NEUCORE: Neural Concept Reasoning for Composed Image Retrieval

Shu Zhao
Pennsylvania State University
University Park, USA
smz5505@psu.edu

Huijuan Xu
Pennsylvania State University
University Park, USA
hkx5063@psu.edu

Editors: Marco Fumero, Emanuele Rodolà, Clementine Domine, Francesco Locatello, Gintare Karolina Dziugaite, Mathilde Caron

Abstract

Composed image retrieval which combines a reference image and a text modifier to identify the desired target image is a challenging task, and requires the model to comprehend both vision and language modalities and their interactions. Existing approaches focus on holistic multi-modal interaction modeling, and ignore the composed and complimentary property between the reference image and text modifier. In order to better utilize the complementarity of multi-modal inputs for effective information fusion and retrieval, we move the multi-modal understanding to fine-granularity at concept-level, and learn the multi-modal concept alignment to identify the visual location in reference or target images corresponding to text modifier. Toward the end, we propose a **NEU**ral **CO**ncept **RE**asoning (NEUCORE) model which incorporates multi-modal concept alignment and progressive multi-modal fusion over aligned concepts. Specifically, considering that text modifier may refer to semantic concepts not existing in the reference image and requiring to be added into the target image, we learn the multi-modal concept alignment between the text modifier and the concatenation of reference and target images, under multiple-instance learning framework with image and sentence level weak supervision. Furthermore, based on aligned concepts, to form discriminative fusion features of the input modalities for accurate target image retrieval, we propose a progressive fusion strategy with unified execution architecture instantiated by the attended language semantic concepts. Our proposed approach is evaluated on three datasets and achieves state-of-the-art results. Code is available at `https://github.com/VisionLanguageLab/NEUCORE`.

1 Introduction

Composed image retrieval [Vo et al., 2019, Chen et al., 2020b, Liu et al., 2021, Delmas et al., 2022] aims to identify target image, corresponding to the input query composed of a reference image and a text modifier describing how the reference image should be modified, as illustrated in Figure 1. Compared to traditional image-to-image retrieval task [Gu et al., 2022, Zhao et al., 2020, 2021, Wu et al., 2022] and text-to-image retrieval task [Ma et al., 2022a, Lu et al., 2022] where single modality is provided as input, composed image retrieval is a challenging task as it requires joint vision and language understanding to retrieve the corresponding target image.

Existing works tackle this problem by directly fusing the multi-modal features after single modality encoding [Liu et al., 2021, Baldrati et al., 2022]. This type of approaches first processes input

Proceedings of the I edition of the Workshop on Unifying Representations in Neural Models (UniReps 2023).

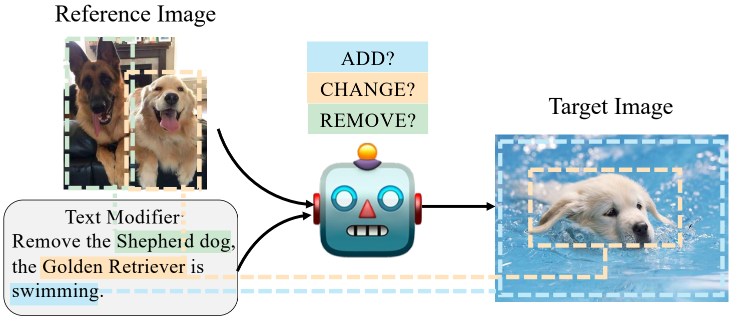

Figure 1: Example of composed image retrieval. Visual concepts are mined from images and aligned with semantic concepts from the text modifier. Based on aligned concepts, the reference image feature is fused with the text modifier in a sequential way to identify the target image feature. Different colors denote different concepts. Note that a concept from text modifier may appear in the reference image (Shepherd Dog), the target image (Swim), or both (Golden Retriever).

modality as a whole and lacks fine-grained multi-modal understanding at concept-level, preventing the model from performing concept-level composition, while most of the time the text modifier specifies partial semantic editing for the reference image. Therefore, we propose a **NEU**ral **CO**ncept **RE**asoning (NEUCORE) model to mine and align the visual concepts in reference and target images with the semantic concepts in text modifier, to enhance the semantic consistency between the multi-modal feature composition and the target image feature. Specifically, our NEUCORE model consists of two main components, i.e., multi-modal concept alignment and progressive multi-modal fusion over concepts.

Firstly, we propose to mine and align visual and semantic concepts under the weak supervision of image-text pair where the ground truth mapping of visual concepts in images and semantic concepts in sentences is unknown, instead of relying on object detectors [Ren et al., 2015] to generate region proposals and object tags to realize multi-modal concept alignment which suffers from the problems of heavy computational load and limited concept label space [Fang et al., 2022]. A multiple-instance learning framework with candidate visual concepts as instance, is designed with practical considerations of concept existence uncertainty, imbalance and noisy optimization, for tackling image-text pair weak supervision. Notably, a text modifier specifies the change to the reference image, and the change operations not only include attribute editing over the existing concepts in reference image, but also involve adding concepts into reference image or removing concepts from the reference image, which will cause the concept existence uncertainty among reference and target images. For example, in Figure 1, the shepherd dog in the text modifier appears in the reference image, but the golden retriever appears in both the reference and target images. To overcome this uncertainty, we concatenate the reference and target image tokens corresponding to candidate visual concepts, and employ a transformer [Vaswani et al., 2017] to jointly encode the reference and target images considering the nice property of patch-level feature encoding in transformer.

