Ask&Confirm: Active Detail Enriching for Cross-Modal Retrieval with Partial Query

Guanyu Cai[1,2*], Jun Zhang[2], Xinyang Jiang[3†], Yifei Gong[2],
Lianghua He[1], Fufu Yu[2], Pai Peng[2], Xiaowei Guo[2], Feiyue Huang[2], Xing Sun[2†]
Tongji University[1], Tencent Youtu Lab[2], Microsoft Research[3]
{caiguanyu,Helianghua}@tongji.edu.cn,xinyangjiang@microsoft.com,pengpai_sh@163.com
{bobbyjzhang,yifeigong,fufuyu,scorpioguo,garyhuang,winfredsun}@tencent.com

Abstract

Text-based image retrieval has seen considerable progress in recent years. However, the performance of existing methods suffers in real life since the user is likely to provide an incomplete description of an image, which often leads to results filled with false positives that fit the incomplete description. In this work, we introduce the partial-query problem and extensively analyze its influence on text-based image retrieval. Previous interactive methods tackle the problem by passively receiving users' feedback to supplement the incomplete query iteratively, which is time-consuming and requires heavy user effort. Instead, we propose a novel retrieval framework that conducts the interactive process in an Ask-and-Confirm fashion, where AI actively searches for discriminative details missing in the current query, and users only need to confirm AI's proposal. Specifically, we propose an object-based interaction to make the interactive retrieval more user-friendly and present a reinforcement-learning-based policy to search for discriminative objects. Furthermore, since fully-supervised training is often infeasible due to the difficulty of obtaining human-machine dialog data, we present a weakly-supervised training strategy that needs no human-annotated dialogs other than a text-image dataset. Experiments show that our framework significantly improves the performance of text-based image retrieval. Code is available at `https://github.com/CuthbertCai/Ask-Confirm`.

1. Introduction

Recently, cross-modal retrieval, especially text-based image retrieval has gained increasing attention [36]. Although significant improvement has been achieved with existing methods [15, 36, 7] for text-based retrieval, we found in practice their retrieval result is barely satisfactory when users only describe some local regions in an image.

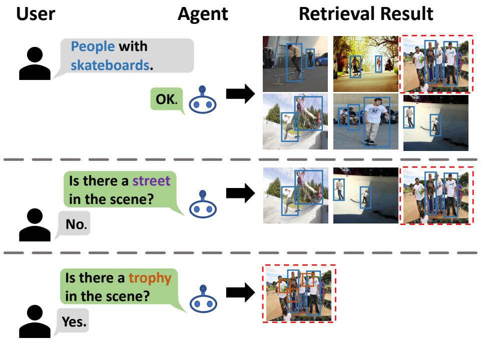
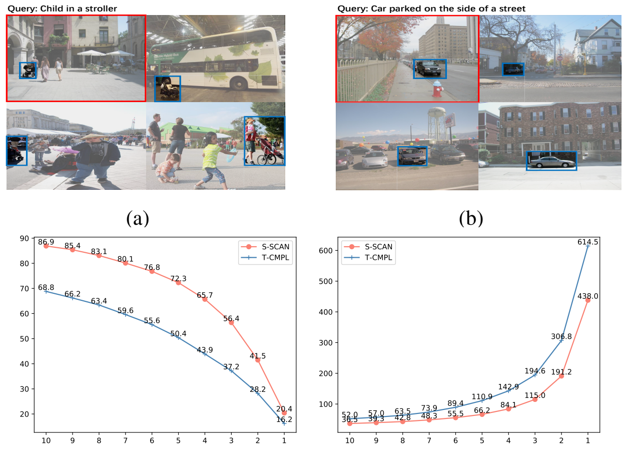

User
Agent
Retrieval Result
People with skateboards.
OK.
Is there a street in the scene?
No.
Is there a trophy in the scene?
Yes.

Figure 1: An illustration of Ask&Confirm. The agent enriches the textual query and narrows down the retrieval scope by iteratively asking users to confirm more information. The target image is highlighted with a red rectangle.

In this work, we introduce a new concept of *partial-query* problem in text-based image retrieval, where the initial text query only describes some objects in the target image. Studies [30, 28] have found that when examining an image, people tend to only focus on the objects that stand out the most. This could lead to problems where the objects that people focus on are not the discriminative objects that can distinguish the target image from similar candidates, thus making the user's input insufficient for retrieving the target image. As shown in Figure 2 (a) and (b), a cross-modal retrieval model performs poorly when a query is only partially given. In both examples, the target image ranks lower than 1000 th, while the other false positives rank top three. A common object (blue box) described by the partial query is presented in all images. However, the rest of images are vastly different. For example, in Figure 2 (a), besides the stroller mentioned in the query, the target image consists of umbrellas, chairs, and so on. Whereas the others consist of different objects like trees and buses. If the retrieval model receives a complete description including all objects, existing methods [15, 17, 32] perform excellently. To show how the partial query hurts retrieval, we test two

[*] Work done during internship at Youtu Lab
[†] Corresponding author: Xinyang Jiang, Xing Sun



(a)

(b)

(c) R@10

(d) Mean Rank

Figure 2: Effect of partial queries. (a) and (b) are visualizations of partial-query retrieval. The target image is surrounded by a red box and the others are the top three ranked scenes. The region that matches the query is surrounded by a blue box. (c) and (d) demonstrates R@10 and Mean Rank of a retrieval model as queries decrease. The horizontal axis represents the query number.

text-image retrieval models, S-SCAN and T-CMPL on Visual Genome [14], which are modified from SCAN [15] and CMPL [36]. The implementation is detailed in Section 4. For each image, its complete description includes 10 captions for different regions. We gradually decrease the number of captions and use them as queries to retrieve the target image. As shown in Figure 2 (c) and (d), for both models, R@10 decreases and Mean Rank increases as the degree of incompletion increases. These results reveal that partial queries should be tackled for a robust retrieval model.

