Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00049-3

Original Article

A Cross-Modal Approach to Enhancing Image Retrieval With Contrastive Language-Image Pretraining (CLIP)-Based Embeddings and Facebook AI Similarity Search (FAISS) Indexing

Johnson T. Fakoya [1] , Micheal O. Ajinaja [2,]

1. *Software Engineering and Information Systems, Federal University of Agriculture, Abeokuta, NGA*
2. *Computer Science, Federal Polytechnic Ile Oluji, Ile Oluji, NGA*

Received: February 28, 2026 | Review began: March 08, 2026 | Review ended: March 28, 2026 | Published: April 02, 2026

Abstract

Cross-modal image retrieval plays an important role in managing large multimedia collections and supporting efficient search across visual and textual data. This study introduces an image retrieval framework based on Contrastive Language-Image Pretraining (CLIP) and Facebook AI Similarity Search (FAISS). The system combines multimodal embedding generation with high-performance vector similarity indexing to support efficient cross-modal search. The framework uses the Contrastive Language-Image Pretraining model to generate shared embeddings for both images and text queries. These embeddings place visual and linguistic information within the same semantic space, which allows the system to connect text descriptions with related images. To support fast and scalable search, the generated embeddings are indexed using the FAISS library. FAISS performs efficient k-nearest neighbor retrieval in high-dimensional vector spaces, which enables rapid similarity search across large datasets. The system supports both text-to-image and image-to-image retrieval tasks. Users search an image database either with descriptive text queries or with reference images. Experimental evaluation shows strong retrieval performance, with effective results based on mean average precision and Recall@k metrics. Additional analyses strengthen these findings. Similarity score distributions and t-distributed stochastic neighbor embeddings show clear grouping of images by conceptual similarity within the embedding space. These results demonstrate how the CLIP representation organizes images based on meaning rather than simple visual patterns. Overall, the results show the value of combining multimodal representation learning with scalable vector indexing. The proposed CLIP-FAISS framework offers a practical solution for image retrieval and supports applications such as visual search engines, digital libraries, and multimedia content management systems.

Categories: AI applications, Algorithm Analysis, Data Mining

Keywords: clip embeddings, faiss indexing, image retrieval, cross-modal search, semantic similarity, contrastive learning, deep learning, content-based image retrieval (cbir), multimodal ai

Introduction

The rapid growth of digital multimedia across social media, e-commerce platforms, medical imaging systems, and digital libraries has made image retrieval an important research area. Large image collections grow every day, and locating relevant images within those collections requires efficient search methods. Effective retrieval systems support information access, content organization, and recommendation services across many applications. Traditional image retrieval methods rely on manually assigned metadata or keywords. Such annotations rarely reflect the true visual or semantic meaning within images, which leads to limited accuracy and poor scalability during search.

How to cite this article:
Fakoya J T, Ajinaja M O (April 02, 2026) A Cross-Modal Approach to Enhancing Image Retrieval With Contrastive Language-Image Pretraining (CLIP)-Based Embeddings and Facebook AI Similarity Search (FAISS) Indexing. Cureus J Comput Sci 3 : es44389-026-00049-3. DOI https://doi.org/10.7759/s44389-026-00049-3

To address these issues, early work in content-based image retrieval (CBIR) focused on extracting visual features directly from images. Techniques examined color histograms, texture patterns, and edge descriptors as ways to represent visual content. These methods improved retrieval compared with keyword-based systems, yet they struggled with high-level meaning. Many CBIR systems matched images with similar visual patterns but failed when users searched for conceptual relationships rather than identical appearances [1]. Later research introduced machine learning and deep neural networks in order to learn visual representations from large datasets. This shift improved feature extraction and strengthened retrieval performance across many tasks [2,3].

Deep learning further advanced image retrieval through the use of convolutional neural networks (CNNs). CNN architectures learned hierarchical visual features and provided stronger representations for tasks such as image classification and object recognition. Feature embeddings produced by CNN models proved useful for similarity search and large-scale image retrieval. Multiple studies reported strong improvements over earlier CBIR approaches based on handcrafted descriptors [4,5]. Despite these gains, many CNN-based retrieval systems focus only on visual similarity. Training also depends on large labeled datasets, which restricts semantic search and cross-modal retrieval.

Multimodal learning introduced new approaches for connecting visual data with natural language. One widely studied model in this area is Contrastive Language-Image Pretraining (CLIP). CLIP learns joint representations of images and text through training on large collections of paired image-text data. Both modalities map into a shared embedding space, which allows direct comparison between textual queries and images. Such alignment supports flexible cross-modal retrieval, including text-to-image search and image-to-image comparison. This capability improves semantic understanding and broadens potential uses of image retrieval systems [6,7]. Research also shows strong performance from CLIP embeddings during text-to-image retrieval tasks, with improved alignment between visual and linguistic representations [8].

Table *1* show comparative analysis table serves to contextualize the proposed framework within the broader evolutionary trajectory of image retrieval systems. It contrasts traditional CBIR, which relies on rigid, handcrafted descriptors like color and texture, with standard supervised CNNs that, while more advanced, often require extensive labeled datasets for visual recognition.

Feature	Traditional CBIR	Standard CNN	Our Approach (CLIP-FAISS)
Feature Type	Handcrafted (SIFT/HOG)	Supervised Visual	Multimodal Semantic
Data Dependency	Low	High (Labeled)	Low (Zero-Shot)
Search Method	Linear Search	Euclidean Distance	FAISS Vector Indexing
Cross-Modal?	No	Limited	Yes (Text & Image)

TABLE 1: Comparative analysis of the proposed CLIP-FAISS framework against traditional and supervised retrieval methodologies

CBIR, Content-Based Image Retrieval; CLIP, Contrastive Language-Image Pretraining; CNN, Convolutional Neural Network; FAISS, Facebook AI Similarity Search; HOG, Histogram of Oriented Gradients; SIFT, Scale-Invariant Feature Transform

By comparing these to the CLIP-Facebook AI Similarity Search (FAISS) approach, the table illustrates a paradigm shift toward "zero-shot" multimodal semantic alignment. It highlights how the integration of foundation models with high-performance vector indexing overcomes the "semantic gap" and the computational bottlenecks inherent in earlier

How to cite this article:
Fakoya J T, Ajinaja M O (April 02, 2026) A Cross-Modal Approach to Enhancing Image Retrieval With Contrastive Language-Image Pretraining (CLIP)-Based Embeddings and Facebook AI Similarity Search (FAISS) Indexing. Cureus J Comput Sci 3 : es44389-026-00049-3. DOI https://doi.org/10.7759/s44389-026-00049-3

methodologies, moving from simple visual matching to sophisticated, scalable cross-modal understanding. While CLIP and FAISS have been individually well-studied, their systematic integration into a unified cross-modal retrieval pipeline with detailed evaluation of both text-to-image and image-to-image search - including qualitative analyses of embedding structure and similarity distributions - remains underexplored. This study addresses that gap by presenting a complete, reproducible framework and evaluating its semantic retrieval capabilities.

Even with improved embeddings, efficient similarity search remains a challenge when datasets contain millions of images. High-dimensional vector representations require fast indexing and search methods. Vector similarity libraries such as FAISS address this need through optimized nearest-neighbor search algorithms. FAISS supports scalable indexing structures and approximate nearest-neighbor search, which reduces computational cost during large-scale retrieval tasks [9]. Combining CLIP embeddings with FAISS indexing forms a practical solution for scalable cross-modal retrieval systems. This integration enables efficient search using either visual examples or natural language queries.

Motivated by these advances, this study presents a cross-modal image retrieval framework built on CLIP embeddings and FAISS indexing. The system extracts multimodal embeddings from images using a pretrained CLIP model, stores those embeddings in a FAISS vector index, and performs similarity search through cosine distance and k-nearest neighbor (k-NN) retrieval. This design joins multimodal representation learning with scalable indexing techniques, producing an efficient and flexible image retrieval approach. The framework supports both text-to-image and image-to-image search across diverse image datasets.

