

# MediaVault: An Intelligent Multimedia Retrieval System Using SigLIP Embeddings and FAISS Indexing

V. Shravan<sup>1</sup>, S. Sameer Thapaswi<sup>2</sup>, A. Saideepak<sup>3</sup>, H. Jayasree<sup>4</sup>

<sup>1,2,3</sup> Former under graduate student, Department of Information Technology Maturi Venkata Subba Rao (MVSR) Engineering College, Nadargul, Hyderabad, India

<sup>4</sup>Professor, Department of Computer Science and Engineering Jyothismathi Institute of Technology and Science, Karimnagar, India.

**Abstract**—The exponential growth of personal and organizational digital media has created a pressing need for intelligent systems capable of managing and retrieving vast multimedia content efficiently. Traditional keyword-based or metadata-driven search methods often fail in scenarios with unstructured or poorly annotated data, lacking the semantic understanding necessary for effective retrieval. To address these limitations, this paper introduces MediaVault, an intelligent multimedia retrieval system that bridges the semantic gap between user intent and visual content. MediaVault integrates SigLIP, a state-of-the-art vision-language transformer model, to generate joint visual-textual embeddings, and employs FAISS for high-speed vector similarity search. This architecture enables context-aware, natural language queries while ensuring privacy through secure, user-specific storage partitions. Experimental evaluations on a curated dataset of diverse images demonstrate that MediaVault significantly improves semantic retrieval accuracy and search efficiency compared to traditional methods. The system's modular and scalable design establishes a robust foundation for future intelligent multimedia applications, including video retrieval, AI-based auto-tagging, and personalized content management.

**Index Terms**—AI-based image search, multimedia indexing, FAISS, CLIP, SigLIP, semantic retrieval, content-based image retrieval.

## I. INTRODUCTION

In the current digital landscape, the rapid proliferation of personal and organizational multimedia content presents significant challenges in organizing, indexing, and retrieving relevant information. Traditional file systems and keyword-based search methodologies often fall short when

managing large-scale datasets, particularly when the data lacks descriptive metadata or structured annotations. These limitations highlight the urgent need for intelligent systems capable of semantically understanding and retrieving multimedia content based on visual and contextual similarities.

To address these challenges, this paper introduces MediaVault, an intelligent multimedia management platform designed to deliver context-aware, semantic retrieval capabilities. MediaVault leverages SigLIP, a transformer-based vision-language model, to encode both visual media and textual queries into a shared semantic embedding space. These embeddings are indexed using the Facebook AI Similarity Search (FAISS) framework, enabling rapid, scalable, and semantically meaningful similarity search operations across large multimedia datasets.

The main contributions of this paper are as follows:

1. Proposing MediaVault as an intelligent system architecture for semantic multimedia retrieval, integrating vision-language embeddings to bridge the gap between user intent and visual content.
2. Implementing FAISS-based indexing to enable high-speed and scalable vector similarity searches, suitable for large-scale multimedia repositories.
3. Demonstrating significant improvements in semantic retrieval accuracy and system performance compared to traditional content-based image retrieval (CBIR) methods, thereby validating MediaVault's effectiveness as an intelligent multimedia retrieval solution.

## II. RELATED WORKS

The field of intelligent multimedia systems has witnessed substantial progress with the integration of content-based retrieval techniques, deep learning models, and advanced feature representation methods. Such developments have significantly enhanced the ability of intelligent systems to understand and retrieve multimedia content based on semantic meaning rather than solely relying on low-level visual features or metadata.

Within this broader domain, Content-Based Image Retrieval (CBIR) has evolved as a critical component of intelligent systems, aiming to bridge the semantic gap between user intent and visual information. Researchers have explored numerous deep learning-based feature extraction techniques to improve retrieval accuracy and efficiency. For instance, Rani et al. [1] demonstrated enhanced retrieval performance by combining features extracted from Big Transfer Networks, emphasizing the potential of deep representations to capture complex semantic information. Similarly, Jain and Gupta [2] highlighted the importance of structural features, suggesting that integrating texture and shape descriptors can significantly improve retrieval relevance in intelligent systems.

Singh and Anand [3] introduced hypergraph-based visual concepts to represent multi-faceted image content, a methodology that informs modular architectures like MediaVault, which can support future extensions such as video segmentation, facial clustering, and object detection. Dubey and Singh [4] showed that combining global and region-based retrieval approaches enhances system performance, reflecting principles adopted in MediaVault's design. Additionally, Agrawal and Anand [5] proposed local motif pattern codes to incorporate fine-grained features, further strengthening semantic retrieval capabilities.

