Cross-modal Image Recommendation for News Articles by Multimodal Foundation Models-based Retrieval-Reranking

Damianos Galanopoulos[1], Andreas Goulas[1,2] and Vasileios Mezaris[1]

[1]Information Technologies Institute (ITI), Centre of Research and Technology Hellas (CERTH), Thessaloniki, Greece
[2]Queen Mary University of London (QMUL), London, UK

Abstract

Retrieving relevant images for a given news article is challenging and can be considered a special version of the cross-modal retrieval problem. This notebook paper presents our solution for the MediaEval NewsImages 2025 benchmarking task. We propose a retrieval-reranking solution based on multimodal foundation models such as VLMs and multimodal LLMs, and utilizing multiple levels of textual granularity. We report the official results of our submitted runs and additional experiments we conducted internally to evaluate our runs.

1. Introduction

In this paper, we, the CERTH-ITI team[1], deal with the text-to-image retrieval task adapted to the needs of the MediaEval NewsImages 2025 task [1, 2]. The lesson learned from our previous participation in NewsImages 2022 [3] and 2023 [4] is that leveraging cross-modal networks performs optimally. Moreover, the recent rise of multimodal large language models (MLLMs) [5, 6] expands the horizons of the text-to-image retrieval task. Based on these outcomes, this year we propose a two-stage retrieval and reranking architecture, leveraging foundation models such as multiple vision-language models (VLMs) for image retrieval and a large multimodal large language model for contextual reranking. Moreover, to go beyond the news article's title and consider richer textual information, we utilize *state-of-the-art* MLLMs to produce multiple representations, such as the article's full text summary and a recommended image caption.

2. Related Work

The text-image association task is similar to the NewsImages benchmarking task and has gained a lot of interest in recent years, focusing on understanding the relationship between textual and visual content in online news articles. Initial approaches were focusing on training networks specifically for the task's needs. In Pivovarova et al. [7], a visual topic model was proposed that aligns image-illustrated topics with textual topics through knowledge distillation training. In Zhang et al. [8], an object detector is pre-trained to encode images and the objects within them, while a cross-modal model is trained to associate visual features with textual features. Moreover, Galanopoulos et al. [3] explored CLIP's capabilities alongside a trainable cross-modal network, concluding that using CLIP was, by a small margin, better than training a custom cross-modal network. The CLIP model was widely used in previous approaches, e.g., [4, 9]. With the

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online
dgalanop@iti.gr (D. Galanopoulos); agoulas@iti.gr (A. Goulas); bmezaris@iti.gr (V. Mezaris)

[1]The code is available at https://github.com/IDT-ITI/NewsImages-MediaEval2025.

introduction and widespread usage of LLMs, more recent approaches choose to utilize the power of these models. Truong-Vinh et al. [10] utilizes prompt engineering techniques to instruct ChatGPT [11] in generating descriptive captions for image candidates to a given news article. Recent studies in similar research fields, such as text-based video retrieval [12, 13], highlight that the use of large foundation models significantly improves performance. Inspired by these works, we proposed an MLLM-based approach to reranking the retrieved potentially-relevant images.

3. Approach

3.1. Data pre-processing

For each news article, the organizers provide a set of metadata, including the article's URL, its title, and the corresponding image URL. The preliminary stage of our methodology is dedicated to the pre-processing of these data, focusing on enhancing the representations associated with each news article. After collecting the available news articles' full text from the web, we consider four distinct representations of each article:

- **Article Title** – The original title provided by the organizers, serving as the primary and most concise textual representation.
- **Article Summary** – A summary generated from the article's full text using the *LLaMA-3.2-1B-Instruct* [14, 15] MLLM, offering a brief overview of the article.
- **Generated Image Caption** – A descriptive caption, generated by applying the *LLaMA-3.2-1B-Instruct* model to the article's full text, and prompting it to express in a single sentence how a related image might be described.
- **Provided Reference Images** – The images originally provided with each article.

3.2. Retrieval

The first stage of our method builds on the video retrieval framework introduced in [12], adapting its multi-model retrieval strategy to news article–image association. Unlike the original trainable pipeline, our approach is fully non-trainable and relies solely on pre-trained VLMs. The Retrieval stage is performed in two phases: the offline image indexing phase and the online article processing phase. At the offline phase, we employ five VLM families — CLIP, BLIP, BLIP-2, SLIP, and BEiT-3 — each with its own pre-trained model variants; all candidate images (from a large pool of images; see Section 4 for the used image test dataset) are indexed by these models to extract their embeddings. At the online phase, one of the article representations described in Section 3.1 (i.e., the title, summary, generated caption or the provided image) is used as query. Using the same VLMs families, for each pre-trained model, the selected representation is encoded into the same shared embedding spaces as the candidate images. Cosine similarity is then computed between the embeddings of a possible article-image pair, and the resulting similarity scores are aggregated within each model family to obtain a single relevance score per family. Finally, the scores across all families are combined using a weighting scheme and used to generate the ranked list of the top-N most relevant images corresponding to the query-article.

