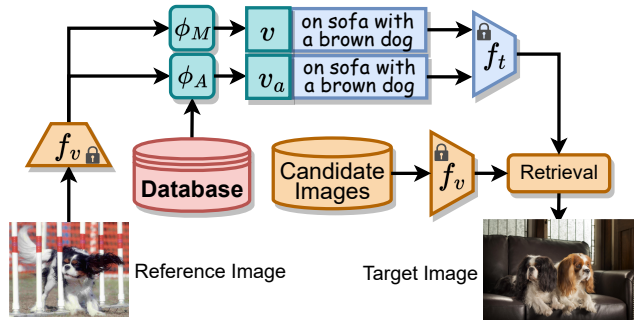


Figure 3. A simple illustration of the inference process of KEDs.

A limitation is that the pseudo-word tokens are not aligned with the textual concepts in the text embedding space, introducing challenges in composing images with text during inference. To this end, we introduce training an extra stream ϕ_A on Pseudo-triplets mined from image-caption pairs to generate pseudo word token v_a aligned with semantics.

Pseudo-Triplets Mining. The salient object in an image is described by the subject noun phrase of the paired caption syntactically. Our intuition is that the pseudo-word token v_a should align with the corresponding subject phrase semantic. Therefore, we extract pseudo-triplets consisting of an image, a piece of context description, and a target caption. To mine a pseudo-triplet, we conduct dependency parsing on a caption T using spacy [26] and replace the subject noun phrase with the pseudo-word token v_a .

Complementary Pseudo-Triplets. Each caption describes the object from a perspective, combining diverse captions stimulates comprehension of the image from different angles. For example, a caption for a photo showing a white husky sitting on a couch can be "Husky on the sofa", while a supplementary caption could be "A sleeping white dog", the two captions together comprehensively describe the different attributes, *e.g.* dog breed, color, action. Therefore, we retrieve two additional captions T_s and T'_s for each caption, generating complementary triplets to enrich semantics.

Semantic registration loss. To train an extra projection ϕ_A on the pseudo-triplets, we introduce a semantic registration loss. The pseudo-word token v_a is injected into the tokenized embedding of the prompt T_o and encoded by a frozen CLIP text encoder f_t . The output feature \hat{v}_a is used for calculating a cosine embedding loss \mathcal{L}_{cos} with the overall caption embedding $t = f_t(T)$:

$$\mathcal{L}_{cos} = 1 - \cos(\hat{v}_a, t). \quad (7)$$

For the complementary triplets, we calculate an averaged supplementary cosine embedding loss \mathcal{L}_{sup} with the retrieved captions embedding $t_s = f_t(T_s)$ and $t'_s = f_t(T'_s)$. The overall semantic registration loss \mathcal{L}_r is calculated by

Supervision	Methods	Cartoon		Origami		Toy		Sculpture		Average	
		R10	R50	R10	R50	R10	R50	R10	R50	R10	R50
ZERO-SHOT	Image-only	0.3	4.5	0.2	1.8	0.6	5.7	0.3	4.0	0.4	4.0
	Text-only	0.2	1.1	0.8	3.7	0.8	2.4	0.4	2.0	0.5	2.3
	Image+Text	2.2	13.3	2.0	10.3	1.2	9.7	1.6	11.6	1.7	11.2
	Pic2word	8.0	21.9	13.5	25.6	8.7	21.6	10.0	23.8	10.1	23.2
	KEDs	14.8	34.2	23.5	34.8	16.5	36.3	17.4	36.4	18.0	35.4
CIRR	Combiner [2]	6.1	14.8	10.5	21.3	7.0	17.7	8.5	20.4	8.0	18.5
Fashion-IQ	Combiner [2]	6.0	16.9	7.6	20.2	2.7	10.9	8.0	21.6	6.0	17.4

Table 1. **Results of the domain conversion experiment using ImageNet-R.** Our method achieves state-of-the-art result.

