Open Access
Full Length Article

ORIGINAL ARTICLE

An efficient Semantic Search System to Retrieve Photos using CLIP, BLIP and FAISS

Imtiaz Ali Dahri[1], Fida Hussain Dahri[2], Nisar Ahmed Dahri[3], Sajjad Hussain Bhutto[4]

[1, 2]Department of Mathematics and Statistics, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah 67480, Pakistan
[3]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[4]Shanghai Dianji University, 2021306, Shanghai, China

ABSTRACT

The rapid growth of personal photo collections has highlighted the limitations of traditional retrieval methods that rely on filenames or manual tagging, creating a persistent semantic gap between image content and user intent. This study proposes an efficient semantic search system that enables natural-language-based retrieval of personal photos through a hybrid multimodal architecture. The system integrates CLIP for vision–language alignment, BLIP for automatic caption generation, and FAISS for high-speed similarity search, combining direct visual semantic matching with text-mediated understanding. Evaluated on a dataset of 2,000 personal images, the system demonstrates strong retrieval performance, achieving mean cosine similarity scores of 0.254-0.282 and perfect Precision@6 (1.000) for attribute-based queries. Results indicate high robustness to variations in lighting, pose, and background, with the strongest performance observed in clothing- and object-related searches, while event-based queries remain more challenging. Key contributions include a privacy-preserving on-device design, a dual-pathway multimodal workflow, and a comprehensive evaluation framework for semantic photo retrieval. Overall, the proposed system effectively bridges the semantic gap in personal photo management and demonstrates the practical value of multimodal AI for intuitive, human-centred image retrieval.

Keywords: *Multimodal Retrieval, Semantic Search, Personal Photo, Management, CLIP, BLIP, FAISS, Privacy-PreservingAI*

Author`s Contribution
[1,2] Data analysis, interpretation, and manuscript writing, Active participation in data collection
[2,3,4] Conception, synthesis, planning of research. Interpretation and discussion

Address of Correspondence
Fida Hussain Dahri
Email: fidahussain@stu.hit.edu.cn

Article info.
Received: June,15, 6, 2025
Accepted: November 13, 2025
Published: December 20, 2025

Cite this article: *Dahari A I, Dahri H f, Dahri A N, Butto H SAn efficient Semantic Search System to Retrieve Photos using CLIP, BLIP and FAISS. J. inf. commun. technol. robot. appl.2025; 14(1):36-53*

Funding Source: *Nil*
Conflict of Interest: *Nil*

INTRODUCTION

The rapid growth of people's personal collections of digital photographs has created unique challenges for information retrieval and human-computer interaction systems. In fact, empirical data show that the volume of photos taken worldwide has increased by 227% from 2013 to 2025, with the number of photos taken today expected to rise from 0.66 trillion to 2.10 trillion annually (see Figure 1; Photutorial, 2025). Tripling the number of photographs taken and shared within 12 years at an accelerating pace, we expect even more ways to engage with and manage these photos in the future.

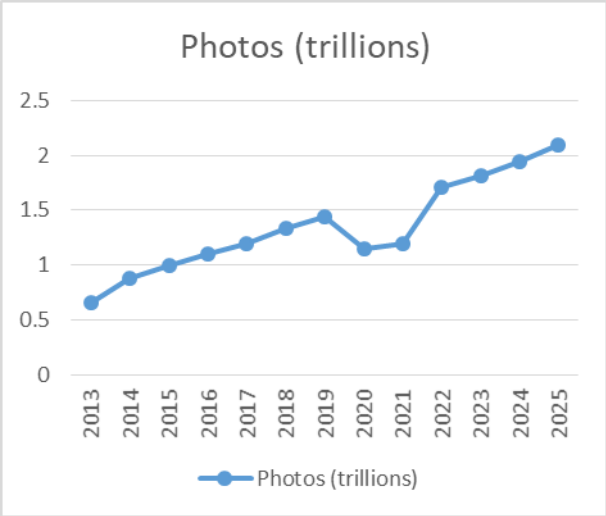

Figure 1. Global annual Photos Captured 2013 to 2025 (Photutorial, 2025). Note: the decline from 2019 to 2021 is due to the COVID-19 pandemic.

Current photo management cannot be separated from traditional photo organisation based on filenames, manual tagging, or basic metadata extraction (Rodden & Wood, 2003). This semantic nature is like human memory for example you should remember the memories of the pictures you take as "the family picnic by the lake" or "the graduation ceremony with friends"; the terms themselves cannot replace the organisational space available for a photo management system to index the content of that memory, so there remains a significant gap between how we conceptualise our memories as humans and the way that standard management systems organise visual content (Smeulders et al., 2000). The challenge, or sometimes referred to as the "semantic gap", illustrates the difference between the low-level visual features the computer understands regarding our memories, despite the high-level human agent knowledge of what we store in our photographs (e.g. content and context on both categories of objects and events). This difference is apparent with personal photos. How we organise and search our own photos is very different from how we search for images online. Photo retrieval presents challenges in personal photo collections that differ from those in general-purpose image search systems. Personal collections may have very few or no structured annotations and exhibit significant semantic variability, even in relatively small collections. Most users organise their photographs chronologically, and everyone expects the folder structure to follow suit. Only basic Exchangeable Image File Format (EXIF) metadata exists; this has also made it harder to find the right images by their meaning. These issues come from a fundamental problem in image search, the "semantic gap." This is the difference between the raw visual data in a photo and the higher-level concepts people understand (Smeulders et al., 2000; Datta et al., 2008). Early image search tools used basic visual features, such as colour patterns, textures, and outlines, to find similar images (Flickner et al., 1995; Datta et al., 2008).

Sometimes, images can look identical in colour or texture, but that doesn't mean they have the same meaning, which is what people actually care about when searching for photos. Imagine determining an image of "celebrating moments": there are complex contextual signs, facial expressions, and social interactions among multiple people, which go far beyond a pixel-level view. These limitations in purely visual approaches have driven the development of more sophisticated multimodal retrieval systems, capable of connecting different types of data at a shared semantic level. Deep learning brought about changes in injection multimodal retrieval systems. It gave rise to cross-modal systems that connect multi-design with multimodal systems at the semantic level through learned representation. Before the advent of deep learning, systems treated images and text separately, through independent pipelines. It was common to combine the two at a later stage of the process with explicit fusion mechanisms and customisation for each domain (Baltrušaitis et al., 2018). The embeddings produced by these systems were

independently derived, resulting in minimal semantic interpretation and poor generalisation across query types.

