2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

An Enhanced Method of Multimodal Information Retrieval Based on Document Segmentation

Jintao Liu[1], Chen Feng[1], Guang Jin[1], Jun Fan[1,✉]

[1] College of Systems Engineering, National University of Defense Technology, Changsha 410073
fanjun891009@163.com

Abstract. With the rapid development of information technology, emerging technologies, typically represented by large language models(LLMs), are powerfully driving profound transformations in multiple industries. However, the LLMs may exhibit the hallucination phenomenon, making it difficult for them to accurately understand and effectively apply relevant industry knowledge in some specific fields. To address this issue, a method called DocColQwen, a multimodal information retrieval enhancement method based on document segmentation in this paper is proposed. First, the large model analyses the user task and then the contextualized late interaction over paligemma 's (Colpalli) idea is used to segment the multimodal experimental document into images. Subsequently, the images and user questions are transformed into vectors through encoding for matching, and the retrieved documents are passed to the Qwen2-VL model for response output. Finally, the method is verified in multimodal experimental documents to validate its effectiveness, providing a solution idea for the analysis and processing of multimodal test documents.
Keywords: Large language model; Multimodality; Vector matching; Retrieval-augmented generation

Keywords: Large language models, Multimodality, Vector matching, Retrieval-augmented generation

1 Introduction

The advent of Large Language Models (LLMs) has instilled novel impetus within the realm of knowledge-based question-answering systems, precipitating substantial transformations. Leveraging their vast parameter magnitudes and profound learning from extensive text corpora, LLMs have manifested exceptional prowess across diverse tasks, including natural language com-prehension, text generation, and dialogue systems [1]. Nonetheless, within specialized domains, such as industrial fault diagnosis and test data analysis, these LLMs may encounter the phenomenon of hallucination. Additionally, LLMs' reliance on static training data gives rise to issues of information obsolescence and circumscribed context awareness. This static characteristic impedes their proficiency in delivering precise and current responses, particularly in dynamic or rapidly evolving scenarios, where ensuring the accuracy and reliability of information becomes a formidable challenge [2].

Retrieval-augmented generation (RAG), as an emerging and highly promising technological approach, enhances the performance of models by introducing external knowledge retrieval, demonstrating unique value in multiple fields such as natural language processing and information retrieval [3]. Benjamin trained bio-clinical models using unstructured and structured data respectively, and constructed an integrated model to reflect machine learning performance, the experiments verified that after introducing RAG and machine-learning probabilities, the performance was significantly improved [4]. Qi et al. constructed an RAG system using the open-source LLaMa-2-7b to generate corresponding LLM responses. By comparing the responses of health coaches and LLMs to sleep-related questions, it was found that both experts and GPT-4 considered the quality of the two to be comparable [5]. Xia et al. created a physical model of bearings by integrating digital twin technology to repair abnormal sensor data. The RAG-assisted LLM generated virtual datasets when sensor failures occurred, and federated continuous learning was adopted to improve the training of the global model in a distributed industrial environment, thereby ensuring data privacy while improving the accuracy of bearing fault diagnosis [6]. Aiming at the problems of data scarcity, high training costs, and catastrophic forgetting in the field of autonomous driving, Yuan et al. achieved efficient explanation generation and control prediction in autonomous driving through in-context learning combined with RAG technology and demonstrated excellent zero-shot generalization ability [7].

