







Maybe you are looking for CroQS 🐸

Cross-modal Query Suggestion for Text-to-Image Retrieval

Giacomo Pacini ^{*1,2} , Fabio Carrara¹ , Nicola Messina¹ ,
Nicola Tonellotto^{1,2} , Giuseppe Amato¹ , and Fabrizio Falchi¹ 

¹ CNR-ISTI, Pisa, Italy - {name.surname}@isti.cnr.it

² University of Pisa, Italy - {name.surname}@unipi.it

Abstract. Query suggestion, a technique widely adopted in information retrieval, enhances system interactivity and the browsing experience of document collections. In cross-modal retrieval, many works have focused on retrieving relevant items from natural language queries, while few have explored query suggestion solutions. In this work, we address query suggestion in cross-modal retrieval, introducing a novel task that focuses on suggesting minimal textual modifications needed to explore visually consistent subsets of the collection, following the premise of “Maybe you are looking for”. To facilitate the evaluation and development of methods, we present a tailored benchmark named CroQS. This dataset comprises initial queries, grouped result sets, and human-defined suggested queries for each group. We establish dedicated metrics to rigorously evaluate the performance of various methods on this task, measuring representativeness, cluster specificity, and similarity of the suggested queries to the original ones. Baseline methods from related fields, such as image captioning and content summarization, are adapted for this task to provide reference performance scores. Although relatively far from human performance, our experiments reveal that both LLM-based and captioning-based methods achieve competitive results on CroQS, improving the recall on cluster specificity by more than 115% and representativeness mAP by more than 52% with respect to the initial query. The dataset, the implementation of the baseline methods and the notebooks containing our experiments are available here: paciosoft.com/CroQS-benchmark/

Keywords: Text-To-Image Retrieval · Query Suggestion · Cross-modal Retrieval · Image Group Captioning · Cross-modal Query Suggestion.

1 Introduction

Query formulation is a central challenge in Information Retrieval (IR), as it consists of aligning user queries with their actual information needs. Query expansion and query suggestion are two families of query reformulation approaches

* Corresponding Author. Email: giacomo.pacini@isti.cnr.it

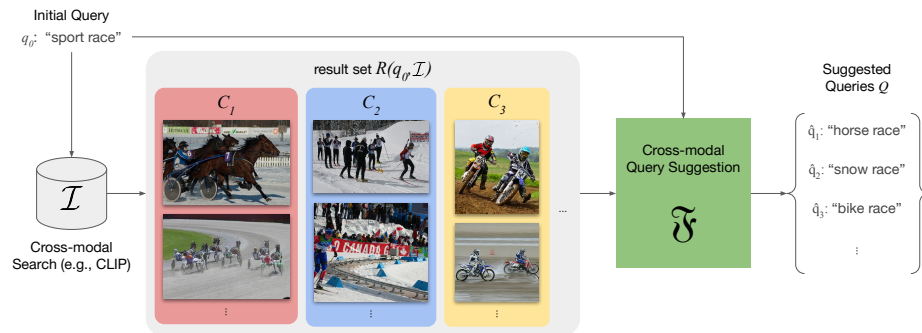


Fig. 1: **Cross-modal Query Suggestion.** Given an initial query q_0 and an image collection \mathcal{I} , a cross-modal query suggestion system \mathcal{F} returns a set of query suggestions \mathcal{Q} based on the visual content of the result set $R(q_0, \mathcal{I})$. Ideally, each suggestion $\hat{q}_i \in \mathcal{Q}$ should represent a semantically coherent group $C_i \subset R(q_0, \mathcal{I})$.

that aim to better align the search results with the user’s information need [19]. The first one is generally automatic and modifies the IR system’s internal representation of the original query. Query suggestion, instead, is an interactive approach that proposes different queries to the user to refine the search. Those techniques either base their reformulations on the initial search results — as happens in Rocchio’s algorithm [23] with its Pseudo Relevance Feedback (PRF) — or on external knowledge sources, such as thesauri, ontologies, or large-scale language models. In the case of PRF, the system uses feedback from the initial set of retrieved documents to automatically adjust the query, while knowledge-based approaches leverage pre-defined relationships between terms to suggest more effective formulations.

Historically, IR and query reformulation approaches have been designed for document retrieval, where a textual query is employed to retrieve textual documents. Therefore, methods, architectures, and datasets that define and solve this task mainly consider the text modality alone. Recently, *cross-modal retrieval* and, in particular, *text-to-image retrieval* is becoming the key to searching large image and video datasets, thanks to the recent development of powerful models that embed images and texts in the same semantic feature space [16,15,21]. In particular, CLIP [21] has simplified research in this area and enabled further possibilities. It embeds text and visual data into the same embedding space and can easily retrieve the most relevant images or videos by computing a maximum inner-product search with the textual query embedding. This model is nowadays the de facto standard in multimedia retrieval, and it has been largely employed in interactive video browsing competitions [13,1,7].

