Dhirubhai Ambani University, Gandhinagar, Gujarat.

A Retrieval-Augmented Multimodal Framework for Image-to-Story and Story-to-Image Generation

MLSE Final Project Documentation

Submitted by:

Name	Student ID
Sujal Dhrangdhariya	202418017
Vedant Dave	202418014
Jatin Sindhi	202418055

Instructor:
Prof. Gopendra V. Singh

December 17, 2025

1. Introduction

Multimodal Artificial Intelligence has emerged as a significant research area, enabling systems to **understand and generate information across multiple modalities** such as **text** and **images**. Among these, **bidirectional generation** between **visual** and **textual domains** has gained considerable attention due to its applications in **visual storytelling**, **creative content generation**, **education**, and **human–AI interaction**.

Recent advances in **vision–language models** and **diffusion-based image generation** have demonstrated strong *zero-shot capabilities*. However, most existing systems operate in a *single-pass manner*, relying solely on **prompt-based conditioning**. Such approaches often struggle to maintain **long-term narrative coherence**, **stylistic consistency**, and **personalization** across multiple generations. Furthermore, **fine-tuning large multimodal models** to encode user preferences is *computationally expensive* and *impractical* for real-world deployment.

To address these challenges, this project adopts a ***Retrieval-Augmented Generation (RAG)*** paradigm for **multimodal storytelling**. Instead of modifying *model parameters*, the system retrieves **semantically relevant prior examples** from an **external memory** and integrates them into the generation process at *inference time*. This **retrieval-based grounding** enhances **narrative coherence**, **stylistic alignment**, and **contextual relevance** without requiring retraining.

The proposed **Retrieval-Augmented Visual Storytelling Generation (RAVSG)** system supports both ***image-to-story*** and ***story-to-image*** generation. By embedding **images** and **text** into a **shared semantic space** and continuously expanding a **multimodal knowledge base**, the system demonstrates *adaptive behavior* over time. This design enables **personalization**, **reuse of prior narrative patterns**, and **improved creative consistency**, making it well-suited for **production-grade multimodal storytelling applications**.

2. System Overview

The **Retrieval-Augmented Visual Storytelling Generation (RAVSG)** system is designed as a **modular, scalable, and asynchronous multimodal architecture** capable of handling both *image-to-story* and *story-to-image* generation tasks. The system integrates **vision models**, **language models**, and a **retrieval mechanism** to enhance contextual grounding, narrative coherence, and stylistic consistency.

At a high level, the system operates by transforming user inputs into a **shared semantic embedding space**, retrieving **semantically relevant prior examples**, and augmenting the generation process using structured prompts. This design allows the system to behave as an *experience-driven generative framework*, where past outputs influence future generations without modifying model parameters.

2.1. Core Functionalities

The system supports the following primary functionalities:

- **Image-to-Story Generation**: Converts visual inputs into descriptive and emotionally coherent narratives.

- **Story-to-Image Generation**: Generates visually consistent images from textual story prompts.
- **Retrieval-Augmented Contextualization**: Enhances generation using previously stored multimodal examples.
- **Continual Memory Update**: Stores newly generated outputs for future retrieval and personalization.

These functionalities are unified under a common retrieval and prompt-conditioning pipeline, ensuring architectural symmetry between both generation directions.

2.2. Architectural Components

The overall architecture of the system consists of the following key components:

- **Input Interface**: Accepts user-provided **text prompts** or **images**.
- **Embedding Module**: Encodes both text and images into a **shared multimodal embedding space**.
- **Retrieval Module**: Uses a **vector database** to identify top-K semantically similar examples.
- **Prompt Construction Engine**: Injects retrieved context into *structured prompts* with stylistic constraints.
- **Generation Workers**: Perform story or image generation using task-specific models.
- **Memory Store**: Maintains a growing **multimodal knowledge base** for retrieval.

Each component is designed to operate independently, enabling flexible scaling and efficient resource utilization.