Supervision	Methods	R1	R5	R10
ZERO-SHOT	Image-only	8.6	15.4	18.9
	Text-only	6.1	15.7	23.5
	Image+Text	10.2	20.2	26.6
	Pic2word	11.5	24.8	33.4
	KEDs	12.0	26.0	34.9
CIRR	Combiner [2]	9.9	22.8	32.2
Fashion-IQ	Combiner [2]	13.2	27.1	35.2

Table 2. **Evaluation on COCO object composition task.**

linearly combining \mathcal{L}_{cos} and \mathcal{L}_{sup} :

$$\mathcal{L}_{sup} = 1 - \frac{1}{2}(\cos(\hat{v}_a, t_s) + \cos(\hat{v}_a, t'_s)), \quad (8)$$

$$\mathcal{L}_r = \mathcal{L}_{cos} + \beta \times \mathcal{L}_{sup}. \quad (9)$$

The semantic alignment stream explicitly matches the pseudo-word token v_a with the textual concepts in the text embedding space. The pseudo-triplet mining process only requires image-text pairs, which is not labor-intensive. Note that the ϕ_A uses the identical Bi-modality Knowledge-guided Projection architecture with ϕ_M .

3.4. Hybrid Inference

During inference, KEDs generates hybrid features of image and text for retrieval. Specifically, as shown in Figure 3, KEDs generate the pseudo-word tokens v and v_a given the reference image, then replace the placeholder token in the tokenized text embedding. In this way, we generate the composed feature using a frozen CLIP text encoder. As introduced in section 3.2 and 3.3, we use a dual-stream training projection network. During inference, the two streams generate two different composed features \hat{v} and \hat{v}_a through the text encoder, we conduct a simple yet effective linear combination to generate a hybrid feature v_h for retrieval:

$$v_h = \alpha \times \hat{v} + (1 - \alpha) \times \hat{v}_a. \quad (10)$$

v_h is used for calculating the similarity with the candidate image features. The image with the highest similarity is selected as the prediction. Under this zero-shot inference process, KEDs is capable of various compositional datasets.

4. Experiments

In this section, we first introduce the benchmarks and experiment setup in section 4.1. Then provide the implementation details and results on the different datasets in section 4.2 and 4.3. We also provide a detailed ablation study and analysis of the experiment results in section 4.4.

4.1. Datasets and Setup

KEDs is trained on the Conceptual Caption Three Million (CC3M) dataset [66]. CC3M contains a wide variety of image-caption pairs and has no overlap with the evaluation datasets. For evaluation, we employ four datasets following previous work, *i.e.* ImageNet-R [12, 24], COCO [46], Fashion-IQ [73], CIRR [50]. The four datasets assess four types of composition ability individually:

Domain conversion: ImageNet-R [12, 24] is used in the domain conversion setup. Specifically, 16983 real images under 200 classes are required to be converted to four styles, *i.e.* cartoon, origami, toy and sculpture. The correct target image should belong to the same class as the reference image while matching the domain description.

Object composition: COCO validation set [46] with 5000 images is employed for evaluating object composition. The reference images are constructed by cropping an object in the image according to the instance mask annotations and the goal is to retrieve the overall image.

Scene manipulation: In terms of the scene manipulation setup, we use the CIRR dataset [50] consisting of crowd-sourced, open-domain images with hand-written descriptions. Following previous works [3, 65], we report the performance on the test split, while conducting ablation studies on the validation set.

Fashion attribute manipulation: Another widely used benchmark is Fashion-IQ [73] designed for fashion images.

