NewsImages: FAISS-Based Image Retrieval and Stable Diffusion Generation for Diverse News Articles

Mirunalini Palaniappan[1], Yuvashree Pudhupadi HariKrishnan[1,*], Harini Jayakumar[1] and Kawvya Muthu Kalyani[1]

[1]*Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India.*

Abstract

This paper presents the VisionX system developed for the MediaEval 2025 NewsImages Challenge, addressing automatic image retrieval and generation for news articles. Our approach combines semantic retrieval using FAISS (Facebook AI Similarity Search) with generative synthesis using Stable Diffusion v1.5 to handle both availability and diversity of visual content. SentenceTransformer embeddings are used to represent article text, while CLIP-based evaluation ensures semantic alignment between text and image. The proposed dual-pipeline framework effectively balances retrieval accuracy and generative creativity, enhancing contextual relevance and diversity in news article illustrations.

1. Introduction

The MediaEval 2025 NewsImages task focuses on improving the alignment between news articles and suitable visual content. While the task overview provides detailed objectives, our work aims to contribute an efficient dual-stage system that integrates large-scale image retrieval and text-guided image generation to enhance the visual representation of news articles.

Our system, developed by **Team VisionX**, is motivated by the need for scalable and semantically grounded techniques that improve visual storytelling in digital journalism. In our approach, retrieval is formulated as a text–image matching problem using semantic embeddings, while generation employs diffusion-based synthesis to create contextually relevant images when suitable visuals are unavailable.

To address the diversity of topics and visual styles in the dataset, we leverage CLIP-based alignment to ensure contextual consistency across diverse domains such as politics, culture, and technology. Efficient similarity matching is achieved using **FAISS (Facebook AI Similarity Search)**[1], which enables high-speed retrieval from large-scale embedding spaces. For image generation, we utilize Stable Diffusion v1.5 [1] to produce non-photorealistic, ethically compliant illustrations conditioned on article text. Our retrieval pipeline builds upon OpenCLIP-based methods successfully applied in previous MediaEval editions [2], demonstrating the effectiveness of multimodal alignment for this task.

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online

*Corresponding author.

†These authors contributed equally.

miruna@ssn.edu.in (M. Palaniappan); yuvashree2370038@ssn.edu.in (Y. P. HariKrishnan); harini2370052@ssn.edu.in (H. Jayakumar); kawvya2370039@ssn.edu.in (K. M. Kalyani)

[1]https://github.com/facebookresearch/faiss

2. Related Work

Research in multimodal learning has explored aligning text and images to enhance media understanding. Bowman et al.[3] introduced mapping between symbolic (text) and perceptual (visual) data, inspiring our article–image alignment. Grootendorst [4] influenced our use of SentenceTransformer embeddings to capture key textual features for retrieval or generation. Heitz et al. [5] highlighted the subjective nature of relevance between news texts and images, motivating our CLIP-based semantic similarity approach. The MediaEval NewsImages 2025 task [6] emphasizes text–image relationships in journalism, aligning with our combined retrieval and diffusion-generation pipeline. Radford et al. [7] proposed CLIP, enabling zero-shot semantic alignment, which we leverage to assess consistency between prompts and generated images. Finally, Wang et al. [8]demonstrated that ensemble multimodal models improve retrieval, motivating our integration of SentenceTransformer, CLIP, and Stable Diffusion for balanced retrieval, diversity, and generation.

3. System Overview and Methodology

The proposed system addresses both subtasks of the NewsImages challenge: retrieving relevant images for a given article and generating new illustrative images when suitable visuals are missing. The system is designed as two complementary pipelines: **image retrieval** and **image generation**, depicted in Figs. 1 and 2. This dual design balances accuracy, semantic relevance, and visual diversity. Retrieval ensures high-fidelity images from existing media, while generation allows flexibility for abstract or rare scenarios.