Materials And Methods

The proposed cross-modal image retrieval framework follows a structured methodological pipeline designed to support efficient multimodal representation learning and scalable similarity search. The overall methodology consists of four major stages: (i) dataset preparation and preprocessing, (ii) multimodal embedding extraction using a pre-trained CLIP model, (iii) vector indexing using FAISS for efficient similarity search, and (iv) retrieval and evaluation using nearest-neighbor search metrics. The complete workflow of the proposed system, including data preparation, embedding generation, indexing, and retrieval, is illustrated in Figure *1*.

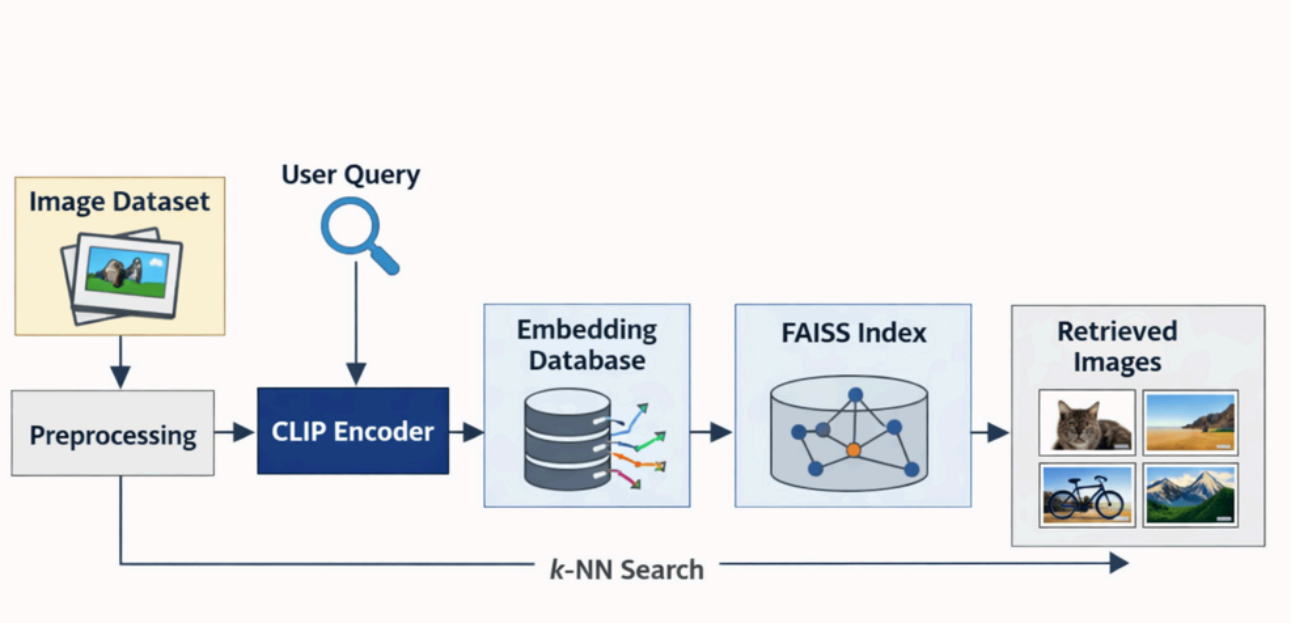

FIGURE 1: System architecture of the proposed CLIP-FAISS image retrieval framework

CLIP, Contrastive Language-Image Pretraining; FAISS, Facebook AI Similarity Search; k-NN, k-Nearest Neighbor

The proposed CLIP-FAISS image retrieval framework supports efficient cross-modal image search by combining multimodal feature extraction with fast similarity indexing. The system includes five main stages: dataset preparation, preprocessing, embedding generation with CLIP, vector indexing with FAISS, and image retrieval through nearest-neighbor search. The process starts with an image dataset, which forms the retrieval database. This dataset contains all images available for search. Before feature extraction begins, each image passes through a preprocessing stage. During this step, images are resized and normalized to match the input format required by the CLIP model. Standardizing image size and format keeps the feature extraction process consistent across the entire dataset.

After preprocessing, the images move into the CLIP encoder. The encoder produces high-dimensional embedding vectors representing the semantic content of each image. These embeddings place images within a shared representation space where related concepts sit close together. As a result, the model captures meaning and context rather than simple pixel similarity. Once generated, the embedding vectors are stored in an embedding database. The system then indexes them with FAISS. FAISS organizes the vectors in a structure built for fast similarity comparison, which allows the system to search large collections of embeddings without heavy computational cost.

During retrieval, the user submits a query in the form of a text description or a reference image. The system converts the query into an embedding vector using the appropriate CLIP encoder. FAISS then performs a k-NN search to locate the most similar vectors within the index using cosine similarity. The system returns the top-k images ranked by similarity score. This architecture supports scalable image retrieval while maintaining strong semantic understanding. CLIP provides the multimodal representation that links images with language, while FAISS handles high-speed vector search across large datasets. Together, these components form an efficient framework capable of supporting both text-to-image and image-to-image retrieval tasks.

While the current dataset comprises 52 images, it serves as a proof-of-concept validation for the CLIP-FAISS integration. The dataset was deliberately selected to include diverse semantic categories (animals, sports objects, artworks, landscapes) to test cross-modal generalization across varied concepts. The architectural design using FAISS is inherently scalable to millions of vectors, and the methodological pipeline established here can be directly transferred to larger benchmark datasets such as MS-COCO or Flickr30k in future work.

Dataset description

The dataset used in this study consists of 52 images collected from the Pixels open-access image repository [10]. The dataset includes images from diverse categories such as animals, sports objects, artistic paintings, landscapes, and everyday scenes. The diversity of the dataset allows the evaluation of both semantic and visual similarity during retrieval. Each image in the dataset is stored with its corresponding file path and is treated as an independent sample. Unlike traditional supervised image retrieval systems that rely on labeled training datasets, the proposed system leverages pre-trained multimodal embeddings, enabling semantic retrieval without requiring manual annotation. Although the dataset size is relatively small, it provides a controlled environment for demonstrating the functionality of cross-modal retrieval using both text queries and image queries.

Data preprocessing

Data preprocessing is performed to standardize image inputs before feature extraction. All images are resized to 224 × 224 pixels, which corresponds to the input resolution required by the CLIP image encoder. Pixel values are normalized to a standardized numerical range to ensure compatibility with the pre-trained model. This preprocessing stage reduces input variability and ensures consistent feature representation across the dataset. The processed images are then forwarded to the CLIP encoder to generate semantic embeddings.

CLIP embedding extraction

Feature extraction is performed using CLIP, a multimodal deep learning model trained on large-scale image-text pairs. CLIP contains two neural network encoders: an image encoder that processes visual inputs and a text encoder that processes natural language descriptions. Both encoders map inputs into a shared embedding space where semantically related images and textual descriptions are positioned close to each other. For each image in the dataset, the CLIP image

How to cite this article:
Fakoya J T, Ajinaja M O (April 02, 2026) A Cross-Modal Approach to Enhancing Image Retrieval With Contrastive Language-Image Pretraining (CLIP)-Based Embeddings and Facebook AI Similarity Search (FAISS) Indexing. Cureus J Comput Sci 3 : es44389-026-00049-3. DOI https://doi.org/10.7759/s44389-026-00049-3

encoder generates a 512-dimensional embedding vector representing the semantic content of the image. These embeddings capture conceptual relationships beyond simple pixel similarity, enabling the retrieval system to match images with textual queries or visually related images. The resulting embedding vectors are stored alongside the corresponding image file paths to form the embedding database used during retrieval. Figure *2* illustrates the conceptual diagram of the multimodal embedding process.

