Journal on Information Technologies & Communication DOI: 10.31276/ICT.2026.0058

An intermediate architecture for secure multimodal search over public web sources and internal scanned documents in Vietnamese organisations

Thanh Cong Bui*, Chinh Dong Xuan, Van Nguyen, Anh Nguyen Van, Phieu Le Van

Institute of Research 486, 1 Alley 805, Nguyen Hoang Ton Street,

Xuan Dinh Ward, Hanoi, Vietnam

Received 5 July 2026; revised 30 September 2026; accepted 3 October 2026

Abstract:

Organisations need search over both public web sources and internal repositories, including scanned portable document format (PDF) files that lack native text and must be retrieved under authorisation constraints. This paper specifies an intermediate architecture for secure multimodal search in Vietnamese organisations. The architecture provides a required integration layer between conventional enterprise keyword search and research-grade multimodal retrieval, rather than introducing a new encoder or retrieval model. The contract has four joints: unified public and internal sources, dual text-page representation of scanned archives, access control as a pre-ranking bitset constraint, and auditable hybrid fusion. Elasticsearch is the execution core, not the contribution. Encoder, optical character recognition (OCR), and late-interaction models remain replaceable. The architecture also defines an evaluation protocol covering Precision@k, nDCG@k, MRR, retrieval latency, OCR coverage, and authorisation completeness. As an architecture-focused study, the paper establishes these evaluation dimensions rather than reporting completed experimental results. The proposed design provides a structured basis for implementing secure multimodal retrieval with component replaceability, traceability, and access-control enforcement across heterogeneous information sources.

Keywords: **Elasticsearch, enterprise search, intermediate architecture, multimodal retrieval, scanned PDF.**

*Corresponding author: Email: congthanhttmt@gmail.com

1. Introduction

Modern organisational information retrieval transcends standard internet web crawling [1, 2]. Institutions now require unified knowledge extraction from a bipartite data ecosystem. The first tier comprises public assets, including official portals, news platforms, and openly accessible documents. The second tier, critical for restricted operations, encompasses proprietary data: secure file servers, document management systems (DMS), and legacy historical archives. Bridging these two distinct environments into a cohesive search experience poses significant architectural challenges [3].

The fundamental divergence between public web search and organisational retrieval lies in access governance [4]. While commercial enterprise platforms support role-based access control (RBAC), they often treat semantic retrieval and security as disjointed processes [5, 6]. Integrating granular, query-time security filtering with deep semantic vector search remains architecturally complex and computationally expensive [7]. A robust system must seamlessly integrate access control within the retrieval pipeline-balancing the trade-offs between early-bound indexing and late-bound filtering - to maintain strict auditability and mitigate information leakage without sacrificing retrieval latency [8].

A pervasive bottleneck globally - particularly within government agencies and legacy institutions in developing nations such as Vietnam - is the heavy reliance on scanned, non-searchable PDF archives [9]. Historically significant files, ratified decisions, and official correspondence frequently exist solely as images. While integrating OCR is a standard approach, relying exclusively on OCR-extracted text is fundamentally flawed for administrative documents [10]. Organisational records frequently embed crucial visual semantics - such as official red stamps, handwritten marginalia, complex data tables, and spatial layouts - which are completely lost or heavily distorted in pure text translation, especially given the complexities of Vietnamese diacritics and low-resolution scans [11].

Consequently, traditional lexical or purely text-based semantic search is insufficient [12, 13]. Recent work on multimodal interaction for semantic web platforms [14] shows the value of non-textual user input at the interface layer. That line of research, however, does not address access-controlled retrieval over organisational repositories. Deploying multimodal search inside restricted environments therefore requires a different

architectural layer. As users increasingly query by abstract concepts, images, or voice, systems must comprehend both textual and visual contexts. Cross-modal retrieval is therefore a necessity [15]. A robust retrieval mechanism must be multimodal, directly mapping the spatial and visual structures of scanned documents to semantic user intents to bypass the limitations of error-prone, text-only pipelines [16-18].

This paper proposes an intermediate architecture for secure multimodal search in Vietnamese organisations. The contribution is not a new encoder or late-interaction ranker. It is a mid-level architectural contract that sits between keyword-oriented enterprise search and research-grade multimodal retrievers. The contract specifies four joints: unified public and internal sources, dual text-page representations of scanned archives, access control as a retrieval primitive rather than an application filter, and transparent hybrid fusion on a production lexical-vector core. Domain constraints of Vietnamese administrative collections - diacritics, low-resolution scans, official stamps, and heterogeneous ACLs - motivate this layer; they are not treated as already-solved empirical results. The paper also provides a gap analysis and an evaluation protocol for a later implementation study.

Intermediate is therefore a placement claim, not a maturity claim. The layer exists because an organisation cannot jump from a BM25 enterprise index to a late-interaction multimodal retriever without losing exact identifiers, query-time authorisation, auditability, and bounded latency. Encoder, OCR, and late-interaction models remain replaceable plugs behind the contract; the four joints do not. If that mid-layer is not stated, the same design is misread either as an integration of existing tools or as an empirical retrieval paper that lacks results.

2. Related work and research gap

2.1. Popular information search frameworks

Keyword-based systems utilise inverted indices and scoring algorithms like BM25 [19]. While computationally efficient and optimal for exact matching (e.g., Elasticsearch, Solr), purely lexical approaches are vulnerable to vocabulary mismatches, OCR noise, and ambiguous queries. Although modern sparse retrieval and late-interaction models have emerged to mitigate some of these lexical limitations, they remain fundamentally text-centric.

Enterprise search systems build upon classical retrieval mechanisms [20] to support organisational environments, extending them with internal data connectors, access control, audit logging, and document-level permission checks [4] - capabilities that are mandatory for organisational deployment. However, conventional solutions remain largely text-centric, often relying on late-bound post-filtering which introduces candidate exhaustion and timing leakage risks. Furthermore, they typically treat scanned document processing, multimodal inputs, and semantic matching as supplementary modules rather than core capabilities.