The arrival of transformer-based architectures (Vaswani et al., 2017) has significantly boosted multimodal learning by enabling joint representations of vision and text. Methods developed by models such as CLIP (Radford et al., 2021) align images and text in a common semantic space via contrastive learning objectives. The CLIP architecture utilised separate encoders for visual and textual elements. For instance, a ViT (Dosovitskiy et al., 2021) for images and a transformer for natural language. The image and text encoders of CLIP are learnt to maximise the cosine similarity of matching image/text pairs and minimise otherwise. CLIP has set the foundation, but it is unified vision-language architectures like BLIP (Bootstrapping Language Image Pre-training) that target bidirectional understanding (Li et al., 2022). BLIP combines contrastive image-to-text learning, image-conditioned language modelling, and language-conditioned image generation into a multi-task learning approach. By automatically generating captions for visual content, BLIP provides rich textual descriptions that can be reused for retrieval based on similarity.

However, despite these advances, multimodal models still face significant scalability and efficiency challenges. Personal photo collections, while not as large as web-scale collections, still require efficient query processing and minimal storage costs for effective real-world use. The embedding spaces from modern models are often high-dimensional (e.g. 512 and 768 dimensions for the best models, Radford et al. (2021)) and require similarity search methods that are fast enough to provide instantaneous responses. One widely used method for similarity search is Facebook AI Similarity Search (FAISS), which has become a standard for large-scale similarity search and provides efficient implementations of ANN algorithms (Johnson et al. 2021). Over the years, it has been adapted to various indexing needs. This includes flat L2 search, which gives exact results; inverted file (IVF), which sacrifices some accuracy for less memory usage; and HNSW, built for fast retrieval and efficient memory use. Deciding which strategy to use is not straightforward. It often means weighing speed, memory use, and accuracy. It's particularly hard to do this when working with a device with limited processing power and storage (Johnson et al. 2021). When considering performance factors, designers must also consider privacy and user experience. Leading photo platforms appear to be at an inflexion point. It enables rich searches, but customers give up control over their data. When individuals have private pictures on servers, as they do in most commercial and industrial systems, they lose control over them by default. Security breaches are not the only vulnerabilities that can occur (Armbrust et al., 2010). Undocumented algorithmic bias also counts. Users do not know how search rankings affect things. Many researchers call the use of AI a "black box problem", as it does not show how results are achieved (Adadi & Berrada, 2018)

In my opinion, many photo upload sites are still outdated in terms of look and functionality. While modern AI systems seem advanced, behind the scenes, they rely on old systems (like folder hierarchies, timeline views, and manual tagging) rather than better ones to help you catalogue and find your pics. Rodden and Wood (2003) showed decades ago that these structures are not aligned with human memory organisation. A more natural solution would align with cognitive psychology: users should be able to engage with the system through conversational search for "Maya's birthday cake disaster" rather than having to remember the date or guess a keyword (Radlinski & Craswell, 2017). But, it is not an easy task to make them everyday tools. Using laboratory-grade multimodal AI on personal devices shows hard engineering limitations. These computer systems require lots of power.

The specialised hardware needed to run these systems is costly. They really limit consumer use. Developers have to make the most of their small collection when they want to make thousands of images. Techniques like model quantisation or Knowledge distillation can reduce computational costs. But these generally give results that are unacceptably inaccurate for user applications. It also makes the overall situation more complicated, making it harder to integrate into existing systems. A typical search system is composed of unstable components, such as the image encoder, text

model, caption generators, and search index. Each of these can fail in its own way. Failure to manage this carefully can lead to significant technical debt from overspending. Real-world images are messier than research images with labels. There can be considerable variation in content and quality depending on the text type. As a result, many data-cleaning steps are often ignored, especially in early-stage prototypes.

This study presents an efficient, semantic photo retrieval system. And the photomanagement system has certain drawbacks that our system addresses. The proposed system uses Multimodal Search (MPEG-7), a Bidirectional LSTM, and a multiclass SVM. The key contributions of this work are.

- This novel study proposes workflows integrating CLIP for image-text alignment, BLIP for automatic caption generation, and FAISS for efficient search of photo collections.
- A cloud-free, on-device processing approach that ensures privacy while maintaining strong retrieval performance and optimised computational efficiency.
- An approach that applies two evaluation methods to personal photo datasets is reported. The first method tests the effectiveness of a quantitative evaluation of the photos. This involves using pre-fixed metrics, including precision, recall, and cosine similarity. The second method implements a qualitative assessment that is based on the user's experience.
- The system is demonstrated with a dataset of over 2,000 personal photos, showcasing its scalability, computational efficiency, and real-world usability.

Further, our research paper is structured as follows. Section 2 presents a comprehensive literature review of multimodal retrieval, captioning models, and similarity search methods. Section 3 outlines the key research gaps and motivations for the proposed system. Section 4 details the methodology, including data preprocessing, CLIP–BLIP feature extraction, dual-pathway representation, and FAISS indexing. Section 5 describes the evaluation framework and metrics used for quantitative and qualitative assessment. Section 6 reports and discusses the experimental results. Finally, Section 7 concludes the study and highlights directions for future work

LITERATURE REVIEW

The field of image retrieval has been transformed by the advancements in deep learning and multimodal AI systems. Conventional content-based image retrieval (CBIR) frameworks mostly rely on low-level visual descriptions, such as colour, texture, and shape features (Datta et al., 2008). However, these techniques were limited by the "semantic gap." This means the computer could analyse basic image features, such as colours and shapes, but failed to understand higher-level ideas, such as "winning moments" or "celebration" (Smeulders et al., 2000). To overcome this shortcoming, advances in deep convolutional neural networks (CNNs) led to a paradigm shift in image retrieval, enabling the computation of more semantically grounded visual representations (Krizhevsky et al., 2012). Initial deep learning methods were dedicated to learning discriminative representations of images by using directly supervised learning on datasets of scale such as ImageNet (Deng et al., 2009). Also, although these approaches were more efficient than classical CBIR systems, they were still hindered by their reliance on predefined visual categories and their inability to handle natural language queries directly. This limitation led to the development of vision-language models, which represented a paradigm shift in language by narrowing the semantic gap between visual content and natural-language comprehension. Radford et al. (2021) introduced the practice of contrastive learning on a massive-scale training regime underpinning joint embeddings of images and text via CLIP (Contrastive Language-Image Pre-training). Training on 400 million image-text pairs, CLIP showed impressive zero-shot transfer across many visual recognition tasks. Building on this success, a considerable number of studies have built on this idea, significantly improving how computers understand both images and language. ALIGN pre-trained the contrastive scaling on 1.8 billion paired pictures and texts, demonstrating that scaling up delivers better performance on several benchmarking tasks (Jia et al., 2021). Florence proposes a unified vision-language foundation model that achieves state-of-the-art performance on both discriminative and generative vision-

language tasks (Yuan et al., 2021). Recent studies have addressed the issue of enhancing the efficiency and effectiveness of contrastive learning in vision-language models following these developments. DeCLIP introduced supervised semantic softmax to alleviate data inefficiency in contrastive learning (Li et al., 2022), and RegionCLIP extended CLIP to region-level understanding, enabling object detection (Zhong et al., 2022). All of these developments have led to the emergence of contrastive vision-language models as the preferred paradigm for multimodal understanding tasks.

