

Comparing CLIP-Based Image Retrieval and Stable Diffusion Turbo for News Article Thumbnails

Sakthi Mukesh Thanga Mariappan^{1,*†}, Muthulakshmi Ramasamy^{2,†} and Beulah Arul^{1,†}

¹Rajalakshmi Engineering College, Chennai, Tamil Nadu, India

²Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India

Abstract

This paper presents Team CodingSoft-REC's submission to the NewsImages challenge at MediaEval 2025, which explores matching news articles with appropriate thumbnail images. We implemented two distinct approaches: (1) a retrieval-based method using CLIP embeddings with FAISS indexing to search the YFCC100M dataset, and (2) a generation-based method using Stable Diffusion Turbo to synthesize images from article text. Our system processed 8,500 news articles, creating automated workflows for both small-scale and large-scale image recommendation tasks. Evaluation results from crowd-sourced assessments indicate that our generated images using SD-Turbo achieved higher perceived fit scores (average: 2.68) compared to retrieved images using CLIP (average: 1.92) on a 5-point Likert scale. This suggests that text-to-image generation models may offer promising alternatives to traditional retrieval methods for news illustration, particularly when paired with appropriate style prompting to ensure editorial standards.

1. Introduction

Visual elements are crucial for engaging readers with online news content, with images serving as primary attention drivers on digital platforms [1, 2]. However, manually selecting appropriate thumbnail images for news articles is time-consuming and challenging, particularly given the rapid pace of news cycles and the need for images that accurately represent diverse story types without misleading readers [3].

The NewsImages challenge at MediaEval 2025 addresses this challenge by comparing two automated approaches: retrieval-based methods that search existing image databases, and generation-based methods that synthesize new images using AI. This comparison is particularly relevant given recent advances in vision-language models like CLIP and text-to-image diffusion models like Stable Diffusion [4]. A key question remains: **Can AI-generated stylized illustrations match or exceed the semantic relevance of retrieved photographs for news article thumbnails?**

Our approach explores this question through two parallel implementations. For retrieval, we employ CLIP embeddings with FAISS indexing to efficiently search the YFCC100M database of Creative Commons images. For generation, we utilize Stable Diffusion Turbo with carefully engineered prompts that emphasize non-photorealistic styles to address ethical concerns about synthetic news imagery. Our central hypothesis is that generation methods, despite creating

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

*Corresponding author.


† These authors contributed equally.

✉ sakthimukesh@gmail.com (S. M. T. Mariappan); muthulakshmiir@gmail.com (M. Ramasamy);

beulah.a@rajalakshmi.edu.in (B. Arul)

ORCID: 0009-0005-3044-6841 (S. M. T. Mariappan); 0009-0009-9911-5812 (M. Ramasamy); 0000-0002-3891-0806 (B. Arul)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

synthetic content, may achieve better semantic alignment with article content due to their flexibility in representing abstract concepts that lack suitable photographic matches.

We participated in both the large-scale fully automated task (8,500 articles) and small-scale human-in-the-loop task (30 articles) [5]. This paper documents our implementation, presents evaluation results comparing both approaches, and analyzes the performance gap with human-selected baseline images.

2. Related Work

Recent advances in vision-language models have transformed cross-modal retrieval capabilities. CLIP [6], developed by Radford et al., learns joint embeddings of images and text through contrastive pre-training on 400 million image-text pairs, achieving remarkable zero-shot transfer without task-specific fine-tuning [7]. For news image selection, CLIP’s ability to match semantic concepts across modalities is particularly valuable, though its training on general web content may not capture news-specific visual conventions.

The NewsImages challenge context presents unique complications for image-text matching. Unlike standard caption-image datasets, news articles often require illustrative imagery—stock photos, archived content, or symbolic graphics—rather than literal depictions of events. Previous NewsImages iterations revealed that CLIP-based retrieval struggles when the YFCC100M database lacks semantically relevant photographs for abstract topics like economic policy or diplomatic relations, suggesting a fundamental mismatch between user-generated travel photography and news illustration needs.