2.3. Asynchronous Execution Flow

To support real-time responsiveness and scalability, the system employs an **asynchronous processing model**. Incoming requests are treated as independent jobs and routed to appropriate generation pipelines based on task type. This design ensures that:

- Only relevant models are activated per request
- Computational resources are efficiently utilized
- Vision and diffusion workloads can scale independently

By decoupling request handling from generation execution, the system achieves improved throughput and robustness, making it suitable for **production-grade multimodal applications**.

3. Multimodal Embeddings and Retrieval-Augmented Generation

A core component of the **Retrieval-Augmented Visual Storytelling Generation (RAVSG)** system is the use of **multimodal embeddings** and a **Retrieval-Augmented Generation (RAG)** framework. This combination enables effective alignment between visual and textual modalities while providing contextual grounding during generation.

3.1. Multimodal Embedding Space

To enable unified processing of images and text, the system projects both modalities into a **shared semantic embedding space**. This representation allows semantic similarity to be measured directly between visual and textual inputs, facilitating cross-modal retrieval.

Formally, let $\mathcal{I}$ denote the image space and $\mathcal{T}$ denote the text space. Two encoding functions are defined as:

$$f_{\mathcal{I}} : \mathcal{I} \to R^d, \quad f_{\mathcal{T}} : \mathcal{T} \to R^d$$

where d denotes the embedding dimensionality. The resulting embeddings are normalized and compared using **cosine similarity**.

This shared embedding space enables the system to:

- Perform **cross-modal similarity search**
- Index both images and stories using a **unified representation**
- Retrieve context relevant to either generation direction

3.2. Retrieval-Augmented Generation Framework

The ***Retrieval-Augmented Generation (RAG)*** framework enhances the generative process by incorporating **semantically similar prior examples** at inference time. Rather than relying solely on user prompts, the system retrieves top-K nearest neighbors from an external memory based on embedding similarity.

The retrieval process provides the following advantages:

- **Improved narrative coherence** through contextual grounding
- **Stylistic consistency** across multiple generations
- **Reduced hallucination** by anchoring generation to prior examples

Retrieved items are treated as *soft constraints* and are not enforced as factual truth. Instead, they serve as inspiration that guides the generative models toward more context-aware outputs.

3.3. Vector Database and Similarity Search

To support efficient retrieval at scale, the system utilizes a **vector database** optimized for *approximate nearest-neighbor search*. All multimodal embeddings are indexed using cosine similarity, enabling fast retrieval even as the knowledge base grows.

The use of an external vector store provides several benefits:

- **Decoupling of memory from model parameters**
- **Scalability** without retraining large models
- Support for **continual learning behavior**

By treating retrieval as an external memory mechanism, the system exhibits adaptive behavior over time while preserving the generalization capabilities of pre-trained generative models.

4. Model Selection and Generation Framework

The effectiveness of the **Retrieval-Augmented Visual Storytelling Generation (RAVSG)** system relies on the careful selection of **state-of-the-art generative models** that support multimodal reasoning, instruction following, and high-quality synthesis. Each model is chosen based on its suitability for a specific task within the bidirectional generation pipeline.

4.1. Multimodal Alignment Model

To establish a strong alignment between visual and textual representations, the system employs a **multimodal embedding model** trained using a *contrastive learning objective*. This model maps both images and text into a shared semantic space, enabling meaningful similarity comparisons across modalities.

The key advantages of this alignment model include:

- **Cross-modal semantic consistency**
- Efficient embedding computation for large-scale retrieval
- Compatibility with vector-based similarity search

This shared representation is essential for enabling retrieval-augmented generation in both image-to-story and story-to-image tasks.

4.2. Image-to-Story Generation Model

For generating narratives conditioned on visual inputs, the system utilizes a **vision–language model** capable of jointly reasoning over images and text. This model integrates visual features with a language decoder, allowing it to generate detailed and coherent stories grounded in visual content.

The model exhibits strong *instruction-following behavior*, which enables effective utilization of structured prompts enriched with retrieved context. This capability allows narrative tone, emotional depth, and stylistic elements to be controlled through prompt design rather than model fine-tuning.