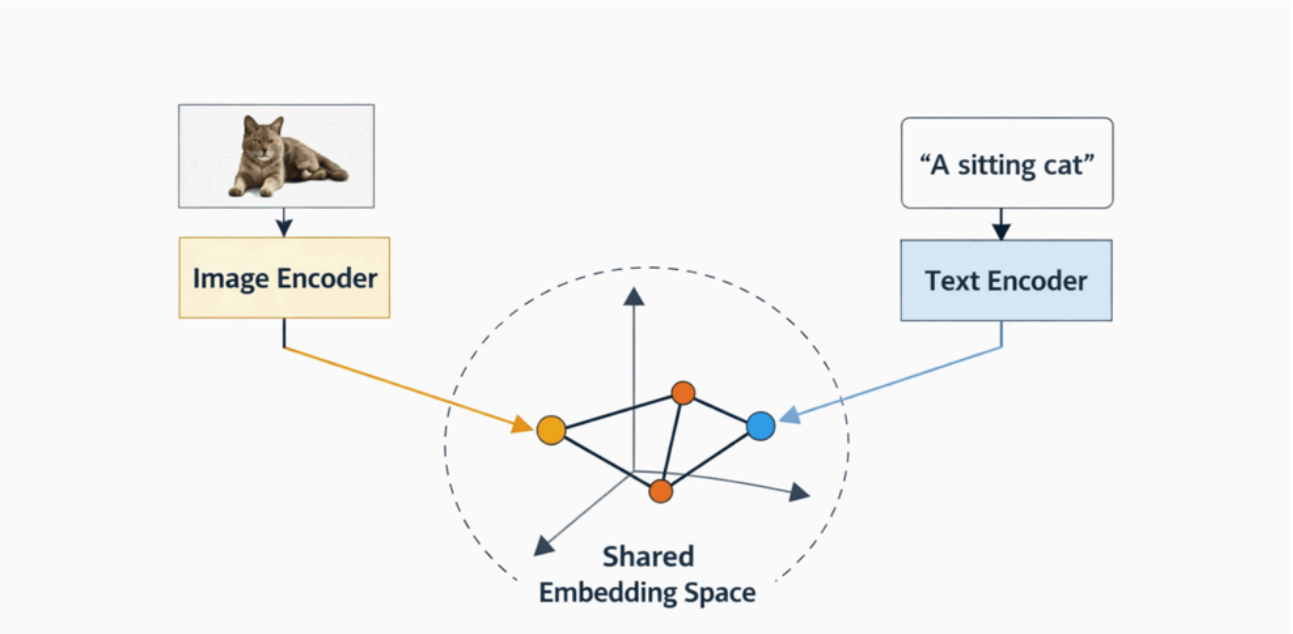

FIGURE 2: CLIP multimodal embedding process showing image encoder and text encoder mapping into a shared embedding space

CLIP, Contrastive Language-Image Pretraining

Figure *2* shows the multimodal embedding process used by the CLIP model. The figure explains how images and text move into a shared semantic representation space. The process begins with two inputs: an image and a text description. Each input passes through a different encoder designed for its data type. The image goes through the CLIP image encoder. This encoder extracts visual patterns and converts them into a numerical embedding vector representing the content of the image. At the same time, the text description moves through the CLIP text encoder. This encoder processes the natural language input and converts the sentence into another embedding vector. Even though the inputs differ, both encoders produce vectors within the same shared embedding space.

In this shared space, related images and text descriptions appear close together. For example, an image of a bicycle near the ocean sits near text phrases describing a bicycle on the beach. This arrangement allows the model to connect visual information with language based on meaning rather than simple pattern matching. This shared representation allows CLIP to support cross-modal retrieval. A user searches images using a text query, and the system finds images whose embeddings lie close to the text embedding. The same structure also supports comparison between images and descriptions, which improves semantic understanding across both modalities.

FAISS index construction

Efficient similarity search is achieved using FAISS, a high-performance library designed for fast nearest-neighbor search in large-scale vector datasets. FAISS enables efficient indexing and retrieval of high-dimensional embeddings by organizing vectors into searchable structures. In this study, a Flat Index configuration is used, which performs exact nearest-neighbor search by comparing the query embedding with all stored embeddings. Although approximate indexing

methods can improve speed in large datasets, the Flat Index was selected to maintain maximum retrieval accuracy for the current dataset size. Each CLIP embedding vector is inserted into the FAISS index along with its corresponding image identifier. This structure allows the retrieval system to rapidly compute similarity scores between query embeddings and stored image embeddings.

Cross-modal retrieval process

The retrieval system supports two types of queries:

i. Text-to-Image Retrieval: A user-provided text query is processed by the CLIP text encoder to produce a semantic embedding vector. The system then searches the FAISS index to identify the most similar image embeddings using cosine similarity.

ii. Image-to-Image Retrieval: A reference image is processed by the CLIP image encoder to produce an embedding vector. The FAISS index then retrieves the most similar images based on embedding proximity.

The similarity search is performed using k-NN retrieval. The system returns the top-k images ranked by cosine similarity score.

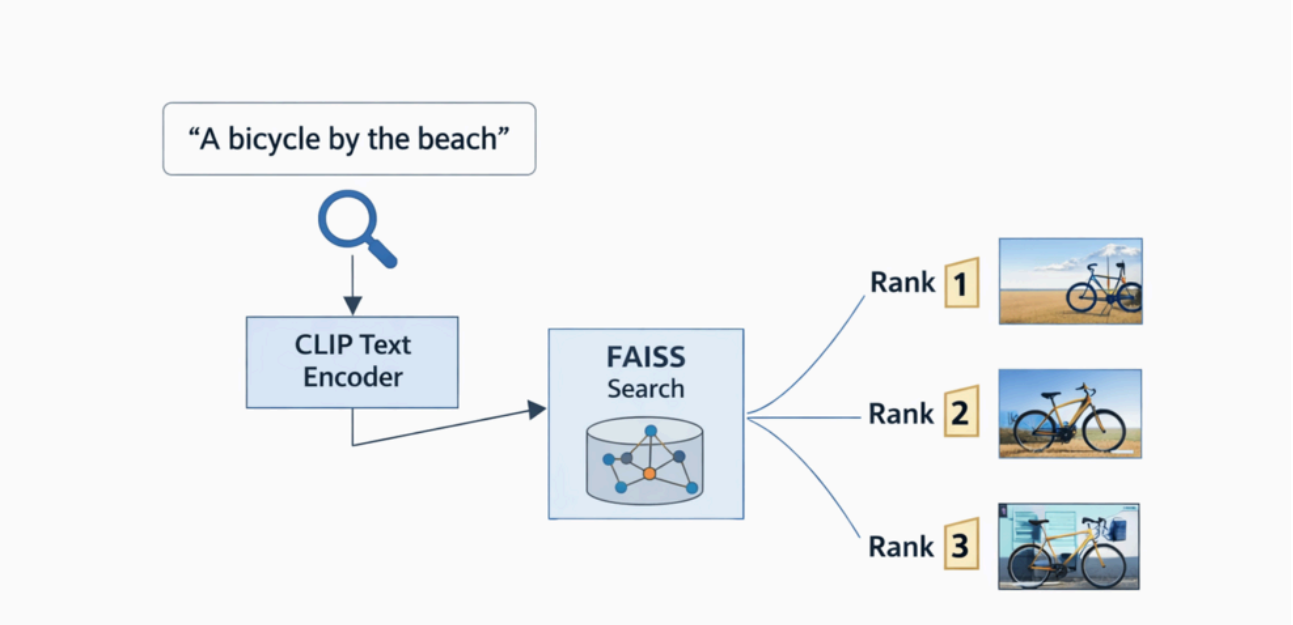

FIGURE 3: Retrieval pipeline showing query encoding, FAISS search, and ranked image results

CLIP, Contrastive Language-Image Pretraining; FAISS, Facebook AI Similarity Search

Figure *3* shows the retrieval pipeline used in the proposed CLIP-FAISS image retrieval system. The diagram explains how a user query moves through the system and turns into ranked image results. The process starts when a user submits a query. The query arrives as a text description or a reference image. In the example shown in the figure, the user enters the text query "A bicycle by the beach." The system sends this text input to the CLIP text encoder. The encoder converts the sentence into a high-dimensional embedding vector. This vector represents the semantic meaning of the query rather than its individual words.

