# Aircraft-Centric Multimodal Retrieval-Augmented Generation System

Meenakshi Iyer(202418031), Shashwat Sharma(202418050)

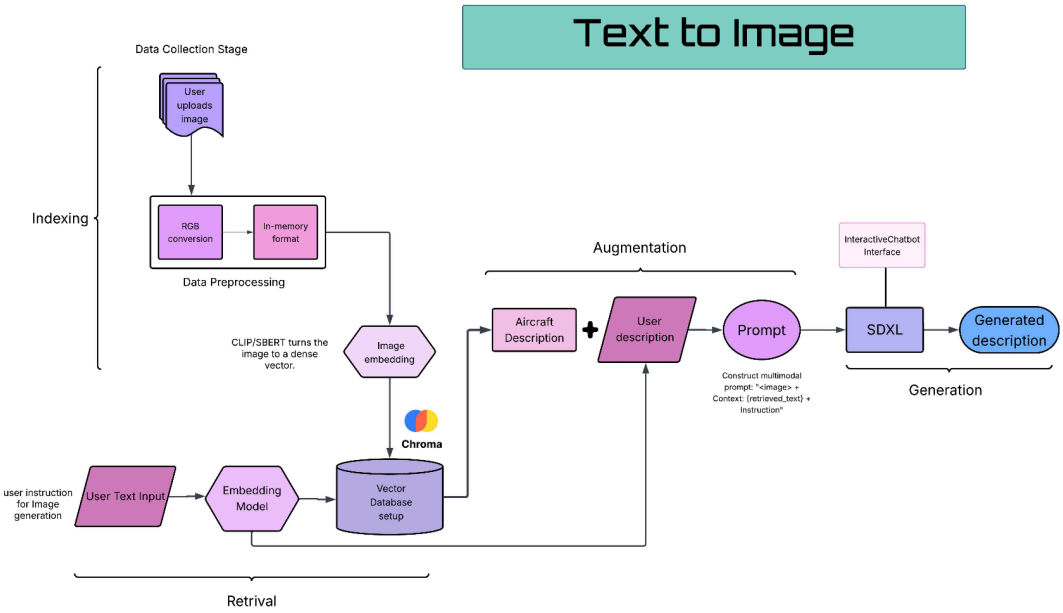
---

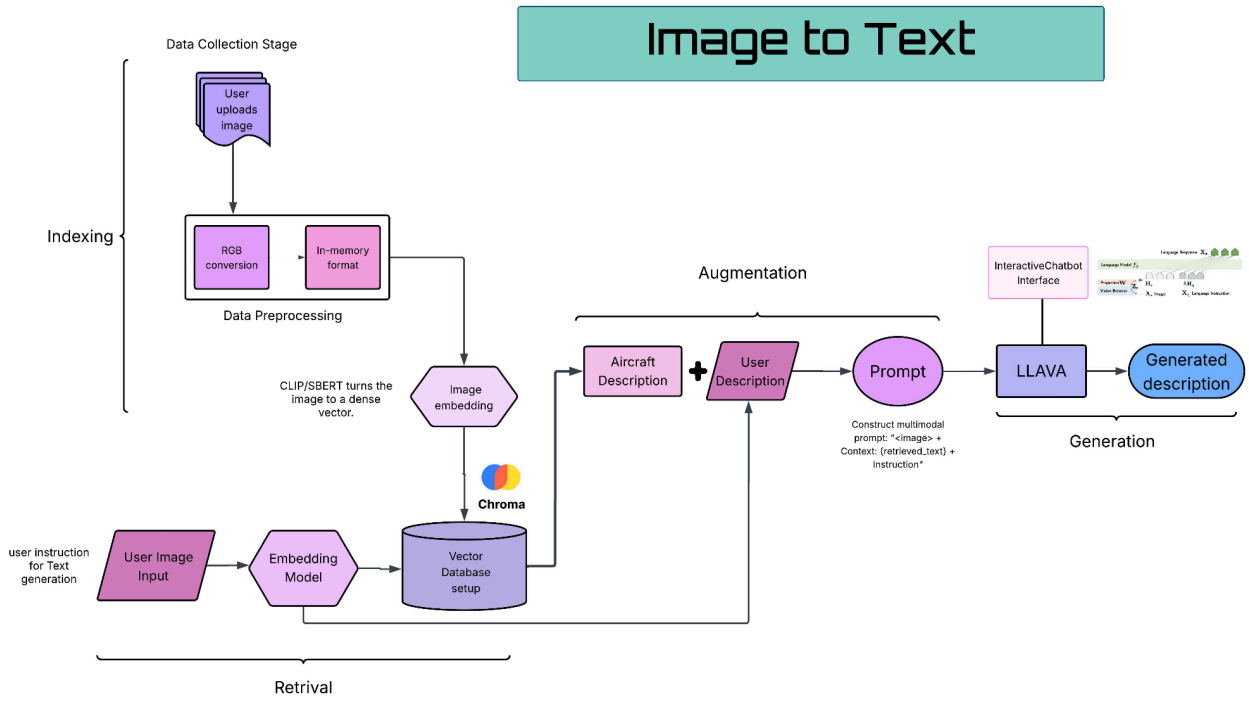
## Introduction:

Recent advances in deep learning have enabled powerful multimodal systems capable of processing both visual and textual data. Vision–language models are increasingly used for tasks such as image captioning, visual question answering, and text-to-image generation. However, when applied to highly specialized domains, such as military aviation, generic models often fail to produce accurate, grounded, and technically meaningful outputs.