Diffusion-based generative models [4], particularly Stable Diffusion and its accelerated variant SD-Turbo, offer an alternative by synthesizing images directly from text prompts. This flexibility potentially addresses retrieval limitations for abstract concepts. However, photorealistic generation raises critical concerns for journalism: authenticity, editorial transparency, and the risk of misleading audiences [8]. The MediaEval 2023 NewsImages challenge first explored generation approaches, finding that stylized generated images achieved comparable or better perceived fit than retrieved photographs for certain topics.

Our work directly addresses this retrieval-generation trade-off in the NewsImages context by systematically comparing both methods at scale while emphasizing non-photorealistic generation styles to maintain editorial standards and reader trust.

3. Dataset

The challenge organizers provided a comprehensive dataset comprising 8,500 English-language news articles collected by GDELT during 2022-2023 from various international publishers and media outlets.

For the retrieval task, participants were required to source images from the Yahoo Flickr Creative Commons 100 Million (YFCC100M) dataset [9], one of the largest public multimedia collections. This dataset contains approximately 100 million images with associated metadata, organized in a hierarchical directory structure. The images represent diverse subjects, locations, and visual styles captured by users worldwide and released under Creative Commons licenses. The dataset’s hierarchical organization and the absence of detailed textual descriptions for many images necessitate robust content-based retrieval methods [10].

4. Approach

4.1. Image Retrieval with CLIP

We employed the OpenAI CLIP model with the ViT-B/32 [11] architecture for the retrieval task. CLIP consists of two encoder networks trained jointly using a contrastive objective, enabling aligned embeddings where semantically similar images and texts are positioned close together [12]. This paper presents Team CodingSoft-REC’s submission to the NewsImages challenge at MediaEval 2025.¹

For each news article, we constructed a comprehensive text representation by combining the article title and tags with the format: “Title. Keywords: tag1 tag2 tag3.” This text was encoded using CLIP’s text encoder to generate 512-dimensional embeddings.

To enable efficient retrieval from YFCC100M, we pre-computed image embeddings for approximately 300,000 images using CLIP’s image encoder. We employed the FAISS [13] library with IndexFlatIP for fast similarity search, using L2-normalized vectors for cosine similarity computation.

4.2. Image Generation with Stable Diffusion Turbo

For the generation task, we selected Stable Diffusion Turbo (SD-Turbo), an optimized variant designed for rapid inference with minimal quality degradation. Our pipeline was configured with FP16 precision, 464×264 pixel generation resolution, and 2 inference steps.

We developed a prompt engineering strategy that incorporates style guidance:

Prompt Template:

[Article Title] [Article Tags], in style of digital illustration, high detail, non photo-realistic

This template combines semantic content from the article with style modifiers that encourage stylized rather than photo realistic output, addressing ethical concerns about generated images being mistaken for real photographs.

5. Results

The challenge employed crowd-sourced evaluation where participants rated image-article fit on a 5-point Likert scale ranging from (1) Very Poor Fit to (5) Very Good Fit. Ratings were aggregated by computing group averages across all raters.

Table 1 presents overall ratings for our four submission runs. Generation consistently outperformed retrieval (2.67-2.68 vs. 1.90-1.93), representing a 40% improvement. Performance remained stable between SMALL and LARGE subtasks, demonstrating successful scalability. Generated images showed lower standard deviation (0.73-0.75 vs. 0.89-0.91), indicating more consistent quality—likely because generation tailors images to each article while retrieval is constrained by database availability.

However, all automated approaches scored below 3.0 (Average Fit). Table 2 compares against the BASELINE of original editorial images (2.99 overall). Generation trails by only 0.31 points, whereas retrieval lags by 1.07 points. This narrower gap is noteworthy given our deliberately non-photorealistic generation style to comply with editorial transparency guidelines. The retrieval approach’s poor performance likely stems from database mismatch—YFCC100M contains