On top of visual encoding of reference and target images using transformer, the optimization for visual and semantic concept alignment is achieved by attention based multiple instance learning [Ilse et al., 2018] under the supervision of the semantic concepts parsed from the text modifier. Practically, the semantic concepts from each text modifier are represented by the multi-label representation in the concept label space which is constructed by the semantic concepts from all text modifiers in training data. Considering that each text modifier only mentions very limited semantic concepts in each pair of reference and target images compared to the large concept label space, there exists the problem of positive-negative imbalance during the optimization. Besides, the partial semantic editing property of the text modifier for the reference image may cause the situation that, some visual concepts are not referred by the text modifier and mislabeled as negative labels in the multi-label representation, according to the labeling rule that only the semantic concepts mentioned in the text modifier will be set as positive in the concept label space for that example. An asymmetric loss is applied to alleviate the imbalance and mislabeling problems during the concept alignment optimization. After the optimization, the visual tokens are assigned with semantic meaning and aligned with the semantic concepts referred in the text modifier.

Secondly, after aligning visual and semantic concepts, we are able to fuse multi-modal features at fine granularity instead of holistic visual and text features [Delmas et al., 2022, Liu et al., 2021] for final target image retrieval. A progressive multi-modal fusion module is proposed to gradually fuse the aligned concepts from the reference image and the text modifier in a sequential way with each step having its own focus. Progressive multi-modal fusion over concepts involves two sub-problems of how to generate the fusion operation sequence, and the specific design for each fusion operation. We propose a unified fusion operation design which can be instantiated by different sequence indicators to realize different fusion operations, through leveraging the advantage that the normalization layer can fuse its own preserved features, obtained from sequence indicators, with input visual features [Ulyanov et al., 2017]. The unified design overcomes the time-consuming drawback of previous hand-crafted fusion designs and removes the dependence on expert knowledge [Andreas et al., 2016b, Mao et al., 2019, Yi et al., 2018]. The fusion operation sequence is sequentially generated by the co-attention of global encoding over individual word embedding in the text modifier, and each sequence step focuses on local semantic concept feature guided by the global semantic context of text modifier. The attended local semantic concept feature is the sequence indicator at each sequence step, and is used to drive the instantiation of the unified fusion operation, taking on the role of a meta-learner. Our proposed fusion sequence generation method is optimized with the final retrieval loss without single step supervision needed as in sequence-to-sentence methods [Hu et al., 2017, Chen et al., 2020a, Mao et al., 2019, Yi et al., 2018, Johnson et al., 2017], and is able to deal with diverse sentences compared to language parser based approaches [Andreas et al., 2016b,a].

To summarize our contributions, we introduce a model NEUCORE for composed image retrieval consisting of a multi-modal concept alignment module and a multi-modal fusion module over aligned concepts. Our NEUCORE model learns fine-grained multi-modal concept alignment under image and sentence level weak supervision with actual influencing factors considered. Reference image and text modifier are progressively fused using a unified fusion operation over aligned concepts and under the sequence guidance of attended local semantic concepts, to gain a representation for target image retrieval with more semantic consistency. We validate our proposed model on three datasets. The results show that our method consistently outperforms the state-of-the-art, demonstrating the effectiveness of our approach.

2 Related Work

Composed Image Retrieval. Composed image retrieval task receives a reference image and a text modifier describing how the reference image should be modified to obtain the desired target image. Previous approaches [Vo et al., 2019, Chen et al., 2020b, Kim et al., 2021, Lee et al., 2021, Liu et al., 2021, Delmas et al., 2022] encode the reference image or text modifier into a holistic feature representation, and ignore the fine-grained information in composed image retrieval where the composed and complimentary property between the reference image and text modifier inputs happens. In this paper, our approach moves the multi-modal understanding to fine granularity at concept-level, and models the interactions between visual and semantic concepts for composed image retrieval.

Multi-Modal Concept Alignment. Multi-modal concept alignment aims to align visual concept space with semantic concept space to enable the fine-grained understanding and interaction between visual and semantic concepts [Li et al., 2022, Luo et al., 2022, Xu et al., 2019]. A line of methods [Liu et al., 2022, Li et al., 2021, Zhou et al., 2020, Li et al., 2020, Tan and Bansal, 2019] employ a pre-trained object detector [Ren et al., 2015] to generate region proposals and their object tags as the visual and semantic concepts, and then align them for downstream tasks. However, the object detectors limit the number of concepts as they are typically trained on limited pre-defined object categories. Some works investigate detector-free-based methods that use grid (patch) features [Fang et al., 2022, Jiang et al., 2020, Ma et al., 2022b] to predict visual concepts under weak supervision. Collaboration among foundation models has recently emerged as a promising direction to obtain multi-modal concepts [Zhao and Xu, 2023]. In this paper, our model mines and aligns multi-modal concepts under the setting that the text modifier only contains partial semantic editing property, which is challenging to find correspondence to align concepts for composed image retrieval.

