RAVSG: Retrieval-Augmented-Visual-Storytelling-Generation

A Studio Ghibli-inspired Multimodal Generation System with Retrieval-Augmented Generation

Ashish Kar Misty Roy Yashraj Sinh Pragnya Dandavate
202418007 202418033 202418064 202418065
Dhirubhai Ambani University, Gandhinagar, India

***Abstract*—Multimodal generation systems capable of translating between visual and textual modalities have gained significant attention due to advances in large-scale vision–language models. This work presents a production-grade *Story–Image Bidirectional Generator*, a system that supports both image-to-story and story-to-image generation. The architecture integrates state-of-the-art open-source models, including CLIP for multimodal embeddings, LLaVA for vision-to-language generation, and Stable Diffusion for text-to-image synthesis. To improve coherence, consistency, and generation quality over time, the system employs a Retrieval-Augmented Generation (RAG) framework powered by FAISS vector databases. The system is deployed using an asynchronous microservice architecture with FastAPI and Redis-based job queues, enabling scalability and efficient GPU utilization. This document details the system design, mathematical foundations, and methodological choices, and discusses future directions for improvement.**

I. Introduction

Bidirectional generation between images and text is a central problem in multimodal artificial intelligence, with applications in visual storytelling, creative media, and human–AI interaction. Recent advances in vision–language models and diffusion-based image generators have demonstrated remarkable zero-shot capabilities. However, such models are typically optimized for single-pass generation and lack mechanisms for adaptation, personalization, and long-term contextual consistency.

In practical storytelling and creative systems, generation quality is not solely determined by model capacity, but by the ability to incorporate prior examples, stylistic preferences, and user-specific context. Pure prompt-based conditioning is often insufficient for maintaining thematic coherence across generations or for adapting to evolving user intent. Moreover, fine-tuning large multimodal models to encode personalization is computationally expensive and infeasible in many production scenarios.

To address these limitations, we adopt a Retrieval-Augmented Generation (RAG) paradigm. Rather than modifying model parameters, RAG augments inference by retrieving semantically relevant prior examples from an external memory. This design allows the system to (i) reuse previously generated stories and images, (ii) gradually accumulate stylistic and narrative preferences, and (iii) provide contextual grounding that improves coherence and personalization without retraining.

Unlike deployment-driven motivations such as latency reduction or scalability, the use of RAG in this work is motivated by *representational and cognitive considerations*. By grounding generation in a growing multimodal memory, the system emulates a form of experiential recall, enabling stylistic continuity and narrative consistency over time. This is particularly important in creative domains, where users expect outputs to align with specific artistic styles or emotional tones.

Furthermore, the proposed system explicitly integrates prompt engineering with retrieval. Retrieved examples are injected into structured prompts that encode stylistic intent, narrative voice, and aesthetic constraints. This combination enables controllable personalization while preserving the generalization capabilities of large pre-trained models.

The main contributions of this work are:

- A retrieval-augmented bidirectional story–image generation framework,
- A unified multimodal memory using CLIP embeddings and FAISS indexing,
- A personalization mechanism based on prompt templates enriched with retrieved context,
- An empirical demonstration of how architectural choices in modern multimodal models align with retrieval-based augmentation.

II. Preliminaries

A. Multimodal Embeddings

Let $\mathcal{I}$ denote the image space and $\mathcal{T}$ denote the text space. CLIP learns joint embeddings via two encoders:

$$f_I : \mathcal{I} \to \mathbb{R}^d, \quad f_T : \mathcal{T} \to \mathbb{R}^d$$

where $d = 512$ in our system.

The embeddings are normalized to lie on the unit hypersphere:

$$\hat{z} = \frac{z}{\|z\|_2}$$

B. Retrieval-Augmented Generation (RAG)

RAG enhances generative models by retrieving semantically similar examples from a knowledge store. In this work, FAISS is used to index CLIP embeddings for efficient nearest-neighbor search.