Existing interactive retrieval models [6, 33, 26, 13, 11, 35, 7, 31] tackle the partial query by involving feedback of users in the retrieval process. Given the initial queries from users, these methods first give several relevant candidates that could potentially be the target image. By comparing the target image with these reference images, users give the retrieval method different forms of feedback to describe the difference between them, such as scores [26, 33], tags [13, 11, 12] or descriptions [7, 29]. The models then refine the retrieval results according to the user feedback and continue next round of iteration until the target image is found. Previous methods only passively receive additional information from users, so users need to have substantial practice and expert knowledge on the retrieval system to give discriminative feedback that can quickly narrow down the retrieval range. Hence, to free users from the burden of analyzing the retrieval results and looking for the discriminative information, we propose that the retrieval model itself should be able to actively search for the discriminative information the current query misses. Another problem of previous interactive retrieval models is time-consuming. For example, description-based methods [7, 29] require users to input long sentence feedback and tag-based methods [13, 11, 12] require users to input a bunch of attributes. Hence, we propose a framework where users only need to make simple yes/no confirmation on AI's question.

In this paper, we propose a novel interactive retrieval framework called *Ask&Confirm* as shown in Figure 1. The agent first retrieves a set of relevant candidates from the gallery based on initial text queries. Then, it will analyze the retrieval results and the overall status of gallery, and actively select discriminative object candidates for users to confirm their presence. Based on users' confirmation, the agent narrows down the range of candidates and eventually gathers enough information to locate the target image. Instead of passively receiving user feedback, a reinforcement learning (RL) based policy is trained to actively search for the discriminative objects missed in the query, and use these objects to distinguish the target image from the rest of gallery. In this active object-based interaction, users only need to confirm the existence of the proposed objects in the target image, no expert knowledge on the retrieval task and extra effort is needed. Moreover, unlike previous RL based interactive methods [7, 19] that require human-annotated dialogs which is impractical to widely collect, our Ask&Confirm framework is trained in a weakly-supervised manner, where only text-image pairs are needed.

The contributions of our framework are as follows: 1) To our knowledge, this is the first work that formally addresses and analyzes the problem of partial query in cross-modal retrieval. 2) Instead of passively receiving missing details from user feedback, we propose a novel interactive retrieval framework *Ask&Confirm* that introduces an active object-based interaction to actively select the most discriminative objects for users to confirm. 3) Rather than using human-annotated dialogs, we propose a weakly-supervised reinforcement learning framework to optimize the interactive policy that explores the statistical characteristics of the gallery. Experiments show that our framework is effective and robust with partial queries.

2. Related Work

2.1. Text-based Image Retrieval

Most text-based image retrieval approaches are based on deep neural networks [36, 15, 17, 9, 32, 5]. The objective of them is to accurately measure the similarity between the inputs from two different modalities. Cross-Modal Projection Learning (CMPL) [36] is proposed to pull image and text embeddings into an aligned space. To further enhance the retrieval in a fine-grained way, [15, 17, 9, 32] proposed different attention-based approaches, applying visual attention between every image region and word.

2.2. Query Expansion

Query expansion tackles incomplete information. Different from partial queries that are complete sentences of local regions, it focuses on queries that are incomplete sentences. An incomplete sentence as the query leads to poor retrieval. Thus, query expansion methods are proposed [37, 18, 22, 4, 8]. [37] learns users' searching history to generate expansion. [18] explores expansion by calculating similarity distance in thesaurus indexed collections. Other methods [22, 4, 8] that focus on image or video retrieval provide expansion based on knowledge bases.

2.3. Visual Dialog

Visual dialog aims to let the machine understand the visual content and have a natural conversation with the user about it. After examining the image, the agent can answer the user's questions on different aspects. Mainstream approaches are based on policy-based reinforcement learning to achieve good question-answer performance [24, 2, 3]. However, the dialogs are purely text-based for both the questioner and answer agent, and a manually annotated dialog dataset is needed to train a visual dialog system.

2.4. Interactive Image Retrieval

The retrieval model is hard to locate the target image with the initial query. Inspired by visual dialog, interactive image retrieval systems [33, 26, 13, 12, 11, 25, 16, 20, 21] are proposed to solve this problem. In these systems, users give feedback to an agent according to a reference image. There are two types of feedback: relevance and difference. For the former one [33, 26], users give relevance scores for the current retrieval results. Then the system re-ranks its retrieval results by using the user's feedback. For the latter one [13, 12, 11, 25, 34], users tell the difference between the target image and a reference image to the system with tags or descriptions. The system then whittles away the irrelevant images and ranks the correct one to the top.

3. Method

3.1. Object-based Interaction

In a partial-query problem, one of the most important tasks for an interactive retrieval model is to obtain the missing discriminative information that can distinguish the target image from others. Generally, the demands of more discriminative information and less user effort are contradictory, because more information usually means that the user has to pay more effort to think about what is the most discriminative thing and to input more descriptions. For example, tag-based methods [13, 11, 12] only require the user to point out a different attribute between the target image and a reference image, but they hardly filter out many negative images per round because too little discriminative information is provided. On the contrary, description-based methods [7, 29] require the user to give long sentence feedback that enriches more details but pays more user effort.

Method	R@1	R@5	R@10	MR
S-SCAN	4.5	13.6	20.4	416.0
S-SCAN+Objects	46.4	70.2	78.4	28.4

Table 1: Retrieval improvements over S-SCAN with ground-truth object descriptions. MR means Mean Rank.

In Ask&Confirm, we propose an object-based interaction where a RL-based policy actively searches for discriminative object candidates for users to confirm, then users just need to confirm whether objects are in the target image. Under the auxiliary of the active policy, the demands of more discriminative information and less user effort are simultaneously satisfied.

We choose object-based interaction based on two main reasons: (1) objects in an image are discriminative enough to distinguish different images, (2) objects can be easily obtained with a pre-trained detector such as RCNN [1].