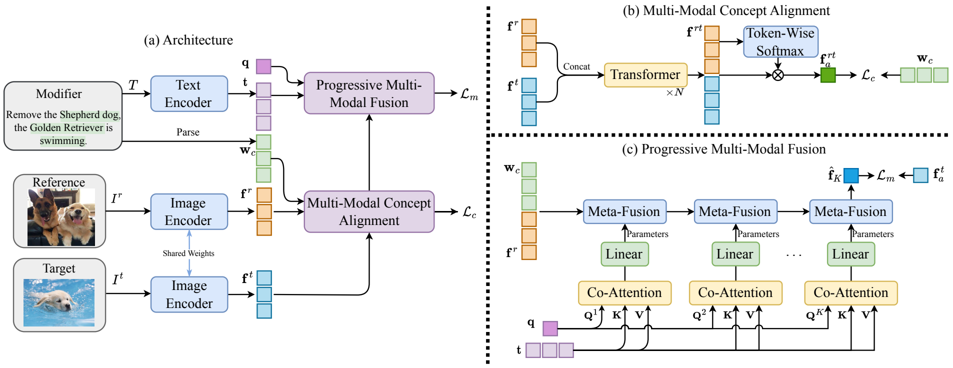

Figure 2: (a) The overall architecture of our proposed NEUCORE model. It mines and aligns multi-modal concepts, then fuses the reference image feature with the text modifier over aligned concepts to identify the target image feature. (b) Multi-modal concept alignment module mines visual concepts from images under image-level supervision and aligns them with semantic concepts from text modifiers. (c) Progressive multi-modal fusion module decomposes the text modifier to a fusion sequence by co-attention and progressive fuses the reference image and text modifier over aligned concepts.

3 Method

Given a reference image I^r and a text modifier T, the composed image retrieval task aims to combine them to identify the target image I^t. Previous approaches holistically process each input modality and then fusion, and is lack of the fine-grained compositional understanding. In this paper, we propose Neural Concept Reasoning (NEUCORE) to tackle the composed image retrieval by mining and aligning multi-modal concepts, and progressively fusing input modalities over concepts, as illustrated in Figure 2. The explanation of the symbols used in this paper is listed in supplementary material.

For feature encoding, the text modifier feature $\mathbf{q}$ and contextualized word features $\mathbf{t}$ are encoded by a text encoder E_T. An image encoder E_I is employed to extract reference and target image features, and the reference tokens $\mathbf{f}^r$ and target tokens $\mathbf{f}^t$ are obtained by flattening encoded visual features:

$$\mathbf{q}, \mathbf{t} = \mathrm{E}_T(T), \mathbf{f}^r = \mathrm{E}_I(I^r), \mathbf{f}^t = \mathrm{E}_I(I^t), \tag{1}$$

where superscripts ^{r} and ^{t} indicate that a feature belongs to the reference or target image, respectively.

3.1 Multi-Modal Concept Alignment

To model the fine-grained vision and language alignment between reference image and text modifier, we propose to mine and align visual concepts with semantic concepts from image-text pair data, which surpasses previous detector-based methods limited to small pre-defined label space. Due to the lack of concept-level supervision, we extract semantic concepts from text modifiers using a language parser as pseudo labels and apply the pseudo labels as image-level supervision. However, it is still challenging to learn multi-modal concept alignment with pseudo semantic concept labels at image level. Specifically, given a semantic concept, we cannot determine whether the correspondent visual concept appears in reference image or target image. For example, a text modifier "`Remove a dog`" means a `dog` concept exists in the reference image, not in the target image. And a text modifier "`Add a cat`" denotes a `cat` concept belonging to the target image, not in the reference image. Formally, given an input (I^r, I^t, or T), a concept set contains all concepts in the input, denoted as $C(\cdot)$. However, we cannot determine $c \in C(I^r)$ or $c \in C(I^t)$, where $c \in C(T)$. To resolve this ambiguity, we combine visual tokens of I^r and I^t to form a larger token set $I^{rt} = [I^r, I^t]$ and $C(T) \subset C(I^{rt})$, where $[\cdot, \cdot]$ denotes concatenation, considering that a concept c described in T must exist in I^r or I^t (or both).

Specifically, we parse[1] semantic concepts from the text modifier and embed each semantic concept $\mathbf{w}_c$ via GloVe [Pennington et al., 2014] word embedding:

$$\mathbf{w}_c = \mathrm{Embedding}(c). \tag{2}$$

where $c \in \mathbb{M}$ and $\mathbb{M}$ is the concept vocabulary constructed by semantic concepts from all text modifiers.

