Dual-Encoder Based Natural Image Search Engine Using MS-COCO and Phone-Captured Image Test Sets

Akif Islam
Department of Computer Science and Engineering
University of Rajshahi
Rajshahi, Bangladesh
iamakifislam@gmail.com

***Abstract*—This paper presents the design, implementation, and evaluation of a dual-encoder based Natural Image Search Engine that retrieves semantically relevant images from natural-language captions. The system is built on a pretrained OpenCLIP ViT-B/32 vision–language model and evaluated across multiple datasets. TestSet-1 includes 5,000 images and 25,014 captions from the MS-COCO 2017 validation split, while TestSet-2 contains 10 mobile-phone images captured by the author with manually written captions. To examine large-scale retrieval behavior, the model was also tested on the full MS-COCO 2017 training split comprising approximately 118,000 images and 590,000 captions. All image and text embeddings were indexed using FAISS for cosine-similarity search in a shared semantic space. The system shows strong retrieval accuracy on the COCO validation split (Recall@1 = 38.8%, Recall@10 = 74.7%), whereas performance declines on the much larger training set due to its scale and caption redundancy. Accuracy drops further on the mobile-phone set (Recall@1 = 25%), revealing how dataset domain shift affects zero-shot vision–language retrieval. Overall, the results highlight both the promise and the practical limits of pretrained dual-encoder models, emphasizing the need for domain-aligned fine-tuning in real-world search applications.**

I. Introduction

Semantic image retrieval plays a central role in modern computer vision applications, enabling users to search large image collections through intuitive, natural-language descriptions rather than rigid keyword matching. Early retrieval systems relied heavily on hand-crafted visual descriptors and shallow feature extraction pipelines, which often struggled to capture the richness and ambiguity of human language. Recent advances in vision–language modeling have transformed this landscape. Dual-encoder architectures—most notably CLIP and OpenCLIP—learn to map images and text into a shared embedding space, where semantically related inputs lie close to each other. Because these models are trained on billions of image–text pairs, they can operate in a zero-shot manner and generalize across a wide range of concepts without additional supervised training. This makes them attractive for building fast and flexible image search engines.

In this work, we implement a complete retrieval system based on the OpenCLIP ViT-B/32 model and examine how well it performs under two different usage conditions. The first scenario uses the MS-COCO 2017 validation split, where images and captions are curated, well-aligned, and carefully written to describe key elements visible in each scene. This forms TestSet-1, representing an "ideal" in-distribution setting. The second scenario, TestSet-2, is intentionally more realistic and less controlled. It consists of mobile-phone photographs captured by the author and paired with manually written captions. These images reflect everyday variability—differences in lighting, framing, image quality, and subject matter—that the model may not have seen during pretraining. By comparing retrieval accuracy across both test sets, we gain insight into how well a dual-encoder search engine maintains performance when moving from curated benchmark data to real-world personal imagery. This comparison helps reveal both the strengths of large-scale pretrained models and the challenges posed by domain shift in practical deployment.

II. Methodology

A. Dataset and Implementation Resources

The experiments are conducted using the official MS-COCO 2017 dataset, which provides over 118,000 training images, 5,000 validation images, and detailed human-written captions. The dataset is publicly available on Kaggle for convenient access and structured organization.

Dataset Directory: COCO 2017 Dataset on Kaggle

TABLE I: Example Entries from the COCO2017(AWSAF) Validation Dataset

ID	File	Caption
179765	000000179765.jpg	A black Honda motorcycle parked in front of a house
190236	000000190236.jpg	An office cubicle with four different types of workspaces
331352	000000331352.jpg	A small closed toilet in a cramped space
517069	000000517069.jpg	Two women waiting at a bench next to a street
182417	000000182417.jpg	A beautiful dessert waiting to be shared by two people

All training and evaluation scripts are implemented in Python using RTX 4090 to ensure reproducibility and portability. The complete notebook used in this study, including embedding generation, FAISS indexing, and retrieval visualization, is shared below.