Given the growing interest in cross-modal search applications, there is a pressing need to develop a tool that interactively suggests queries in cross-modal scenarios — an area that is under-explored in the literature. For this reason, in

this work, we aim to explore the query suggestion task in the scenario where queries are provided in natural language, and the search database comprises a potentially large set of unannotated images. We refer to this task as *cross-modal query suggestion*, which consists of proposing suitable modifications to the initial user query based on visual clues present in the retrieved images. Ideally, each suggestion should represent different aspects of the original result set and allow the user to interactively understand the semantic groups of images that populate the search collection. Figure 1 shows an overview of the task with an example.

The proposed task is challenging for many reasons. In a cross-modal retrieval scenario, document collections are usually made of plain images that do not carry captions, textual descriptions, or any additional metadata of the documents. Furthermore, the information to extract from an image document can change depending on the context. For this reason, textual features statically provided with the images — such as the captions of image-caption datasets — are not effective for the query suggestion task since they lose some of the possible *meanings* of an image. To solve this problem, it is necessary to define a mechanism that, given an initial query and a semantic representation of the result images, constructs a suggested query. Considering the novelty of this problem in cross-modal retrieval, it is also necessary to define a method and a benchmark for method comparison. To this aim, we introduce the CroQS benchmark. We developed CroQS through a semi-automatic, human-supervised process. We generated every semantic cluster of images from an initial query and a clustering step, further tuning the obtained sets through careful human judgment. A human annotator also defined a reference query suggestion per group to have a solid ground truth. In order to objectively evaluate different methods with the benchmark, we defined a set of metrics, and we adapted several baseline methods from related fields — such as image captioning and content summarization — to solve the task.

To summarize, the contributions of this paper are threefold:

- We define and tackle the novel task of cross-modal query suggestion, which extends classical text-based query suggestion by directly incorporating latent high-level information extracted from the retrieved images.
- We provide the CroQS benchmark, accompanied by proper evaluation metrics, to quantitatively evaluate models on this new challenging task.
- We propose some reasonable baselines for this benchmark by adapting some captioning-based and LLM-based methods.

Our results show that the methods derived from captioning are more effective in building queries that are specific to their group of images, while LLM-based approaches are balanced throughout the properties and tend to build suggestions very similar to the initial queries.

2 Related Works

Cross-modal query suggestion is a task spanning different research domains. Besides classical IR query suggestion, this task borrows from methods extracting

image content as textual descriptions, thus relating to image group captioning and image captioning in general. While image group captioning is about building a natural language sentence that describes a set of images, cross-modal query suggestion is about generating a more detailed query that better explains a group of query results.

2.1 Query Suggestion

In dense retrieval [9], query reformulation has a lower impact in improving retrieval metrics, while it can still be exploited to improve retrieval systems’ interactivity and explorability. Query suggestion provides to the user several queries that the system deems related to the user’s interest according to assumptions made by the retrieval system [19]. Various works proposed solutions to build query suggestions based on large-scale graphs of queries and their click-through rates [5,14]. More recent approaches for solving this task in the textual domain employed RNN [28] and transformer networks [18] or leveraged different learning schemas such as reinforcement learning [4]. Nowadays, the interest is shifting towards the use of Large Language Models (LLMs) [3,2], which show remarkable abilities in suggesting queries given the user intention or the search history as a prompt. The work mostly resembling our objective is the one in Zha et al. [30], which proposed Visual Query Suggestion, a system for collections of captioned images that groups the result set of an initial query and suggests a reformulated query for each group. However, they built the queries from the texts associated with the images, leveraging a keyword selection algorithm. In contrast, our scenario assumes no textual metadata is available for the images. Zeng et al. [29] proposed a method to build query suggestions in a cross-modal retrieval setting by creating a hierarchy of categories (12 high-level categories and 500 low-level nouns) employed as the statically defined terms for the query suggestions. While having a similar objective to ours, this approach strongly depends on (and is limited to) the static pre-defined categories.

2.2 Image Group Captioning

The choice of image captioning as a foundational field is motivated by several factors. In cross-modal query suggestion, the goal is to generate a suggested query that effectively represents a set of images while maintaining alignment with the initial query. Similarly, image captioning methods aim to generate a descriptive sentence for a given image. Although these two tasks are not identical, they share important commonalities in their input-output structures. Both tasks start from visual inputs — images — and generate textual outputs. This overlap suggests that techniques developed for image captioning may be adapted to handle cross-modal query suggestion. The advent of CLIP [21] introduced the possibility of representing texts and images in a cross-modal space and paved the way for various works in the image captioning field exploiting it. Mokady et al. introduced ClipCap [17], an architecture that, given a CLIP image embedding, generates a caption for it through the use of a mapping network and a large

language model. In 2023, Li et al. [10] introduced DeCap, which has the same goal but leverages a simpler architecture and achieves higher captioning results.

Wang et al. proposed a method to build image captions optimizing the *distinctive* aspects of the input image with respect to the characteristics of a group of similar images [25,26,27], but they still focus on captioning a single image. The method proposed by Phueaksri et al. [20] aims at summarizing a collection of images through their scene graphs and an external knowledge graph into a scene graph of the whole collection.

