CLIP-Based News Image Retrieval with Keyword-Enriched Prompts for MediaEval 2025 NewsImages Task

Ramcharan Swaminathan[1,†], Sakthivel Thangapandiyan[1,†], Mirunalini Palaniappan[1,*,†] and Saravanan Esakki[1,†]

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

Abstract

News articles are increasingly published with accompanying images, yet the process by which these visuals are chosen, generated, or matched to the text is often opaque. This results in unclear relationships between textual and visual content. This paper presents the **SSN-CSE** team's approach for the **MediaEval 2025 NewsImages** task, focusing on the retrieval subtask that recommends suitable images for news articles by learning the relationships between text and visuals.

The proposed system leverages the **CLIP ViT-B/32** model for cross-modal retrieval between article text and candidate images from the YFCC100M dataset. To enhance textual representations, Named Entity Recognition (spaCy) and keyword extraction (KeyBERT) are integrated to generate enriched prompts. Cosine similarities between enriched text embeddings and pre-computed image embeddings are then used to retrieve the most relevant image per article.

The method was evaluated on both the **SMALL** and **LARGE** subtasks. The system achieved an **average rating of 2.68 / 5** on the SMALL task and **2.67 / 5** on the LARGE task (8 500 articles). These results highlight practical scalability of a CLIP-based retrieval pipeline with prompt enrichment and provide insights into bridging abstract textual concepts with visually meaningful representations.

1. Introduction

Modern online news articles are multimodal, combining textual and visual elements. Images play a crucial role in illustrating stories and attracting readers. However, image selection is often ambiguous—recent news may lack dedicated visuals, leading to the use of stock or AI-generated images. The **MediaEval 2025 NewsImages** task challenges participants to recommend relevant images for news articles. The **SSN-CSE** approach aims to generate visually meaningful thumbnails that enhance comprehension by bridging the gap between abstract textual concepts and concrete visual representations.

2. Related Work

Image–text retrieval has advanced through contrastive vision–language models such as **CLIP** [1], which enable zero-shot transfer across domains. Prior MediaEval studies [2] demonstrated that CLIP-based retrieval surpasses simpler baselines. The 2025 NewsImages overview [3] further compared retrieval and generative systems.

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

*Corresponding author.

†These authors contributed equally.

ramcharan2310608@ssn.edu.in (R. Swaminathan); sakthivel2310758@ssn.edu.in (S. Thangapandiyan); miruna@ssn.edu.in (M. Palaniappan); saravanan2310681@ssn.edu.in (S. Esakki)

The proposed approach differs from previous work by combining CLIP with **text enrichment** through NER (spaCy) and keyword extraction (KeyBERT), thereby improving contextual grounding for abstract headlines. Keyword and entity extraction have been applied in information retrieval [4] and multimedia annotation [5]. Studies comparing BLIP-2 and CLIP [6] and those employing VSE++ [7] indicate that prompt quality strongly affects alignment accuracy. These insights are extended here to large-scale media retrieval.

3. Proposed Approach

The system consists of three main stages: (1) semantic text enrichment, (2) multimodal embedding extraction, and (3) similarity-based retrieval. The overall workflow is shown in Figure 1.