Aircraft recognition and description is a fine-grained visual reasoning problem. Many aircraft share similar silhouettes, configurations, and structural components, where small visual details—such as wing sweep angle, engine placement, or tail geometry—carry significant semantic importance. General-purpose models frequently hallucinate features or confuse visually similar aircraft types.

To address these challenges, this project proposes a domain-specific multimodal Retrieval-Augmented Generation (RAG) system focused on military aircraft. The system combines aircraft image data with a structured aircraft knowledge base and uses retrieval mechanisms to ground generation in verified domain information.






# 2. Dataset Description

## 2.1 Aircraft Image Dataset

The visual component of the system is built using a military aircraft image dataset that contains labeled images of a wide range of military aircraft classes. The dataset includes aircraft from multiple operational categories such as fighter aircraft, bombers, transport aircraft, and unmanned aerial vehicles, reflecting the diversity present in real-world military aviation.

A key challenge addressed by this dataset is the high inter-class visual similarity among aircraft types, where subtle structural differences distinguish one class from another. Additionally, the dataset exhibits substantial variability across multiple dimensions, including:

- Viewing angles and perspectives
- Aircraft scale and distance from the camera
- Background environments and operational settings
- Lighting and atmospheric conditions

The images represent real-world operational scenarios rather than synthetic or artificially generated data, making the task of visual understanding more realistic and challenging.

This dataset is used as the primary source for visual embedding generation and plays a critical role in retrieval-based grounding, enabling the system to associate aircraft images with relevant textual knowledge during inference.

---

### 2.2 Aircraft Knowledge Dataset

The textual component of the system is provided by the file aircraft_descriptions.json, which serves as the domain-specific knowledge base for the Retrieval-Augmented Generation pipeline. Each entry in this dataset corresponds to a distinct aircraft type and contains a detailed technical description capturing key structural and design characteristics. These descriptions typically include information related to:

- Engine configuration and placement
- Wing design, geometry, and sweep
- Tail and stabilizer configuration
- Fuselage shape and structural layout
- Landing gear arrangement
- Other distinguishing structural features

This structured aircraft knowledge dataset functions as the retrieval corpus within the RAG framework. During inference, relevant entries are retrieved and injected into the generation process, ensuring that both image-to-text and text-to-image outputs remain grounded in accurate, aircraft-specific domain knowledge.

## 3. Problem Formulation

The project addresses the following core problems:

1. **Image-to-Text Generation**
   Given an aircraft image, generate an accurate and technically grounded textual description.
2. **Text-to-Image Generation**
   Given a textual aircraft description, generate a corresponding aircraft image that adheres to structural and visual constraints.
3. **Hallucination Reduction**
   Ensure that generated outputs do not introduce features inconsistent with real aircraft designs.

These problems are addressed through multimodal retrieval and generation, rather than purely generative modeling.

## 4. Methodology

### 4.1 Overall Approach

The system follows an end-to-end multimodal Retrieval-Augmented Generation pipeline, implemented in the provided Python file exported from the notebook

The methodology consists of:

1. Data ingestion
2. Multimodal embedding generation
3. Vector-based retrieval
4. Context-augmented generation

---

## 4.2 Dataset Processing

The visual data used in this project is derived from the Illia56/Military-Aircraft-Detection dataset, which provides a comprehensive and diverse collection of military aircraft imagery. The dataset consists of 12,008 images spanning 43 distinct aircraft types, including but not limited to *A-10, F-16, F-35, Su-57*, and other aircraft from multiple nations and operational categories such as fighters, transport aircraft, VTOL/SVTOL platforms, and surveillance systems.

Each image in the dataset is accompanied by bounding box annotations in PASCAL VOC format, enabling precise localization of aircraft within complex visual scenes. These annotations allow the preprocessing pipeline to focus on the aircraft region of interest while minimizing background noise, which is critical for learning discriminative visual features in high-variance environments.

During preprocessing, aircraft images are:

- Parsed along with their corresponding PASCAL VOC annotation files
- Cropped or region-focused using bounding box information where applicable
- Resized to consistent spatial dimensions required by the vision encoder
- Normalized to ensure stable numerical input distributions during embedding generation

The dataset covers a wide range of viewing perspectives, including different altitudes, orientations, scales, and operational environments, improving the robustness and generalization capability of the learned visual representations.

In parallel, aircraft descriptions are loaded from the structured JSON-based aircraft knowledge corpus, where each aircraft type is associated with a detailed technical description. These textual entries capture aircraft-specific attributes such as engine configuration, wing geometry, tail structure, and fuselage design.

Each aircraft type is treated as a shared semantic entity across visual and textual modalities, enabling consistent cross-modal alignment. This structured pairing forms the foundation for multimodal embedding generation and subsequent retrieval-based grounding in the RAG system.