N. Trieu et al. [24] proposed a transformer model for *multi-image summarization* captioning a set of similar images in the specific domain of e-commerce product images. Also Li et al. [11] studied image group captioning and proposed a method that contrastively builds a caption for a target group of images comparing against an additional *context group* of images, which serves as a reference point to distinguish the target group’s unique characteristics. However, those methods do not consider an initial text (or query) to condition the generation of the caption as would be needed in cross-modal query suggestion.

3 Cross-Modal Query Suggestion

3.1 Problem Definition

Let $q_0 \in \mathcal{T}$ be the initial query prompted by the user in natural language (being \mathcal{T} the set of text strings) and $R(q_0, \mathcal{I}) \subset \mathcal{I}$ be the result set obtained by searching \mathcal{I} with q_0 , where $\mathcal{I} = \{I_i\}_{i=1}^D$ is the image collection. Indicating with $2^{\mathcal{X}}$ the power set of \mathcal{X} (the set of subsets of \mathcal{X}), we define a cross-modal query suggestion system $\mathfrak{F} : \mathcal{T} \times 2^{\mathcal{I}} \rightarrow 2^{\mathcal{T}}$ as

$$\mathcal{Q} = \mathfrak{F}(q_0, R(q_0, \mathcal{I})), \tag{1}$$

where $\mathcal{Q} = \{\hat{q}_i\}_{i=1}^N \subset \mathcal{T}$ is the set of suggested queries, the output of the system. All the suggested queries \hat{q}_i should be *variations of the same initial query* q_0 , oriented to disambiguate and better explain the initial result set $R(q_0, \mathcal{I})$.

3.2 Proposed Framework

The core of the cross-modal query suggestion task revolves around formulating suggested queries based on the original query and the main concepts represented in the original result set. We assume that an initial result set $R(q_0, \mathcal{I})$ contains one or more disjoint groups $\mathcal{C} = \{C_i\}_{i=1}^M \subseteq R(q_0, \mathcal{I}), C_i \cap C_j = \emptyset \forall i \neq j$ where each C_i comprises images sharing semantic content with higher specificity (e.g., “a bike race” or “a skiing race”) than the original query q_0 (e.g., “a sport race”).

In the proposed framework, the process begins with the user submitting an original query q_0 . Then, the result set $R(q_0, \mathcal{I})$ provided by the retrieval system is partitioned into M distinct semantic groups C_i by a clustering algorithm. At this point, the cross-modal query suggestion system is tasked to generate a suggested query \hat{q}_i for each group C_i . The suggestion system runs M times,

receiving as input q_0 and one semantic group at a time. Finally, a *user interface* presents the set of suggested queries $\{\hat{q}_i\}_{i=1}^M$ to the user for further exploration.

Under this assumption, we simplify the formulation as

$$\hat{q}_i = \mathfrak{F}(q_0, C_i), \quad (2)$$

where the cross-modal query suggestion system $\mathfrak{F} : \mathcal{T} \times 2^{\mathcal{I}} \rightarrow \mathcal{T}$ now operates on a set of semantically coherent images C_i and produces a single suggested query \hat{q}_i for that set. We delegate the partitioning of the initial result set $R(q_0, \mathcal{I})$ into different semantic groups $\{C_1, C_2, \dots, C_M\}$ to off-the-shelf clustering algorithms operating on the semantic image representations of the collection and focus our investigation on how to implement \mathfrak{F} .

The proposed methodology, therefore, operates only after the semantic clusters have been established, focusing exclusively on generating refined queries based on the original input and the external clustering system’s output.

Whereas the proposed framework applies to any cross-modal representation space, in our experiments, we used CLIP [21] to build the semantic representations of the images of the tested collection and to perform cross-modal searches. Thanks to the properties of the CLIP space, image search results can be sorted by decreasing cosine similarity to the textual query. We used the same image representations extracted from CLIP to also build the semantic clusters.

In the following, we introduce the CroQS benchmark to evaluate different query suggestion methods, and then we present a set of baseline methods.

4 The CroQS Benchmark

4.1 Dataset

Defining an objective way to compare different cross-modal query suggestion methods is not trivial. One of the most critical variables for a fair comparison is the definition of the clusters \mathcal{C} for the given result set of images $R(q_0, \mathcal{I})$. Two systems could come up with completely different and perhaps equally valid clusters, leading to difficult comparability of the results.

Consequently, through manual work of collection exploration and query definition, we built a benchmark for cross-modal query suggestion, named CroQS, to factor out cluster variability in evaluating the generation of suggested queries. To develop CroQS, we selected a pool of annotators and assigned one initial query to each of them without overlap. CroQS comprises 50 initial textual queries. A group of expert judges chose the initial queries, evaluating the variability and the level of ambiguity in the associated result sets. The initial queries belong to various categories, i.e., 6 are about *animals*, 17 about *people and actions*, 10 about *objects*, 14 about *places*, and 3 are from other contexts. For each query, the top 200 images in the result set were manually validated and partitioned into a varying number of semantic clusters of images. When a set of images could not be attributed to any distinguishable cluster of the initial query, it was labeled as noise and discarded. Each validated cluster is associated with a

human-annotated suggestion for the target refined query. We built CroQS on top of the *train split* of COCO [12], a dataset of more than 118,000 captioned images. The image captions provided by COCO only assisted us in defining the human-annotated suggestions and were otherwise discarded for our purposes. Figure 3 shows samples from CroQS, and Table 1 summarizes its statistics.