Then, we obtain $\mathbf{f}^{rt}$ by concatenating reference and target tokens, and employ a transformer Trans to exchange context and find correspondence:

$$\mathbf{f}^{rt} = \mathrm{Trans}([\mathbf{f}^r, \mathbf{f}^t]). \tag{3}$$

After modeling relation between reference and target tokens, we adopt a token-wise softmax to acquire attention weights $\mathbf{a}$, and then use the weights to summarize visual tokens to get a visual concept feature $\mathbf{f}_a^{rt}$ which contains visual foreground information:

$$\mathbf{a} = \mathrm{Softmax}(\mathbf{f}^{rt}), \mathbf{f}_a^{rt} = \mathbf{a} \cdot \mathbf{f}^{rt}. \tag{4}$$

Finally, multi-modal alignment score $\mathbf{s}$ is calculated by the concept features and embeddings of semantic concepts:

$$\mathbf{s} = \mathbf{f}_a^{rt} \cdot \mathbf{w}_c. \tag{5}$$

However, employing vanilla classification loss functions, such as the binary cross-entropy loss function, to optimize the score $\mathbf{s}$ leads to poor alignment. Text modifier only explicitly describes partial concepts compared to abundant visual concepts in the pair of input images, denoted as $\overline{C}(I^{rt}) = C(I^{rt}) - C(T)$ and $C(T) \subsetneqq C(I^{rt})$. However, all concepts belonging to $\overline{C}(I^{rt})$ are viewed as negative labels, leading to an increased risk of misclassifying visual concepts as background. On the other hand, the number of concept categories $|\mathbb{M}|$ is much larger than positive concepts described in each text modifier, *i.e.*, $|C(I^{rt})| \ll |\mathbb{M} - C(I^{rt})|$, causing high imbalance between positive and negative concept labels in the multi-label representation. As a result, the problems of mislabeling and imbalance hurt the training process, leading to incorrect concept alignment.

Therefore, we introduce an asymmetric loss [Ridnik et al., 2021] for multi-modal concept alignment to balance the visual and semantic concepts dynamically and discard possibly mislabeled concepts:

$$\mathcal{L}_{\mathrm{c}} = -\frac{1}{N}\left(\sum_{i\in\mathbb{P}}(1-\mathbf{s}_i^{'})^{\beta+}\log(\mathbf{s}_i^{'}) + \sum_{j\in\mathbb{N}}(\mathbf{s}_j^{'})^{\beta-}\log(1-\mathbf{s}_j^{'})\right), \tag{6}$$

where $\mathbf{s}^{'} = \mathrm{sigmoid}(\mathbf{s})$; $\mathbb{P}$ and $\mathbb{N}$ are the positive and negative set, respectively. $\beta+$ and $\beta-$ are two hyper-parameters that control the degree of focus on positive and negative concepts.

3.2 Progressive Multi-Modal Fusion over Concepts

After obtaining aligned multi-modal concepts, the reference image feature $\mathbf{f}^r$ and text modifier feature $\mathbf{q}$ are progressively fused in a sequential way over aligned concepts to identify the target concept feature $\mathbf{f}_a^t$. We propose to decompose the sequential fusion steps from the text modifier without step level supervision. Specifically, to focus on distinct semantic contexts of a text modifier, K independent fully connected layers $\mathrm{FC}_i, i = 1, 2, \cdots, K$ are employed to project text modifier feature $\mathbf{q}$. We adopt the Multi-Head Attention (MHA) [Vaswani et al., 2017] to extract the indicator vector for each semantic fusion step in the fusion sequence $\mathbb{S}$:

$$\mathbb{S} = (\mathrm{MHA}(\mathrm{FC}_1(\mathbf{q}), \mathbf{t}, \mathbf{t}), \mathrm{MHA}(\mathrm{FC}_2(\mathbf{q}), \mathbf{t}, \mathbf{t}), \cdots, \mathrm{MHA}(\mathrm{FC}_K(\mathbf{q}), \mathbf{t}, \mathbf{t})) \tag{7}$$

To progressively combine the reference image and text modifier over aligned concepts, we propose to instantiate specific operators from a meta-fusion architecture according to the generated fusion steps, which surpasses previous methods with time-consuming hand-crafted fusion operator design and the limit to expert knowledge. Our basic idea is that fusion steps can be clustered into semantic fusion groups, like "`ADD`" and "`REMOVE`," although the expressions within each group may vary in natural language. Therefore, we devise a transformer-based [Vaswani et al., 2017] meta-fusion

[1] https://spacy.io/

module $\mathrm{MetaFusion}$ and allow it to be instantiated for specific semantic fusion groups. Specifically, we employ a fully connected layer to generate parameters $\mathrm{FC}(\mathbf{S}_i)$ according to fusion steps' indicators, where $i = 1, 2, \cdots, K$ and $\mathbf{S}_i \in \mathbb{S}$, and initialize the normalization layers [Huang and Belongie, 2017] in the transformer with these parameters:

$$\begin{aligned} \mu_i &= \mathrm{FC}(\mathbf{S}_i), \\ \sigma_i &= \mathrm{FC}(\mathbf{S}_i). \end{aligned} \tag{8}$$