To address lexical limitations, semantic search employs dense vector embeddings to capture underlying intent. General-purpose models such as Sentence-BERT [6] enable semantic similarity matching in English, while language-specific models like PhoBERT [10] adapt this paradigm for Vietnamese, and multilingual retrieval models such as BGE-M3 [21] extend coverage to cross-lingual settings. Nevertheless, pure semantic retrieval struggles with exact alphanumeric identifiers and exhibits lower interpretability compared to keyword matching.

OCR-based document search has advanced through models such as TrOCR [9] and LayoutLMv2 [11], which enable text and layout extraction from scanned images. End-to-end approaches like Donut [17] eliminate explicit OCR dependency entirely, while architectures like DocFormer [18] deeply fuse visual and spatial features. However, these models focus on recognition accuracy and information extraction rather than retrieval pipeline integration or access governance.

Expanding beyond text, recent architectures leverage models like CLIP [15] for joint text-image representations, ColPali [22] for preserving spatial layouts in raw document images typically lost in OCR, and Whisper [23] for voice-activated querying. Deploying such systems requires complex orchestration of heterogeneous channels, hybrid ranking algorithms, and stringent security protocols.

Within restricted organisational contexts, architecting a comprehensive search platform necessitates the fusion of diverse retrieval paradigms - ranging from standard full-text matching and OCR-augmented indexing to access-aware semantic retrieval. Table 1 provides a comparative analysis of these prevailing frameworks, delineating their

respective strengths, inherent limitations, and specific contributions to the hybrid search architecture proposed in this research.

Table 1. Comparison of popular information search frameworks.

Framework	Strengths	Limitations in organisations	Implication for this work
Public web search	Very large scale, link analysis, fast ranking	Cannot index private internal files; does not enforce local access policies	Useful as inspiration, but not sufficient for internal search
Keyword enterprise search	Mature indexing, exact search, metadata filters	Weak for semantic mismatch, image queries, noisy OCR and scanned PDFs	Needed as the stable lexical core
Semantic search	Retrieves paraphrases and related meanings	May miss exact codes; needs embeddings and vector index; lacks native ACL support	Added to improve recall and user intent matching
OCR-based document search	Makes scanned PDFs searchable	Recognition errors; layout and tables remain difficult	Must be integrated with confidence and page-level evidence
Multimodal retrieval	Supports image and page visual evidence	Harder evaluation; higher computational cost	Required for images, diagrams, page scans and image queries
Access-aware search	Protects internal documents	Requires identity, ACL, audit and secure filtering	Mandatory for organisational deployment

2.2. Research gap

Building upon the analysis of prevailing literature, this study identifies three critical research gaps:

First, existing retrieval frameworks are predominantly designed for either open-domain web corpora [19] or closed organisational repositories [20, 4], but not both simultaneously. The challenges of orchestrating open-domain web data alongside highly restricted, proprietary corpora - which operate under fundamentally different update lifecycles, metadata schemas, and governance constraints - within a unified indexing and ranking framework remain largely unaddressed in the literature.

Second, while OCR and document understanding have been studied extensively as standalone tasks [9, 11, 17, 18], this body of work focuses on recognition accuracy and

document comprehension rather than retrieval integration or access governance. Current systems often treat scanned PDF processing as an optional pre-processing step, extracting flat text while discarding critical visual semantics such as institutional stamps, spatial layouts, and handwritten signatures [9]. Consequently, there is a distinct lack of frameworks designed to natively index and rank multimodal representations of legacy scanned documents within secure organisational search architectures.

Third, a fundamental disconnect exists between advancements in multimodal retrieval and the rigid security constraints of enterprise systems. While access control in organisational information systems has been well studied [4] and semantic vector retrieval has advanced significantly [6, 21], multimodal retrieval is often not treated as a first-class component of secure enterprise architectures. Conversely, state-of-the-art research in cross-modal semantic retrieval - including joint text-image models [15] and page-level visual retrievers [22], and voice-activated querying [23] - typically assumes open-domain environments with unrestricted access. The architectural integration of granular, query-time security filtering with high-dimensional multimodal vector search remains insufficiently addressed.

Therefore, despite significant advances in lexical, semantic, multimodal, and enterprise search technologies, there remains a lack of an intermediate architecture that simultaneously specifies source unification, OCR-aware dual representation, multimodal indexing, and access-aware retrieval within restricted organisational environments.

To bridge these gaps, this study specifies an intermediate architecture rather than a new retrieval model or a mere aggregation of existing tools. The missing piece is the mid-level contract that connects heterogeneous sources, multimodal evidence channels, and authorisation as a first-class retrieval constraint. Encoder, OCR, and late-interaction components remain replaceable modules behind that contract.

3. Problem formulation and scientific basis

3.1. Problem formulation

Let $D = D_{\text{pub}} \cup D_{\text{int}}$ be a heterogeneous collection, where D_{pub} contains public web data and public files, while D_{int} contains internal files, scanned PDF archives, images, and documents stored in organisational repositories. Each document $d \in D$ has content

features, source metadata, and an access policy $A(d)$. A user $u \in U$ submits a query q through text, voice, or image. For voice-activated queries, an automatic speech recognition (ASR) module at the client interface immediately transcribes the spoken input into a textual format prior to retrieval. The system must return a ranked list $R(q,u)$ containing only documents for which u satisfies $A(d)$.

The retrieval problem is therefore not simply to maximise relevance. It must maximise relevance under permission constraints. In other words, the system must find documents that are both relevant and authorised. This distinguishes the proposed framework from public search systems and makes authentication, authorisation, metadata, and auditing part of the retrieval model.