Employing CroQS, it is possible to compare different methods only by means of the suggested queries they generate. In the following, we present the properties to be evaluated and a set of metrics that we can exploit to measure them.

Table 1: Statistics of the CroQS benchmark

# Queries	# Images	Images per Query	# Clusters	Clusters per Query
		(min/avg/max)		(min/avg/max)
50	8929	109/178.6/200	295	2/5.9/10

4.2 Evaluation Metrics

Due to the novelty of the cross-modal query suggestion task, defining the properties to be pursued is also necessary. We distinguished three main desired properties of the generated suggested queries: **Cluster Specificity**, **Representativeness**, and **Similarity to Original Query**.

Cluster Specificity measures how the suggested query \hat{q}_i describes the images belonging to its cluster C_i . In particular, it must highlight the properties that distinguish the images in C_i from those in other clusters $C_j, i \neq j$. To measure it, we quantify how well the suggested query \hat{q}_i recalls the element of C_i among the initial result set $R(q_0, \mathcal{I})$. Specifically, we compute the following:

$$\text{Recall}_{\text{Cluster}} = \frac{|C_i \cap R_{[:k]}(\hat{q}_i, R(q_0, \mathcal{I}))|}{k}, \quad (3)$$

where C_i acts as the set of relevant documents, $R_{[:k]}(\hat{q}_i, R(q_0, \mathcal{I}))$ are the top-k elements of $R(q_0, \mathcal{I})$ when sorted by the suggested query \hat{q}_i and act as the search collection, and $k = |C_i|$. A $\text{Recall}_{\text{Cluster}}$ of 1 is obtained when all elements of C_i are ranked before images belonging to other clusters.

Representativeness of the suggested query \hat{q}_i for a group of results C_i indicates how much \hat{q}_i better represents the images in C_i than the original query q_0 . To measure Representativeness, we look at how results in C_i are ranked when searching with the suggested query \hat{q}_i over the whole collection \mathcal{I} . We compute standard retrieval metrics, i.e., Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG), and Recall, considering the images belonging

to the starting group of images C_i as relevant.

$$\text{Recall} = \frac{|C_i \cap R(\hat{q}_i, \mathcal{I})|}{|C_i|} \quad (4)$$

Then, while Representativeness Recall exclusively measures the ability of a method to retrieve the images of its cluster from the whole collection, NDCG and MAP also measure if those images are placed among the first results.

A suggested query can be specific but not representative, and vice versa. Consider the case where the query q_0 is "A dog running", the current cluster C_i contains images of *black* dogs running, and the result set $R(q_0, \mathcal{I})$ shows images of dogs running outside. A suggested query \hat{q}_i like "black" could be specific but not representative. Conversely, if \hat{q}_i is "dog running outside", it would be representative but not specific at all.

Similarity to the Original Query measures how much the suggested query \hat{q}_i diverges from the original query q_0 , as the task requires suggesting corrections to a user-written query rather than re-writing it from scratch. It is interesting to evaluate both the syntactic and the semantic similarity of the suggested queries. For syntactic similarity, we adopted Jaccard score $\text{Jaccard}(q_0, \hat{q}_i)$, while for semantic similarity, we propose to adopt a similarity measure computed in an embedding space, such as the CLIP textual space $\text{CLIP}(q_0, \hat{q}_i)$. We refer to this latter metric as CLIP Query.

5 Baseline Methods

We propose two baseline methods to tackle the cross-modal query suggestion task inspired by related fields — image captioning and content summarization.

5.1 Prototype Captioning

As analyzed in Section 2, image captioning is closely related to cross-modal query suggestion but with significant differences that must be addressed. While traditional image captioning methods are designed to take a single image as input and generate a descriptive sentence, cross-modal query suggestion methods work with a set of images and aim to generate a single suggested query that encompasses all of them. To bridge this gap, we propose utilizing continuous and semantically meaningful representations of the images rather than relying on discrete descriptions of individual images. By representing a set of images in such a semantic space, we can compute a meaningful prototype point \hat{p}_i representing an entire group C_i . In particular, the CLIP’s image space [21] is able to place semantically similar images close, making it an ideal choice for this task.

