Towards Semantic Searching in Diverse Multimodal Collections

Václav Kučera, Martin Bulín, Jan Švec, Pavel Ircing
Department of Cybernetics @ Faculty of Applied Sciences @ University of West Bohemia in Pilsen
Univerzitní 8, 301 00 Pilsen, Czech Republic
vaclavk@students.zcu.cz, {bulinm, honzas, ircing}@kky.zcu.cz

Abstract

Digital humanities projects increasingly rely on heterogeneous collections of multimodal data, including video testimonies, scanned documents, and photographs. Despite the growing availability of such archives, researchers face challenges in efficiently locating relevant content due to the diversity of formats and the lack of unified retrieval methods. In this work, we present a general framework for semantic search over collections of multiple modalities. The framework integrates specific parsers and transforms all inputs into textual representations leveraging services like automatic speech recognition (ASR), optical character recognition (OCR), and generative-AI-based image captioning. Text is subsequently segmented into overlapping chunks, indexed in a vector database, and enriched through an automatic question generation (AQ) pipeline to create ground-truth queries for evaluation. We evaluate the framework on a constructed dataset derived from Holocaust-related archives, comparing two retrieval strategies (pure vector search vs. hybrid semantic-lexical search) under two chunking scenarios. Results demonstrate that hybrid search consistently outperforms vector-only retrieval, achieving high recall across modalities, and that semantic search is feasible even with diverse and noisy input sources. This framework provides a robust foundation for exploring complex multimodal archives, facilitating access to content that would otherwise remain difficult to discover.

Keywords: multimodal data, semantic search, digital humanities, vector retrieval

1. Introduction

Large multimodal archives encompassing diverse data types began to be amassed on a significant scale during the 1990s, when recording technologies and storage capacities became sufficiently affordable for institutions to preserve substantial volumes of material.

However, the challenge of efficiently retrieving relevant information from these extensive corpora emerged almost immediately thereafter. One prominent example is the USC Shoah Foundation's Visual History Archive (USC Shoah Foundation), which preserves authentic testimonies from Holocaust survivors and witnesses. This collection comprises approximately 52,000 interviews conducted between 1994 and 1999, totaling over 115,000 hours of video material recorded in 32 languages. The sheer volume of this material renders the identification of relevant content exceedingly challenging.

This challenge prompted the MALACH project (2001–2007), which sought to enhance access to these oral histories by advancing automatic speech recognition (ASR) and information retrieval (IR) technologies.

The initial approach in the MALACH project treated ASR and IR as independent tasks: audio was transcribed into text using state-of-the-art ASR systems, segmented into documents, and subjected to standard document-oriented IR. This strategy rapidly revealed significant limitations. Besides the poor performance of the ASR systems (roughly 40% Word-Error-Rate – WER – across languages), the IR systems had is own issues stemming from oversimplified designs, notably fixed-length sliding windows for segmenting continuous transcripts into pseudo-documents, bypassing the more complex task of topical coherence detection. Evaluated in CLEF campaigns in 2005–2007 using detailed topics specifying user information needs, these systems achieved dismal mean Generalized Average Precision (mGAP) scores (Pecina et al., 2007), attributable to both ASR errors and inadequate segmentation. Standard bag-of-words methods failed to leverage distinctions between relevant and non-relevant material mentioned in the search topic specifications. Ultimately, this era yielded disjointed ASR-IR pipelines with poor results and no user-friendly graphical interface for non-experts.

In the second "epoch" of research, we redefined the paradigm for ASR and IR system design to better align with the practical demands of searching continuous speech transcripts. Recognizing the absence of discrete documents in automatically transcribed streams, we shifted focus to identifying precise replay points—specific timestamps marking the onset of topic-relevant discussion—enabling direct playback of corresponding video segments. Departing from prior document-oriented IR systems, which relied solely on lexical overlap without semantic processing, we adopted a spoken term detection (STD) paradigm. In STD, queries comprise single words or short phrases submitted against a fixed collection, inverting the traditional keyword search model.