After transformer fusion instantiation, reference image tokens exchange information with aligned concepts in multi-head attention and fuse with text modifier information in the normalization layer:

$$\begin{aligned} \hat{\mathbf{f}}^{'}_{i-1} &= \mathrm{NL}(\hat{\mathbf{f}}_{i-1}; \mu_i, \sigma_i), \\ \mathbf{Q}, \mathbf{K}, \mathbf{V} &= \mathbf{W}\hat{\mathbf{f}}^{'}_{i-1}, \\ \hat{\mathbf{f}}^{''}_{i-1} &= \mathrm{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \hat{\mathbf{f}}^{'}_{i-1}, \\ \hat{\mathbf{f}}^{'''}_{i-1} &= \mathrm{NL}(\hat{\mathbf{f}}^{''}_{i-1}; \mu_i, \sigma_i), \\ \hat{\mathbf{f}}_i &= \mathrm{FFN}(\hat{\mathbf{f}}^{'''}_{i-1}) + \hat{\mathbf{f}}^{'''}_{i-1}, \end{aligned} \tag{9}$$

where NL is a normalization layer; FFN is a feedforward network; $\mathbf{W}$ is a weight matrix; $\hat{\mathbf{f}}_0 = \mathbf{f}^r$. After K fusion steps, we obtain $\hat{\mathbf{f}}_K$ as the modified image feature to match the target images.

Finally, the concept matching score $\mathbf{m}$ can be calculated by:

$$\mathbf{m} = \hat{\mathbf{f}}_K \cdot \mathbf{f}^t_a, \tag{10}$$

where $\mathbf{f}^t_a$ is the visual concept feature of target image.

To train our proposed NEUCORE model, given mini-batch data, the model is optimized by the batch-based classification loss, which is demonstrated to be an efficient optimization loss function for the composed image retrieval task in previous approaches [Vo et al., 2019, Lee et al., 2021, Delmas et al., 2022]:

$$\mathcal{L}_m = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\{\gamma \mathbf{m}(\mathbf{I}^r_i, \mathbf{T}_i, C(I^r_i), \mathbf{I}^t_i)\}}{\sum_j \exp\{\gamma \mathbf{m}(\mathbf{I}^r_i, \mathbf{T}_i, C(I^r_i), \mathbf{I}^t_j)\}}, \tag{11}$$

where γ is a temperature parameter. $\mathbf{m}(\mathbf{I}^i_r, \mathbf{T}^i, C(I^i_r), \mathbf{I}^i_t)$ is the matching score (cosine similarity), consisting of the concept matching score which is described in Equation (10) and the context matching score which is described in [Delmas et al., 2022]. Combining with the loss function of concept alignment, the final loss function is:

$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_c, \tag{12}$$

where α control the trade-off between the two loss functions.

The learning algorithm of our NEUCORE model is summarized in supplementary material.

4 Experiments

4.1 Datasets

To evaluate our NEUCORE model, we extensively conduct experiments on three widely used datasets: Shoes [Guo et al., 2018], FashionIQ[Wu et al., 2021], and CIRR [Liu et al., 2021].

Shoes is constructed from the Attribute Discovery Dataset [Berg et al., 2010]. In [Guo et al., 2018], authors additionally label natural language query sentences for the composed image retrieval task based on attribute labels in the original dataset. The dataset consists of 9k training triplets and 1.7k test queries. **FashionIQ** covers three fashion categories: Dress, Top tee, and Shirt. It contains 46k images for training, and 15k images for validation and testing. There are 18k queries for training, 12k queries for validation, and 12k queries for testing. Each query has two captions describing how to modify from the reference image to the target image. **CIRR** consists of over 36k open-domain images with human-generated text modifier. It is a more challenging dataset due to the richness of visual information and the diversity of language queries. Following [Liu et al., 2021], 36k triplets are split into 80% for training, 10% for validation, and 10% for testing in our experiments.

Table 1: **Results (%) on CIRR dataset.** CIRPLANT* employs a large pre-trained vision-language model.

Method	$R@K$				$R_s@K$			$\frac{(R@5+R_s@1)}{2}$
	$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$	
CIRPLANT* (init. OSCAR) Liu et al. [2021]	19.55	52.55	68.39	92.38	39.20	63.03	79.49	45.88
CIRPLANT Liu et al. [2021]	15.18	43.36	60.48	87.64	33.81	56.99	75.40	38.59
TIRG Vo et al. [2019]	10.01	38.31	54.59	84.69	37.36	59.31	72.51	37.84
ARTEMIS Delmas et al. [2022]	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
NEUCORE	**18.46**	**49.40**	**63.57**	**89.35**	**44.27**	**67.06**	**78.92**	**46.84**

4.2 Evaluation Protocol

Following [Delmas et al., 2022], we report composed image retrieval performance in Recall within top-K ($R@K$). Particularly, for CIRR dataset, following [Liu et al., 2021], we additionally report Recall within top-K on the visually similar subset ($R_s@K$), where the subset of candidate target images is visually similar to the correct target image, and it requires the fine-grained understanding ability of both vision and language modalities and their interactions. Following previous works [Delmas et al., 2022], we evaluate NEUCORE model on the test set of CIRR dataset, the validation set of Shoes dataset, and the validation set of FashionIQ dataset.