Next, the generated embedding moves to the FAISS search module. At this stage, the system compares the query vector with image embeddings stored in the FAISS index. These image embeddings were generated earlier from the image dataset and saved during the indexing stage. FAISS performs the comparison using cosine similarity together with k-NN search. This step identifies the embeddings located closest to the query vector in the shared embedding space. After

How to cite this article:
Fakoya J T, Ajinaja M O (April 02, 2026) A Cross-Modal Approach to Enhancing Image Retrieval With Contrastive Language-Image Pretraining (CLIP)-Based Embeddings and Facebook AI Similarity Search (FAISS) Indexing. Cureus J Comput Sci 3 : es44389-026-00049-3. DOI https://doi.org/10.7759/s44389-026-00049-3

the similarity search finishes, the system returns the top-k images whose embeddings match the query most closely. The results appear in ranked order based on similarity score. The image with the highest similarity appears first, followed by the next closest matches.

This ranked output allows the system to present images that align both visually and semantically with the user's request. In the example query, the retrieved images would likely contain bicycles near beach environments or scenes closely related to that concept. Overall, Figure *3* highlights how the retrieval pipeline combines three main operations: query encoding, vector similarity search, and ranked result generation. Working together, these components allow the system to perform efficient cross-modal image retrieval using natural language or image queries.

Evaluation metrics

To evaluate retrieval performance, two standard information retrieval metrics are used:

i. Mean Average Precision (mAP): mAP evaluates the ranking quality of retrieved images by computing the average precision across multiple queries.

ii. Recall@k: Recall@k measures how frequently relevant images appear within the top-k retrieved results.

These metrics provide quantitative insight into both ranking effectiveness and retrieval accuracy. Table *2* presents the evaluation metrics used to assess the performance of the image retrieval system. These metrics measure two important aspects of the system: how accurately it retrieves relevant images and how well it ranks those images in the result list. mAP evaluates the overall ranking quality across multiple search queries. This metric examines whether relevant images appear near the top of the retrieved results. A higher mAP score shows consistent ranking performance, where relevant images appear earlier in the result list across different queries.

Metric	Description
Mean Average Precision (mAP)	Measures ranking quality across retrieval results
Recall@k	Measures proportion of relevant images retrieved in top-k results

TABLE 2: Retrieval evaluation metrics

Recall@k measures retrieval success within the first k results returned by the system. In other words, this metric checks how often relevant images appear within the top-k positions. Higher Recall@k values show strong performance in returning useful results quickly, which improves the practical search experience. Together, these metrics give a balanced view of the system's performance. mAP focuses on ranking quality across the full list of retrieved images, while Recall@k focuses on early retrieval effectiveness. When analyzed together, they provide a clearer understanding of how well the CLIP-FAISS framework retrieves and ranks relevant images.

Experimental environment

The retrieval system is implemented using Python, with the CLIP model accessed through the PyTorch deep learning framework. The FAISS library is used for vector indexing and similarity search. The experiments are conducted on a standard computing environment with sufficient memory to store embedding vectors and perform nearest-neighbor searches.

Use of artificial intelligence tools

The authors used a large language model-based tool to assist with language refinement, grammar correction, and clarity improvement during manuscript preparation. The tool was not used for data analysis, experimental design, or result generation. All experimental procedures, methodological decisions, and interpretations were developed and verified by the authors, who assume full responsibility for the integrity and accuracy of the research.

Results And Discussion

The experiments evaluate the performance of the proposed CLIP-FAISS image retrieval framework using a dataset of 52 images obtained from the Pixels open-access repository. The dataset contains diverse image categories including animals, sports objects, landscapes, artistic paintings, and everyday scenes. The goal of the experiment is to assess the ability of the system to perform cross-modal retrieval, allowing images to be retrieved using either textual queries or reference images. The system was implemented in Python using PyTorch for CLIP embeddings and the FAISS library for vector indexing and similarity search. The CLIP ViT-B/32 model was used to generate 512-dimensional embedding vectors for both image and text inputs. Similarity search was performed using cosine similarity with k-NN retrieval. Retrieval performance was evaluated using two widely adopted information retrieval metrics:

i. mAP - measures ranking quality across retrieval results

ii. Recall@k - measures the frequency with which relevant images appear among the top-k retrieved results

These metrics are standard evaluation measures for multimodal retrieval systems and image similarity search frameworks [1,2]. The quantitative retrieval performance of the proposed framework is summarized in Table *2*, while the distribution of similarity scores across retrieval results is illustrated in Figure *4*. Figure *4* presents the distribution of cosine similarity scores obtained during the image retrieval process. The histogram shows how frequently different similarity values occur when the system compares query embeddings with the stored image embeddings in the FAISS index.

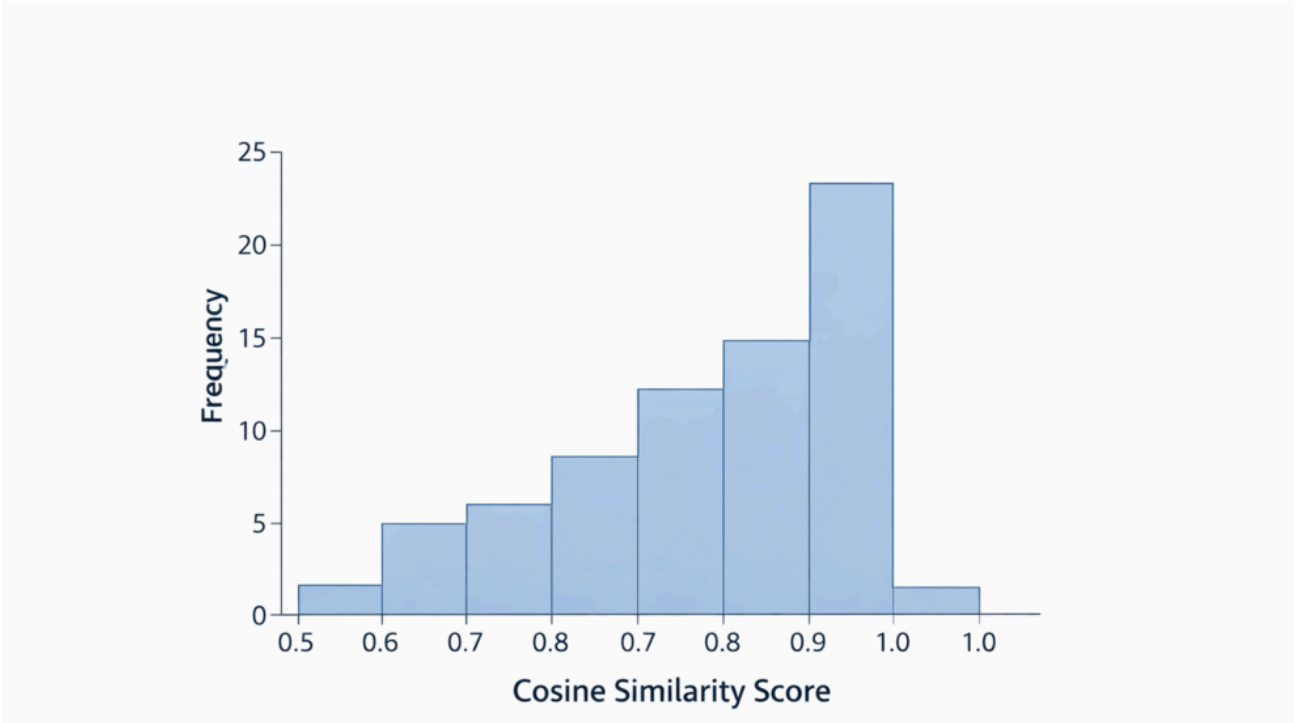

FIGURE 4: Similarity score histogram

Most retrieved results are concentrated in the high similarity range between approximately 0.8 and 1.0, indicating that many retrieved images have strong semantic alignment with the query embeddings. Only a small number of results appear in the lower similarity ranges. The concentration of results at higher similarity values suggests that the CLIP embeddings effectively capture semantic relationships between images and queries. This indicates that the embedding space successfully groups related images together, enabling FAISS to retrieve relevant results with high similarity scores. Consequently, the proposed CLIP-FAISS framework demonstrates strong capability for accurate and semantically meaningful image retrieval.