Aiming at the issues of the gap between vector similarity and knowledge reasoning relevance in RAG, the advantages of knowledge graphs (KG) and vector retrieval were made full use of by Lei et al. By introducing knowledge graphs into LLMs, the knowledge-logic reasoning ability of LLMs was enhanced [8]. An interpretable road-user behavior prediction system that integrates KG reasoning with the expressiveness of large language models was developed by Mohamed et al. using RAG technology. This was achieved by combining KG embed-ding and Bayesian reasoning based on historical information in the knowledge graph and real-time data from vehicle-mounted sensors [9]. A high-performance LLMs framework focused on specific domains was proposed by Barron et al. By preserving relevant sub-graphs of information in a structured manner, the accuracy of question-answer pairs in RAG can be enhanced by the KG. The experiments show that in this framework, a very restrictive role is played in reducing hallucinations in LLMs and the need for fine-tuning [10]. A novel RAG method was proposed by Matsumoto et al. by converting the KG into a vector database and using RAG to retrieve relevant facts from it, the structured data retrieval and the powerful nuanced understanding ability of LLMs were fully utilized, which ensures that the model can continuously adapt to user expectations and improve its performance and applicability enhanced [11].

However, the single-modal data processing can no longer meet the increasingly complex task requirements. With the rapid development of information technology, data has shown multi-modal characteristics, and multi-modal retrieval enhancement has become a popular and important research field [12]. For a long time, most RAG systems rely on retrieving text fragments from large knowledge bases and then using generative models to expand their response content. When faced with complex documents, especially those containing a large number of visual elements, it is difficult to analyze them

effectively. The multimodal large language models(MLLMs) can handle multiple types of data simultaneously and have demonstrated great potential in fields such as content understanding, generation, and recommendation. Applying retrieval enhancement to MLLMs not only improves the model's ability to understand data but also enhances its practicality in real-world application scenarios. Su et al. proposed a hybrid RAG-empowered medical MLLMs framework, which ensures data security and improves output quality through a hierarchical cross-chain architecture. At the same time, it uses multi-modal metrics to optimize retrieval results and introduces the theory of information age and contract theory to promote the sharing of fresh data, addressing the key challenges of data security and freshness in the healthcare field [13]. Lin et al. drew inspiration from RAG and designed a visual prompt-based multi-modal question-answering method, integrating fine-grained external knowledge collected from specialized visual models into MLLMs [14]. Li et al. enhanced the adaptability to applications such as parsers, text, and visual descriptions by constructing a flexible action space and using RAG technology to record the functions of user interface elements in a customized knowledge base during the exploration and deployment phases respectively and efficiently retrieve updates, demonstrating precision and flexibility in handling complex multi-step operations and customized task workflows [15]. Xue et al. designed RAG by introducing a structured scene graph, which greatly improved the accurate recognition ability of objects in the image, accurately analyzed the semantic relationship between objects, and deeply understood the spatial topology of the scene, so as to provide more effective solutions for the analysis and processing of complex scenes [16]. However, the existing retrieval enhancement methods based on MLLMs require parsing the entire document, which not only increases the retrieval time but also causes certain interference in the target positioning of the retrieved results [17]. The technology of Qwen2-VL provides new ideas and methods for solving the cross-modal problem of vision and language. By mapping image features and text features to the same embedding space, Qwen2-VL can effectively capture the association between them, thus supporting more complex tasks [18].The birth of the contextualized late interaction over paligemma (ColPali) technology carries out document retrieval by directly embedding the image of the document page. It can generate high-quality multi-vector embedding from the image of the document page, which promotes the development of retrieval enhancement technology towards the field of multimodal image analysis [19]. However, in the field of test data, there are many documents, professional content, strong data privacy, and high requirements for the accuracy of the retrieval results. The existing model cannot meet the real-time fast query and has the problem of inaccurate retrieval positioning. In this paper, by integrating the vision-language model Qwen2-VL and using the contextualized late interaction over paligemma (ColPali) technology to generate embedding vectors from the images of document pages, a novel multi-modal experimental document detection model, the DocColqwen model, is proposed. This enables the system to not only parse text but also analyze and understand the visual information in the document. DocColqwen skillfully combines the advantages of visual document retrieval technology and vision-language models, enabling it to respond to user queries in a more comprehensive and accurate manner. This experimental document retrieval system can process image documents, generate corresponding

embedding vectors, retrieve the most relevant con-tent based on these vectors, and then provide comprehensive and accurate answers.