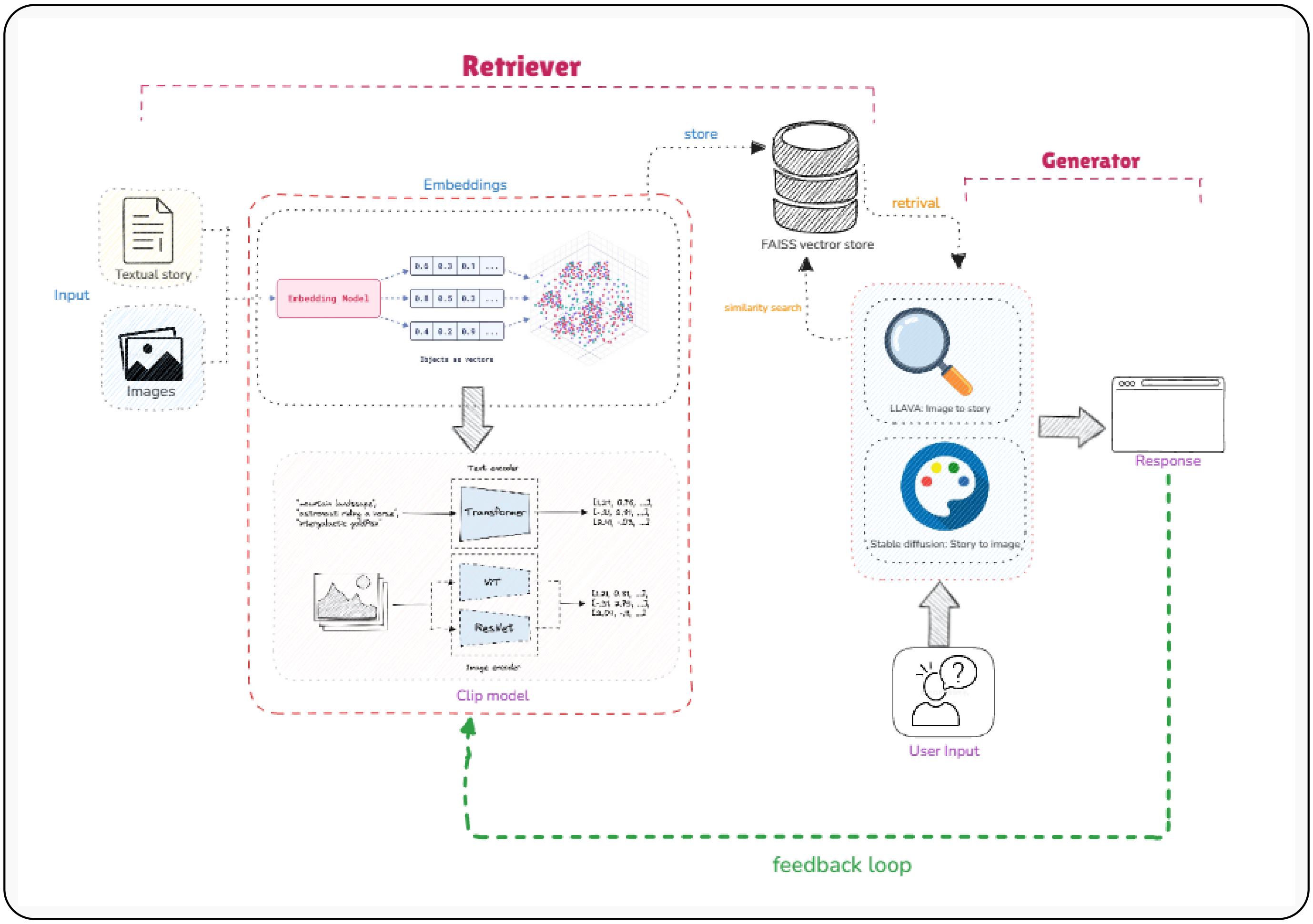


Fig. 1: Overall architecture of the Story–Image Bidirectional Generator. User requests are routed to either image-to-story or story-to-image pipelines. CLIP embeddings provide a shared multimodal space for retrieval via FAISS, while generation is handled asynchronously by LLaVA and Stable Diffusion workers. LLava for image to story and Stable Diffusion for story to image.

