Unified Multimodal Interleaved Document Representation for Retrieval

Anonymous ACL submission

Abstract

Information Retrieval (IR) methods aim to identify documents relevant to a query, which have been widely applied in various natural language tasks. However, existing approaches typically consider only the textual content within documents, overlooking the fact that documents can contain multiple modalities, including images and tables. Also, they often segment each long document into multiple discrete passages for embedding, which prevents them from capturing the overall document context and interactions between paragraphs. To address these two challenges, we propose a method that holistically embeds documents interleaved with multiple modalities by leveraging the capability of recent vision-language models that enable the processing and integration of text, images, and tables into a unified format and representation. Moreover, to mitigate the information loss from segmenting documents into passages, instead of representing and retrieving passages individually, we further merge the representations of segmented passages into one single document representation, while we additionally introduce a reranking strategy to decouple and identify the relevant passage within the document if necessary. Then, through extensive experiments on diverse IR scenarios considering both the textual and multimodal queries, we show that our approach substantially outperforms relevant baselines, thanks to the consideration of the multimodal information within documents.

1 Introduction

Information Retrieval (IR) is the task of fetching relevant documents from a large corpus in response to an input query, which plays a critical role in various real-world applications including web search engines and question-answering systems (Shah et al., 2019; Lewis et al., 2020; Guu et al., 2020). Over the years, IR methods have evolved significantly, with approaches broadly categorized into sparse and dense retrieval paradigms. Specifically, sparse retrieval methods (Robertson et al., 1994; Jones, 2004) focus on lexical overlap between queries and documents; meanwhile, dense retrieval methods (Karpukhin et al., 2020; Xiong et al., 2021) utilize neural embeddings to represent queries and documents in a continuous vector space. Note that, recently, dense retrieval methods have gained more popularity over sparse methods due to their capability to capture semantic nuances and context beyond simple keyword matching, leading to multiple successes with improved performance.

Despite their huge successes, existing (dense) retrieval methods face a couple of severe challenges. First, they primarily rely on the textual data for document embedding and retrieval, overlooking the fact that modern documents often contain multimodal content, such as images and tables (beyond the plain text), which can carry critical information that may be essential for accurately understanding and retrieving the relevant documents. To be specific, a diagram within a medical article can more effectively represent the structure of a molecule or the progression of a disease, offering more clarity that would be difficult to achieve with text alone, and omitting such multimodal content can lead to an incomplete understanding (and potentially inaccurate retrieval) of the documents. Also, the segmentation of long documents into discrete passages, which is commonly employed by existing retrieval models to handle the length limitation for embeddings, may prevent models from capturing the full context and the intricate relationships between different parts of the document, ultimately leading to suboptimal retrieval performance. It is worthwhile noting that, concurrent to our work, while there has been recent work that screen captures the document and then embed its screenshots (to consider different modalities in a unified format) (Faysse et al., 2024; Ma et al., 2024), not only its content (such as paragraphs, images, and tables) can be fragmented into different sub-images, lead-

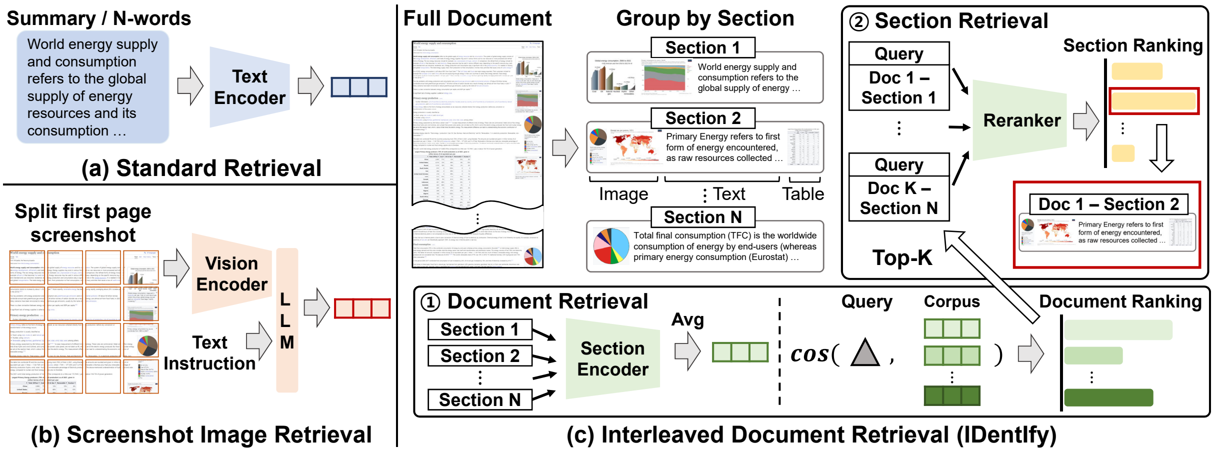


Figure 1: Comparison of different IR approaches. **(a)**: Conventional methods use a small portion of the text within the document for its representation. **(b)**: Recent methods use first-page screenshot images to represent the document. **(c)**: Our approach leverages the full contextual information within documents interleaved with multiple modalities by considering them in their original format, and is further capable of pinpointing relevant sections for the query.

ing to the loss of contextual coherence across the entire document, but also the visual representation of text may hinder the model's ability to capture the semantic relationships present in the original textual data, while increasing the image resolution leads to the concern on the memory requirements.

In this work, we introduce a novel approach to holistically represent documents for IR, which addresses the aforementioned challenges by representing and retrieving the documents interleaved with multiple modalities in a unified manner (See Figure 1). Specifically, our method revolves around the recent advance of Vision-Language Models (VLMs), which enable the processing and integration of multimodal content (such as text, images, and tables) directly into a single token sequence, thereby preserving the context and relationships between various parts of the document, unlike the previous approaches that rely on the fragmented visual representations. Also, in cases where the number of tokens in a document is large and exceeds the capacity of a single context window of VLMs, we propose a strategy to segment the document into passages, each represented within the token limit, and combine these passage embeddings into a unified document representation. This strategy differs from existing IR approaches that independently represent and retrieve at the passage level, potentially losing the overall document context. Lastly, to accurately identify only the relevant sections within the retrieved lengthy document, we introduce a reranking mechanism that is trained to pinpoint the passage most pertinent to the query (among all the other passages within the document), allowing for both the coarse-grained document-level matching and fine-grained passage-level retrieval. We refer to our overall framework as **I**nterleaved **D**ocument **In**formation Retrieval S**y**stem (IDentIfy).

We experimentally validate the effectiveness of IDentIfy on four benchmark datasets, considering both the text-only and multimodal queries. On a battery of tests conducted, we observe that our approach substantially outperforms relevant baselines that consider only the uni-modality for document representations, thanks to the consideration of multimodal content. Further, we find that the strategy to represent the whole document with its single representation (by merging embeddings of its splits) is superior to the approach of individually representing them for document retrieval, but also performing reranking over the sections of the retrieved document is superior to the approach of directly retrieving those sections, which confirm the efficacy of the proposed retrieval and reranking pipeline for document and passage retrieval, respectively.

2 Related Work

Information Retrieval Information Retrieval (IR) involves finding documents relevant to a query, which plays a crucial role in applications such as search and question-answering (Zhu et al., 2023; Gao et al., 2023; Ram et al., 2023; Shi et al., 2024; Jeong et al., 2024a). Earlier IR approaches measured the similarity between queries and documents based on their lexical term matching, such as BM25 and TF-IDF (Robertson et al., 1994; Jones, 2004). Yet, these methods often struggled to capture the semantic nuances beyond surface-level term overlaps. To overcome this, along with advancements in language models (Devlin et al., 2019; Liu et al., 2019), there have been dense retrieval approaches that embed both the queries and documents into a shared dense vector space (Karpukhin et al., 2020; Xiong et al., 2021), enabling the calculation of semantic similarity between them more effectively by capturing the deeper contextual information. Yet, previous studies have mainly focused on enhancing the textual representations of queries and documents, while overlooking the multimodal nature of documents beyond text, which can potentially provide richer context and aid in more accurate retrieval (Liu et al., 2021; Jeong et al., 2024b).

Multimodal Information Retrieval Recent studies in IR have expanded the focus from purely text-based retrieval models to those that consider other modalities, such as images (Radford et al., 2021; Xiao et al., 2024), tables (Herzig et al., 2021; Chen et al., 2024) and graphs (Baek et al., 2023); however, the majority of these approaches (Zhou et al., 2024; Long et al., 2024; Lerner et al., 2024; Nowak et al., 2024; Caffagni et al., 2024) have primarily explored how to process the multimodal *queries*, meanwhile, they often overlook the equally important multimodal characteristics of the *documents* being retrieved. In efforts to handle diverse multimodal elements within documents, there are concurrent studies that have proposed to capture screenshots of documents, such as PDFs (Faysse et al., 2024; Cho et al., 2024) or Wikipedia web pages (Ma et al., 2024), and subsequently encoding them through vision models (Ding et al., 2024). However, these methods are not only limited by factors, such as image resolution and computational memory, constraining their application to documents longer than a single page[1], but also fall short by treating the diverse modalities within a document as a single visual entity, leading to sub-optimal document representations that may fail to effectively capture the nuanced interdependence between text and images. Also, while there are concurrent studies (Jiang et al., 2024b; Lin et al., 2024) that consider images and text as retrieval targets, they primarily focus on representing image-text pairs and their retrieval, rather than addressing the holistic representation of documents that include multiple images and another modality (tables). Finally, all the aforementioned work does not address the issue of splitting documents into smaller fragments (passages or sub-images), which may disrupt the holistic contextual view of the entire document.

[1]For instance, Ma et al. (2024) requires processing 9.8k image tokens just to process a single-page document, and it results in 2TB of storage for handling the entire Wikipedia corpus, which may not be practical.