(1) The DocColQwen breaks through the traditional document parsing paradigm that requires OCR extraction first, and greatly improves the retrieval speed.

(2) Through the construction of the "input-coding-retrieval-generation" whole process automation system, the advantages of visual document segmentation and visual language model technology are organically integrated to realize the joint modelling of graphic semantic space, and significantly improve the understanding depth of complex layout documents.

(3) As far as we know, this paper is the first to use the visual retrieval method of document segmentation in the field of equipment test documents, which provides a reference scheme for complex equipment test documents to provide fast, effective and accurate retrieval, reduce manual operation and improve decision-making efficiency.

2 Introduction of the method

2.1 Large language models(LLMs)

Due to the emergence of deep learning, especially the transformer architecture, breakthroughs have been achieved in various fields such as semantic segmentation. In this paper, the segmented documents are recognized and reasoned by Qwen2-VL. Images with different resolutions and aspect ratios can be understood by it. The high flexibility and adaptability are demonstrated when it deals with such images. Globally leading performance has been obtained in multiple visual understanding benchmark tests. As shown in the following Fig.1, it is the network structure diagram of Qwen2-VL[18].

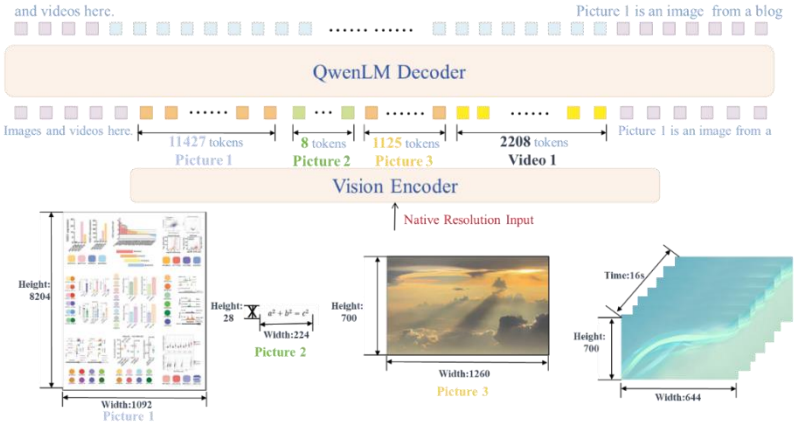


Fig. 1. The network architecture of the Qwen2-VL

From the above Fig.1, it can be seen that the input image size of Qwen2-VL can be independently defined, and after passing through the vision encoder, it becomes a

sequence of tokens to be added to the QwenLM decoder. The core structure of vision encoder is vision transformer (VIT)[20], as shown in the following Fig.2.