This transition also fostered tighter integration be-

Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026, pages 12–19
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tween ASR and IR components. STD indexes incorporated not only the highest-probability ASR transcriptions but also competing hypotheses weighted by their estimated probabilities, enhancing detection robustness. Numerous ML methods were employed for STD over the years, details about the latest incarnation using the Transformer architecture can be found in (Švec et al., 2023). In this era we have also developed several iteratively improved versions of the graphical user interface (GUI), empowering non-expert users to submit queries and instantly replay pertinent testimony segments.

The latest set of techniques — named the "Asking Questions" (AQ) framework shifts to proactive, generative content enrichment. STD enabled efficient pinpointing of exact-term matches via integrated ASR-IR indexing but remained limited to user-initiated lexical queries, yielding timestamps for replay without semantic expansion or contextual guidance. In contrast, AQ generates contextually grounded, time-aligned question-answer pairs directly from transcripts, filtered for semantic coherence, to create navigable "open-set topics" that supplement lengthy monologues. This transforms passive listening into interactive exploration, anticipating user needs rather than reacting to explicit terms, while preserving testimony authenticity; it yields sparse, high-quality questions (one every 2 minutes post-filtering), outperforming STD in facilitating thematic discovery across unstructured speech (Bulin et al., 2025).

Advances in state-of-the-art optical character recognition (OCR) algorithms and large language models (LLMs) now enable the integration of previously overlooked data sources within these collections, including scanned textual documents and photographs. The unification of such diverse modalities within a single retrieval framework constitutes the primary contribution of this paper.

1.1. Related Work

Work on semantic retrieval has evolved from dense neural models for open-domain question answering, which replace keyword matching with learned vector representations of text (Karpukhin et al., 2020), to more fine-grained interaction mechanisms that improve semantic matching within purely textual collections (Khattab and Zaharia, 2020). Subsequent research extended retrieval beyond text by learning shared embedding spaces for images and language (Radford et al., 2021), and more recently by training multimodal large language models to act as universal retrievers across mixed text–image inputs (Lin et al., 2024). While these approaches advance semantic and cross-modal search, they typically assume relatively clean data and jointly trained embedding models. In contrast, our framework is designed for arbitrary settings: it is capable of processing heterogeneous and potentially noisy archival materials (e.g., video testimonies, scanned documents, photographs) by converting all modalities into textual form and combining semantic vector search with lexical matching, thereby prioritizing transparency, robustness, and practical applicability, for instance, in real-world cultural heritage collections.

2. Evaluation Data and Methodology

The proposed framework is designed to be domain-independent and applicable to heterogeneous multimodal collections. For demonstration purposes, we constructed an evaluation dataset grounded in the Holocaust domain. All experiments were conducted in English; however, the framework itself is language-agnostic.

We selected 17 publicly accessible testimony recordings from the public part of the *USC Shoah Foundation Dataset* (USC Shoah Foundation), each approximately 2.5 hours in length. The recordings were processed using our automatic speech recognition (ASR) engine, producing time-aligned transcripts. This constituted the first modality: video testimonies represented as textual segments.

During the interviews, witnesses frequently present photographs or documents to the camera. These were automatically extracted, resulting in 299 images. To further diversify the dataset, we selected 108 scanned historical documents from the *Arolsen Archives* (Arolsen Archives). Hence, in total, we obtained 407 images constituting the second modality.

All images were processed using parsers described in Sec. 3: Optical Character Recognition (OCR) was applied to extract textual content, and Large Language Model (LLM) captioning was used to generate semantic descriptions. After this step, all modalities were transformed into textual form that was subsequently segmented according to the chunking strategy described in Sec. 3.2. In total, we obtained 2,420 chunks serving as retrieval units.

