

Beyond Similarity: Two-Stage Retrieval for News Image Search

Bulat Khaertdinov^{1,*}, Aashutosh Ganesh¹, Mirela Popa¹ and Nava Tintarev¹

¹Maastricht University, Netherlands

Abstract

Thumbnail images for news articles must reflect content, meet editorial standards, and appeal to readers. Prior work shows that even editor-chosen thumbnails often diverge from user preferences, underscoring the subjectivity of relevance. We propose a two-stage framework that first retrieves candidates with pre-trained vision-language models (VLMs) and then re-ranks them using one of the three modules: (1) an aesthetics-based scorer, (2) an LLM-rewriting component that enriches article titles with semantically related captions, and (3) a VLM-Judge evaluator for editorial quality. Offline experiments with CLIP-base, CLIP-large, and SigLIP-2 show consistent improvements with VLM-Judge and LLM-rewriting re-ranking strategies, with CLIP-large achieving the best performance. Online user studies further reveal that VLM-Judge scores approach those of professional editors, suggesting that AI-assisted re-ranking can effectively support editorial workflows by improving relevance and aligning with reader expectations. The code is available via the following link: https://github.com/bulatkh/newsimages_um_rtl


1. Introduction


Selecting suitable thumbnails to accompany news articles is a non-trivial task. Beyond illustrating the content, chosen images are expected to meet editorial standards and align with user preferences. Prior work has shown that even editor-selected thumbnails do not always match with reader preferences [1], underscoring the difficulty of capturing relevance to end users. Vision-language models (VLMs) such as CLIP [2] and SigLIP [3] enable zero-shot image retrieval by embedding text and images in a shared space, and have shown promise for news thumbnail retrieval [4]. Retrieval quality can be improved in two-stage architectures, where an initial VLM-based retrieval is followed by re-ranking. Prior re-ranking strategies have incorporated aesthetic and editorial cues [5] and query expansion with large language models (LLMs) [6]. Another research line investigated model-centric approaches that refine multimodal representations through additional training [7, 8]. However, such approaches often depend on ground-truth annotations, which can be noisy, subjective, and biased in the context of news image recommendations [1].

Inspired by the potential of two-stage image search systems, our team (DACS-UM-RTL) presents a retrieval framework that combines large-scale candidate retrieval with modular re-ranking strategies. Our system leverages a pre-trained CLIP-like VLM as a candidate generator and incorporates three complementary re-ranking modules: (1) an aesthetics-based scorer that captures human-perceived appeal, (2) an LLM-Rewriting component that enriches article titles with semantically related captions, and (3) a VLM-Judge evaluator that scores candidate images along editorial dimensions. We evaluate our approach both offline and in online user studies. Results show that LLM-Rewriting and VLM-Judge strategies can improve over candidate generation baselines, and that VLM-based evaluators approach the quality of human editorial

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

 b.khaertdinov@maastrichtuniversity.nl (B. Khaertdinov)

 © 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)

selection, despite the inherent subjectivity of the task.

2. Approach

Our system follows a two-stage retrieval pipeline. In the first stage, we retrieve candidate images from the source dataset using a VLM backbone. In the second stage, we refine the ranking of the top- k retrieved images using a set of re-ranking strategies that incorporate semantic, editorial, and aesthetic dimensions. Both stages rely exclusively on pre-trained models rather than supervision from editor-selected images. Prior work shows that editorial choices do not necessarily align with reader preferences [1], making direct optimization potentially misaligned with the *fit and relevance* as judged by readers. By using general-purpose pre-trained models and modular re-ranking, we aim to capture a broader notion of relevance and presentation quality, while acknowledging that models trained on web-scale data may still reflect representational, societal and cultural biases.

2.1. VLM-based Retrieval for Candidate Generation

Given a news article, we embed its title into a multimodal embedding space using the text encoder of a VLM. In our experiments, we evaluated two families of VLMs, such as CLIP [2] and SigLIP [3]. Candidate retrieval is performed by computing the cosine similarity between the article embedding and pre-computed image embeddings from the YFCC100M dataset [9]. This results in a ranked list of images, ordered by similarity score. To ensure efficiency, we retain only the top- k candidates for further processing.

2.2. Re-Ranking Framework

While similarity-based retrieval provides an efficient initial approximation, it does not account for many subtle aspects of news-image matching, such as editorial quality, bias, or visual appeal. To address this, we propose a modular re-ranking framework, where each module takes the top- k candidates and refines their order based on additional criteria. Each re-ranking module updates the similarity score as follows:

$$s_{\text{final}} = (1 - \lambda) \cdot s_{\text{sim}} + \lambda \cdot s_{\text{reranking}} \quad (1)$$

where s_{sim} is the similarity-based score from the retrieval stage, $s_{\text{reranking}}$ is the auxiliary score from a re-ranking strategy, and λ is a weighting parameter. In our study, we experimented with three re-ranking strategies described in the following paragraphs. As a baseline, we use the output of the first retrieval stage, i.e. similarity search results produced by a backbone vision-language model without re-ranking.

