Hybrid Retrieval and Generative Image Recommendation for News Articles: The FAST-MS(DS) Approach at MediaEval 2026

Aqsa Khan Jadoon[1,*], Dr. Muhammad Rafi[1,*]

[1]FAST National University of Computer and Emerging Sciences, Karachi, Pakistan

k248047@nu.edu.pk

Abstract

This paper describes the FAST-MS(DS) submission to the News Image Recommendation task at MediaEval 2026. Given a news article text, the goal is to recommend a visually fitting image from a large candidate pool. We present four distinct approaches ranging from lightweight text-based retrieval to a generative image synthesis pipeline. Our retrieval runs use MPNet cosine similarity, BGE-Large with CrossEncoder re-ranking, and a type-aware hybrid fusion of CLIP, BGE, and TF-IDF signals. Additionally, we explore AI-driven image generation using RealVisXL V4.0, a photorealistic Stable Diffusion XL model, prompted with article-aware context. Our generative approach achieves the highest final score of 3.209 on the crowdsourced Likert evaluation, substantially outperforming all retrieval baselines, which ranged from 2.112 to 2.205.

1. Introduction

The News Image Recommendation task at MediaEval 2026 challenges participating teams to build a pipeline that, given a news article, selects or generates a fitting image to accompany it [1]. Participants receive 8,500 training articles with associated images sourced from GDELT, and must produce recommendations for approximately 800 test articles drawn from MIND and historical newspaper archives. Submissions are rated by crowd workers on a five-point Likert scale, and the winning team is determined by the highest mean rating across 40 held-out evaluation articles.

The task requires navigating a fundamental tension: retrieved images are semantically grounded but may not exist for rare or novel news events, while generated images can be tailored precisely to article content but risk producing unrealistic or contextually inconsistent results. Our group, FAST-MS(DS), explores both directions by submitting three retrieval-based runs and one generative run, with the aim of understanding where each paradigm succeeds and where it falls short.

This paper is organized as follows. Section 2 discusses related work. Section 3 describes each of our four runs in detail. Section 4 presents and analyzes our results. Section 5 summarizes our findings and outlines future directions.

2. Related Work

Prior work on news image recommendation has largely relied on cross-modal retrieval, embedding article text and candidate images into a shared semantic space and selecting the nearest neighbor [2, 3]. Models such as CLIP [4] enable direct image–text similarity scoring and have become a strong baseline in this space. More recent work has augmented single-encoder retrieval with two-stage pipelines that use a fast bi-encoder for recall followed by a slow cross-encoder for re-ranking [5].

The generative direction is newer and less explored in the news domain. Text-to-image diffusion models such as Stable Diffusion [6] and its photorealistic extensions (e.g., RealVisXL) can produce high-quality, contextually conditioned images, but prior MediaEval editions have shown that prompt quality and style conditioning critically affect human ratings [7].

Our work builds on both lines, contributing a type-aware fusion retrieval strategy and a prompt-engineering scheme for generative news images.

[0]*MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online*

*Corresponding authors.

Aqsa Khan Jadoon (k248047@nu.edu.pk); Dr. Muhammad Rafi (muhammad.rafi@nu.edu.pk)

3. Approach

All runs share a common candidate pool that merges the 8,500 training images with the evaluation set images, giving a richer retrieval corpus. Article text is enriched with NLP-derived metadata (article type classification and named entity extraction) before any embedding step.

