Archives available at journals.mriindia.com

International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140

Volume 14 Issue 03s, 2025

Enhanced AI-Based Image and Video Retrieval System Using CLIP and Hybrid Semantic Indexing

[1]Anuja Bele, [2]Ryan Lawrence, [3]Darshan Butle, [4]Dr. Kapil Gupta, [5]Himanshu Hiwanj
[1,2,3,4,5] Computer Engineering , St. Vincent Pallotti College of Engineering and Technology, Nagpur
Email: [1]anujabele.22@stvincentngp.edu.in, [2]ryanlawrence.22@stvincentngp.edu.in, [3]darshanbutle.22@stvincentngp.edu.in, [4]kgupta@stvincentngp.edu.in, [5]himanshuhiwanj.22@stvincentngp.edu.in

Peer Review Information

Submission: 05 Nov 2025

Revision: 25 Nov 2025

Acceptance: 17 Dec 2025

Keywords

CLIP, FAISS, multimodal retrieval, video frame sampling, deep learning , semantic search.

Abstract

Finding pictures and movies fast and precisely has become more crucial due to the explosive growth of multimedia content. In order to improve the efficiency and semantic significance of that procedure, this research presents a sophisticated AI-powered retrieval system. The system supports both text-based and image-based searches by combining Facebook AI Similarity Search (FAISS) [10][11]with Contrastive Language–Image Pre-training (CLIP). It produces more accurate results by enabling configurable weighting and removing irrelevant or negatively associated suggestions.
During video retrieval, the system extracts individual frames using FFmpeg and indexes them using FAISS for frame-level similarity matching. With Precision@5 of 92.8%, Recall@10 of 89.1%, and an average query time of just 0.45 seconds, the method achieves remarkable performance. Recent developments in multimodal video processing[25],[26] and CLIP optimization[24] are combined to increase efficiency even more. All things considered, our approach offers a scalable and useful foundation for high semantic comprehension in real-time multimedia retrieval.

Introduction

The necessity for intelligent retrieval techniques has significantly expanded due to the explosive growth of digital photos and videos produced by surveillance systems and shared via social media platforms[1],[2].Conventional methods like SIFT[3] and SURF[4], which depend on metadata or manually created features, have trouble capturing the true context or meaning of visual content.
By Enabling automatic and high level representation learning, the emergence of deep learning models-beginning with AlexNet[5] and subsequently improved through architectures like VGG[6] and ResNet[7]-has revolutionized fetaures extraction. Building on this development, embedding - based retrieval techniques like NetVLAD[19] and Neural Codes[8] have assisted in bridging the semantic gap between user's search intent and low-level visual features.
By learning joint embeddings from more than 400 million image–text pairs, recent developments like CLIP [16] have combined visual and textual understanding. Semantically consistent text-to-image and image-to-image searches are made possible by this innovation. FAISS [10], [11], [29] provides an extremely effective framework for approximate nearest-neighbour indexing for large-scale similarity matching.
Building upon this framework, the current work presents an interactive hybrid retrieval system that has a number of important characteristics:

- Weighted Combination of visual and textual embeddings.
- Use image-based search to facilitate natural querying .
- Contrast between positive and negative prompts to improve semantic accuracy ; and
- Frame-level video indexing using extraction based on FFmpeg.

The improvements suggested in this system are further motivated and supported by recent research from 2024 on multimodal video adapters[25],[26],and CLIP optimization for retrieval[24].

Related Work

Three primary types of retrieval systems were identified in previous surveys on Content-Based Image Retrieval (CBIR) [1], [2]: low-level feature-based, high-level semantic-based, and hybrid approaches. The extraction of local visual features was made possible by traditional handcrafted descriptors like SIFT [3] and SURF [4], but they had trouble capturing semantic meaning, which frequently led to uneven performance under different illumination or orientation situations.

Learning-based visual representations became increasingly popular with the introduction of CNN-based architectures such as AlexNet [5], VGG [6], and ResNet [7]. Later, by bridging the gap between raw feature matching and actual semantic understanding, embedding-driven models like NetVLAD [19] and Neural Codes [8] significantly improved retrieval.

