Balancing Relevance and Compliance: Text-to-Image Methods for News Articles Visualization

Mahrukh Khan[1], Alishba Subhani[1], Muhammad Rafi[1], and Atif Tahir[2]

[1] *National University of Computer and Emerging Sciences, Pakistan*
[2] *Institute of Business Administration, Pakistan*

Abstract

The paper discusses the approach followed by team CVG-IBA for the task on NewsImage of MediaEval 2025. The shared tasks require to get images for a given news article title through retrieval and generation of images. For retrieval task we proposed two separate methods for small and large subset of data known as SEEK and TRACE, respectively. The main emphasis on retrieval task is on relating textual patterns into image content. The proposed methods for image generation are Virtual Environment for Visual Deep Learning (Vivid) and PromptForge for small and large subset, respectively. The main idea here is to generate realistic images by using unambiguous prompting. Overall, the average score from results suggested that the image generation methods have outperformed the retrieval methods in both the small and large datasets. Best rating of 3.401 was achieved by VIVID exploits the uses illustrations and non-photorealistic image generation. The tasks are evaluated using the Mean Reciprocal Rank (MRR) and Precision@K metrics.

1 INTRODUCTION

The NewsImages task in MediaEval 2025 asks participants to create images that fit news articles while avoiding misleading realism [1].

This NewsImages challenge 2025, is all about pairing news articles with the relevant images doing both retrieval and Generation of images. Online news articles recently more of multimodal to convey the very essences of the news to the reader. A wide spectrum of news domains poses many challenges for pairing up both text and image contents in such a way to serve the best purpose of the news. Previous research has shown that people prefer AI-generated and retrieved images over editorial selection [2], highlighting the importance of automated approaches in news image selection. To identify integral latent connection between the multimodal content is the central idea of this year's competition, both the retrieval and generation subtasks investigate how a combination of image and text features impact the future online journalism and personalized news content delivery to the users. Our team CVG-IBA participated in both shared tasks. For the retrieval approach, we utilized SEEK and TRACE, leveraging the CLIP model [3] to fine-tune image-text alignment for news articles. FAISS (Facebook AI Similarity Search) indexing was employed for efficient image retrieval, enabling quick access to the most relevant images based on article descriptions. On the other hand, for the image generation task, we first used KeyBERT to extract phrases from the article data and based on the frequency and relevance we selected keywords and used a mix of extracted phrases and article title to generate positive prompt for the image generation model. Later we passed the prompts to the custom trained SDXL model for image generation at native resolution. This way we combined the power of contextual prompts with image generation to generate high quality images.

2 APPROACHES

2.1 Retrieval

This section describes the methodology employed in the development of a news image retrieval system, utilizing a fine-tuned CLIP (Contrastive Language-Image Pretraining) model and FAISS for efficient image retrieval. The system aims to match news article descriptions with relevant images by leveraging the multimodal capabilities of CLIP [4].

Fine-Tuning the CLIP Model: To optimize the CLIP model for news image retrieval, the model is fine-tuned on the dataset of news article-image pairs. The training objective is to align image and text representations in a shared feature space, using a contrastive loss function. The CLIP model is trained using the AdamW optimizer, with a learning rate of 5e-5. The model undergoes training for 3 epochs, with contrastive loss computed for each batch. The model's weights are saved after each epoch.

Image Embedding and FAISS Index Construction: Following fine-tuning, the next step is to generate image embeddings. These embeddings are calculated by feeding each image through the CLIP model, which outputs a feature vector

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

mahrukh.khan@nu.edu.pk (M. Khan); alishba.subhani@nu.edu.pk (A. Subhani); muhammad.rafi@nu.edu.pk (M. Rafi); atiftahir@iba.edu.pk (A. Tahir)

CEUR Workshop Proceedings (CEUR-WS.org)

representing the image in the shared embedding space. The embeddings are then normalized to unit length to facilitate cosine similarity-based retrieval.

The FAISS library is utilized to index the image embeddings, allowing for fast nearest-neighbour searches [5]. The indexing process proceeds as follows:

1. **Embedding Computation:** Each image is passed through the fine-tuned CLIP model to generate its embedding.
2. **Indexing:** The embeddings are indexed using a FAISS HNSW (Hierarchical Navigable Small World) index, chosen for its efficient memory usage and fast retrieval capabilities.
3. **Index Storage:** The FAISS index, along with the image IDs, is saved to disk for later retrieval.

Image Retrieval: The image retrieval process involves generating embeddings for a given query and searching for the most similar images in the FAISS index. The process can be triggered either by a custom text query or by providing an article ID. The retrieval process is as follows:

1. **Text Query Embedding:** The input query, which consists of the article title and tags, is tokenized and passed through the fine-tuned CLIP model to produce a text embedding.
2. **Similarity Search:** The text embedding is then used to query the FAISS index. The top-k most similar images are retrieved based on their cosine similarity to the text embedding.
3. **Result Retrieval:** The retrieved image IDs are mapped to their corresponding images, which are then returned as the output. Each result includes the image ID and a similarity score, indicating how closely the image matches the query.