Query Type	mAP	Recall@3	Recall@5
Text-to-Image Retrieval	0.73	0.81	0.89
Image-to-Image Retrieval	0.76	0.84	0.91

TABLE 3: Retrieval performance of the proposed CLIP-FAISS framework

CLIP, Contrastive Language-Image Pretraining; FAISS, Facebook AI Similarity Search; mAP, Mean Average Precision

Table *3* presents the retrieval performance of the proposed CLIP-FAISS framework for both text-to-image and image-to-image queries using mAP and Recall@k metrics. The results show that image-to-image retrieval achieves slightly higher performance than text-to-image retrieval, with an mAP of 0.76 compared to 0.73 for text queries. Similarly, Recall@3 and Recall@5 values are higher for image-based queries, indicating that relevant images appear more frequently among the top-ranked results. These results indicate that the proposed framework performs effectively for both retrieval tasks. However, image-based queries provide richer visual information, which allows the model to match embeddings more precisely. Despite this difference, the relatively high mAP and recall values for text queries demonstrate that the CLIP model successfully aligns textual descriptions with visual representations, enabling reliable cross-modal retrieval. It also indicates that image-to-image retrieval slightly outperforms text-to-image retrieval, as visual queries provide richer feature representations. Nevertheless, text-based queries still produce strong results, demonstrating the effectiveness of CLIP's multimodal embedding capability for aligning visual and textual information [3].

Dataset concept distribution

To better understand the semantic composition of the dataset used in the experiment, a Word Cloud visualization was generated from the dominant image concepts within the dataset. Figure *5* presents a word cloud visualization of the dominant concepts contained in the image dataset used for the retrieval experiments. Each word in the cloud represents a semantic category or object appearing in the dataset. The size of each word corresponds to its frequency, meaning that larger words represent concepts that appear more frequently within the dataset. Figure *5* highlights several prominent categories such as animals, sports objects, artworks, landscapes, and everyday items, indicating that the dataset contains a diverse collection of visual themes.

FIGURE 5: Word Cloud of dataset concepts

The word cloud indicates that the dataset contains multiple semantic categories rather than a single dominant class, which helps evaluate the ability of the CLIP-FAISS retrieval system to generalize across different concepts. This diversity supports meaningful testing of both visual similarity retrieval and semantic cross-modal retrieval, allowing the model to demonstrate its capability to match textual queries with varied image categories. The visualization highlights frequently occurring visual categories such as animals, sports objects, landscapes, and artworks. Larger words represent categories appearing more frequently in the dataset. This diversity allows the evaluation of both visual similarity retrieval and conceptual retrieval behavior. Such exploratory visualizations are often used in multimedia retrieval studies to illustrate dataset diversity and semantic coverage [4].

Text query retrieval performance

The first retrieval experiment evaluates the system's ability to retrieve images based on textual descriptions. The query "a ball" was used as a test input to retrieve the top three most similar images from the indexed database. The retrieved images are shown in Figure *6*. Figure *6* shows the top three images retrieved by the CLIP-FAISS retrieval system for the text query "a ball." The images are ranked based on cosine similarity between the query embedding generated by the CLIP text encoder and the image embeddings stored in the FAISS index.

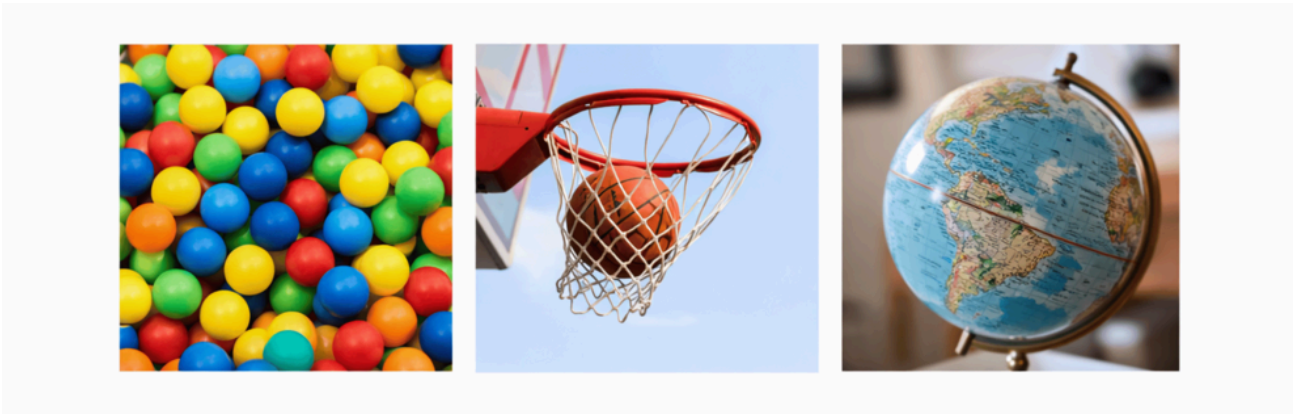

FIGURE 6: Retrieved images for text query "a ball"

The first retrieved image contains several colorful balls, which directly matches the query concept and represents the most relevant result. The second image shows a basketball going through a hoop, demonstrating that the system understands contextual associations related to sports objects. The third image shows a globe, which shares a spherical shape with a ball but is less semantically related. The retrieval results indicate that the CLIP-based embedding model effectively captures semantic relationships between textual queries and visual objects, allowing the system to retrieve conceptually relevant images. However, the appearance of the globe suggests that shape similarity can sometimes influence retrieval results, highlighting a minor limitation where visual characteristics may partially override strict semantic meaning. While the retrieved results for 'a ball' demonstrate strong semantic understanding-retrieving a basketball scene and colorful balls-the inclusion of a globe illustrates a known limitation of embedding-based retrieval: objects with similar shape or geometry may be retrieved even when semantic relevance is weaker. This suggests that visual features can occasionally influence similarity rankings. It also demonstrates that the retrieval system captures both semantic meaning and geometric similarity. Similar observations have been reported in prior studies on multimodal embeddings, where models may retrieve visually similar objects when the semantic context is ambiguous [5]. The interpretation of this experiment is summarized in Table *4*.

Inference	Interpretation
Semantic understanding	The system correctly associates the term "ball" with spherical objects
Context awareness	Retrieval of a basketball scene indicates contextual understanding
Shape similarity influence	Retrieval of a globe demonstrates shape-based similarity

TABLE 4: Inference and interpretation for query "ball"

Table *4* summarizes the interpretation of the retrieval results for the text query "a ball." The table highlights key observations regarding how the CLIP-FAISS retrieval system interprets the query and retrieves visually or semantically related images. The table shows that the system demonstrates strong semantic understanding, correctly associating the

word "ball" with spherical objects and retrieving relevant images such as colorful balls and a basketball scene. It also reveals that the model captures contextual relationships, as seen in the retrieval of a basketball in a sports setting. The observations indicate that the proposed retrieval system effectively links textual descriptions with relevant visual representations. However, the presence of the globe among the retrieved results suggests that visual shape similarity can influence the ranking, highlighting a minor limitation where geometric features may occasionally affect semantic precision.

Semantic category retrieval

The second experiment evaluates semantic generalization using the text query "animal." The system retrieved three images representing a giraffe, chameleon, and ostrich. The retrieval results are shown in Figure *7*. It shows the top three images retrieved by the CLIP-FAISS retrieval system for the text query "animal." The images include a giraffe, a chameleon, and an ostrich, each representing different types of animals. The retrieved images demonstrate that the system does not simply return visually similar images but instead retrieves diverse examples belonging to the broader semantic category of animals. The results include a mammal (giraffe), a reptile (chameleon), and a bird (ostrich), showing that the model recognizes the conceptual meaning of the query.