Formally, the scoring function gates relevance with an access-control indicator:

$$\text{Score}(q,d,u) = \text{rel}(q,d) \cdot \mathbb{I}(\text{permitted}(u,d)) \quad (1)$$

Based on this gated score, the final retrieval and ranking function is defined as:

$$R(q,u) = \text{rank}(d \in D \mid \text{Score}(q,d,u) > 0) \quad (2)$$

where $\text{rel}(q,d)$ is the continuous relevance score estimated by lexical, semantic, visual, and metadata evidence; $\mathbb{I}(.)$ is an indicator function that returns 1 if the access condition is met and 0 otherwise; and $\text{permitted}(u,d)$ is determined from user identity, roles, groups, source policies, and document-level access-control lists.

To avoid the waste and leakage risk of sequential post-filtering, the architecture uses early-binding in the information-retrieval sense [24]. Document access-control lists (ACLs) are materialised as filter bitsets at index time (Table 3). At query time, lexical and vector retrieval intersect their candidate sets with the bitset for user u before relevance scores are computed. This is compatible with Lucene/Elasticsearch document-level security: unauthorised documents are excluded from the candidate set rather than removed after ranking. The paper does not claim a cryptographic isolation layer or a document graph.

3.2. Hybrid multimodal scoring

The scientific hypothesis of the proposed framework is that hybrid multimodal scoring is more robust than single-channel retrieval on Vietnamese organisational

collections. Lexical evidence is reliable for exact names, numbers, codes, and official terms. Dense semantic evidence improves retrieval when the query and document use different wording. An OCR pipeline makes scanned PDFs lexically and semantically searchable; it is not an extra scoring channel. Visual evidence supports image queries and visually rich documents. Metadata evidence supports source, date, file type, and authority signals.

Each evidence channel produces a score. Scores must be normalised before fusion because BM25, cosine similarity and metadata features have different ranges. A linear fusion model is transparent and easy to tune. When a production system uses Elasticsearch, an alternative is to retrieve lexical and vector candidates separately and combine ranked lists using reciprocal rank fusion or another rank-fusion method.

The composite final score $rel(q,d)$ utilises a linear combination approach, which has become the standard methodology in modern hybrid retrieval systems to fuse disparate sparse and dense signals [25]:

$$rel(q,d) = w_1 \cdot S_{\text{lex}}(q,d) + w_2 \cdot S_{\text{sem}}(q,d) + w_3 \cdot S_{\text{vis}}(q,d) + w_4 \cdot S_{\text{meta}}(q,d) \quad (3)$$

where w_1, w_2, w_3, and w_4 are the respective weights tuned by validation queries or administrative policy. To prevent double-counting, OCR output has no independent scoring weight; it populates the lexical and semantic feature spaces only. If a channel is absent for a document or query (no page image, no OCR text, or a text-only query), its weight is set to zero and the remaining weights are renormalised to sum to one.

For voice-activated queries, ASR transforms the audio input into text representations, which are subsequently routed to the textual scoring pipelines. The individual semantic textual and visual scores are computed via cosine similarity between their respective dense embeddings:

$$S_{\text{sem}}(q,d) = \cos(E_{\text{text}}(q_{\text{vec}}), E_{\text{text}}(d_{\text{text}})) \quad (4)$$

$$S_{\text{vis}}(q,d) = \cos(E_{\text{vis}}(q_{\text{vec}}), E_{\text{vis}}(d_{\text{vis}})) \quad (5)$$

These evidence channels and their specific roles within the multimodal retrieval pipeline are summarised in Table 2.

Table 2. Evidence channels and their roles in the proposed framework.

Evidence channel	Main model or method	Data handled	Purpose
Lexical	Inverted index, BM25	Public pages, internal text, OCR text	Exact and explainable search
Semantic text	Sentence or multilingual embeddings	Chunks, titles, summaries, OCR passages	Paraphrase and intent matching
Visual	CLIP/page-image embeddings	Images, figures, PDF page renderings	Image-to-document and visually rich page retrieval
Metadata	Deterministic heuristics	Source, timestamps, document type	Authority and filtering signals

4. Secure multimodal search framework

The proposed framework contains four layers: data sources, ingestion and security gate, processing and indexing, and authenticated search service. It is designed for organisations that need to combine public data collection with internal document search. The overall design is shown in Fig. 1.

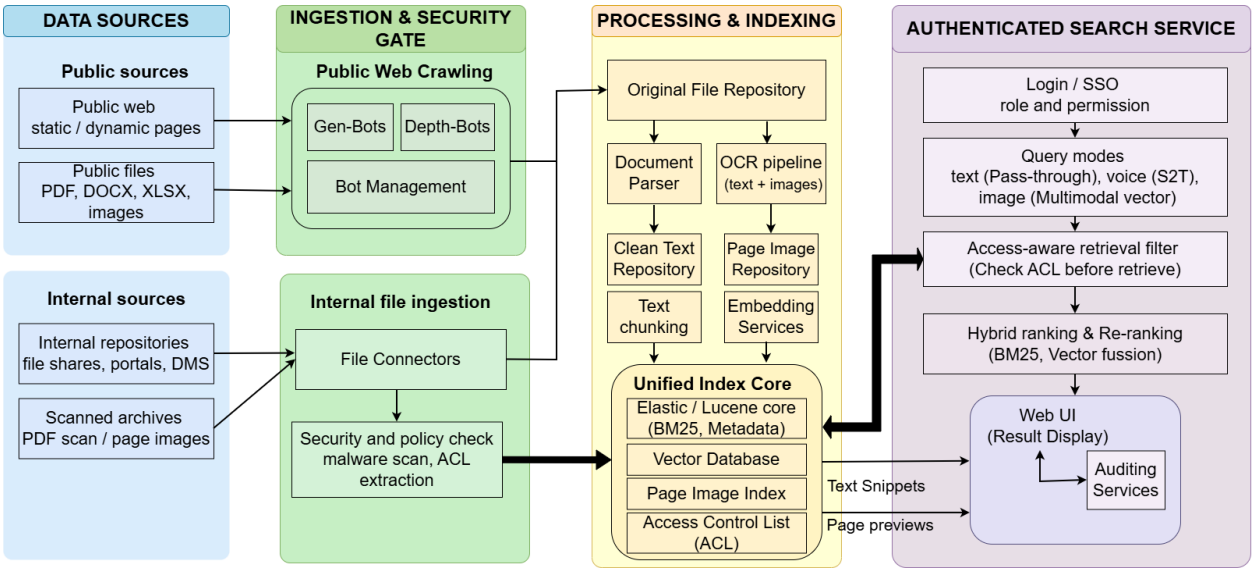


Fig. 1. Proposed framework for public sources and internal scanned documents.

