Implementation and Analysis of a Multimodal RAG System for Iranian Tourism

Amir Malekhosseini, Amin Fadaee
Sharif University of Technology

September 13, 2025

Abstract

This report presents the results of a comprehensive project on the implementation and evaluation of a Multimodal Retrieval-Augmented Generation (RAG) system. The project's goal was to build an intelligent system to answer both text and image-based queries in the specialized domain of Iranian tourism. The system architecture comprises two main components: a powerful Retriever, utilizing separate models for text and image processing, and a Generator based on the `Qwen/Qwen2.5-VL-7B-Instruct` large multimodal language model. Both retriever models—the text-based `glot500-base` and the image-based `CLIP`—were fine-tuned on the project's custom dataset using Contrastive Learning. Final evaluation demonstrated that the RAG system, with an accuracy of 76.25%, significantly outperformed the baseline model (28.75%). This result clearly validates the critical importance of augmenting large language models with retrieved information and the dramatic impact of domain-specific fine-tuning.

Key Insights: The project's findings underscore the system's heavy reliance on data quality, the pivotal role of prompt engineering, and the remarkable performance gains achieved through fine-tuning the retriever models.

The complete project is available at: `https://drive.google.com/drive/folders/1JinjakzZPmynelkl_xWyOoxN7khPZa8m?usp=sharing`.

1 Introduction

In recent years, Large Vision Language Models (LVLMs) have shown impressive capabilities in understanding and generating content from a combination of text and images. However, their reliance on internal parametric knowledge limits their effectiveness in specialized domains. The Retrieval-Augmented Generation (RAG) architecture addresses this gap by providing relevant, external context to the model. This project focuses on implementing a complete Multimodal RAG system within the domain of Iranian tourism. In this report, we first detail the project methodology, including the data collection process, model architecture, and evaluation framework. We then present and analyze the quantitative and qualitative results, concluding with a summary of our key findings.

2 Methodology

The project workflow consisted of data preparation, model design and fine-tuning, and a comprehensive system evaluation.

2.1 Dataset and Corpus

The project's corpus consists of textual and visual data related to the tourist, historical, and cultural attractions of various Iranian provinces. Textual data was extracted from structured JSON files, and a collection of relevant images was gathered for each location. A critical observation during development was the high dependency of the retrieval system on data quality. To enhance the text

retriever's performance, the textual data was preprocessed to ensure that the name of each location was explicitly repeated within its description, thereby creating a stronger semantic link for the model.

2.2 RAG System Architecture

The system was designed with a two-stage retrieve-then-generate architecture.

2.2.1 Retriever Component

The retriever is responsible for finding the most relevant documents (text and images) and consists of two parallel subsystems:

1. **Text Retriever:** The `cis-lmu/glot500-base` model was used. Initial performance was found to be very weak and unreliable. Therefore, the model was fine-tuned using **Contrastive Learning** with the `MultipleNegativesRankingLoss` function on pairs of (description, location name) from our corpus. This process yielded a significant improvement in the quality of the embeddings after just 2-3 epochs.

2. **Image Retriever:** The `CLIP` model was fine-tuned on (image, text description) pairs from the tourism dataset. This specialization enabled the model to develop a better understanding of the visual features specific to Iranian landscapes and architecture.

For both models, the generated embeddings were indexed in a **FAISS** vector database to enable efficient, high-speed similarity searches.

2.2.2 Generator Component

The `Qwen/Qwen2.5-VL-7B-Instruct` multimodal model was selected for the generation stage. This choice was made after comparing it with alternatives like Llama; the Qwen model demonstrated superior fluency in Persian, produced higher-quality outputs, and was notably more "prompt-receptive"—adhering more closely to instructions. To manage memory, the model was loaded with 4-bit quantization.
Prompt Engineering: The prompt structure was pivotal in guiding the model. It was discovered that adding a simple example of the desired response format (a few-shot technique) within the prompt made the model's answers significantly more targeted and compliant with the required output format.

2.3 Evaluation Framework

The final evaluation was conducted on a test set of 80 text-based and 21 multimodal multiple-choice questions. System performance was measured using standard RAG metrics and was compared against a baseline. The baseline consisted of the same Qwen model answering questions without access to any retrieved documents.

3 Results and Analysis

3.1 Overall System Performance

Comparing the full RAG system with the baseline model highlights the profound impact of the retrieval-augmentation process.

Table 1: Overall accuracy comparison of the RAG system vs. the baseline model.

Model Configuration	Final Accuracy
`Qwen (Baseline - No RAG)`	28.75%
`Our RAG System (Text-only Questions)`	**76.25%**

As shown in Table 1, the RAG system improved accuracy from 28.75% to 76.25%, a relative improvement of 165%. This result confirms that the Qwen model's parametric knowledge alone was insufficient for this specialized domain, and the system's success was almost entirely dependent on the retrieved context.

3.2 Retriever Performance Analysis

To better understand the system's behavior, the retriever component was analyzed for both text-only and multimodal inputs.

Table 2: Key retriever metrics for text-only vs. multimodal systems.

System Type	Precision@3	Answer Recall Rate	Hit Rate@2
RAG (Text-only)	89.17%	77.50%	77.50%
RAG (Multimodal)	80.95%	**85.71%**	85.71%

The analysis of Table 2 reveals several key insights:

- **Precision vs. Recall Trade-off:** The text-only system was slightly more precise in retrieving relevant documents. However, the multimodal system had a significantly higher **Answer Recall Rate**. This demonstrates that the image modality successfully compensated for the shortcomings of text search in numerous cases, increasing the overall chance of finding the correct answer.
- **Ranking Quality:** The exact match between the Answer Recall Rate and Hit Rate@2 is a critical finding. It indicates that whenever the retriever found a document containing the correct answer, it **always** placed it in the top two results. This confirms that the system's primary challenge is one of recall (finding the document) rather than ranking (ordering the documents).

4 Conclusion

This project successfully implemented an efficient multimodal RAG system for the domain of Iranian tourism. The results clearly demonstrated that the RAG architecture, especially when paired with retriever models fine-tuned via contrastive learning, dramatically improves accuracy, elevating performance from near-random guessing to a reliable and precise level.

The most significant finding was that adding the image modality, while slightly reducing retrieval precision, substantially boosted the overall Answer Recall Rate. This proves that text and image modalities can act in a complementary fashion to create a more robust system. Furthermore, the detailed analysis identified that the primary bottleneck is the retriever's ability to recall the correct document, while its ranking and the generator's extraction capabilities are already highly effective. Future work should therefore focus on improving the core recall mechanisms to further enhance system performance.