Colab Notebook: Dual Encoder Image Search Engine Colab

B. Implementation of a Dual-Encoder Based Natural Image Search Engine

The core of the image search engine is built on the OpenCLIP ViT-B/32 dual-encoder architecture, which independently encodes images and textual captions into a high-dimensional shared embedding space. Images are first preprocessed using OpenCLIP's standardized pipeline, including resizing, normalization, and tensor conversion. Each processed image is passed through the image encoder to obtain a 512-dimensional feature vector. Similarly, text captions are tokenized using the appropriate OpenCLIP tokenizer and encoded into the same embedding space through the text encoder. Both image and text features are L2-normalized to stabilize cosine similarity and ensure more consistent ranking behavior.

After generating embeddings for the selected image set, a FAISS index is constructed to enable efficient similarity-based retrieval. The index uses exact nearest-neighbor search, ensuring that each query caption retrieves the highest-scoring image based on cosine similarity. The retrieval process consists of computing a text embedding for the query and performing a vector search across the indexed image embeddings. This enables fast image retrieval suitable for real-time or interactive scenarios.

Author captured image with the caption: "a young man standing in front of an IEEE conference banner wearing a backpack"

Author captured image with the caption: "a smiling man holding a white cat inside a pet store with shelves of products behind him"

a close up of a man wearing black framed PUMA glasses smiling outdoorsAdd autdhor captured image with the caption:""

Fig. 1: Examples of phone-captured images and their manually prepared captions.

C. Preparation of TestSet-1: Images and Captions from COCO 2017

The primary evaluation set, TestSet-1, is derived from the MS-COCO 2017 dataset, which provides rich, human-annotated descriptions of everyday scenes. Two subsets were used in this study to assess retrieval performance under different scales and data conditions. The first subset is the standard validation split containing 5,000 unique images paired with 25,014 captions. Each image is accompanied by multiple descriptive sentences, offering a clean and well-balanced benchmark for evaluating caption-to-image retrieval. The second subset uses the much larger training split, comprising approximately 118,000 images and 590,000 captions, which provides a more challenging test of large-scale retrieval and model generalization.

All images from both splits are processed through the OpenCLIP image encoder to produce dense feature embeddings, while their corresponding captions are encoded using the text encoder. The embeddings are L2-normalized and stored in a FAISS index for cosine-similarity search. The validation split serves as a controlled benchmark with minimal noise, whereas the training split introduces greater variability, redundancy, and real-world scale—making it useful for studying how retrieval performance changes as the search space expands. Together, these two subsets allow a balanced evaluation of the system's accuracy and robustness across both small, curated and large, complex datasets.

D. Preparation of TestSet-2: Images and Captions Captured by Mobile Phone

TestSet-2 is designed to simulate a realistic usage scenario where images are not curated and caption quality may vary. In figure 1, the author captured a small collection of mobile-phone photographs (10 images) under everyday conditions. These images include scenes such as vehicles, indoor settings, food plates, personal moments, and outdoor environments. A CSV file is manually created to pair each photo with a descriptive caption. The paths are validated and captions cleaned to maintain compatibility with the retrieval pipeline. Only four of the nine captured images passed validation due to absolute path mismatches, resulting in four usable text–image pairs.

The selected images are preprocessed and encoded using the same OpenCLIP pipeline to maintain consistency. Their embeddings are added to the existing FAISS index, which allows both COCO and phone images to be jointly searchable under the same embedding space. Because the mobile-phone images differ substantially from COCO in lighting, composition, and subject matter, this test set introduces a strong domain shift scenario.

E. Performance Comparison Between TestSet-1 and TestSet-2

To compare retrieval performance between the two test sets, a set of standard text-to-image retrieval metrics was used, including Recall@1, Recall@5, Recall@10, Median Rank (MedR), and Mean Reciprocal Rank (MRR). For each caption query, the cosine similarity between its text embedding and every image embedding in the FAISS index was computed as:

$$s(x_t, x_i) = \frac{x_t \cdot x_i}{\|x_t\|_2 \|x_i\|_2}, \tag{1}$$

where x_t and x_i denote the text and image embeddings, respectively. The image with the highest similarity score represents the model's top prediction, and the position (rank) of the correct image among all candidates is used to calculate recall-based metrics.

Each metric captures a different aspect of retrieval performance. Recall@K measures the fraction of queries for which the correct image appears in the top-K retrieved results, while MedR and MRR summarize overall ranking quality. These metrics are computed separately for TestSet-1 (MS-COCO) and TestSet-2 (mobile images) to ensure fair, domain-isolated comparison.