Supervision	Methods	Dress		Shirt		TopTee		Average	
		R10	R50	R10	R50	R10	R50	R10	R50
ZERO-SHOT	Image-only	5.4	13.9	9.9	20.8	8.3	17.7	7.9	17.5
	Text-only	13.6	29.7	18.9	31.8	19.3	37.0	17.3	32.9
	Image+Text	16.3	33.6	21.0	34.5	22.2	39.0	19.8	35.7
	Pic2word	20.0	40.2	26.2	43.6	27.9	47.4	24.7	43.7
	SEARLE-XL	20.3	43.2	27.4	45.7	29.3	50.2	25.7	46.3
	KEDs	21.7	43.8	28.9	48.0	29.9	51.9	26.8	47.9
CIRR	Combiner [2]	17.2	37.9	23.7	39.4	24.1	43.9	21.7	40.4
Fashion-IQ	Combiner [2]	30.3	54.5	37.2	55.8	39.2	61.3	35.6	57.2
Fashion-IQ	Combiner* [2]	31.6	56.7	36.4	58.0	38.2	62.4	35.4	59.0
Fashion-IQ	CIRPLANT [50]	17.5	40.4	17.5	38.8	21.6	45.4	18.9	41.5
Fashion-IQ	ALTEMIS [11]	27.2	52.4	21.8	43.6	29.2	54.8	26.1	50.3
Fashion-IQ	MAAF [14]	23.8	48.6	21.3	44.2	27.9	53.6	24.3	48.8

Table 3. Results on Fashion-IQ validation set. * is the result using ResNet50x4 backbone.

Supervision	Methods	R1	R5	R10	R50
ZERO-SHOT	Image-only	7.4	23.6	34.0	57.4
	Text-only	20.9	44.8	56.7	79.1
	Image+Text	12.4	36.2	49.1	78.2
	Pic2word	23.9	51.7	65.3	87.8
	SEARLE-XL	24.2	52.4	66.3	88.6
	KEDs	26.4	54.8	67.2	89.2
CIRR	Combiner [2]	30.3	60.4	73.2	92.6
Fashion-IQ	Combiner [2]	20.1	47.7	61.6	85.9
CIRR	Combiner* [2]	33.6	65.4	77.4	95.2
CIRR	TIRG [71]	14.6	48.4	64.1	90.0
CIRR	ARTEMIS [11]	17.0	46.1	61.3	87.7
CIRR	CIRPLANT [50]	19.6	52.6	68.4	92.4

Table 4. Evaluation on CIRR test set. * is the result using ResNet50x4 backbone.

The reference images indicate the type of clothes while the text describes the expected attributes. We report the results on the validation set.

In section 4.3, we quantitatively compare KEDs with the following baseline methods:

Image-only uses the similarity between the target image feature and reference image feature for retrieval.

Text-only retrieves the target image using the text features.

Image+text takes the average of the visual and text features to retrieve the target image.

Pic2word [65] is a method that learns a mapping network to project the image feature to the text embedding space. A frozen CLIP [61] text encoder fuses the mapped visual feature and text feature to retrieve the target image.

SEARLE-XL [3] generates descriptions using a large language model based on the class name and uses the descriptions to train a textual inversion network. The pseudo-word tokens are distilled from the inversion network.

Supervised methods [2, 11, 14, 50] are trained on the la-



Figure 4. Qualitative results on Fashion-IQ dataset. Images with green borders represent the ground truth.

beled triplet datasets including CIRR and Fashion-IQ. The performance is reported following previous work.

4.2. Implementation Details

We adopt the ViT-L/14 CLIP[61] backbone and use CC3M as the training dataset. The database is constructed by randomly selecting 0.5M image-text pairs and encoding the image and text by the pre-trained CLIP. We employ the GPU version faiss library [34] for efficient real-time retrieval over the database. The Bi-modality Knowledge-guided projection module contains three layers of multi-head cross-attention with a hidden dim of 768. We use AdamW [52] optimizer with a learning rate of $5e^{-5}$ and 0.1 weight decay. We conduct a linear warmup of 10000 steps and cosine learning rate decay to smooth the optimization. The model is trained for 30 epochs with a batch size of 512 on 8 RTX 4090 GPUs within one day.