¹Code available at: <https://github.com/SakthiMukesh7905/Mediaeval-Newsimage>

Table 1
Overall Average Ratings for CodingSoft-REC Submissions

Run Name	Method	Task	Avg. Rating	Std. Dev.
CodingSoft-REC_CLIP	Retrieval	SMALL	1.90	0.89
CodingSoft-REC_SDTURBO	Generation	SMALL	2.67	0.75
CodingSoft-REC_CLIP	Retrieval	LARGE	1.93	0.91
CodingSoft-REC_SDTURBO	Generation	LARGE	2.68	0.73

primarily travel and lifestyle photography rather than news-relevant imagery. For abstract concepts (economic policy, diplomatic tensions), semantically appropriate photographs simply do not exist, forcing CLIP to select weakly related images. The generation approach’s relative success suggests that flexibility in creating symbolic or illustrative imagery matters more than photographic authenticity for news thumbnails. Our prompt engineering strategy—explicitly requesting "digital illustration, non photo-realistic" styles—appears effective at both improving semantic relevance and addressing ethical concerns about misleading synthetic imagery.

Table 2
Comparison with Editorial Baseline

Approach	SMALL Task	LARGE Task	Overall
BASELINE (Editorial)	3.041	2.956	2.99
CodingSoft-REC_SDTURBO	2.67	2.68	2.68
CodingSoft-REC_CLIP	1.90	1.93	1.92

6. Discussion and Outlook

This paper compared two automated approaches for news thumbnail selection in the MediaEval 2025 NewsImages challenge: CLIP-based retrieval from YFCC100M and Stable Diffusion Turbo generation with style-guided prompts. Our evaluation on 8,500 articles reveals that text-to-image generation fundamentally outperforms retrieval when matching images to news content.

In our contribution, we demonstrate that deliberate non-photorealistic styling addresses ethical concerns about misleading synthetic imagery while maintaining competitive performance, offering a practical path for deploying generation in journalistic contexts.

Future work should focus on hybrid systems that leverage retrieval’s photographic authenticity when appropriate matches exist while defaulting to generation for abstract topics, alongside interactive tools that enable journalists to iteratively refine AI-generated candidates through prompt adjustment.

Declaration on Generative AI

During the preparation of this work, the author used generative AI tools solely for grammar and spelling checks. All content including approach, analysis, and discussion has been prepared and critically reviewed by the authors to ensure originality and clarity. Also no sentence has been generated by AI to write up this paper.

References

- [1] Y. De Haan, S. Kruikemeier, S. Lecheler, G. Smit, R. Van der Nat, When does an infographic say more than a thousand words? audience evaluations of news visualizations, *Journalism Studies* 19 (2018) 1293–1312.
- [2] H. Caple, Visual media: The importance of visuals as partners in the news, in: *The Routledge Handbook of Language and Media*, Routledge, 2017, pp. 230–243.
- [3] T.-P. Chang, T.-C. Hsiao, T.-L. Chen, T.-C. Lo, A framework for detecting fake news by identifying fake text messages and forgery images, *Enterprise Information Systems* 19 (2025) 2436495.
- [4] 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*, 2022, pp. 10684–10695.
- [5] 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.
- [6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: *International conference on machine learning*, PmLR, 2021, pp. 8748–8763.
- [7] A. Khandelwal, L. Weihs, R. Mottaghi, A. Kembhavi, Simple but effective: Clip embeddings for embodied ai, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14829–14838.
- [8] 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.
- [9] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L.-J. Li, YFCC100M: The new data in multimedia research, *Communications of the ACM* 59 (2016) 64–73. URL: <http://cacm.acm.org/magazines/2016/2/197425-yfcc100m/fulltext>.
- [10] A. Gordo, J. Almazán, J. Revaud, D. Larlus, Deep image retrieval: Learning global representations for image search, in: *European conference on computer vision*, Springer, 2016, pp. 241–257.
- [11] S. Hong, J. Wu, L. Zhu, W. Chen, Brain tumor classification in vit-b/16 based on relative position encoding and residual mlp, *Plos one* 19 (2024) e0298102.
- [12] 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.
- [13] B. Sriman, S. A. Silviya, Y. Mouleesh, S. Vinod, S. Nishanthini, P. Nikitha, Intelligent document interaction with advanced vector embeddings and faiss-cpu indexing, in: *2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, 2024, pp. 1–6.