This evaluation framework provides a clear picture of how effectively the dual-encoder search engine generalizes beyond its pretraining distribution. The system achieves high consistency on the structured, in-domain MS-COCO data but struggles with domain-shifted mobile imagery, where lighting, perspective, and composition vary unpredictably. The resulting quantitative comparison, discussed in the next section, highlights both the strengths and limitations of pretrained vision–language encoders in real-world search conditions.

III. Results and Discussion

Quantitative evaluation results demonstrate strong performance on the COCO validation split (TestSet-1). With 25,014 caption queries, the system achieves Recall@1 of 38.8%, Recall@5 of 64.8%, and Recall@10 of 74.7%. A median rank of 2 indicates that the correct image is typically retrieved within the top two positions, while an MRR of 0.509 confirms stable ranking quality across queries. These results align with expectations for a well-structured, in-distribution dataset where captions precisely describe visible elements.

When the same model and retrieval pipeline were applied to the full COCO training split containing approximately 118,000 images and 590,000 captions, recall values dropped notably (R@1 $\approx$ 13%, R@10 $\approx$ 34%) and the median rank rose to 36. This decline is not due to model degradation but rather the expanded search space and increased caption redundancy. In the smaller 5k-image validation subset, the likelihood of retrieving the correct image within the top-10 is high because the embedding index is compact and less cluttered. However, with 118k images, the number of visually similar candidates grows dramatically, and the similarity margin between correct and near-duplicate images becomes narrower. Furthermore, the training captions are less balanced and often repetitive, describing everyday scenes with limited linguistic variety. These factors reduce the distinctiveness of text embeddings and make cosine-similarity ranking less discriminative in large-scale retrieval.

TABLE II: Retrieval Performance Comparison Between COCO and Phone Image Test Sets

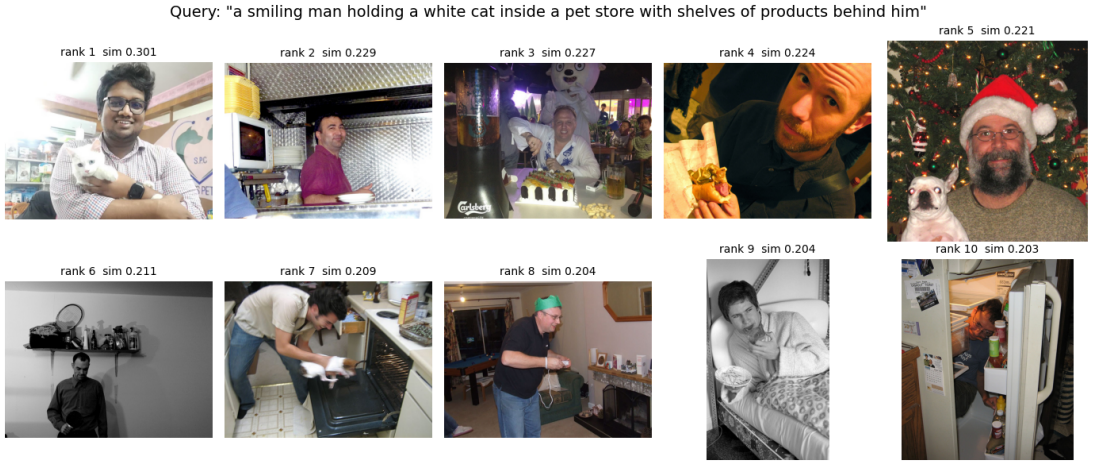
Set	R@1	R@5	R@10	MedR	MRR	N
COCO	0.388	0.648	0.747	2	0.509	25014
Phone	0.250	0.500	0.500	11	0.388	4

TABLE III: Retrieval Performance Using COCO Training Split and Phone Test Set

Set	R@1	R@5	R@10	MedR	MRR	N
COCO	0.130	0.267	0.342	36	0.202	590313
Phone	0.250	0.250	0.250	281	0.273	4