2.1. Automatic Question Generation

To enable scalable evaluation without manual annotation, we employed the Asking-Questions (AQ) framework (Švec et al., 2024). For each out of the original 2,420 chunks, the framework generates multiple semantically grounded questions (approximately three per chunk) and may internally create sub-chunks aligned with each generated query, as shown in Table 1.

This process resulted in 7,249 query – sub-chunk pairs. Each generated query is considered relevant to its corresponding (sub-)chunk as well as to the

original chunk	*One notable development is the organic connection which now exists between the SS and the Police. In 1956, when he was @ppointed Chief of the German Police, HIMMLUR was enabled to effect that fusion between the two forces he controlled which is & marked feature of Germany's internal security organisation today, fs J All senior police officers and many of the junior officers a@re also members of the SS, holding rank in both organisations. The official policy is to recruit new menbers of the police force solely from the SS. 'hus the integration of the state security organisation (the police) and the party security organisation (the SS) is, for all practical purposes, complete.*
sub-chunk 1	*One notable development is the organic connection which now exists between the SS and the Police. In 1956, when he was @ppointed Chief of the German Police, HIMMLUR was enabled to effect that fusion between the two forces he controlled which is & marked feature of Germany's internal security organisation today, fs J All senior police officers and many of the junior officers a@re also members of the SS, holding rank in both organisations. The official policy is to recruit new menbers of the police force solely from the SS.*
query 1	`When was HIMMLUR appointed Chief of the German Police?`
sub-chunk 2	*The official policy is to recruit new menbers of the police force solely from the SS. 'hus the integration of the state security organisation (the police) and the party security organisation (the SS) is, for all practical purposes, complete.*
query 2	`According to the official policy, from what organization were new police force members recruited?`
sub-chunk 3	*'hus the integration of the state security organisation (the police) and the party security organisation (the SS) is, for all practical purposes, complete.*
query 3	`What organizations had effectively merged?`

Table 1: Example output of the AQ framework on a selected ASR result sample, illustrating the generation of evaluation queries and the extraction of relevant sub-chunks (used in scenario B).

original chunk, forming the ground truth for a known-item retrieval task.

2.1.1. Retrieval Setup

Chunks were indexed in a vector database using the proposed framework described in Sec. 3. For each generated query, we computed its embedding and performed top-k retrieval.

We evaluated two scenarios:

(A) *Original chunk indexing:* Only the 2,420 original chunks were indexed. The AQ framework was used solely to generate evaluation queries (three per original chunk on average), which are all considered to semantically correspond to the same original chunk.

(B) *Sub-chunk indexing:* 7,249 AQ-generated sub-chunks were indexed, resulting in a one-to-one mapping between queries and indexed items.

Scenario A better reflects realistic deployment, where question generation is not part of the production pipeline.

2.2. Evaluation Metrics

We report Recall@k (for $k \in \{1, 3, 5, 10\}$) and Mean Reciprocal Rank (MRR@10).

Recall@k measures whether at least one relevant item appears among the top-k retrieved results and so can be interpreted as the probability of finding the correct segment within the first k returned results.

$$\text{Recall@}k = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}\{\text{rank}_q \leq k\}, \qquad (1)$$

where rank_q denotes the rank of the first relevant item for query q.

Mean Reciprocal Rank (MRR@10) evaluates how highly the first relevant result is ranked, considering only the top 10 retrieved items. If no relevant item appears within the top 10 results, the reciprocal rank is defined as zero.

$$\text{MRR@10} = \frac{1}{|Q|} \sum_{q \in Q} \begin{cases} \frac{1}{\text{rank}_q}, & \text{if } \text{rank}_q \leq 10, \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

2.3. Metrics Limitations

Although each generated query is guaranteed to be relevant to its originating chunk, it may also be semantically relevant to other chunks in the collection. Since no exhaustive manual relevance annotation was performed, such additional relevant matches remain undetected. Consequently, the reported metrics may slightly underestimate the true semantic retrieval performance.