(1) Aesthetics-based re-Ranking. First, we propose to incorporate an aesthetic quality of images to re-rank the candidates. Specifically, we utilized an aesthetics prediction model¹ provided in [10]. That is a linear model built on top of embeddings generated by CLIP. For each candidate, the model outputs an aesthetic score ranging from 1 to 10, which is normalized to $[0, 1]$ and used as $s_{\text{reranking}}$ in Eq. 1.

(2) LLM-Rewriting. Second, we propose a re-ranking strategy that uses semantic enrichment through query rewriting with a large language model (LLM). Instead of relying solely on the original text article, we utilize the generative abilities of LLMs to produce captions of potential thumbnail images based on article titles. The model was prompted to generate 5

¹<https://github.com/LAION-AI/aesthetic-predictor>

Table 1

Offline evaluation. Hits@5 (%) for the one-stage baseline and proposed re-ranking strategies. Δ indicates the absolute change relative to the baseline (positive values denote improvement).

Model ID	Top- k	λ	Baseline	Aesthetics		LLM-Rewriting		VLM-Judge	
				Hits@5	Δ	Hits@5	Δ	Hits@5	Δ
CLIP-base	25	0.2	74	73.76	-0.24	75.06	+1.06	76.59	+2.59
CLIP-large	25	0.2	82.47	81.88	-0.59	82.82	+0.35	82.82	+0.35
SigLIP-2	25	0.2	78.59	78.00	-0.59	80.12	+1.53	79.76	+1.17

diverse captions describing suitable thumbnails based on the article’s title. Each caption is then embedded using the same VLM text encoder employed in the retrieval stage. For each candidate image, we compute cosine similarities with all generated caption embeddings and retain the maximum similarity value $s_{\text{reranking}} = \max_{c \in \mathcal{C}} [\text{sim}(f_{\text{text}}(c), f_{\text{img}}(i))]$, where \mathcal{C} denotes the set of paraphrased captions, $f_{\text{text}}(c)$ and $f_{\text{img}}(i)$ are the text and vision representations from VLM encoders, and i is the candidate image. This approach enriches the textual representation of the query and promotes relevant images that might not have been aligned with the title alone.

(3) VLM-Judge. Finally, we employ a generative vision-language model, namely Qwen2-7B-Instruct [11], as a *judge* that evaluates shortlisted candidate images with respect to the input article. This model exploits four editorial dimensions, highlighted by the challenge organizers in the *quest for insight*: 1) *image quality*: visual clarity and absence of technical flaws; 2) *trustworthiness*: whether the image appears to originate from a credible journalistic source. 3) *bias- and stereotype-free*: whether the image avoids reinforcing stereotypes or biased visual representations; 4) *clickbait-free*: whether the image avoids clickbait elements. The model is prompted to generate a numerical score $s_{\text{dim}} \in [1, 5]$ for each dimension. For each candidate image, scores are averaged across criteria to form $s_{\text{reranking}}$ that is then used in Eq. 1. This allows the system to combine the efficiency of embedding similarity with the nuanced judgments of a generative model that reflects media-relevant evaluation criteria.

2.3. Development and Inference Design

We used 10% of the MediaEval 2025 NewsImages dataset (850 queries and 850 matching images) as the validation set for offline experimentation [12]. For online evaluations, we applied the proposed two-stage image search system to test news titles and used a subset of the YFCC100M dataset containing approximately 1.5 million images. In detail, candidate images from YFCC100M are embedded with vision-language backbones and indexed using a FAISS vector database for efficient similarity search. The retrieval is performed by computing cosine similarity between article embeddings and indexed image embeddings, yielding the top- k candidates. Three re-ranking modules (Section 2) share a unified interface and fuse candidate similarity scores with the produced re-ranking scores (Eq. 1).

3. Results and Analysis

3.1. Offline Experiments

We evaluate three VLM backbones: CLIP-base (clip-vit-b/32), CLIP-large (clip-vit-l/14) [2], and SigLIP-2 (siglip2-so400m-patch14-384) [13]. We also tuned the number of candidates $k \in \{10, 25, 50\}$ and weight $\lambda \in \{0.2, 0.4, 0.6, 0.8\}$ for re-ranking on the validation set.

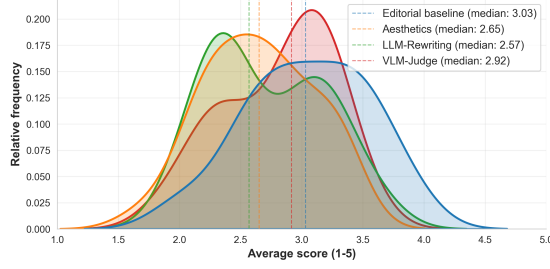


Figure 1: Online evaluation. Distribution of average user scores for the proposed re-ranking strategies and the editorial baseline.