Vision-Language Models Recently developed Vision-Language Models (VLMs) have emerged as a powerful tool for jointly processing visual and textual data, which combine the image understanding capabilities of visual encoders (Radford et al., 2021; Zhai et al., 2023) with the advanced reasoning abilities of language models (OpenAI, 2022, 2023a). These models have achieved remarkable performance across diverse vision-language (VL) tasks (such as image captioning and visual question answering) (Dai et al., 2023; OpenAI, 2023b), with the substantially limited attention on their applications to IR. We note that the latest developments in this field have particularly focused on enabling VLMs to handle interleaved, multimodal content, which involves a mixed sequence of images and text (Zhang et al., 2023; Li et al., 2024b). In particular, LLaVA-NeXT-Interleave (Li et al., 2024b) introduces a fine-tuning approach that specifically enhances the VLMs' capacity to understand complex interleavings of multiple images and text within a single context. Drawing inspiration from these advances, we propose to harness the capabilities of VLMs to create unified embeddings for documents interleaved with text and images (as well as tables).

3 Method

We present IDentIfy to holistically represent documents interleaved with multimodal elements.

3.1 Preliminaries

We begin with preliminaries, formally explaining information retrieval and vision-language models.

Information Retrieval IR is formally defined as the task of identifying a set of relevant documents $\{\boldsymbol{d}_1, \boldsymbol{d}_2, \dots, \boldsymbol{d}_k\} \subseteq \mathcal{D}$ from a large corpus $\mathcal{D}$, given an input query $\boldsymbol{q}$. Here, each query $\boldsymbol{q}$ and document $\boldsymbol{d}$ are represented as a sequence of tokens: $\boldsymbol{q} = [q_1, q_2, \dots, q_n]$ and $\boldsymbol{d} = [d_1, d_2, \dots, d_m]$, and traditional IR approaches typically consider these tokens as purely textual elements. However, we propose to extend this assumption to have the tokens of both the textual and visual content, to capture the multimodal nature of many real-world documents. Then, this new extension raises important questions of how can both the textual and visual content be represented within a unified token framework, and how can these multimodal tokens be seamlessly integrated and encoded for document representations. To answer them, we harness the power of recent vision-language models below.

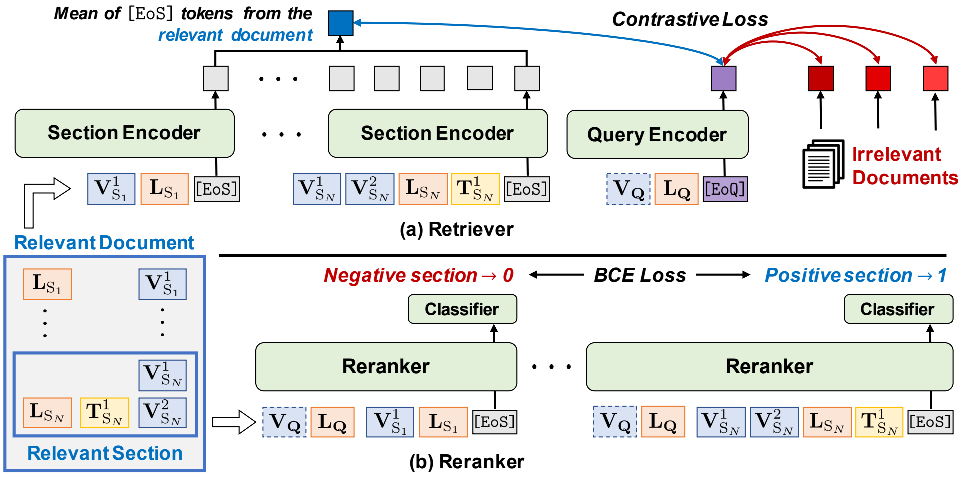


Figure 2: Overview of the proposed IDentIfy. **(a)**: In our document retriever, a query encoder represents a query (purple), and sections are encoded with a section encoder whose embeddings averaged to form a document representation (blue). Contrastive learning loss (red) is used for training the document retriever. **(b)**: Reranker scores query-section relevance with the concatenation of the query and section, trained using Binary Cross-Entropy loss.

Vision-Language Models We now turn to describing Vision-Language Models (VLMs), which are designed to jointly encode the textual and visual information in a unified token framework. These models are generally comprised of two main components: a visual encoder and a language model, interconnected through a projection layer. Specifically, given the document that may contain interleaved modalities (e.g., text and images), the visual encoder extracts high-level visual features from (multiple) images embedded within the document, mapping them into a latent space. Then, these visual features are transformed into a sequence of visual tokens via the projection layer, represented as follows: $\mathbf{V} \in \mathbb{R}^{V \times d_{\text{emb}}}$, where V denotes the visual token length and d_{emb} is the token dimension size. Similarly, for the textual content embedded within the document, the language model uses a word embedding layer to convert the input text into a sequence of tokens, as follows: $\mathbf{L} \in \mathbb{R}^{L \times d_{\text{emb}}}$, where L denotes the token length of text.

In this work, we also propose to account for tables that are an integral modality for holistically representing the full content of documents. However, in contrast to text and images that have dedicated processing layers within the VLM architectures, tables do not have a specific representation layer. Nevertheless, we argue that recent VLMs are pre-trained on diverse web data, and subsequently they are implicitly learned to handle the table structures formatted in HTML. Consequently, we treat HTML-format table data as a linearized sequence of HTML words, applying the same word embedding layer as is used for plain text. To be formal, this process converts the table content into table tokens, as follows: $\mathbf{T} \in \mathbb{R}^{T \times d_{\text{emb}}}$, where T is the token length of the table. Lastly, once extracted, the visual tokens, text tokens, and table tokens are concatenated (to form a unified token sequence) and then passed through the remaining layers of VLMs, to capture both uni- and cross-modal relationships across different modalities, ultimately enabling the comprehensive understanding of the documents.

3.2 Retriever

We now explain how we design a retriever specifically tailored for multimodal interleaved document retrieval. In particular, our approach leverages a VLM capable of processing text, images, and tables within a single document. Further, following the standard practice of existing retrieval architectures (Karpukhin et al., 2020; Xiong et al., 2021), we use a dual-encoder structure, which consists of a query encoder and a section encoder, both are based on the VLM, illustrated in Figure 2 (a).

Specifically, thanks to the use of the VLM, our query encoder can take either purely textual queries $\boldsymbol{q} = \mathbf{L}_{\text{Q}}$ or multimodal queries consisting of text and corresponding visual elements $\boldsymbol{q} = [\mathbf{V}_{\text{Q}}, \mathbf{L}_{\text{Q}}]$. Also, to obtain the final query representation, we introduce a learnable token called 'End of Query', $\texttt{[EoQ]} \in \mathbb{R}^{d_{\text{emb}}}$. This token is appended to the end of the sequence of query tokens $\boldsymbol{q}$, and the final concatenated tokens $[\boldsymbol{q}, \texttt{[EoQ]}]$ are then passed through the query encoder. Then, the model output corresponding to [EoQ] is used as the final query representation, as follows: $\mathbf{Z}_{\text{Q}} \in \mathbb{R}^{d_{\text{emb}}}$.

For documents, we first represent each document $\boldsymbol{d}$ as a sequence of sections $\boldsymbol{d} = [\boldsymbol{s}_i]_{i=1}^{S}$ (with a total of S sections), where each section $\boldsymbol{s}_i$ is derived by dividing the document according to its subtitles. $\boldsymbol{s}_i$ can contain a combination of text tokens $\mathbf{L}_{\mathrm{S}i}$, visual tokens from embedded images $\mathbf{V}_{\mathrm{S}i}$, and table tokens $\mathbf{T}_{\mathrm{S}i}$, denoted as follows: $\boldsymbol{s}_i = [\mathbf{V}_{\mathrm{S}_i}, \mathbf{L}_{\mathrm{S}_i}, \mathbf{T}_{\mathrm{S}_i}]$. Then, to obtain a section-level representation, similar to the query representation, we introduce a learnable token, called 'End of Section': `[EoS]`$\in \mathbb{R}^{d_{\mathrm{emb}}}$, which is similarly appended at the end of each section. We then forward concatenated tokens $[\boldsymbol{s}_i,$ `[EoS]`$]$ to the section encoder, and, after that, the output corresponding to `[EoS]` is used to form the section representation, as follows: $\mathbf{Z}_{\mathrm{S}_i} \in \mathbb{R}^{d_{\mathrm{emb}}}$. Additionally, the overall document representation is obtained by averaging the representations of all sections within the document, defined as follows: $\mathbf{Z}_{\mathrm{D}} = \frac{1}{S}\sum_{i=1}^{S} \mathbf{Z}_{\mathrm{S}_i}$.

The remaining step to discuss here is how to train those two query and section encoders for IR. Recall that the goal of the retriever is to assess a relevance score between the query and the document. To achieve this goal, we use a contrastive learning loss based upon the query and document representations, whose objective is to assign higher similarity scores to relevant documents (positive samples) and lower scores to irrelevant ones (negative samples) for the query, formulated as follows:

$$\mathcal{L}_{\text{retriever}} = -\frac{1}{B}\sum_{i=1}^{B} \log\left(\frac{\phi\left(\mathbf{Z}_{\mathrm{Q}_i}, \mathbf{Z}_{\mathrm{D}_i}\right)}{\sum_{j=1}^{B} \phi\left(\mathbf{Z}_{\mathrm{Q}_i}, \mathbf{Z}_{\mathrm{D}_j}\right)}\right),$$

$$\phi\left(\mathbf{a}, \mathbf{b}\right) = \exp\left(\frac{\mathbf{a}^{\top}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}\right), \tag{1}$$

where B is the batch size during the training phase. Here, by minimizing $\mathcal{L}_{\text{retriever}}$, the retriever learns to optimize the similarity between queries and their relevant documents, enabling the retrieval of the most pertinent documents (among all) for the given input query during inference.