4.1. Data sources

To manage heterogeneous organisational information, the proposed framework categorises data sources into two primary groups.

The first group comprises public data sources, including official websites, open PDFs, images, and dynamic web pages. These resources are collected through a controlled

crawling infrastructure. Specifically, Gen-Bots extract content and metadata from static HTML/XML pages, while Depth-Bots render client-side scripts to capture dynamic content. A centralised manager coordinates seed URLs, domain policies, crawling schedules, politeness constraints, deduplication, and crawl logs.

The second group encompasses internal repositories, such as shared file systems, internal portals, document management systems (DMS), manually uploaded files, and database-backed storage. Due to their proprietary nature, these assets are accessed using specialised connectors rather than standard web crawlers. These connectors extract and preserve critical native metadata, including file hierarchies, document ownership, timestamps, classification labels, and access-control policies. Crucially, the extraction of these access-control policies at the source level forms the foundational prerequisite for the early-bound security mechanisms executed during the retrieval phase.

Within this internal group, scanned PDF archives represent a critical subset of organisational records, frequently containing official forms, stamps, and signatures. The framework applies a dual representation strategy to these documents: OCR-extracted text is stored for keyword and semantic matching, while rendered page images are retained to preserve visual layouts. This dual representation keeps both textual content and page layout available for retrieval. Localised visual cues that matter in Vietnamese administrative files - official red stamps, signatures, and handwritten notes - remain in the page-image channel rather than being discarded after OCR. Concrete detector or OCR-model choices are left to the implementation protocol in Section 5.

4.2. Ingestion, authentication, and access control

For public data, the security gate checks file types, content safety, source reliability, and duplication. For internal data, the security gate additionally extracts and normalises access policies. Each indexed item is associated with source-level and item-level permissions. The access policy may be derived from directory permissions, document management metadata, manual classification, or organisational role definitions.

User-facing search is not anonymous. Users must log in through a local account, single sign-on (SSO) service, or token-based mechanism. The authenticated identity is mapped to roles, groups, units, and permission scopes. Crucially, rather than relying on

vulnerable late-bound post-filtering, the system enforces early-bound pre-filtering during retrieval. As formalised in Eqs. (1)-(2) and Algorithm 1, the permitted bitset of user u is applied during candidate generation, before scores, titles, snippets, thumbnails, or cached summaries are produced. This is early-binding relative to ranking, not a claim of index-time cryptographic isolation.

Auditing is also part of the framework. The system records query time, user identity, query type, returned document identifiers, clicked documents, and access-denied events. Audit logs help administrators detect abnormal search behaviour and support accountability in internal environments. The overall workflow of authenticated and access-aware multimodal query processing is illustrated in Fig. 2.

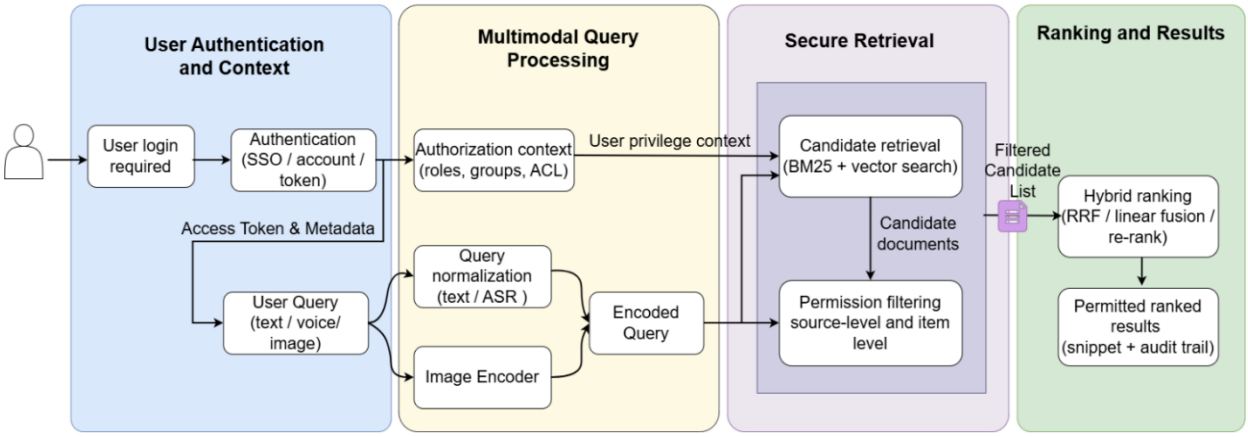


Fig. 2. Authenticated and access-aware multimodal query processing.

4.3. Processing and indexing core

The processing layer transforms raw data into structured, indexable units. HTML pages are stripped of boilerplate and scripts. Digital office documents are parsed for text, tables, and metadata. Crucially, text-based PDFs are parsed directly, while scanned PDFs are rendered into images and processed via OCR. Page renderings are retained so that layout-sensitive cues (stamps, signatures, forms) can be matched in the visual channel without requiring a separate stamp detector in this architecture paper. Embedded images are extracted alongside their contextual captions.