The field of automatic image captioning has grown tremendously and has since surpassed encoder-decoder architectures, incorporating multimodal transformers. BLIP (Bootstrapping Language-Image Pre-training), proposed by Li et al. (2022), can be considered another significant breakthrough in this field, as the authors introduce a unified approach that simultaneously maximises both understanding and generation goals. BLIP's most important innovation is its bootstrapping approach. This means the model generates its own training data using synthetic captions. On this basis, several significant extensions and enhancements have been made to BLIP's success. The latest updates to that architecture in BLIP-2 (Li et al., 2023) improved efficiency by using frozen pre-trained image encoders and language models, enabling better performance with much less computation. Instruct BLIP (Dai et al., 2023) improved upon BLIP-2 by allowing it to follow user instructions, making image captioning more flexible and practical. Simultaneously, a few other prominent changes have been made in image captioning, such as CoCa (Yu et al., 2022), which combines a contrastive objective with a captioning objective in a single framework, and Flamingo (Alayrac et al., 2022), which demonstrates few-shot learning in vision-language tasks. All of these developments have enhanced the quality and range of automatically-generated image descriptions, specifically by producing descriptions that are higher quality and easier to use in downstream applications such as semantic search.

Efficient similarity search in high-dimensional vector spaces is needed to scale the use of these capabilities. Johnson et al. (2021) created FAISS (Facebook AI Similarity Search), which has become the default large-scale similarity search implementation, addressing several limitations of traditional string matchers. FAISS offers a comprehensive set of indexing algorithms that enable developers to optimise for search precision, memory efficiency, or query speed. On these grounds, recent trends in vector databases have focused on enhancing scalability and reducing computational overhead. The use of approximate nearest neighbour (ANN) algorithms, such as HNSW (Hierarchical Navigable Small World) (Malkov & Yashunin, 2020), IVF (Inverted File) (Jegou et al., 2011), and others, has been widely investigated and optimised for various applications. Similarity search operations have been further enhanced through specialised hardware acceleration, such as GPU-based approaches and custom silicon (Johnson et al., 2021). In combination with these hardware innovations, similarity search has also been the focus of significant effort in integrating it into machine learning pipelines. New product offerings such as Pinecone (Pinecone Systems Inc., 2025), Weaviate (Weaviate B.V., 2025), and Milvus (Wang et al., 2021) provide complete end-to-end solutions for embedding-based applications, including real-time indexing, distributed search, and integration with popular machine learning libraries.

Personal photo management has been revealed as an essential field for the application of semantic search technologies. Manual tagging and folder-based hierarchies of visual materials have shown that traditional methods of organising photos are impractical for managing extensive personal collections. This problem is exacerbated by the fact that most users associate photos with semantic elements, such as events, names, and objects, rather than metadata and temporal elements (Rodden & Wood, 2003). To combat this problem, a variety of commercial systems have used AI-based solutions. Google Photos offered object recognition-based clustering and spot search features through face identification (Anil, 2015). The same features were applied to Apple Photos, which aimed to process images on-device to ensure privacy (Apple Inc., 2024). However, such systems mainly depend on predefined visual categories. This constitutes a substantial limitation because they are unable to process arbitrary, open-ended natural-language requests (Radford et al., 2021).

Developing this option further, a new direction in exploring multimodal semantic image retrieval has emerged, combining visual, textual, and structural attributes (Pustu-Iren et al. 2021). They showed that combining deep visual features (CLIP) with OCR-extracted text and patent document context yields better access to the image content of patents. Continuing this strategy of multimodal approaches, integrating a large language model (LLM) with photo search has yielded the best results. The advances of impressive visual language models that achieve the few-shot learning capability, Flamingo (Alayrac et al., 2022), and big language models that can follow instructions through the task of InstructBLIP (Dai et al., 2023), lay the groundwork here to finally interact in a truly conversational manner with photo collections finally. Extending this functionality, models such as Visual ChatGPT have shown users how to engage in sophisticated dialogue to edit, query, and reason over their visual media (Wu et al., 2023). These trends and advances indicate a significant leap toward a more natural and adaptable management of the photo interface.

Testing how well semantic image search systems work for personal photos is tricky because deciding if a result is truly relevant is very subjective. What feels like a perfect match to one person might not be as meaningful to another. While standard evaluation metrics such as precision, recall, and average precision are still used, they often fall short of capturing the subtle and personal ways people interpret meaning and similarity in their own photos (Pustu-Iren et al., 2021). These standard metrics can't measure the full depth of what makes images semantically similar to users on a personal level. Recent efforts have introduced CLIP-Score, which uses CLIP embeddings to evaluate image-text alignment (Radford et al., 2021).

Nevertheless, these automated methods can inherit one's bias from their models. However, human evaluation is time-consuming and does not work well for large systems. To address this, researchers have begun developing hybrid methods that combine feedback from both crowd workers and expert annotators. This helps create high-quality evaluation data more efficiently (Hodosh et al., 2013).

Research is still ongoing on developing standard benchmarks for personal photo retrieval. Although these developments have been made, several issues and boundaries still exist regarding the creation of these kinds of photo search systems. The problems of privacy and data security are considerable in the semantic search systems. A trade-off between on-device and cloud-based processing is a key challenge in managing personal photo collections, as each new option introduces new vulnerabilities, including data leakage and attacks by adversaries (Das et al., 2024). Furthermore, the high computational cost of training and inference for large multimodal models makes them unsuitable for limited-resource devices and raises concerns about their substantial energy consumption (Schwartz et al., 2019; Strubell et al., 2019). Another big question, besides all these computational issues, is bias and fairness. Vision-language models frequently exhibit systemic bias by race, gender, and culture, thereby introducing search bias that can harm all user communities (Ferrara, 2024; Saleha & Tabatabaeib, 2025).