3.3 Reranker

To enable fine-grained retrieval within documents beyond the retrieval of documents themselves, we introduce a section-level reranking mechanism that identifies the section most relevant to the query. In particular, once the document is retrieved, the objective of the reranker f_{R} is to pinpoint the specific sections within the document that best match the query. We also note that this reranker is similarly operationalized with the VLM along with a binary classifier on top of it, which directly measures the relevance of each query-section pair (Figure 2 (b)).

Formally, for a retrieved document, we take each of its sections $\boldsymbol{s}_i$ with a learnable token for section embedding `[EoS]` attached to the end and concatenate it with query $\boldsymbol{q}$, forming the input sequence of $[\boldsymbol{q}, \boldsymbol{s}_i,$ `[EoS]`$]$. The concatenated tokens are then processed through the reranker, and its output corresponding to `[EoS]` captures the relevance between the query and section, which is further subsequently passed to a binary classifier consisting of a linear layer followed by a Sigmoid function. Through this, the classifier outputs a probability score indicating the likelihood of the section being relevant to the query, *i.e.*, a score close to one denotes a high relevance (positive section), meanwhile, a score near zero indicates irrelevance (negative section).

To train this reranker, we use the binary cross-entropy loss, formulated as follow:

$$\mathcal{L}_{\text{reranker}} = \sum_{i=1}^{B}\sum_{j=1}^{S_i} \frac{1}{BS_i} \ell\left(\mathbf{y}_{\boldsymbol{s}_{i,j}}, f_{\mathrm{R}}\left([\boldsymbol{q}, \hat{\boldsymbol{s}}_{i,j}]\right)\right),$$

$$\ell\left(y, \hat{y}\right) = -\left[y \log \hat{y} + (1-y)\log(1-\hat{y})\right], \tag{2}$$

where S_i is the number of sections in the i-th document, $\mathbf{y}_{\boldsymbol{s}_{i,j}}$ is the label for the j-th section of the i-th document $\boldsymbol{s}_{i,j}$ (with its value of one if relevant to the query $\boldsymbol{q}$, otherwise zero), $\hat{\boldsymbol{s}}_{i,j} = [\boldsymbol{s}_{i,j},$ `[EoS]`$]$, and B is the batch size during training. Also, during training, the sections not labeled as relevant to the query are considered negative samples. Then, by minimizing $\mathcal{L}_{\text{reranker}}$, the reranker learns to predict section relevance for any query, thus refining our overall retrieval process by allowing the retrieval of not just whole documents but also their most relevant sections, for multiple use cases of IR.

4 Experiments

4.1 Experimental Setups

Datasets We evaluate IDentIfy on four benchmark datasets designed for multimodal IR tasks that require understanding of both the textual and visual cues within queries and documents, as follows: **Encyclopedic-VQA** (Mensink et al., 2023) is a large-scale benchmark for multimodal Visual Question Answering (VQA) with queries linked to specific Wikipedia sections and includes both text-only and multimodal queries; **InfoSeek** (Chen et al., 2023) is a knowledge-intensive VQA dataset with multimodal questions generated from Wikidata triples that include diverse entities such as landmarks, animals, and food; **ViQuAE** (Lerner et al., 2022) involves both text-based and multimodal queries about human entities, linked to annotated Wikipedia sections, making it ideal for evaluating

Table 1: Results with different document formats for retrieval.

Format	R@1	R@10	R@100	MRR@10
Entity	3.1	15.5	39.7	6.1
Summary	13.4	41.3	66.5	21.6
Text-document	12.5	37.8	68.7	19.8
+ Single-image	16.4	45.4	77.1	25.3
+ Interleaved (Ours)	**20.5**	**50.0**	**78.0**	**29.4**

Table 2: Results with different section retrieval strategies. Document (Ours) performs document retrieval and section reranking, whereas Passage performs section retrieval and reranking. * denotes the model without reranking.

Granularity	R@1	R@10	R@20	MRR@10
Passage*	3.9	16.9	22.0	7.5
Passage	28.6	36.4	37.8	31.2
Document (Ours)	**35.1**	**50.8**	**53.6**	**40.3**

Table 3: Performance on document retrievals. **(a)**: Results of document retrieval for multimodal queries on InfoSeek and ViQuAE. **(b)**: Results of document retrieval for textual queries on Encyclopedic-VQA (Enc-VQA) and ViQuAE.

(a) Document Retrieval with Multimodal Queries

Format	Dataset	R@1	R@10	R@100	MRR@10
Text-document	InfoSeek	6.8	23.6	52.5	11.2
+ Interleaved	InfoSeek	**10.2**	**30.4**	**57.3**	**15.7**
Text-document	ViQuAE	13.5	40.4	67.4	20.9
+ Interleaved	ViQuAE	**17.5**	**46.0**	**69.4**	**26.3**

(b) Document Retrieval with Textual Queries

Format	Dataset	R@1	R@10	R@100	MRR@10
Text-document	Enc-VQA	62.7	76.3	87.4	67.0
+ Interleaved	Enc-VQA	**65.4**	**76.8**	**87.8**	**69.0**
Text-document	ViQuAE	55.8	71.5	**83.0**	60.9
+ Interleaved	ViQuAE	**56.5**	**72.2**	**83.0**	**61.6**

section retrieval; **Open-WikiTable** (Kweon et al., 2023) extends WikiSQL (Zhong et al., 2017) and WikiTableQuestions (Pasupat and Liang, 2015), targeting open-domain table QA by identifying documents or sections containing relevant tables. We provide more details on datasets in Appendix A.

Baselines We compare our approach against diverse baselines that capture different document representations. First, the **Entity** and **Summary** baselines retrieve documents based on their titles and summary sections, respectively, leveraging high-level textual cues. Also, the **Text-document retriever** baseline utilizes the full textual content of documents for retrieval. We further include the **Single-image** baseline that additionally leverages the first image of each document. **IDentIfy** is our model that holistically represents multimodal content (text, images, and tables) within documents.

Evaluation Metrics To evaluate our approach, we use standard metrics: Recall@K (R@K) measures whether the relevant document or section appears within the top-K results; MRR@K measures how early the first relevant item is ranked (within top-K) by averaging its inverse rank across queries.

Implementation Details We use LLaVA-NeXT-Interleave (Li et al., 2024b) of 0.5B parameters as the basis VLM for both the retriever and reranker. During training, documents are represented using randomly selected four sections, while in inference, we consider all sections within each document. For section-level retrieval, all sections within the top 25 retrieved documents are reranked. Experiments are conducted on a single H100 GPU.

4.2 Experimental Results and Analyses

Main Results We report retrieval performance on the Encyclopedic-VQA dataset in Table 1, where queries include both text and images. We observe that IDentIfy achieves the best performance, improving R@1 scores by 53.0%, 64.0%, and 25.0% over Summary, Text-document, and Single-image retrieval baselines, respectively, with similar trends observed for other metrics. These results demonstrate the effectiveness of integrating multimodal content holistically into a unified representation. To further illustrate the advantages of our approach, we provide case studies in Appendix C.

We further examine the impact of our document retrieval and section reranking pipeline. In Table 2, the passage retriever represents individual sections as separate retrieval units, whereas the document retriever (ours) aggregates multiple section representations into a single representation. Then, we perform reranking over the retrieved sections or the sections from the retrieved documents, and then report the results in Table 2 (where * denotes the model without reranking). From this, we observe that the passage retriever without reranking (Passage*) achieves suboptimal retrieval performance, highlighting the challenge in pinpointing the most relevant section within a document using traditional retrieval methods. In contrast, when the reranker is used alongside the document retriever, the performance significantly surpasses that of the passage retrieval. These results confirm the importance of leveraging holistic context from multiple, interrelated sections within documents.

Interleaved format enhances document retrieval across modalities. We further expand our experiments to two additional datasets, InfoSeek and ViQuAE, and report document retrieval results. As shown in Table 3, our model consistently outperforms the Text-document baseline for both multimodal and text-only queries. We attribute these

Table 4: Performance on section reranking. **(a)**: Results of section reranking for multimodal queries on Encyclopedic-VQA (Enc-VQA) and ViQuAE. **(b)**: Results of section reranking for textual queries on Enc-VQA and ViQuAE. For fair comparison between different methods, the reranking target (i.e., candidate sections from retrieved documents) are the same across them.

(a) Section Reranking with Multimodal Queries

Format	Dataset	R@1	R@10	R@20	MRR@10
Text-document	Enc-VQA	40.7	52.8	55.5	44.8
+ Interleaved	Enc-VQA	**42.4**	**53.6**	**55.7**	**46.3**
Text-document	ViQuAE	**12.6**	31.7	37.7	**18.2**
+ Interleaved	ViQuAE	11.4	**32.1**	**39.2**	17.5

(b) Section Reranking with Textual Queries

Format	Dataset	R@1	R@10	R@20	MRR@10
Text-document	Enc-VQA	68.1	79.4	80.2	72.3
+ Interleaved	Enc-VQA	**69.7**	**80.1**	**80.6**	**73.6**
Text-document	ViQuAE	27.8	50.2	57.7	35.0
+ Interleaved	ViQuAE	**29.9**	**50.9**	**59.8**	**36.7**

Table 5: Retrieval results for tables, where Zero-shot denotes a model trained on Encyclopedic-VQA but not on the target dataset. Finetuned refers to additional training of the model on the target dataset. **(a)**: Results for tabular document retrieval on Open-WikiTable (OWT). **(b)**: Textual and tablular section retrieval results on ViQuAE and OWT datasets, respectively. **(c)**: Reranker accuracy of a classification task that identifies the section containing the query-associated table given a gold document.