The above methodology **TRACE** is used to retrieve images for the large dataset shared. Whereas, for the retrieval of subset of data (30 images), **SEEK** was implemented which used pretrained CLIP model for retrieval of images based on the article title. The code is available at GitHub repository [6].

2.2 Generation

News thumbnails must capture article semantics without asserting ethical concerns for photo-realistic images; thus, we aim for digital art, illustration, painting, cartoon, and sketches. The approach combines:

Prompt generation: Extract article-level keywords and short phrases (KeyBERT + engineered prompt templates) from the supplied links of the news articles, where the links are invalid the title phrase used instead.

Image generation: Use ComfyUI and a text-to-image stable-diffusion style checkpoint to synthesize images from the generated prompts, the seeds are randomized and native resolution of 1384×784 is used for image generation. This ensures attention to details and high frequency details are embedded into the generated image. This specific resolution is used to ensure the native divide by 8 rule and provide a scaling factor of ~3 to the final image. The step size of 20 was used to provide realistic steps for image generation with CLIP text encoder with 77 tokens and Euler sampler. Finally, VAE decoder was used to convert images from latent space to image space

Post-processing & packaging: The higher resolution image is then center-cropped and rescaled to the required 460×260 image. Finally for reproducibility the seed information is saved as a CSV file.

The method Vivid is the version of PromptForge where the prompts comprise only the titles of the articles. The code is made available at the GitHub repository [7]

3 RESULTS AND DISCUSSION

The retrieved/generated images were evaluated during an online evaluation event where all the participating teams ranked the images using a Likert scale 1-5 (where 1 being the lowest value which is marked as “Very Poor Fit” and 5 being the highest value which is marked as “Very Good Fit”. The results are based on the average rating received for a sample of 50 articles out of 8,500 (selected by the organizers) during the online evaluation event. Some of the images were selected from the released subset of 30 articles. Thus, the dataset is further labelled as small (30 sample articles selected from the full databank of 8,500) and large dataset (full dataset of 8,500 news articles). Table 1 summarizes the achieved results based on the evaluation dataset of 50 prompts and we have observed that the generated images received slightly higher score than the retrieved images.

Table 1: **Comparison of results with baseline**

Dataset	Method	Average Score
Small	Baseline	3.041
	SEEK	3.074
	TRACE	3.080
	PromptForge	3.233
	VIVID	**3.401**
Large	Baseline	2.956
	TRACE	3.114
	PromptForge	**3.194**

VIVID achieved a score of 3.40 for the small dataset while PromptForge achieved an average result of 3.19, this was mainly because of: a) quality of images and attention of details that the generated images carry and b) the power of text-to-image generators which generated the image in an iterative manner in accordance with the prompt. Retrieval task is limited to the training dataset and its labelling, thus for every prompt there can only be a single and this cannot be modified to generate a new image. However, the generated images target the text and generate images accordingly. Furthermore, we have also noticed that retrieval task is not sensitive to the ethical aspects of the challenge, i.e., non-photorealistic images were the main target of this challenge while the retrieval tasks use pre-trained models which heavily rely on the realistic images, thereby generating realistic images which are not suitable for use in news articles.