Beyond purely visual methods, hybrid approaches have gained attention. Sadique and Haque [6] proposed a CBIR technique utilizing Color Layout Descriptors (CLD) and Gray-Level Co-Occurrence Matrices (GLCM) for effective color and texture feature extraction. Fouad [7] developed a hybrid retrieval system combining textual annotations with visual features, illustrating how intelligent systems can leverage multi-modal data for improved retrieval

accuracy. Xu [8] reviewed self-supervised learning (SSL) in medical imaging, underscoring the value of SSL for intelligent systems when labeled data is limited—a relevant insight for future enhancements of MediaVault.

Mansourvar et al. [9] examined fusion-based color descriptors, revealing the benefits of integrating multiple feature extraction techniques to enhance retrieval performance in intelligent systems. Collectively, these works establish a foundation upon which MediaVault builds, integrating state-of-the-art vision-language models and scalable similarity search to create an intelligent, context-aware multimedia retrieval platform capable of addressing the challenges inherent in modern multimedia management.

## III. PROPOSED APPROACH

The MediaVault system adopts a modular, AI-driven architecture—depicted in Figure 1—designed to facilitate efficient content-based multimedia retrieval through natural language queries. The system addresses the semantic gap between user-provided textual input and visual media content by mapping both modalities into a shared high-dimensional embedding space. This transformation is accomplished using SigLIP, a state-of-the-art transformer-based vision-language model capable of generating semantically rich embeddings for both images and text. These embeddings are subsequently indexed using the Facebook AI Similarity Search (FAISS) framework, which enables rapid and scalable similarity search operations across large multimedia datasets.

To ensure secure and individualized media storage, each user is assigned a token-authenticated directory that maintains data privacy and isolation. The backend infrastructure is developed using Flask, providing support for image upload, retrieval, and search functionalities. The embedding and indexing processes are encapsulated within a dedicated pipeline that transforms user-uploaded media into a searchable vector database. Upon receiving a search query, the system tokenizes and encodes the input using the same SigLIP model, after which the resulting query vector is compared against the FAISS index to retrieve the top-k most semantically similar images.

In contrast to traditional retrieval approaches that rely primarily on metadata or static labels, MediaVault's semantic search mechanism enables flexible, context-aware media retrieval through descriptive natural language input. This intelligent retrieval capability offers a highly personalized user experience and demonstrates significant applicability across various domains—including education, journalism, and digital archiving—where efficient access to unstructured multimedia content is essential.

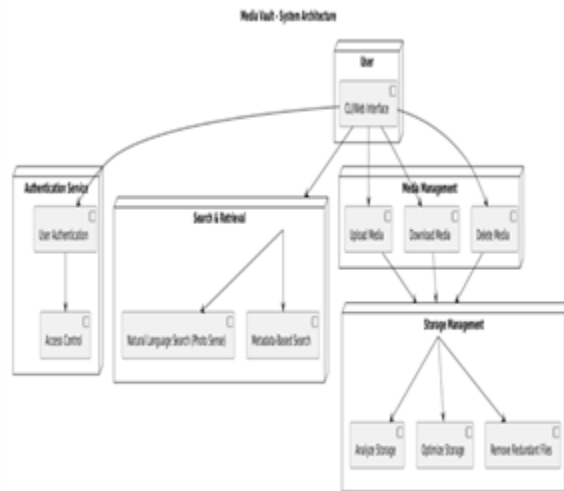


Fig 1: System Architecture

### 3.1 System Architecture

The architecture of MediaVault adopts a modular, service-oriented design intended to ensure scalability, maintainability, and user-centric functionality. The system comprises five primary components: the User Interface, Authentication Service, Search and Retrieval Module, Media Management, and Storage Management. These components are integrated to provide a secure, efficient, and intelligent multimedia retrieval environment.

Users interact with the system through either a command-line interface (CLI) or a web-based interface, enabling operations such as media upload, download, deletion, and content search. The Authentication Service manages user validation via secure login credentials and enforces access control mechanisms to safeguard data integrity and privacy.

Central to MediaVault is the Search and Retrieval Module, which consists of two subsystems: Natural Language Search (Photo Sense) and Metadata-Based Search. The Natural Language Search subsystem

employs SigLIP-generated vector embeddings to semantically align textual queries with stored image content, thereby delivering a more intuitive and context-aware retrieval experience. In contrast, the Metadata-Based Search supports conventional keyword-based retrieval when metadata—such as filenames or tags—is available.

The Media Management module facilitates user operations for uploading, downloading, and managing media files within personalized repositories. Each user's media is securely stored within a token-authenticated directory to ensure privacy and data isolation. To optimize storage efficiency, the Storage Management subsystem includes functionalities for monitoring usage statistics, restructuring file storage for improved performance, and eliminating redundant files. These measures collectively enhance system responsiveness and scalability, even as the volume of multimedia data grows.

