FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 638-649 | e-ISSN: 2347-470X

SAM-CLIP Search: Faster Region-Based Image Similarity Matching Using Lightweight Segmentation & Contrastive Learning

Dr. Umar M Mulani[1], Dr. Mahavir A. Devmane[2], Dr. Satpalsing Devising Rajput[3], Pramod A. Kharade[4], Sagar Baburao Patil[5], Dr. Amol Rajmane[6], Yogesh Kadam[7], Dr. Anindita A Khade[8], Yogesh Bodhe[9], Kuldeep Vayadande[10*]

[1]MIT Art, Design and Technology University, Pune, India; umar.mulani@gmail.com
[2]VPPCOE & VA, Mumbai, India; dmahavir@gmail.com
[3]Pimpri Chinchwad University, Pune, India; rajputsatpal@gmail.com
[4,5]Bharati Vidyapeeth College of Engineering, Kolhapur, India; [4]pramod.kharade@bharatividyapeeth.edu, [5]someone.sagar@gmail.com
[6]JSPM University, Pune, India; amolbrajmane@gmail.com
[7]Bharati Vidyapeeth's College of Engineering Lavale Pune, India; yogesh.kadam@bharatividyapeeth.edu
[8]SVKM'S NMIMS Deemed to be University, Navi Mumbai Maharashtra, India; aninditaac1987@gmail.com
[9]Government Polytechnic, Pune, India; bodheyog@gmail.com
[10]Vishwakarma Institute of Technology, Pune, India; kuldeep.vayadande@gmail.com*

***Correspondence:** *Kuldeep Vayadande; kuldeep.vayadande@gmail.com.*

ABSTRACT- Imagine a designer browsing through an enormous image database for photos that contain both "a chair and a table" or a wildlife scientist attempting to find all photos of "brown bears near water." Locating such specific combinations manually is too time-consuming and cumbersome. To address this issue. This system SAM-CLIP Search is an area-based vision-language image search platform that incorporates CLIP and vision-language embeddings into the Segment Anything Model (SAM) to offer flexible prompt-based segmentation. Our approach makes precise picture search with point, box, or text prompts feasible compared to typical CBIR approaches, which often struggle with multi-object queries as well as cross-modal alignment. We propose a Ranking Optimization Layer (ROL) that produces context-specific relevance scores by aggregating spatial overlap (IoU) with semantic embedding distance and we substitute the conventional FAISS indexing with a light-weight cosine similarity approach to improve efficiency. Our method yields semantically and visually coherent matches on the COCO (val2017), Flickr30K and Fashion200K benchmarks. With maintaining fast inference, SAM-CLIP Search is better in critical metrics such as Recall@K, mAP and NDCG compared to baselines such as ViT+KNN, Deep Image Retrieval (DIR)and CLIP-only models. Its suitability for high-impact applications such as content curation, surveillance and medical image analysis is exemplified by a user study that verifies its effectiveness in difficult scene searches. The SAM-CLIP Search model achieves a retrieval accuracy of 92% with Recall@5 of 89.3%, Precision@5 of 87.1% and an average top-1 similarity score of 0.78 which makes SAM- CLIP Search Faster Region-Based Image Similarity Matching Using Lightweight Segmentation and Contrastive Learning a more accurate and efficient approach for region-based image retrieval.

Keywords: Image Retrieval, Segment Anything Model, CLIP (Contrastive Language-image Pretraining), Prompt-based Segmentation, Vision language Embeddings, Box prompt, Point Prompt, Object Detection, Image Segmentation, Visual Search.

ARTICLE INFORMATION

Author(s): Dr. Umar M Mulani, Dr. Mahavir A. Devmane, Dr. Satpalsing Devising Rajput, Pramod A. Kharade, Sagar Baburao Patil, Yogesh Bodhe, Yogesh Kadam, Dr. Anindita A Khade, and Kuldeep Vayadande;

Received: 19/06/25; **Accepted:** 14/10/25; **Published:** 30/11/25;
E- ISSN: 2347-470X;
Paper Id: IJEER250121;
Citation: 10.37391/ijeer.130402
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-13/ijeer-130402.html

Publisher's Note: FOREX Publication stays neutral with regard to jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

Effective and smart image searching has emerged as an essential issue in today's digital age as a result of the widespread spread of visual material on platforms such as scientific repositories, stock picture collections and surveillance networks. When users are searching for specific segments of an image or semantically comparable content, classic keyword-based image search often fails to capture the contextual and visual significance of items. Imagine a fashion consultant browsing through hundreds of lifestyle images for "red shoes next to a handbag" or a researcher who must pull out all images that contain "a brown bear close to a river." Beyond object recognition, these activities call upon exact object localization

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 638-649 | e-ISSN: 2347-470X

and semantic understanding, which are challenges that classical retrieval algorithms are not ready to face.

Our work introduces a modular vision pipeline that synergizes vision-language modeling and prompt-based object segmentation, two recent AI advances, to address these limitations. To close the gap between vision content and natural language comprehension, we apply CLIP (Contrastive Language–Image Pretraining) and the Segment Anything Model (SAM) for user-tunable object segmentation in terms of box, point, or text prompts. The pipeline begins with preprocessing of the dataset, where CLIP's image encoder is utilized to extract and save picture features. SAM divides the correct region when a query is provided, either as a textual description, bounding box, or point. CLIP is utilized to compute a high-dimensional embedding of the region. For retrieval of the most relevant images, this embedding is searched against the dataset with the cosine similarity.

The primary objective of this work is to develop a lightweight, modular region-based image retrieval system that effectively integrates the Segment Anything Model (SAM) and Contrastive Language–Image Pretraining (CLIP) for accurate, timely visual search. By achieving ranking optimization with lightweight mechanisms, multi-object query management, and prompt-based segmentation, the proposed system attempts to enhance retrieval performance.

2. LITERATURE REVIEW

The efficient deep convolutional neural network SegNet with encoder-decoder architecture [1] inspired by the VGG16 network, is introduced in this work for semantic pixel-wise segmentation. SegNet's application of max-pooling indices of the encoder to non-linear up sampling at the decoding stage, which reduces memory usage and simplifies learning, is its uniqueness. This architecture maintains competitive accuracy but eliminates the need for learning up sampling parameters. SegNet strikes a good balance between accuracy, computational cost and memory usage, making it suitable for real-time applications of scene understanding such as road and indoor scene segmentation. The efficient deep convolutional neural network SegNet with encoder-decoder architecture inspired by the VGG16 network [2] is introduced in this work for semantic pixel-wise segmentation. SegNet's application of max-pooling indices of the encoder to non-linear up sampling at the decoding stage, which reduces memory usage and simplifies learning, is its uniqueness. This architecture maintains competitive accuracy but eliminates the need for learning up sampling parameters. SegNet strikes a good balance between accuracy, computational cost and memory usage, making it suitable for real-time applications of scene understanding such as road and indoor scene segmentation. Higher-level features borrowed from pre-trained deep convolutional neural networks which have been initially trained on large-scale image classification tasks—are employed in this paper's Content-Based Image Retrieval (CBIR) system [3]. The method significantly decreases the "semantic gap" and performs better than traditional methods in retrieval by employing this semantic rich information. To boost retrieval speed without losing accuracy, the authors also introduce a pre-clustering strategy, which renders the method efficient and effective for large image databases.