Another major problem is that these AI systems are not transparent. It's challenging to understand how they decide which photo to retrieve. Because users need to trust the results, it's essential to know the reason behind a decision. However, the complex "black-box" nature of deep learning poses a significant challenge. This is why recent research strongly recommends creating an AI system that is easier to understand and explain. (Arrieta et al., 2019; Joshi et al., 2021).

RESEARCH METHODOLOGY

Research Gaps and Opportunities

In our literature survey, we identified significant research gaps that will inform the proposed work. Individual components of CLIP, BLIP, and FAISS have been thoroughly analysed; however, little has been proposed for systematically combining them for personal photo retrieval. What is more, most current evaluation-related studies focus on standard image retrieval benchmarks rather than on the peculiarities of working with individual photos. There is another gap: the complementary strengths and weaknesses of various

multimodal AI methods when integrated into hybrid systems are not adequately explored. To fill these gaps, the present work proposes a new hybrid architecture that incorporates the power of contrastive vision-language models, automated image captioning, and efficient similarity search in a systematic manner. This study advances knowledge of the effective operation of these technologies in real-world applications by conducting a detailed analysis of a large dataset of personal photos, which opens the door to more effective semantic search and retrieval within the system for managing individual images.

This study follows a structured four-stage multimodal retrieval pipeline integrating visual–textual feature extraction, dual-pathway representation, vector indexing, and natural-language query processing. Figure 2 illustrates the overall workflow. The framework is designed to bridge the semantic gap by aligning images and text in a shared embedding space and enabling fast nearest-neighbour retrieval.

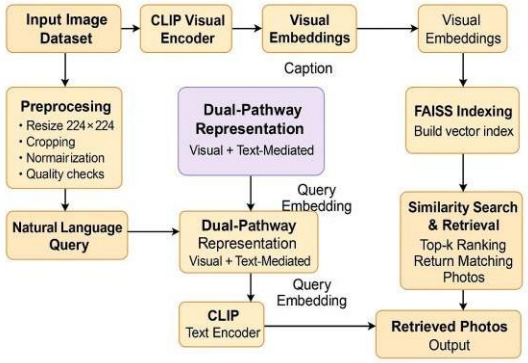

Figure 2. Methodology flow diagram of the proposed CLIP-BLIP-FAISS semantic search system for personal photo retrieval.

1.1. Dataset Description and Preprocessing

This dataset consists of 2000 high-resolution personal photos (1080p JPEG). We checked the image file format using metadata contained in the JPG file. Images were resized to 224×224, centre-cropped, and normalised with ImageNet statistics:

$$\text{Normalize}(x) = \frac{x-\mu}{\sigma}, \mu = [0.485, 0.456, 0.406], \sigma = [0.229, 0.224, 0.225]. \quad (1)$$

All computations were performed locally to comply with privacy regulations.

1.2. Multimodal Feature Extraction

1.3. Visual Embeddings (CLIP)

Each image I is encoded via the CLIP ViT-B/32 visual transformer:

$$v = f_{\text{CLIP}}(I) \in \mathbb{R}^{512}, \quad (2)$$

Followed by L2-normalisation:

$$\hat{v} = \frac{v}{\|v\|_2}. \quad (3)$$

CLIP learns image-text alignment using a contrastive objective:

$$\mathcal{L}_{\text{CLIP}} = -\log \frac{\exp(\text{sim}(v,t)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v,t_j)/\tau)}, \quad (4)$$

Where t is the paired caption embedding, and τ is a learnable temperature parameter.

1.4. Textual Semantics via BLIP Captioning

BLIP generates a textual description c for each image:

$$c = g_{\text{BLIP}}(I), \quad (5)$$

Followed by a text encoder that produces a semantic embedding:

$$t = f_{\text{text}}(c), t \in \mathbb{R}^{512}. \quad (6)$$

This provides high-level semantics complementing the visual pathway.

1.5. Dual-Pathway Representation

Two parallel embeddings represent each image:

Visual pathway:

$$\hat{v} \quad (7)$$

Text-mediated pathway:

$$\hat{t} \quad (8)$$

The hybrid representation enables compositional and abstract semantic retrieval:

$$h = \alpha\hat{v} + (1-\alpha)\hat{t}, 0 \le \alpha \le 1. \quad (9)$$

In this study, retrieval is performed independently on both pathways, and results are merged.

1.6. Vector Indexing Using FAISS

All embeddings are stored in a FAISS IndexFlatIP structure, optimised for cosine similarity:

$$\text{sim}(x, y) = \frac{x \cdot y}{\|x\|\|y\|}$$

(10)

Index construction includes:

L2 normalisation

Index initialisation

Adding embeddings

Serialisation

Given a query embedding q, FAISS returns top- k neighbours:

$$\{I_1, I_2, \ldots, I_k\} = \text{FAISS}(q, k).$$

(11)

This supports efficient $O(n)$ similarity search accelerated via GPU computation.

1.7. Query Processing and Retrieval

A natural-language query Q is encoded using CLIP:

$$q = f_{\text{CLIP-text}}(Q), \hat{q} = \frac{q}{\|q\|_2}.$$

(12)

Retrieval is performed in the shared embedding space:

$$\text{score}(i) = \text{sim}(\hat{q}, \hat{v}_i), \text{ or } \text{sim}(\hat{q}, \hat{t}_i).$$

(13)

Top results are ranked by similarity score and returned to the user.

Further, Algorithm 1 details the full processing pipeline for operationalising the multimodal retrieval framework, which incorporates CLIP-based visual encoding (Radford et al., 2021), caption generation using BLIP (Li et al., 2021), and FAISS vector indexing for efficient nearest-neighbour search.

1.7.1. Dataset Description and Preprocessing

This dataset consists of 2000 high-resolution personal photos (1080p JPEG). We checked the image file format using metadata contained in the JPG file. Images were resized to 224×224, centre-cropped, and normalised with ImageNet statistics:

$$\text{Normalize}(x) = \frac{x-\mu}{\sigma}, \mu = [0.485, 0.456, 0.406], \sigma = [0.229, 0.224, 0.225].$$

(1)

All computations were performed locally to comply with privacy regulations.