To enhance semantic retrieval precision, documents are segmented into granular chunks (e.g., paragraphs, PDF pages, or OCR blocks). This ensures that highly relevant,

localised passages are matched effectively within multi-topic documents. These chunks are subsequently mapped to lexical, dense vector, metadata, and access-control fields.

For practical deployment, Elasticsearch serves as an execution environment, supporting near-real-time ingestion, BM25 lexical matching, and hybrid vector search [26, 27]. Elasticsearch is the execution core, not the scientific contribution. The contribution is the intermediate contract - source adapters, dual scanned-page representation, ACL bitsets, and auditable fusion - placed on top of that engine. To formalise this design, Table 3 delineates the recommended index schema for both public and internal search.

Table 3. Recommended index fields for public and internal document search.

Field group	Examples	Used for	Security note
Source metadata	URL, repository, file path, crawl time, hash	Traceability, freshness, deduplication	Internal paths shown only to permitted users
Text fields	Title, body, OCR text, captions	BM25 and snippets	Snippets are generated only from documents inside the permitted bitset
Vector fields	Text chunk embedding, image embedding, page embedding	Semantic and visual retrieval	Vector hits inherit the parent document ACL bitset and are excluded before ranking
Document structure	Page number, section, table, image index	Passage-level result and highlighting	No page preview without permission
Access fields	Owner, group, role, classification, ACL	Authorisation filtering	Mandatory for internal data; translated to bitsets prior to indexing

4.4. Query modes and ranking

Queries are preprocessed according to their modality. Text inputs undergo Unicode normalisation, tokenisation, spelling correction, and expansion. Voice queries are transcribed via automatic speech recognition (ASR) before standard text processing, while image queries are converted into visual embeddings for vector matching. For scanned PDFs, queries are evaluated against both extracted OCR text and visual page representations.

The ranking layer synthesises candidates by fusing the top-k results retrieved independently from lexical (BM25), dense semantic, and visual indices. Although

advanced late-interaction models or cross-encoder re-ranking can be integrated if sufficient interaction logs exist, the framework defaults to a transparent, context-constrained hybrid rank fusion. Linear fusion or reciprocal rank fusion (RRF) is the default because it is interpretable, tunable, and auditable. Cross-encoders and other late-interaction models are optional plugs and evaluation baselines (Table 5), not required components of the intermediate architecture. The complete secure multimodal query-processing workflow is summarised in Algorithm 1.

Algorithm 1. Secure multimodal query processing.

Input: Query q, User u, Index I, Access Bitsets B

Output: Ranked list R of permitted documents with snippets

1: P(u) ← Authenticate(u) // obtain roles, groups, permission scope
2: IF authentication fails THEN Log event & RETURN Error
3: m ← Modality(q)
4: IF m = voice THEN
5: q_norm ← Normalize(ASR(q))
6: q_vec ← TextEncoder(q_norm)
7: ELSE IF m = text THEN
8: q_norm ← Normalize(q)
9: q_vec ← TextEncoder(q_norm)
10: ELSE //Image query
11: q_norm ← null
12: q_vec ← ImageEncoder(q)
13: // Intersect ACL bitsets before scoring. BM25 returns ∅ if q_norm is null.
14: C_permitted ← BM25(q_norm, I_lex, P_u, B) ∪ VectorSearch(q_vec, I_sem, I_vis, P_u, B)
15: // Score normalisation & context-constrained hybrid rank fusion
16: FOR each d ∈ C_permitted DO
17: $S(d) \leftarrow w_1 \cdot s_lex(d) + w_2 \cdot s_sem(d) + w_3 \cdot s_vis(d) + w_4 \cdot s_meta(d)$
18: R ← SortDescending(C_permitted, S)
19: Log(u, q, m, results=|R|, denied-probes, timestamp)
20: RETURN GenerateSnippets(R)

5. Evaluation protocol

Because this paper specifies an architecture rather than a completed experiment, Section 5 defines a protocol for later empirical work on two collections: public Vietnamese web sources and internal scanned documents. The public part tests crawling and open search. The internal part tests scanned-PDF retrieval, access control, and authenticated search. The evaluation design is illustrated in Fig. 3.

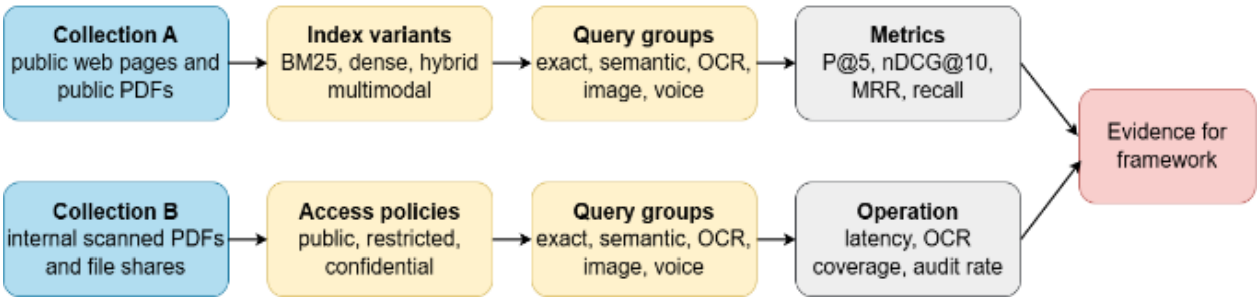


Fig. 3. Evaluation protocol for public and internal Vietnamese collections.

5.1. Dataset and query groups

The experimental dataset comprises public web pages, digital office documents, and legacy scanned PDFs. Each document is annotated with metadata and explicit ACLs. To systematically evaluate each architectural module, the query set is categorised into lexical, semantic, OCR-dependent, visual, and voice queries.

Table 4. Example evaluation protocol configuration.