C. *Personalization through Retrieval and Prompt Conditioning*

Personalization in generative systems is traditionally achieved through fine-tuning or reinforcement learning from human feedback. In contrast, the proposed system achieves personalization at inference time by combining retrieval-augmented context with carefully structured prompts.

Let $\mathcal{C} = \{c_1, \ldots, c_K\}$ denote retrieved contextual examples, where each c_i may represent a previously generated story, image description, or stylistic annotation. These examples are embedded into the prompt as natural language context:

$$P = \langle \text{System Prompt}, \mathcal{C}, \text{User Input} \rangle \tag{1}$$

For image-to-story generation, the system prompt encodes a specific narrative style and emotional tone, such as a Studio Ghibli-inspired aesthetic characterized by softness, nostalgia, and subtle melancholy. The retrieved context reinforces this style by providing concrete narrative exemplars, enabling the model to maintain stylistic consistency across generations.

Similarly, for story-to-image generation, the retrieved examples influence visual attributes such as color palette, composition, and atmosphere. Negative prompts are used to explicitly suppress common diffusion artifacts, further refining output quality.

This approach allows the system to adapt to user preferences over time without modifying model parameters, effectively treating the vector database as an external, continuously evolving memory.

III. Model Selection

A. *CLIP for Multimodal Alignment*

CLIP is chosen as the embedding backbone due to its contrastive training objective, which explicitly aligns visual and textual representations in a shared latent space. This property is essential for retrieval-augmented multimodal systems, as it allows both images and stories to be indexed and queried using a unified similarity metric:

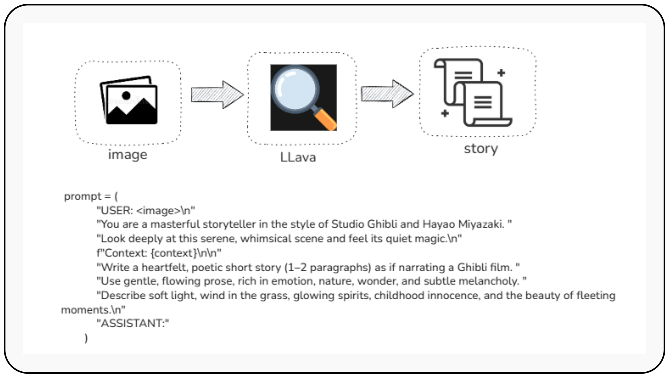
$$\mathrm{sim}(I, T) = f_I(I)^\top f_T(T)$$

The 512-dimensional embedding space provides a favorable balance between representational richness and computational efficiency, making it suitable for large-scale FAISS indexing.

B. LLaVA for Image-to-Story Generation

LLaVA extends large language models with visual grounding by integrating a vision encoder with a language decoder. Its architecture enables cross-attention between visual tokens and textual representations, making it well-suited for narrative generation conditioned on images.

Crucially, LLaVA demonstrates strong instruction-following behavior, allowing it to effectively utilize structured prompts containing retrieved context and stylistic instructions. This capability is central to the system's personalization strategy, as narrative voice and emotional tone are primarily controlled through prompt design rather than model fine-tuning.



```
prompt = (
        "USER: <image>\n"
        "You are a masterful storyteller in the style of Studio Ghibli and Hayao Miyazaki. "
        "Look deeply at this serene, whimsical scene and feel its quiet magic.\n"
        f"Context: {context}\n\n"
        "Write a heartfelt, poetic short story (1–2 paragraphs) as if narrating a Ghibli film. "
        "Use gentle, flowing prose, rich in emotion, nature, wonder, and subtle melancholy. "
        "Describe soft light, wind in the grass, glowing spirits, childhood innocence, and the beauty of fleeting
moments.\n"
        "ASSISTANT:"
    )
```

Fig. 2: Prompt construction for image to story generation with LLava.

C. Stable Diffusion for Story-to-Image Generation

Stable Diffusion is selected for text-to-image synthesis due to its latent diffusion architecture, which separates semantic reasoning from pixel-space generation. This design allows high-level textual conditioning, including stylistic modifiers and retrieved visual cues, to influence the generation process effectively.