1.8. Multimodal Feature Extraction

1.9. Visual Embeddings (CLIP)

Each image I is encoded via the CLIP ViT-B/32 visual transformer:

$$v = f_{\text{CLIP}}(I) \in \mathbb{R}^{512},$$

(2)

Followed by L2-normalisation:

$$\hat{v} = \frac{v}{\|v\|_2}.$$ (3)

CLIP learns image-text alignment using a contrastive objective:

$$\mathcal{L}_{\text{CLIP}} = -\log \frac{\exp(\text{sim}(v,t)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v,t_j)/\tau)},$$

(4)

Where t is the paired caption embedding, and τ is a learnable temperature parameter.

1.9.1. Textual Semantics via BLIP Captioning

BLIP generates a textual description c for each image:

$$c = g_{\text{BLIP}}(I),$$

(5)

Followed by a text encoder that produces a semantic embedding:

$$t = f_{\text{text}}(c), t \in \mathbb{R}^{512}.$$

(6)

This provides high-level semantics complementing the visual pathway.

1.9.2. Dual-Pathway Representation

Two parallel embeddings represent each image:

Visual pathway:

$\hat{v}$ (7)

Text-mediated pathway:

$\hat{t}$ (8)

The hybrid representation enables compositional and abstract semantic retrieval:

$h = \alpha\hat{v} + (1-\alpha)\hat{t}, 0 \le \alpha \le 1.$

(9)

In this study, retrieval is performed independently on both pathways, and results are merged.

1.9.3. Vector Indexing Using FAISS

All embeddings are stored in a FAISS IndexFlatIP structure, optimised for cosine similarity:

$$\text{sim}(x, y) = \frac{x \cdot y}{\|x\|\|y\|}$$

(10)

Index construction includes:

L2 normalisation

Index initialisation

Adding embeddings

Serialisation

Given a query embedding q, FAISS returns top- k neighbours:

$$\{I_1, I_2, \ldots, I_k\} = \text{FAISS}(q, k).$$

(11)

This supports efficient $O(n)$ similarity search accelerated via GPU computation.

1.9.4. Query Processing and Retrieval

A natural-language query Q is encoded using CLIP:

$$q = f_{\text{CLIP-text}}(Q), \hat{q} = \frac{q}{\|q\|_2}.$$

(12)

Retrieval is performed in the shared embedding space:

$$\text{score}(i) = \text{sim}(\hat{q}, \hat{v}_i), \text{ or } \text{sim}(\hat{q}, \hat{t}_i).$$

(13)

Top results are ranked by similarity score and returned to the user.

Further, Algorithm 1 details the full processing pipeline for operationalising the multimodal retrieval framework, which incorporates CLIP-based visual encoding (Radford et al., 2021), caption generation using BLIP (Li et al., 2021), and FAISS vector indexing for efficient nearest-neighbour search.

Algorithm 1. Proposed Multimodal Semantic Retrieval Algorithm

Input:

I = {I1, I2, …, In} Image dataset

Q Natural language query

Output:

R: Ranked set of retrieved photos

1:---------- Preprocessing ----------

2: For each image Ii ∈ I do

3: Ii_resized ← Resize(Ii, 224×224)

4: Ii_clean ← Normalise(Ii_resized)

5: Ii_valid ← QualityCheck(Ii_clean)

6: End For

7:---------- Multimodal Feature Extraction ----------

8: For each image Ii ∈ I do

9: v_i ← f_CLIP(Ii) Visual embedding (Radford et al., 2021)

10: c_i ← g_BLIP(Ii) Caption generation (Li et al., 2022)

11: t_i ← f_text(c_i) Text embedding
12: v̂_i ← Normalize(v_i)
13: t̂_i ← Normalize(t_i)
14: End For
15:---------- Dual-Pathway Representation ----------
16: For each image Ii ∈ I do
17: h_i ← α·v̂_i + (1−α)·t̂_i Hybrid representation
18: End For
19:---------- FAISS Indexing ----------
20: index ← Init_FAISS()
21: index.add(h_i for all i) Build vector index
22:---------- Query Processing ----------
23: q ← f_CLIP_text(Q) Encode query
24: q̂ ← Normalize(q)
25:---------- Similarity Search & Ranking ----------
26: scores, R ← index.search(q̂, k) Top-k retrieval
27: R ← SortDescending(scores) Rank by cosine similarity
28: Return R

1.9.5. Evaluation Methodology

Evaluation looks at numbers and other things. Queries were divided into attribute, relationship, and activity categories.

1.9.6. Quantitative Metrics

(1) Mean Cosine Similarity

Measures the average semantic alignment of the top-6 retrieved images:

$$\text{MeanSim} = \frac{1}{6}\sum_{i=1}^{6} \text{sim}_i.$$

(14)

(2) Precision@6

Fraction of top-6 relevant results:

$$\text{Precision @6} = \frac{\text{Number of Relevant Results in Top 6}}{6}.$$

(15)

(3) Recall@6

Fraction of all relevant images retrieved in top-6:

$$\text{Recall@6} = \frac{\text{Relevant in Top 6}}{\text{Total Relevant in Dataset}}$$

(16)

(4) F1-Score@6

Balanced measure of accuracy and completeness:

$$\text{F1@6} = \frac{2\times(\text{Precision @6}\times\text{Recall @6})}{\text{Precision @6}+\text{Recall @6}}.$$

(17)

1.9.7. Qualitative Evaluation

Qualitative analysis examined:
Top-6 semantic relevance (manual annotation)
Error taxonomy:

- false positives
- false negatives

partial matches
Robustness to lighting, pose, and background variation
Further, the similar reliability for complex multi-attribute inquiries
The purpose of this layer is to evaluate abstract queries that are not based on numeric ones.
Ethical and Validity Controls
Privacy-by-design: all processing performed locally
Bias mitigation: cross-checking results for demographic sensitivity

- Reproducibility: fixed random seeds, version-controlled environment

- External validity: tested across diverse subcategories of personal photos

RESULTS AND DISCUSSION

1.9.8. Quantitative Evaluation

In this section, we evaluate our semantic search system, enhanced with artificial intelligence, for retrieving images from the collection. We take a look at how systems perform on quantitative metrics and qualitative user experience. We highlight system capacity and limits across queries. The quantitative evaluation demonstrates the effectiveness of our hybrid multimodal architecture across diverse personal photo queries. Table 1 and Figure 3 present comprehensive performance metrics for 10 representative queries, demonstrating the system's ability to handle diverse semantic search scenarios.