3. Multimodal Search Framework

The concept of the proposed framework, illustrated in Fig. 1, is designed to provide a unified and extensible system for retrieving information from heterogeneous collections of data. Users interact with the system through a user-friendly web-based interface, allowing both conventional textual queries and exploration of the indexed multimodal content.

3.1. General Data Parser

The foundation of the proposed multimodal search framework is a modular data parsing pipeline designed to transform heterogeneous source files into a unified semantic representation. This pipeline, implemented in Python, uses a routing mechanism based on file extensions to dispatch documents to specialized parsers. Each parser is responsible for extracting textual information and, where applicable, spatial or temporal metadata, ensuring that the semantic context is preserved across different

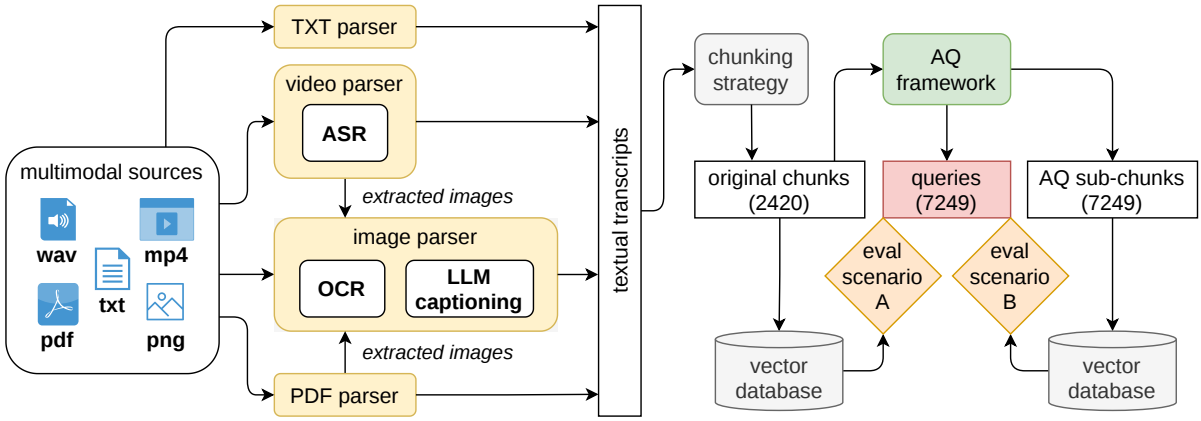

Figure 1: Overall concept and evaluation of the presented multimodal search framework.

modalities. The PDF parser (Section 3.1.3) is included in the pipeline described here to support future applications, but it is not evaluated on the dataset considered in this study.

3.1.1. Audio Recordings

Audio files, typically in WAV or MP3 formats, are processed through an Automated Speech Recognition (ASR) module. The system utilizes the *UWebASR* API (Lehečka et al., 2023) to perform transcription of spoken content. The resulting transcript is processed according to the chunking strategy described in Section 3.2. Each resulting segment preserves its temporal boundaries (start and end timestamps), enabling precise localization within the audio stream during retrieval. This enables the search engine to pinpoint the exact segment within the audio file during retrieval.

3.1.2. Images

Images are processed using a method that captures both their literal details and semantic content.

- *Optical Character Recognition (OCR):* We use *Tesseract OCR* (Smith, 2007) to extract any textual information present within the image. This is particularly crucial for scanned documents, infographics, or slides. Extracted text blocks are further processed using the unified chunking strategy described in Section 3.2. Each chunk retains its associated bounding box coordinates, enabling spatial localization during retrieval.
- *Semantic Captioning:* To capture the visual content of the image, we leverage a Large Language Model (LLM), specifically OpenAI's GPT-4o (OpenAI, 2024). The model generates a comprehensive textual description of the image content and identifies key objects along with their relative positions (top, center, bottom, etc.). The generation is guided by the constraints defined in the chunking strategy (Section 3.2), ensuring that each description forms a single indexable unit compatible with the embedding model.