Based on this approach, we selected two captioning models — **ClipCap** [17] and **DeCap** [10] — which are built on CLIP’s semantic space. These models can be fed with a prototype point from the CLIP space, without the need for

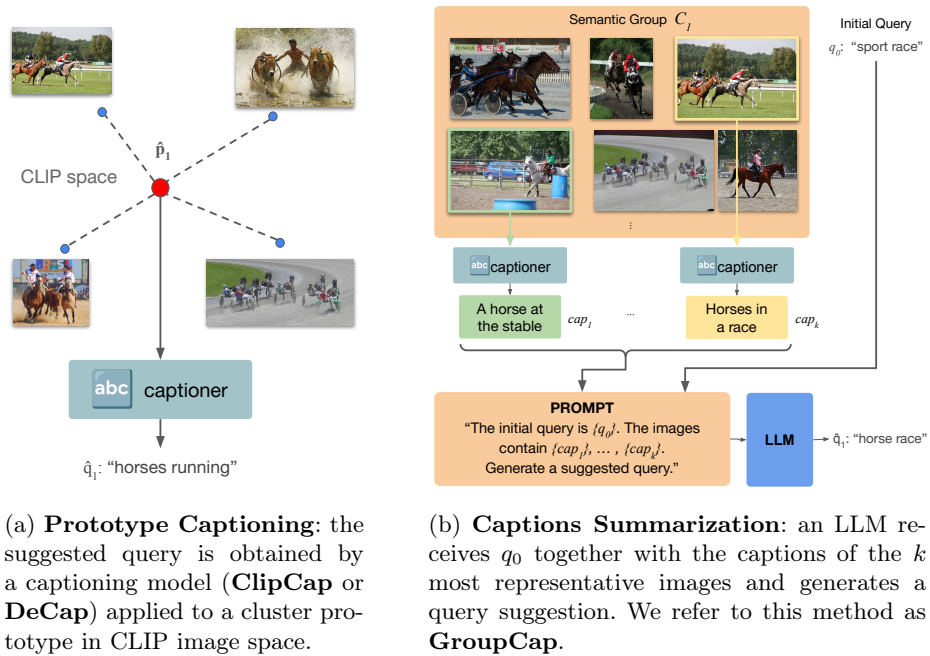


Fig. 2: Architectures of the baseline methods proposed.

an associated image, enabling them to generate text that is representative of the entire group of images. The core idea is to select a prototype point that is close enough to each image in the group, as in the example of Figure 2a. This prototype is then decoded to generate a representative text for the image group.

However, this strategy has a main limitation: off-the-shelf captioning models cannot be conditioned or guided by the initial query q_0 and provide an independent caption that usually deviates from q_0 . This is reflected in lower scores in terms of similarity to the original query. To overcome this limitation, we modified the ClipCap architecture to make it *query-aware*. As explained by Mokady et al. [17], ClipCap is made of two main components: a network that maps a CLIP image embedding to a set of GPT2 tokens, and a transformer decoder — GPT2 [22] in particular — which decodes that set of tokens to a natural language image caption. This two-step architecture allowed us not only to build a tailored input as explained above but also to provide the initial query q_0 to the second module of the architecture. We will refer to this model as ClipCap ^{q_0} .

5.2 Captions Summarization

Given the significant advancements in large language models (LLMs) and their remarkable ability to generate and understand text, we decided to explore how these models could be adapted for the cross-modal query suggestion problem.

We used a few-shot learning approach, allowing the model to grasp the task requirements and generate the desired text without needing extensive retraining.

To achieve this, we propose the **GroupCap** architecture shown in Figure 2b, which is designed to generate query suggestions through a series of steps. First, an algorithm selects the most relevant images from the cluster. These images are then processed by an image captioning method, which generates captions for each one. A carefully crafted prompt is created to instruct the LLM on how to use the captions and the initial query to generate a suggested query. The LLM then processes the formatted prompt and produces the final query suggestion. This approach leverages the ability of modern LLMs to effectively understand and generalize tasks using few-shot prompts, making it an ideal method for adapting LLMs to the cross-modal query suggestion task.

6 Experiments

6.1 Quantitative Results

This subsection reports the results obtained over our benchmark dataset by the adapted baseline methods. For reference, the scores obtained by the human-annotated query suggestions are also reported. Each score is computed as the Macro Average of the scores obtained on each cluster, for each initial query q_0 . Representativeness scores are calculated based on the top 100 documents of the result sets for the evaluated suggestions, and NDCG is assessed at rank 10.

Table 2: Comparison of the best cross-modal query suggestion configurations for each architecture. We report macro-averaged mean and standard deviation.

Method	Similarity to q_0		Cluster Specificity	Representativeness		
	CLIP	Jaccard	Recall Cluster	Recall	NDCG	MAP
q_0	1.00 \pm 0.00	1.00 \pm 0.00	0.19 \pm 0.07	0.52 \pm 0.03	0.18 \pm 0.05	0.21 \pm 0.05
human	0.87 \pm 0.03	0.44 \pm 0.11	0.59 \pm 0.13	0.62 \pm 0.12	0.52 \pm 0.14	0.45 \pm 0.11
ClipCap [17]	0.74 \pm 0.06	0.17 \pm 0.11	0.54 \pm 0.16	0.40 \pm 0.15	0.32 \pm 0.15	0.28 \pm 0.12
DeCap [10]	0.77 \pm 0.06	0.16 \pm 0.12	0.53 \pm 0.18	0.40 \pm 0.15	0.31 \pm 0.15	0.28 \pm 0.12
GroupCap	0.90 \pm 0.04	0.56 \pm 0.12	0.39 \pm 0.17	0.47 \pm 0.12	0.35 \pm 0.18	0.32 \pm 0.14