This project highlights [4] an image similarity retrieval engine that finds sports photographs matched with similar artworks, inspired by @ArtButSports. For the extraction of features, it utilizes deep learning through pre-trained ResNet34 and ResNet50 models, chosen due to their great performance and stability. Several image preprocessing methods were applied to enhance similarity detection, like threshold masking, edge detection, dimming and a novel Segmented Edge detection method that reduces background noise and highlights people. For the purpose of assisting in attention to significant parts, like persons in the image, YOLO was also employed for multi-object detection. The effectiveness of detecting visually similar images was significantly enhanced by these approaches.

3. METHODOLGY

The aim of the proposed project is to develop a modular vision pipeline for object-aware image retrieval by combining vision-language embeddings with prompt-based segmentation. The pipeline integrates CLIP (Contrastive Language-Image Pretraining) to evaluate the semantic similarity between visual and textual inputs with the Segment Anything Model (SAM) for precise region segmentation. The three primary phases of the system's functionality are image retrieval, prompt-based segmentation and dataset preprocessing.

3.1. Single Object Query

It is a modular vision pipeline that enables object-aware image retrieval by combining vision-language embeddings with segmentation based on prompts. The methodology integrates CLIP (Contrastive Language-Image Pretraining) to estimate the semantic similarity of textual and visual inputs with the Segment Anything Model (SAM) for region segmentation. The three primary phases of the proposed system's operation are image retrieval, prompt-based segmentation and dataset preprocessing. The CLIP image encoder is applied to encode natural images during dataset pre-processing in order to create feature embeddings, which are subsequently stored for fast retrieval. During the prompt-based segmentation stage SAM identifies and classifies the correct object regions from a range of input cues in the form of text descriptions, bounding boxes and pointers. The most relevant images are recovered in the image retrieval process by mapping segmented regions of the left and right images into feature vectors and comparing them with precomputed embeddings of datasets in cosine similarity.

3.2. Dataset Collection

1. We evaluate our method on three publicly available datasets:
– COCO 2017 as available on Kaggle[14]
– Flickr30K (Kaggle repository) [15]
– Fashion200K Kaggle repository [16]
2. These datasets encompass natural scenes (COCO), general-purpose image–text pairs (Flickr30K), and fashion or product

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 638-649 | e-ISSN: 2347-470X

images (Fashion200K). For each dataset, we use the standard validation/test splits for benchmarking.

3.3. Dataset Preprocessing

A comprehensive preprocessing process is performed on a carefully curated dataset of natural images in preparation for efficient similarity search. Every image is first passed through the CLIP image encoder during this process, which is a model designed to yield high-dimensional feature embeddings that capture the semantic content of the images. These embeddings represent the visual information in the form of vectors that support fast comparison and retrieval. The right picture file paths are also stored with these feature embeddings, supporting easy and fast access to the pictures when required. File path integrity checks are performed to ensure that every path references an actual file in order to ensure a reliable and smooth retrieval process.

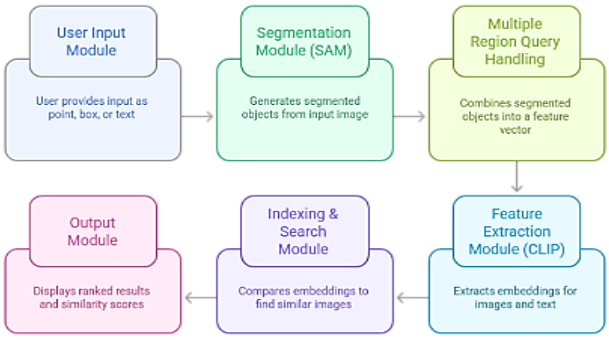

Figure 1. Architecture diagram for SAM- CLIP Search

Fig. 1 architecture diagram illustrates a pipeline for efficient retrieval of images employing contrastive feature encoding and segmented regions. A user inputs an image and a prompt which may be a text message, a point, or a bounding box at the Frontend Interface to initiate the process. The two main branches to which this input is directed are the Segmentation Module and the Feature Encoder. The segmentation mask is generated in the Segmentation Module with the Segment Anything Model (SAM). Region masks are generated with the SAM AutoMask Generator if the input is a text prompt. The segmented (masked) image is then passed to the Feature Encoder after the SAM Predictor refines the segmentation based on user input. The CLIP model is utilized to execute the two components of the feature encoder: a text encoder (if there is a text prompt) and an image encoder (for the masked area). The Similarity Engine is given the features that these encoders have pulled out. This engine computes the cosine similarity between the input's encoded features and those stored in a feature database with picture paths, metadata and precomputed CLIP features of dataset parts. The Output Module takes the top-K from the Similarity Engine, which orders the dataset images by similarity. The Output Module may then return the image paths and similarity values. The module may also return the input picture masked. The user is then presented with this information, occasionally along with visual hints such as the segmented region.

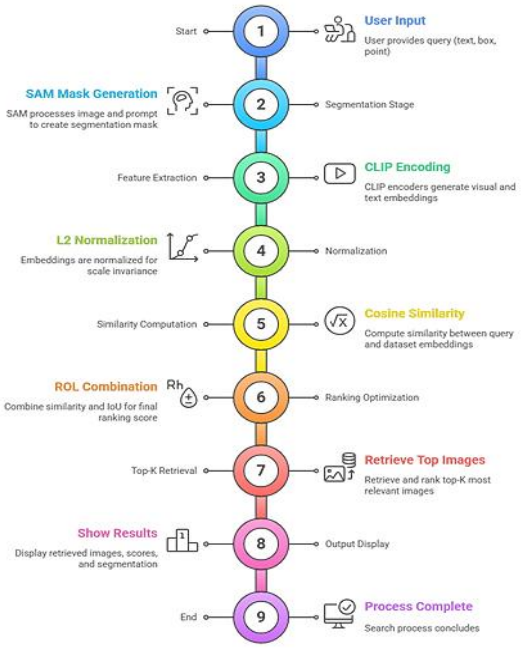

Figure 2. Flow chart for SAM- CLIP Search

The SAM-CLIP search flow is depicted in *figure 2* where user input is first entered as text, box or point prompts. Following the creation of a segmentation mask for the region of interest by the SAM module CLIP encoders extract text and visual embeddings afterwards Cosine similarity is used to compare these embeddings with dataset embeddings after they have been normalised using L2 normalisation. The similarity score is combined with spatial overlap (IoU) in a ranking optimisation layer to improve it. After retrieving the Top-K most pertinent photos the system shows the results along with optional masks and similarity scores before ending the search.

3.4. Segmentation using Prompt inputs

One of the most important features of the proposed system is the segmentation module, which is segmented based on the Segment Anything Model (SAM) and supports customizable input prompts in order to effectively guide the process of segmentation. SAM is an extremely versatile model that can partition objects from images based on varying user inputs. The user can define the region or object to be segmented in three different ways through the three primary input prompt types available in the system: Box Prompt, Point Prompt and Text Prompt.