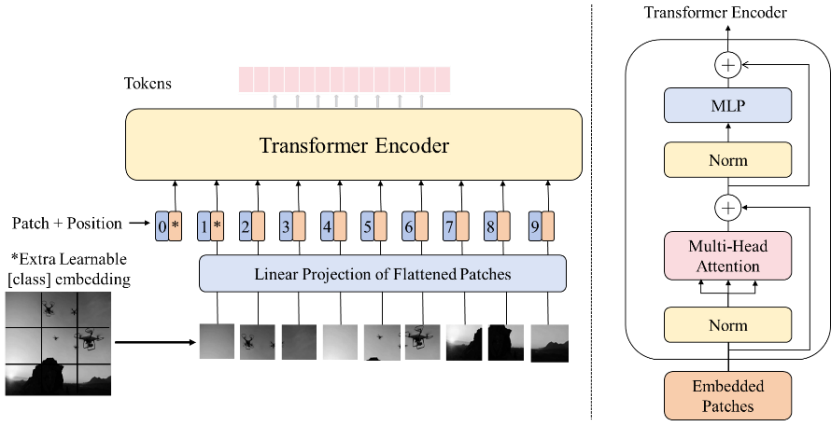


Fig.2. The network architecture of the VIT

For each input 2D image I $\in R^{H*W*C}$ in visual encoding, which, (H, W) is the resolution of the input image and C is the number of channels, VIT reshapes it into a series of tiled 2D image patches $P \in R^{N*(P^2*C)}$. which, (P, P) is the resolution of the image and N $= \frac{HW}{P^2}$ is the number of image patches. The VIT adds a [class] token in front of the image sequence and uses a learnable ID positional embedding $V_{pos} \in R^{(N+1)*D_v}$. Which the D_v is the dimension of visual encoding. The input visual representation can be calculated as follows.

$$V_0 = [E_{[class]}; p_1 W_p; \ldots; p_N W_p] + V_{pos} \quad (1)$$

Which, $W_p \in R^{N*(P^2*C)}$ is a trainable linear projection layer and $V_0 \in R^{(N+1)*D_v}$. Simplifying it as $Encoder^V$, the $l-th$ layer can be expressed as Eq.(2).

$$V_l = Encoder_l^V(V_{l-1}), l = 1,2 \ldots \ldots L_V \quad (2)$$

Which, L_V is the number of layers of visual encoding. Specifically, the VIT extracts the image into N non-overlapping patches (patch), and the patch is represented as $x_i \in \mathbb{R}^{h \times w \times c}$. Then, these patches are linearly projected into a token sequence. The QwenLM decoder is responsible for generating the target sequence based on the context information provided by the encoder. It predicts the output sequence word by word through a multi-layer self-attention mechanism and an encoder-decoder attention mechanism.

2.2 Contextualized Late Interaction over PaliGemma (ColPali)

The ColPali is an advanced visual retrieval model that seamlessly integrates visual elements with text information. By incorporating visual language models, it enhances the efficiency and accuracy of document retrieval, thus significantly improving retrieval performance. Through multi - vector embedding transformation of document pages, the ColPali can not only effectively capture the visual features in the documents

but also fully consider their text attributes, ensuring both speed and precision in the retrieval process. Its structure is shown in Fig.3 below.

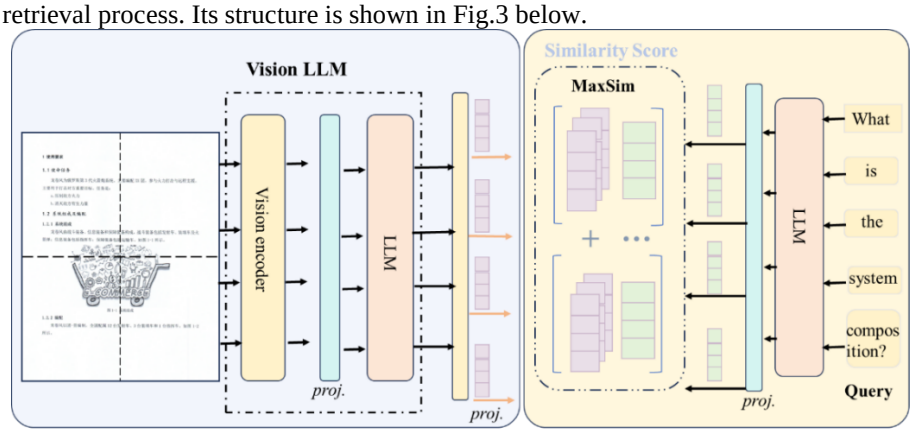


Fig.3. The network architecture of the ColPali

The ColPali model utilizes visual language models to construct efficient and diverse vector embeddings in visual space, significantly enhancing the accuracy and efficiency of document retrieval. By processing the output block of the vision encoder through a linear projection layer, multi-dimensional vectorized expression of document content is achieved. The natural language of the que3ry is projected onto a linear space using LLMs, and then focused on improving the similarity score between document embedding and query embedding to achieve the goal of optimizing retrieval performance.

