CroQS: Cross-modal Query Suggestion for Text-to-Image Retrieval in Image Archives

Giacomo Pacini[1,2], Nicola Messina[1], Nicola Tonellotto[1,2], Giuseppe Amato[1] and Fabrizio Falchi[1]

[1]Institute of Information Science and Technologies (ISTI), National Research Council (CNR), Pisa, Italy
[2]University of Pisa, Pisa, Italy

Abstract

This paper describes a dataset and framework for cross-modal query suggestion in text-to-image retrieval, originally published at ECIR 2025. As Digital Libraries and cultural heritage archives increasingly manage vast collections of unannotated images, it is essential to enable the retrieval of specific content without metadata. Many legacy collections were digitized without corresponding item-level descriptions due to the costs of manual annotation. While modern cross-modal retrieval models, such as CLIP, enable users to search these collections using natural language, users frequently struggle to formulate queries that perfectly align with the latent visual semantics of the database. To address this, we introduce the novel task of *cross-modal query suggestion*, which interactively guides users by suggesting textual refinements based on visual clusters identified in the retrieval results. We describe the creation of CroQS, a benchmark dataset comprising 50 diverse queries and 295 semantic clusters in generic domain. Furthermore, we evaluate two distinct methodological approaches: adapting image captioning models to describe cluster prototypes, and leveraging Large Language Models (LLMs) for few-shot summarization. Our findings indicate a trade-off where captioning models offer high specificity at the cost of natural phrasing, whereas LLM-based methods provide superior representativeness and align more closely with human query formulation.

Keywords

Text-To-Image Retrieval, Query Suggestion, Cross-modal Retrieval, Generative AI

1. Introduction

The formulation of effective search queries remains a central bottleneck in Information Retrieval (IR). This challenge is particularly present in large multimedia archives, where users must navigate extensive repositories of visual data, such as digitized photographs, artworks, or historical manuscripts that lack granular textual metadata or consistent indexing. In recent years, the field has shifted toward cross-modal retrieval, driven by vision-language models like CLIP [1] that map images and text into a shared semantic space. These models allow users to search unannotated image collections using natural language queries. However, a significant gap remains between a user's initial, often generic intent (e.g., searching for "a sport race") and the specific, rich visual content available in the archive (e.g., distinct clusters of "bike races" or "skiing competitions" that the user may not know exist).

To bridge this gap, we draw inspiration from text-based Query Suggestion (QS), a technique widely adopted in commercial search engines to refine user intent. While traditional QS relies on query logs or textual ontology mining, resources rarely available for raw image collections, we proposed a purely visual approach. In this work, summarized from our ECIR 2025 paper [2], we define the task of *Cross-modal Query Suggestion*. Unlike standard image captioning, which aims to describe a single image accurately, this task requires the system to analyze a set of retrieved images, identify coherent semantic clusters representing different facets of the result set, and generate a refined textual query for

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy
giacomo.pacini@isti.cnr.it (G. Pacini)
https://paciosoft.com/CroQS-benchmark/ (G. Pacini)
0009-0007-7745-4456 (G. Pacini); 0000-0003-3011-2487 (N. Messina); 0000-0002-7427-1001 (N. Tonellotto); 0000-0003-0171-4315 (G. Amato); 0000-0001-6258-5313 (F. Falchi)