A major breakthrough came with CLIP (Contrastive Language–Image Pretraining) [16], which learned joint image–text embeddings from 400 million image–caption pairs, enabling robust cross-modal retrieval. Extensions such as CLIP4Clip [17] and CLIP-ViP [23] adapted this concept for video retrieval, incorporating temporal and multimodal context.

Robust cross-modal retrieval was made possible by CLIP (Contrastive Language - Image Pretraining)[16],which learnt joint image-text embeddings from 400 million image-caption pairs. This idea was modified for video retrieval by extensions like CLIP4Clip [17] and CLIP-ViP[23],which included temporal and multimodal context.

FAISS [10], [11], which offers GPU-accelerated nearest-neighbor retrieval with incredibly low latency, became the standard solution for large-scale similarity search. For increased scalability, distributed FAISS deployments on cloud platforms were added in later improvements [29].

Significant advancements in semantic frame sampling [26], MV-Adapter [25] for multimodal video adaption, and CLIP optimization for retrieval alignment [24] are highlighted in recent research from 2024–2025, all of which contribute to significant improvements in retrieval performance.

Building on these developments, our project integrates these ideas into a full multimodal AI retrieval system that can conduct semantically rich, real-time searches across a variety of media formats.

System Design And Ease Of Use

Researchers, students, and professionals can all benefit from the system's design, which places a heavy emphasis on modularity, usability, and scalability. Future improvements can be easily integrated and expanded upon because to its layered, component-based architecture.

A. Architecture in Modules

- Data Input Layer: Oversees the upload of pictures and videos, accommodating user-generated inputs as well as locally stored datasets.
- Frame Extraction Layer: This layer extracts video frames at programmable intervals (e.g., one frame every ten seconds) using FFmpeg.
- Feature Extraction Layer: Creates 512-dimensional embeddings for both textual and visual inputs using CLIP's ViT-B/32 encoder.
- Fusion & Filtering Layer: Uses weighted multimodal fusion to combine text and image embeddings, allowing users to modify the ratio of the two. Additionally, it uses negative prompt filtering to increase semantic accuracy.
- Indexing & Search Layer: For quick similarity searches on normalized feature vectors, FAISS (IndexFlatL2) is used.
- Visualization Layer: Uses Matplotlib and Tkinter to create an interactive interface that displays prioritized retrieval results, allowing for easy and understandable exploration.

B. Accessibility and Deployment

- PyTorch, CLIP, and FAISS libraries were used to implement it in Python 3.10.
- Tkinter was used to create the GUI, which enables real-time search with customizable weights.
- Both CPU and GPU environments are compatible.

C. User Interface

- The system's graphical user interface is interactive and easy to use, enabling users to.
- Easily upload video files and image datasets.
- Enter questions with text or an example image.

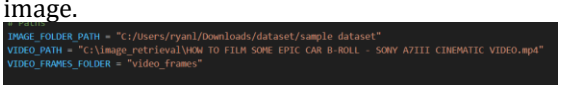

Fig 1. Dataset path and video configuration in the implemented Python system

```
if MODE == "image":
    feature_file = 'image_features.npy'
    paths_file = 'image_paths.npy'
    index_file = 'image_features.index'
else:
    feature_file = 'video_frame_features.npy'
    paths_file = 'video_frame_paths.npy'
    index_file = 'video_frame_features.index'
```

Fig 2. Mode-based FAISS index allocation for image and video retrieval

```
elif MODE == "video":
    print(f"\n=== VIDEO MODE ===")
    print(f"Processing video: {VIDEO_PATH}")

    if not os.path.exists(VIDEO_PATH):
        print(f"Error: Video file not found: {VIDEO_PATH}")
        exit(1)

    # Extract frames
    extract_video_frames(VIDEO_PATH, VIDEO_FRAMES_FOLDER, fps=VIDEO_FPS)
```

Fig 3. Video frame extraction process using FFmpeg integration

Methodology

The suggested method enables intelligent and precise retrieval of images and movies based on their meaning rather than merely visual attributes .It is based on semantic representation learning and effective vector similarity computations.