The use of both positive and negative prompts enables fine-grained control over image quality and aesthetic attributes. Retrieved contextual information complements the textual story by reinforcing visual consistency and reducing mode collapse across generations.

D. FAISS as External Memory

FAISS provides efficient approximate nearest-neighbor search, enabling real-time retrieval from a growing multimodal dataset. By decoupling memory from model parameters, FAISS serves as a scalable external knowledge store that supports continual learning behavior without catastrophic forgetting.

IV. Dataset Descriptions

Our initial knowledge base for RAVSG was created by combinig images from two different datasets containing Ghibli-Art style images which are Ghibli Art Dataset (Kaggle) & Ghibli-Style 100 Dataset (Hugging Face)

Ghibli Art Dataset (Kaggle) The Ghibli Art dataset hosted on Kaggle provides a collection of AI-generated images in a whimsical, Studio Ghibli-inspired style. This dataset pairs visual artworks with long-form narrative captions, enabling multimodal learning tasks such as image-to-story generation and cross-modal retrieval. The paired image–text structure is suitable for training models that map between visual content and narrative descriptions. *Dataset Size*

- Number of unique rows: 973
- Average caption length: 250–400 words

Collection Process

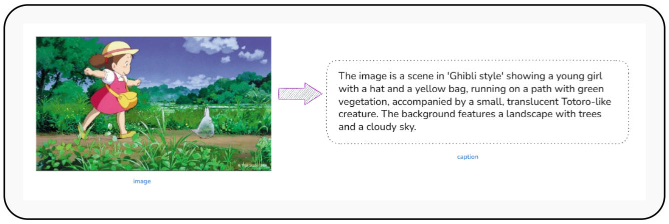
- **Image Generation:** Images were sourced from Hugging Face.
- **Caption Generation:** Captions were generated using image captioning models and refined prompts to produce detailed, story-like narratives. They were then checked and filtered for relevance and descriptive richness.

Preprocessing

- Cleaned for grammar, fluency, and duplication
- Standardized metadata and image paths

Ghibli-Style 100 Dataset (Hugging Face) The Ghibli-Style 100 dataset on Hugging Face consists of approximately 100 Ghibli-style images and corresponding descriptive captions. The dataset includes a variety of character scenes, landscapes, and stylistic illustrations in the Studio Ghibli aesthetic. Each image is accompanied by a text caption that describes its content, supporting multimodal tasks and benchmarking for story-to-image and image-to-story generation. This dataset contains 100 unique images.

Example



The image is a scene in 'Ghibli style' showing a young girl with a hat and a yellow bag, running on a path with green vegetation, accompanied by a small, translucent Totoro-like creature. The background features a landscape with trees and a cloudy sky.

Fig. 3: Example of the dataset used. (See Section IX for dataset details.)

V. Methodology

System Routing and Execution Flow

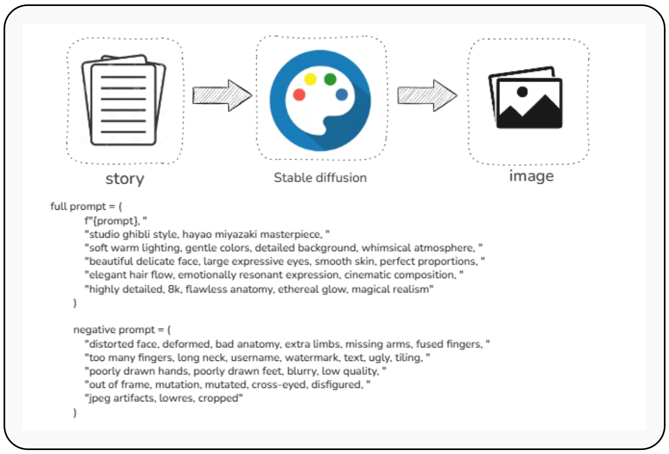
As illustrated in Fig. 1, the system operates as a unified multimodal service that dynamically routes requests based on task intent. Each incoming request is classified as either image-to-story or story-to-image using lightweight metadata inspection at the API layer.