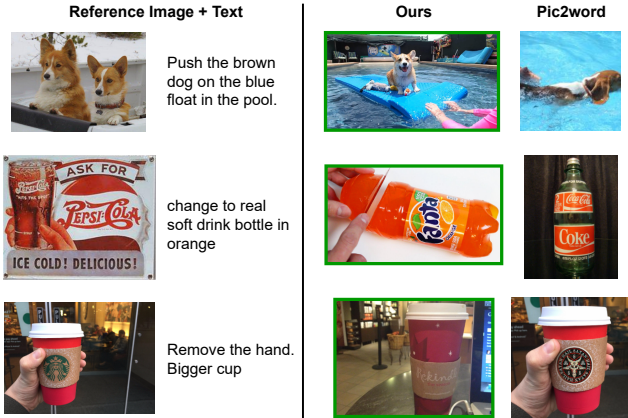


Figure 5. Qualitative results on CIRR dataset. Images with green borders represent the ground truth.

Part	Method	CIRR			Fashion-IQ	
		R1	R5	R10	R10	R50
All	Pic2word	22.6	52.6	66.6	24.7	43.7
	KEDs	27.3	56.4	69.2	26.8	47.9
ϕ_M	w/o Top-K img	26.5	54.5	67.0	24.0	43.6
	w/o Top-K cap	26.3	54.5	66.7	22.5	42.2
	w/o linear	21.7	49.0	60.9	15.9	31.6
ϕ_A	w/o ϕ_A	24.0	53.2	66.9	24.5	44.4
	w/o context	25.0	54.2	67.8	25.0	44.8
	w/o extra	27.2	56.0	68.5	26.0	46.3
DB	CIRR	27.1	56.3	68.9	25.8	46.4
	Fashion-IQ	26.7	55.7	68.5	26.0	46.4

Table 5. Ablation studies on CIRR and FashionIQ validation sets. For FashionIQ, we consider the average recall. ϕ_M is the Bimodality Knowledge-guided projection module, ϕ_A is the textual concept alignment branch, while DB represents the Database.

4.3. Quantitative and Qualitative Results

Table 1 shows the performance of the ImageNet-R dataset, KEDs consistently outperforms all previous methods with a substantial margin, *i.e.*, +7.9% Recall@10 and +12.2% Recall@50 on average. The boost is attributed to the alignment between pseudo-word tokens and object semantics.

On the COCO object composition task, KEDs also outperforms previous methods (+1.5% Recall@10) as reported in Table 2. Notably, KEDs implicitly captures fine-grained object information, leveraging multi-modal knowledge from the database to recognize objects and deduce scene layouts from neighboring images.

In terms of the fashion attribute manipulation task reported in Table 3, KEDs shows impressive results. For instance, there is a 1.5% improvement on Recall@10 and 2.3% on Recall@50 in the shirt domain compared with the

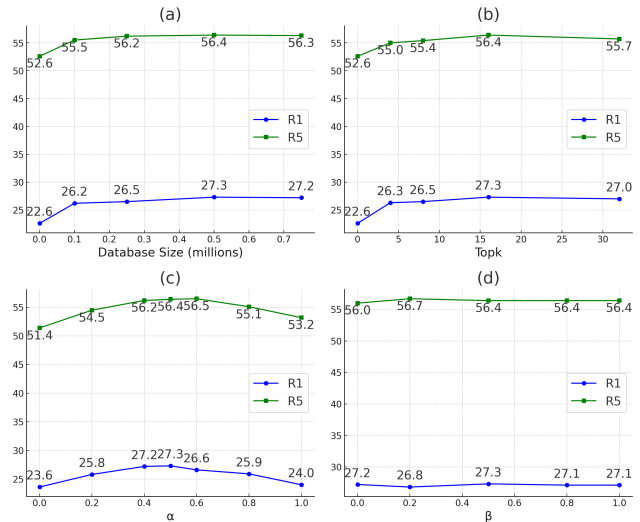


Figure 6. Visualization of how (a) the item amount of the database, (b) the number of retrieved neighbors, (c) the weight α for mixture feature during inference, and (d) the loss weight β influence the performance respectively on the CIRR validation set.

previous state-of-the-art method SEARLE-XL [3]. The results demonstrate the generalization ability of KEDs. Even in a narrow domain, KEDs still captures the attributes of the reference image from neighbors.

