Dual-Pipeline Multimedia Enrichment for News Articles via Image Generation and Retrieval*

Dhannya Santhakumari Madhavan[1,*,†], Lakshmi Priya Swaminatha Rao[2,†], Kavinsoorya Kaliyanan Subramani[3,†], Jeeva Govindaraj[4,†], Arivuchezhiyan Elanchezhiyan[5,†] and Ashwath Ram Thavasi[6,†]

Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India

Abstract

This working note reports on a dual-pipeline system for multimedia enrichment in news articles. The approach combines automatic image generation using Stable Diffusion XL (SDXL) with image retrieval based on CLIP embeddings and FAISS similarity search. Experiments were carried out on both large-scale (8500 records) and small-scale (30 records) datasets. Outputs were standardized through resizing and formatting, with results prepared for submission and further evaluation in the shared task context.

1. Introduction

News publishers and recommender systems depend on images and thumbnails to engage readers with news articles [1]. The increasing scale of digital journalism demands automation in multimedia content creation and management. Articles often require relevant images for improved readability, audience engagement, and contextual understanding. However, manual image curation is resource-intensive. Technological progress now enables the automatic retrieval of matching images (image retrieval) or even the generation of suitable visuals for news articles using generative AI (image generation). While these advancements offer valuable opportunities for the news media, they also introduce significant technical and ethical challenges. Ensuring that images accurately correspond to the accompanying news content is essential. It is equally important to prevent visuals from misleading or deceiving readers into believing they depict real events when they do not.

Online news articles are inherently multimodal. The text of an article is often accompanied by an image and/or other multimedia items. This image, however, is not only important for illustrating and complementing the text of news articles. It plays a critical role in capturing the readers' attention; it is often the first thing readers see when browsing a news platform.

2. Background

Research in multimedia and recommender systems generally assumes a simple relationship between images and text occurring together. For example, in image captioning [2], the caption is often assumed to describe the literally depicted content of the image. In contrast, when images accompany news articles, the relationship becomes less clear [3]. Since there are often no images available for the most recent news messages, stock images, archived photos, or even generated images are used. Here, preliminary studies showed that users prefer AI-generated content over stock images [4, 5]. In [6], the authors show that diffusion models (DMs) can achieve state-of-the-art image synthesis by sequentially applying

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

*Corresponding author.

†These authors contributed equally.

dhannyasm@ssn.edu.in (D. S. Madhavan); lakshmipriyas@ssn.edu.in (L. P. S. Rao); kavinsoorya2470049@ssn.edu.in (K. K. Subramani); jeeva2470041@ssn.edu.in (J. Govindaraj); arivu2470054@ssn.edu.in (A. Elanchezhiyan); ashwathram2470063@ssn.edu.in (A. R. Thavasi)

0000-0002-0302-7458 (D. S. Madhavan); 0000-0002-9923-4020 (L. P. S. Rao)

denoising autoencoders. They address the high computational cost of pixel-based DMs by training them in the latent space of pretrained autoencoders, forming Latent DMs. This approach strikes an optimal balance between efficiency and detail preservation. By integrating cross-attention layers, they enable flexible conditioning on inputs like text or layouts, allowing high-resolution generation while achieving superior performance across tasks such as inpainting, super-resolution, and scene synthesis with significantly reduced computational demands.

The goal of this task is to investigate these intricacies in more depth, in order to understand the implications that it may have for the areas of journalism and news personalization. Our work aims to address these challenges by automating image generation by producing images aligned with news articles using generative diffusion models and automating image retrieval by retrieving the most contextually relevant images from a large dataset using semantic similarity techniques. By implementing both tasks in large and small datasets, we demonstrate scalability, efficiency, and adaptability of our approach.