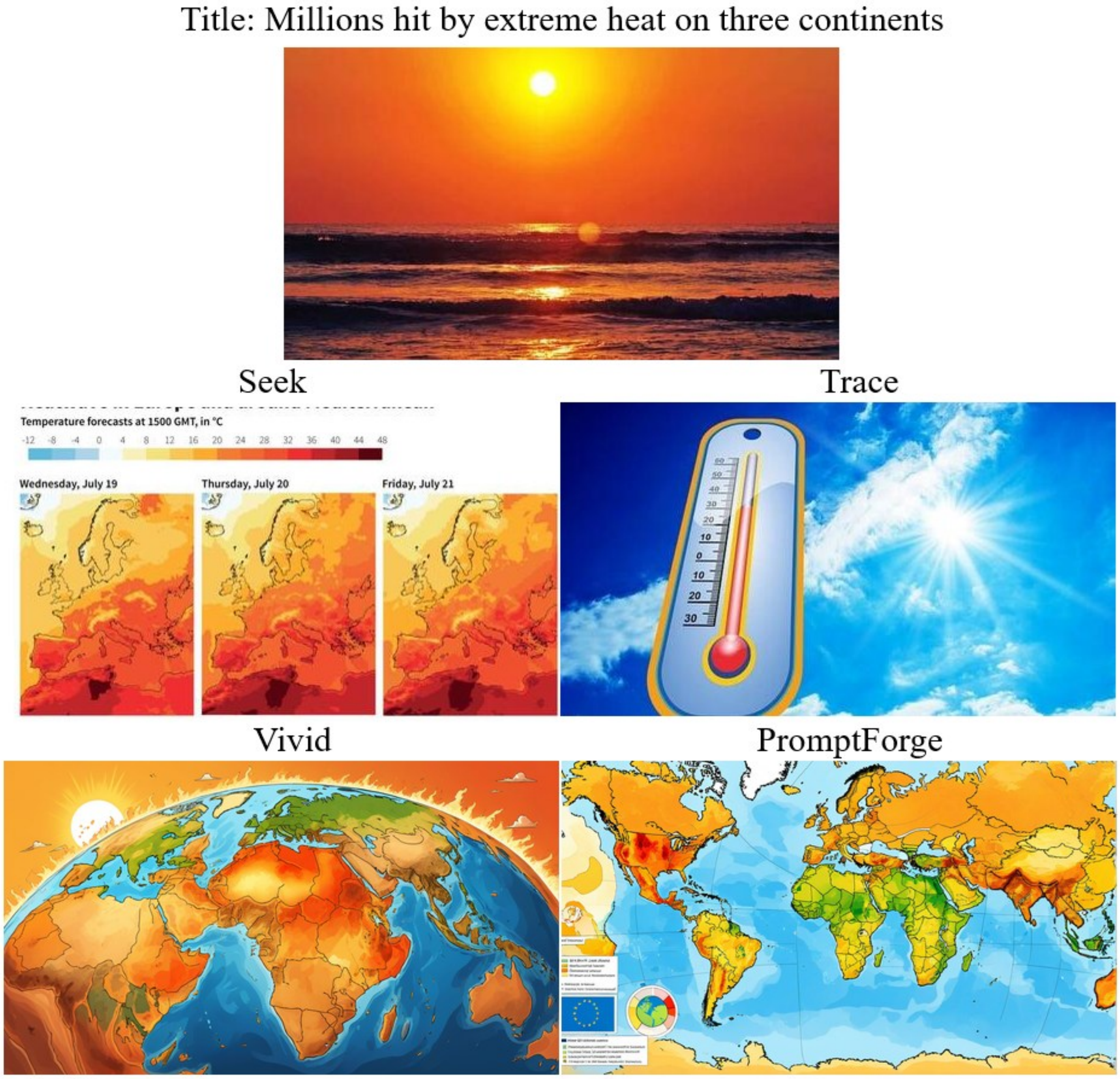

Figure 1: **Comparison of retrieved/generated images with the reference image for article id: 7304.**

Figure 1 shows a sample image with title and reference image on top and the retrieved and generated images at the bottom. Full image submissions are made available at the GitHub repositories [6, 7].

4 CONCLUSION

This paper outlines our approach for the MediaEval 2025 NewsImages task, where we participated in generating and retrieving images for news articles. For image retrieval, we employed SEEK and TRACE, utilizing the CLIP model and FAISS indexing for efficient image matching. In the image generation task, we implemented VIVID and PromptForge, which use Stable Diffusion to create non-photorealistic images. Our results indicate that the generative methods, particularly VIVID, delivered the highest scores, showcasing their ability to produce contextually accurate and ethically appropriate images. These findings highlight the effectiveness of generative models in producing visually compelling thumbnails while adhering to task constraints

ACKNOWLEDGMENTS

We would like to acknowledge the support of Video Surveillance Lab and NLP Lab at National University of Computer and Emerging Sciences, Karachi Campus for providing computer labs and resources to carry out this research activity. We also like to thank the task organizers for managing this shared task very smoothly and providing help and support at every time we posed any query.

DECLARATION ON GENERATIVE AI

During the preparation of this work, the authors used Grammarly for grammar and spelling checks. Further, the authors used AI for generation tasks that were a requirement for the task. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the publication’s content.

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: MediaEval 2025.

[2] L. Heitz, A. Bernstein, L. Rossetto, An Empirical Exploration of Perceived Similarity between News Article Texts and Images, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.

[3] 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, pp. 8748-8763.

[4] Lahajal NK. Enhancing image retrieval: A comprehensive study on photo search using the clip mode. arXiv preprint arXiv:2401.13613. 2024 Jan 24.

[5] Rahman MD, Rabbi SM, Rashid MM. Optimizing Domain-Specific Image Retrieval: A Benchmark of FAISS and Annoy with Fine-Tuned Features. arXiv preprint arXiv:2412.01555. 2024 Dec 2.

[6] A. Subhani, Trace: text-to-image retrieval for Mediaeval, (2025) NewsImage task, *GitHub repository*, https://github.com/AlishbaSubhani/Newsimages-Image-Retrieval-Mediaeval-Competition-2025.

[7] M. Khan, PromptForge: text-to-image generator using task specific attention-based keywords, (2025), *GitHub repository*, https://github.com/mahrukhkhan0992/MediaEval2025.