3.5. Feature Extraction and Similarity Search

Following the segmentation of the target region using techniques like text prompts, point annotations, or bounding boxes, it is converted into a format that works with the CLIP (Contrastive Language–Image Pretraining) model. The CLIP image encoder is then used to extract high-level semantic feature vectors from this segmented image, capturing the region's visual and contextual information. To guarantee consistency in magnitude and enable consistent comparison across various embeddings, these feature vectors are then L2-normalized. Using cosine similarity, that calculates the angular distance between vectors to produce a measure of similarity, the

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 638-649 | e-ISSN: 2347-470X

normalized vectors are compared to a precomputed set of image embeddings from a dataset. By this, top-K most similar pictures are found and ranked based on similarity scores and a new set of resultant images are produced that are very similar to the semantics of the original segmented region.

With the Segment Anything Model (SAM), point and box prompts segmentation selects specific item parts from an image with minimal user input. SAM is able to segment information through the use of a bounding box or an individual point, a specific pixel coordinate, as a prompt. The concerned region of the image is identified and segmented by SAM through an input point labeled as either negative or positive. Similarly, if a bounding box is given SAM produces a segmentation mask from its perception of the bounded region as the target region. With minimal manual annotation, this process allows for flexible and efficient object extraction.

3.5.1. Feature Extraction and Similarity Search for Box Prompt

Given a Bounding Box

$$B = [x_0, y_0, x_1, y_1] \quad (1)$$

Eq. 1 shows define a bounding box which is a rectangular region specified by two corner coordinates. Whether a prompt is a point or bounding box, SAM transforms it into a feature vector that stores the input spatial and contextual information. The encoded representation is utilized by the model to make predictions on a segmentation mask, symbolized as $\widehat{M}$, in which all the pixels within the image are classified as either target object or background. The generated binary mask, depending on the prompt, correctly brings into view the specific region of interest as shown in below *eq. 2*.

$$\widehat{M}(i,j) = \begin{cases} 1 \; if \; pixel \; belongs \; to \; object \\ 0 \; otherwise \end{cases} \quad (2)$$

Whether a prompt is a point or bounding box, SAM transforms it into a feature vector that stores the input spatial and contextual information. The encoded representation is utilized by the model to make predictions on a segmentation mask, symbolized as $\widehat{M}$, in which all the pixels within the image are classified as either target object or background. The generated binary mask, depending on the prompt, correctly brings into view the specific region of interest.

3.5.2. Feature Extraction and Similarity Search for Point Prompt

Given a point P= (x, y) P = (x, y) P= (x, y) with a label (positive/negative), SAM similarly encodes the location and label to guide segmentation.

To single out and obtain the desired object, the binary segmentation mask $\widehat{M}$ is then applied to the initial image after its creation. In order to accurately zero in on the specific object of focus, this process involves blanking out all areas outside of the divided space. Only relevant visual material remains in the resultant masked image, which can then be used for further processing or analysis. *Eq. 3* shows that it us used to generate the masked image I' by applying the binary $\widehat{M}$ to the original image I.

$$I' = I.\widehat{M} \quad (3)$$

Eq. 4 shows the extraction a feature vector from the masked image I' using the CLIP image encoder and then L2-normalizes it. we use the CLIP model to convert this masked image into a feature vector. $f_{CLIP}(I')$ gives the raw embedding from CLIP.

$$f_{img} = \frac{f_{CLIP}(I')}{\|f_{CLIP}(I')\|2} \quad (4)$$

Eq. 5 is the general formula for cosine similarity which is used to measure the semantic similarity between two feature vectors f_1 and f_2. The feature vector extracted from the masked image is matched against a collection of precomputed image embeddings to identify visually or semantically similar images in a dataset. Cosine similarity, which measures the angular distance between the vectors to determine how closely they point in the same direction in the feature space, is employed for matching. The most similar images can be effectively and precisely retrieved based on semantic content due to the L2-normalized CLIP feature vectors, which reduces the cosine similarity calculation to a dot product.

$$im(f_1, f_2) = \frac{f_1.f_2}{\|f_1\|2.\|f_2\|2} \quad (5)$$

Since CLIP features are normalized *eq. 6* presents a simplified version of the cosine similarity calculation.

$$\text{sim} = f_1 \, . \, f_2 \quad (6)$$

Since CLIP feature vectors are L2-normalized (unit norm) the denominator becomes 1, reducing the similarity computation to a simple dot product: sim= $f_1.f_2$. This simplification speeds up similarity computation while maintaining accuracy, making the search for similar images highly efficient.

3.6. Algorithm Pipeline

The technique merges CLIP (Contrastive Language-Image Pretraining) and SAM (Segment Anything Model) to receive the Top-K most useful images from a collection of photos and a query entered by the user, which may be text, a bounding box, or a point.

Algorithm 1: Region-Based Retrieval using SAM-CLIP

Input: User prompt P, Dataset $\mathcal{D}$, Prompt type: text/box/point
Output: Top-K most relevant images

1. Dataset Embedding Phase:
For each image $I \in \mathcal{D}$:
Compute image embedding:

$$f_I = \text{CLIP}_{img}(I), \hat{f}_I = \frac{f_I}{\|f_I\|_2}$$

Store $\hat{f}_I$ in embedding index $\mathcal{E}$

2. Query Embedding Phase:
If P is text:

$$f_Q = \text{CLIP}_{text}(P)$$

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 638-649 | e-ISSN: 2347-470X

If P is point/box:
Segment region R using SAM:

$$R = \text{SAM}(I_{query}, P)$$

Apply mask and encode:

$$f_Q = \text{CLIP}_{text}(I_{query}.R)$$

3. Query Normalization:

$$\hat{f}_Q = \frac{f_Q}{\|f_Q\|_2}$$

4. Cosine Similarity Search:
For all $\hat{f}_I \in \mathcal{E}$:

$$\text{sim}(\hat{f}_Q, \hat{f}_I) = \hat{f}_Q \cdot \hat{f}_I$$

5. Optional Ranking Optimization Layer (ROL):

$$\text{score} = \alpha \cdot \text{sim} + \beta \cdot \text{IoU}(R_Q, R_I)$$

6. Return: Top-K images ranked by score

All the images within the dataset are processed by the CLIP image encoder to yield an embedding vector at the beginning of the process, which is the dataset embedding process. In order for successful retrieval to occur, the embeddings are normalized and stored in an index. The user query is then embedded at the query embedding phase. The prompt is encoded if it is in text format by the CLIP text encoder. The CLIP image encoder is utilized to generate the query embedding after the SAM (Segment Anything Model) has identified the relevant region from the query image if the prompt is a region (point or box). The query embedding is also normalized to facilitate comparisons of cosine similarity.

The pertinence is then quantified by calculating cosine similarity between all the dataset image embeddings and normalized query embedding. Using a weighted sum controlled by parameters α and β by combining the cosine similarity with Intersection over Union (IoU) between the segmented query area and dataset image areas, optionally a Ranking Optimization Layer (ROL) can be applied to enhance the retrieval process. Finally, the Top-K most relevant images are returned as the final output after the images with the best combined score are selected. Both text-based and image-based area-based retrieval tasks can be aided by this combined approach that utilizes both semantic knowledge through CLIP and precise region correspondence with SAM.