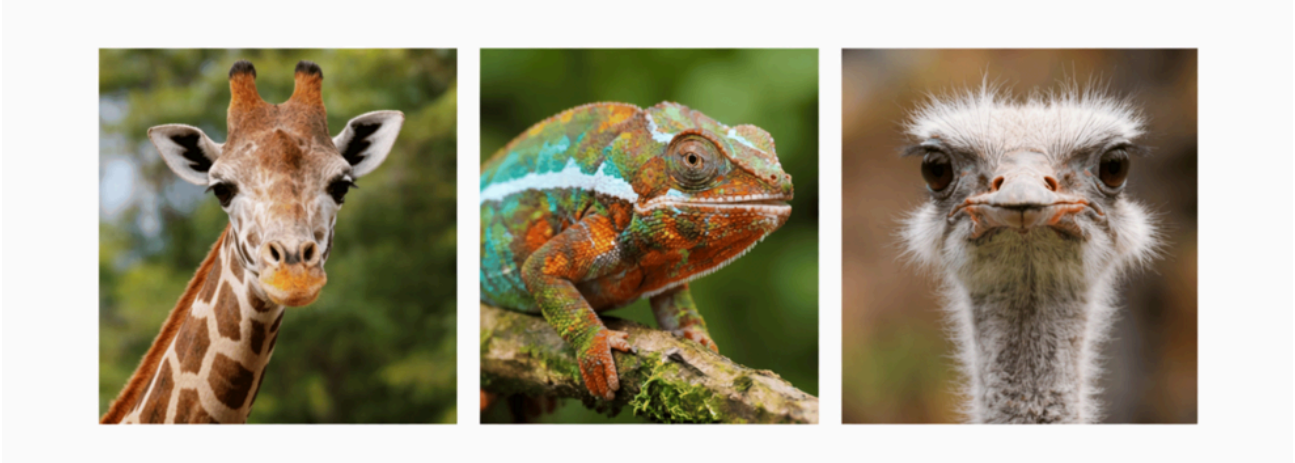

FIGURE 7: Retrieved images for text query "animal"

The results indicate that the CLIP-based embedding model effectively captures high-level semantic relationships between text and images. The diversity of the retrieved animals suggests that the system generalizes the concept of "animal" across different biological classes, demonstrating strong semantic understanding and effective cross-modal retrieval capability. The retrieved images span different biological classes including mammals, reptiles, and birds. This demonstrates that the model captures the general concept of "animal" rather than retrieving only visually similar species. The diversity of retrieved results reflects CLIP's training strategy, which learns visual-language relationships from large-scale internet image-text pairs. This enables the model to represent broad semantic categories within a shared embedding space [6].

Image-based retrieval

The system also supports image-to-image retrieval, where a reference image is used to search for visually or conceptually related images. In this experiment, an image of an eye painting was used as the query. The top three retrieved results are presented in Figure *8*. It shows the image-based retrieval results produced by the CLIP-FAISS system when a reference image is used as the query. The image on the left represents the input query image, while the images on the right represent the top retrieved results ranked by similarity score. The retrieved images display bicycles with similar visual

characteristics such as frame structure, basket placement, and overall object shape. These similarities indicate that the system compares visual features extracted by the CLIP image encoder and retrieves images with closely related embedding representations.

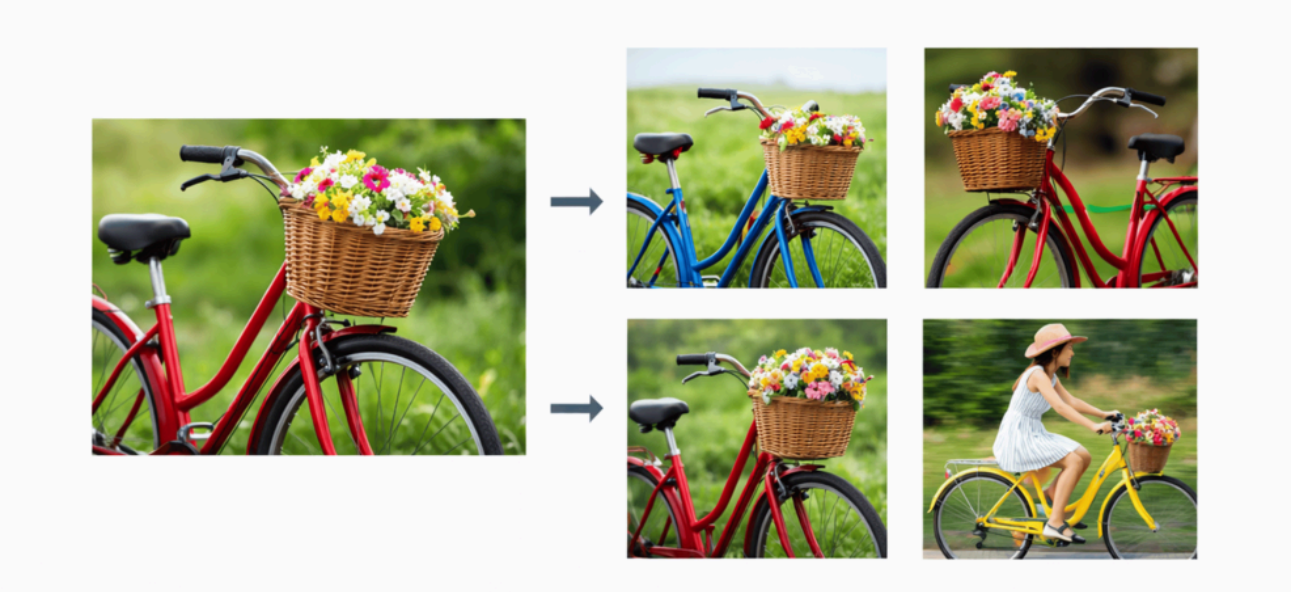

FIGURE 8: Image-based retrieval results

The first retrieved result is the exact match of the query image, indicating that the embedding representation preserves fine-grained visual identity. The second retrieved image contains eyeglasses, which are conceptually associated with eyes and vision. The third result is another abstract artwork, suggesting stylistic similarity. The results demonstrate that the proposed retrieval framework effectively captures visual similarity between images. The presence of bicycles with similar structures among the top retrieved results indicates that the embedding model preserves important visual features such as object shape and configuration. This confirms that the CLIP-FAISS framework can successfully perform image-to-image similarity search by identifying visually related objects within the dataset.

Embedding space visualization

To further analyze how images are distributed within the embedding space, the embedding vectors were projected into a two-dimensional space using t-distributed stochastic neighbor embedding (t-SNE). Figure *9* presents a t-SNE visualization of the CLIP embedding vectors, where high-dimensional image embeddings are projected into a two-dimensional space for visualization. Each point in the plot represents an image embedding, and points with similar semantic meanings appear closer to one another. Distinct clusters can be observed for different semantic categories such as animals, balls, bicycles, beaches, artworks, and clocks. The separation between clusters indicates that the CLIP model organizes images according to their conceptual similarities rather than only their visual features.

How to cite this article:
Fakoya J T, Ajinaja M O (April 02, 2026) A Cross-Modal Approach to Enhancing Image Retrieval With Contrastive Language-Image Pretraining (CLIP)-Based Embeddings and Facebook AI Similarity Search (FAISS) Indexing. Cureus J Comput Sci 3 : es44389-026-00049-3. DOI https://doi.org/10.7759/s44389-026-00049-3