Key characteristics of the image-to-story generation model include:

- **Visual grounding** for narrative generation
- Support for **long-form descriptive outputs**
- Robust handling of **retrieval-augmented prompts**

4.3. Story-to-Image Generation Model

For visual synthesis from textual descriptions, the system employs a **diffusion-based text-to-image model**. This model operates in a latent space, enabling high-quality image generation while maintaining efficient computation.

The story-to-image pipeline leverages:

- **Positive prompts** to reinforce desired visual attributes
- **Negative prompts** to suppress common visual artifacts
- Retrieved visual context to maintain *stylistic consistency*

By combining retrieved examples with structured prompting, the model generates images that align closely with the narrative intent and aesthetic style of the input story.

4.4. Role of Retrieval in Generation

Retrieval plays a central role across both generation directions. Instead of acting as a rigid constraint, retrieved examples function as *soft guidance* that influences tone, structure, and visual composition. This design allows the system to balance creativity with consistency, enabling expressive storytelling while maintaining contextual relevance.

Through this modular model selection and retrieval-integrated design, the RAVSG system achieves a flexible and scalable multimodal generation framework suitable for real-world creative applications.

5. Dataset Description and Preprocessing

The performance and quality of the **Retrieval-Augmented Visual Storytelling Generation (RAVSG)** system are strongly influenced by the **multimodal datasets** used to initialize its external memory. The datasets are selected to support both *image-to-story* and *story-to-image* generation while maintaining stylistic coherence and narrative richness.

5.1. Dataset Sources

The initial knowledge base of the system is constructed using **Studio Ghibli–inspired multimodal datasets**, which consist of visually expressive images paired with long-form textual descriptions. These datasets are particularly suitable for visual storytelling tasks due to their emphasis on emotional tone, atmosphere, and narrative depth.

The datasets used in this project include:

- **Ghibli Art Dataset**: A collection of Ghibli-style images paired with detailed narrative captions, enabling rich image-to-story mapping.
- **Ghibli-Style 100 Dataset**: A smaller curated dataset containing diverse Ghibli-inspired scenes with corresponding descriptive text.

Together, these datasets provide a balanced combination of visual diversity and narrative expressiveness.

5.2. Data Characteristics

The datasets exhibit the following characteristics:

- **High visual diversity**, including landscapes, characters, and atmospheric scenes
- **Long-form textual descriptions** suitable for narrative generation
- Strong *stylistic consistency* inspired by Studio Ghibli aesthetics

Such characteristics are essential for training and evaluating systems focused on creative and emotionally grounded storytelling.

5.3. Preprocessing Pipeline

Before indexing the data into the multimodal memory, a preprocessing pipeline is applied to ensure data quality and consistency. The preprocessing steps include:

- **Text cleaning**: Removal of grammatical errors, redundant phrases, and low-quality captions
- **Normalization**: Standardization of text formatting and metadata
- **Image validation**: Verification of image resolution and format consistency
- **Embedding computation**: Conversion of images and text into shared semantic embeddings

These preprocessing steps ensure that only high-quality multimodal representations are stored in the vector database.

5.4. Knowledge Base Initialization

After preprocessing, the embeddings are indexed into the **vector database**, forming the system's initial *external memory.* This memory serves as the foundation for retrieval-augmented generation and evolves over time as new outputs are generated and appended.

By maintaining a curated and continuously expanding knowledge base, the system supports improved retrieval precision, narrative coherence, and stylistic consistency across generations.

6. Methodology and System Workflow

This section describes the **methodology** and end-to-end **workflow** adopted by the **Retrieval-Augmented Visual Storytelling Generation (RAVSG)** system. The methodology is designed to ensure **modularity**, **scalability**, and **context-aware generation** for both supported tasks.

6.1. Request Routing and Task Identification

Each user request is processed through a lightweight **routing mechanism** at the API layer. Based on the input type and metadata, the system determines whether the task corresponds to *image-to-story* or *story-to-image* generation.

Formally, each request can be represented as:

$$J = \langle x, \tau, s \rangle$$

where x denotes the input payload, $\tau \in \{\mathrm{I2T}, \mathrm{T2I}\}$ represents the task type, and s indicates the execution state.