Data group	Example content	Main technical issue	Evaluation focus
Public web	Static pages, dynamic pages, public PDFs	Crawling, cleaning, duplicate handling	Crawl success, precision, freshness
Internal files	DOCX, XLSX, PDF text, uploaded reports	Connector, metadata, ACL mapping	Permission filtering and metadata search
Scanned PDFs	Signed documents, old reports, archived forms	OCR, page rendering, Vietnamese diacritics	OCR coverage, recall, page-level retrieval
Images	Embedded figures, photos, page thumbnails	Image embedding and caption linking	Image-to-document retrieval
Voice queries	Spoken Vietnamese search requests	ASR, normalisation, spelling correction	Voice-to-result accuracy and latency

To reflect realistic organisational scenarios, internal queries involve the cross-modal retrieval of localised visual semantics. For security validation, identical queries are executed under varying user permission scopes. The same queries are run under different roles so that later experiments can measure authorisation completeness and can report latency and candidate-set size by role. Those measurements test leakage risk; they do not by themselves prove that every side-channel is closed.

Furthermore, the evaluation protocol establishes comparative baselines—encompassing purely lexical (BM25), dense semantic, and late-interaction cross-encoder models. This benchmarking is designed to quantify the trade-offs among retrieval accuracy, query latency, and architectural security in restricted environments. The overall configuration of this experimental protocol is summarised in Table 4.

5.2. Baselines and metrics

The proposed system should be compared with several baselines: keyword-only search, semantic-only search, OCR keyword search, hybrid text search, and the full multimodal access-aware framework. The evaluation must report both retrieval quality and operational quality. Retrieval metrics include Precision@k, Recall@k, MRR, and nDCG@k (where $k \in \{5, 10, 20\}$). Operational metrics include latency, index size, OCR coverage, crawler success rate, and access-control correctness. Access-control correctness is measured by verifying that unauthorised documents do not appear in results, snippets, thumbnails, or previews. The baseline systems used for validating the proposed framework are summarised in Table 5.

Table 5. Baselines for validating the framework.

Method	Index used	Supported queries	Expected limitation
B1: Keyword only	BM25/inverted index	Text query	Weak semantic recall; cannot use image query
B2: OCR keyword	BM25 over OCR text	Text query for scanned PDFs	Sensitive to OCR errors
B3: Dense semantic	Text vector index	Text query	May miss exact identifiers
B4: Cross-encoder (SoTA)	Bi-encoder + Cross-encoder reranker	Text query	Prohibitive computational latency; lacks native visual integration
B5: Full proposed	BM25 + dense + visual + metadata + ACL	Text, voice, image, scanned PDF	Requires rigorous weight tuning and higher initial indexing complexity

5.3. Ablation strategy

An ablation study is structured to systematically isolate and quantify the contribution of each architectural component. For instance, bypassing the dense retrieval module evaluates the system's dependency on semantic formulations, while removing the BM25 component assesses its precision degradation on exact administrative codes.

To ensure credible empirical validation, human relevance judgements are required. Assessors must label query-document pairs utilising a predefined four-tier scale (not relevant, partially relevant, relevant, and highly relevant). For proprietary documents, assessors evaluate data strictly within their authorised clearance, or utilise cryptographically anonymised metadata. The anticipated outcomes of the ablation protocol are detailed in Table 6.

Table 6. Ablation plan for scientific validation.

Setting	Removed component	Affected query group	Evidence expected
A1	Dense semantic retrieval	Semantic/paraphrase queries	Lower recall and nDCG@10
A2	BM25 retrieval	Exact document number/name queries	Lower Precision@5
A3	OCR text pipeline	Scanned PDF queries	Reduced text-based recall for legacy archives.
A4	Visual embedding	Image and page-layout queries	Lower image-to-document accuracy
A5	ACL filtering	Internal search	Simulated offline to quantify timing leakage and unauthorised candidate exhaustion.

6. Discussion

6.1. Scientific contribution

The paper contributes an intermediate architecture for organisational retrieval, with four concrete joints. First, public web sources and restricted internal archives, including scanned PDFs, share one indexing and ranking contract. Second, scoring combines lexical, semantic, visual, and metadata evidence, while OCR is a text pipeline rather than an additional relevance term. Third, authorisation is applied as a pre-ranking bitset constraint,

so snippets and previews are generated only for permitted documents. Fourth, an evaluation protocol is given so that a later implementation can measure each joint instead of treating the design as already validated. Section 6.3 orders that implementation along the same four joints.

The framework also clarifies the role of Elasticsearch. Elasticsearch is suitable as a core retrieval infrastructure because it provides mature full-text indexing, BM25 ranking, structured filtering, aggregations, near-real-time indexing, and vector search support. However, Elasticsearch alone does not solve the scientific problem of Vietnamese scanned PDF retrieval, multimodal query processing, or access-aware fusion. These functions must be designed around the core and validated experimentally.

6.2. Practical benefits

For public data, the system helps organisations build a domain-specific search index from selected websites instead of depending entirely on commercial search engines. For internal data, the system helps users find reports, decisions, scanned documents, images, and file-based knowledge that are otherwise hidden in folders and archives. The ability to search scanned PDFs is particularly valuable because it recovers access to historical documents that currently exist only as images or localised visual features like stamps.

For security management, authenticated search reduces the risk of information leakage. A user receives only results that match both the query and the permission policy. Audit logs provide evidence of search behaviour and can be integrated with organisational monitoring. This is different from public search, where the main question is relevance, not internal permission.

6.3. Limitations and future work

The framework still has limitations. OCR quality may be low for degraded scans, handwritten annotations, skewed pages, and dense tables. Linear fusion still requires validation-time weight calibration; Section 3.2 only specifies proportional renormalisation when a modality is missing, not learned adaptive weights. Image retrieval may produce visually similar but semantically unrelated results if captions and surrounding text are

ignored. Access-control mapping can also be difficult when internal repositories use inconsistent permission structures.