3.1.3. PDF Documents

PDF documents are handled by a structure-aware parser based on the *PyMuPDF* library (Artifex Software, Inc., 2026). The parser decomposes the document into text blocks and images.

- *Text Blocks:* Extracted text blocks are processed according to the chunking strategy described in Section 3.2. Structural metadata, including page numbers and bounding boxes, are preserved for each resulting unit. Short, non-informative segments (e.g., page numbers or artifacts) are filtered out to maintain the quality of the index.
- *Embedded Images:* Images embedded within the PDF are extracted and processed through the image parsing pipeline described in Section 3.1.2. This ensures that diagrams, charts, and illustrations within a document are fully indexed both by their textual content (via OCR) and their visual semantics (via captioning).

3.1.4. Plain Text

Plain text files are parsed in a way that preserves their structural organization. Logical paragraphs are detected using line breaks and subsequently processed using the chunking strategy described in Section 3.2. The parser maintains the absolute line numbers associated with each resulting chunk, allowing direct referencing to the original source.

3.1.5. Video Recordings

Video files are treated as complex multimodal data requiring both temporal and visual analysis.

- *Audio Transcription:* The audio stream is extracted using *FFmpeg* (FFmpeg Developers, 2026)and processed through the ASR module, identical to the standalone audio parser described in Section 3.1.1.

- *Visual Frame Analysis:* Visual content is obtained by extracting keyframes at a specified sampling rate (selected to be 0.1 frame per second). Each frame is then processed by the OpenAI GPT-4o model to generate semantic descriptions, as described in Section 3.1.2. This approach enables the system to index video content based on both spoken information and visual information over time.

3.2. Chunking Strategy

To ensure compatibility with the limited context window of embedding models, the pipeline employs a unified token-aware chunking strategy across all modalities. All textual data are segmented using a sliding window approach with a fixed length of 256 tokens and an overlap of 32 tokens. The tokenizer of the underlying embedding model, *all-MiniLM-L6-v2* (Reimers and Gurevych, 2019), is used to guarantee that each chunk remains within the model's context window.

This strategy is applied across all modalities while respecting their logical structure:

- *Plain Text and OCR:* Chunks are created within logical boundaries such as paragraphs or OCR blocks to maintain semantic coherence.

- *Audio and Video:* Transcripts are segmented temporally, where each textual chunk preserves its corresponding start and end timestamps.

- *Semantic Captions:* For images and video frames, the chunk size acts as a constraint for the generative model, ensuring that descriptions are concise and ready for indexing as single units.

This multi-layered approach preserves necessary context and metadata (spatial and temporal) while providing the level of detail required for precise semantic retrieval.

3.3. Retrieval with Semantic Vectors

The core of our retrieval system is based on dense vector representations. We use the *all-MiniLM-L6-v2* sentence-transformer model (Reimers and Gurevych, 2019), which maps text chunks into a 384-dimensional dense vector space. This model provides a favorable trade-off between embedding quality and computational efficiency, making it suitable for large-scale semantic search.

For retrieval, we explore two approaches: a purely vector-based search using the Hierarchical Navigable Small World (HNSW) algorithm (Malkov and Yashunin, 2018) for efficient nearest-neighbor search, and a hybrid strategy that combines dense embeddings with traditional lexical features based on TF–IDF (Spärck Jones, 1972) to also capture exact keyword matches.

(i) *HNSW:* The index is implemented using the `IndexHNSWFlat` class from *Faiss* library (Douze et al., 2024). It is constructed with a connectivity parameter `M`=32 and a construction expansion factor `efConstruction`=200. During retrieval, the search expansion factor is set to `efSearch`=64 to balance efficiency and recall.

(ii) *Hybrid search:* The index combines retrieval scores from the dense HNSW index and a sparse TF-IDF index using a weighted sum:

$$S_{\text{hybrid}} = \alpha \cdot S_{\text{dense}} + (1 - \alpha) \cdot S_{\text{sparse}}, \quad (3)$$

where S_{dense} and S_{sparse} are the scores from the dense and sparse retrievers, respectively, and $\alpha \in [0, 1]$ is a weighting parameter (set to 0.5 in our experiments). Both indices apply L_2 normalization, ensuring comparable cosine similarity scores in the range $[0, 1]$.