---

## 4.3 Embedding Generation

To enable multimodal retrieval and generation, both textual and visual data are transformed into dense vector representations within a common embedding framework.

### Text Embeddings

Aircraft descriptions from the JSON knowledge base are converted into dense embeddings using a language embedding model. These embeddings are designed to capture semantic meaning as well as technical aircraft-specific attributes, such as engine configuration, wing geometry, tail structure, and fuselage design.
Rather than relying on surface-level lexical similarity, the embedding process emphasizes conceptual and structural information, allowing the system to distinguish between visually and functionally similar aircraft types.

Each textual embedding is associated with its corresponding aircraft identifier, ensuring a consistent mapping between descriptive knowledge and visual representations.

### Image Embeddings

Aircraft images are processed using a vision encoder to extract high-dimensional visual feature embeddings. These embeddings encode structural and geometric characteristics, including overall silhouette, wing placement, engine positioning, and tail configuration.
Where bounding box annotations are available, the embedding process focuses on aircraft regions of interest, reducing background interference and improving feature discrimination.

The resulting image embeddings represent aircraft visual patterns in a compact form suitable for similarity-based retrieval.

### Cross-Modal Compatibility

Both text and image embeddings are generated in a manner that supports cross-modal similarity comparison, enabling retrieval between images and descriptions. This compatibility is essential for linking visual inputs with relevant textual context in the RAG pipeline.

---

## 4.4 Vector Store Construction

All generated embeddings are stored in a vector database that serves as the retrieval backbone of the system. Each entry in the vector store includes:

- The embedding vector
- Aircraft identity or class label
- Modality type (image or text)
- Associated metadata for traceability

The vector store enables efficient nearest-neighbor similarity search, allowing the system to retrieve the most relevant aircraft descriptions for a given image query, or the most relevant images for a given textual prompt.

This retrieval layer is a critical component of the RAG architecture, as it ensures that downstream generation is conditioned on domain-relevant, aircraft-specific information, rather than relying solely on the generative model's internal parameters.

# 5. Retrieval-Augmented Generation (RAG) System

## 5.1 Motivation for Using RAG

In specialized domains such as military aviation, purely generative vision–language models often produce inaccurate or fabricated details due to the absence of explicit domain knowledge at inference time. Aircraft recognition and description require precise understanding of structural and functional attributes, where even minor inconsistencies can lead to incorrect interpretations.

Retrieval-Augmented Generation (RAG) addresses this limitation by explicitly incorporating external, domain-specific knowledge during inference. Instead of relying solely on learned model parameters, the system retrieves relevant aircraft information from a curated knowledge base and uses it to guide the generation process. This approach significantly reduces hallucination and improves factual grounding.

---

## 5.2 Retrieval Phase

The retrieval phase serves as the knowledge selection mechanism of the system. Queries are first transformed into embeddings and then matched against the vector store to identify the most relevant entries.

- Image-based Queries
  When an aircraft image is provided as input, its visual embedding is used to retrieve the most similar aircraft descriptions from the vector database. Similarity is computed based on structural and geometric visual features encoded in the embedding space.

- Text-based Queries
  When a textual prompt is provided, the corresponding text embedding is used to retrieve both aircraft descriptions and related aircraft images that are semantically aligned with the query.

This retrieval process ensures that only aircraft-specific, contextually relevant information is selected and passed to subsequent stages.

---

## 5.3 Context Injection

The retrieved aircraft descriptions are aggregated and structured into a contextual input that is supplied to the generation model. This retrieved context acts as an external constraint, guiding the model toward outputs that are consistent with known aircraft characteristics.

By grounding generation in retrieved knowledge rather than relying solely on internal representations, the system maintains alignment with verified aircraft attributes such as configuration, design features, and category-specific properties.

---

## 5.4 Generation Phase

The generation phase uses the retrieved and injected context to produce outputs that are both coherent and domain-consistent.

### Image-to-Text Generation

For image-to-text tasks, the system generates aircraft descriptions that are explicitly grounded in the retrieved textual knowledge. The generated output emphasizes structural and visual attributes observed in the image, such as wing configuration, engine placement, and overall aircraft geometry.

### Text-to-Image Generation

For text-to-image tasks, the input prompt is enriched with retrieved aircraft-specific information prior to generation. This grounding process helps maintain consistency between the generated image and the intended aircraft class, reducing visual ambiguity and improving class fidelity.

# 9. Future Work and Conclusion

Several extensions can further enhance the system:

- Fine-grained aircraft subtype classification
- Attribute-level grounding (engine count, wing geometry)
- Integration with real-time inference APIs
- Quantitative evaluation using multimodal similarity metrics
- Expansion of the knowledge corpus with operational and historical data

This project demonstrates the effectiveness of Retrieval-Augmented Generation for domain-specific multimodal learning. By grounding image and text generation in a curated aircraft knowledge base, the system achieves higher accuracy, interpretability, and reliability than standalone generative models. The proposed approach highlights the importance of retrieval-based grounding when deploying generative AI in specialized, high-stakes domains such as military aviation.