3. Approach

3.1. Image generation

Our approach for Image Generation leverages SDXL due to its high image fidelity, accurate text-to-image alignment, and flexibility for prompt engineering, implemented using the Hugging Face diffusers library with GPU acceleration. To optimize prompts, we truncate article titles to 70 characters and retain up to three tags, while ensuring they remain within CLIP's 77-token limit. The prompt follows the format: "Editorial news photo illustrating: <title>. Keywords: <tags>. Realistic photojournalism, no text, no watermark", with additional modifiers enhancing realism and editorial suitability. Images are generated at a standardized resolution of 460 × 260 pixels in PNG format, embedding the prompt as metadata for traceability. The workflow consists of loading the dataset (full or subset), generating the prompt, creating the image through the SDXL pipeline, and finally resizing and saving the image with metadata.

3.2. Image retrieval

For image retrieval, CLIP (ViT-B/32) is used to generate joint text-image embeddings, while FAISS enables efficient large-scale nearest-neighbour search. In the large retrieval task, precomputed CLIP image features are loaded to build a FAISS index. Article titles are encoded using CLIP's text encoder, and a Top-1 similarity search retrieves the best-matching image, which is then resized to 460 × 260 pixels and saved with standardized naming. For the small retrieval task, a subset of the dataset is used where images are directly matched using image_id, followed by resizing and saving. This pipeline ensures fast and accurate text-to-image retrieval.

The implementation relies on several key dependencies, including PyTorch with CUDA for GPU acceleration, the diffusers library for Stable Diffusion XL (SDXL), OpenAI's clip for embedding generation, and faiss for efficient retrieval indexing. Supporting libraries like pandas and numpy handle data processing, while Pillow manages image operations. Error handling is robust, with missing images logged, loading errors caught through exceptions, and SDXL generation failures skipped with detailed logs. Performance is optimized through GPU acceleration, batch processing for FAISS operations, and prompt truncation to speed up inference, ensuring faster and more reliable generation and retrieval processes.

3.3. Dataset description

The project utilizes two main datasets: newsarticles.csv, a full dataset containing approximately 8,500 records, and subset.csv, a smaller version with 30 records for testing or quick evaluation. The image directory *newsimages_25_v1.1/newsimages/* stores all existing article images, while *features_batches/* contains precomputed CLIP feature vectors used for large-scale retrieval tasks. Each article entry in the dataset includes essential metadata fields such as *article_id, article_url, article_title, article_tags,*

image_id, and image_url, providing structured links between textual content and corresponding images for efficient retrieval and generation.
The complete codebase is available at the following link: https://github.com/jeeva2470041/MediaEval

4. Results and Analysis

The results obtained demonstrate the effectiveness of our retrieval and generation approaches. Figure 1 shows some sample images retrieved and their description. The CLIP-based image retrieval model consistently achieved higher relevance scores, confirming its robustness in aligning textual and visual semantics. The scores are ratings from an online crowdsourced event using a 5-point Likert scale from 1-5 (with 1 being the worst and 5 being the best rating). Specifically, the retrieval system attained an average evaluator score of 2.96 on the small dataset and 2.44 on the large dataset, indicating strong performance even as data complexity increased.

In contrast, SDXL image generation runs achieved average scores of 2.63–2.89 on the small dataset and 2.74 on the large dataset, producing visually appealing and contextually relevant images but occasionally diverging from the precise article semantics. Figure 2 shows some sample images generated and their description.

Table 1 shows the average scores for various runs. Table 2 gives a short description of each run.

Table 1
Average Score for different runs

Run ID	Run Name	Run Type	Task Type	Model Run	Average
21	ELITE_CODERS_STABLE	GEN	SMALL	Stable Diffusion	2.89
20	ELITE_CODERS_SDXL	GEN	SMALL	SDXL	2.63
10	ELITE_CODERS_SDXL	GEN	LARGE	Stable Diffusion XL	2.74
19	ELITE_CODERS_CLIP	RET	SMALL	CLIP (ViT-B/32)	2.96
9	ELITE_CODERS_CLIP	RET	LARGE	CLIP (ViT-B/32)	2.44

Table 2
Description of runs

Run ID	Description
21	SDXL variant with prompt tokenization and random-seed variation for diversity.
20	Generated editorial image using SDXL with title-and-tags prompt; 25 steps, guidance 7.5.
10	Generation on large dataset; identical SDXL parameters, varied input prompts.
19	Retrieval using CLIP embeddings (title + tags) indexed in FAISS; top-1 match returned.
9	Retrieval on large dataset; same CLIP configuration applied at scale.

