We thanks reviewers for their feedback. First, we address some common concerns.

Disparity between the region caption based queries in our experiments and the user queries in real scenario.
As emphasized in our manuscript (L53), leveraging region captions for weak supervision can been seen as one of the advantages of our method, as no annotation is required. While we agree strong supervisory signals such as real user queries could bridge the domain gap and would like to explore further in this direction, we choose at this stage to use only "weak but free" signals to see to what extent they can be generalized to practical applications. We demonstrate the effectiveness of the proposed method on simulated data with extensive experiments, on real scenarios via human subject studies on Amazon Mechanical Turk. Our experiments show the region caption based queries can potentially be generalized to real scenarios with high promise.

Patterns of real user queries
Compared with the simulated caption data, the real user queries we collected show distinct patterns, e.g. long descriptions of the target scenes, less informative sentences (e.g. 'good!'). That's why we observe smaller performance gaps between our method and the alternative approaches in the user study. We will include more details of the user study and examples of the real user queries in the revised paper.

Reviewer 1

#Q2: **Baselines**
With relatively less prior works on this research, we compare the proposed method with variants of state-of-the-art approaches for the most related topics, e.g. dialog-based interactive product search. We will incorporate the suggested baselines (e.g. late score/rank fusion vs early feature fusion, linear query encoding vs hierarchical query encoding) in the revised paper.

#Q4: **Applications of the proposed method**
We envision the proposed method could generally help with natural image search. Potential applications include retrieving very specific images of complex scenes the users encountered before, or exploring inspiring images for creative content generation (e.g. Adobe Stock Image).

Reviewer 2

#Q1: **Incorporating advanced language models**
We have explored using bidirectional language encoders and found it performs similar with unidirectional encoders in this task. We conjecture that unidirectional and bidirectional encoders provide comparable contextual signals when encoding the per-turn query as a single feature vector for downstream modules. In the current manuscript, we focus more on the sequential encoding of multiple sentences, and would like to explore and incorporate more advanced language models such as BERT in the future.

Reviewer 3

#Q1: **Distinguishing the paper's contribution with memory networks**
In contrast to the previous sentence encoding methods which perform **query** and possibly **update** operations on a predefined external memory space (e.g. the agenda items in Kiddon et al. 2016, neural checklist models), we focus on a more challenging scenario where the model needs to **create** and **update** the memory module (the state vectors in our case) **on-the-fly** so as to maintain the dynamic states of multiple-turn queries. We will elaborate more to distinguish our method with memory networks in the revised paper.

#Q2: **Experimental details**
(1) The region captions and their orders are randomly sampled. We keep the captions and their orders of the validation and test sets unchanged for all our experiments; (2) We use ten turns in all our simulated experiments as we'd like to track and demonstrate the performance of the proposed method in both short-term and long-term scenarios, as shown in Fig. 3. In the user study, we start with ten-turn queries but observe the users are less willing to continue and finish the tasks if they could not succeed in five turns, so we evaluate the five-turn queries in our experiment; (3) We use different image sets for training, validation and evaluation (L208), where the images retrieved are from the corresponding sets at different stages respectively. All the evaluations (including the user study) are performed on the test set, which contains 9896 images (L208); (4) All images in the candidate set are ranked in each turn. (5) Faster RCNN is NOT finetuned in our experiments.

#Q3: **Number of the placeholders**
We confirm in our experiments that more placeholders result in better performance but also lead to scalability difficulties. One of the main goal of the proposed work is to model multiple-turn queries with dynamic lengths using a fixed set of hidden states (fixed computional budget accordingly). We're also happy to include the suggested experiments and improve the presentation of the paper in the revised version.

#Q4: **Questions about the human evaluation**
As answered in $\#Q2$, we start experimenting with ten queries but discover it cannot fit in the real scenario. We agree that using less placeholders will be more convincing in this case and will rerun the experiment in the revised paper.