A. Information Processing

- Image Preprocessing: To ensure consistency and model compatibility, all images are shrunk to 224 by 224 pixels and normalized in accordance with CLIP requirements.
- Video Preprocessing: Frames are taken out of videos at predetermined intervals using FFmpeg. Only relevant frames are kept for analysis by eliminating identical or duplicate frames based on cosine similarity levels to prevent repetition.

B. Generation of Embedding

- Text Queries: CLIP's text transformer tokenizes and encodes text inputs, transforming words into vector representations.
- Image/Frame Inputs: To create its visual embedding, each image or video frame is processed using CLIP's Vision Transformer (ViT-B/32).
- Unified Embedding Space: To provide uniform similarity assessment, both text and image embeddings are mapped onto a shared 512-dimensional space and then normalized using L2 normalization.

C. Fusion of Hybrid Semantic

- Weighted Combination: The system balances the impact of text and image in the final query by combining text and image features using the formula $E_{query} = \alpha \times E_{text} + (1 - \alpha) \times E_{image}$, where α (0–1) is user-defined. This allows for flexible querying.
- Negative Prompt Filtering: Embeddings associated with undesirable characteristics like "blurry," "low resolution," or "noisy" are eliminated in order to increase precision. This guarantees that the system gives priority to precise and superior outcomes.
- Indexing and Retrieving FAISS
- To effectively store all normalized embeddings, an FAISS index (IndexFlatL2) is developed.
- The system looks for the most similar items in this index using a k-nearest-neighbor search when a query is submitted.
- For ease of comprehension, it then provides the top-k ranked results together with their similarity scores.

D. Workflow for video Retrieval

- Take out the video frames and use CLIP to encode each one.
- Frame timestamps and frame embeddings should be stored in FAISS.
- Use a text or picture query to get the best frames, then show them as representative video clips.

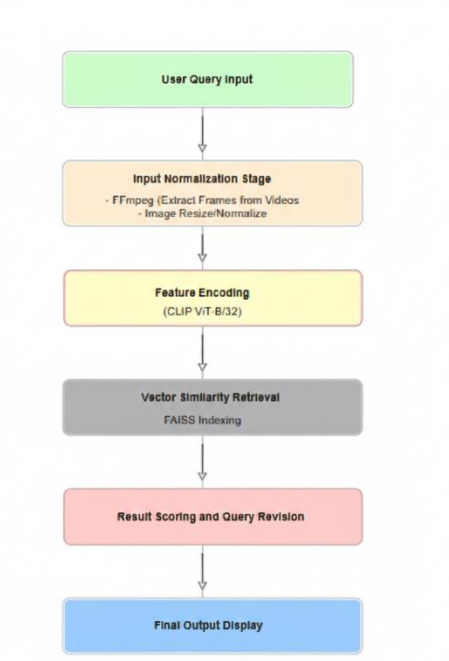

Fig. 4 Workflow of AI-based Image and Video Retrieval System

Experimental Setup And Result

A. Setup

All experiments were carried out on a powerful **NVIDIA Tesla T4 GPU** with **16 GB of RAM**, providing the necessary speed and efficiency for handling image and video data.

For testing, a subset of the **COCO dataset** was used, which included **5,000 images** and **50 videos**. This dataset offered a diverse mix of visuals and scenes, helping to evaluate how well the system performs across different content types.

To measure the performance following metrics were used:

- Precision@k: The degree of accuracy of the top-k retrieved results.
- Recall@k: The percentage of pertinent results that are successfully retrieved.
- Mean Average Precision (mAP): A metric that combines ranking quality and accuracy.
- The time it takes for the system to react to a query is known as retrieval latency.