Table 1: Quantitative Performance Metrics for Representative Query Types

Query	Mean Score	Precision@6	Recall@6	F1_Score@6	Relevance
Wearing Tie	0.274	0.833	0.051	0.096	98
In an orange-coloured Shirt sitting on a Tree	0.262	1.0	0.016	0.032	372
Winning Moments	0.258	0.0	0.0	0.0	0
Wearing Glasses	0.271	1.0	0.052	0.099	115
Celebrating Moments	0.281	0.0	0.0	0.0	0
I am with two people only	0.278	1.0	0.015	0.029	403
Study time	0.254	0.0	0.0	0.0	0
I am with more than three people	0.277	0.833	0.012	0.024	403
I am with the Kids	0.282	0.833	0.012	0.025	401
I am in white clothes	0.282	1.0	0.015	0.029	401

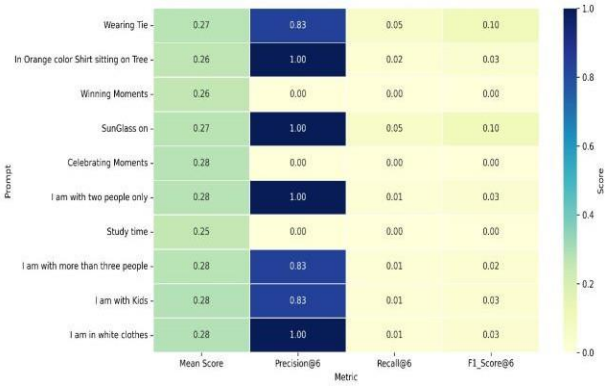

Figure 3. Evaluation Metrics Heatmap Per Prompt

The results demonstrate that the proposed hybrid architecture successfully achieves its primary objective: enabling photo retrieval from a personal collection using natural language prompts. Even though recall values remain relatively low across most queries, the consistently high precision shows that when the system gives results, they are almost always relevant to the user's intent. This confirms the feasibility of retrieving meaningful or required images directly through descriptive queries, which is the central contribution of this research.

For queries based on clear visual attributes, such as “Wearing Tie”, “Wearing Glasses”, and “I am in white clothes”, the system produced accurate and valuable results, with precision values ranging from 0.833 to 1.0. These results indicate that the system reliably identifies and returns accurate photos. In fact, this means that a user searching for specific images in their collection is getting the required pictures with correct results, with minimal irrelevant output.

Although recall values were low from 0.012 to 0.015, particularly for relational prompts like “I am with two people only” or “I am with kids”, the system still retrieved at least some relevant examples for these cases because these queries are achieving relatively high precision from 0.833 to 1.0. This shows that the framework can understand and respond to more complex details, even if it does not capture the complete set of possible outcomes. From the user’s perspective, this remains valuable, as it ensures that prompts involving social contexts yield meaningful outputs.

In contrast, prompts related to activities and events, such as “Winning Moments,” “Celebrating Moments,” and “Study time,” did not score highly in the quantitative evaluation. Yet, they still returned meaningful images in practice. For instance, queries about winning or celebrating often brought up photos of people holding a cup or trophy (See Figure 8), while the study time prompt frequently surfaced images (See Figure 9) with books either in hand or somewhere in the scene. This shows that even if the system struggles to capture the whole abstract idea behind these events. The system can still pick up on visual hints that act as substitutes for the intended meaning. This partial success is encouraging, as it shows the framework is not limited to simple attribute recognition but is beginning to move toward higher-level, more contextual interpretations. With further refinements, such as incorporating metadata like timestamps, applying scene classification, or fine-tuning on rich datasets, the system could become much more effective at consistently retrieving event-focused content.

To further strengthen the evaluation, we have used 50 diverse natural language queries that included attributes, relationships, and activities. The precise boxplot for 50 queries is shown in Figure 4. It summarises the distribution of evaluation metrics for all queries.

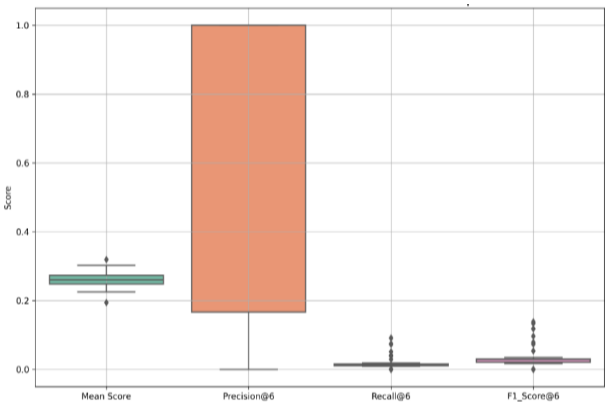

Figure 4. Illustrates the allocation of the performance metrics to the 50 queries. The results show great precision but limited recall. Performing retrievals is selective.

The results show that the mean similarity scores remain consistent at 0.25-0.28, indicating stable feature matching across all query types. Precision values are generally high, with many queries achieving near-perfect retrieval accuracy, confirming that when the system returns results, they are usually relevant/helpful. But, recall and F1 scores remain low, highlighting the system’s limitation in capturing the full range of relevant images. This pattern is evident in event-based queries, where the framework often relies on strong visual cues (e.g., trophies for winning, books for studying) rather than fully understanding the event’s abstract semantics.

The system does not perform equally well for all types of queries, but it still achieves its main goal: allowing people to find personal photos using

natural language. Queries about clear visual attributes work best; relational queries yield results that are still useful, and event-based prompts show early signs of understanding context. Overall, this confirms the practical value of the hybrid model using CLIP, BLIP and FAISS approach, while also highlighting areas for improvement. Future work, such as adding metadata, using scene classification, or fine-tuning datasets rich in activities and events, could make the system even better at handling more complex, abstract searches.

1.9.9. Qualitative Evaluation

A qualitative evaluation of specific queries further illustrates the system's strengths. For example, the "Wearing Tie" query demonstrates the system's ability to recognise formal attire. As illustrated in Figure 5, the system successfully retrieved images featuring neckties across diverse environmental contexts, including outdoor settings (trees, buildings) and indoor environments (tables, offices). The consistent similarity scores (0.24-0.29) demonstrate robust feature extraction despite significant background variations.

Figure 5. Top retrieval results for" Wearing Tie" query demonstrating robustness to background variations (Precision@6: 0.833).