Firstly, we discover that the distribution of objects in an image gallery is generally low-entropy, making it a discriminative feature for retrieving the target image. For example, in Visual Genome [14], some objects, such as "trophy" and "skateboard", rarely appear. If an image includes them, they are discriminative enough to narrow down the retrieval scope quickly. To verify this observation, two types of queries are compared using the same retrieval method S-SCAN : partial query only and supplement partial query with the name of the objects. As shown in Table 1, remarkable improvements are achieved by adding object words, verifying that objects contain discriminative information to distinguish the target image from the rest of the gallery.

Second, the convenience of obtaining objects of an image also makes the object-based interaction practical. Previous text-based image retrieval methods [5, 15, 17, 32] extract image features by an object detector [1]. By reusing the detector, we can directly obtain objects of each image.

3.2. Interactive Retrieval Agent

By adopting the proposed object-based interaction, we propose an interactive retrieval agent to tackle the partial-query problem. It takes the charge of extracting features, interacting with the user and retrieving the target image. In this section, we illustrate how the agent works, especially how it actively searches for object candidates for the user to confirm, which greatly reduces the user effort.

Define a set of captions $Q = \{q_n\}_{n=1}^{N_Q}$ that composes descriptions of an image i, where each q_n describes a region. By regarding Q as queries, the goal of a retrieval agent R is to retrieve the target image i_* from a gallery $I = \{i_n\}_{n=1}^{N}$ through T rounds interaction with the user.

The partial-query problem considers that Q only describes parts of an image instead of the full image.

The interactive retrieval agent R includes four main components: **Text Encoder**, **Image Encoder**, **Candidate Generator** and **Ranker**. As the interactive workflow 3 illustrated, Text Encoder and Image Encoder embed partial queries and images to a textual-visual feature space as textual features and visual features respectively. At each round, Candidate Generator actively searches for the most discriminative objects as candidates for the user to confirm. Given the objects, the user confirms them as the positive or negative, where positive objects refer to the ones that exist in the target image, vise versa. Then, names of positive objects are added to the partial query and the new query's feature is updated by the Text Encoder. Finally, based on positive objects, negative objects and the features of queries and images, Ranker retrieves the target image. In detail, Ranker first computes an initial similarity between the textual query and visual features. Secondly, the initial similarity is further refined by the user-confirmed objects. If a gallery image contains the negative objects, the similarity between the image and queries would be refined to a lower value. The retrieval result of the current round is given by the refined similarity. Below we provide details on the specific design of each component.

Text Encoder. At t th round, the input partial queries are denoted as $Q_t = \{q_n\}_{n=1}^{N_Q^t}$. They are embedded into textual features by Text Encoder (TE):

$$x_n^T = TE(q_n), q_n \in Q_t \tag{1}$$

where x_n^T denotes a texture feature. The set of all textual features of Q_t is denoted as $X_t^T = \{x_n^T\}_{n=1}^{N_Q^t}$. In detail, we use a gated recurrent unit as TE just like [29].

Image Encoder. Given an image gallery $I = \{i_n\}_{n=1}^{N}$, Image Encoder (IE) extracts the visual feature and detects objects for each image:

$$(x_n^I, A_n) = IE(i_n) \tag{2}$$

where x_n^I denotes the visual feature of i_n and A_n denotes the objects $\{a_1, a_2, ...\}$ that appear in i_n. The set of all visual features of I is denoted as $X^I = \{x_n^I\}_{n=1}^{N}$.

Candidate Generator. At t th round, Candidate Generator actively searches for the most discriminative objects as candidates for the user to confirm positive objects that appear in the target image i_*. These candidates are denoted as $A_t = \{a_n\}_{n=1}^{N_A}$. The user confirms positive objects $A_t^p = \{a_n^p\}_{n=1}^{N_A^p}$, thus, the rest of A_t are negative objects. They are denoted as $A_t^q = \{a_n^q\}_{n=1}^{N_A^q}$ where $N_A^q + N_A^p = N_A$.

The text of A_t^p is used as the additional description of i_*. It is denoted as $Q_t^c = \{\mathbb{T}(a_n^p)\}_{n=1}^{N_A^p}$, where $\mathbb{T}(a_n^p)$ is the word of a_n^p. To enrich details of the target image, the additional description is added into queries where $Q_t = Q_{t-1} \cup Q_t^c$.

Ranker. Given X_t^T, X^I, and A_t^q, Ranker gives a retrieval result of t th round. Firstly, Ranker computes the similarity between queries and each image, where $S_{t,n}(X_t^T, x_n^I)$ denotes the similarity between X_t^T and x_n^I. Secondly, if i_n contains negative objects belong to A_t^q, we refine $S_{t,n}$ with a lower value where $S_{t,n} := S_{t,n} \times 0.9$. With the refined similarity, Ranker gives a retrieval result.

3.3. Weakly-supervised Policy Learning

The key of Ask&Confirm to satisfy the demand of more discriminative information and less user effort is an active search policy. It selects the most discriminative objects as candidates for users to confirm, according to the textual feature and the object distribution of an image gallery. Thus, it frees the user to think about what are the most discriminative objects and input long sentence feedback.

In this work, the active search policy is learned with a weakly-supervised RL-based training. The weakly-supervised policy learning automatically finds an optimal policy by letting the agent iteratively interact with users and self-update based on the users' feedback. The whole policy learning is very concise and can be easily conducted in a weakly-supervised manner, because we only need to know objects in each image and users' feedback can be mimicked by ground-truth objects in the target image. We can even just reuse the detector for extracting image features to detect objects. On the contrary, previous dialog-based retrieval methods [7, 19, 3] require burdensome collections of chatting sessions. The superiority that our method needs no extra data collections makes it more practical.

Reinforcement Learning. The policy obtained by Candidate Generator is modeled as a policy net π, parameterized with ϕ_π, which outputs each object's probability $P(a)$ of getting selected. The five components in our policy learning *action*, *state*, *policy*, *value* and *reward* are as follows:

Actions refers to the objects selected by Candidate Generator at each round, i.e., $a \in \mathcal{A}$, $\mathcal{A}$ is the set of all objects.

State s^t is defined as a concatenation of $s_1^t = \sum_{n=1}^{N_Q^t} x_n^T / N_Q^t$ and $s_2^t = P_r(a)$, where $P_r(a)$ is the distribution of a among the top 100 images generated by the Ranker. We utilize such design to make π aware of information both from partial queries and the ranking list.