This layered and modular architecture not only enables secure and efficient multimedia interaction but also establishes a robust foundation for future extensibility. Potential enhancements include integration with cloud storage solutions, implementation of AI-driven automated tagging, and support for collaborative multi-user environments.

### 3.2 Methodology

The methodology employed in the development of MediaVault is structured as a modular, data-driven pipeline designed to facilitate intelligent media search, secure storage, and seamless user interaction. The system architecture integrates techniques from computer vision, deep learning, and vector similarity search to enable natural language-based content retrieval.

The process begins with user-specific authentication, wherein each user is assigned a unique token-based access credential. This credential serves as the basis for creating and managing isolated media directories, thereby ensuring privacy and security for uploaded content. Upon successful authentication, users can upload media files through either a command-line interface (CLI) or a web-based frontend.

Each uploaded image undergoes preprocessing and is subsequently passed through the SigLIP embedding model, which generates high-dimensional vector representations capturing the semantic features of the

media content. These embeddings are normalized and indexed using the Facebook AI Similarity Search (FAISS) framework, facilitating efficient nearest-neighbor retrieval during search operations.

For query processing, users submit natural language queries through the interface. The system tokenizes and encodes these textual inputs using the same SigLIP model, ensuring that both the query and stored media reside within a shared semantic vector space. FAISS then computes similarity metrics between the query embedding and the stored image embeddings to identify and retrieve the most relevant media files.

In addition to semantic search, MediaVault also supports metadata-based retrieval for scenarios where users prefer conventional keyword filtering. This functionality enables searches based on filenames, tags, or other metadata attributes when semantic embeddings are not applicable. To further enhance performance and manage storage efficiently, the system continuously monitors usage metrics and storage utilization, enabling proactive optimization of system resources.

## IV. RESULTS

### 4.1 Experimental setup

To assess the effectiveness of the MediaVault system, a series of experiments was conducted in a controlled environment. The backend infrastructure was developed using Python 3.11 in conjunction with the Flask framework and integrated with the Facebook AI Similarity Search (FAISS) library to enable high-speed vector similarity searches. The SigLIP vision-language model was employed to generate semantic embeddings for both image content and natural language queries, ensuring consistent representation within a shared embedding space. Data storage was implemented using a local file system, organized into hashed user-specific directories to maintain data privacy and separation. Embedding mappings corresponding to each media file were stored in serialized index files to facilitate efficient retrieval operations. For experimental purposes, a dataset comprising over 500 images was assembled, encompassing diverse categories such as architecture, nature, events, and abstract art, thereby simulating a realistic personal media archive scenario. Each image in the dataset was processed through the SigLIP

model and indexed within the FAISS framework to support subsequent retrieval evaluations.

Hardware Configuration: Intel Core i7 Processor, 16 GB RAM

Software Configuration: Python 3.11, Flask, SigLIP (ViT-B-16-SigLIP2), FAISS, Torch, OpenCLIP, Pillow, SQLite

### 4.2 Results

The MediaVault system exhibited strong performance during experimental evaluation. Utilizing a dataset of 500 diverse images, the system consistently achieved low-latency retrieval and high semantic accuracy, effectively aligning natural language queries with visually and contextually relevant media. The integration of SigLIP-generated embeddings with FAISS-based indexing enabled typical query responses to execute in under 150 milliseconds. Additionally, the use of token-authenticated directories ensured secure and isolated media storage for individual users. The modular architecture demonstrated scalability, maintaining performance consistency with increasing dataset size. These findings underscore the system's effectiveness for applications requiring fast, accurate, and intelligent multimedia retrieval.