Table 2 reports the metrics obtained by the best configuration of each architecture. Human-annotated suggestions are getting the best results in all fields, except for similarity to the initial query, where the GroupCap model gets a higher average score. The methods based on DeCap and ClipCap are biased towards the Cluster Specificity metric, where those are getting the highest scores, while they rank last on the other metrics. This can be due to the original application field of these models, the captioning task, where the outputs need to be very specific and descriptive for the given image. This enabled them to effectively identify

the characteristics that distinguish one cluster from the others for the same initial query, leading to high scores in Cluster Specificity. On the other hand, the higher level of detail of the suggestions generated by these methods often causes semantic shifts from the initial query, lowering their Representativeness scores.

The baseline methods achieve lower Representativeness Recall scores compared to q_0 . By the way, it is important to note that the methods should find a trade-off between collection exploration and Representativeness.

Table 3: Comparison of different captioning-derived cross-modal query suggestion methods. Macro-averaged mean scores and associated standard deviations.

Method	Similarity to q_0		Cluster Specificity	Representativeness		
	CLIP	Jaccard	Recall Cluster	Recall	NDCG	MAP
DeCap _{centroid}	0.77 \pm 0.06	0.16 \pm 0.12	0.53 \pm 0.18	0.40 \pm 0.15	0.31 \pm 0.15	0.28 \pm 0.12
DeCap _{repr}	0.73 \pm 0.06	0.13 \pm 0.10	0.52 \pm 0.16	0.36 \pm 0.15	0.30 \pm 0.16	0.27 \pm 0.12
ClipCap _{centroid}	0.72 \pm 0.06	0.14 \pm 0.10	0.55 \pm 0.17	0.40 \pm 0.15	0.30 \pm 0.15	0.27 \pm 0.11
ClipCap _{repr}	0.70 \pm 0.06	0.12 \pm 0.11	0.53 \pm 0.15	0.35 \pm 0.14	0.31 \pm 0.15	0.27 \pm 0.11
ClipCap _{centroid} ^{q0}	0.74 \pm 0.06	0.17 \pm 0.11	0.54 \pm 0.16	0.40 \pm 0.15	0.32 \pm 0.15	0.28 \pm 0.12
ClipCap _{repr} ^{q0}	0.72 \pm 0.07	0.15 \pm 0.10	0.52 \pm 0.16	0.37 \pm 0.16	0.32 \pm 0.15	0.28 \pm 0.12

Table 4: Comparison of different GroupCap LLM-based cross-modal query suggestion methods. We report macro-averaged mean and standard deviation.

Method	Similarity to q_0		Cluster Specificity	Representativeness		
	CLIP	Jaccard	Recall Cluster	Recall	NDCG	MAP
GroupCap _{Mistral}	0.81 \pm 0.05	0.24 \pm 0.09	0.46 \pm 0.19	0.43 \pm 0.13	0.35 \pm 0.16	0.31 \pm 0.12
GroupCap _{LLama3}	0.90 \pm 0.04	0.56 \pm 0.12	0.39 \pm 0.17	0.47 \pm 0.12	0.35 \pm 0.18	0.32 \pm 0.14

6.2 Ablation Study

Prototype Selection We experimented with two ways to select the prototype \hat{p} of the cluster C in CLIP space as input for the prototype captioning methods: a) selecting the *centroid* of the image cluster

$$\hat{p} = \frac{1}{|C|} \sum_{I \in C} \text{CLIP}(I), \quad (5)$$

and b) selecting the most *representative* element as the one having the highest mean cosine similarity to the other images in CLIP space

$$\hat{p} = \arg \max_{I \in C} \sum_{I' \in C} \cos(\text{CLIP}(I), \text{CLIP}(I')). \quad (6)$$

We distinguish the two versions of the same method through the notations $\text{DeCap}_{\text{centroid}}$ and $\text{DeCap}_{\text{repr}}$, and $\text{ClipCap}_{\text{centroid}}$ and $\text{ClipCap}_{\text{repr}}$ respectively. Considering the scores obtained by the same captioning method (i.e., $\text{DeCap}_{\text{centroid}}$ versus $\text{DeCap}_{\text{repr}}$), we can observe that applying the methods over the clusters’ centroids generally leads to higher scores. In fact, $\text{DeCap}_{\text{centroid}}$ beats its version applied on representatives in all the metrics, and we reported it in the comparison of Table 2 for the *DeCap* method.

ClipCap: Query Aware vs Vanilla Table 3 shows that the query-aware versions of ClipCap (ClipCap^{qa}) achieve higher scores both in Similarity to Initial Query and Representativeness metrics with respect to vanilla ClipCap, but a lower Cluster Specificity. However, query awareness generally leads these models to obtain more balanced scores among the measured properties, which is usually preferred. For this reason, we picked the query-aware variant as the best configuration we reported in Table 2 for ClipCap.