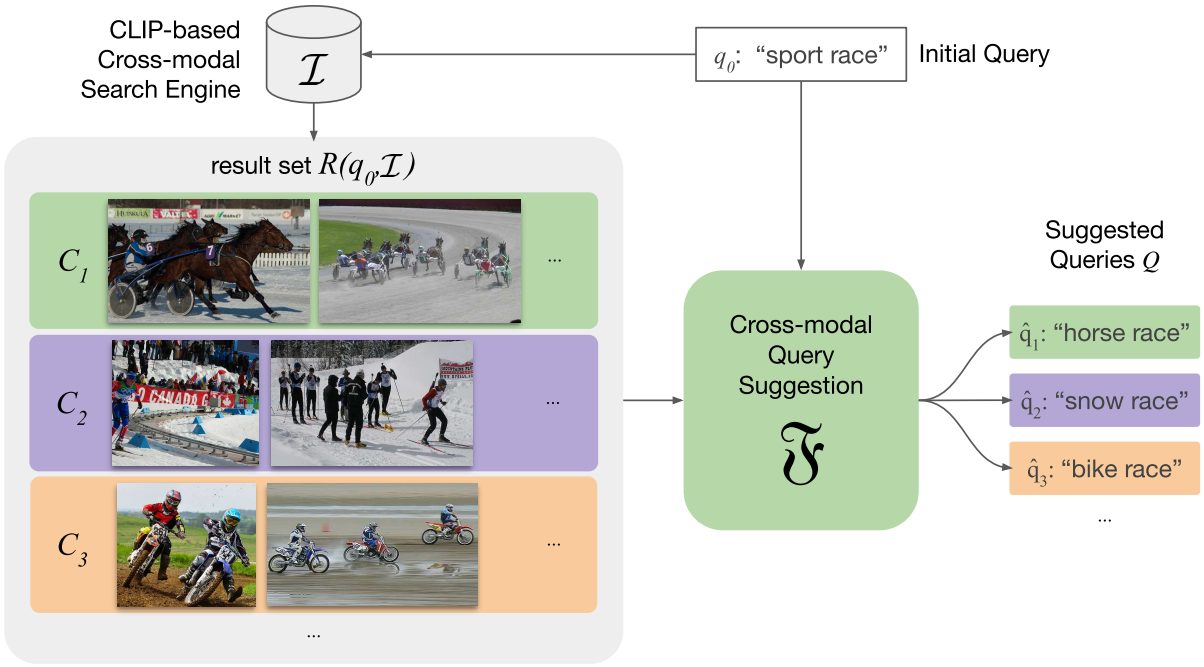

Figure 1: The Cross-modal Query Suggestion Framework. The system processes the visual results from an initial query q_0, partitions them into semantic clusters using visual embeddings, and generates a refined textual query $\hat{q}_i$ for each cluster to facilitate exploratory search.

each cluster. Ideally, these suggestions act as interactive pivots, enabling the user to disambiguate their search and explore the collection's diversity without needing prior knowledge of its specific contents.

2. Problem Formulation

Task Definition. We formalize the cross-modal query suggestion task as a function dependent on an initial user query and the latent visual structure of the content it retrieves. Let q_0 be the initial natural language query and $R(q_0, I)$ be the result set retrieved from an unannotated image collection I. We assume that this result set is not visually monolithic but contains multiple latent semantic groups, denoted as $\mathcal{C} = \{C_1, \ldots, C_M\}$, where each C_i represents a visually and semantically coherent subset of images.

Assumptions. The goal of the cross-modal query suggestion system, $\mathfrak{F}$, is to generate a specific suggested query $\hat{q}_i$ for each identified group C_i. We make the simplifying assumption that each suggested query can be generated by looking only at its corresponding group of images. In general, the information about other groups can be useful to contrast the current suggestion against other similar ones, but we leave this case for future work. Thus, we can express each suggested query as $\hat{q}_i = \mathfrak{F}(q_0, C_i)$.

Desired Output Properties. The generated query $\hat{q}_i$ must satisfy a difficult balance of conflicting properties to be useful in a real-world search scenario. First, it must be *specific* enough to distinguish the images in cluster C_i from the rest of the original results, effectively filtering out noise. Second, it must be *representative*, acting as an effective, standalone search query that can retrieve relevant items from the global collection, not just re-rank the current view. Finally, it should maintain high semantic and syntactic *similarity* to the original query q_0. A suggestion that diverges too much from the original intent runs the risk of disorienting the user; the ideal suggestion is a refinement, not a complete rewriting, of the user's expressed information need.

3. The CroQS Benchmark

A major roadblock in advancing this novel field is the lack of standardized datasets for evaluating query suggestions based purely on image clusters. To address this, we constructed the CroQS benchmark through a semi-automatic process involving human supervision. We utilized images from the COCO [3] dataset as a proxy for a generic image collection but deliberately discarded its rich captions to simulate a realistic, unannotated library scenario.