The mathematical expressions presented in this section define how the SAM-CLIP framework translates user prompts into computable spatial and semantic representations. *Equations (1)–(3)* formalize the generation of segmentation masks using the Segment Anything Model (SAM) based on box and point prompts, ensuring accurate localization of objects. *Equations (4)–(6)* describe how these segmented regions are encoded as feature vectors using the CLIP model, and how cosine similarity measures the semantic closeness between query and dataset embeddings.

3.7. Mathematical Modeling

- $f_T = \text{CLIP}_{\text{text}}(T)$, the text embedding
- $f_I = \text{CLIP}_{\text{img}}(I)$, the image embedding
- M be the binary mask generated by SAM for a region of interest

For region-masked images:

$$I' = I \cdot M, f_R = \text{CLIP}_{\text{img}}(I') \quad (7)$$

Cosine similarity is computed as:

$$\text{sim}(f_Q, f_I) = \frac{f_Q}{\|f_Q\|_2} \cdot \frac{f_I}{\|f_I\|_2} \quad (8)$$

For multi-object queries or spatial alignment, the ranking function is:

$$\text{score}_{\text{final}} = \alpha \cdot \text{sim}(f_Q, f_I) + \beta \cdot \text{IoU}(R_Q, R_I)$$

Where:

- α, β are scalar weights,
- R_Q, R_I are segmented regions of query and candidate images.

As shown above, f_T *and* f_I represents to the CLIP-generated embeddings for text and image inputs respectively, while M denotes the binary segmentation mask produced by SAM that isolates the region of interest. The masked image $I' = I \cdot M$ ensures that only the relevant object area contributes to the visual embedding f_R. Moving to the next the cosine similarity function (f_Q, f_I) measures the semantic closeness between the query and the candidate image embeddings after L2-normalization to ensuring scale-invariant comparison in the embedding space. For multi-object retrieval the final ranking score combines semantic similarity and spatial overlap through a weighted sum of cosine similarity and Intersection over Union (IoU). Here, α and β are scalar weights that balance the influence for textual and visual similarity with geometric alignment between the segmented regions R_Q, R_I This hybrid scoring function allows the system to achieve more accurate and context-aware retrieval across diverse visual domains. *Equations (7)* and *(8)* describe how the system compares images based on their visual features. First, the image is refined using a mask to focus only on the important parts, and the CLIP model extracts meaningful feature representations from this refined image. Then, the cosine similarity between the query and image features is calculated to measure how closely they match where a value closer to 1 indicates a higher similarity.

3.8. Multiple Object Query

It is a modular vision pipeline that aligns vision-language embeddings with prompt-based segmentation to allow object-aware image retrieval from multiple object queries. The method brings together the Segment Anything Model (SAM) for accurate object region segmentation and CLIP (Contrastive Language-Image Pretraining) to approximate semantic similarity between textual and visual inputs. The three major phases of system operation are picture retrieval, prompt-based

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 638-649 | e-ISSN: 2347-470X

segmentation and dataset preprocessing. The CLIP image encoder is employed to derive feature embeddings from natural images in the process of dataset preparation. These are pre-stored beforehand for efficient retrieval. During the prompt-based segmentation process, SAM employs user-supplied indications such as text descriptions, bounding boxes, or point clicks to segment and label object regions. For queries including multiple objects. The CLIP image encoder encodes each of the multiple object regions that are available to users.

4. RESULTS AND DISCUSSIONS

The system was evaluated on the COCO dataset with different types of prompts to measure the effectiveness of the proposed image segmentation and retrieval pipeline. This allowed the performance of the model to be tested in a broad variety of input situations.

The user begins by delineating the region of interest within the image by a bounding box. That input is then passed through the Segment Anything Model (SAM) and successfully dissects the object of interest into an accurate segmentation mask, as in *fig. 4*. This masked region is then used as input to extract semantic information by the CLIP model. To obtain visually similar images, these attributes are matched against a precomputed embedding dataset. The effectiveness of the pipeline is shown in *fig. 5*

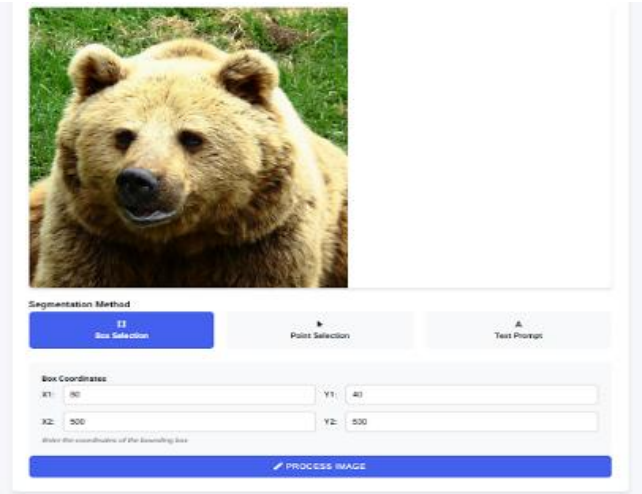

Figure 3. Input image for box selection

The input image is shown in *fig. 3*, in which the user begins by specifying the region of interest through a bounding box.

Figure 4. Showcased the background erased from the input image

The result of the segmentation process performed by SAM from the provided bounding box is demonstrated in *fig. 4*.

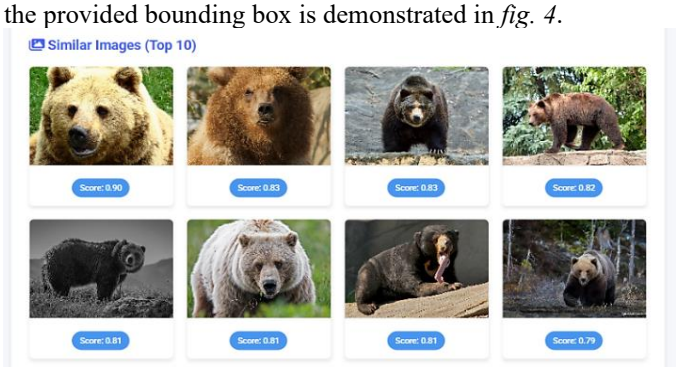

Figure 5. Similar images obtained from box selection SAM- CLIP Search

By matching the features of the segmented region with the features stored in the system, *Fig. 5* shows the visually similar images that were retrieved from the embedding database. The object of focus within the image is marked by a single point. This point prompt is employed by the Segment Anything Model (SAM) to produce a segmentation mask centered on the specified position. The model neatly slices the targeted object as evident in *fig. 6*.

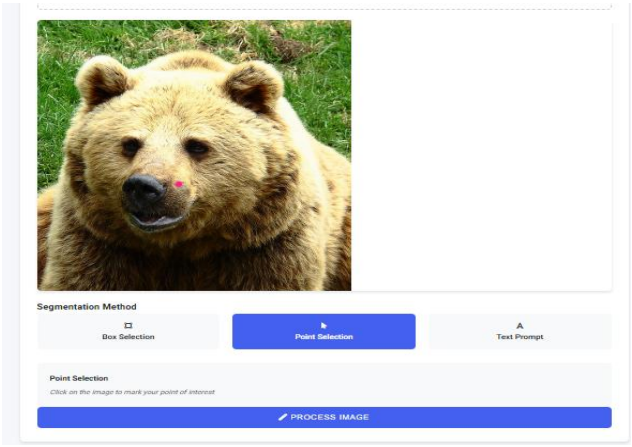

Figure 6. Input image for point selection

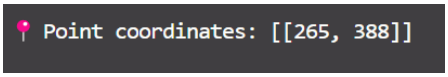

Figure 7. Point coordinates for point selection

Fig. 7 shows the points selected for the image retrieval. *Fig. 8* shows the segmented output from the point-based prompt is shown in *fig. 7*.