Future work should focus on building a Vietnamese benchmark for internal and scanned-document retrieval, fine-tuning Vietnamese embedding models on organisational terminology, improving page-level visual retrieval, and adding retrieval-augmented question answering while preserving source citation and access control. Additionally, future empirical studies should benchmark the proposed transparent rank fusion against computationally heavy late-interaction models to formally quantify the trade-offs between retrieval accuracy, query latency, and architectural security. A production deployment should also evaluate scalability, update latency, fault tolerance, and security against adversarial queries.

Further development should follow the intermediate contract rather than begin with a new encoder. The first increment is a production lexical core with index-time ACL bitsets and snippet generation only on the permitted set. The second increment adds the scanned-page pipeline: OCR text into lexical and semantic fields, plus retained page images. The third increment adds visual/page embeddings and the default auditable fusion (normalised linear scores or reciprocal rank fusion), including proportional weight renormalisation when a modality is missing. Only after those joints are in place should a prototype be measured with the protocol in Section 5- Precision@k, nDCG@k, MRR, authorisation completeness, and latency by role - and only then should late-interaction or Vietnamese-specific models be plugged in as optional baselines. Learned adaptive weights, retrieval-augmented question answering, and a full cryptographic threat model remain later extensions, not conditions for the architecture to stand.

6.4. Deployment architecture and use of Elasticsearch

For real-world integration, translating this theoretical framework into a deployable prototype requires organising it as a set of decoupled microservices. The crawler service collects public web data under a predefined domain policy. The connector service ingests internal files from approved folders or document management systems. The document processing service performs file parsing, OCR, page rendering, image extraction, chunking, summarisation, and embedding generation. The indexing service writes lexical fields, vector fields, metadata, and access-control fields into the retrieval core. The search

API receives authenticated requests from the web interface and coordinates candidate retrieval, permission filtering, rank fusion, snippet generation, and audit logging.

Elasticsearch is suitable for the retrieval core because it supports full-text search, metadata filters, near-real-time indexing, and production monitoring. In this framework, Elasticsearch can store one index for document-level metadata and another index for chunk-level passages. A scanned PDF can be represented by multiple page chunks, each containing OCR text, page number, OCR confidence, page-image identifier, and access policy inherited from the parent document. Vector fields can be used for semantic text retrieval and image/page retrieval, while BM25 remains available for exact term matching.

A practical hybrid search strategy is to retrieve top-k results from BM25, top-k from dense vectors, and top-k from visual vectors, then merge them by reciprocal rank fusion or by a normalised linear scoring function. This design is useful because each retrieval channel has a different strength. BM25 is reliable for identifiers and exact names, dense retrieval improves recall for semantic queries, and visual retrieval helps when the query is an image or when the target page contains important visual layout. Ranking is computed only on the bitset-permitted candidate set; unauthorised documents never enter the fused list or the snippet stage.

6.5. Security, governance, and operational requirements

Internal search systems must be governed more strictly than public search applications. The system should define clear source registration procedures, including who is allowed to add a new public website, who approves an internal repository connector, and how often each source is re-indexed. Every document should have a stable identifier, a hash value for change detection, and a record of when and how it entered the system. For scanned documents, the system should retain links between the original page image, the OCR output, and the indexed chunk so that users can verify the evidence.

Access control must hold for retrieval, display, preview, download, and logging. The security scope of this architecture paper is limited. Protected assets are document text, snippets, thumbnails, vectors, and audit records. The in-scope adversary is an authenticated user who should not learn unauthorised content, including through titles, snippets, or previews. The intended properties are authorisation completeness and the

absence of post-filter ranking. A full threat model covering embedding inversion, OCR poisoning, encrypted search, and all timing side-channels is outside this design paper; Section 5 only requires those risks to be measured. Audit logs should be append-only and should record access-denied events because repeated denials may indicate probing.

Operational evaluation should include not only search relevance but also service-level indicators. These include average and percentile latency, queue length for OCR processing, failed OCR pages, index update delay, storage growth, number of indexed internal documents, and number of access-control mismatches found during testing. These indicators are important because an internal search engine is a long-running information infrastructure, not only an experimental retrieval model.

7. Conclusions

This paper proposed an intermediate architecture for secure multimodal search in Vietnamese organisations that must use both public web sources and internal file repositories. The architecture treats scanned PDF archives as a first-class source, keeping page images alongside OCR text. It also treats login, authorisation, pre-ranking access filtering, and audit logging as parts of retrieval rather than as an external application shell.

The architecture combines controlled crawling, internal connectors, OCR, text and image embeddings, lexical and vector indexes, hybrid ranking, and a web interface. Elasticsearch/Lucene is a practical core for full-text search, filtering, and hybrid retrieval; OCR, embedding, visual, and speech modules remain replaceable. The evaluation protocol and ablation plan define how a later prototype should be measured. Future work should implement that prototype, build a Vietnamese cross-modal collection, and report results by query group. Those experiments should also compare transparent fusion with late-interaction rankers on accuracy, latency, and authorisation-error rates—not treat this paper as that empirical comparison.

CRediT author statement

Thanh Cong Bui: Conceptualisation, Methodology, Supervision, Writing - Review and Editing; Chinh Dong Xuan: Methodology, Software, Validation; Van Nguyen: Conceptualisation, Literature Review, Writing - Original Draft Preparation; Anh Nguyen Van: Investigation, Validation, Writing - Review and Editing, Phieu Le Van: Supervision, Project Administration, Writing - Review & Editing.

ACKNOWLEDGEMENTS

The authors would like to thank the research group members who contributed to the design of the initial prototype consistent with this intermediate architecture.

COMPETING INTERESTS

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] W.X. Zhao, J. Liu, W. Wu, et al. (2023), "Dense text retrieval based on pre-trained language models: A survey", *ACM Computing Surveys*, **55(13s)**, pp.1-39, DOI: 10.1145/3590138.