3.1. Input and Preprocessing

The system processes the NewsImages dataset, which includes a full set of news articles (`newsarticles.csv`) and a curated evaluation subset (`subset.csv`). Each article contains an ID, title, tags, and associated images. Input text for embeddings is formed as:

```
text = article_title + ". " + article_tags
```

Missing values are replaced by empty strings. Text is converted into embeddings using `SentenceTransformer (all-MiniLM-L6-v2)`, producing 384-dimensional vectors that are L2-normalized. GPU acceleration is applied to speed up embedding computation for large-scale datasets.

3.2. Image Retrieval Pipeline

Article embeddings are stored in a FAISS inner-product index (`Index FlatIP`)

- Its embedding is computed and normalized.
- Top-K most similar embeddings are retrieved from the FAISS index.
- Retrieved articles' metadata (article ID, image ID, image URL), similarity scores, and ranks are collected.

The top-K most similar images are retrieved from the FAISS index along with their metadata (article ID, image ID, image URL) and presented as output images. The retrieval pipeline leverages large-scale similarity search to ensure accurate, contextually relevant image selection.

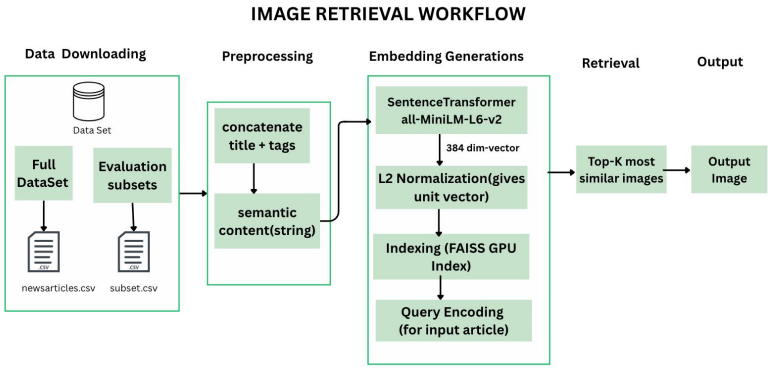

Figure 1: Flow diagram for Image Retrieval Workflow.

3.3. Image Generation Pipeline

When retrieval does not yield suitable images, the generation pipeline constructs prompts by concatenating the article title and tags. Stable Diffusion v1.5 [1] with the EulerDiscreteScheduler generates illustrative images conditioned on these prompts. Generation parameters include 30 inference steps and a guidance scale of 7.5. Generated images are saved locally, with metadata (prompts and file paths) recorded in CSV files. This pipeline allows the system to handle articles lacking pre-existing visuals while maintaining semantic alignment.

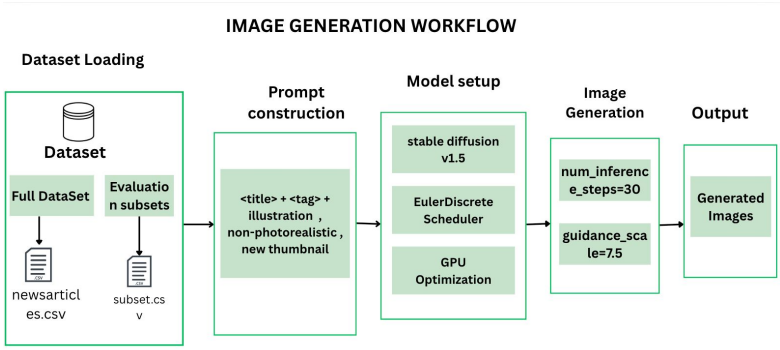

Figure 2: Flow diagram for Image Generation Workflow.

3.4. Evaluation and Motivation

The dual-pipeline design is motivated by the need to balance accuracy, diversity, and flexibility:

- Retrieval ensures high-quality, real-world images when available.
- Generation provides coverage for abstract, emerging, or rare concepts.
- CLIP-based alignment assesses semantic consistency between article text and generated/retrieved images.