Reward is defined as the similarity between textual features and the target visual feature, i.e., $S(X_t^T, x_*^I)$ where x_*^I is the visual feature of the target image.

Policy π is implemented with a three-layer MLP. The object sampling distribution $P(a)$ is approximated with $\pi(s^t)$.

Value is estimated with $V(s^T)$. The value net V is implemented with a two-layer MLP, parameterized with ϕ_v.

Given the actions, state, reward, value and policy, a Proximal Policy Optimization (PPO) [27] is applied to optimize the policy net ϕ_π and ϕ_v. Please refer to the original paper of PPO for more details.

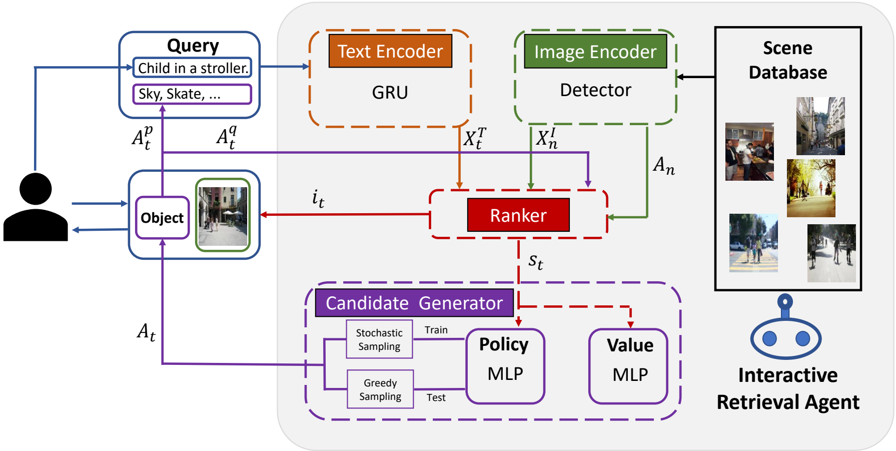


Figure 3: The proposed interactive cross-modal retrieval framework of Ask&Confirm. The interactive retrieval agent gradually enriches details of an image by heuristically providing users with object candidates.

Shaping. RL is hard to converge for dialog agents [23, 7], thus, previous RL-based dialog agents [7, 24, 3] adopt a supervised learning with annotated dialogs for shaping the RL training. To avoid the burdensome human annotation, we propose a weakly-supervised shaping method without annotated dialogs. Our motivation is that objects in the target image should have a high probability to be selected, because adding these objects into the queries could potentially significantly increase the similarity between queries and the target image. However, this probability is infeasible to obtain during test time, because the target image is unknown. As a result, instead of obtaining the probability of objects existing in the target image, we approximate it with the probability of objects that semantically relevant to the target image's corresponding query. For example, if the target image's corresponding query is "a man is surfing", we can infer that objects relevant to this query (e.g., "man", "sea" and "surfboard") should have a high probability to appear in the target image. The semantic relevance between an object and a query can be estimated by the conditional probability of an object a_j existing in the query's corresponding target image, given the query Q_t, denoted as $P(a_j|Q_t)$. $P(a_j|Q_t)$ can be estimated by computing the frequency of a_j and Q_t both appearing in the same target image i_k:

$$P(a_j|Q_t) = \frac{\sum_{k=1}^{N} \mathbb{1}(a_j \in i_k || Q_t \in i_k)}{\sum_{m=1}^{|\mathcal{A}|} \sum_{k=1}^{N} \mathbb{1}(a_m \in i_k || Q_t \in i_k)} \quad (3)$$

where $\mathbb{1}(\cdot)$ is an indicator function. $Q_t \in i_k$ is a corresponding query of i_k and $a_j \in i_k$ denotes an object in i_k.

A practical problem is that Q_t hardly appears in different i_k, which causes $\sum_{k=1}^{N} \mathbb{1}(a_j \in i_k || Q_t \in i_k)$ always being 1. Thus, we use a set of words $\{w_n\}_{n=1}^{N_w}$ to represent Q_t, where w_n is a tokenized word in Q_t. The tokenized word w_n could appear in different images. $\mathbb{1}(Q_t \in i_k)$ is replaced with $\sum_{n=1}^{N_w} \mathbb{1}(w_n \in i_k)$. $P(a_j|Q_t)$ is then modified to:

$$P(a_j|Q_t) = \frac{\sum_{k=1}^{N} \sum_{n=1}^{N_w} \mathbb{1}(a_j \in i_k || w_n \in i_k)}{\sum_{m=1}^{|\mathcal{A}|} \sum_{k=1}^{N} \sum_{n=1}^{N_w} \mathbb{1}(a_m \in i_k || w_n \in i_k)} \quad (4)$$

Guiding with $P(a|Q_t)$, We then train π by optimizing

$$\mathcal{L}_s = \sum_{t=1}^{N_s} (P(a|Q_t) - \pi(s^t))^2 \quad (5)$$

where N_s means that $\mathcal{L}_s$ is optimized for every N_s rounds.

Combining RL with the shaping, loss of the policy learning process is $\mathcal{L} = \mathcal{L}_p + \alpha \cdot \mathcal{L}_s$, where $\mathcal{L}_p$ denotes the loss of PPO and coefficient α is used to balance the RL learning and shaping. The shaping is crucial in our method otherwise the training process cannot converge.

4. Experiments

Dataset. There is no existing benchmark for interactive partial-query retrieval and we build a new dataset based on Visual Genome [14]. In Visual Genome, multiple regions are detected by an object detector [1] for each image, and each of the object region is annotated with a description. We preprocess the data by following the protocol in [29], resulting in 105,414 images. Images are split into 92,105/5,000/9,896 for training/validation/testing. To perform an interactive partial-query retrieval without extra data collection, we regard a region caption as a partial query offered by users and objects in the target image as feedback from users. All evaluations are performed on the test split.