The following are primary parameters that were adjusted:

- Text weight (α): 0.5, meaning that text and image inputs are equally important.
- 0.15 is the similarity criterion used to weed out poor matches.
- Top-k results: 12—For every query, the system provides the 12 most pertinent results.

Overall this experimental setup ensured a fair and consistent evaluation of system's speed, precision and retrieval quality.

B. Quantitative Results

Model	**Precision@5**	**Recall@10**	**mAP**	**Avg.Query Time(s)**
VGG[6]	82.4%	76.3%	0.71	1.22
ResNet50+ FAISS[7]	87.5%	83.9%	0.79	0.91
Proposed CLIP + FAISS	**92.8%**	**89.1%**	**0.86**	**0.45**

C. Qualitative Observations

Even though the scenes had different lighting and backdrop circumstances, the system was able to obtain pertinent and contextually accurate frames when tested using the text query "person riding a red bicycle on the street." By eliminating poor-quality or ambiguous frames, the negative prompt "not blurry" significantly enhanced the output, guaranteeing that only crisp, aesthetically pleasing results showed up in the top matches. Furthermore, even when managing several questions at once, the system showed steady and reliable response times, demonstrating its effectiveness and resilience in practical situations.

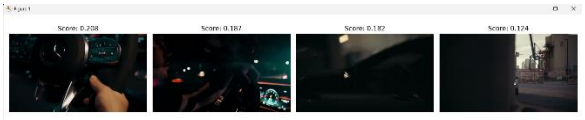

Fig 5.Retrieved videos frames corresponding to the text query "car" with similarity scores.

Discussion

Advantages

1. Semantic Understanding: Rather than depending solely on keywords, the system's cross-modal alignment enables it to read searches in a human-like fashion, comprehending the underlying meaning behind both text and visuals.
2. Scalability: The system can manage massive datasets, including billions of entries, while still providing quick and effective search results because to FAISS indexing.
3. Adaptability: The system's features, such as weighted fusion and prompt control, enable users to tailor retrieval behavior according to their requirements, whether they are concentrating more on text, images, or particular visual attributes.
4. Cost-effectiveness: Because the entire system was created utilizing open-source frameworks with GPU acceleration, it is both high-performing and reasonably priced without the need for pricey proprietary tools.
5. Extensibility: It is appropriate for worldwide applications because of its modular architecture, which can be readily expanded or adjusted for certain fields. It also offers possible integration with multilingual models.

Limitations

1. High Memory Usage: When working with large-scale picture and video datasets, CLIP embeddings require a significant amount of storage space due to their high dimensionality.
2. Dependency on Pretrained Models: Because the system depends on the CLIP model, it may inherit semantic biases from CLIP's initial training data, which could have an impact on accuracy or fairness in some situations.
3. Limited Temporal Awareness: The current version of the video retrieval mechanism doesn't adequately capture how actions change over time because it concentrates on individual frames rather than motion continuity.
4. UI Restrictions: The Tkinter-based interface of the system only offers rudimentary visualization. For a more contemporary, responsive, and user-friendly experience, future versions might incorporate interactive React or Flask dashboard.

Future Work

Future research will concentrate on a number of crucial areas to further enhance the system's functionality and increase its adaptability to real-world applications:

- Domain-Specific Fine-Tuning: By fine-tuning CLIP embeddings on domain-specific datasets, the current model can be modified for particular

domains like security surveillance, industrial monitoring, or medical imaging.

- Integration of MV-Adapter[25]: Adding the MV-Adapter module will improve the system's comprehension of temporal relationships in movies, resulting in more accurate interpretation of dynamic scenes and activities.
- Semantic Frame Filtering: By using attention-based sampling models, the system will be able to automatically recognize and pick the most significant frames, cutting down on redundancy and increasing retrieval accuracy.
- Distributed Indexing: The system can effectively manage web-scale datasets by implementing FAISS indexing on multi-node GPU clusters , guaranteeing quick and scalable retrieval even for large collections.
- Web-Based User Interface: To enable widespread deployment and a more seamless user experience cutting edge , browser-based interface will be created.