2.3 The Multi modal information retrieval method based on document segmentation(DocColQwen)

A novel multimodal retrieval method called DocColQwen in this paper, which integrated Qwen2-VL with Colpali, addresses the problem that existing multimodal document retrieval methods cannot effectively locate the source of documents and are difficult to perform joint query retrieval of multiple documents. The specific implementation process is shown in the following Fig.4.

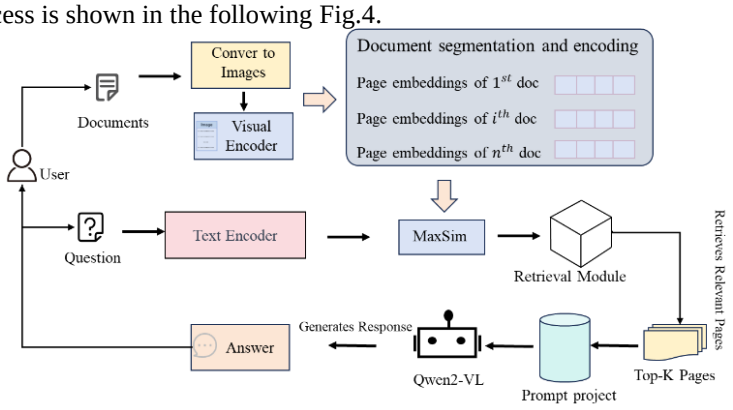


Fig.4. The flowchart of the DocColQwen

The method proposed in this article mainly consists of three steps.

Step 1: Document embedding vectorization stage. This stage mainly involves inputting the original document into the system to convert the page document into block-based image formats and then extracting visual information from the image for feature representation.

Step 2: Document retrieval and query phase. The main task of this stage is to utilize the maximum similarity matching algorithm to efficiently search for the top K pages with high similarity to user input text queries in large-scale document collections, providing more accurate and fast page information support for subsequent Q&A.

Step 3: Q&A stage. In this stage, the Qwen2-VL will be used for visual Q&A. Using the visual information extracted in the early stage and the retrieved page information, generate corresponding masks based on the user's question, and then predict these masks to obtain the final answer.

3 Case study

In this study, the proposed scheme was comprehensively verified using experimental data documents with privacy removed and publicly available data documents. The verification process was strictly carried out on a professional computer equipped with an RTX 4090 high-performance graphics card in an orderly manner under the Linux operating system environment. This fully ensured a high degree of efficiency and stability in the calculation process, effectively avoiding calculation errors and operational interruptions caused by insufficient hardware performance or system compatibility issues.

The ROUGE-L, and BERTScore are used in this work to measure the performance of the DocColQwen model.

$$ROUGE - L = \frac{(1+\beta^2)R_{LCS}P_{LCS}}{R_{LCS}+\beta^2 P_{LCS}} \tag{3}$$

Where, $R_{LCS} = \frac{LCS(C,S)}{\text{len }(S)}$ and $P_{LCS} = \frac{LCS(C,S)}{\text{len }(C)}$ represent the recall rate and the precision rate respectively. $LCS(C,S)$ is the length of the longest common subsequence.

$$\text{BERTScore} = 2 * \frac{P_{BERT} * R_{BERT}}{P_{BERT}+R_{BERT}} \tag{4}$$

Where, $P_{BERT} = \frac{1}{|\hat{x}|}\sum \max\left(sim(x_i, \hat{x}_j)\right)$ and $R_{BERT} = \frac{1}{|x|}\sum \max\left(sim(x_i, \hat{x}_j)\right)$,which are obtained through similarity calculation.

Given that data privacy and security are bound to be key concerns after the paper is publicly published, this paper carefully selected publicly available research paper documents for testing. On the one hand, it significantly enhanced the transparency of the research process and results, enabling other researchers to clearly trace the data sources and analysis steps, thus greatly improving the reproducibility of the research. On the other hand, it strictly complied with relevant international and domestic regulations in the field of data protection, fully demonstrating the author's team's profound understanding and respect for data ethics.