(a) Document Retrieval for Tables

Model	R@1	R@10	R@100	MRR@10
Zero-shot	29.4	58.0	86.0	38.1
Finetuned	**55.8**	**84.1**	**93.5**	**66.1**

(c) Tabular Classification

Model	Random	Zero-shot	Finetuned
Acc@1	11.9	9.3	**56.5**

(b) Section Retrieval for Tables

Model	Modality	Dataset	R@1	R@10	R@20	MRR@10
Zero-shot	Text	ViQuAE	20.3	**49.0**	**57.7**	28.9
Finetuned	Text	ViQuAE	**29.9**	**50.9**	**59.8**	**36.7**
Zero-shot	Table	OWT	5.9	**20.5**	**29.4**	9.1
Finetuned	Table	OWT	**8.4**	**36.7**	**52.8**	**15.2**

gains to the integration of multimodal content, allowing the VLM to capture richer alignments and leverage pre-existing knowledge for more effective document representation (Xu et al., 2024).

Interleaved format is also beneficial in section retrieval. Similarly, we evaluate section retrieval performance on Encyclopedic-VQA and ViQuAE datasets, for both multimodal and textual queries. As shown in Table 4, our model outperforms the Text-document baseline in most cases. However, the performance gains over the baseline are smaller compared to the document retrieval setup. This is likely because section reranking focuses on evaluating the relationship between a single section and a query (rather than leveraging the holistic context of the entire document), and individual sections may lack the diverse multimodal information necessary for fully capturing the intent of queries.

Retrieving tables interleaved within documents is challenging. We explore the retrieval task for tabular data, aiming to identify documents or sections containing query-relevant tables, and compare models trained on Encyclopedic-VQA (Zero-shot) with those additionally trained on Open-WikiTable (Finetuned). As shown in Table 5 (a), the Finetuned retriever outperforms the Zero-shot retriever on retrieving documents containing query-relevant tables. However, more fine-grained section reranking results (identifying sections containing query-relevant tables) in Table 5 (b) may reveal a notable modality-specific challenge: the performance of Zero-shot and Finetuned rerankers is considerably lower on table retrieval compared to their performance on text retrieval, despite both the text and tables being represented with word tokens. To better understand this, we design a classification task, where rerankers are tasked with identifying the correct section containing the target table within the golden document. Then, as shown in Table 5 (c), the Zero-shot reranker performs comparably to random selection, while the Finetuned reranker shows modest improvements. These findings highlight the intrinsic challenge of tabular retrieval, suggesting the need for table-specific modules to more holistically represent multimodal interleaved documents.

More sections enhance document retrieval performance but raise computational costs. To see how the number of sections used for representing each document impacts performance, we evaluate document retrieval on the InfoSeek dataset by varying the sections per document during training. As shown in Figure 3, incorporating more sections improves MRR@10 from 7.5 to 15.7 due to leveraging richer multimodal and contextual information. However, this comes at the cost of increased computational requirements, as processing more sections raises GPU memory consumption.

Sections from the same document act as effective negatives to enhance reranker performance. In training the reranker, we investigate whether considering sections from the same document as negative examples (called In-document) is effective than other strategies, such as Top-K negatives (top-K retrieved sections based on their similarity with the input query) and In-batch negatives (positive sections from other samples in the same batch). As

Figure 3: Trade-off between performance (MRR@10) and training cost (GPU Memory) for retrieval.

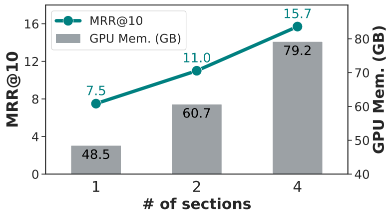


Table 6: Comparison of negative sample selection strategies for reranker training: Top-K (top-k retrieved sections), In-batch (sections from other samples in the batch), and In-document (sections in the same document).

Negative	R@1	R@20	MRR@10
Top-K	38.1	55.3	44.4
In-batch	39.5	55.4	45.0
In-document (Ours)	**42.4**	**55.7**	**46.3**

Table 7: Comparison of different training objectives for the reranker: Contrastive considers sections as retrieval units and uses the one for document retriever training; Document + BCE concatenates the query with multiple sections from the same document and uses the BCL loss; Section + BCE trains the reranker by concatenating the query with each section individually.

(a) Section Retrieval for Multimodal Queries

Train Loss	R@1	R@10	R@20	MRR@10
Contrastive	3.6	15.0	21.3	6.5
Document + BCE	13.6	29.6	32.9	24.1
Section + BCE (Ours)	**42.4**	**53.6**	**55.7**	**46.3**

(b) Section Retrieval for Textual Queries

Train Loss	R@1	R@10	R@20	MRR@10
Contrastive	13.6	37.7	45.1	20.6
Document + BCE	23.8	43.4	47.2	39.1
Section + BCE (Ours)	**69.7**	**80.1**	**80.6**	**73.6**

shown in Table 6, we observe that the In-document approach achieves superior performance especially on R@1, demonstrating its ability to effectively identify the most pertinent section among highly similar sections within the same document, i.e., its training objective can encourage the reranker to focus on fine-grained distinctions between closely related sections (within the same document).

BCE loss is the most effective to train the section reranker. In our reranker design, we use a binary cross-entropy (BCE) loss by concatenating the query with each document section individually (Section + BCE), allowing the model to directly assess query-section relevance. As an alternative, we also explore a contrastive loss (Contrastive), which models section reranking similarly to document retrieval but uses sections as the retrieval units, and a variant of BCE loss (Document + BCE), where the query is concatenated with multiple sections (both positive and negative) from the same document. As shown in Table 7, the Section + BCE reranker outperforms both alternatives. Specifically, contrastive loss performs the worst, suggesting that direct concatenation of query and section provides clearer relevance signals, consistent with conventional reranking approaches. Moreover, while Document + BCE leverages inter-section context, its performance might be hindered by training constraints as the model processes fewer sections during training (Jiang et al., 2024a; Lee et al., 2024), and addressing it would be interesting future work.

Our Approach is Robust with Different Base Models. To ensure the effectiveness of our approach across VLMs, we evaluate its performance with another VLM, LLaVA-OneVision (Li et al., 2024a), with 0.5 billion paramters, in addition to LLaVA-NeXT-Interleave (Li et al., 2024b) used in our main experiments. Results in Table 8 show that ours continues to outperform baselines, achieving a notable 30.1% gain in R@1 over the best baseline.

Table 8: Results with another base model (LLaVA-OneVision) for document retrieval (with different document formats).

Format	R@1	R@10	R@100	MRR@10
Entity	2.3	10.3	29.7	4.3
Summary	7.6	24.7	55.7	12.0
Text-document	7.0	24.1	50.4	11.7
+ Single-image	9.3	31.4	61.9	15.4
+ Interleaved (Ours)	**12.1**	**36.1**	**62.5**	**18.2**

5 Conclusion

In this paper, we introduced IDentIfy, a novel IR framework designed to address the limitations of conventional methods that rely on textual content of documents and their segmented passages. Specifically, our approach sits on top of recent VLMs, which enables integration and representation of diverse multimodal content (including text, images, and tables) into a unified document representation. Also, unlike previous strategies that segment documents at the passage level, our method merges these segments to maintain the document's structural coherence, while further introducing a reranking strategy for precise identification of relevant sections. Extensive experiments across various IR datasets demonstrated that IDentIfy consistently outperforms existing baselines, confirming that the interleaved multimodal representation significantly enhances the quality of the document retrieval. We believe IDentIfy represents a crucial step toward more comprehensive and contextually aware IR systems, capable of handling the increasing multimodality of modern information sources.

Limitations

Due to the limitations of a single H100 GPU, we represent documents by selecting a limited number of sections and averaging their corresponding embeddings. While this reduces the computational demands, our findings suggest that capturing a broader document context leads to improved retrieval performance. Hence, leveraging the long context window of LVLMs could further enhance document retrieval by capturing more comprehensive information from the full document. Moreover, our reranker design follows the conventional approach of concatenating the input query with individual sections. However, we believe that providing the reranker with all the sections together would allow the model to better leverage the contextual information from the entire interleaved document, potentially resulting in improved performance. In order to fully leverage the interleaved format in the IR system, addressing the issues by reducing the GPU load when processing interleaved documents would greatly boost overall IR performance. We leave these explorations for future work.

Ethics Statement

In this work, we use a publicly available retrieval corpus for information retrieval tasks. However, the retrieval corpus may contain private, harmful, or biased content. Such undesirable features could unintentionally be reflected in the behavior of retrievers and rerankers trained on this data, potentially leading to ethical concerns during real-world deployment. However, current information retrieval techniques, including ours, do not address the retrieval of undesirable content. We recognize the critical need for safeguards to mitigate this issue. This is essential to ensure that information retrieval systems are reliable, fair, and safe for deployment.

References

Jinheon Baek, Alham Fikri Aji, Jens Lehmann, and Sung Ju Hwang. 2023. Direct fact retrieval from knowledge graphs without entity linking. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. *arXiv preprint arXiv:2404.15406*.

Peter Baile Chen, Yi Zhang, and Dan Roth. 2024. Is table retrieval a solved problem? exploring join-aware multi-table retrieval. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *Preprint*, arXiv:2411.04952.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024. PDF-MVQA: A dataset for multimodal information retrieval in pdf-based visual question answering. *arXiv preprint arXiv:2404.12720*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Martin Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and

Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024a. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024b. Database-augmented query representation for information retrieval. *arXiv preprint arXiv:2406.16013*.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. 2024a. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2024b. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*.

Karen Spärck Jones. 2004. A statistical interpretation of term specificity and its application in retrieval. *J. Documentation*, 60(5):493–502.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sunjun Kweon, Yeonsu Kwon, Seonhee Cho, Yohan Jo, and Edward Choi. 2023. Open-wikitable : Dataset for open domain question answering with complex reasoning over table. In *Findings of the Association for Computational Linguistics (ACL)*.

Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien M. R. Arnold, Vincent Perot, Siddharth Dalmia, Hexiang Hu, Xudong Lin, Panupong Pasupat, Aida Amini, Jeremy R. Cole, Sebastian Riedel, Iftekhar Naim, Ming-Wei Chang, and Kelvin Guu. 2024. Can long-context language models subsume retrieval, rag, sql, and more? *arXiv preprint arXiv:2406.13121*.