The sparse TF–IDF index is constructed using the `TfidfVectorizer` class from the scikit-learn library (Pedregosa et al., 2011), configured with `min_df`=1, `max_df`=0.9, `max_features`=10000, and a list of stemmed English stopwords obtained using the Snowball stemmer provided by the *NLTK* library (Bird et al., 2009).

4. Results

From a broader perspective, three evaluation layers can be identified for the proposed framework: (1) content extraction quality (ASR, OCR, LLM-based captioning), (2) chunking strategy and information representation, and (3) cross-modal semantic retrieval.

In this paper, we focus exclusively on the third layer, i.e., the ability of the system to retrieve the correct multimodal segment given a textual query. It should be noted, however, that retrieval performance is inherently influenced by the quality of upstream processing stages.

We evaluated two retrieval strategies described in Section 3.3: (i) vector-based retrieval using

HNSW indexing and (ii) hybrid search combining semantic vectors with keyword-based matching. Furthermore, we report results for the two evaluation scenarios introduced in Section 2.1.1: (A) original chunk indexing and (B) sub-chunk indexing.

4.1. Scenario A: Original Chunk Indexing

Results for the more realistic setup (7,249 queries vs. 2,420 original chunks) are shown in Table 3.

In this scenario, multiple queries map to the same larger chunk, resulting in a slight decrease in performance compared to Scenario B, where each (shorter) sub-chunk corresponds to a single query. As expected, hybrid search consistently outperforms HNSW across all metrics.

Importantly, Recall@10 remains above 0.60 for video transcripts and above 0.95 for OCR documents in the hybrid configuration. This indicates that even under realistic indexing conditions, the system is capable of retrieving the correct multimodal segment within a small set of top-ranked results.

4.2. Scenario B: Sub-chunk Indexing

Table 2 reports results for the AQ-level indexing setup (7,249 queries vs. 7,249 indexed sub-chunks).

Across all modalities, hybrid search consistently outperforms pure HNSW vector retrieval. The improvement is particularly visible in Recall@1 and MRR, indicating that hybrid search more frequently ranks the correct segment at the very top of the result list. This suggests that combining lexical matching with semantic similarity helps stabilize retrieval when queries are closely aligned with the original wording of the source segment. This behavior is captured in Table 4, which presents the top-5 retrieved chunks for the query. In the table, the desired chunk is highlighted in bold. When using the pure HNSW index, this chunk does not appear even among the top-10 results. In contrast, under the hybrid setup, the required chunk is ranked second.

OCR-based documents achieve the highest scores overall (e.g., Recall@10 above 0.94 and MRR around 0.80 in the hybrid setup). This can likely be attributed to the relatively well-structured and information-dense nature of scanned documents, where generated questions often correspond to explicit factual statements.

In contrast, video transcripts (ASR modality) show lower Recall@1 and MRR. This may reflect the more narrative and less structurally explicit nature of spoken testimonies or finding semantically close passages from another testimonies, as the structure of the interviews is unified. Although the required chunk according to the evaluation protocol is not retrieved, several returned passages indicate clear semantic relevance to the query, as assumed in Section 2.3.

Caption-based image representations provide useful semantic summaries for pictures with unique activities, however, LLM-based descriptions of scanned documents, for instance, can confuse the retrieval system significantly. Therefore, we additionally report an evaluation of the LLM-based captioning parser excluding scanned documents, denoted as *Capt.** in Tables 2 and 3. The results show a clear improvement in performance under this setting.

4.3. Efficiency Considerations

Due to offline preprocessing (ASR transcription, OCR, caption generation, chunking, and indexing), online retrieval is computationally lightweight. On a standard CPU machine, average query latency remains below 500 ms for both retrieval strategies.