Baselines. Ask&Confirm is a simple framework compatible to any cross-modal retrieval methods. We implement variants of SCAN [15] and CMPL [36], which are named Simplified SCAN (S-SCAN) and CMPL with Triplet loss (T-CMPL) respectively, as the basic retrieval models and build the proposed interactive retrieval agent on them. Both of the variants adopt the text and image encoder in Section 3.2 to obtain textual features $X^T = \{x_j^T\}_{j=1}^{J}$ and visual features $X^I = \{x_{k,m}^I\}_{k,m=1}^{K,M}$. (a) **S-SCAN**: We modify the bidirectional attention mechanism in SCAN to a unidirectional one to adopt multi-query inputs. Thus, the similarity between x_j^T and x_k^I is modified as

$$S_{j,k}(x_j^T, x_k^I) = \frac{1}{M} \sum_{m=1}^{M} \gamma_{j,k} \cdot cos(x_j^T, x_{k,m}^I) \quad (6)$$

Method	R@1	R@5	R@10	MR	Q	A
S-SCAN	4.5	13.6	20.4	416.0	1	10
S-SCAN+AC	8.6	33.9	59.8	96.0	1	10
S-SCAN	14.7	31.8	41.7	166.7	2	5
S-SCAN+AC	16.8	43.3	67.7	70.7	2	5
S-SCAN	33.5	56.2	65.9	59.0	4	3
S-SCAN+AC	34.1	61.4	80.1	37.8	4	3

Table 2: Results of Ask&Confirm on S-SCAN after 10 rounds. AC denotes the Ask&Confirm framework.

where $\gamma_{j,k} = \frac{exp(cos(x_j^T, x_{k,m}^I))}{\sum_{m=1}^{M} exp(cos(x_j^T, x_{k,m}^I))}$ and cos denotes the cosine similarity. The similarity between X^T and x_k^I is the average of $S_{j,k}$ among all x_j^T. (b) **T-CMPL**: Similar to CMPL, we adopt global alignment to match textual and visual features without any attention mechanisms. Thus, the similarity between x_j^T and x_k^I is

$$S_{j,k}(x_j^T, x_k^I) = cos(x_j^T, \frac{1}{M}\sum_{m=1}^{M} x_{k,m}^I) \tag{7}$$

The similarity between X^T and x_k^I is the average of $S_{j,k}$ among all x_j^T.

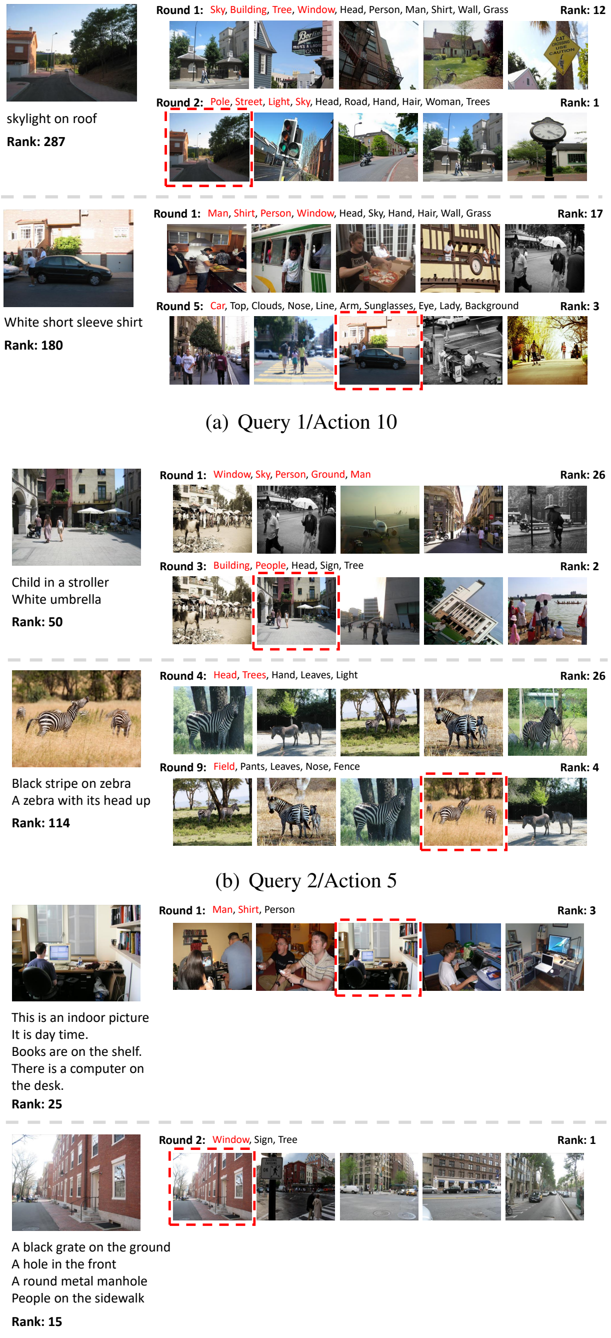
Both S-SCAN and T-CMPL are optimized with a common ranking loss. It is clear that Ask&Confirm focuses on the interactive mode and is independent of the network architecture and similarity computing. Thus, Ask&Confirm can adopt any existing cross-modal retrieval models.

Implementation Details. During training, T is set to 20 to conduct twenty-round interaction. In each round, we set $N_A = 10$ which means sampling 10 objects from the object sampling distribution $P(a)$. During testing, we vary T and N_A and apply a greedy sampling to choose objects with the highest probabilities. Similar to [1], we utilize a Faster RCNN pretrained on Visual Genome with 1600 object categories to extract features of the top 36 regions and predict objects of regions. Textual and visual features are mapped into vectors with a dimension of 256. For the optimization of policy learning, we update all parameters for every 600 rounds and adopt Adam [10] as the optimizer. Learning rates of ϕ_π, ϕ_v are $3e^{-4}$ and $1e^{-3}$. Coefficient α is set to 1000. All models are trained for 500 epochs.