Paul Lerner, Olivier Ferret, and Camille Guinaudeau. 2024. Cross-modal retrieval for knowledge-based visual question answering. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval*.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G. Moreno, and Jesús Lovón-Melgarejo. 2022. Viquae, a dataset for knowledge-based visual question answering about named entities. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zheyuan Liu, Cristian Rodriguez Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Xinwei Long, Jiali Zeng, Fandong Meng, Zhiyuan Ma, Kaiyan Zhang, Bowen Zhou, and Jie Zhou. 2024. Generative multi-modal knowledge retrieval with large language models. In *Proceedings of the AAAI National Conference on Artificial Intelligence (AAAI)*.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251*.

Thomas Mensink, Jasper R. R. Uijlings, Lluís Castrejón, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araújo, and Vittorio Ferrari. 2023. Encyclopedic VQA: visual questions about detailed properties of fine-grained categories. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Averi Nowak, Francesco Piccinno, and Yasemin Altun. 2024. Multimodal chart retrieval: A comparison of text, table and image based approaches. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`.

OpenAI. 2023a. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2023b. GPT-4V(ision) system card. `https://openai.com/index/gpt-4v-system-card/`.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Trans. Assoc. Comput. Linguistics*, 11:1316–1331.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: knowledge-aware visual question answering. In *Proceedings of the AAAI National Conference on Artificial Intelligence (AAAI)*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: retrieval-augmented black-box language models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Zilin Xiao, Ming Gong, Paola Cascante-Bonilla, Xingyao Zhang, Jie Wu, and Vicente Ordonez. 2024. Grounding language models for visual entity recognition. *arXiv preprint arXiv:2402.18695*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jialiang Xu, Michael Moor, and Jure Leskovec. 2024. Reverse image retrieval cues parametric memory in multimodal llms. *arXiv preprint arXiv:2405.18740*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2023. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. VISTA: visualized text embedding for universal multi-modal retrieval. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

A Details of Experimental Setups

Dataset configuration Table 9 summarizes the key properties of the datasets used in our experiment, including query modality, target item, entity domain, number of entities, and whether a section ID is provided to indicate the section containing the answer. Additionally, we provide the number of samples in the training, evaluation, and test splits, as well as the size of the corpus. We provide a more detailed explanation of the datasets below.

- **Encyclopedic-VQA** (Mensink et al., 2023) is a large-scale visual question-answering (VQA) benchmark dataset, widely used for measuring the performance of multimodal IR models. Each query is linked to a specific section of a Wikipedia document (containing an answer for it) and is manually annotated by humans. Also, this dataset offers both text-only and multimodal queries. In addition to this, the queries are related to fine-grained properties of species and landmarks. Our experiments focus on the single-hop category where questions can be answered in a single retrieval step.
- **InfoSeek** (Chen et al., 2023) is a dataset designed for knowledge-intensive VQA, covering a wide range of entities (such as landmarks, animals, and food). Questions are generated by filling human-written templates with knowledge triples (subject, relation, object) available from Wikidata, which involve only the multimodal queries. As the test dataset is not available, we use the validation set as our test set, and split the training set into training and validation subsets with a 9:1 ratio.
- **ViQuAE** (Lerner et al., 2022) is a dataset focused about human entities. It provides both textual and multimodal queries, with each query linked to a specific section of a Wikipedia document that contains an answer annotated by humans, which makes it an ideal benchmark for section retrieval.
- **Open-WikiTable** (Kweon et al., 2023) is an extension of WikiSQL (Zhong et al., 2017) and WikiTableQuestions (Pasupat and Liang, 2015), designed for open-domain table question answering that requires retrieval of the most relevant table from a broader corpus. For our experiments, we adapt the WikiTableQuestions subset of Open-WikiTable, aiming at identifying the document or document section containing the target table.

Dataset pre-processing In our study, we leverage interleaved multimodal content from Wikipedia documents. However, existing corpora associated with IR datasets often lack this content, typically only including the first few words of each document. Therefore, we download the HTML file of each Wikipedia document for corpus augmentation.

If the dataset provides Wikipedia URLs for its corpus, we use them to download the HTML files. Alternatively, if only entity names are provided, we generate Wikipedia URLs using those names. If a Wikipedia URL is deprecated, we remove the corresponding document from the corpus along with any associated queries. From the HTML files, we extract text, image URLs, and tables. We then split the contents by subtitles in the document where each chunk corresponds to a section. For the images, we use the image URLs to download the corresponding images, removing any invalid URLs. This process produces a dictionary that organizes text, images, and tables by section.

Since downloading contents for all documents across datasets is time- and memory-intensive, we preprocess subsets of each corpus, including documents relevant to queries in the training, evaluation, and test splits, along with unrelated documents.

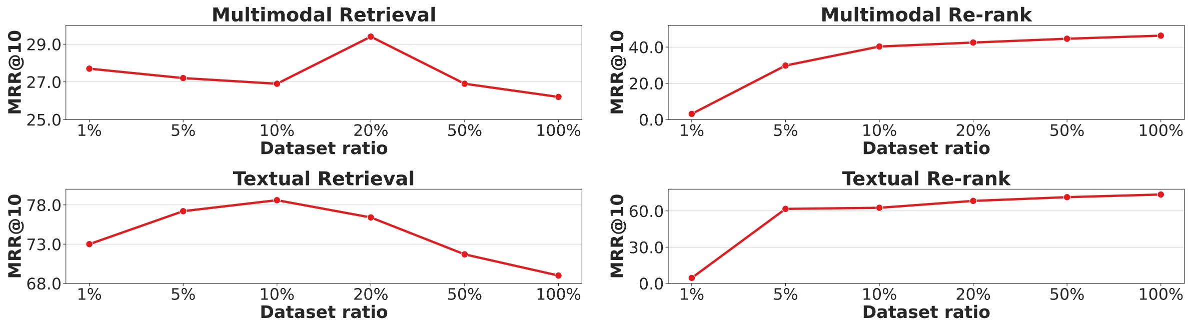
Implementation Details To take advantage of larger batch sizes (while reducing GPU memory usage), we apply LoRA (Hu et al., 2022). Also, to further optimize the GPU usage, we scale each image down to half of its original height and width and then combine four scaled-down images into a single composite image. All experiments are conducted using a single H100 GPU.

B Additional Experimental Results

Data Requirements for Models We analyze the effect of different dataset sizes for training on retriever and reranker performance. To achieve this, we randomly prune samples in the Encyclopedic-VQA dataset at various ratios and report the performance of models trained on these subsets. In Figure 4 (a), we observe that too many samples can degrade retrieval performance. Also, retrieval of textual queries requires fewer samples to reach its optimal performance compared to multimodal retrieval. Similarly, in Figure 4 (b), section retrieval for multimodal queries requires 10% of the dataset to achieve 80% of the full-dataset performance, while section retrieval for textual queries needs only 5%. These observations suggest that addi-

Table 9: Information retrieval datasets summary.

Dataset	Query Modality	Target	Domain	Entities	Section ID	Train	Eval	Test	Corpus size
Encyclopedic-VQA	Text, Text-Image	Text	Species, Landmarks	17k	○	177k	2.2k	3.8k	100k
InfoSeek	Text-Image	Text	Diverse	11k	×	209k	23k	74k	500k
ViQuAE	Text, Text-Image	Text	Human	1k	○	1.2k	1.2k	1.2k	100k
Open-WikiTable	Text	Table	Table	-	○	3.3k	0.4k	0.4k	1.8k



(a) Retriever performance **(b) Reranker performance**

Figure 4: Retrieval performance with different dataset sizes for training. **(a)**: When training a retriever, large datasets rather deteriorate the retrieval performance as it may be overfitted, resulting in low generalization. **(b)**: On the other hand, a larger dataset size is beneficial to training a re-ranker.

tional modalities increase the need for more data. This accounts for the inferior performance of the interleaved format in the ViQuAE experiments (Table 4 (a)). The ViQuAE dataset, at only 2.2% of the size of Encyclopedic-VQA, may be small for the reranker to effectively learn multimodal query-section alignments. We also observe that section retrieval is more challenging, with more samples improving the reranker's performance. This explains why the ViQuAE reranker has much lower section retrieval scores compared to the one trained on the Encyclopedic-VQA (Table 4 (b)). Given the challenge of obtaining large query-section pair samples, exploring more effective reranker training pipelines is necessary.

C Case Studies of Document Retrieval

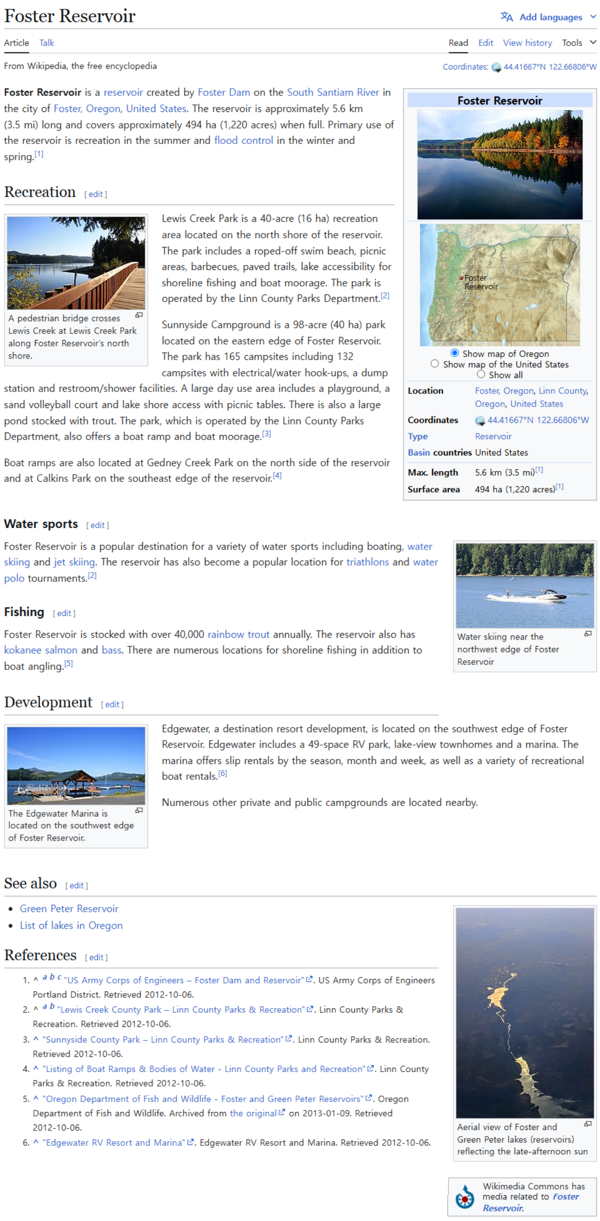
We conduct case studies to demonstrate the advantages of our approach in document retrieval with textual and multimodal queries. In Figure 5 and Figure 6, we illustrate the instances where our approach, which leverages interleaved multimodal contents (e.g., images, tables, and text) within documents, retrieved correct documents for given queries, while the conventional one, which represents documents using only textual data, retrieved documents that appeared to be relevant but were not actually related to the queries.