Table 1 reports the highest Hits@5 values per backbone and re-ranking strategy on the validation set. As can be seen, the *VLM-Judge* strategy yields improvements with all backbone retrievers. The ranking strategy *LLM rewriting* also provides positive gains across models. In contrast, aesthetics-based re-ranking has a small negative effect (negative delta). Among the evaluated backbones, CLIP-large achieves the highest absolute performance (82.82% Hits@5) for both LLM-rewriting and VLM-Judge re-ranking mechanisms. In contrast, the smallest CLIP-base model benefits the most from VLM-Judge re-ranking in terms of relative improvements. We used CLIP-large as a backbone for the online evaluation stage.

3.2. Online Evaluation

The online evaluations were conducted on 30 test queries (titles) corresponding to distinct news articles, provided by the NewsImages MediaEval challenge organizers [12]. As described in Section 2.3, we used 1.5 million images from YFCC100M as a source of candidate images. Each retrieved image was rated by 31 users on a Likert scale between 1 and 5. Figure 1 shows the distribution of average user scores per query for the proposed re-ranking strategies, as well as the editorial baseline. The baseline scores were obtained using the images that originally appeared as thumbnails selected by news editors. The editorial baseline received the highest median rating (3.02), followed by the VLM-Judge variant, while the LLM-Rewriting and Aesthetics-based strategies achieved lower scores. A Wilcoxon signed-rank test comparing per-query average scores revealed no significant difference ($p = 0.052$) between the VLM-Judge variant and the editorial baseline. This result suggests that the automated re-ranking strategy closely approaches the quality of editors’ selection. Notably, the median score for the professional editors is only around 3 (out of 5), suggesting that human selections are rated by users as moderately relevant. This underscores the difficulty of news-image sourcing and points to opportunities for AI-assisted retrieval. In line with the growing interest in interactive retrieval [14, 15, 16], our re-ranking modules could be integrated into editorial workflows to support exploration, mitigate bias, and help editors select images better aligned with reader expectations.

4. Conclusion

This paper presents a two-stage framework for news thumbnail retrieval that combines pre-trained vision-language models with modular re-ranking. We show that LLM-rewriting and VLM-Judge improve over baseline retrieval, with the latter approaching the quality of human editorial choices. Ground-truth thumbnails received only moderate user ratings, underscoring the complexity in aligning with readers. These findings highlight the potential of AI-assisted tools to support editorial workflows and motivate future work on interactive retrieval methods.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT-5 for: grammar and spelling check, improving writing style.

References

- [1] 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.
- [2] 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.
- [3] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 11975–11986.
- [4] 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. URL: <http://ceur-ws.org/Vol-3658/paper7.pdf>.
- [5] M. Zhang, Y. Wei, Z. Xing, Y. Ma, Z. Wu, J. Li, Z. Zhang, Q. Dai, C. Luo, X. Geng, B. Guo, Aligning vision models with human aesthetics in retrieval: Benchmarks and algorithms, in: A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (Eds.), Advances in Neural Information Processing Systems, volume 37, Curran Associates, Inc., 2024, pp. 86399–86434. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/9d3faa41886997cfc2128b930077fa49-Paper-Conference.pdf.
- [6] H. Zhu, J.-H. Huang, S. Rudinac, E. Kanoulas, Enhancing interactive image retrieval with query rewriting using large language models and vision language models, in: Proceedings of the 2024 International Conference on Multimedia Retrieval, 2024, pp. 978–987.
- [7] Z. Liu, W. Sun, D. Teney, S. Gould, Candidate set re-ranking for composed image retrieval with dual multi-modal encoder, arXiv preprint arXiv:2305.16304 (2023).
- [8] G. Zhan, Y. Liu, K. Han, W. Xie, A. Zisserman, Elip: Enhanced visual-language foundation models for image retrieval, arXiv preprint arXiv:2502.15682 (2025).
- [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, Commun. ACM 59 (2016) 64–73. URL: <https://doi.org/10.1145/2812802>. doi:10.1145/2812802.
- [10] C. Schuhmann, R. Beaumont, LAION-Aesthetics | LAION, 2022. URL: <https://laion.ai/blog/laion-aesthetics/>.
- [11] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, J. Lin, Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution, arXiv preprint arXiv:2409.12191 (2024).
- [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: Working Notes Proceedings of the MediaEval 2025 Workshop, 2025.
- [13] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, et al., Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, arXiv preprint arXiv:2502.14786 (2025).
- [14] S. Lee, S. Yu, J. Park, J. Yi, S. Yoon, Interactive text-to-image retrieval with large language models: A plug-and-play approach, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 791–809.
- [15] B. Khaertdinov, M. Popa, N. Tintarev, Visualref: Interactive image search prototype with visual relevance feedback, in: Proceedings of the Nineteenth ACM Conference on Recommender Systems, 2025, pp. 1353–1356.
- [16] Z. Long, K. Liang, G. Aragon Camarasa, R. McCreddie, P. Henderson, Diffusion augmented retrieval: A training-free approach to interactive text-to-image retrieval, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2025, pp. 823–832.