Evaluation Metrics. We adopt the common R@K (K=1, 5, 10) metric and Mean Rank (MR) to measure the retrieval performance. R@K indicates the percentage of the queries where at least one ground truth is retrieved among the top-K candidates.

4.1. Results

Based on S-SCAN and T-CMPL, we build two interactive retrieval models with the proposed Ask&Confirm framework. To prove the effectiveness of Ask&Confirm, we test them in three settings: (1) Q1/A10, (2) Q2/A5, and (3) Q4/A3. QK means K queries are given by users in the beginning, and AK means K objects are provided by an agent in each round. All results are recorded after 10 rounds.


Round 1: Sky, Building, Tree, Window, Head, Person, Man, Shirt, Wall, Grass
Rank: 12
skylight on roof
Rank: 287
Round 2: Pole, Street, Light, Sky, Head, Road, Hand, Hair, Woman, Trees
Rank: 1
Round 1: Man, Shirt, Person, Window, Head, Sky, Hand, Hair, Wall, Grass
Rank: 17
White short sleeve shirt
Rank: 180
Round 5: Car, Top, Clouds, Nose, Line, Arm, Sunglasses, Eye, Lady, Background
Rank: 3

(a) Query 1/Action 10

Round 1: Window, Sky, Person, Ground, Man
Rank: 26
Child in a stroller
White umbrella
Rank: 50
Round 3: Building, People, Head, Sign, Tree
Rank: 2
Round 4: Head, Trees, Hand, Leaves, Light
Rank: 26
Black stripe on zebra
A zebra with its head up
Rank: 114
Round 9: Field, Pants, Leaves, Nose, Fence
Rank: 4

(b) Query 2/Action 5

Round 1: Man, Shirt, Person
Rank: 3
This is an indoor picture
It is day time.
Books are on the shelf.
There is a computer on
the desk.
Rank: 25
Round 2: Window, Sign, Tree
Rank: 1
A black grate on the ground
A hole in the front
A round metal manhole
People on the sidewalk
Rank: 15

(c) Query 4/Action 3

Figure 4: Visualization of Ask&Confirm based on S-SCAN. We show examples in three settings. Positive objects in each round are highlighted in red. The target image is surrounded with a red bounding box.

Results are illustrated in Table 2 and 3. For both basic retrieval models in three different settings, Ask&Confirm strengthens their performance in all evaluation metrics. Ask&Confirm enhances R@10 of S-SCAN from 20.4% to 59.8% and strengthens R@10 of T-CMPL by 20.7% with Q1/A10. In the other two settings, the advantage of R@10 brought from Ask&Confirm recedes a bit but at

Method	R@1	R@5	R@10	MR	Q	A
T-CMPL	3.1	10.5	16.3	593.4	1	10
T-CMPL+AC	5.2	20.4	37.0	313.8	1	10
T-CMPL	7.3	19.5	28.3	283.3	2	5
T-CMPL+AC	8.6	26.9	47.2	211.3	2	5
T-CMPL	14.5	33.5	44.0	118.2	4	3
T-CMPL+AC	15.1	38.6	59.5	98.7	4	3

Table 3: Results of Ask&Confirm on T-CMPL after 10 rounds. AC denotes the Ask&Confirm framework.

least achieves 14.2%. As for R@5, Ask&Confirm based on S-SCAN achieves 61.4% and the one based on T-CMPL achieves 38.6% with Q4/A3. In other settings, the enhancement of Ask&Confirm is more obvious and even achieves 11.5% with Q2/A5 based on S-SCAN. Both basic retrieval models are improved by Ask&Confirm of R@1 in all settings. In particular, Ask&Confirm based on S-SCAN achieves R@1=34.1% with Q4/A3. With Ask&Confirm, MR of both basic retrieval models in three settings is moved up by a large margin. These results demonstrate the effectiveness of Ask&Confirm.

4.2. Visualizations

Examples of interactive retrieval are shown in Figure 4. Several interesting discoveries are found out in visualizations. Firstly, the agent tends to offer several objects in the first few round regularly, such as “window”, “man”, “sky”, “head”, “tree” and so on. These are the objects that come up most frequently in Visual Genome This is a reasonable choice because it either has a large possibility to add a ground-truth object to queries or eliminates plenty of images that include these objects. Secondly, the agent can offer objects that are not common but related to the semantics of given queries and images in latter rounds. For example, to retrieve the image that includes zebras, the agent offers “field” and “fence” in round 9 which rarely occur but are related to zebras. To retrieve the image with a query “White short sleeve shirt”, the agent offers “sunglasses” and “top” in round 5 which belong to clothing just like the query, and offers “car” which shows in the image. We ascribe these properties to our policy learning approach. The statistic-based shaping guides the agent to give priority to the most frequent objects and the reinforcement learning promotes objects related to the semantics of images.

4.3. Ablation Studies

Number of Query and Action. To verify Ask&Confirm is robust to the number of queries and actions, we test it based on S-SCAN with different query numbers $N_Q^1 \in \{1, 2, 4\}$ where users input 1, 2, or 4 queries and different action numbers $N_A \in \{3, 5, 10\}$ where the agent provides 3, 5, or 10 object candidates in each round. Results on R@5, R@10, and MR in each round are shown in Figure 5.

In detail, with the same queries, the performance of Ask&Confirm gradually improves when N_A increases, which shows that more actions in each round facilitate the retrieval. On the other hand, when N_A is fixed, queries with higher $N_Q^1 = 4$ outperform the ones with lower N_Q^1, which is consistent with our discovery in Figure 2. Although models with fewer queries achieve worse performance, improvements over them are even more. Especially, when $N_Q^1 = 1$ and $N_A = 10$, Ask&Confirm achieves the largest improvement. We conclude that fewer queries leave more space for the agent to optimize the basic model’s retrieval. Despite the change of the number of queries and actions, Ask&Confirm consistently enhances S-SCAN on all metrics. It examines the robustness of Ask&Confirm which facilitates retrieval stably in all situations.