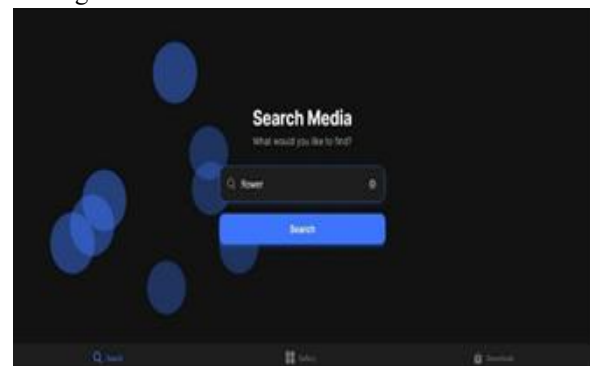


Fig 2: Media Vault Landing Page

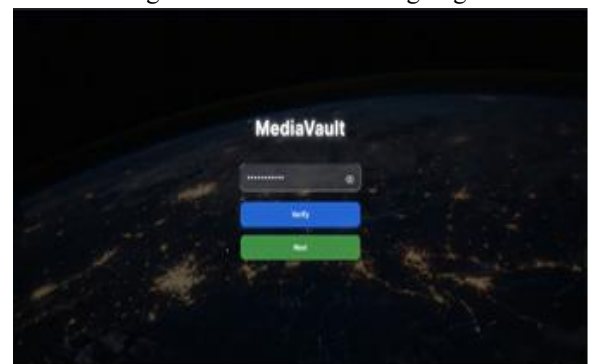


Fig 3: Media Vault Search Interface



Fig 4: Media Vault Search Results

#### 4.3 Discussion

- Semantic Retrieval Enhances User Experience: MediaVault enables users to perform image searches using natural language queries, leveraging SigLIP embeddings and FAISS indexing to facilitate context-aware retrieval without reliance on manual tags or metadata.
- Privacy-Preserving Architecture Supports Personalization: User media is securely stored within hashed, token-authenticated directories, ensuring data isolation and privacy. This architectural approach also supports future enhancements, such as advanced access control mechanisms and shared vault functionalities.
- Scalability and Modularity Facilitate Future Expansion: MediaVault's modular architecture enables seamless integration of additional intelligent capabilities, including video frame indexing, facial clustering, and AI-driven automated tagging, thereby enhancing the system's adaptability to evolving multimedia retrieval needs.

#### V. CONCLUSION

The MediaVault system effectively demonstrates the integration of deep learning techniques and semantic similarity search for efficient multimedia retrieval. By leveraging SigLIP embeddings in conjunction with FAISS indexing, the platform enables natural language-based image searches, delivering an intuitive and context-aware user experience. Its privacy-centric architecture ensures secure, user-specific media storage, while its modular design supports scalability and facilitates future feature

enhancements. MediaVault thus offers a practical and intelligent solution for personal media organization, educational content management, and archival systems. Potential future developments include extending the system's capabilities to video indexing, implementing automated tagging through generative models, and enabling cross-device synchronization.

This work was conducted as part of the academic final-year project undertaken by Authors 1, 2, and 3 in the Department of Information Technology at MVSR Engineering College, under the supervision of Author 4 during association with MVSR Engineering College.

#### REFERENCES

- [1] M. S. Rani, "A Novel Technique for Image Retrieval based on Concatenated Features of Big Transfer Networks," *International Journal of Image, Graphics and Signal Processing (IJIGSP)*, vol. 15, no. 2, pp. 1–9, 2023. DOI: 10.5815/ijigsp.2023.02.01
- [2] A. K. Jain and B. B. Gupta, "Image Retrieval by Utilizing Structural Connections within an Image," *IJIGSP*, vol. 8, no. 1, pp. 67–74, 2016. DOI: 10.5815/ijigsp.2016.01.08
- [3] M. K. Singh and R. S. Anand, "Image Retrieval using Hypergraph of Visual Concepts," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 9, no. 12, pp. 45–53, 2017. DOI: 10.5815/ijitcs.2017.12.05
- [4] S. R. Dubey and S. K. Singh, "Efficient Global and Region Content Based Image Retrieval," *IJIGSP*, vol. 4, no. 5, pp. 41–49, 2012. DOI: 10.5815/ijigsp.2012.05.05
- [5] R. K. Agrawal and R. S. Anand, "Image Retrieval based on Local Motif Patterns Code," *IJIGSP*, vol. 10, no. 6, pp. 56–64, 2018. DOI: 10.5815/ijigsp.2018.06.07
- [6] Md. Farhan Sadique and S. M. Rafizul Haque, "Content-Based Image Retrieval Using Color Layout Descriptor, Gray-Level Co-Occurrence Matrix and K-Nearest Neighbors," *IJITCS*, vol. 12, no. 3, pp. 19–25, 2020. DOI: 10.5815/ijitcs.2020.03.03
- [7] Mohamed M. Fouad, "Content-based Search for Image Retrieval," *IJIGSP*, vol. 5, no. 11, pp. 46–52, 2013. DOI: 10.5815/ijigsp.2013.11.05

- [8] Jiashu Xu, “A Review of Self-supervised Learning Methods in the Field of Medical Image Analysis,” IJIGSP, vol. 13, no. 4, pp. 20–27, 2021. DOI: 10.5815/ijigsp.2021.04.03
- [9] M. Mansourvar, S. Shamshirband, R. G. Raj, R. Gunalan, and I. Mazinani, “A Genesis of a Meticulous Fusion based Color Descriptor to Analyze the Supremacy between Machine Learning and Deep Learning,” International Journal of Intelligent Systems and Applications (IJISA), vol. 12, no. 2, pp. 33–40, 2020. DOI: 10.5815/ijisa.2020.02.03