The high precision@6 score (0.833) indicates robust distinction between tie-wearing and non-tie images, with only one result being a false positive in the top 6. Notably, this system was not susceptible to position variance (standing vs. sitting) or environmental clutter, indicating that the CLIP visual encoder can capture the relevant areas of clothing without off-task side distractors. That complex Multi-Attribute Query Processing can be performed.

The query "In Orange color Shirt sitting on a Tree" characterises a complex multi-attribute search case that involves recognising colour, the type of clothing, and the environment. Figure 6 shows how the system can help integrate various semantic dimensions.

Figure 6. Retrieval results for complex multi-attribute query "In Orange colour Shirt sitting on Tree" (Pre- cision@6: 1.000).

Manual assessment showed that the system successfully integrated colour, clothing recognition, and environmental context

recognition, and 3/4 of the top results matched the query description perfectly. At the same time, the 4th image had a tree-based context but lacked the orange colour specification. This relevance rate of 80% (precision@4 = 0.80) indicates a high level of expression in the compositional semantics of the system, resulting from the joint caption-embeddings architecture. The ideal precision@4 suggests that the sixth image retrieved during the retrieval process met the query requirement, demonstrating the superiority of a dual-pathway processing strategy.

The Query "I am with Kids" demonstrates the system's high level of comprehension of familial and social relations. The system successfully identified the variety of adult-child interactions, as shown in Figure 7.

Figure 7. Results for" I am with Kids" query demonstrating sophisticated relationship recognition across diverse contexts (Precision@6: 0.833).

The system achieved a precision@6 of 83.3%, correctly recognising 6 out of 6 results as representing meaningful adult-child interactions. Such high performance can be explained by the following main strengths: good contextual knowledge, with proper identification of various contexts like classroom, playgrounds, and home settings; ability to pick up the diversity in interactions such as teaching, play, and caregiving, and age discrimination, which provides clear differentiation of children and adults belonging to various ages. In addition, there was a very high clustering of the similarity scores between 0.28 and 0.29, showing robust feature representation.

The Winning Moments question shows how the system responds to higher-level activity and event-based questions. People celebrating are always shown in the images of trophy and/or medal holders (figure 8). Many other things were strongly indicative of 'to win' as well. This includes visual markers, trophies, medals, group poses, and more. These have a similarity score of 0.25-0.26, indicating stable cross-recognition.

Figure 8. "Winning Moments" is a top result for the query, aligning closely with trophies and medals as visual symbols of success.

Still, this result is not very satisfactory for this query. The numerical results showed a low recall and F1 value. On the other hand, a qualitative inspection suggests that the system is likely associating the prompt with the presence of

visible trophies and medals. The model may not generalise winning as an event perfectly, but it does lock onto some concrete, specific, symbolic objects which we usually regard as winning.

Analysis of the "Study Time" Question. The query "Study Time" shows that the system understands activity-focused questions as it focuses on strong indicators. According to Figure 9, most of the retrieved images show books, notebooks, or other study materials on a table or on an individual. The results' similarity scores ranged from 0.25 to 0.28 in some cases, indicating consistency with the study context.

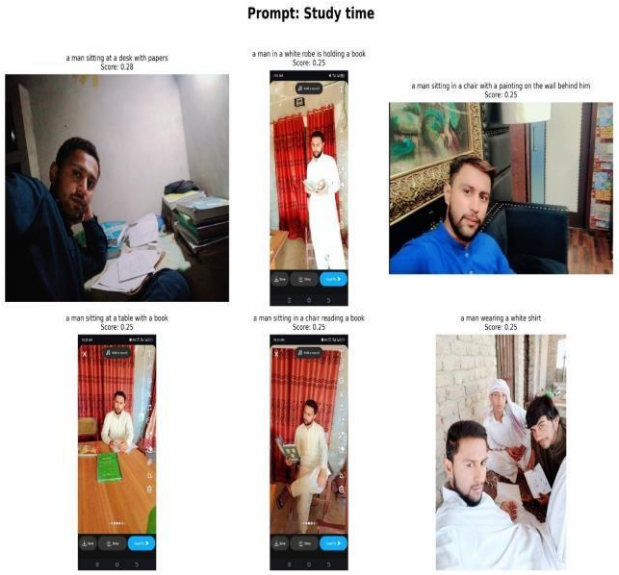

Figure 9. "Study Time," another top retrieval query result, showing high alignment with books and study materials as visual markers of the activity.

Even though the quantitative assessment showed low recall for this query, qualitative inspection indicates that the system is still capable of capturing the gist of the activity. The use of visual indicators, such as books and writing materials, provides evidence of this. These signify strongly for the broader conception of 'studying'; even when the whole abstract meaning of which is not represented completely.

In general, the qualitative results reaffirmed what we saw in the numbers. This system works best when the query relates to specific brand attributes, such as clothes or colours. The system can produce beneficial results when the query involves relations such as 'is with' or 'kids'. When the prompts are related to activities or events, the system is not perfect yet. However, it has shown promise in correctly interpreting details such as trophies or books. This means that users can successfully find their personal photos using natural language, even in more abstract or complex queries.

CONCLUSION

This paper presented a hybrid multimodal architecture integrating CLIP, BLIP, and FAISS into a privacy-preserving semantic search system for personal photographs. The use of a dual-pathway approach, with processing that includes both direct visual semantic alignment and text-mediated understanding, demonstrated good performance on concrete attribute-based and relationship-based queries, with complete or near-complete precision in most cases, even with plausible distractors. However, with event-based queries, we began to see limitations in abstract semantic reasoning, although cosine similarity scores were highly consistent across all query types, suggesting solid embedding quality and stable retrieval capabilities.

The proposed design works entirely on-device, with no reliance on cloud-based infrastructure, supporting both user privacy and processing efficiency. The research also developed a new evaluation framework comprising quantitative similarity and precision values, as well as qualitative analyses of errors, to capture the system's utilisation in real-world retrieval tasks and to help establish a benchmark for assessing personal photo search. We further narrowed the semantic gap between human memory recall and automated retrieval by enabling natural language interaction with personal digital archives. Future work will include: developing event-

level semantic understanding, detecting and characterising bias in the vision–language model, and optimising the dual-pathway multimodal pipelines for deployment on resource-constrained devices. By combining efficiency, accuracy, and privacy, the current study contributes to the practical viability of an efficient photo search system. It sets expectations for a human-centred multimodal retrieval system.