LLM Backbone Table 4 reports the metrics of GroupCap when using different LLM models for caption summarization. We tested Mistral-7B [8] and LLama3-8B [6] in the GroupCap architecture using DeCap as the captioning model. While $\text{GroupCap}_{\text{Mistral}}$ builds more cluster-specific suggestions, $\text{GroupCap}_{\text{LLama3}}$ achieves the best results both on the Representativeness measures and on Similarity to Initial Query measures, where it even gets a slightly better macro-averaged score than the human-annotated suggestions due to its high capabilities of understanding the few-shot prompt. We picked LLama3 as the LLM backbone for the GroupCap configuration reported in Table 2.

7 Conclusions

In this paper, we formalized the cross-modal query suggestion task, introducing a set of suitable properties and metrics for quantitatively evaluating different query suggestion methods. We proposed the CroQS benchmark, composed of a set of 50 diverse queries and a total of 295 image clusters, each having a human-annotated query suggestion. We adapted two different classes of methods — two captioning methods and one LLM-based method — as baselines for this novel task. We observed that all the methods achieve important gains over the results obtained using just the initial query, meaning that the proposed baselines can effectively suggest queries that better discriminate clusters, without diverging too much from the original query formulation. In the future, we aim to propose a method to assess the effectiveness of the end-to-end query suggestion pipeline, therefore including the definition of clusters and the detection of noisy ones.

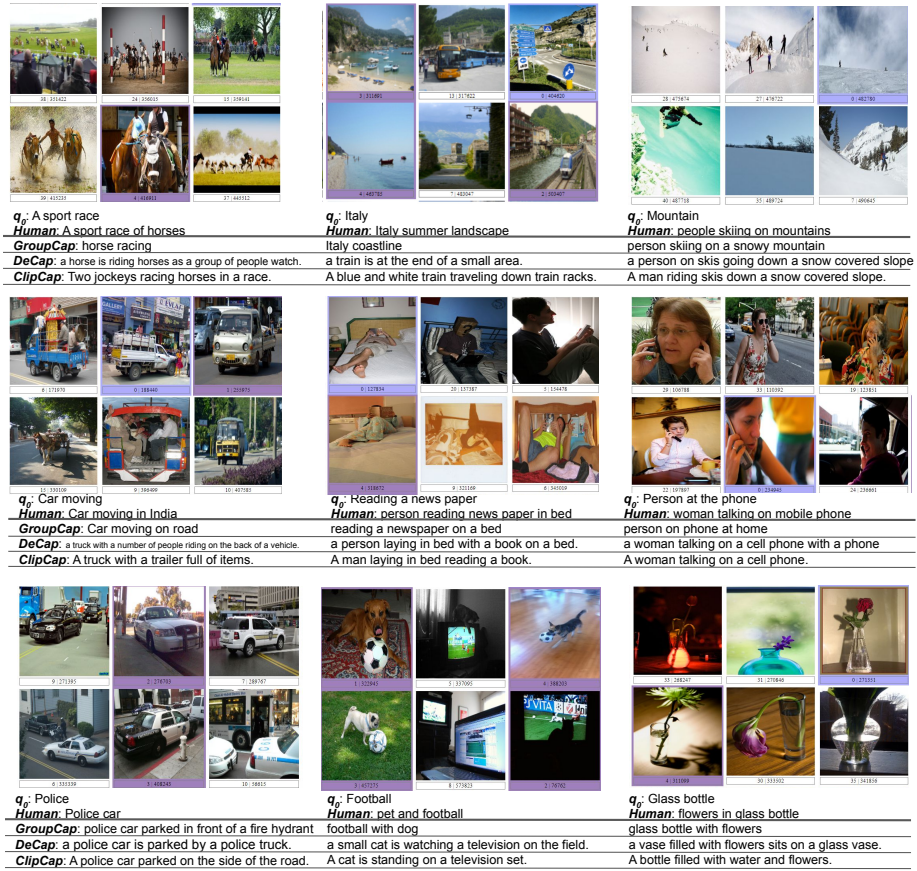


Fig. 3: CroQS samples and predictions. Each panel reports a sample of a semantic cluster with its initial query q_0 , its annotation, and the query suggested by the tested methods. Representative images (Eq. 6) have a colored outline.