Retrieval is evaluated using standard metrics (**precision@K**, **recall@K**, **Mean Reciprocal Rank (MRR)**, **nDCG**), while image generation is assessed via CLIP-based similarity scores. Comparative plots and distributions visualize performance and text-image alignment, demonstrating the effectiveness of combining retrieval and generation pipelines. All experiments

were conducted on an NVIDIA RTX 3060 GPU (12 GB VRAM) with CUDA acceleration. The full retrieval indexing and image generation processes required approximately 4 hours for the complete dataset.Stable Diffusion v1.5 model from StabilityAI (via Hugging Face Diffusers library).

4. Results

Performance evaluation was conducted during the official MediaEval 2025 crowdsourced event. The evaluation measured the perceived fit and relevance of retrieved and generated images across both small and large task subsets.

Table 1
Official Evaluation Ratings from MediaEval 2025

Run Name	Run Type	Task Type	Average Rating
VisionX_CLIP	RET	SMALL	2.74
VisionX_SD	GEN	SMALL	2.61
VisionX_SD	GEN	LARGE	1.71
VisionX_CLIP	RET	LARGE	2.61

The VisionX_CLIP retrieval approach achieved the highest average rating (2.74) on the small task, confirming the reliability of FAISS-based retrieval for contextual matching. VisionX_SD demonstrated competitive results for generative subtasks, producing diverse and visually appealing images, albeit with slightly reduced textual alignment. The results validate that combining retrieval and generation yields complementary strengths for news visualization.

5. Conclusion

This paper presented the VisionX system for the MediaEval 2025 NewsImages task, integrating FAISS-based semantic retrieval with Stable Diffusion generation to improve visual relevance in news media. The retrieval approach achieved strong contextual matching in the official evaluation, while the generation pipeline provided flexible visual coverage for missing or abstract content. Future work will focus on integrating multimodal CLIP embeddings for cross-domain retrieval, improving prompt design for Stable Diffusion, and exploring hybrid reranking of retrieved and generated images.

Declaration of Generative AI

During the preparation of this work, the following generative AI tools were used:

- ChatGPT (OpenAI, GPT-5) was used for language editing and structuring.
- Stable Diffusion v1.5 was used for generating illustrative images as part of the image generation subtask.

Code and Workflow

The full implementation and workflow are available at our GitHub repository:
https://github.com/yuvashreeph/Mediaeval

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.

[2] L. Heitz, Y. K.Chan, H. Li, K. Zeng, A. Bernstein, L. Rossetto, Prompt-based alignment of headlines and images using openclip, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.

[3] M. Bowman, S. K. Debray, L. L. Peterson, Reasoning about naming systems, ACM Trans. Program. Lang. Syst. 15 (1993) 795–825. URL: https://doi.org/10.1145/161468.161471. doi:10.1145/161468.161471.

[4] M. Grootendorst, Keybert: Minimal keyword extraction with bert, 2020. https://github.com/MaartenGr/KeyBERT.

[5] L. Heitz, A. Bernstein, L. Rossetto, An empirical exploration of perceived similarity between news article texts and images, in: MediaEval 2023 Workshop, 2024.

[6] L. Heitz, L. Rossetto, B. Kille, A. Lommatzsch, M. Elahi, D.-T.Dang-Nguyen, Newsimages in mediaeval 2025 - comparing image retrieval and generation for news articles., in: MediaEval 2025 Workshop, 2025.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning (ICML), 2021.

[8] T. Wang, J. Tian, X. Li, X. Xu, Y. Jiang, Ensemble pre-trained multimodal models for image-text retrieval in the newsimages mediaeval 2023, in: S. Hicks, A. Lommatzsch, A. Hürriyetoglu, R. Vuillemot, M. G. Constantin, V. Thambawita, M. A. Larson (Eds.), Working Notes Proceedings of the MediaEval 2023 Workshop, Amsterdam, The Netherlands and Online, 1-2 February 2024, volume 3658 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3658/paper11.pdf.