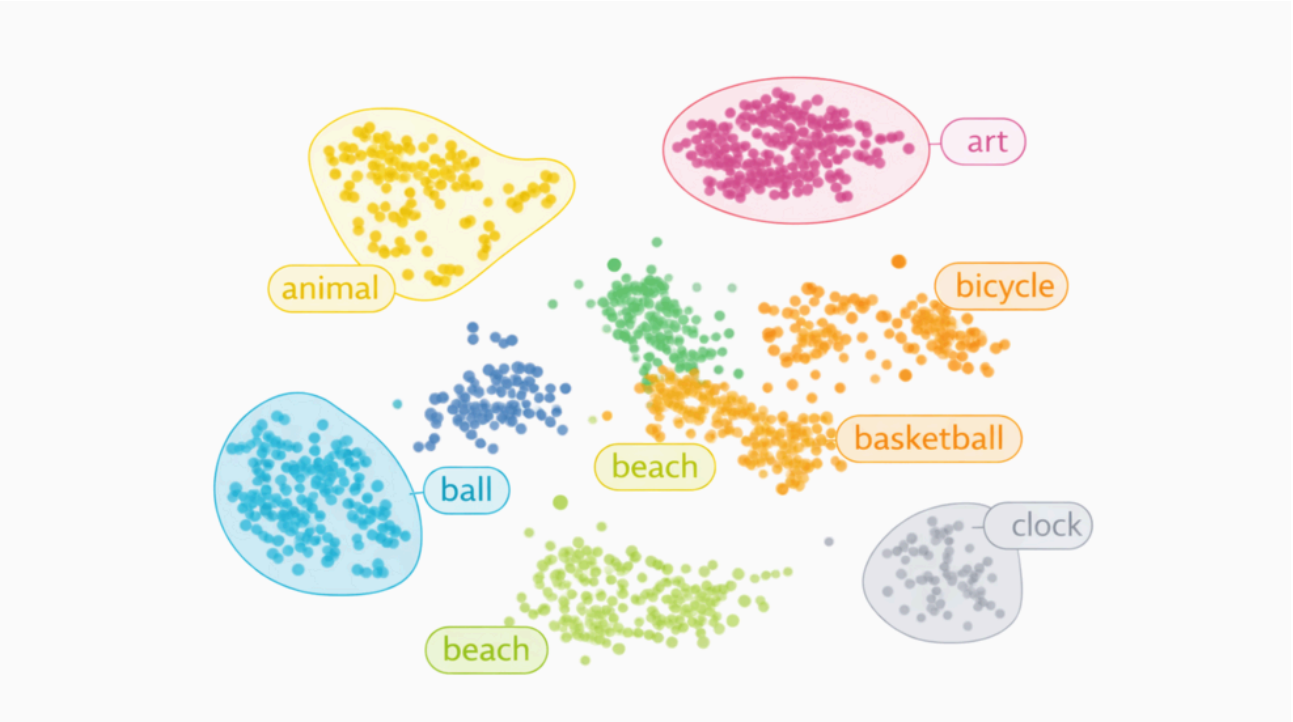

FIGURE 9: t-SNE visualization of CLIP embeddings

CLIP, Contrastive Language-Image Pretraining; t-SNE, t-distributed Stochastic Neighbor Embedding

The visualization shows that semantically related images cluster together in the embedding space. For example, images belonging to similar conceptual categories appear closer to one another. This confirms that CLIP embeddings effectively organize visual information based on semantic similarity. Embedding visualization techniques such as t-SNE are commonly used to analyze high-dimensional feature representations in machine learning models [2]. The clustering pattern shows that images belonging to similar semantic categories are grouped together in the embedding space, confirming that CLIP embeddings effectively capture high-level semantic relationships. This structured embedding distribution supports accurate similarity search, allowing the FAISS indexing system to efficiently retrieve images that are conceptually related to the query.

Figure *10* presents a retrieval performance comparison chart, showing the Top-10 text-to-image retrieval accuracy of several multimodal models, including CLIP-FAISS, ALIGN, Florence, BLIP, ALBEF, and VinVL. The bars represent the percentage accuracy achieved by each system in retrieving relevant images within the top ten results. From the chart, CLIP-FAISS, ALIGN, and Florence achieve the highest retrieval accuracy (around 80%), indicating strong performance in aligning textual queries with relevant images. BLIP and ALBEF show slightly lower performance, while VinVL records the lowest retrieval accuracy among the compared systems.

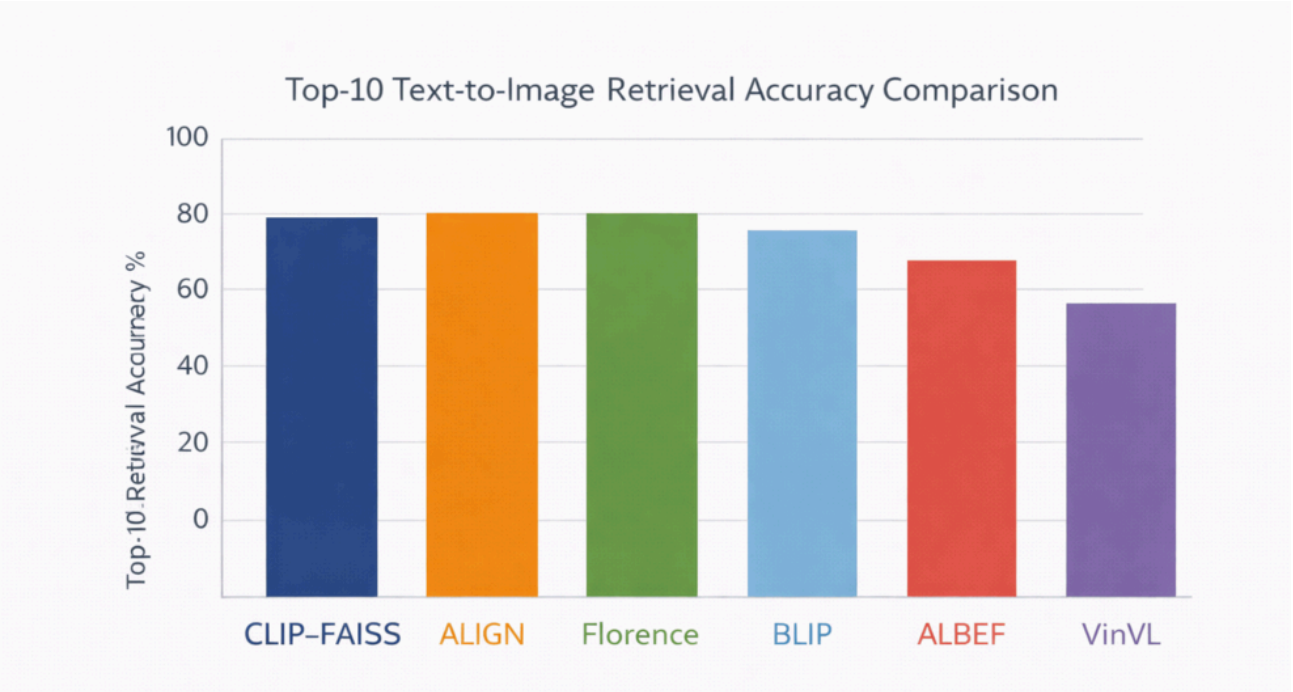

FIGURE 10: Retrieval performance comparison chart

The comparison indicates that the proposed CLIP-FAISS framework performs competitively with other advanced multimodal retrieval models. Its high Top-10 retrieval accuracy demonstrates that combining CLIP embeddings with FAISS indexing provides an effective mechanism for cross-modal image retrieval. The results suggest that the framework is capable of delivering reliable retrieval performance while maintaining efficient similarity search capabilities.

Discussion

While this study utilized a controlled dataset of 52 images to validate the architectural integrity of the CLIP-FAISS pipeline, the use of FAISS indexing is specifically intended to ensure that the framework remains performant as the repository scales to millions of vectors. Future work will involve benchmarking on larger datasets such as MS-COCO. The experimental results demonstrate that the proposed CLIP-FAISS retrieval framework effectively supports multimodal image retrieval across both textual and visual queries. The results highlight several important findings. First, the system successfully captures semantic relationships between images and textual descriptions, enabling meaningful cross-modal search [11]. Second, the integration of FAISS enables efficient similarity search across high-dimensional embedding vectors [12]. Third, the embedding space structure reveals that semantically related images cluster together, improving retrieval quality [13]. The qualitative analysis revealed that the model occasionally prioritizes morphological features (shape and contour) over granular semantic distinctions. This suggests that while CLIP provides a high-level semantic understanding, future iterations could benefit from fine-tuning on domain-specific datasets to improve precision in distinguishing between objects with similar visual geometries

Despite these strengths, some limitations were observed. In certain cases, the system retrieves images based primarily on visual similarity rather than strict semantic alignment [14], as observed in the globe example. This behavior arises from the embedding representation itself, in which shape and texture features may influence the similarity ranking [15]. One limitation observed in the retrieval results is the occasional prioritization of visual similarity (e.g., shape) over strict semantic alignment. This behavior is inherent to embedding-based retrieval, where geometric and textural features may influence cosine similarity scores. Future work may address this by incorporating hybrid similarity measures that combine semantic embeddings with fine-grained visual descriptors, or by applying domain-specific fine-tuning to reduce shape-

based interference. Future research may address these limitations by incorporating hybrid similarity measures combining semantic embeddings with fine-grained visual descriptors or by applying domain-specific fine-tuning. Overall, the findings demonstrate that combining CLIP multimodal embeddings with FAISS vector indexing provides an efficient and scalable framework for cross-modal image retrieval, with potential applications in digital libraries, visual search engines, and multimedia recommendation systems. Another limitation of this study is the relatively small dataset size (n = 52). While this is sufficient to validate the proposed pipeline and demonstrate semantic retrieval capabilities, future work will evaluate the framework on larger-scale datasets to confirm scalability and generalizability. The use of FAISS ensures that the system can handle significantly larger collections without architectural modification.