Acknowledgments. This work was partially supported by FAIR – Future Artificial Intelligence Research - Spoke 1 (PNRR M4C2 Inv. 1.3 PE00000013) and MUCES – a MULTimedia platform for Content Enrichment and Search in audiovisual archives (PRIN 2022 PNRR P2022BW7CW - CUP: B53D23026090001) funded by the NextGenerationEU, by the Spoke “FutureHPC & BigData” of the ICSC – Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing funded by the Italian Government, the FoReLab and CrossLab projects (Departments of Excellence), the NEREO PRIN project (Research Grant no. 2022AEFHAZ) funded by the Italian Ministry of Education and Research (MUR).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Amato, G., Bolettieri, P., Carrara, F., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: Visione 5.0: Enhanced user interface and ai models for vbs2024. In: International Conference on Multimedia Modeling. pp. 332–339. Springer (2024)
2. Bacciu, A., Palumbo, E., Damianou, A., Tonellotto, N., Silvestri, F.: Generating query recommendations via llms. arXiv preprint arXiv:2405.19749 (2024)
3. Baek, J., Chandrasekaran, N., Cucerzan, S., Herring, A., Jauhar, S.K.: Knowledge-augmented large language models for personalized contextual query suggestion. In: Proceedings of the ACM on Web Conference 2024. pp. 3355–3366 (2024)
4. Bodigutla, P.K.: High quality related search query suggestions using deep reinforcement learning. arXiv preprint arXiv:2108.04452 (2021)
5. Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., Li, H.: Context-aware query suggestion by mining click-through and session data. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 875–883 (2008)
6. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
7. Gurrin, C., Zhou, L., Healy, G., Bailer, W., Dang Nguyen, D.T., Hodges, S., Jónsson, B.P., Lokoč, J., Rossetto, L., Tran, M.T., et al.: Introduction to the seventh annual lifelog search challenge, lsc’24. In: Proceedings of the 2024 International Conference on Multimedia Retrieval. pp. 1334–1335 (2024)
8. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023), <https://arxiv.org/abs/2310.06825>
9. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. pp. 39–48 (2020)
10. Li, W., Zhu, L., Wen, L., Yang, Y.: Decap: Decoding clip latents for zero-shot captioning via text-only training (2023)
11. Li, Z., Tran, Q., Mai, L., Lin, Z., Yuille, A.L.: Context-aware group captioning via self-attention and contrastive features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3440–3450 (2020)
12. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2014)
13. Lokoč, J., Andreadis, S., Bailer, W., Duane, A., Gurrin, C., Ma, Z., Messina, N., Nguyen, T.N., Peška, L., Rossetto, L., et al.: Interactive video retrieval in the age of effective joint embedding deep models: lessons from the 11th vbs. *Multimedia Systems* **29**(6), 3481–3504 (2023)
14. Mei, Q., Zhou, D., Church, K.: Query suggestion using hitting time. In: Proceedings of the 17th ACM conference on Information and knowledge management. pp. 469–478 (2008)
15. Messina, N., Amato, G., Esuli, A., Falchi, F., Gennaro, C., Marchand-Maillet, S.: Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **17**(4), 1–23 (2021)

16. Messina, N., Stefanini, M., Cornia, M., Baraldi, L., Falchi, F., Amato, G., Cucchiara, R.: Aladin: Distilling fine-grained alignment scores for efficient image-text matching and retrieval. In: International Conference on Content-based Multimedia Indexing. pp. 64–70 (2022)
17. Mokady, R., Hertz, A., Bermano, A.H.: Clipcap: Clip prefix for image captioning (2021)
18. Mustar, A., Lamprier, S., Piwowarski, B.: On the study of transformers for query suggestion. *ACM Transactions on Information Systems (TOIS)* **40**(1), 1–27 (2021)
19. Ooi, J., Ma, X., Qin, H., Liew, S.C.: A survey of query expansion, query suggestion and query refinement techniques. In: 2015 4th International Conference on Software Engineering and Computer Systems (ICSECS). pp. 112–117. IEEE (2015)
20. Phueaksri, I., Kastner, M.A., Kawanishi, Y., Komamizu, T., Ide, I.: Image-collection summarization using scene-graph generation with external knowledge. *IEEE Access* (2024)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
22. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. *OpenAI blog* **1**(8), 9 (2019)
23. Rocchio, J.J.: Relevance feedback in information retrieval. In: *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall (1971)
24. Trieu, N., Goodman, S., Narayana, P., Sone, K., Soricut, R.: Multi-image summarization: Textual summary from a set of cohesive images. *arXiv preprint arXiv:2006.08686* (2020)
25. Wang, J., Xu, W., Wang, Q., Chan, A.B.: Compare and reweight: Distinctive image captioning using similar images sets. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I* 16. pp. 370–386. Springer (2020)
26. Wang, J., Xu, W., Wang, Q., Chan, A.B.: Group-based distinctive image captioning with memory attention. In: *Proceedings of the 29th ACM International Conference on Multimedia*. pp. 5020–5028 (2021)
27. Wang, J., Xu, W., Wang, Q., Chan, A.B.: On distinctive image captioning via comparing and reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(2), 2088–2103 (2022)
28. Wu, B., Xiong, C., Sun, M., Liu, Z.: Query suggestion with feedback memory network. In: *Proceedings of the 2018 World Wide Web Conference*. pp. 1563–1571 (2018)
29. Zeng, Z., Fu, J., Chao, H., Mei, T.: Searching personal photos on the phone with instant visual query suggestion and joint text-image hashing. In: *Proceedings of the 25th ACM international conference on Multimedia*. pp. 118–126 (2017)
30. Zha, Z.J., Yang, L., Mei, T., Wang, M., Wang, Z.: Visual query suggestion. In: *Proceedings of the 17th ACM international conference on Multimedia*. pp. 15–24 (2009)