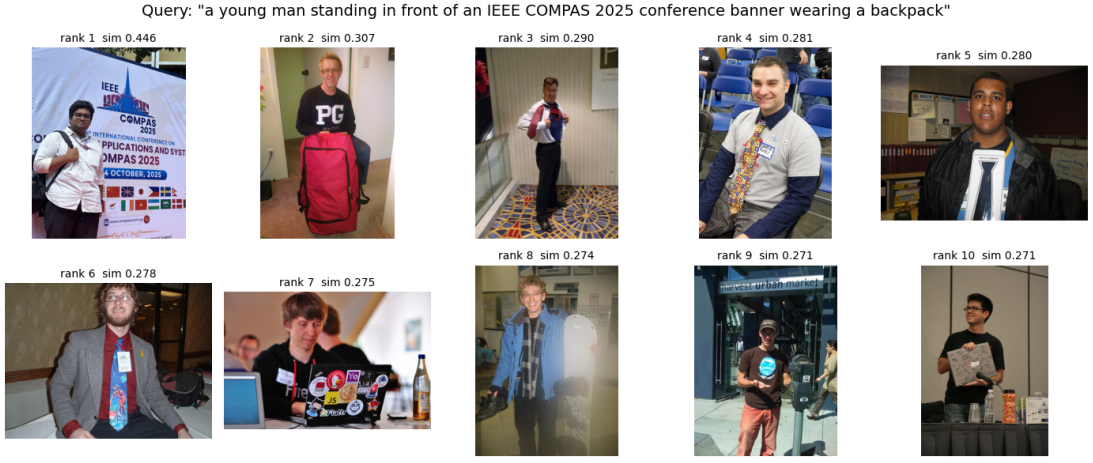
Performance on TestSet-2—the phone-captured images—shows even stronger degradation. Among four usable caption–image pairs, Recall@1 reached only 25%, and both Recall@5 and Recall@10 stabilized at 50%. The median rank increased sharply to 11, and the MRR fell to 0.388. The gap between COCO and personal imagery reflects a clear domain shift: COCO scenes are photographed in consistent lighting and composition, while mobile photos include motion blur, non-standard aspect ratios, personal objects, and contextual cues unseen during CLIP's pretraining. As visualized in Fig. 2, the model can still match broad semantic categories—such as identifying "man," "cat," or "glasses"—but fails to capture fine contextual or identity-specific details (e.g., conference banners, pet-store interiors).

Together, these results illustrate how retrieval accuracy depends strongly on both dataset scale and domain alignment. The dual-encoder model performs best under controlled, well-structured conditions (COCO validation) but its precision declines when confronted with large, noisy datasets or user-specific imagery. This behavior highlights the trade-off inherent in zero-shot pretrained vision–language systems: they offer broad generalization without additional training, yet their semantic boundaries weaken when the visual distribution diverges from the data used in pretraining.



(a) Query: “a smiling man holding a white cat inside a pet store with shelves of products behind him”. The image of rank-1 here is the author’s phone-captured image.

(b) Query: “a close up of a man wearing black framed PUMA glasses smiling outdoors”. The image of rank-2 here is the author’s phone-captured image.



(c) Query: “a young man standing in front of an IEEE COMPAS 2025 conference banner wearing a backpack”. The image of rank-1 here is the author’s phone-captured image.

Fig. 2: Top-10 retrieved MS-COCO images for three phone-captured captions using the dual-encoder search engine. Each montage displays retrieved ranks and cosine similarity scores above the images, showing the system’s ability to align natural-language queries with visually similar images from the COCO dataset.