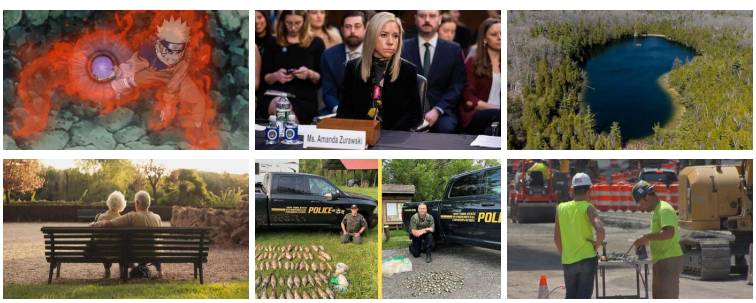

Figure 1: Sample images for retrieval, clockwise from left-top: **1**.Naruto Explains the Inspiration Behind the Rasengan. **2**.Women suing Texas over abortion bans give emotional testimony. **3**.Canadian Lake Chosen as Site to Mark Beginning of Anthropocene Epoch. **4**.Construction on State Street impacts businesses and workers in the area. **5**.Here's What Happens to Confiscated Fish in New York. **6**.Making capital gains in retirement.

Figure 2: Sample images for retrieval, clockwise from left-top: **1**.Helena Agri-Enterprises hosts Evolve Innovations Expo in Memphis. **2**.No Evidence Of Toddler On Highway Before Carlee Russell Vanished: Cops. **3**.'Undesired' air brake release behind B.C. train derailment near Alberta: report. **4**.Why England CAN win the Women's World Cup 2023 - despite missing key players. **5**.Hero Dog Gets $2,000 Reward for Locating Escaped Inmate Michael Burham. **6**.Global stocks, dollar gain as UK cooling inflation lifts sentiment.

These findings highlight that retrieval remains the more reliable method for news-image recommendation due to its semantic precision, whereas generative approaches like SDXL show potential for enhancing visual diversity and realism. Future enhancements will focus on improving multimodal understanding and retrieval accuracy. Experiments will involve using multimodal prompts that combine titles, tags, and summaries to enrich image generation. Integrating CLIP-guided SDXL generation is planned to achieve higher semantic alignment between text and images. Faster CLIP variants such as ViT-L/14 will be explored to enhance retrieval precision, along with extending the retrieval process from Top-1 to Top-K for multi-image results. Additionally, prompt robustness will be improved by incorporating contextual embeddings from models like BERT or GPT to capture deeper semantic relationships.

5. Conclusion

This work presents two pipelines - one for image generation and and the other for retrieval in news media applications. The SDXL-based generation framework produces realistic and contextually relevant images through optimized prompts, while the CLIP–FAISS retrieval system enables fast and accurate image matching, even at large scale. Both pipelines are optimized for GPU performance and reliability. Future efforts will focus on improving semantic precision and include ethical considerations.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to check spelling and re-phrase text. After using ChatGPT, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content. The scientific insights, conclusions, and recommendations have been obtained by human authors.

References

[1] 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: Working Notes Proceedings of the MediaEval 2025 Workshop, 2025.

[2] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga., A comprehensive survey of deep learning for image captioning., ACM Computing Surveys (CSUR) 51 (2019) 1–36.

[3] N. Oostdijk, H. van Halteren, M. L. Erkan Başar, The connection between the text and images of news articles: New insights for multimedia analysis., In Proceedings of The 12th Language Resources and Evaluation Conference (2020) 4343–4351.

[4] L. Heitz, A. Bernstein, , L. Rossetto, An empirical exploration of perceived similarity between news article texts and images., MediaEval 2023 Working Notes Proceedings. 3658 (2024).
[5] L. Heitz, Y. K. Chan, H. Li, K. Zeng, A. Bernstein, L. Rossetto, Prompt-based alignment of headlines and images using openclip., MediaEval 2023 Working Notes Proceedings. 3658 (2024).
[6] 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.