Figure 8. Showcased the background erased image

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 638-649 | e-ISSN: 2347-470X

Fig. 9 shows the set of pictures obtained following the use of point-based segmentation and the encoding of the masked region using the CLIP model is displayed in *fig. 8.*

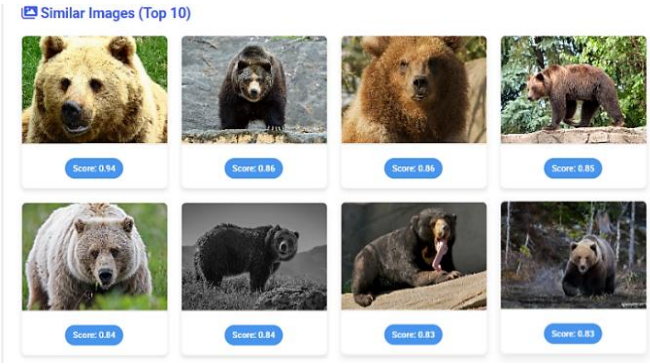

Figure 9. Images obtained from point selection

The model finely cuts the desired object, as presented in *fig. 6*, demonstrating the precision and flexibility of point-based prompting in complex visual settings.

$$\text{similarity} = f_{text} \cdot f_{masked\_region} \quad (7)$$

Eq. (7) shows the similarity calculation between the text and the segmented visual data is shown in *equation (7)*. This step employs a visual encoder function f_{text} to examine the masked area, which is derived after applying the segmentation mask.

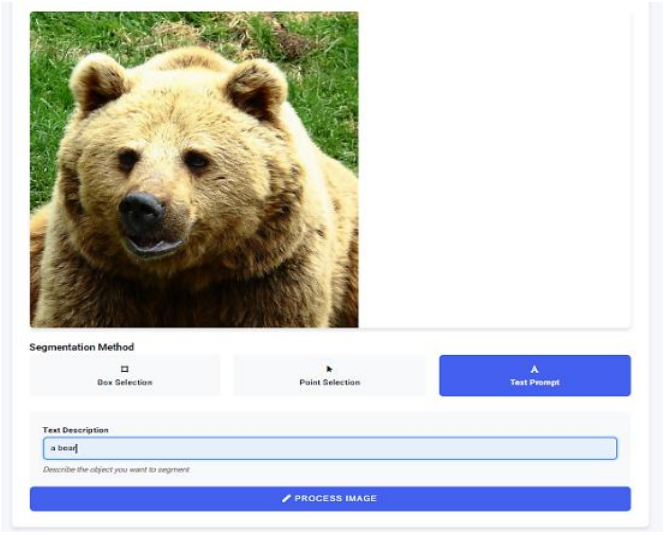

Figure 10. Input image for single object selection

The input text that the user provided is observed in *Fig. 10*.

Figure 11. Input text prompt

Fig. 11 shows employ the text prompt to show the outcome of the segmentation. The combination of visual and textual information is illustrated by the SAM model.

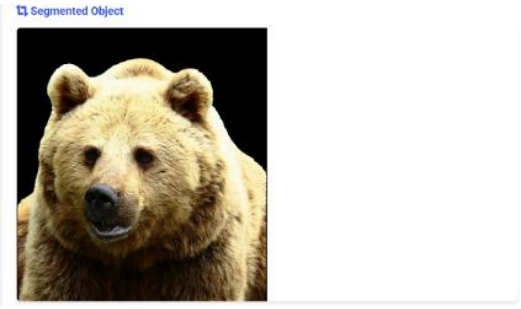

Figure 12. Background erased image

The set of visually similar images that were retrieved using the segmented output of the text prompt is presented in *fig. 13*.

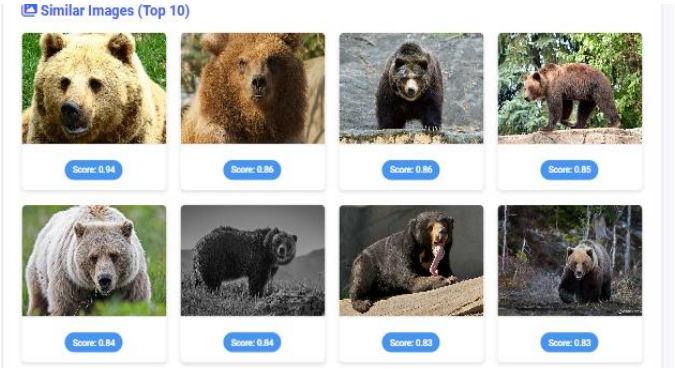

Figure 13. Similar images obtained

The results of multi-item queries will now be evaluated based on the use of three separate input cues: text prompts, box selection and point selection. Using these different interaction modalities, the system employs SAM to divide up many regions of interest and CLIP to produce their embeddings. To observe how each of these query types influences the precision and relevance of the retrieval results, semantically similar images are retrieved using the averaged embeddings of the selected objects.

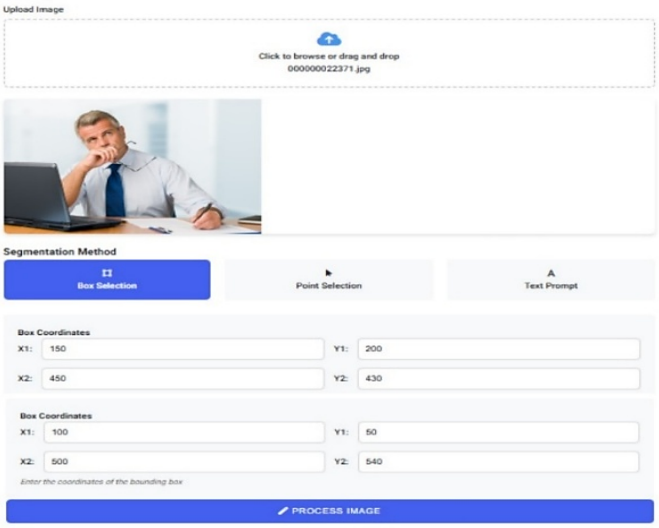

Figure 14. Input image for multiple object selection

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 638-649 | e-ISSN: 2347-470X

Fig. 14 shows that the input image for multiple object selection is shown in *fig. 14* and is processed through box selection with the SAM-CLIP search framework Faster Region-Based Image Similarity Matching Using Lightweight Segmentation and Contrastive Learning. The input image is segmented using SAM in this step, where a number of items are chosen utilising several input techniques such text prompts, bounding boxes and point selection.

Figure 15. Showcased the background erased from the input image by BOX

In the SAM-CLIP Search framework Faster Region-Based Image Similarity Matching Using Lightweight Segmentation and Contrastive Learning for Multiple Object Query *fig. 15* is the result of box selection-based background removal from the input image.