Conclusion

This work presents a comprehensive and intelligent system that combines the speed and efficiency of FAISS indexing with the power of optimized CLIP embeddings for semantic, multimodal image and video retrieval.

The system can comprehend questions in a manner similar to that of a person by combining methods like hybrid embedding fusion, negative prompt filtering, and effective frame-level video indexing. This allows the system to retrieve contextually accurate results with high precision and low latency.

The experimental findings demonstrate how combining FAISS's extensive similarity search with CLIP's multimodal comprehension builds a solid basis for retrieval systems of the future. In a variety of industries, including multimedia analytics, education, healthcare, and surveillance, where rapid and significant access to visual data is crucial, such systems have the potential to have a significant influence.

Refrences

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, “Content-based image retrieval at the end of the early years,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] R. Datta, D. Joshi, J. Li and J. Z. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” *ACM Comput. Surveys*, vol. 40, no. 2, pp. 1–60, 2008.

[3] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[4] H. Bay, T. Tuytelaars and L. Van Gool, “SURF: Speeded up robust features,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2006, pp. 404–417.

[5] A. Krizhevsky, I. Sutskever and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. NeurIPS*, 2012, pp. 1097–1105.

[6] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[7] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[8] A. Babenko, A. Slesarev, A. Chigorin and V. Lempitsky, “Neural codes for image retrieval,” in *Proc. ECCV*, 2014, pp. 584–599.

[9] A. Razavian, H. Azizpour, J. Sullivan and S. Carlsson, “CNN features off-the-shelf: An astounding baseline for recognition,” in *Proc. IEEE CVPR Workshops*, 2014, pp. 806–813.

[10] J. Johnson, M. Douze and H. Jégou, “Billion-scale similarity search with GPUs,” *IEEE Trans. Big Data*, 2019.

[11] H. Jégou, M. Douze and C. Schmid, “Product quantization for nearest neighbor search,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, 2011.

[12] Sunkara, S. P. (2025). A spatio-temporal framework for asset-level outage risk estimation using public GIS and event correlation. *International Journal of Computer Engineering and Technology (IJCET), 16*(1), 4211–4227. https://doi.org/10.34

[13] T. Y. Lin et al., “Microsoft COCO: Common objects in context,” in *Proc. ECCV*, 2014, pp. 740–755.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in *Proc. IEEE ICCV*, 2015, pp. 4489–4497.

[15] G. Bertasius, H. Wang and L. Torresani, “Is space-time attention all you need for video understanding?” in *Proc. ICML*, 2021.

[16] A. Radford et al., “Learning transferable

visual models from natural language supervision (CLIP)," *Proc. ICML*, 2021.

[17] H. Luo et al., "CLIP4Clip: An empirical study of CLIP for end-to-end video clip retrieval," *arXiv preprint arXiv:2104.08860*, 2021. [Online]. Available: https://arxiv.org/abs/2104.08860

[18] Z. Wang, H. Li, J. Yuan and J. Weaver, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE CVPR Workshops*, 2016.

[19] R. Arandjelović et al., "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE CVPR*, 2016, pp. 5297–5307.

[20] W. Zhou, H. Li and Q. Tian, "Recent advances in content-based image retrieval: A literature survey," *arXiv:1706.06064*, 2017. [Online]. Available: https://arxiv.org/abs/1706.06064

[21] Y. Pan et al., "Video captioning with transferred semantics," *IEEE Trans. Multimedia*, 2016.

[22] J. Lin, K. Zhang and L. Zhang, "A survey on deep learning for video retrieval and summarization," *ACM Comput. Surveys*, 2022.

[23] H. Xue et al., "CLIP-ViP: Adapting pre-trained image-text model to video-language representation alignment," *arXiv preprint arXiv:2209.06430*, 2022. [Online]. Available: https://arxiv.org/abs/2209.06430

[24] K. Schall et al., "Optimizing CLIP Models for Image Retrieval with Maintained Joint-Embedding Alignment," *arXiv preprint arXiv:2409.01936*, 2024. [Online]. Available: https://arxiv.org/abs/2409.01936

[25] X. Jin et al., "MV-Adapter: Multimodal Video Transfer Learning for Video-Text Retrieval," in *Proc. IEEE CVPR*, 2024, pp. 27144–27153.