Formally, each request is represented as:

$$J = \langle x, \tau, s \rangle \tag{2}$$

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

This routing mechanism ensures that only the relevant retrieval and generation pipelines are activated, minimizing unnecessary computation and enabling independent scaling of vision–language and diffusion workers.



```
full prompt = (
        f"{prompt}, "
        "studio ghibli style, hayao miyazaki masterpiece, "
        "soft warm lighting, gentle colors, detailed background, whimsical atmosphere, "
        "beautiful delicate face, large expressive eyes, smooth skin, perfect proportions, "
        "elegant hair flow, emotionally resonant expression, cinematic composition, "
        "highly detailed, 8k, flawless anatomy, ethereal glow, magical realism"
    )

    negative prompt = (
        "distorted face, deformed, bad anatomy, extra limbs, missing arms, fused fingers, "
        "too many fingers, long neck, username, watermark, text, ugly, tiling, "
        "poorly drawn hands, poorly drawn feet, blurry, low quality, "
        "out of frame, mutation, mutated, cross-eyed, disfigured, "
        "jpeg artifacts, lowres, cropped"
    )
```

Fig. 4: Prompt construction for story to image generation with stable diffusion.

Retrieval and Prompt Construction

Regardless of task direction, all inputs are first projected into a shared multimodal embedding space using CLIP encoders:

$$z_I = f_I(I), \quad z_T = f_T(T)$$

The resulting embeddings are used to retrieve top-K nearest neighbors from FAISS indices using cosine similarity:

$$\mathcal{R}(z_q) = \arg\max_{z_i}^{K} z_q^\top z_i$$

Retrieved items serve as contextual exemplars rather than factual constraints. As shown in Fig. 2 and Fig. 4, they are injected into structured prompt templates alongside stylistic instructions and, where applicable, negative constraints.

This design allows retrieval to influence tone, composition, and narrative structure without overpowering the generative model's prior knowledge.

Image-to-Story and Story-to-Image Generation

For image-to-story requests, retrieved textual exemplars are combined with the input image to form the final prompt:

$$\hat{T} = g_{\mathrm{LLaVA}}(\langle I, \mathcal{C}_T, \mathrm{Style} \rangle) \tag{3}$$

For story-to-image requests, retrieved images guide visual attributes during latent diffusion:

$$\hat{I} = g_{\mathrm{SD}}(T, \mathcal{C}_I) \tag{4}$$

Although the generation models differ, both pipelines share the same retrieval, prompting, and memory update logic, reinforcing architectural symmetry.

Continual Memory Update

After generation, outputs are re-embedded and appended to the multimodal memory:

$$\mathcal{D} \leftarrow \mathcal{D} \cup \{f_I(\hat{I}), f_T(\hat{T})\}$$

This continual update loop allows the retrieval distribution to evolve with user interaction, gradually biasing generation toward previously expressed stylistic and narrative preferences without modifying model parameters.

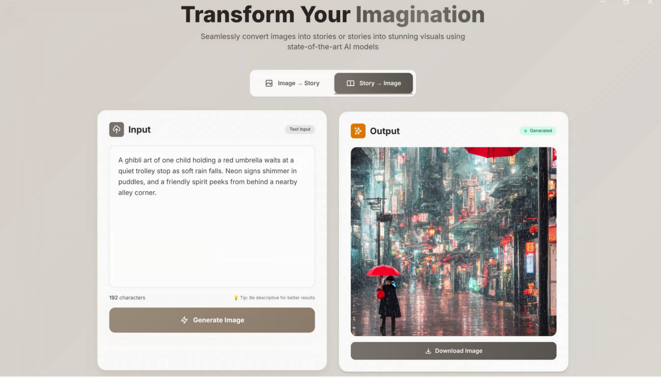
Asynchronous Processing

Each request is handled as an asynchronous job:

$$J = \langle x, \tau, s \rangle \tag{5}$$

where x is the input, τ denotes the task type, and s represents job status. This design supports high throughput and enables independent scaling of workers.

VI. Results



Transform Your Imagination

Seamlessly convert images into stories or stories into stunning visuals using state-of-the-art AI models

Input — Text Input

A ghibli art of one child holding a red umbrella waits at a quiet trolley stop as soft rain falls. Neon signs shimmer in puddles, and a friendly spirit peeks from behind a nearby alley corner.