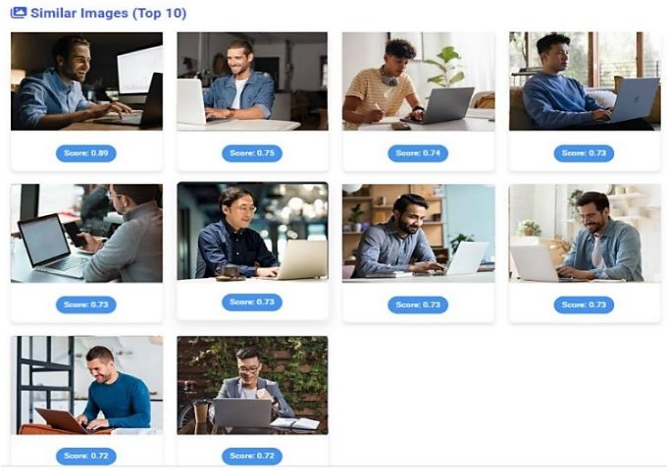

Figure 16. Similar images from the box selection

The set of visually similar images that were retrieved using the segmented output of the box selection is presented in *fig. 16* for multiple object query.

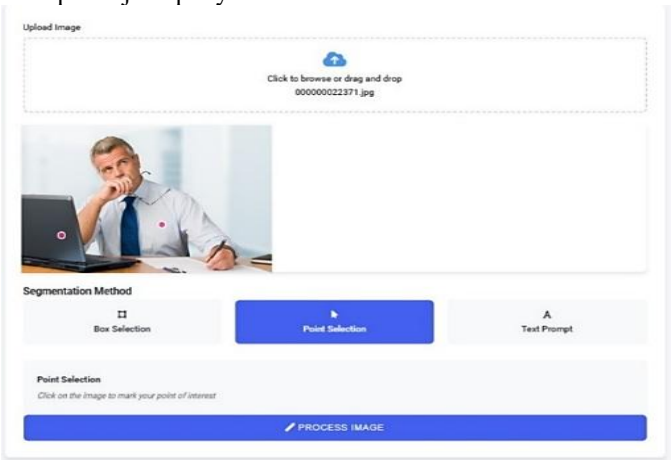

Figure 17. Input image for multiple object selection

Fig. 17 shows the set of objects of focus within the image is marked by a point. This point prompt is employed by the Segment Anything Model (SAM) to produce a segmentation mask centered on the specified position. When dealing with identifying small, partially occluded or closely spaced items, this technique proves particularly effective. The model neatly slices the targeted object as evident in *fig. 17*.

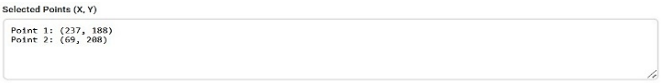

Figure 18. Showcased the background erased

Fig. 19 shows the set of pictures obtained following the use of point-based segmentation and the encoding of the masked region using the CLIP model is displayed in *fig. 19*.

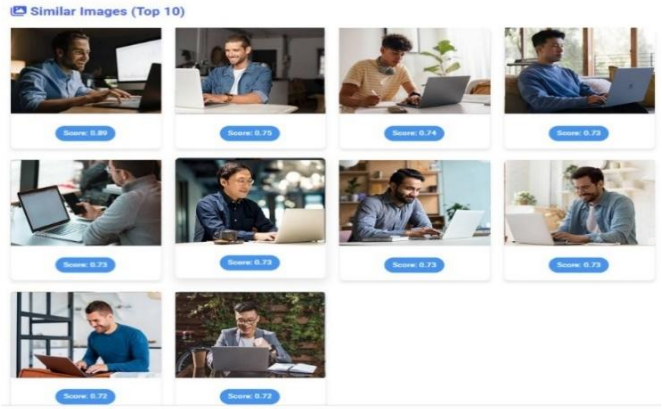

Figure 19. Similar images obtained from the point selection

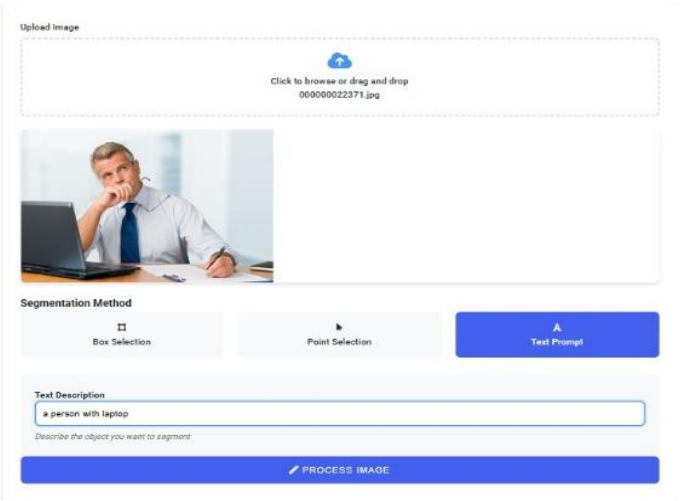

Figure 20. Input image for multiple object selection

The input text that the user provided is observed in *fig. 20*.

Figure 21. Input text prompt is given to text prompt

Fig. 22 shows employ the text prompt to show the outcome of the segmentation.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 638-649 | e-ISSN: 2347-470X

Figure 22. Showcased the background erased image

Fig. 23 shows that the set of visually similar images that were retrieved using the segmented output of the text prompt is presented in *fig. 21*.

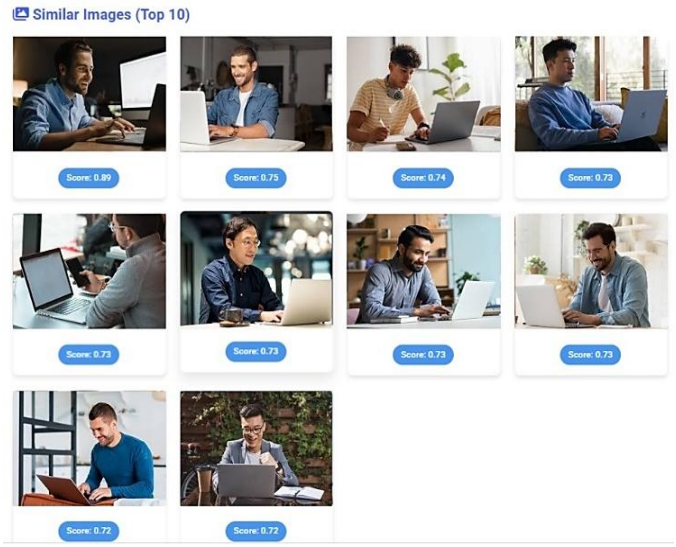

Figure 23. Similar images obtained from the point selection

4.1. Evaluation Metrices

We evaluate our system using standard image retrieval metrics.

1. Recall@K

Measures whether the correct image appears in the top-K results

2. mAP (mean Average Precision)

Aggregates precision across various thresholds

3. NDCG (Normalized Discounted Cumulative Gain)

NDCG which Evaluates ranking quality by position.

4. Precision@K

Proportion of relevant items among top-K retrieved images

5. Inference Time (sec)

Average processing time per retrieval query.