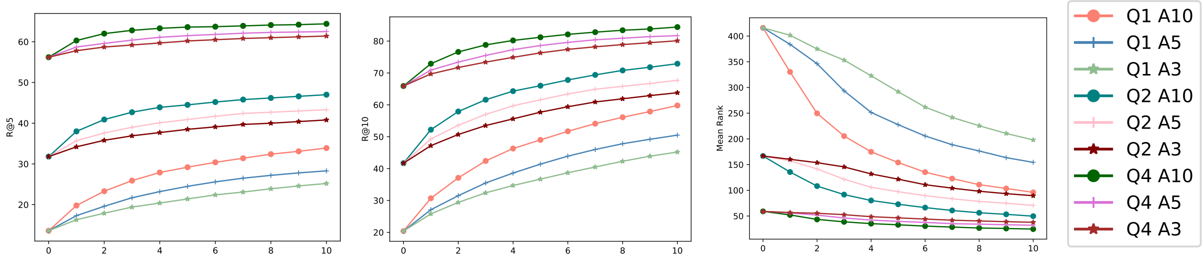
Policy. Finding a policy that guides the agent to choose discriminative objects is essential to Ask&Confirm. As a result, we compare our policy learning method with three pre-defined policies: (1) **Random**: In each round, the agent samples objects from a uniform distribution. (2) **QASim**: Inspired by [18], objects that have similar textual features with a query are preferred. We use cosine similarity between textual features of queries and objects to indicates their similarity. (3) **QACohe**: Considering that some objects tend to occur coherently, such as “building” and “window”, we compute a joint distribution $P_c(a_i, a_j)$ in the train split, where a_i and a_j are in the same image. Then, we use $P_c(a^*, a_j)$ where $a^* = \underset{a \in \mathcal{A}}{\text{argmax}} \frac{1}{N_Q^t} \sum_{n=1}^{N_Q^t} cos(x_n^T, TE(\mathbb{T}(a)))$ to sample objects.

Experiments based on S-SCAN are conducted in three settings just like Section 4.1. As shown in Figure 6, under all settings, the proposed policy learning outperforms the other by a large margin in terms of R@10. After 10 rounds, our policy learning strategy outperforms the second-best policy by 12.1%, 7.2%, and 5.0% in three settings. We also observe that a good policy increases R@10 rapidly in the first several rounds and slows down in subsequent rounds. Such a policy provides better interactive experiences because users retrieve the target image with fewer interactions.

Model Agnostic. By comparing the improvements on S-SCAN and T-CMPL as shown in Table 4, we examine that the proposed framework is model-agnostic. Although the implementation and performance of T-CMPL and S-SCAN are different, Ask&Confirm strengthens both of them on all evaluation metrics. In detail, the two models’ improvements of MR are very close. As for R@K metrics, improvements are more obvious on S-SCAN due to its better original performance. These results demonstrate that Ask&Confirm can easily cooperate with a common text-based retrieval model to boost the retrieval performance.

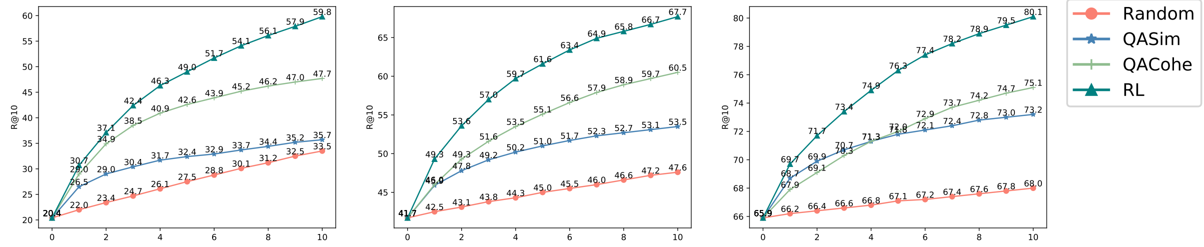
4.4. User Study

To demonstrate the advantage of the active object-based interaction over tag-based and description-based interac-



(a) R@5 (b) R@10 (c) Mean (d) Setting

Figure 5: Results of Ask&Confirm based on S-SCAN. The horizontal axis represents the query turn. Q denotes the number of queries and A denotes the action number in each round.



(a) Query 1/Action 10 (b) Query 2/Action 5 (c) Query 4/Action 3 (d) Policy

Figure 6: Results of different policies.The horizontal axis represents the query turn. The vertical axis represents R@10. The proposed RL-based policy learning approach outperforms others.

Method	R@1	R@5	R@10	MR	Q	A
T-CMPL+AC	+1.9	+9.9	+10.7	-279.6	1	10
S-SCAN+AC	+4.1	+20.3	+39.4	-320.0	1	10
T-CMPL+AC	+1.3	+7.4	+8.9	-72.0	2	5
S-SCAN+AC	+2.1	+11.5	+26.0	-96.0	2	5
T-CMPL+AC	+0.6	+5.1	+15.5	-19.5	4	3
S-SCAN+AC	+0.6	+15.2	+14.2	-21.1	4	3

Table 4: Results of Ask&Confirm on different basic retrieval models after 10 rounds.

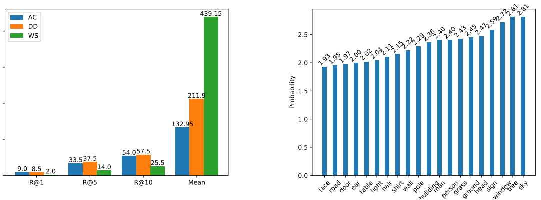
tion, we compare Ask&Confirm (AC) with Drill-Down (DD) [29] and WhittleSearch (WS) [13] where DD is a description-based method and WS is a tag-based method. To make a fair comparison of interactive mode, we re-implement DD and WS based on S-SCAN and adopt their interactive mode. 50 images are sampled from the test set. For each image, 4 different users (details in supplementary) are required to retrieve it in 5 rounds with 3 different methods. The retrieval performance in terms of R@1, R@5, R@10 and Mean Rank (Mean) are shown in Figure 7 (a).

To evaluate users' effort on different methods, we record the average time users take to retrieve each image. AC costs **37.67s**, DD costs **53.60s** and WS costs **35.18s**.