4.3 Implementation Details

We employ ResNet [He et al., 2016] pre-trained on ImageNet as the image encoder. For the text encoder, we adopt BiGRU [Cho et al., 2014] to encode sentence $\mathbf{q}$ and obtain the hidden states as contextualized word features $\mathbf{t}$. The concept alignment module consists of 2 transformer layers [Dosovitskiy et al., 2021]. The batch size is 32. Following [Delmas et al., 2022], we freeze the image encoder during the first 8 epochs. Then the model is trained for 50 epochs. We use the AdamW optimizer [Loshchilov and Hutter, 2019] and set the initial learning rate to 5×10^{-4} with a decay of 0.5 every 10 epochs. β_+ and β_- are 1 and 4, respectively. K is 3. γ is 2.65926. α is 200. The model is trained on one NVIDIA RTX A5000 GPU.

4.4 Main Results

We compare the performance of our proposed NEUCORE model with previous SOTA works on three benchmarks.

Results on CIRR Dataset. Table 1 shows the results. Our proposed NEUCORE model achieves 3.79% average improvement compared to the SOTA approach, ARTEMIS. Specifically, NEUCORE model improves 1.5%, 3.3%, 2.26%, and 1.62% in $K=1$, $K=2$, $K=10$, and $K=50$ of $R@K$, respectively. It demonstrates that NEUCORE model can better retrieve target images according to reference images and text modifiers. Moreover, NEUCORE model improves 4.28%, 4.86%, and 3.25% in $K=1$, $K=2$, and $K=3$ of $R_s@K$, indicating the recall of top-K on a visual similar subset, which is a more challenging metric and evaluates a model's fine-grained understanding ability on a visual similar target subset. Under the same pre-training condition without using the vision-language pre-trained weights, denoted as CIRPLANT, our model outperforms CIRPLANT significantly on all metrics. Furthermore, when the CIRPLANT model is initialized with the pre-trained weights from the OSCAR model [Li et al., 2020] trained on 6.5 million image-caption pairs, our model can still outperform it in $R_s@1$ and $R_s@2$ by 5.07 and 4.03, respectively. It indicates our proposed multi-modal concept alignment module can effectively mine and align fine-trained multi-modal concepts and the multi-modal fusion can fuse the reference image feature and text modifier feature over concepts to identify the target image feature.

Results on Shoes Dataset. The results are shown in Table 2. Our proposed NEUCORE model improves 1.04%, 2.37%, and 1.44% in $K=1$, $K=10$, and $K=50$ of Recall compared to the SOTA method ARTEMIS [Delmas et al., 2022] on this dataset. And it achieves 1.62% improvement on average, $(R@1 + R@10 + R@50)/3$.

Table 2: **Results (%) on Shoes dataset.**

Method	$R@1$	$R@10$	$R@50$	$\frac{R@1+R@10+R@50}{3}$
TIRG Vo et al. [2019]	14.46	47.51	75.17	45.71
VAL Chen et al. [2020b]	16.49	49.12	73.53	46.38
CoSMo Lee et al. [2021]	16.72	48.36	75.64	46.91
ARTEMIS Delmas et al. [2022]	18.72	53.11	79.31	50.38
NEUCORE	**19.76**	**55.48**	**80.75**	**52.00**

Table 3: **Results (%) on Fashion IQ dataset.**

Method	$R@10$	$R@50$	$\frac{R@10+R@50}{2}$
CIRPLANT Liu et al. [2021]	14.82	35.52	25.17
TIRG Vo et al. [2019]	23.17	47.48	35.32
CoSMo Lee et al. [2021]	19.87	42.65	31.26
ARTEMIS Delmas et al. [2022]	26.05	50.29	38.17
NEUCORE	**26.45**	**51.75**	**39.15**

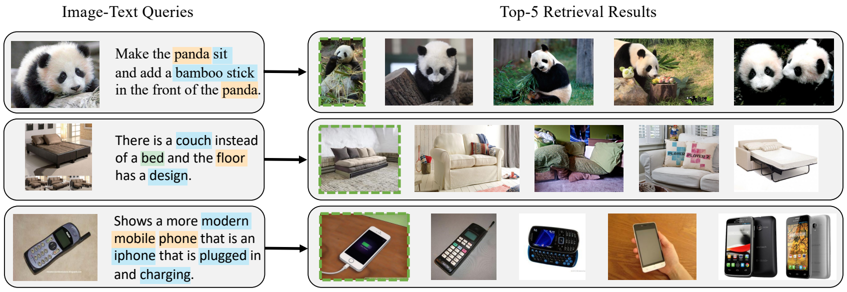
Image-Text Queries
Top-5 Retrieval Results
Make the panda sit and add a bamboo stick in the front of the panda.
There is a couch instead of a bed and the floor has a design.
Shows a more modern mobile phone that is an iphone that is plugged in and charging.

Figure 3: Qualitative examples of image-text queries on CIRR validation set and its Top-5 retrieval results. Green dotted boxes denote the ground-truth target images, and semantic concepts in the input text modifiers are highlighted by different colors. Green indicates that a concept appears only in the reference image. Yellow denotes that a concept appears both in the reference and target images. Blue means that a concept appears only in the target image.

Results on FashionIQ Dataset. Table 3 illustrates the main results. Our proposed NEUCORE model still obtains state-of-the-art results of $R@10$ and $R@50$. Due to the limited space, detailed results with more evaluation metrics on the FashionIQ dataset can be found in supplementary material.