Table 1. The comparison between different algorithms

Model	ROUGE-L	BERTScore
Llama-3.2	0.222	0.654
Qwen-VL	0.480	0.636
DocColQwen	0.632	0.712

As shown in the table above, the comparison between different algorithms shows that DocColQwen can match the retrieved content better than other large models. As can be seen from the figure below, when the user asks "What are the components of Tornado?". The DocColQwen proposed in this paper can explain "Tornado" instead of simply translating its literal meaning.

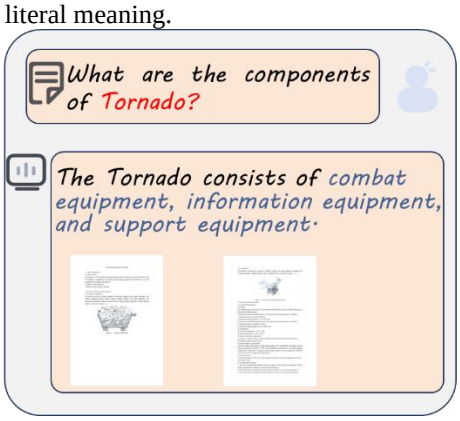


Fig.5. The test results of the test documents(a)

In addition, the DocColQwen will retrieve two images that best match the content of the text, and summarize based on the best-matched images. Compared with traditional methods, it can provide users with document location and reduce the time they need to search for documents. The content answered by DocColQwen comes from the red box in the following Fig.6.

General Requirements for Tornado

1. Usage Requirements
1.1 Mission Tasks
The Tornado is the third-generation rocket launcher system of Russia. It is mainly allocated to the XX Regiment to participate in fire strikes and long-range support. Its main purpose is to strike important enemy targets. The tasks are:
a. Suppress enemy firepower.
b. Eliminate enemy effective strength.

1.2 System Composition and Allocation
1.2.1 System Composition
The Tornado consists of combat equipment, information equipment, and support equipment. The combat equipment includes launch vehicles, loading vehicles, and rocket projectiles. The information equipment includes command vehicles, and the support equipment includes transport vehicles. As shown in Figure 1 - 1.

Figure 1 – 1 system composition

Fig.6. The location image matched by the query(b)

In order to further explore DocColQwen's ability to understand document details on the same document, the results of questioning the document details is shown in Fig.7.

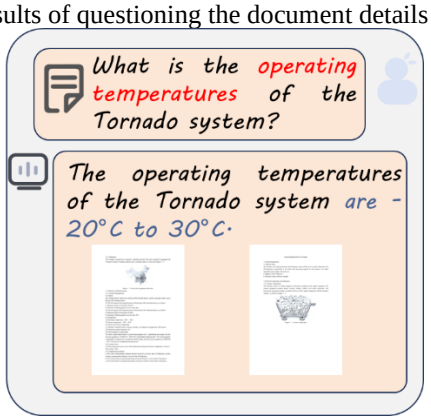
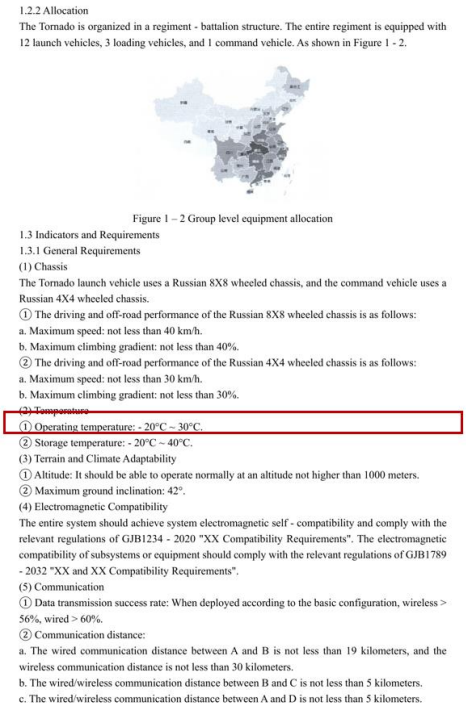


Fig.7. The test results of the test documents(a)

The image document with the highest matching degree is shown in Fig.8.