In Figure 5, a textual query asks for the name of the park located on the north shore of Foster Reservoir. The conventional approach retrieved a document containing unrelated information about a different reservoir. While this document includes terms such as "Peak District National Park" and "North America farm," which make the document superficially relevant, it fails to answer the query. In contrast, our approach identified the document containing the correct answer to the given query.

The advantages of integrating multimodal content into document representation become more apparent in document retrieval with multimodal queries, as shown in Figure 6. For a query consisting of an image of a town hall in Hanover and a textual question about its designer, both our approach and the conventional one retrieved documents about town halls in Germany. However, our approach pinpointed the exact document about the town hall in Hanover, indicating that Hermann Eggert designed the building. The conventional method retrieved a document about a town hall in Munich, which is somewhat related but not an exact match to the query image or question.

These cases underscore the benefits of leveraging multimodal content in information retrieval. Integrating interleaved multimodal elements, our approach aligns more effectively with the input query, resulting in more accurate and fine-grained retrieval. This superiority is supported by Xu et al. (2024), which highlights that models perform better when prompted with rich multimodal information, enabling them to capture alignments across modalities and enhance the representation of given inputs.

Q: What is the name of the park on the north shore of foster reservoir?



Foster Reservoir

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Coordinates: 44.41667°N 122.66806°W

Foster Reservoir is a reservoir created by Foster Dam on the South Santiam River in the city of Foster, Oregon, United States. The reservoir is approximately 5.6 km (3.5 mi) long and covers approximately 494 ha (1,220 acres) when full. Primary use of the reservoir is recreation in the summer and flood control in the winter and spring.[1]

Foster Reservoir	
Location	Foster, Oregon, Linn County, Oregon, United States
Coordinates	44.41667°N 122.66806°W
Type	Reservoir
Basin countries	United States
Max. length	5.6 km (3.5 mi)[1]
Surface area	494 ha (1,220 acres)[1]

Show map of Oregon / Show map of the United States / Show all

Recreation [edit]

Lewis Creek Park is a 40-acre (16 ha) recreation area located on the north shore of the reservoir. The park includes a roped-off swim beach, picnic areas, barbecues, paved trails, lake accessibility for shoreline fishing and boat moorage. The park is operated by the Linn County Parks Department.[2]

A pedestrian bridge crosses Lewis Creek at Lewis Creek Park along Foster Reservoir's north shore.

Sunnyside Campground is a 98-acre (40 ha) park located on the eastern edge of Foster Reservoir. The park has 165 campsites including 132 campsites with electrical/water hook-ups, a dump station and restroom/shower facilities. A large day use area includes a playground, a sand volleyball court and lake shore access with picnic tables. There is also a large pond stocked with trout. The park, which is operated by the Linn County Parks Department, also offers a boat ramp and boat moorage.[3]

Boat ramps are also located at Gedney Creek Park on the north side of the reservoir and at Calkins Park on the southeast edge of the reservoir.[4]

Water sports [edit]

Foster Reservoir is a popular destination for a variety of water sports including boating, water skiing and jet skiing. The reservoir has also become a popular location for triathlons and water polo tournaments.[2]

Water skiing near the northwest edge of Foster Reservoir

Fishing [edit]

Foster Reservoir is stocked with over 40,000 rainbow trout annually. The reservoir also has kokanee salmon and bass. There are numerous locations for shoreline fishing in addition to boat angling.[5]

Development [edit]

Edgewater, a destination resort development, is located on the southwest edge of Foster Reservoir. Edgewater includes a 49-space RV park, lake-view townhomes and a marina. The marina offers slip rentals by the season, month and week, as well as a variety of recreational boat rentals.[6]

Numerous other private and public campgrounds are located nearby.

The Edgewater Marina is located on the southwest edge of Foster Reservoir.

See also [edit]

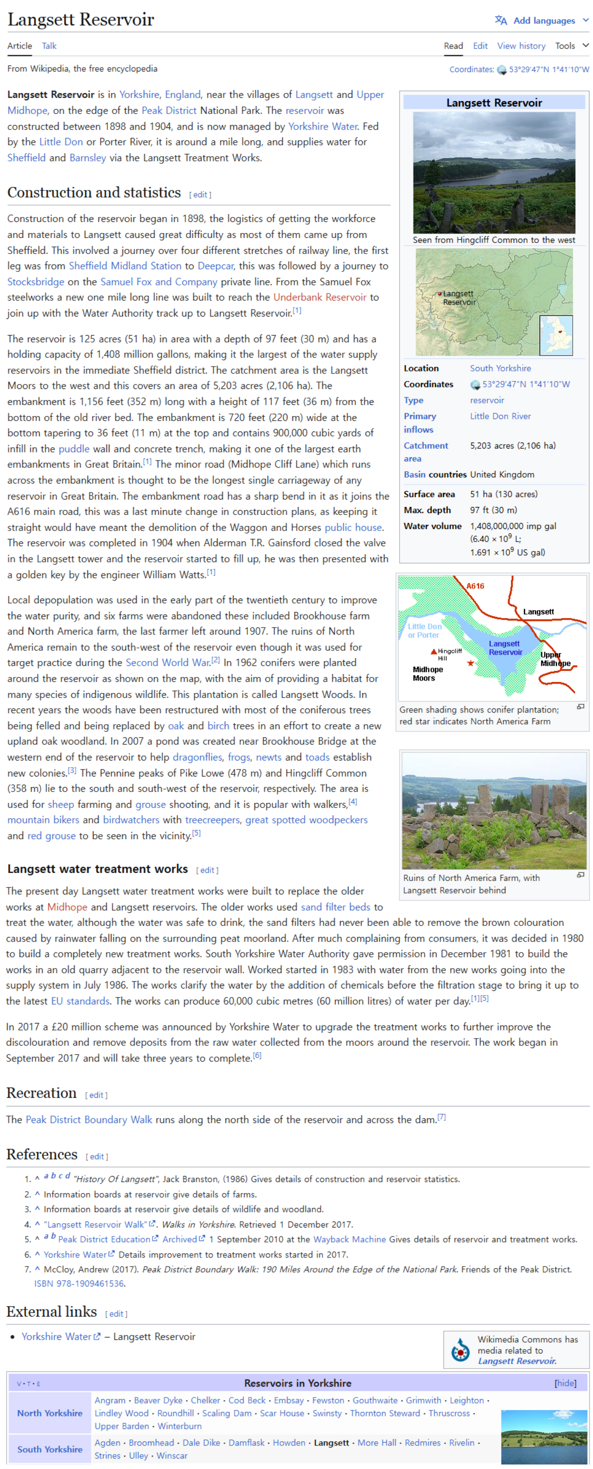
- Green Peter Reservoir
- List of lakes in Oregon

References [edit]

1. ^ a b c "US Army Corps of Engineers – Foster Dam and Reservoir". US Army Corps of Engineers Portland District. Retrieved 2012-10-06.
2. ^ a b "Lewis Creek County Park – Linn County Parks & Recreation". Linn County Parks & Recreation. Retrieved 2012-10-06.
3. ^ "Sunnyside County Park – Linn County Parks & Recreation". Linn County Parks & Recreation. Retrieved 2012-10-06.
4. ^ "Listing of Boat Ramps & Bodies of Water – Linn County Parks and Recreation". Linn County Parks & Recreation. Retrieved 2012-10-06.
5. ^ "Oregon Department of Fish and Wildlife – Foster and Green Peter Reservoirs". Oregon Department of Fish and Wildlife. Archived from the original on 2013-01-09. Retrieved 2012-10-06.
6. ^ "Edgewater RV Resort and Marina". Edgewater RV Resort and Marina. Retrieved 2012-10-06.

Aerial view of Foster and Green Peter lakes (reservoirs) reflecting the late-afternoon sun

Wikimedia Commons has media related to *Foster Reservoir*.



Langsett Reservoir

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Coordinates: 53°29′47″N 1°41′10″W

Langsett Reservoir is in Yorkshire, England, near the villages of Langsett and Upper Midhope, on the edge of the Peak District National Park. The reservoir was constructed between 1898 and 1904, and is now managed by Yorkshire Water. Fed by the Little Don or Porter River, it is around a mile long, and supplies water for Sheffield and Barnsley via the Langsett Treatment Works.