3.3. Reranking

The second stage of our approach focuses on refining the initial top-N images returned from the Retrieval stage for each news article. To this end, we employ the *Qwen2.5-VL-7B* model [16],

an instruction-tuned MLLM designed to assess fine-grained semantics between textual and visual content. For each article, the article's title (regardless of which representation of the articles was utilized in the Retrieval stage) and each of the top-N images are jointly processed by *Qwen2.5-VL-7B*, which generates contextual relevance scores reflecting the degree of semantic alignment between the text and each image. These scores are used to reorder the candidate images, resulting in the final ranking that more accurately reflects the contextual relationship between the news content and the candidate images.

4. Submitted Runs and Results

Following the task rules, we utilize, as a test dataset, a subset of the YFCC100M dataset used by OpenAI for CLIP [17, 18]. Furthermore, to internally evaluate our approach, we test it on the NewsImages 2022 training data [19]. We submitted six runs on this year's dataset as follows:

- **Run #1**: Full pipeline using article title.
- **Run #2**: Full pipeline using the article-provided image instead of the article title.
- **Run #3**: Full pipeline using the generated summary.
- **Run #4**: Full pipeline using the generated image captions.
- **Run #5**: Only retrieval stage using the article title without reranking.
- **Run #6**: Fusion of Run #1 and Run #4, where the same MLLM as in Section 3.3 determined which of the two images is most relevant to the article.

Table 1
Evaluation results (Average ratings) of the official collaborative online evaluation event. The best / second-best results are in bold / underline, respectively; higher values are better.

	Baseline	Run #1	Run #2	Run #3	Run #4	Run #5	Run #6
AVG rating ↑	3.041	2.857	2.703	2.644	2.709	<u>2.860</u>	**2.893**

We present the official results (Table 1) and results from the internal experiments (Table 2) we conducted in order to evaluate our methods and select our final runs. Regarding the official results, the evaluation metric is the "average rating", which expresses human ratings from a crowdsourced event on a 5-point Likert scale rating. Differently from this, in our internal experiments the Recall@K, where $K = 1, 5, 10, 50, 100$ and Mean Reciprocal Rank (MRR) are used as evaluation metrics. Each of these two experimental setups presents its own challenges; and, the official evaluation relies on human judgment, while internal experiments exploit existing ground-truth text-image associations and standard retrieval metrics. Please note that Run #2 can not be applied to our internal experiments, since in these experiments the set of article-provided images also serves as the image test dataset (in place of the YFCC100M subset used in the official runs).

Table 1 presents the results as concluded from the online evaluation event, where members of the participation teams evaluated the submitted images for every query-article. We submitted results only for the SMALL dataset [1]. We also compare our runs with the provided *Baseline*, where the original images that are part of the articles were assessed as if they were another submitted run [1]. Run #1, where we utilize only the article's title and perform reranking, performs similarly to Run #5, where only the retrieval stage is used (AVG ratings of 2.857 and 2.860, respectively). This suggests that the retrieval stage using the title alone is already

Table 2
Results on our internal evaluation dataset (NewsImages 2022 training data) in R@1,5,10,50,100 and Mean Reciprocal Rank (MRR) terms. The best / second-best results are in bold / underline, respectively. "–" means results are not available. Higher values are better.

Run_id	R@1 ↑	R@5 ↑	R@10 ↑	R@50 ↑	R@100 ↑	MRR ↑
Our best 2023 [4] run	-	0.456	0.54	<u>0.716</u>	<u>0.783</u>	0.357
Run #1	<u>0.446</u>	<u>0.529</u>	<u>0.568</u>	0.67	0.736	<u>0.49</u>
Run #2	-	-	-	-	-	-
Run #3	0.428	0.498	0.537	0.639	0.694	0.47
Run #4	0.430	0.514	0.542	0.680	0.746	0.48
Run #5	0.282	0.477	0.563	**0.748**	**0.802**	0.38
Run #6	**0.458**	**0.531**	**0.574**	0.703	0.765	**0.51**

reasonably effective, and the full pipeline in Run #1 (i.e., the addition of reranking) does not lead to improvements. Run #2, where the article-provided images are used, shows a slight decrease in AVG rating (2.703), indicating that directly using the article's image as input is less effective than using the title. This means that the retrieved images may visually resemble the original image of each article, but the meaning conveyed can be different, leading to lower performance. Similarly, in Run #3, where richer textual information is used, the performance also decreases compared to Run #1 and Run #1. Finally, combining Run #1 and Run #4 by letting the MLLM determine which run decision is most suitable for the given article increases performance to 2.893, as it adjusts some mismatches and leverages complementary strengths from both runs.