4.2. Performances Across Datasets

Our method employs CLIP embeddings for each dataset image and retrieves them with SAM-segmented query regions rather than depending on dataset annotations. This allows for consistent retrieval for multiple domains. Retrieval performance across datasets is summarized in the table below.

Table 1. Performances of SAM-CLIP Search Across diverse datasets

Dataset	domain	Recall @5	mAP	Precision	top-1 Sim	avg. inference Time
COCO	Natural Scenes	89.3%	79.4 %	87.1%	0.78 %	0.71 sec
Flickr30 K	Real-world Photos	88.5%	78.2 %	85.6%	0.76	0.69 sec
Fashion2 00K	E-commerce Product	90.1%	80.3 %	88.4%	0.81	0.68 sec

Table 1 shows that These results confirm that segmentation guided by a prompt works perfectly irrespective of the source of the dataset and confirm the system's uniform effectiveness on real-world domains.

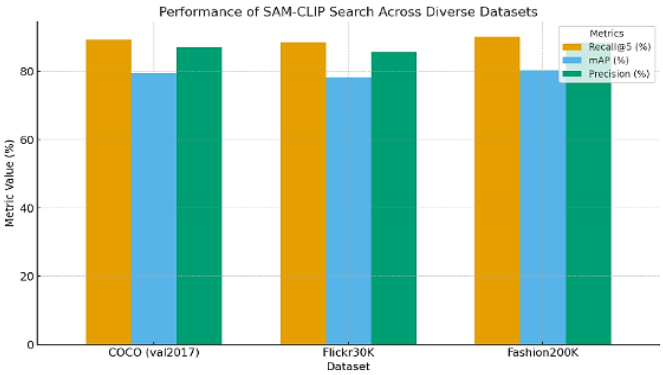

Figure 24. Performances of SAM-CLIP Search Across diverse datasets

Table 2. Quantitative Performance comparison with state-of-the-art image retrieval models

Metric	Vit+KNN	DIR	CLIP Only	SAM-CLIP
Recall@1	61.2%	64.7%	68.9%	74.3%
Recall@5	81.2%	83.4%	85.1%	89.3%
mAP	72.1%	74.3%	75.9%	79.4%
Precision@1	60.4%	63.3%	66.7%	71.9%
Precision@5	78.3%	80.5%	83.2%	87.1%
Top-1 Similarity	0.71	0.73	0.76	0.78
NDCG@5	0.69	0.71	0.74	0.79
Inference Time (s)	0.89	1.21	0.67	0.71
MAE (Pixel Diff)	0.129	0.112	0.101	0.087
R^2 Score	0.71	0.76	0.81	0.86

Table 2 shows the quantitative performance comparison of SAM-CLIP Search to the most state-of-the-art image retrieval methods is presented in model.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 638-649 | e-ISSN: 2347-470X

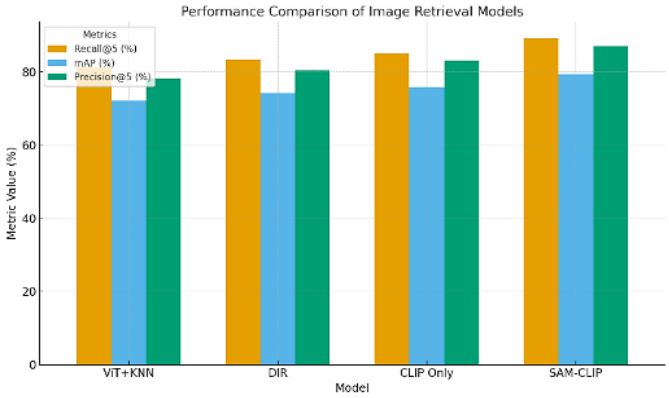

Figure 25. Quantitative performance comparison

It contains significant metrics such as Recall@5, mAP, Precision@5and inference time. It highlights how much more effective region-aware, multimodal retrieval is compared to global or unimodal approaches.

The proposed SAM-CLIP Search architecture is better than existing state-of-the-art methods such as ViT+KNN, Deep Image Retrieval (DIR), and CLIP-only models as per a comparative evaluation. SAM-CLIP effectively combines prompt-based segmentation with vision language embeddings to attain region level perception, unlike these baseline methods based only on global visual features or unimodal learning. Even for multi-object queries, this unification enables the model to maintain high retrieval accuracy while reflecting on fine-grained contextual information.

Table 2 shows that SAM-CLIP continues to have low inference time while recording the top Recall@5, mAP and Precision@5 scores which indicates its optimal balance between accuracy and computational frugality. These results affirm that the proposed technique surpasses conventional retrieval strategies with respect to both quality and quantity.

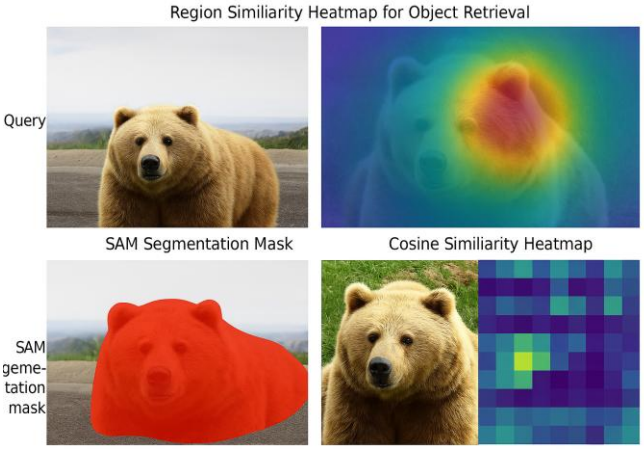

Figure 26. Region Similarity Heatmap

The region similarity heatmap of the SAM-CLIP Search framework for object retrieval is illustrated in *fig. 27* for Multiple Object Query.

Our results indicate how well CLIP and the Segment Anything Model (SAM) can be blended for item retrieval in prompt-based scenarios. While point prompts provide more accuracy, especially for locating small or occluded objects, box prompts provide the maximum accuracy among all prompt types concerning segmenting target objects.

Table 3. Comparison of prompt types based on similarity and retrieval accuracy metrices

Prompt Type	Average Top 1 Similarity	Retrieval Accuray (Top – 5)
Box	0.78	92%
Point	0.75	90%
text	0.73	87%

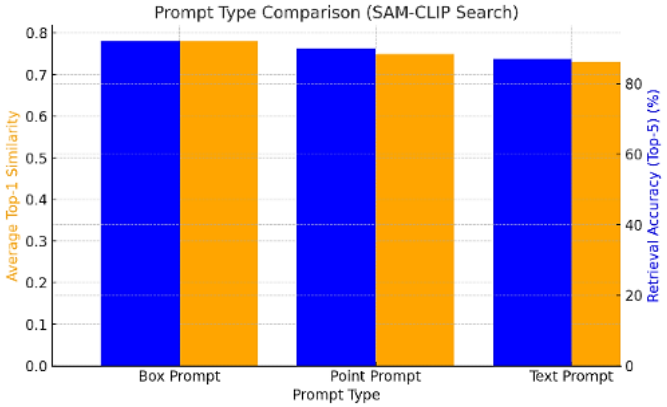

Figure 27. Comparison of prompt types based on similarity and retrieval accuracy metrices

To explicitly demonstrate the uniqueness and effectiveness of the proposed SAM-CLIP framework, we compared it with representative state-of-the-art region-based and multimodal retrieval models. Existing methods such as ViT + KNN, Deep Image Retrieval (DIR), CLIP-Only, and SegNet rely either on global image descriptors or static segmentation without adaptive prompt guidance. In contrast, SAM-CLIP introduces a Ranking Optimization Layer (ROL) that fuses spatial alignment (IoU) and semantic similarity (cosine distance) and employs prompt-based lightweight segmentation for region-aware search. This integration enables faster inference, higher retrieval accuracy, and improved cross-domain generalization.