Dataset Construction. The construction of the benchmark involved a multi-stage validation process to ensure high quality. We selected 50 initial queries spanning diverse categories relevant to general collections, such as animals, people, actions, objects, and places. For each query, the top 200 images retrieved by CLIP were analyzed. Automated clustering in the CLIP embedding space provided initial groupings, which were then expertly validated by human annotators. These experts discarded noise and ensured that each retained cluster represented a distinct, human-interpretable visual concept. Crucially, the annotators provided a "gold standard" suggested query for each valid cluster. The resulting dataset contains 8,929 analyzed images organized into 295 validated semantic clusters, averaging 5.9 distinct semantic groups per initial query. This structure allows us to factor out the variability of different clustering algorithms and focus the evaluation strictly on the quality of the generated query suggestions.

Evaluation Protocol. To quantify performance, we established a evaluation framework mapping to the desired properties defined above. First, to ensure the suggested query accurately targets its visual group, we measure *Cluster Specificity* via Recall, assessing how effectively the new query re-ranks the specific cluster members to the top of the initial result set. Second, to ensure the query remains useful for broadening the search beyond the current context, we measure *Representativeness* via standard IR metrics (Recall, NDCG, and MAP) on the global collection, treating the verified cluster members as pseudo-relevant items. Finally, to ensure continuity with the user intent, we measure syntactic and semantic *Similarity* using the Jaccard index on text tokens and cosine similarity between CLIP textual embeddings of the original and suggested queries, respectively.

4. Methodology

We explored two primary architectural strategies to solve the cross-modal query suggestion task, adapting state-of-the-art models from related domains to the specific constraints of our problem.

Latent Prototype Captioning. The first strategy, **Latent Prototype Captioning**, adapts image captioning models to the query suggestion task. Standard latent captioning models, such as ClipCap [4] and DeCap [5], are designed to generate text from a single image embedding. To apply them to a *set* of images, we compute a single "prototype" embedding representing the entire cluster. Specifically, we used the centroid of the cluster's images in the CLIP semantic space. This prototype is fed into the captioning network to generate a description representing the group's average visual content.

Since standard captioning ignores the user's input, we further enhanced the ClipCap architecture to be "query-aware". Standard ClipCap uses a mapping network to translate a CLIP image embedding into a sequence of prefix tokens for a frozen GPT-2 decoder. We modified this by injecting the embedding of the original textual query q_0 directly into the decoding process, thereby conditioning the generated caption on the user's original phrasing.

Captions Summarization. The second strategy, **Captions Summarization (GroupCap)**, leverages the impressive few-shot reasoning capabilities of Large Language Models (LLMs). This approach operates in two steps. First, it extracts descriptive captions for the most representative images within a cluster using an off-the-shelf captioning model. Second, these captions, along with the initial query q_0, are fed into an LLM (such as Llama3 [6] or Mistral [7]) within a few-shot prompting setup. The LLM is

Table 1
Comparison of the best cross-modal query suggestion configurations for each strategy. We report macro-averaged mean and and standard deviation. The reported GroupCap strategy uses DeCap as image captioner and LLaMa3-8B as LLM.

	Similarity to q_0		Cluster Specificity	Representativeness		
Method	**CLIP**	**Jaccard**	**Recall Cluster**	**Recall**	**NDCG**	**MAP**
Baselines						
original q_0	1.00 ± 0.00	1.00 ± 0.00	0.19 ± 0.07	0.52 ± 0.03	0.18 ± 0.05	0.21 ± 0.05
human	0.87 ± 0.03	0.44 ± 0.11	0.59 ± 0.13	0.62 ± 0.12	0.52 ± 0.14	0.45 ± 0.11
Latent Prototype Captioning						
ClipCap [4]	0.74 ± 0.06	0.17 ± 0.11	$\mathbf{0.54 \pm 0.16}$	0.40 ± 0.15	0.32 ± 0.15	0.28 ± 0.12
DeCap [5]	0.77 ± 0.06	0.16 ± 0.12	0.53 ± 0.18	0.40 ± 0.15	0.31 ± 0.15	0.28 ± 0.12
Captions Summarization						
GroupCap	$\mathbf{0.90 \pm 0.04}$	$\mathbf{0.56 \pm 0.12}$	0.39 ± 0.17	$\mathbf{0.47 \pm 0.12}$	$\mathbf{0.35 \pm 0.18}$	$\mathbf{0.32 \pm 0.14}$

instructed via examples to synthesize the common visual themes found in the captions and the user's original intent into a single, cohesive, and natural search query. This method aims to capitalize on the LLM's pre-trained world knowledge and linguistic flexibility to generate high-quality search phrases.