Langsett Reservoir	
Location	South Yorkshire
Coordinates	53°29′47″N 1°41′10″W
Type	reservoir
Primary inflows	Little Don River
Catchment area	5,203 acres (2,106 ha)
Basin countries	United Kingdom
Surface area	51 ha (130 acres)
Max. depth	97 ft (30 m)
Water volume	1,408,000,000 imp gal (6.40 × 10⁹ L; 1.691 × 10⁹ US gal)

Seen from Hingcliff Common to the west

Construction and statistics [edit]

Construction of the reservoir began in 1898, the logistics of getting the workforce and materials to Langsett caused great difficulty as most of them came up from Sheffield. This involved a journey over four different stretches of railway line, the first leg was from Sheffield Midland Station to Deepcar, this was followed by a journey to Stocksbridge on the Samuel Fox and Company private line. From the Samuel Fox steelworks a new one mile long line was built to reach the Underbank Reservoir to join up with the Water Authority track up to Langsett Reservoir.[1]

The reservoir is 125 acres (51 ha) in area with a depth of 97 feet (30 m) and has a holding capacity of 1,408 million gallons, making it the largest of the water supply reservoirs in the immediate Sheffield district. The catchment area is the Langsett Moors to the west and this covers an area of 5,203 acres (2,106 ha). The embankment is 1,156 feet (352 m) long with a height of 117 feet (36 m) from the bottom of the old river bed. The embankment is 720 feet (220 m) wide at the bottom tapering to 36 feet (11 m) at the top and contains 900,000 cubic yards of infill in the puddle wall and concrete trench, making it one of the largest earth embankments in Great Britain.[1] The minor road (Midhope Cliff Lane) which runs across the embankment is thought to be the longest single carriageway of any reservoir in Great Britain. The embankment road has a sharp bend in it as it joins the A616 main road, this was a last minute change in construction plans, as keeping it straight would have meant the demolition of the Waggon and Horses public house. The reservoir was completed in 1904 when Alderman T.R. Gainsford closed the valve in the Langsett tower and the reservoir started to fill up, he was then presented with a golden key by the engineer William Watts.[1]

Local depopulation was used in the early part of the twentieth century to improve the water purity, and six farms were abandoned these included Brookhouse farm and North America farm, the last farmer left around 1907. The ruins of North America farm remain to the south-west of the reservoir even though it was used for target practice during the Second World War.[2] In 1962 conifers were planted around the reservoir as shown on the map, with the aim of providing a habitat for many species of indigenous wildlife. This plantation is called Langsett Woods. In recent years the woods have been restructured with most of the coniferous trees being felled and being replaced by oak and birch trees in an effort to create a new upland oak woodland. In 2007 a pond was created near Brookhouse Bridge at the western end of the reservoir to help dragonflies, frogs, newts and toads establish new colonies.[3] The Pennine peaks of Pike Lowe (478 m) and Hingcliff Common (358 m) lie to the south and south-west of the reservoir, respectively. The area is used for sheep farming and grouse shooting, and it is popular with walkers,[4] mountain bikers and birdwatchers with treecreepers, great spotted woodpeckers and red grouse to be seen in the vicinity.[5]

Green shading shows conifer plantation; red star indicates North America Farm

Ruins of North America Farm, with Langsett Reservoir behind

Langsett water treatment works [edit]

The present day Langsett water treatment works were built to replace the older works at Midhope and Langsett reservoirs. The older works used sand filter beds to treat the water, although the water was safe to drink, the sand filters had never been able to remove the brown colouration caused by rainwater falling on the surrounding peat moorland. After much complaining from consumers, it was decided in 1980 to build a completely new treatment works. South Yorkshire Water Authority gave permission in December 1981 to build the works in an old quarry adjacent to the reservoir wall. Worked started in 1983 with water from the new works going into the supply system in July 1986. The works clarify the water by the addition of chemicals before the filtration stage to bring it up to the latest EU standards. The works can produce 60,000 cubic metres (60 million litres) of water per day.[1][5]

In 2017 a £20 million scheme was announced by Yorkshire Water to upgrade the treatment works to further improve the discolouration and remove deposits from the raw water collected from the moors around the reservoir. The work began in September 2017 and will take three years to complete.[6]

Recreation [edit]

The Peak District Boundary Walk runs along the north side of the reservoir and across the dam.[7]

References [edit]

1. ^ a b c d "History Of Langsett", Jack Branston, (1986) Gives details of construction and reservoir statistics.
2. ^ Information boards at reservoir give details of farms.
3. ^ Information boards at reservoir give details of wildlife and woodland.
4. ^ "Langsett Reservoir Walk". Walks in Yorkshire. Retrieved 1 December 2017.
5. ^ a b Peak District Education Archived 1 September 2010 at the Wayback Machine Gives details of reservoir and treatment works.
6. ^ Yorkshire Water Details improvement to treatment works started in 2017.
7. ^ McCloy, Andrew (2017). *Peak District Boundary Walk: 190 Miles Around the Edge of the National Park*. Friends of the Peak District. ISBN 978-1909461536.

External links [edit]

- Yorkshire Water – Langsett Reservoir

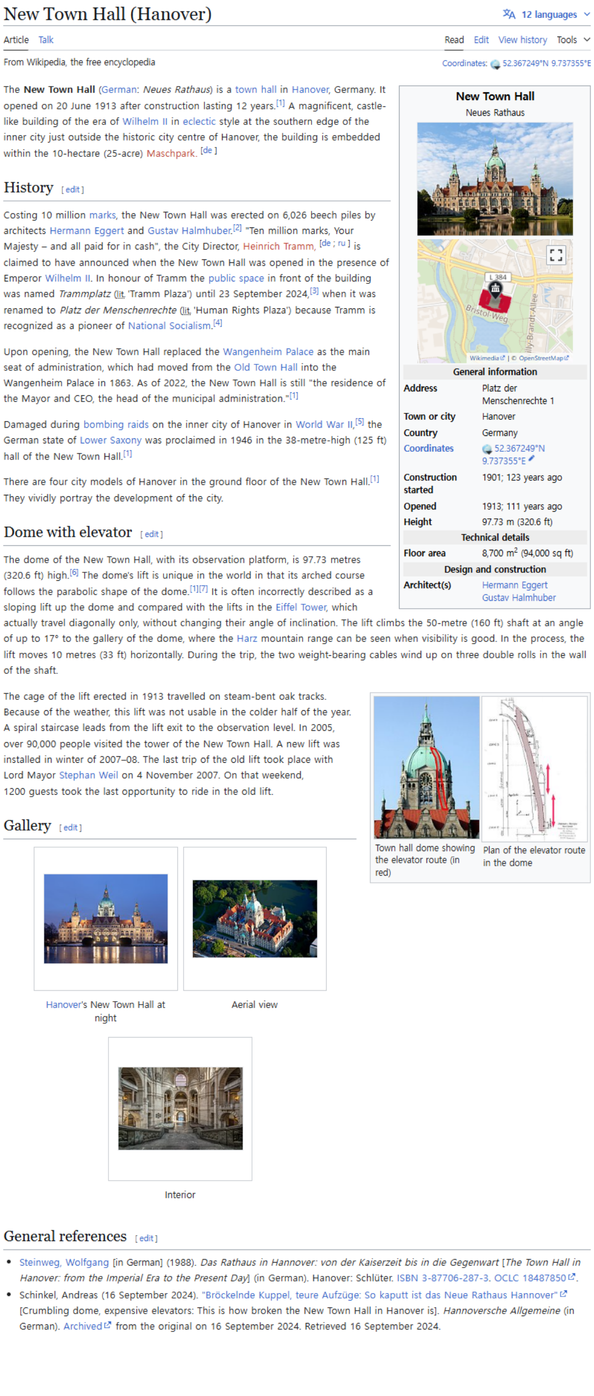
Wikimedia Commons has media related to *Langsett Reservoir*.

Reservoirs in Yorkshire	
North Yorkshire	Angram · Beaver Dyke · Chelker · Cod Beck · Embsay · Fewston · Gouthwaite · Grimwith · Leighton · Lindley Wood · Roundhill · Scaling Dam · Scar House · Swinsty · Thornton Steward · Thruscross · Upper Barden · Winterburn
South Yorkshire	Agden · Broomhead · Dale Dike · Damflask · Howden · **Langsett** · More Hall · Redmires · Rivelin · Strines · Ulley · Winscar

(a) Interleaved Multimodal Document Retrieval **(b) Text-only Document Retrieval**

Figure 5: Retrieved documents across different document formats for document retrieval with a given textual query. **(a)**: A document retrieved when represented leveraging interleaved multimodal contents within documents (ours). **(b)**: A document retrieved when using only textual format

Q: Who designed this building?



New Town Hall (Hanover)

12 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

Coordinates: 52.367249°N 9.737355°E

The **New Town Hall** (German: *Neues Rathaus*) is a town hall in Hanover, Germany. It opened on 20 June 1913 after construction lasting 12 years.[1] A magnificent, castle-like building of the era of Wilhelm II in eclectic style at the southern edge of the inner city just outside the historic city centre of Hanover, the building is embedded within the 10-hectare (25-acre) Maschpark. [de]

New Town Hall	
Neues Rathaus	
General information	
Address	Platz der Menschenrechte 1
Town or city	Hanover
Country	Germany
Coordinates	52.367249°N 9.737355°E
Construction started	1901; 123 years ago
Opened	1913; 111 years ago
Height	97.73 m (320.6 ft)
Technical details	
Floor area	8,700 m² (94,000 sq ft)
Design and construction	
Architect(s)	Hermann Eggert Gustav Halmhuber

History [edit]

Costing 10 million marks, the New Town Hall was erected on 6,026 beech piles by architects Hermann Eggert and Gustav Halmhuber.[2] "Ten million marks, Your Majesty – and all paid for in cash", the City Director, Heinrich Tramm, [de; ru] is claimed to have announced when the New Town Hall was opened in the presence of Emperor Wilhelm II. In honour of Tramm the public space in front of the building was named *Trammplatz* (lit. 'Tramm Plaza') until 23 September 2024,[3] when it was renamed to *Platz der Menschenrechte* (lit. 'Human Rights Plaza') because Tramm is recognized as a pioneer of National Socialism.[4]

Upon opening, the New Town Hall replaced the Wangenheim Palace as the main seat of administration, which had moved from the Old Town Hall into the Wangenheim Palace in 1863. As of 2022, the New Town Hall is still "the residence of the Mayor and CEO, the head of the municipal administration."[1]

Damaged during bombing raids on the inner city of Hanover in World War II,[5] the German state of Lower Saxony was proclaimed in 1946 in the 38-metre-high (125 ft) hall of the New Town Hall.[1]

There are four city models of Hanover in the ground floor of the New Town Hall.[1] They vividly portray the development of the city.