REFERENCE

[1] Photutorial. (2025). Photo statistics: How many photos are taken every day? Retrieved August 4, 2025, from https://photutorial.com/photos-statistics/

[2] Rodden, K., & Wood, K. R. (2003). How do people manage their digital photographs? In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 409–416). Association for Computing Machinery. https://doi.org/10.1145/642611.642682

[3] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12), 1349–1380. https://doi.org/10.1109/34.895972

[4] Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys, 40(2), 1–60. https://doi.org/10.1145/1348246.1348248

[5] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., & Yanker, P. (1995). Query by image and video content: The QBIC system. Computer, 28(9), 23–32. https://doi.org/10.1109/2.410146

[6] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR).

[8] Li, J., Li, D., Savarese, S., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the 39th International Conference on Machine Learning (ICML) (pp. 12763–12780). PMLR.

[9] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML) (pp. 8748–8763). PMLR.

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (NIPS 2017) (pp. 5998–6008).

[11] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[12] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50–58. https://doi.org/10.1145/1721654.1721672

[13] Chen, J., & Ran, X. (2019). Deep learning with edge computing: A review. Proceedings of the IEEE, 107(8), 1655–1674. https://doi.org/10.1109/JPROC.2019.2921977

[14] Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2018). Model compression and acceleration for deep neural networks: [15] The principles, progress, and challenges. IEEE Signal Processing Magazine, 35(1), 126–138. https://doi.org/10.1109/MSP.2017.2765695

[16] Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. IEEE Transactions on Knowledge and Data Engineering, 33(7), 2836–2849. https://doi:10.1109/TBDATA.2019.2921572

[17] Radlinski, F., & Craswell, N. (2017). A theoretical framework for conversational search. Proceedings of the 2017 Conference on Conference on Human Information Interaction and Retrieval (CHIIR' 17), 117–126. https://doi.org/10.1145/3020165.3020183

[18] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., & Young, M. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems 28 (NIPS 2015), 2503–2511.

[19] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on

Computer Vision and Pattern Recognition, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[20] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25 (NIPS 2012), 1097–1105.

[21] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML) (pp. 4904–4916). PMLR.

[22] Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., & Yan, J. (2022). Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208. https://doi.org/10.48550/arXiv.2110.05208

[23] Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., & Jégou, H. (2024). The FAISS library. arXiv. https://arxiv.org/abs/2401.08281

[24] Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Lv, M., Wang, L., Wu, J., Yu, J., Zhang, L., Zhou, C., & Zuo, W. (2021). Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432.

[25] Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y., & Gao, J. (2022). RegionCLIP: Region-based language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 16793–16803).

[26] Jégou, H., Douze, M., & Schmid, C. (2011). Product quantisation for nearest neighbor search. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(1), 117–128. https://doi.org/10.1109/TPAMI.2010.57

[27] Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(4), 824–836. https://doi.org/10.1109/TPAMI.2018.288947Pinecone Systems Inc. (2025). Pinecone vector database. https://www.pinecone.io

[28] Wang, C., He, X., Guo, R., Li, M., Wang, X., Yan, J., Li, X., Pan, J., & Lin, Z. (2021). Milvus: A purpose-built vector data management system. In Proceedings of the 2021 International Conference on Management of Data (SIGMOD' 21) (pp. 2614–2627). https://doi.org/10.1145/3448016.3457550

[29] Weaviate B.V. (2025). Weaviate open source vector database. https://weaviate.io

[30] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. Advances in Neural Information Processing Systems (NeurIPS), *35*, 1–54. https://doi.org/10.48550/arXiv.2204.14198

[31] Anil, A. (2015, May 28). Picture this: A new way to share, search and store your photos. The Keyword | Google. https://blog.google/products/photos/picture-this-fresh-approach-to-photos/

[32] Apple Inc. (2024, May). Apple Platform Security. https://support.apple.com/guide/security/welcome/web

[33] Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., & Hoi, S. (2023). InstructBLIP: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500. https://doi.org/10.48550/arXiv.2305.06500

[34] Pustu-Iren, K., Bruns, G., & Ewerth, R. (2021). A multimodal approach for semantic patent image retrieval. In Proceedings of the 2nd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2021) (Vol. 2909, pp. CEUR Workshop Proceedings. https://doi.org/10.15488/16876

[35] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning (ICML), *139*, 8748–8763. https://doi.org/10.48550/arXiv.2103.00020

[36] Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., & Duan, N. (2023). Visual ChatGPT: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 https://doi.org/10.48550/arXiv.2303.04671

[37] Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task. JAIR, 47, 853-899. https://doi.org/10.1613/jair.3994

[38] Clark, A., & Contributors. (2025). Pillow (Version 9.0.0) [Computer software]. Python Software Foundation. https://pypi.org/project/Pillow/

[39] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. arXiv. https://arxiv.org/abs/1912.01703

[40] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. M. (2019). HuggingFace's Transformers: State-of-the-art natural language processing. arXiv. https://doi.org/10.48550/arXiv.1910.03771

[41] Kaggle Inc. (2025). Kaggle: Your machine learning and data science community. https://www.kaggle.com

[42] Zauner, C. (2010). Implementation and benchmarking of perceptual image hash functions [Master's thesis, University of Applied Sciences Hagenberg].

[43] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

[44] Amigó, E., Gonzalo, J., Artiles, J. et al. A comparison of extrinsic clustering evaluation metrics based on formal constraints. Inf Retrieval 12, 461–486 (2009). https://doi.org/10.1007/s10791-008-9066-8

[45] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 1–15. http://proceedings.mlr.press/v81/buolamwini18a.html

[46] Cavoukian, A. (2010). Privacy by design: the definitive workshop. A foreword by Ann Cavoukian, Ph.D. Identity in the Information Society, 3, 247–251. https://doi.org/10.1007/s12394-010-0062-y

[47] Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 107–113. https://doi.org/10.1145/1327452.1327492

[48] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. arXiv. https://arxiv.org/abs/1803.09010

[49] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

[50] ISO. (2014). ISO 19115-1:2014 — Geographic information — Metadata — Part 1: Fundamentals. International Organization for Standardization.

[51] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

[52] Markham, A., & Buchanan, E. (2012). Ethical decision-making and internet research: Recommendations from the AoIR ethics working committee (Version 2.0). Association of Internet Researchers. https://aoir.org/reports/ethics2.pdf

[53] Peng, R. D. (2011). Reproducible research in computational science. Science, 334(6060), 1226–1227. https://doi.org/10.1126/science.1213847