The computationally intensive stages are data parsing and index construction, which are performed only once during preprocessing. Parsing the complete dataset of 424 files including `.txt`, `.jpg`, and `.png` files) required 36 minutes in total on a standard CPU machine. Index construction scales linearly with the number of text chunks, with an average build time of approximately 2.25 s per 100 chunks.

5. Conclusion

In this paper, we have presented a general framework for semantic search across heterogeneous multimodal collections. Our evaluation, conducted on a constructed dataset from the Holocaust domain, demonstrates that the system is capable of retrieving relevant segments across multiple modalities, including ASR transcripts of video testimonies, OCR-processed documents, and LLM-generated image captions.

Two retrieval strategies were compared (HNSW vector search and hybrid semantic-lexical search), and two evaluation scenarios were explored (original chunk vs. sub-chunk indexing). The results indicate that hybrid search consistently outperforms pure vector-based retrieval, and that even under realistic indexing conditions, the system retrieves the correct segment within a small set of top-ranked results with high reliability. Overall, the results demonstrate that semantic search over heterogeneous multimodal collections is feasible and reasonably robust, even when different modalities exhibit varying levels of textual quality and structural consistency.

Source	Queries	Chunks	HNSW					Hybrid search				
			k1	k3	k5	k10	MRR	k1	k3	k5	**k10**	MRR
ASR	5700	1530	0.17	0.30	0.36	0.44	0.25	0.32	0.49	0.56	**0.64**	0.42
OCR	716	533	0.60	0.75	0.80	0.87	0.69	0.73	0.88	0.92	**0.96**	0.81
Capt.	833	357	0.19	0.31	0.38	0.47	0.27	0.30	0.43	0.52	**0.61**	0.39
Capt.*	601	249	0.21	0.34	0.42	0.51	0.30	0.32	0.47	0.56	**0.66**	0.41
All	7249	2420	0.19	0.31	0.36	0.44	0.26	0.32	0.48	0.54	**0.62**	0.41

Table 2: Retrieval performance by modality – Scenario A (2420 original chunks)

Source	Queries	Chunks	HNSW					Hybrid search				
			k1	k3	k5	k10	MRR	k1	k3	k5	**k10**	MRR
ASR	5700	5700	0.22	0.36	0.41	0.49	0.30	0.32	0.50	0.57	**0.64**	0.43
OCR	716	716	0.60	0.77	0.82	0.88	0.69	0.71	0.89	0.91	**0.95**	0.80
Capt.	833	833	0.24	0.35	0.42	0.51	0.31	0.34	0.46	0.52	**0.63**	0.42
Capt.*	601	601	0.26	0.38	0.46	0.55	0.34	0.34	0.49	0.55	**0.67**	0.43
All	7249	7249	0.25	0.38	0.44	0.51	0.33	0.34	0.51	0.57	**0.65**	0.44

Table 3: Retrieval performance by modality – Scenario B (7249 chunks from the AQ framework)

5.1. Future Work

Several directions for future work can further enhance and generalize the proposed framework:

- *Evaluation on different domains:* Extending the evaluation beyond Holocaust-related data to other cultural heritage collections or entirely different domains.
- *Cross-modal query capabilities:* Enabling retrieval not only via textual queries but also through queries based on images or audio segments. For example, a user could search for video segments similar to a given sound recording, or find documents semantically related to a sample image.
- *Enhanced evaluation metrics:* Incorporating more sophisticated metrics capturing partial relevance or cross-modal semantic similarity, especially for collections where multiple relevant segments may exist for a single query.

These extensions would further improve the usability and applicability of the framework in digital humanities contexts, providing researchers with flexible and semantically aware access to diverse multimodal archives.

6. Acknowledgements

The work has been supported by the grant of the University of West Bohemia, project No. SGS-2025-011.

7. Bibliography

Arolsen Archives. Arolsen archives: International center on nazi persecution. Online archive of more than 30 million historical documents related to Nazi persecution.