Dome with elevator [edit]

The dome of the New Town Hall, with its observation platform, is 97.73 metres (320.6 ft) high.[6] The dome's lift is unique in the world in that its arched course follows the parabolic shape of the dome.[1][7] It is often incorrectly described as a sloping lift up the dome and compared with the lifts in the Eiffel Tower, which actually travel diagonally only, without changing their angle of inclination. The lift climbs the 50-metre (160 ft) shaft at an angle of up to 17° to the gallery of the dome, where the Harz mountain range can be seen when visibility is good. In the process, the lift moves 10 metres (33 ft) horizontally. During the trip, the two weight-bearing cables wind up on three double rolls in the wall of the shaft.

The cage of the lift erected in 1913 travelled on steam-bent oak tracks. Because of the weather, this lift was not usable in the colder half of the year. A spiral staircase leads from the lift exit to the observation level. In 2005, over 90,000 people visited the tower of the New Town Hall. A new lift was installed in winter of 2007–08. The last trip of the old lift took place with Lord Mayor Stephan Weil on 4 November 2007. On that weekend, 1200 guests took the last opportunity to ride in the old lift.

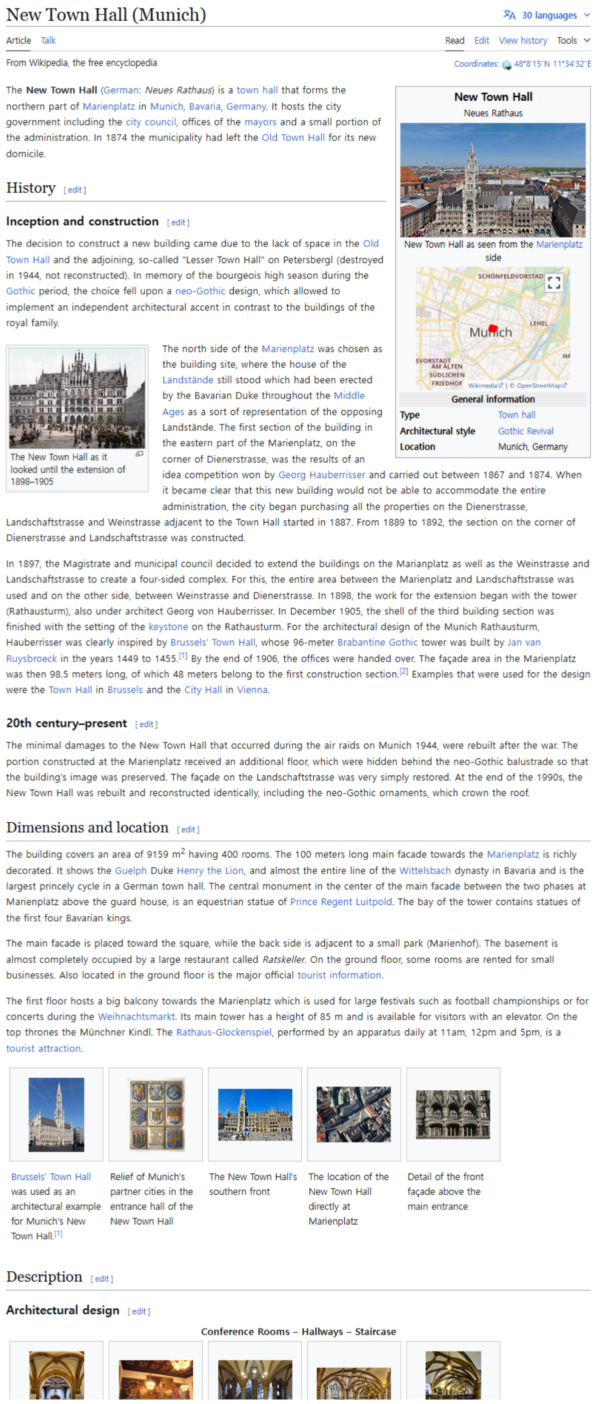
Town hall dome showing the elevator route (in red) | Plan of the elevator route in the dome

Gallery [edit]

Hanover's New Town Hall at night | Aerial view | Interior

General references [edit]

- Steinweg, Wolfgang (in German] (1988). *Das Rathaus in Hannover: von der Kaiserzeit bis in die Gegenwart* [*The Town Hall in Hanover: from the Imperial Era to the Present Day*] (in German). Hanover: Schlüter. ISBN 3-87706-287-3. OCLC 18487850.
- Schinkel, Andreas (16 September 2024). "Bröckelnde Kuppel, teure Aufzüge: So kaputt ist das Neue Rathaus Hannover" [Crumbling dome, expensive elevators: This is how broken the New Town Hall in Hanover is]. *Hannoversche Allgemeine* (in German). Archived from the original on 16 September 2024. Retrieved 16 September 2024.



New Town Hall (Munich)

30 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

Coordinates: 48°8′15″N 11°34′32″E

The **New Town Hall** (German: *Neues Rathaus*) is a town hall that forms the northern part of Marienplatz in Munich, Bavaria, Germany. It hosts the city government including the city council, offices of the mayors and a small portion of the administration. In 1874 the municipality had left the Old Town Hall for its new domicile.

New Town Hall	
Neues Rathaus	
New Town Hall as seen from the Marienplatz side	
General information	
Type	Town hall
Architectural style	Gothic Revival
Location	Munich, Germany

History [edit]

Inception and construction [edit]

The decision to construct a new building came due to the lack of space in the Old Town Hall and the adjoining, so-called "Lesser Town Hall" on Petersbergl (destroyed in 1944, not reconstructed). In memory of the bourgeois high season during the Gothic period, the choice fell upon a neo-Gothic design, which allowed to implement an independent architectural accent in contrast to the buildings of the royal family.

The New Town Hall as it looked until the extension of 1898–1905

The north side of the Marienplatz was chosen as the building site, where the house of the Landstände still stood which had been erected by the Bavarian Duke throughout the Middle Ages as a sort of representation of the opposing Landstände. The first section of the building in the eastern part of the Marienplatz, on the corner of Dienerstrasse, was the results of an idea competition won by Georg Hauberrisser and carried out between 1867 and 1874. When it became clear that this new building would not be able to accommodate the entire administration, the city began purchasing all the properties on the Dienerstrasse, Landschaftstrasse and Weinstrasse adjacent to the Town Hall started in 1887. From 1889 to 1892, the section on the corner of Dienerstrasse and Landschaftstrasse was constructed.

In 1897, the Magistrate and municipal council decided to extend the buildings on the Marianplatz as well as the Weinstrasse and Landschaftstrasse to create a four-sided complex. For this, the entire area between the Marienplatz and Landschaftstrasse was used and on the other side, between Weinstrasse and Dienerstrasse. In 1898, the work for the extension began with the tower (Rathausturm), also under architect Georg von Hauberrisser. In December 1905, the shell of the third building section was finished with the setting of the keystone on the Rathausturm. For the architectural design of the Munich Rathausturm, Hauberrisser was clearly inspired by Brussels' Town Hall, whose 96-meter Brabantine Gothic tower was built by Jan van Ruysbroeck in the years 1449 to 1455.[1] By the end of 1906, the offices were handed over. The façade area in the Marienplatz was then 98.5 meters long, of which 48 meters belong to the first construction section.[2] Examples that were used for the design were the Town Hall in Brussels and the City Hall in Vienna.

20th century–present [edit]

The minimal damages to the New Town Hall that occurred during the air raids on Munich 1944, were rebuilt after the war. The portion constructed at the Marienplatz received an additional floor, which were hidden behind the neo-Gothic balustrade so that the building's image was preserved. The façade on the Landschaftstrasse was very simply restored. At the end of the 1990s, the New Town Hall was rebuilt and reconstructed identically, including the neo-Gothic ornaments, which crown the roof.

Dimensions and location [edit]

The building covers an area of 9159 m² having 400 rooms. The 100 meters long main facade towards the Marienplatz is richly decorated. It shows the Guelph Duke Henry the Lion, and almost the entire line of the Wittelsbach dynasty in Bavaria and is the largest princely cycle in a German town hall. The central monument in the center of the main facade between the two phases at Marienplatz above the guard house, is an equestrian statue of Prince Regent Luitpold. The bay of the tower contains statues of the first four Bavarian kings.

The main facade is placed toward the square, while the back side is adjacent to a small park (Marienhof). The basement is almost completely occupied by a large restaurant called *Ratskeller*. On the ground floor, some rooms are rented for small businesses. Also located in the ground floor is the major official tourist information.

The first floor hosts a big balcony towards the Marienplatz which is used for large festivals such as football championships or for concerts during the Weihnachtsmarkt. Its main tower has a height of 85 m and is available for visitors with an elevator. On the top thrones the Münchner Kindl. The Rathaus-Glockenspiel, performed by an apparatus daily at 11am, 12pm and 5pm, is a tourist attraction.

Brussels' Town Hall was used as an architectural example for Munich's New Town Hall.[1] | Relief of Munich's partner cities in the entrance hall of the New Town Hall | The New Town Hall's southern front | The location of the New Town Hall directly at Marienplatz | Detail of the front façade above the main entrance

Description [edit]

Architectural design [edit]

Conference Rooms – Hallways – Staircase

(a) Interleaved Multimodal Document Retrieval **(b) Text-only Document Retrieval**

Figure 6: Retrieved documents across different document formats for document retrieval with a given multimodal query. **(a)**: A document retrieved when represented leveraging interleaved multimodal contents within documents (ours). **(b)**: A document retrieved when using only textual format