This routing strategy ensures that only the **relevant generation pipeline** is activated, reducing unnecessary computation.

6.2. Embedding and Retrieval Process

Regardless of task direction, all inputs are first transformed into a **shared multimodal embedding space**. These embeddings are then used to retrieve semantically similar examples from the **vector database**.

The retrieval process involves:

- Encoding the input using multimodal encoders
- Performing **similarity search** using cosine similarity
- Selecting the top-K most relevant prior examples

The retrieved examples provide *contextual grounding* that guides the subsequent generation stage.

6.3. Prompt Construction and Conditioning

Retrieved context is injected into **structured prompt templates** along with the user input and stylistic instructions. This prompt construction step plays a critical role in controlling narrative tone, emotional depth, and visual aesthetics.

Key elements of prompt conditioning include:

- **System-level instructions** defining style and constraints
- **Retrieved contextual exemplars** for grounding
- *Optional negative constraints* to suppress undesired outputs

This approach allows retrieval to influence generation without overpowering the model's inherent creativity.

6.4. Generation and Output Synthesis

Once the final prompt is constructed, the request is forwarded to the appropriate **generation worker**. For image-to-story tasks, the system generates a coherent narrative grounded in visual input. For story-to-image tasks, the system synthesizes images that align with the semantic and stylistic intent of the story.

6.5. Continual Memory Update

After generation, the newly produced outputs are re-embedded and appended to the **multimodal knowledge base**. This *continual update loop* enables the system to adapt over time by gradually biasing retrieval toward previously generated styles and narratives.

6.6. Asynchronous Execution Model

To support efficient and scalable operation, all generation tasks are handled using an **asynchronous job-based execution model**. This design allows:

- Independent scaling of generation components
- Improved throughput under high request loads
- Efficient utilization of computational resources

Through this methodology, the RAVSG system achieves a balanced integration of retrieval, prompting, and generation, enabling robust and adaptive multimodal storytelling.

7. Experimental Results and Analysis

This section presents the **experimental results** and qualitative analysis of the **Retrieval-Augmented Visual Storytelling Generation (RAVSG)** system. The evaluation focuses on assessing the system's ability to generate **coherent narratives**, **visually consistent images**, and **stylistically aligned outputs** across both generation tasks.

7.1. Evaluation Criteria

Due to the creative and subjective nature of visual storytelling, the system is evaluated using a combination of *qualitative analysis* and observational metrics rather than strict numerical benchmarks. The following criteria are considered:

- **Narrative Coherence**: Logical flow and consistency in generated stories
- **Visual Fidelity**: Alignment between generated images and textual descriptions
- **Stylistic Consistency**: Adherence to the intended aesthetic and tone
- **Contextual Relevance**: Effective utilization of retrieved examples

These criteria reflect real-world expectations of creative AI systems.

7.2. Story-to-Image Generation Results

In the *story-to-image* task, the system successfully generates images that closely match the semantic content and emotional tone of the input stories. Retrieved visual context reinforces consistent color palettes, lighting conditions, and compositional structure.

Key observations include:

- Strong preservation of **atmospheric elements** described in the story
- Reduced occurrence of common *diffusion artifacts*

- Improved stylistic alignment across multiple generations

These results demonstrate the effectiveness of combining retrieval with structured prompt conditioning.

7.3. Image-to-Story Generation Results

For the *image-to-story* task, the system produces narratives that are emotionally rich and stylistically expressive. Visual cues such as lighting, scene composition, and mood are effectively translated into descriptive language.

Notable observations include:

- High-quality **descriptive storytelling** grounded in visual content
- Consistent narrative tone aligned with system prompts
- Occasional *creative divergence*, where imaginative elements are introduced

While minor semantic drift may occur, it often enhances creative expressiveness without significantly compromising coherence.

7.4. Impact of Retrieval-Augmented Generation

The integration of **Retrieval-Augmented Generation** has a measurable impact on output quality. Compared to prompt-only generation, retrieval-augmented outputs exhibit improved consistency, richer descriptions, and better alignment with user intent.