1.2.2 Allocation

The Tornado is organized in a regiment - battalion structure. The entire regiment is equipped with 12 launch vehicles, 3 loading vehicles, and 1 command vehicle. As shown in Figure 1 - 2.

Figure 1 – 2 Group level equipment allocation

1.3 Indicators and Requirements

1.3.1 General Requirements

(1) Chassis

The Tornado launch vehicle uses a Russian 8X8 wheeled chassis, and the command vehicle uses a Russian 4X4 wheeled chassis.

① The driving and off-road performance of the Russian 8X8 wheeled chassis is as follows:

a. Maximum speed: not less than 40 km/h.

b. Maximum climbing gradient: not less than 40%.

② The driving and off-road performance of the Russian 4X4 wheeled chassis is as follows:

a. Maximum speed: not less than 30 km/h.

b. Maximum climbing gradient: not less than 30%.

(2) Temperature

① Operating temperature: - 20°C ~ 30°C.

② Storage temperature: - 20°C ~ 40°C.

(3) Terrain and Climate Adaptability

① Altitude: It should be able to operate normally at an altitude not higher than 1000 meters.

② Maximum ground inclination: 42°.

(4) Electromagnetic Compatibility

The entire system should achieve system electromagnetic self - compatibility and comply with the relevant regulations of GJB1234 - 2020 "XX Compatibility Requirements". The electromagnetic compatibility of subsystems or equipment should comply with the relevant regulations of GJB1789 - 2032 "XX and XX Compatibility Requirements".

(5) Communication

① Data transmission success rate: When deployed according to the basic configuration, wireless > 56%, wired > 60%.

② Communication distance:

a. The wired communication distance between A and B is not less than 19 kilometers, and the wireless communication distance is not less than 30 kilometers.

b. The wired/wireless communication distance between B and C is not less than 5 kilometers.

c. The wired/wireless communication distance between A and D is not less than 5 kilometers.

Fig.8. The location image matched by the query(b)

Of course, the existing large models can be used for retrieval enhancement, such as the results obtained by Alibaba's Tongyi Qianwen when operating on the same document.

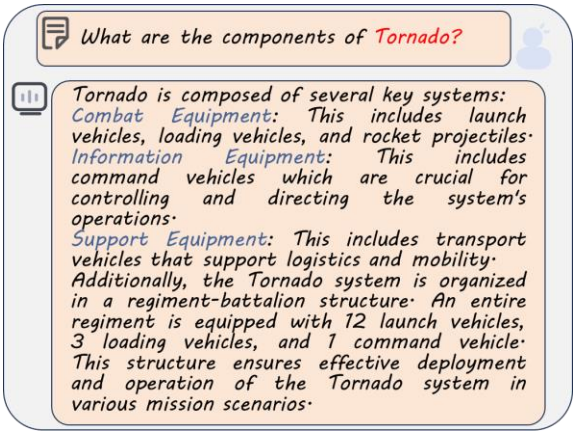


Fig.9. The same question‘s answer of Tongyi Qianwen

It can be seen that the answers to Tongyi Qianwen look more comprehensive and detailed. However, unlike traditional open source models, DocColQwen not only can quickly segment and match documents to understand user meaning, but also can recommend the most matching images to help users locate their positions, which helps users to under-stand and summarize them for inspection, and further reduces the losses caused by the illusion of LLM. It has certain advantages in offline conditions and scenarios with high data privacy.