[2] J. Lin, R. Nogueira, A. Yates (2021), "Pretrained transformers for text ranking: BERT and beyond", *Synthesis Lectures on Human Language Technologies*, **14(4)**, pp.1-325, DOI: 10.2200/S01123ED1V01Y202108HLT053.

[3] J. Wang, X. Yi, R. Guo, et al. (2021), "Milvus: A purpose-built vector data management system", *Proc. SIGMOD*, pp.2614-2627, DOI: 10.1145/3448016.3457550.

[4] U. Kruschwitz, C. Hull (2017), "Searching the enterprise", *Foundations and Trends in Information Retrieval*, **11(1)**, pp.1-142, DOI: 10.1561/1500000053.

[5] O. Khattab, M. Zaharia (2020), "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT", *Proc. SIGIR*, pp.39-48, DOI: 10.1145/3397271.3401075.

[6] N. Reimers, I. Gurevych (2019), "Sentence-BERT: Sentence embeddings using Siamese BERT-networks", *Proc. EMNLP-IJCNLP*, pp.3982-3992, DOI: 10.18653/v1/D19-1410.

[7] J. Johnson, M. Douze, H. Jégou (2021), "Billion-scale similarity search with GPUs", *IEEE Transactions on Big Data*, **7(3)**, pp.535-547, DOI: 10.1109/TBDATA.2019.2921572.

[8] R. Bost, B. Minaud, O. Ohrimenko, et al. (2017), "Forward and backward private searchable symmetric encryption", *Proc. CCS*, pp. 1465-1482, DOI: 10.1145/3133956.3134014.

[9] M. Li, T. Lv, J. Chen, et al. (2023), "TrOCR: Transformer-based optical character recognition with pre-trained models", *Proc. AAAI*, **37(11)**, pp.13094-13102, DOI: 10.1609/aaai.v37i11.26538.

[10] D.Q. Nguyen, A.T. Nguyen (2020), "PhoBERT: Pre-trained language models for Vietnamese", *Findings of EMNLP*, pp.1037-1042, DOI: 10.18653/v1/2020.findings-emnlp.92.

[11] Y. Xu, M. Li, L. Cui, et al. (2021), "LayoutLMv2: Multi-modal pre-training for visually-rich document understanding", *Proc. ACL*, pp.2579-2591, DOI: 10.18653/v1/2021.acl-long.201.

[12] L. Phan, H. Tran, T. Nguyen, et al. (2022), "ViT5: Pretrained text-to-text transformer for Vietnamese language generation", *Proc. NAACL-HLT*, pp.136-147, DOI: 10.18653/v1/2022.naacl-main.10.

[13] V. Karpukhin, B. Oguz, S. Min, et al. (2020), "Dense passage retrieval for open-domain question answering", *Proc. EMNLP*, pp.6769-6781, DOI: 10.18653/v1/2020.emnlp-main.550.

[14] X. Zhang, G. Flumeri, A. Vozzi, et al. (2026), "Multimodal assessment of mental states and visual search for a user-centred design of semantic web platforms", *Applied Sciences*, **16(10)**, DOI: 10.3390/app16104756.

[15] A. Radford, J.W. Kim, C. Hallacy, et al. (2021), "Learning transferable visual models from natural language supervision", *Proc. ICML*, pp.8748-8763, DOI: 10.48550/arXiv.2103.00020.

[16] Y. Huang, T. Lv, L. Cui, et al. (2022), "LayoutLMv3: Pre-training for document AI with unified text and image masking", *Proc. ACM Multimedia*, pp.4083-4091, DOI: 10.1145/3503161.3548112.

[17] G. Kim, T. Hong, M. Yim, et al. (2022), "OCR-free document understanding multimodal transformer", *Proc. ECCV*, pp.498-517, DOI: 10.1007/978-3-031-19815-1_29.

[18] S. Appalaraju, B. Jasani, B.U. Singh, et al. (2021), "DocFormer: End-to-end transformer for document understanding", *Proc. ICCV*, pp.993-1003, DOI: 10.1109/ICCV48922.2021.00104.

[19] S. Robertson, H. Zaragoza (2009), "The probabilistic relevance framework: BM25 and beyond", *Foundations and Trends in Information Retrieval*, **3(4)**, pp.333-389, DOI: 10.1561/1500000019.

[20] C.D. Manning, P. Raghavan, H. Schütze (2008), *Introduction to Information Retrieval*, Cambridge University Press, 506pp, DOI: 10.1017/CBO9780511809071.

[21] J. Chen, S. Xiao, P. Zhang, et al. (2024), "BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation", *arXiv preprint*, DOI: 10.48550/arXiv.2402.03216.

[22] M. Faysse, H. Sibille, T. Wu, et al. (2025), "ColPali: Efficient document retrieval with vision language models", *Proc. ICLR*, DOI: 10.48550/arXiv.2407.01449.

[23] A. Radford, J.W. Kim, T. Xu, et al. (2023), "Robust speech recognition via large-scale weak supervision", *Proc. ICML*, **202**, pp.28492-28518.

[24] C. Büttcher, C.L.A. Clarke, G.V. Cormack (2010), *Information Retrieval: Implementing and Evaluating Search Engines*, MIT Press, 632pp.

[25] Y. Luan, J. Eisenstein, K. Toutanova, et al. (2021), "Sparse, dense, and attentional representations for text retrieval", *Transactions of The Association for Computational Linguistics*, **9**, pp.329-345, DOI: 10.1162/tacl_a_00369.

[26] Elastic (2026), "Hybrid search", *Elasticsearch Documentation*, https://www.elastic.co/docs/solutions/search/hybrid-search, accessed 28 September 2026.

[27] Elastic (2026), "Reciprocal rank fusion", *Elasticsearch Reference*, https://www.elastic.co/docs/reference/elasticsearch/rest-apis/reciprocal-rank-fusion, accessed 28 September 2026.