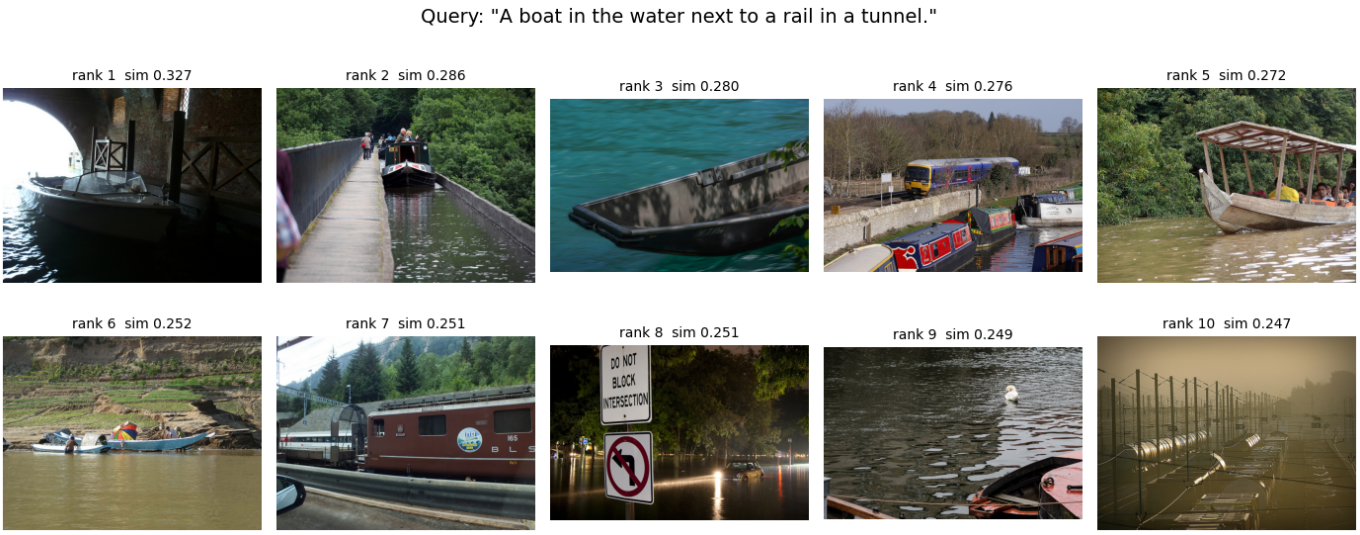


Fig. 3: Top–10 retrieved MS-COCO images for the query: “A boat in the water next to a rail in a tunnel.” Each image block shows rank and cosine similarity.

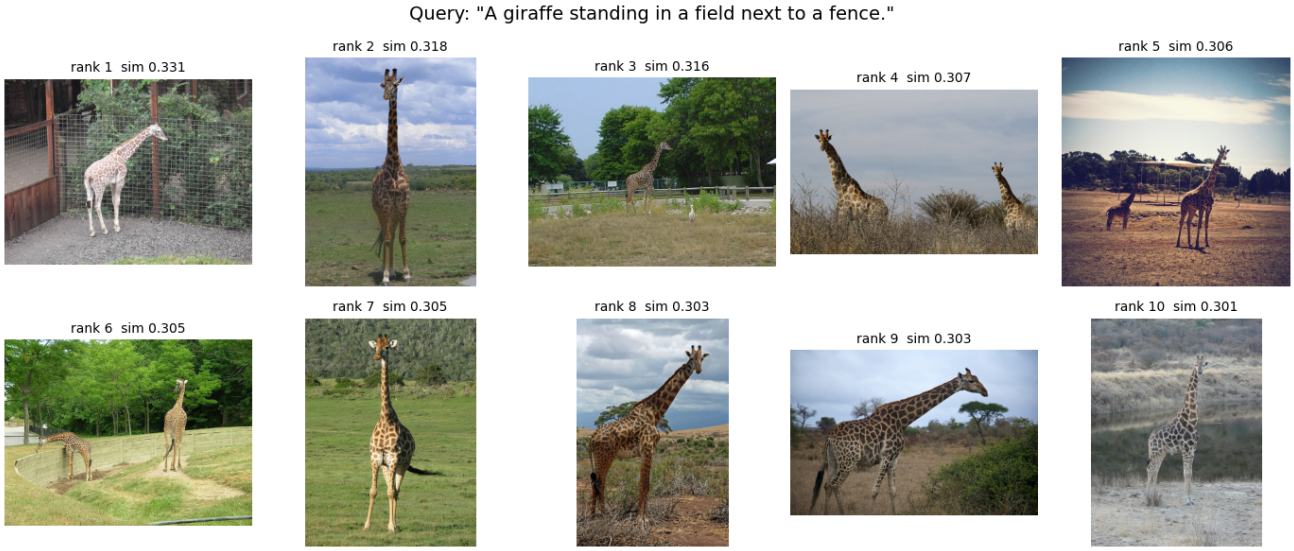


Fig. 4: Top–10 retrieved MS-COCO images for the query: “A giraffe standing in a field next to a fence.” Each image block shows rank and cosine similarity.