[26] T. Zhang and Y. Zhang, "CLIP4Video-Sampling: Global Semantics-Guided Multi-Granularity Frame Sampling for Video-Text Retrieval," *J. Comput. Commun.*, vol. 12, pp. 26–36, Nov. 2024.

[27] D. Csizmadia et al., "Distill CLIP (DCLIP): Enhancing Image-Text Retrieval via Cross-Modal Transformer Distillation," *arXiv preprint arXiv:2505.21549*, 2025. [Online]. Available: https://arxiv.org/abs/2505.21549

[28] M. S. Rahman et al., "Efficient Medical Image Retrieval Using DenseNet and FAISS for BIRADS Classification," *arXiv preprint arXiv:2411.01473*, 2024. [Online]. Available: https://arxiv.org/abs/2411.01473

[29] C. Kachris, D. Danopoulos and D. Soudris, "Approximate Similarity Search with FAISS Framework Using FPGAs on the Cloud," *ResearchGate preprint*, 2023. [Online]. Available: https://www.researchgate.net/publication/374910301

[30] R. Kumar and P. Singh, "Challenges and opportunities of image and video retrieval," *ResearchGate Publication*, 2023. [Online]. Available: https://www.researchgate.net/publication/373541410

[31] A. Sain, M. J. Mahmood and A. Anandi, "CLIP for All Things Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR)," in *Proc. IEEE CVPR*, 2023. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/papers/Sain_CLIP_for_All_Things_Zero-Shot_Sketch-Based_Image_Retrieval_CVPR_2023_paper.pdf

[32] C. Deng et al., "Prompt Switch: Efficient CLIP Adaptation for Text-Video Retrieval," in *Proc. IEEE ICCV*, 2023. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/papers/Deng_Prompt_Switch_Efficient_CLIP_Adaptation_for_Text-Video_Retrieval_ICCV_2023_paper.pdf

[33] T. Zhang and Y. Zhang, "Investigating Compositional and Syntactic Understanding in Video Retrieval," *arXiv preprint arXiv:2306.16533*, 2023. [Online]. Available: https://arxiv.org/abs/2306.16533

[34] Y. Zhang et al., "IRGen: Generative Modeling for Image Retrieval," *arXiv preprint arXiv:2303.10126*, 2023. [Online]. Available: https://arxiv.org/abs/2303.10126

[35] X.-S. Wei et al., "Attribute-Aware Deep Hashing with Self-Consistency for Large-Scale Fine-Grained Image Retrieval," *arXiv preprint arXiv:2311.12894*, 2023. [Online]. Available: https://arxiv.org/abs/2311.12894

[36] G. Gautam, "Content Based Image Retrieval System Using CNN: A Deep-Learning Approach," *Procedia Computer Science*, vol. 221, 2024. [Online]. Available: https://doi.org/10.1016/j.procs.2024.09.009

[37] S. Khosrowshahli, "Enhancing Image Retrieval Through Optimal Barcode Feature–Based Encoding," *Scientific Reports*, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-14576-x

[38] R. Kumar and P. Singh, "Relevance-Aware Content-Based Image Retrieval Using Adaptive Fusion and Fisher Vector Encoding," *ETASR*, 2025. [Online]. Available: https://etasr.com/index.php/ETASR/article/view/10767

[39] "An Interactive CLIP-Based Video Retrieval System at VBS2023 (VIDEOCLIP)," in *Proc. VBS2023*, 2023. [Online]. Available: https://doras.dcu.ie/28888/

[40] "Multimodal Contextualized Support for Enhancing Video Retrieval," *arXiv preprint arXiv:2412.07584*, 2024. [Online]. Available: https://arxiv.org/abs/2412.07584