Table 4 presents the result on the CIRR test set, we observe a 2.2% and 2.4% improvement in Recall@1 and Recall@5 respectively. KEDs achieves 26.4% Recall@1, surpassing many supervised methods. The result reveals the real-world application potential of KEDs. Across these four datasets, KEDs consistently demonstrates its effectiveness in various compositional tasks.

We showcase qualitative example predictions by KEDs in Figure 5 and Figure 4. Examples illustrate the ability of KEDs to comprehend the semantic meaning of modification description while preserving visual information from the reference image.

4.4. Ablation Studies

To verify the influence of the modules in KEDs, we conduct a detailed ablation study on several aspects, as reported in Table 5 and illustrated in Figure 6.

Effectiveness of the proposed modules. We train KEDs with variants to evaluate the effectiveness of each module. Results are presented in Table 5. For the BKP module ϕ_M , we observe a consistent performance drop without the Top-K image or caption feature, indicating the importance of the knowledge modality. The performance also decreases when training without the shared linear block, emphasizing the importance of using a shared embedding space of ϕ_M to learn from external knowledge. Furthermore, the importance of the textual concept alignment branch ϕ_A is demonstrated by the performance boost. As mentioned in 3.3,






















Reference Image	Visual Knowledge and Text Knowledge						
	 old pepsi vending machine ... this is the exact one they had outside the center	 second image of the vintage bottle.	 a bottle of soft drink on wooden table	 illustrative image of a classic bottle	 suddenly the coke disappears and the bottle is left with a murky substance	 their bottle is very recognizable , as a matter of fact they included in the logo	
	 point a camera at a diver and watch them go from graceful o awkward !	 simply diving : octopus are encountered on almost every dive	 discover a new dimension to underwater , new wildlife and an adventure you will never forget	 local wildlife ... an image captured by underwater cameras that aim to show off natural beauty	 you might think you know these enormous inland seas , but watery depths are full of surprises .	 symptom is one of the major problems scuba divers deal with .	
	 check out all the new additions in baby and kids furniture , decor , toys , bedding and more	 this new butterfly piece crib bedding set with all the bundle you will need .	 baby 's room with a cot	 it 's time to prepare the best nursery for your prospective baby !	 if you are welcoming a little princess , you will want to consider our crib	 cream & beige nursery perfect for a little prince charming .	

Figure 7. Visualization of visual and text knowledge. The reference image is picked from the CIRR validation set and the visual and text knowledge is retrieved from CC3M.

the captions contain redundant semantics. Therefore the mapped features should specifically align with the subject phrases. This is achieved by offering contexts in ϕ_A . When the context semantics are omitted, the performance dramatically decreases. Additionally, there is a slight performance drop without extra captions, aligning with the intuition that extra captions depict objects from diverse perspectives, aiding semantic alignment. We compare different weights α for mixing dual-stream features during inference, a consistent improvement over single-stream features is observed in Figure 6 (c). In terms of the weights β for the extra caption loss during training in Figure 6 (d), results show that KEDs is robust to varying weights.

Analysis of the Top-K values. To find the optimal number of nearest neighbors to retrieve during training, we conduct experiments on the CIRR validation set and visualize the result in Figure 6 (b). A small value of K provides limited information while a large value affects efficiency. To strike a balance, we set K to 16. We also visualize the retrieved items in the Figure 7. The retrieved images and captions are closely relevant to the reference image in certain attributes, assisting KEDs in extracting common information. The bi-modality knowledge in the relevant images and captions enriches the semantics of the generated pseudo-tokens.

Design of the database. Figure 6 (a) visualizes how the item amount in the database influences performance. While more items improve the performance, the benefit is marginal. Therefore we set the number of items to 0.5 Million image-caption pairs. In the bottom part of Table 5, we additionally test two variants of visual and text features in the database during inference, *i.e.* substitute CC3M features with CIRR features or Fashion-IQ features. However, re-

sults indicate the CIRR features or Fashion-IQ features are not compatible with CC3M features. We believe the open-domain knowledge contributes to the zero-shot performance and the proposed Bi-modality Knowledge-guided Projection network learned the knowledge within the database.