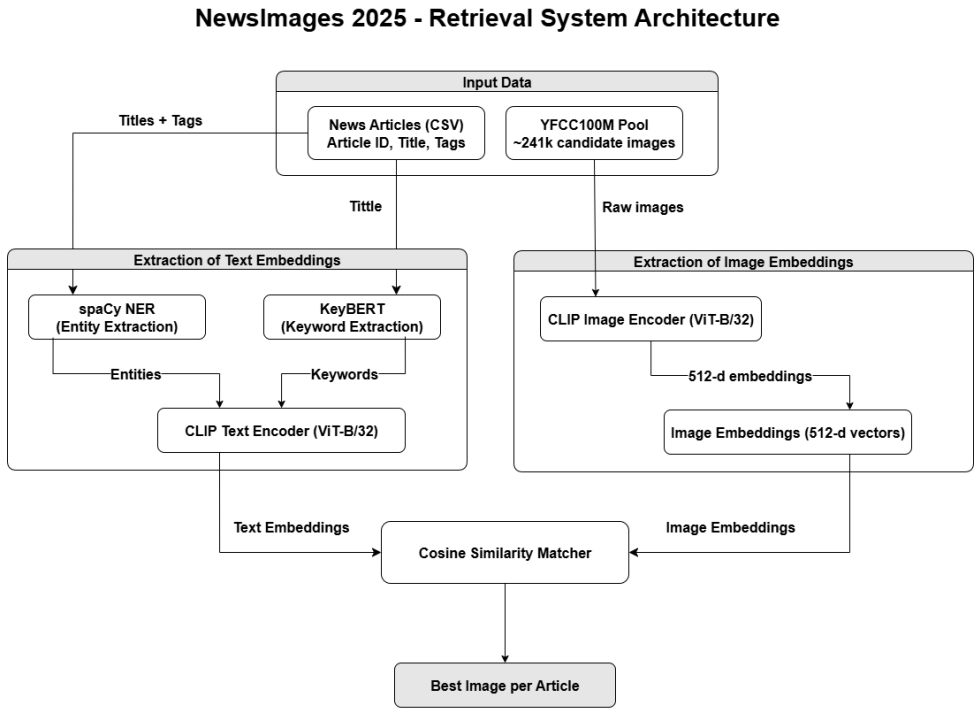

Figure 1: Architecture of the proposed SSN-CSE system.

3.1. Data Preparation and Preprocessing

The Proposed System uses the MediaEval 2025 NewsImages dataset containing about 8 500 articles with titles, tags, and metadata. Candidate images are drawn from 241 000 entries of YFCC100M, providing diverse real-world visuals.

3.2. Semantic Text Enrichment

Named Entity Recognition (spaCy `en_core_web_sm`) extracts visually relevant entities such as organizations, locations, and events. KeyBERT [4] complements this by selecting the top five

representative keywords. The enriched prompt is defined as:

$$P_{\text{enriched}} = T_{\text{title}} + \text{“. Keywords: ”} + \{e_1, e_2, \ldots, k_1, k_2, \ldots\} \tag{1}$$

3.3. Embedding Extraction and Retrieval

CLIP ViT-B/32 [1] is used to extract 512-dimensional embeddings. Image and text embeddings are L2-normalized, and cosine similarity is computed as:

$$\text{sim}(\hat{v}_i, \hat{t}) = \hat{v}_i^T \hat{t}, \qquad I^* = \arg\max_{I_i} \text{sim}(\hat{v}_i, \hat{t}) \tag{2}$$

This enables efficient retrieval while preserving semantic alignment.

4. Results and Analysis

The system was evaluated on both SMALL and LARGE subtasks using the official organizer ratings. **Note:** The baseline corresponds to *editor-selected reference images*, not a separate algorithmic system.

Table 1
Performance comparison of SSN-CSE (CLIP + KeyBERT + YFCC) vs. baseline (editor-selected).

System	Task	Average Score	Consistency
SSN-CSE	SMALL	2.68	0.003
SSN-CSE	LARGE	2.67	0.003
Baseline (editor)	SMALL	3.04	0.085
Baseline (editor)	LARGE	2.96	0.085

Figure 2 shows a qualitative example for the article “National Park Service Warning After Bison Attack.” The method retrieves contextually appropriate wildlife imagery.

(a) Proposed system (b) Baseline (editor)

Figure 2: Qualitative comparison for “National Park Service Warning After Bison Attack.”

Although the averages are below the editor baseline, the pipeline is straightforward to reproduce and scales to larger datasets. The “consistency” values in Table 1 are reported descriptively from the original analysis and are not intended as a superiority claim. Future iterations will incorporate retrieval-oriented ranking metrics (e.g., Recall@k, nDCG@k, MAP@k) alongside task ratings and explore re-ranking with concept expansion.

5. Conclusion

A CLIP-based news-image retrieval framework with keyword and entity enrichment has been presented. While average ratings are below the human-curated baseline, the approach remains simple, reproducible, and scalable. The qualitative analysis indicates clear next steps: add retrieval metrics tailored to ranking quality, adapt to news domains, and apply lightweight re-ranking to better handle abstract content and culturally bound concepts.

Code Availability

The workflow and scripts are available at: https://github.com/Ramcharan-Swaminathan/SSN-CSE.git

Declaration on Generative AI

During the preparation of this work, GPT-5 was used for grammar, spelling, and formatting assistance. All text was manually reviewed, and the authors take full responsibility for the final content.

References

[1] 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.

[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] 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.

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

[5] L. Heitz, L. Rossetto, B. Kille, A. Lommatzsch, M. Elahi, D.-T. Dang-Nguyen, Mediaeval 2023 newsimages task overview, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2023.

[6] 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.

[7] A. Elliah, M. P, K. V, H. Bharathi, A. Bhaskar, Vithula, Connecting text and images in news articles using vse++, 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/paper26.pdf.