4.5 Ablation Study

To show the effectiveness of each component of our model design, ablation study for different variants are conducted on the CIRR validation set as CIRR dataset contains more concepts and is better to evaluate the fine-grained understanding ability of models.

Multi-modal Concept Alignment Module. The ablation study results of the concept alignment module are illustrated in Table 4. "*Reference Only*" and "*Target Only*" denote that we do not concatenate the reference and target image tokens and only use reference or target image tokens to align with semantic concepts. The performance decrease of these two variants demonstrates that our concept alignment module with the concatenation of reference and target images can help alleviate the ambiguity problem of corresponding visual concepts, and improve multi-modal concept mining and alignment for composed image retrieval task. "*Cross-Entropy Loss*" means that the asymmetric loss for optimizing the concept alignment described in Equation (6) is replaced with the binary cross-entropy loss. The ablation results show the performance decrease, and confirm the existence of the problems about positive-negative imbalance and mislabeling in multi-label concept classification. Our proposed concept alignment module with asymmetric loss can help alleviate these problems.

Table 4: **Ablation results (%)** of multi-modal concept alignment module on CIRR validation set.

Variants	$R@5$	$R_s@1$	$\frac{(R@5+R_s@1)}{2}$
NEUCORE	**51.10**	**45.35**	**48.22**
Reference Only	50.74	42.76	46.75
Target Only	49.77	42.48	46.12
Cross-Entropy Loss	47.52	42.71	45.11

Table 5: **Ablation results (%)** of multi-modal fusion module on CIRR validation set.

Variants	$R@5$	$R_s@1$	$\frac{(R@5+R_s@1)}{2}$
NEUCORE	**51.10**	**45.35**	**48.22**
Remove Progressive Fusion Module	50.29	43.92	47.10
Layer Norm	49.32	44.73	47.03

Table 6: **Analysis of zero-shot novel concept generalization** on the new data split CIRR$_{\text{zs}}$ with novel concepts in test data. K denotes $R_s@K$ (%).

Method	$K = 1$	$K = 2$	$K = 3$
ARTEMIS Delmas et al. [2022]	30.86	55.71	68.29
NEUCORE	**37.43**	**63.14**	**72.57**
Remove Concept Module	31.71	53.71	66.29

Progressive Multi-Modal Fusion Module. The ablation study results of the progressive multi-modal fusion module are shown in Table 5. It confirms the effectiveness of our progressive multi-modal fusion module with automatic fusion sequence generation and unified fusion module design. "*Remove Progressive Fusion module*" means that we remove the multi-modal progressive fusion module and directly fuse concatenation of the reference image feature and text modifier feature. The decreased ablation results demonstrate that our proposed fusion module can better identify the target feature by progressively fusing the reference image feature and text modifier feature over aligned concepts with each step having focus. "*Layer norm*" denotes we use vanilla layer normalization instead of adaptive instance normalization without adaptive instantiation. The results show adaptive instance normalization can fuse features better in composed image retrieval task.

4.6 Analysis of Zero-Shot Concepts

Compared with previous approaches, our proposed NEUCORE model mines and aligns visual concepts and semantic concepts. It aligns the visual embedding space and semantic embedding space and transfers knowledge from language to vision, which can improve the zero-shot concept recognition ability. To demonstrate it, we create a data split from CIRR validation set, named CIRR$_{\text{zs}}$. Specifically, we parse the concepts from the training and validation sets. Next, we compute their difference set to obtain zero-shot concepts, *i.e.*, not seen during training time. Then we only keep these samples that contain zero-shot concepts to create zero-shot data split $\mathcal{D}_{\text{zs}}$, resulting in 316 zero-shot concepts and 350 samples.

Results are reported in Table 6. Compared to the SOTA approach ARTEMIS which fuses holistic multi-modal features for composed image retrieval, NEUCORE model improves 6.57%, 7.43%, and 4.28% in $K = 1$, $K = 2$, and $K = 3$ of recall within top-K of subset R_s, respectively. Moreover, we also present the results of removing the multi-modal concept alignment module, which results in performance drops of 5.72%, 9.43%, and 6.28% in $R_s@K$, respectively. It illustrates that the concept alignment module improves the zero-shot concept recognition ability by aligning visual concepts with semantic concepts represented by word embedding, which transfers the knowledge from word embedding to visual concepts.

4.7 Qualitative Results

Figure 3 shows the retrieval examples from a restricted subset of the CIRR validation set Liu et al. [2021], where candidate target images are visually similar. It requires learning fine-grained vision and language features and their interactions. The results show that our proposed NEUCORE model can understand the content of text modifier and compose the reference image feature and text modifier feature to identify the target image feature.

5 Conclusion

In this paper, we propose a model named NEUCORE to tackle the composed image retrieval task, which consists of multi-modal concept alignment module and progressive multi-modal fusion module. Multi-modal concept alignment module mines and aligns visual concepts from images with semantic concepts from text modifiers, and the progressive multi-modal fusion module progressively fuses the reference image feature with the text modifier feature over aligned concepts to identify the target image feature. Extensive experiments demonstrate our proposed NEUCORE model learns fine-grained multi-modal alignment and their interactions at concept-level from image-text paired data.