5. Conclusion

In this paper, we propose a Knowledge Enhanced Dual Stream zero-shot composed image retrieval framework (KEDs) for Zero-shot Composed Image Retrieval. KEDs includes a Bi-modality Knowledge-guided Projection network. In particular, the network incorporates a database to provide relevant images and captions for the reference images. In this way, KEDs generates pseudo-word tokens with attribute information. In addition, KEDs integrates an extra stream to generate pseudo-word tokens aligned with textual concepts in the text embedding space. During inference, we combine pseudo-word tokens from two streams for retrieval. Extensive experiments show that KEDs surpasses previous methods on four datasets including ImageNet-R, COCO, Fashion-IQ and CIRR. Future work may leverage a Large language model to generate detailed descriptions.

6. Acknowledgement

This work is partially supported by National Science and Technology Major Project (2022ZD0117802). This work is also partially supported by the Fundamental Research Funds for the Central Universities (Grant Number: 226-2023-00126, 226-2022-00051) and China Postdoctoral Science Foundation (524000-X92302).

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. [3](#)
- [2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21466–21474, 2022. [1](#), [2](#), [5](#), [6](#)
- [3] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. *arXiv preprint arXiv:2303.15247*, 2023. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#)
- [4] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022. [3](#)
- [5] Wenhui Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. [3](#)
- [6] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 136–152. Springer, 2020. [2](#)
- [7] De Cheng, Gerong Wang, Bo Wang, Qiang Zhang, Jungong Han, and Dingwen Zhang. Hybrid routing transformer for zero-shot learning. *Pattern Recognition*, 137:109270, 2023. [3](#)
- [8] Niv Cohen, Rinon Gal, Eli A Meir, Gal Chechik, and Yuval Atzmon. “this is my unicorn, fluffy”: Personalizing frozen vision-language representations. In *European Conference on Computer Vision*, pages 558–577. Springer, 2022. [3](#)
- [9] Marcos V Conde and Kerem Turgutlu. Clip-art: Contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3956–3960, 2021. [3](#)
- [10] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):1–60, 2008. [1](#)
- [11] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. *arXiv preprint arXiv:2203.08101*, 2022. [1](#), [2](#), [6](#)
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [2](#), [5](#)
- [13] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021. [3](#)
- [14] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020. [6](#)
- [15] Yu Du, Fangyun Wei, Ziheng Zhang, Miaoqing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. [3](#)
- [16] Sepideh Esmailpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6568–6576, 2022. [3](#)
- [17] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. [3](#)
- [18] Lluís Gomez, Yash Patel, Marçal Rusinol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4230–4239, 2017. [3](#)
- [19] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VI 14*, pages 241–257. Springer, 2016. [1](#)
- [20] Anirudh Goyal, Abram Friesen, Andrea Banino, Theophane Weber, Nan Rosemary Ke, Adria Puigdomenech Badia, Arthur Guez, Mehdi Mirza, Peter C Humphreys, Ksenia Konyushova, et al. Retrieval-augmented reinforcement learning. In *International Conference on Machine Learning*, pages 7740–7765. PMLR, 2022. [3](#)
- [21] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoon Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023. [2](#)
- [22] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020. [2](#)
- [23] Sunan He, Taian Guo, Tao Dai, Ruizhi Qiao, Xiujuan Shu, Bo Ren, and Shu-Tao Xia. Open-vocabulary multi-label classification via multi-modal knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 808–816, 2023. [3](#)
- [24] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. [2](#), [5](#)
- [25] Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. Fid-light: Efficient and effective retrieval-augmented text generation. In *Proceedings of the 46th Inter-*