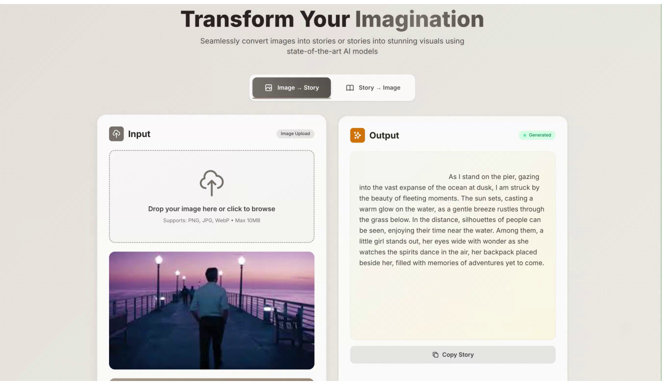
192 characters

💡 Tip: Be descriptive for better results

Output — ● Generated

Fig. 5: Story to Image Generation

The generated image closely aligns with the input prompt, accurately depicting a solitary child holding a red umbrella in a rain-soaked, neon-lit environment with a calm, cinematic atmosphere. The Studio Ghibli–inspired style is consistently reflected through soft lighting, gentle color tones, and emotionally resonant composition. Notably, the output is free from common visual artifacts, demonstrating the effectiveness of the structured positive and negative prompt design.



Transform Your Imagination

Seamlessly convert images into stories or stories into stunning visuals using state-of-the-art AI models

As I stand on the pier, gazing into the vast expanse of the ocean at dusk, I am struck by the beauty of fleeting moments. The sun sets, casting a warm glow on the water, as a gentle breeze rustles through the grass below. In the distance, silhouettes of people can be seen, enjoying their time near the water. Among them, a little girl stands out, her eyes wide with wonder as she watches the spirits dance in the air, her backpack placed beside her, filled with memories of adventures yet to come.

Image → Story
Story → Image
Input
Image Upload
Drop your image here or click to browse
Supports: PNG, JPG, WebP • Max 10MB
Output
Generated
Copy Story

Fig. 6: Image to Story Generation

The generated story largely adheres to the system prompt's stylistic and tonal constraints, rendering the dusk-lit pier and ocean horizon in a poetic, Ghibli-inspired narrative that conveys emotional depth, gentle melancholy, and a sense of fleeting beauty. Key visual elements such as the reflective mood, silhouettes, and evening atmosphere are effectively captured. However, the introduction of imaginative elements like a little girl and glowing spirits, which are not present in the image, indicates semantic drift, showing that the model prioritizes stylistic expressiveness over strict visual fidelity while maintaining strong narrative coherence.

VII. Conclusion

We presented a bidirectional multimodal generation system that integrates retrieval-augmented generation with state-of-the-art vision–language and diffusion models. By embedding both images and stories into a shared representation space and continuously expanding a multimodal knowledge base, the system exhibits emergent personalization and improved coherence over time.

Unlike fine-tuning-based approaches, personalization in the proposed framework arises from continual retrieval and prompt conditioning, allowing the system to adapt dynamically to user preferences while preserving the generalization capabilities of large pre-trained models.

VIII. Future Improvements

Potential directions for future work include:

- Domain-adaptive fine-tuning of embedding models to improve retrieval precision,
- Hierarchical or user-specific vector indices for stronger personalization,
- Confidence-aware retrieval weighting during prompt construction,
- Multi-turn narrative memory for long-form storytelling,
- Human-in-the-loop feedback for curated memory updates.

IX. Citations

This project utilizes two publicly available Studio Ghibli–style datasets to support the bidirectional image-to-story and story-to-image generation tasks. The datasets are listed below:

- Munir, H. (2024). *Studio Ghibli Art Dataset*. Kaggle. Available at: https://www.kaggle.com/datasets/humairmunir/ghilbi-art
- Moving-J. (2024). *Ghibli-Style 100 Dataset*. Hugging Face. Available at: https://huggingface.co/datasets/moving-j/ghibli-style-100/tree/main/dataset_ghibli