Conclusion on Performance: Ask&Confirm achieves similar R@k accuracy and much better Mean Rank compared to DD. Ask&Confirm significantly outperforms WS.

Conclusion on User Effort: Ask&Confirm takes significantly less time to complete the retrieval compared to DD and takes similar time compared to WS.

Overall, Ask&Confirm achieves similar performance with description-based interaction and similar retrieval time with tag-based interaction. It examines that Ask&Confirm not only achieves a friendly user experience, but also achieves excellent retrieval performance. Furthermore, Figure 7 (b) shows the percentage of objects provided by Ask&Confirm in the user study. It demonstrates what the RL-based policy learns from the image gallery.



(a) User Study (b) Object Distribution

Figure 7: (a) User study. AC denotes Ask&Confirm. (b) Object distribution during the user study.

5. Conclusion

We firstly introduce the partial-query problem that easily makes cross-modal retrieval models collapsed and propose Ask&Confirm, an interactive retrieval framework, to tackle this problem. Ask&Confirm heuristically guides users to enrich details of images by actively searching for discriminative objects of the target image for users to confirm. A weakly-supervised RL-based policy is proposed to conduct the active search, which leverages the characteristics of the image gallery. Experimental results demonstrate the effectiveness and robustness of Ask&Confirm. The weakly-supervised training procedure also makes it more practical than other dialog-based retrieval models.

6. Acknowledgement

This work is supported by Joint Funds of the National Science Foundation of China under Grant U18092006, the Shanghai Municipal Science and Technology Committee of Shanghai Outstanding Academic Leaders Plan under Grant 19XD1434000, the Projects of International Cooperation of Shanghai Municipal Science and Technology Committee under Grant 19490712800, the National Natural Science Foundation of China under Grant 61772369, Grant 61773166, Grant 61771144, National Key R&D Program of China under Grant 2020YFA0711400, Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100), Shanghai Municipal Commission of Science and Technology Project (19511132101), the Changjiang Scholars Program of China, and the Fundamental Research Funds for the Central Universities.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[2] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.

[3] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2951–2960, 2017.

[4] Maaike de Boer, Klamer Schutte, and Wessel Kraaij. Knowledge based query expansion in complex multimedia event detection. *Multimedia Tools and Applications*, 75(15):9025–9043, 2016.

[5] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. 2018.

[6] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, et al. Query by image and video content: The qbic system. *Computer*, 28(9):23–32, 1995.

[7] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In *Advances in Neural Information Processing Systems*, pages 678–688, 2018.

[8] Yuanfeng He, Yuanxi Li, Jiajia Lei, and Clement HC Leung. A framework of query expansion for image retrieval based on knowledge base and concept similarity. *Neurocomputing*, 204:26–32, 2016.

[9] Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. Saliency-guided attention network for image-sentence matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5754–5763, 2019.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Adriana Kovashka and Kristen Grauman. Attribute pivots for guiding relevance feedback in image search. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 297–304, 2013.

[12] Adriana Kovashka and Kristen Grauman. Attributes for image retrieval. In *Visual Attributes*, pages 89–117. Springer, 2017.

[13] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2973–2980, 2012.

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[15] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[16] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 801–809, 2018.

[17] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 3–11, 2019.

[18] Yongli Liu, Chao Li, Pin Zhang, and Zhang Xiong. A query expansion algorithm based on phrases semantic similarity. In *2008 International Symposiums on Information Processing*, pages 31–35, 2008.

[19] Sho Maeoki, Kohei Uehara, and Tatsuya Harada. Interactive video retrieval with dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 952–953, 2020.

[20] Nils Murrugarra-Llerena and Adriana Kovashka. Image retrieval with mixed initiative and multimodal feedback. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 310. BMVA Press, 2018.

[21] Nils Murrugarra-Llerena and Adriana Kovashka. Image retrieval with mixed initiative and multimodal feedback. *Computer Vision and Image Understanding*, 207:103204, 2021.

[22] Apostol Natsev, Alexander Haubold, Jelena Tešić, Lexing Xie, and Rong Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 991–1000, 2007.

[23] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pages 278–287, 1999.

[24] Aishwarya Padmakumar and Raymond J Mooney. Dialog policy learning for joint clarification and active learning queries. *arXiv preprint arXiv:2006.05456*, 2020.

[25] Devi Parikh and Kristen Grauman. Relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 503–510, 2011.

[26] Yong Rui, Thomas S Huang, Michael Ortega, and Sharad Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

[27] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[28] Sarah Shomstein and Steven Yantis. Control of attention shifts between vision and audition in human cortex. *Journal of Neuroscience*, 24(47):10702–10706, 2004.

[29] Fuwen Tan, Paola Cascante-Bonilla, Xiaoxiao Guo, Hui Wu, Song Feng, and Vicente Ordonez. Drill-down: Interactive retrieval of complex scenes using natural language queries. In *Advances in Neural Information Processing Systems*, pages 2651–2661, 2019.

[30] Alex HC van der Heijden. *Selective Attention in Vision*. Routledge, 2003.

[31] 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 Conference on Computer Vision and Pattern Recognition*, pages 6439–6448, 2019.

[32] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3792–3798. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[33] Hong Wu, Hanqing Lu, and Songde Ma. Willhunter: Interactive image retrieval with multilevel relevance. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 1009–1012, 2004.

[34] Xinru Yang, Haozhi Qi, Mingyang Li, and Alexander Hauptmann. From a glance to" gotcha": Interactive facial image retrieval with progressive relevance feedback. *arXiv preprint arXiv:2007.15683*, 2020.

[35] Aron Yu and Kristen Grauman. Fine-grained comparisons with attributes. In *Visual Attributes*, pages 119–154. Springer, 2017.

[36] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 686–701, 2018.

[37] Z. Zhu, J. Xu, X. Ren, Y. Tian, and L. Li. Query expansion based on a personalized web search model. In *Third International Conference on Semantics, Knowledge and Grid (SKG 2007)*, pages 128–133, 2007.