- national ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1437–1447, 2023. 2
- [26] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, 2015. Association for Computational Linguistics. 4
- [27] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4555–4564, 2016. 3
- [28] Zhengdong Hu, Yifan Sun, and Yi Yang. Suppressing the heterogeneity: A strong feature extractor for few-shot segmentation. In *The Eleventh International Conference on Learning Representations*, 2023. 3
- [29] Peiliang Huang, Junwei Han, De Cheng, and Dingwen Zhang. Robust region feature synthesizer for zero-shot object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7622–7631, 2022. 3
- [30] Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Retrieval-enhanced contrastive vision-text models. *arXiv preprint arXiv:2306.07196*, 2023. 2, 3
- [31] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022. 2
- [32] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023. 3
- [33] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3
- [34] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 4, 6
- [35] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9336–9345, 2022. 3
- [36] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1771–1779, 2021. 2
- [37] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 802–812, 2021. 2
- [38] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 2
- [39] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3
- [40] 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. 3
- [41] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3
- [42] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429, 2022. 3
- [43] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1405–1413, 2023. 3
- [44] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*, 2023. 3
- [45] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 3
- [46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 5
- [47] Qi Liu, Dani Yogatama, and Phil Blunsom. Relational memory-augmented language models. *Transactions of the Association for Computational Linguistics*, 10:555–572, 2022. 2
- [48] Rui Liu, Wenguan Wang, and Yi Yang. Volumetric environment representation for vision-language navigation. In *CVPR*, 2024. 3
- [49] Yikun Liu, Jiangchao Yao, Ya Zhang, Yanfeng Wang, and Weidi Xie. Zero-shot composed text-image retrieval. *arXiv preprint arXiv:2306.07272*, 2023. 2
- [50] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the*

- IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021. 1, 2, 5, 6
- [51] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6959–6969, 2022. 3
- [52] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [53] Fan Ma, Xiaojie Jin, Heng Wang, Jingjia Huang, Linchao Zhu, Jiashi Feng, and Yi Yang. Temporal perceiving video-language pre-training. *arXiv preprint arXiv:2301.07463*, 2023. 3
- [54] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reliable video narrator via equal distance to visual tokens. *arXiv preprint arXiv:2312.08870*, 2023. 3
- [55] Christian M Meyer and Iryna Gurevych. *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*. na, 2012. 3
- [56] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 3
- [57] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022. 3
- [58] Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodola, and Francesco Locatello. Asif: Coupled data turns unimodal models to multimodal without training. *arXiv preprint arXiv:2210.01738*, 2022. 3
- [59] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. In *ICCV*, 2019. 1
- [60] Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing raindrops and rain streaks in one go. In *CVPR*, 2021. 2
- [61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3, 6
- [62] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*, 2023. 2
- [63] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 251–260, 2010. 1
- [64] Yong Rui, Thomas S Huang, Shih-Fu Chang, et al. Image retrieval: Past, present, and future. *Journal of Visual Communication and Image Representation*, 10(1):1–23, 1999. 1
- [65] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314, 2023. 1, 2, 3, 4, 5, 6
- [66] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernamed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 5
- [67] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, et al. K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems*, 35:15558–15573, 2022. 3
- [68] Cheng Shi and Sibe Yang. Edadet: Open-vocabulary object detection using early dense alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15724–15734, 2023. 3
- [69] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016. 3
- [70] Yoav Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3
- [71] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval—an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448, 2019. 1, 2, 6
- [72] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 154–162, 2017. 1
- [73] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. 1, 2, 5
- [74] Junda Wu, Rui Wang, Handong Zhao, Ruiyi Zhang, Chaochao Lu, Shuai Li, and Ricardo Henao. Few-shot composition learning for image retrieval with prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4729–4737, 2023. 2
- [75] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19265–19274, 2023. 2, 3
- [76] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 3

- [77] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 3
- [78] Youngjae Yu, Seunghwan Lee, Yuncheol Choi, and Gunhee Kim. Curlingnet: Compositional learning between images and text for fashion iq data. *arXiv preprint arXiv:2003.12299*, 2020. 2
- [79] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023. 3
- [80] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pages 493–510. Springer, 2022. 3
- [81] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13041–13049, 2020. 3