Overall, the experimental results indicate that retrieval-based contextual grounding plays a crucial role in enhancing both narrative quality and visual coherence in multimodal storytelling systems.

8. Conclusion and Future Scope

This project presented the **Retrieval-Augmented Visual Storytelling Generation (RAVSG)** system, a **multimodal AI framework** designed to support *bidirectional generation* between images and text. By integrating **retrieval-augmented generation**, **multimodal embeddings**, and **state-of-the-art generative models**, the system addresses key limitations of traditional prompt-only generation approaches.

The proposed architecture demonstrates how an *external memory mechanism* can significantly enhance **narrative coherence**, **stylistic consistency**, and **contextual relevance** without modifying model parameters. Through unified embeddings and continual memory updates, the system exhibits *adaptive behavior* over time, enabling personalization and reuse of prior narrative patterns.

Experimental observations indicate that retrieval-augmented prompting improves both *story-to-image* and *image-to-story* generation by grounding outputs in semantically relevant examples while preserving creative expressiveness. The modular and asynchronous design further supports scalability, making the system suitable for **production-grade creative applications**.

8.1. Future Scope

While the proposed **Retrieval-Augmented Visual Storytelling Generation (RAVSG)** system demonstrates strong performance, several enhancements can be explored to further improve **scalability**, **accuracy**, and **user experience**.

8.1.1 Cloud-Based Deployment and Scalability

A key future direction involves deploying the system on a **cloud-based infrastructure**. Utilizing cloud servers enables:

- **Faster execution** through access to high-performance GPUs and TPUs
- **Horizontal scalability** to handle multiple concurrent users
- **Efficient storage** of large multimodal vector databases

Cloud-native services can be leveraged to host the **vector database**, generation models, and asynchronous workers, ensuring low-latency responses and improved system reliability in real-world deployments.

8.1.2 Persistent Memory and Historical Context

The current system maintains a growing multimodal memory for retrieval. This concept can be extended by storing **user-specific past history** as a persistent memory layer. Such a memory would:

- Preserve long-term narrative preferences
- Enable consistent storytelling across sessions
- Allow the system to learn evolving user intent

By treating historical interactions as a form of *experiential memory*, the system can generate outputs that are increasingly personalized and context-aware over time.

8.1.3 Hybrid Model Architecture

Future improvements may include adopting a **hybrid generation approach** that combines:

- Large vision–language models for high-level reasoning
- Lightweight specialized models for stylistic refinement
- Rule-based or heuristic constraints for narrative structure

Such a hybrid architecture can balance **creative expressiveness** with **controlled generation**, improving both accuracy and interpretability while reducing computational overhead.

8.1.4 Accuracy and Model Enhancement

The accuracy of both retrieval and generation can be improved through:

- Fine-grained domain adaptation of embedding models
- Confidence-aware retrieval weighting mechanisms
- Integration of feedback-driven ranking strategies

Additionally, ensemble-based prompting strategies and adaptive prompt optimization can further enhance generation quality across diverse storytelling scenarios.

8.1.5 Advanced Interaction and Feedback Integration

Incorporating **human-in-the-loop feedback** represents another promising extension. User feedback can be used to:

- Curate and filter stored memory entries
- Reinforce preferred styles and themes
- Suppress undesired narrative or visual patterns

This interactive learning loop would allow the system to evolve in alignment with user expectations while maintaining robustness and creative diversity.

Overall, these future enhancements highlight the potential of retrieval-augmented multimodal systems to evolve into **intelligent, memory-driven creative assistants** capable of delivering personalized and scalable storytelling experiences.

Thank You

We would like to sincerely thank **Sir** for assigning us this project. Working on this project helped us gain deeper practical understanding and hands-on experience with real-world concepts. The knowledge and skills we learned through this project are highly relevant to industry applications and have strengthened our technical foundation. We truly hope that you appreciate our project and the sincere effort we have put into completing it.

Project Team Members:

Sujal Dhrangdhariya (202418017)
Vedant Dave (202418014)
Jatin Sindhi (202418055)