5. Experimental Results and Discussion

Our experimental analysis revealed distinct performance trade-offs between the two proposed strategies, as reported in Table 1 and visualized in the qualitative examples in Figure 2.

Latent captioners are more specific. The captioning-based models demonstrated exceptional performance in *Cluster Specificity*, achieving a Recall of approximately 0.54. By focusing entirely on the deep visual features of the cluster prototype, these models successfully generated highly detailed descriptions that differentiated subtle visual variations (e.g., distinguishing a cluster of "cats on a bed" from another cluster of "cats on a sofa" within results for "cat"). However, this specificity often came at the cost of retrieval performance as a general search query. The generated text was frequently too descriptive, resulting in lower Representativeness scores compared to human annotations.

LLM summarization is closest to human baseline. Conversely, the LLM-based GroupCap approach achieved a superior balance for practical IR applications. It obtained the highest scores in semantic *Similarity* to the original query (0.90 CLIP score vs 0.77 for DeCap) and *Representativeness* (0.35 NDCG vs 0.31 for DeCap), getting the closest to human performance in these retrieval metrics. The LLM's ability to abstract away from specific visual details and its inherent bias towards natural language generation allowed it to formulate queries that were concise, effective, and closer to the human baseline.

6. Conclusion

In this work, we formalized the task of Cross-modal Query Suggestion, addressing a critical need for enhancing explorability in unannotated visual archives within image archives. We introduced CroQS, a novel benchmark designed to standardize evaluation in this emerging field. Our comparative analysis highlights that while high visual specificity can be achieved through direct prototype captioning, the synthesis of effective, natural-sounding search queries is best handled by large language models that can integrate visual context with user intent through few-shot summarization. These findings provide a foundation for future intelligent search interfaces that can guide users through the unseen semantic structure of large-scale image collections. Future directions include developing end-to-end pipelines that

Figure 2: Qualitative examples from CroQS. We show the initial query, the visual cluster, the human ground truth annotation, and suggestions generated by the caption summarization strategu (GroupCap) and the latent prototype captioning strategy (DeCap and ClipCap).

integrate unsupervised cluster detection, validation experiments on real-world, noisier digital libraries and investigating novel strategies based on large multimodal models [8, 9] and group captioning [10].

Declaration on Generative AI

During the preparation of this work, the author(s) used Google's Gemini large language model for drafting, paraphrasing, and rewording content. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Acknowledgments

This work was partially supported by FAIR – Future Artificial Intelligence Research - Spoke 1 (PNRR M4C2 Inv. 1.3 PE00000013) and 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).

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[2] G. Pacini, F. Carrara, N. Messina, N. Tonellotto, G. Amato, F. Falchi, Maybe you are looking for CroQS cross-modal query suggestion for text-to-image retrieval, in: European Conference on Information Retrieval, Springer, 2025, pp. 138–152.

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[4] R. Mokady, A. Hertz, A. H. Bermano, ClipCap: Clip prefix for image captioning, 2021. arXiv:2111.09734.

[5] W. Li, L. Zhu, L. Wen, Y. Yang, DeCap: Decoding clip latents for zero-shot captioning via text-only training, in: The Eleventh International Conference on Learning Representations, 2023.

[6] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[7] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, et al., Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[8] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al., Flamingo: a visual language model for few-shot learning, Advances in neural information processing systems 35 (2022) 23716–23736.

[9] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, Advances in neural information processing systems 36 (2023) 34892–34916.

[10] J. Wang, W. Xu, Q. Wang, A. B. Chan, Group-based distinctive image captioning with memory attention, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 5020–5028.