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. In *HLT-NAACL*, pages 1545–1554. The Association for Computational Linguistics, 2016a.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, pages 39–48. IEEE Computer Society, 2016b.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *CVPR*, pages 21434–21442. IEEE, 2022.

Tamara L. Berg, Alexander C. Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV (1)*, volume 6311 of *Lecture Notes in Computer Science*, pages 663–676. Springer, 2010.

Lichang Chen, Guosheng Lin, Shijie Wang, and Qingyao Wu. Graph edit distance reward: Learning to edit scene graph. In *ECCV*, volume 12364 of *Lecture Notes in Computer Science*, pages 539–554. Springer, 2020a.

Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *CVPR*, pages 2998–3008. Computer Vision Foundation / IEEE, 2020b.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734. ACL, 2014.

Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. ARTEMIS: attention-based retrieval with text-explicit matching and implicit similarity. In *ICLR*. OpenReview.net, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Injecting semantic concepts into end-to-end image captioning. In *CVPR*, pages 17988–17998. IEEE, 2022.

Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with RGB modality only. In *CVPR*, pages 1050–1059. IEEE, 2022.

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogério Schmidt Feris. Dialog-based interactive image retrieval. In *NeurIPS*, pages 676–686, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *ICCV*, pages 804–813. IEEE Computer Society, 2017.

Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1510–1519. IEEE Computer Society, 2017.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2132–2141. PMLR, 2018.

Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik G. Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *CVPR*, pages 10264–10273. Computer Vision Foundation / IEEE, 2020.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. Inferring and executing programs for visual reasoning. In *ICCV*, pages 3008–3017. IEEE Computer Society, 2017.

Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *AAAI*, pages 1771–1779. AAAI Press, 2021.

Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *CVPR*, pages 802–812. Computer Vision Foundation / IEEE, 2021.

Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Unsupervised vision-and-language pre-training without parallel images and captions. In *NAACL-HLT*, pages 5339–5350. Association for Computational Linguistics, 2021.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer, 2020.

Zhi Li, Lu He, and Huijuan Xu. Weakly-supervised temporal action detection for fine-grained videos with hierarchical atomic actions. In *ECCV*, volume 13670 of *Lecture Notes in Computer Science*, pages 567–584. Springer, 2022.

Yongfei Liu, Chenfei Wu, Shao-Yen Tseng, Vasudev Lal, Xuming He, and Nan Duan. KD-VLP: improving end-to-end vision-and-language pretraining with object knowledge distillation. In *NAACL-HLT (Findings)*, pages 1589–1600. Association for Computational Linguistics, 2022.

Zheyuan Liu, Cristian Rodriguez Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, pages 2105–2114. IEEE, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*. OpenReview.net, 2019.

Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. COTS: collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *CVPR*, pages 15671–15680. IEEE, 2022.

Zhekun Luo, Shalini Ghosh, Devin Guillory, Keizo Kato, Trevor Darrell, and Huijuan Xu. Disentangled action recognition with knowledge bases. In *NAACL-HLT*, pages 559–572. Association for Computational Linguistics, 2022.

Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. EI-CLIP: entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *CVPR*, pages 18030–18040. IEEE, 2022a.

Xiaojian Ma, Weili Nie, Zhiding Yu, Huaizu Jiang, Chaowei Xiao, Yuke Zhu, Song-Chun Zhu, and Anima Anandkumar. Relvit: Concept-guided vision transformer for visual relational reasoning. In *ICLR*. OpenReview.net, 2022b.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *ICLR*. OpenReview.net, 2019.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543. ACL, 2014.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

Tal Ridnik, Emanuel Ben Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *ICCV*, pages 82–91. IEEE, 2021.

Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In *EMNLP*, pages 5099–5110. Association for Computational Linguistics, 2019.

Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*, pages 4105–4113. IEEE Computer Society, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

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 *CVPR*, pages 6439–6448. Computer Vision Foundation / IEEE, 2019.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogério Feris. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *CVPR*, pages 11307–11317. Computer Vision Foundation / IEEE, 2021.

Hui Wu, Min Wang, Wengang Zhou, Houqiang Li, and Qi Tian. Contextual similarity distillation for asymmetric image retrieval. In *CVPR*, pages 9479–9488. IEEE, 2022.

Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069, 2019.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In *NeurIPS*, pages 1039–1050, 2018.

Shu Zhao and Huijuan Xu. Less is more: Toward zero-shot local scene graph generation via foundation models. *CoRR*, abs/2310.01356, 2023.

Shu Zhao, Dayan Wu, Wanqian Zhang, Yu Zhou, Bo Li, and Weiping Wang. Asymmetric deep hashing for efficient hash code compression. In *ACM Multimedia*, pages 763–771. ACM, 2020.

Shu Zhao, Dayan Wu, Yucan Zhou, Bo Li, and Weiping Wang. Rescuing deep hashing from dead bits problem. In *IJCAI*, pages 1338–1344. ijcai.org, 2021.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, pages 13041–13049. AAAI Press, 2020.