4 Case study

In the current era of rapid development of information technology, cutting-edge technologies represented by LLMs are driving numerous industrial transformations. However, in highly specialized scientific research fields, when dealing with cutting-edge and complex theoretical knowledge and experimental data, it is easy to encounter misunderstandings, logical confusion, and the inability to accurately apply relevant knowledge, which restricts its application in key areas. To solve this problem, a multimodal retrieval enhancement method based on document segmentation is proposed in this paper, which integrates multimodal processing and retrieval strategies to enhance the knowledge understanding and application ability of large language models in specific fields. Firstly, the Colpali technology is used to finely segment multimodal documents, convert them into structured image data, preserve key information, and separate text and image information. Next, encoding algorithms are used to convert the images and user questions into vectors, and a vector space model is used to retrieve relevant document information. Finally, the retrieved documents were input into the Qwen2-VL model for analysis and understanding of user tasks. The results showed that the DocColQwen method had significant advantages and provided a new solution for multimodal document analysis and processing, helping to expand the application of large language models in specific fields.

References

1. Y. Chi et al., “SELA: Tree-search enhanced LLM agents for automated machine learning.” 2024. [Online].
2. S. Lee, H. Hsu, and C.-F. Chen, “LLM hallucination reasoning with zero-shot knowledge test.” 2024. [Online].
3. C. Li and J. Flanigan, “RAC: Efficient LLM factuality correction with retrieval augmentation.” 2024. [Online].
4. B. S. Glicksberg et al., “Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room,” Journal of the American Medical Informatics Association, vol. 31, no. 9, pp. 1921–1928, May 2024.
5. Q. C. Ong et al., “Advancing health coaching: A comparative study of large language model and health coaches,” Artificial Intelligence in Medicine, vol. 157, p. 103004, 2024.
6. Y. Xia et al., “FCLLM-DT: Enpowering federated continual learning with large language models for digital twin-based industrial IoT,” IEEE Internet of Things Journal, pp. 1–1, 2024.

7. J. Yuan et al., “RAG-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model.” 2024. [Online].
8. L. Liang et al., “KAG: Boosting LLMs in professional domains via knowledge augmented generation.” 2024. [Online].
9. M. M. Hussien, A. N. Melo, A. L. Ballardini, C. S. Maldonado, R. Izquierdo, and M. Á. Sotelo, “RAG-based explainable pre-diction of road users behaviors for automated driving using knowledge graphs and large language models,” Expert Systems with Applications, vol. 265, p. 125914, 2025.
10. R. C. Barron et al., “Domain-specific retrieval-augmented generation using vector stores, knowledge graphs, and tensor factorization.” 2024. [Online].
11. N. Matsumoto et al., “KRAGEN: A knowledge graph-enhanced RAG framework for biomedical problem solving using large language models.,” Bioinformatics, vol. 40, no. 6, pp. 1–4, 2024.
12. S.-C. Lin, C. Lee, M. Shoeybi, J. Lin, B. Catanzaro, and W. Ping, “MM-embed: Universal multimodal retrieval with multi-modal LLMs.” 2024. [Online].
13. C. Su et al., “Hybrid RAG-empowered multi-modal LLM for secure data management in internet of medical things: A diffusion-based contract approach.” 2024. [Online].
14. Y. Lin et al., “Rethinking visual prompting for multimodal large language models with external knowledge.” 2024. [Online].
15. Y. Li et al., “AppAgent v2: Advanced agent for flexible mobile interactions.” 2024. [Online].
16. J. Xue, Q. Deng, F. Yu, Y. Wang, J. Wang, and Y. Li, “Enhanced multimodal RAG-LLM for accurate visual question answering.” 2024. [Online].
17. S. Kawashima, S. Shiramatsu, and T. Mizumoto, “Development of RAG system for digital transformation of local government and considering optimal document-segmentation methods,” *Lecture Notes in Networks and Systems*, pp. 603–617, 2024.
18. P. Wang et al., “Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution,” arXiv preprint arXiv:2409.12191, 2024.
19. M. Faysse et al., “ColPali: Efficient document retrieval with vision language models.” 2024. [Online].
20. A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021.