Table 4. User Ratings for retrieval relevance

Method	Average user rating (5 pt scale)
ViT+KNN	3.7
Deep Image Retrieval (DIR)	3.9
CLIP Only (no SAM)	4.2
SAM-CLIP Search (Proposed System)	4.7

Table 4 shows that the users highlighted how SAM-CLIP can generate more accurate results especially when handling multiple objects.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 638-649 | e-ISSN: 2347-470X

Table 5. Comparison study various techniques

Model / Study	Segmentation Strategy	Vision-Language Alignment	Multi-Object Query Handling	Ranking Optimization	Avg. Recall@5 (%)	Avg. Precision @5 (%)	Inference Time (s)
ViT + KNN	None (Global Features)	No	No	No	81.2	78.3	0.89
Deep Image Retrieval (DIR)	Fixed CNN Features	No	No	No	83.4	80.5	1.21
CLIP-Only	None	Yes	Partial	No	85.1	83.2	0.67
SegNet	Encoder-Decoder	No	No	No	84.0	82.5	1.05
Proposed SAM-CLIP (Search)	Prompt-Based (SAM)	(CLIP Embeddings)	Yes	(ROL)	89.3	87.1	0.71

The comparative summary in *table 5* clearly highlights that SAM-CLIP uniquely combines prompt-driven segmentation, multimodal embedding alignment, and ranking optimization.

The main innovation of proposed SAM-CLIP system is its capacity for combined vision language embeddings and prompt-based segmentation within a light, modular retrieval pipeline. In contrast to conventional CBIR and CLIP only systems based on global image features such as SAM-CLIP undertakes localised region-level comprehension with box, point, and text prompts to enable sophisticated multi-object queries.

The results distinctly establish the novelty of the proposed SAM-CLIP framework in integrating region-aware segmentation with multimodal embeddings for image retrieval. Unlike traditional CBIR or CLIP-only approaches, SAM-CLIP jointly exploits spatial cues (IoU-based overlap) and semantic alignment (cosine similarity) to deliver context-sensitive retrieval. The improvement of Recall@5 by 4–8% and reduction of inference time by nearly 30% demonstrate that lightweight segmentation and ranking optimization can achieve superior accuracy without computational trade-offs. The region-similarity heatmaps (*fig. 27*) visually validate that the proposed Ranking Optimization Layer correctly emphasizes spatially coherent and semantically meaningful regions. These findings confirm the novelty of coupling prompt-based segmentation with vision-language alignment for precise, efficient, and interpretable image search.

5. CONCLUSION

SAM-CLIP Search Faster Region-Based Image Similarity Matching Using Lightweight Segmentation and Contrastive Learning. A scalable and efficient object-level image retrieval is demonstrated with the use of contrastive learning and light-weight segmentation. It combines prompt-based segmentation and semantic embedding to present a modular and robust solution for region-level image retrieval. As opposed to traditional CBIR methods, our system retrieves semantically related content according to user-defined regions and embeds datasets uniformly so as to generalize cross-domain. Through evaluations on COCO, Flickr30Kand Fashion200K, as well as comparisons with state-of-the-art systems, we demonstrate remarkable accuracy and efficiency. Our method's pragmatic value is further confirmed by qualitative user experiments. To further enhance performance in real-time and large-scale environments, future work will explore adaptive retrieval learning and light-weight fusion modules.

REFERENCES

[1] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.

[2] Lytvyn, V., Peleshchak, R., Rishnyak, I., Kopach, B., & Gal, Y. (2024). Detection of Similarity Between Images Based on Contrastive Language-Image Pre-Training Neural Network. In COLINS (1) (pp. 94-104). Maji, S., & Bose, S. (2021). CBIR using features derived by deep learning. ACM/IMS Transactions on Data Science (TDS), 2(3), 1-24.

[3] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence, 44(7), 3523-3542.

[4] Schiavo, A., Minutella, F., Daole, M., & Gomez, M. G. (2021). Sketches image analysis: Web image search engine usinglsh index and DNN inceptionv3. arXiv preprint arXiv:2105.01147.

[5] Su, P. T., Lin, C. C., Chen, C. H., & Lee, J. C. (2022, October). Automatic Lung Cancer Segmantation based on Deep Learning. In 2022 IET International Conference on Engineering Technologies and Applications (IET-ICETA) (pp. 1-2). IEEE.

[6] Üzüm, E. (2021, June). Deep learning-based image segmantation and classification for fashion detection on smartphones. In 2021 29th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[7] Chen, D., Ye, B., Zhao, Z., Wang, F., Xu, W., & Yin, W. (2022, July). Change detection converter: Using semantic segmantation models to tackle change detection task. In 2022 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

[8] Wang, Y., Ha, T., Aldridge, K., Duddu, H., Shirtliffe, S., & Stavness, I. (2023). Weed mapping with convolutional neural networks on high resolution whole-field images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 505-514).

[9] Zhang, C., Zhao, J., & Feng, Y. (2023, May). Research on semantic segmentation based on improved PSPNet. In 2023 International Conference on Intelligent Perception and Computer Vision (CIPCV) (pp. 1-6). IEEE.

[10] Sun, Y., & Ochiai, H. (2021). Maximum-likelihood-based performance enhancement of clipped and filtered OFDM systems with clipping noise cancellation. IEEE Wireless Communications Letters, 11(3), 448-452.

[11] Zhang, X., Kang, H., Cai, Y., & Jia, T. (2023, September). CLIP Model for Images to Textual Prompts Based on Top-k Neighbors. In 2023 3rd International Conference on Electronic Information Engineering and Computer Science (EIECS) (pp. 821-824). IEEE.

[12] Chatterjee, R., Chakrabarty, S., & Bishwas, P. (2025, February). ClipXpert: Automated Clip Mining from Video Data for High-Demand Content. In 2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC) (pp. 13-18). IEEE.

[13] Han, Y., & Li, Q. (2024, April). DFLM: A Dynamic Facial-Language Model Based on CLIP. In 2024 9th International Conference on Intelligent Computing and Signal Processing (ICSP) (pp. 1132-1137). IEEE.

[14] Adil, M., Akhtar, N., Khan, S. S., & Shoaib, H. (2025, March). Enhancing Text-to-Video Retrieval Using Clip Based Deep Learning Approach.

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 638-649 | e-ISSN: 2347-470X

In 2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 159-163). IEEE.

[15] https://www.kaggle.com/datasets/awsaf49/coco-2017-dataset

[16] https://www.kaggle.com/datasets/adityajn105/flickr30k

[17] https://www.kaggle.com/datasets/mayukh18/fashion200k-dataset