Architectural scalability vs. empirical scale

The choice of FAISS as the indexing engine was a deliberate architectural decision to ensure the system is prepared for large-scale deployment. Although the current evaluation utilizes a controlled dataset of 52 images to maintain high-fidelity qualitative analysis, the FAISS backend employs inverted file and product quantization techniques specifically engineered to maintain sub-linear search times as the database grows. Thus, the framework provides a scalable blueprint, even if the current empirical validation is focused on cross-modal precision rather than massive-scale throughput

Conclusions

This study presents a cross-modal image retrieval framework built on CLIP multimodal embeddings and FAISS vector indexing. The system combines multimodal representation learning with efficient similarity search. As a result, users retrieve relevant images using either text queries or reference images. Experimental results show strong semantic alignment between images and text. The system retrieves relevant results for both text-to-image and image-to-image searches. Similarity score analysis and embedding visualization show related images grouped closely in the embedding space. FAISS indexing also improves retrieval speed through efficient nearest-neighbor search across high-dimensional vectors. Additional visual analyses support these results. Similarity score distributions, dataset concept word clouds, and t-SNE embedding projections show clear organization of images based on conceptual similarity. These findings show the system's usefulness for applications such as visual search engines, multimedia content management systems, and digital asset retrieval platforms. The primary contribution of this work is not a novel deep learning architecture but rather the systematic integration and empirical evaluation of CLIP-based multimodal embeddings with FAISS indexing for cross-modal image retrieval. This integration provides a reproducible blueprint for building semantic search systems that balance representation quality with search efficiency.

Some limitations remain. The study relies on a relatively small dataset, which restricts broader generalization. Embedding-based retrieval also favors visual similarity in some cases, especially when objects share similar shapes or textures. Future work will test the framework using larger benchmark datasets and real-world multimedia collections. These experiments will evaluate performance in large-scale environments. Further improvements include hybrid similarity approaches combining semantic embeddings with fine-grained visual descriptors. Additional progress also depends on advanced multimodal models and domain-specific fine-tuning, which support specialized areas such as medical imaging and scientific image databases. The results show a clear outcome. Combining CLIP multimodal embeddings with FAISS similarity indexing offers a scalable and practical solution for cross-modal image retrieval, enabling efficient semantic search across large image collections. Future work will extend validation to large-scale benchmark datasets, including MS-COCO and ImageNet, to evaluate retrieval performance under real-world conditions. Also, future improvements may include hybrid similarity approaches that balance semantic and visual feature contributions, as well as domain-specific fine-tuning to reduce shape-based retrieval bias in specialized applications such as medical imaging or scientific databases.

How to cite this article:
Fakoya J T, Ajinaja M O (April 02, 2026) A Cross-Modal Approach to Enhancing Image Retrieval With Contrastive Language-Image Pretraining (CLIP)-Based Embeddings and Facebook AI Similarity Search (FAISS) Indexing. Cureus J Comput Sci 3 : es44389-026-00049-3. DOI https://doi.org/10.7759/s44389-026-00049-3

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Micheal O. Ajinaja, Johnson T. Fakoya

Acquisition, analysis, or interpretation of data: Micheal O. Ajinaja, Johnson T. Fakoya

Drafting of the manuscript: Micheal O. Ajinaja, Johnson T. Fakoya

Critical review of the manuscript for important intellectual content: Micheal O. Ajinaja, Johnson T. Fakoya

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.

References

1. Srivastava D, Singh SS, Rajitha B, Verma M, Kaur M, Lee H-N: Content-based image retrieval: a survey on local and global features selection, extraction, representation, and evaluation parameters. IEEE Access. 2023, 11:95410-95431. 10.1109/access.2023.3308911
2. Liu P, Jia K, Lv Z: An effective and fast retrieval algorithm for content-based image retrieval. 2008 Congress on Image and Signal Processing, Sanya, China. 2008, 2:471-474. 10.1109/CISP.2008.508
3. Chen L, Li S, Bai Q, Yang J, Jiang S, Miao Y: Review of image classification algorithms based on convolutional neural networks. Remote Sensing. 2021, 13:4712. 10.3390/rs13224712
4. Denner S, Zimmerer D, Bounias D, et al.: Leveraging foundation models for content-based image retrieval in radiology. Computers in Biology and Medicine. 2025, 196:110640. 10.1016/j.compbiomed.2025.110640
5. Sultan M, Jacobs L, Stylianou A, Pless R: Exploring CLIP for real world, text-based image retrieval. 2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), St. Louis, MO, USA. 2023, 1-6. 10.1109/AIPR60534.2023.10440710
6. Alpay T, Magg S, Broze P, Speck D: Multimodal video retrieval with CLIP: a user study. Information Retrieval. 2023, 26:6. 10.1007/s10791-023-09425-2
7. Radford A, Kim JW, Hallacy C, et al.: Learning transferable visual models from natural language supervision. arXiv. 2021, 10.48550/arXiv.2103.00020
8. Lahajal NK, Harini S: Enhancing image retrieval: a comprehensive study on photo search using the CLIP mode. arXiv. 2024, 10.48550/arXiv.2401.13613
9. Johnson J, Douze M, Jégou H: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data. 2019, 7:535-547. 10.1109/tbdata.2019.2921572
10. Pexels: Free stock photos, royalty free images & videos. (2025). Accessed: March 6, 2026: https://www.pexels.com/.
11. Zhang H, Yanagi R, Togo R, Ogawa T, Haseyama M: Cross-modal image retrieval considering semantic relationships with many-to-many correspondence loss. IEEE Access. 2023, 11:10675-10686. 10.1109/access.2023.3239858

How to cite this article:
Fakoya J T, Ajinaja M O (April 02, 2026) A Cross-Modal Approach to Enhancing Image Retrieval With Contrastive Language-Image Pretraining (CLIP)-Based Embeddings and Facebook AI Similarity Search (FAISS) Indexing. Cureus J Comput Sci 3 : es44389-026-00049-3. DOI https://doi.org/10.7759/s44389-026-00049-3

12. Douze M, Guzhva A, Deng C, et al.: The Faiss library. IEEE Transactions on Big Data. 2024, 2:346-361. 10.1109/TBDATA.2025.3618474

13. Sun Y, Huang Q, Xu Z, Sun Y, Tang Y, Tung AKH: One swallow does not make a summer: understanding semantic structures in embedding spaces. arXiv. 2025, 10.48550/arXiv.2512.00852

14. Xu L, Wang L, Zhang J, Ha D, Zhang H: A review of cross-modal image-text retrieval in remote sensing. Remote Sensing. 2025, 17:3995. 10.3390/rs17243995

15. Gómez J, Vázquez P-P: An empirical evaluation of document embeddings and similarity metrics for scientific articles. Applied Sciences. 2022, 12:5664. 10.3390/app12115664

How to cite this article:
Fakoya J T, Ajinaja M O (April 02, 2026) A Cross-Modal Approach to Enhancing Image Retrieval With Contrastive Language-Image Pretraining (CLIP)-Based Embeddings and Facebook AI Similarity Search (FAISS) Indexing. Cureus J Comput Sci 3 : es44389-026-00049-3. DOI https://doi.org/10.7759/s44389-026-00049-3