Artifex Software, Inc. 2026. Pymupdf. Python bindings for the MuPDF PDF toolkit. Version X.Y.Z.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Martin Bulin, Jan Svec, Pavel Ircing, Adam Fremund, and Filip Polak. 2025. Exploring oral history archives using state-of-the-art artificial intelligence methods. *Acta Informatica Pragensia*, 14(2):207–214.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

FFmpeg Developers. 2026. Ffmpeg multimedia framework. Version 7.1 (or latest version).

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Wen-tau Yih, Yann Dauphin, and Wilker Yala. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*.

Rank	Score	Top-5 chunks for query *What was located approximately twenty miles from Berlin?*
1	0.3644	But worst of all was in Berlin. I didn't get this in Kongsberg. I don't think they were organized. Yet that way Berlin they had
2	**0.3519**	**Is father. Her father was born in a little town about twenty miles from Berlin and known now and and they had properties a lot of proves. They had a so call ia**
3	0.3501	have intellectually known something like that was coming. It was inevitable even even the most assimilated would be at some point hit with this. How far it would go? Of course, nobody knew at that time, but this was the first time it really came home to him. There was his sign his name be smerched by these vermin. And what approximate date was that? Thirty four or five, fairly shortly after the Nazis came to power? In thirty three, it was at that time that he met. And now go on to answer your question why did we leave Konigsberg and went to Berlin?
4	0.3466	Her father was born in a little town about twenty miles from Berlin and known now and and they had properties a lot of proves. They had a so call iagod Rita good. It's a huge farm. There were ten thousand actors of land. We're talking about twenty five thousand acres. They. The father had a lot of comments and industry. He was selling coal for heating in Berlin. The grandfather had a raising cows, horses, pigs and supplying meat to Berlin and a dirty all kind of dirty cheese, milk or whatever. Then they had a factory, they made a food for horses and he used to raise horses for the Wehrmacht.
5	0.3450	That was the reason we moved to Berlin. And can you describe briefly what your home was like in Berlin and what the area that you lived in was like? Well, I lived in Bilmastorf, which is West Berlin. We lived in a typical Berlin four story building and I have a picture of it. We had, I don't know, ten twelve room apartment. There were two apartments to each floor, an elevator landing and very conveniently, the other people who lived on the same floor as we did were also Jewish, which was wonderful.

Table 4: Top-5 retrieved chunks using the *(ii) Hybrid search approach* (see Section 3.3)

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of SIGIR*.

Jan Lehečka, Jan Švec, Josef V. Psutka, and Pavel Ircing. 2023. Transformer-based speech recognition models for oral history archives in english, german, and czech. In *Proc. Interspeech 2023*, pages 201–205.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2024. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*.

Yu. A. Malkov and D. A. Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs.

OpenAI. 2024. GPT-4o. `https://openai.com`. Large multimodal language model.

Pavel Pecina, Petra Hoffmannová, Gareth J. F. Jones, Ying Zhang, and Douglas W. Oard. 2007. Overview of the CLEF-2007 cross-language speech retrieval track. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5152 of *Lecture Notes in Computer Science*, pages 674–686. Springer.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633, Washington, DC, USA. IEEE Computer Society.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Jan Švec, Luboš Šmídl, and Jan Lehečka. 2023. Transformer-based encoder-encoder architecture for spoken term detection. In *Pattern Recognition*, pages 346–357, Cham. Springer Nature Switzerland.

USC Shoah Foundation. Visual history archive. Digital archive of video testimonies and associated metadata, maintained by USC Shoah Foundation.

Jan Švec, Martin Bulín, Adam Frémund, and Filip Polák. 2024. Asking questions framework for oral history archives. In *Advances in Information Retrieval – 46th European Conference on Information Retrieval, ECIR 2024, Proceedings, Part III*, volume 14610 of *Lecture Notes in Computer Science*, pages 167–180. Springer.