The above official results differ from the findings of our internal experiments, which were conducted during the submission preparation period and prior to the release of the official results. Table 2 presents these results. From these preliminary experiments, we concluded that Run #6 outperforms the rest of the runs in R@1 and R@5, which are the most suitable Recall depths for making conclusions for this year's task. Moreover, Run #1 and Run #5 perform just as well as Run #6. When we use the Retrieval-only stage (Run #5), the performance starts to increase after a depth of 50. Finally, the utilization of the article's summary (Run #3) is also a well-performing approach. However, these results are different from the official results, probably due to the somewhat different nature of the task and the differences in the evaluation procedure.

5. Conclusion

In this work, we proposed a solution for the MediaEval NewsImages task leveraging state-of-the-art multimodal foundation models within a two-stage retrieval–reranking framework. Our experiments demonstrated that using only the article's title as input is effective, achieving competitive performance compared to more complex approaches. In contrast, incorporating richer textual inputs—such as summaries or the original article-provided images—tended to introduce semantic drift, resulting in weaker alignment with human judgments. Overall, our findings indicate that the complementary title and image-based retrieval strategy combined with an MLLM-driven reranking stage offers a robust and efficient solution for the task.

Acknowledgements

This work was supported by the EU's Horizon Europe programme under grant agreement 101070190 AI4Trust.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take 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: Working Notes Proceedings of the MediaEval 2025 Workshop, 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] D. Galanopoulos, V. Mezaris, Cross-modal Networks and Dual Softmax Operation for MediaEval NewsImages 2022, in: Working Notes Proceedings of the MediaEval 2022 Workshop, volume 3583, CEUR Workshop Proceedings, 2023.

[4] A. Leventakis, D. Galanopoulos, V. Mezaris, Cross-modal Networks, Fine-Tuning, Data Augmentation and Dual Softmax Operation for MediaEval NewsImages 2023., in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.

[5] A. Yang, B. Yang, B. Zhang, et al., Qwen2.5 technical report, arXiv preprint arXiv:2412.15115 (2024).

[6] S. Bai, K. Chen, X. Liu, et al., Qwen2. 5-vl technical report, arXiv preprint arXiv:2502.13923 (2025).

[7] L. Pivovarova, E. Zosa, Visual Topic Modelling for NewsImage Task at MediaEval 2021, in: Working Notes Proceedings of the MediaEval 2021 Workshop, Online, 13-15 December 2021, volume 3181 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021.

[8] P. Zhang, X. Li, et al., VinVL: Revisiting visual representations in vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5579–5588.

[9] 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: MediaEval, 2023.

[10] H.-C. Truong-Vinh, D.-K. Ta, D.-D. Nguyen, L.-T. Nguyen, Q.-V. Nguyen, Beyond Keywords: ChatGPT's Semantic Understanding for Enhanced Media Search, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.

[11] OpenAI, ChatGPT (GPT-4), https://chat.openai.com/, 2024. Large language model, accessed 2024-10-14.

[12] D. Galanopoulos, A. Goulas, A. Leventakis, et al., An LLM Framework for Long-Form Video Retrieval and Audio-Visual Question Answering Using Qwen2/2.5, in: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2025, pp. 3730–3739.

[13] J. Wu, C.-W. Ngo, W.-K. Chan, S.-H. Zhong, Llm-based query paraphrasing for video search, CoRR abs/2407.12341 (2024).

[14] Meta, Llama-3.2-1b-instruct, 2024. URL: https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct, accessed: 2025-10-05.

[15] A. Grattafiori, A. Dubey, A. Jauhri, et al., The Llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[16] S. Bai, K. Chen, X. Liu, et al., Qwen2.5-VL Technical Report, Technical Report, arXiv, 2025. arXiv:2502.13923, Technical Report, Qwen2.5-VL series.

[17] A. Radford, J. W. Kim, C. Hallacy, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[18] A. Ramesh, M. Pavlov, G. Goh, et al., Zero-shot text-to-image generation, in: Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8821–8831. URL: https://proceedings.mlr.press/v139/ramesh21a.html.

[19] A. Lommatzsch, B. Kille, Ö. Özgöbek, M. Elahi, D.-T. Dang-Nguyen, News Images in MediaEval 2022, in: Working Notes Proceedings of the MediaEval 2022 Workshop, volume 3583, CEUR Workshop Proceedings, 2023.