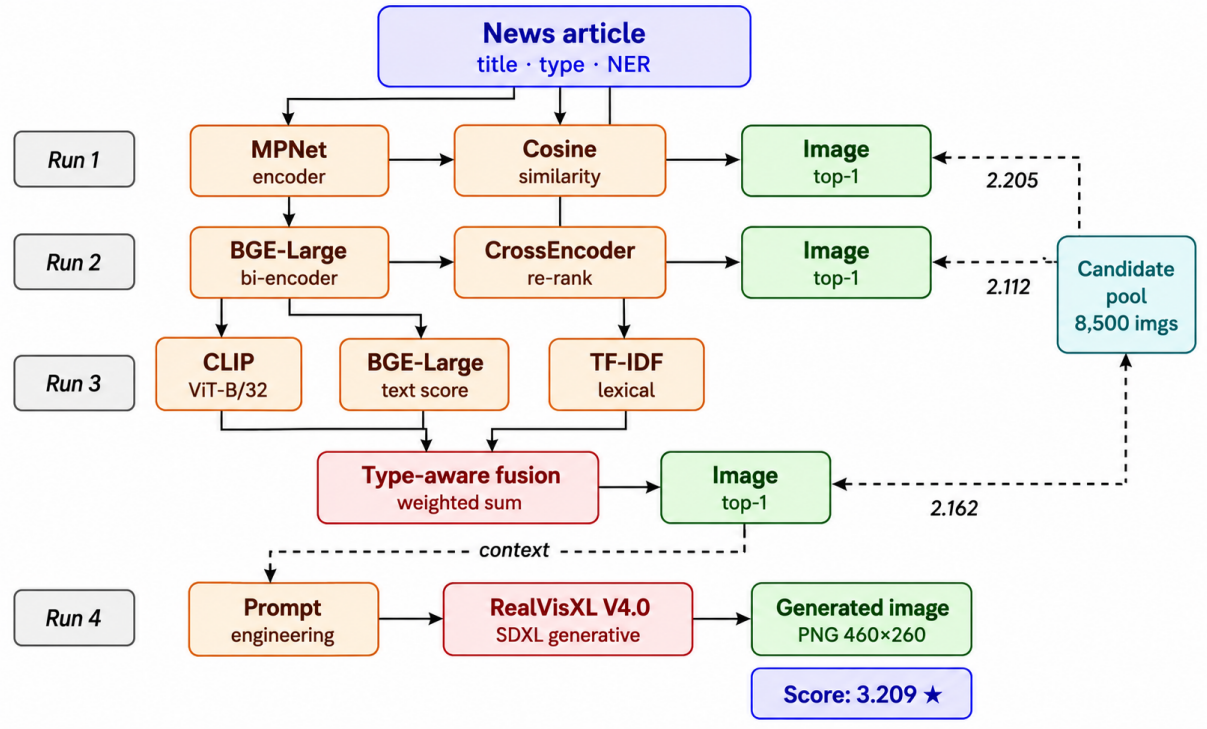

Figure 1: Overview of the four FAST-MS(DS) run pipelines. Dashed arrow from Run 3 to Run 4 indicates that the hybrid-matched candidate title is used as secondary context in the generative prompt.

3.1. Run 1 — Basic Retrieval (MPNet)

The simplest baseline encodes both article titles and candidate titles using `all-mpnet-base-v2` [5], a sentence-level MPNet model. Cosine similarity is computed between the test article embedding and every candidate embedding, and the top-scoring candidate's image is selected. This run serves as a lower-bound reference for semantic text matching.

3.2. Run 2 — Official Retrieval (BGE + CrossEncoder)

The second run implements a two-stage pipeline to improve precision over the MPNet baseline. First, `BAAI/bge-large-en-v1.5` retrieves the top-50 candidate articles by cosine similarity (recall stage). Then a `ms-marco-MiniLM-L6-v2` CrossEncoder scores every (test title, candidate title) pair within the recall set, and the highest-scoring candidate is selected (re-ranking stage). The CrossEncoder provides richer pair-level interaction than independent embeddings, at the cost of higher latency.

3.3. Run 3 — Hybrid Retrieval (CLIP + BGE + TF-IDF)

The hybrid run fuses three complementary signals under a type-aware weighting scheme. Article type (one of *sports*, *visual*, *historical*, *abstract*, *general*) is predicted by a lightweight classifier at pre-processing time. For each test article, three similarity vectors over the full candidate pool are computed: (i) an OpenCLIP ViT-B/32 image–text score, (ii) a BGE-Large text–text score, and (iii) a TF-IDF lexical overlap score. All three vectors are min-max normalized and combined as:

$$s_{\text{hybrid}} = w_c \cdot \hat{s}_{\text{CLIP}} + w_b \cdot \hat{s}_{\text{BGE}} + w_t \cdot \hat{s}_{\text{TF-IDF}} \tag{1}$$

The weights (w_c, w_b, w_t) are conditioned on article type. *Historical* articles are assigned $(0.05, 0.15, 0.80)$, weighting TF-IDF heavily because scanned newspaper text has no strong visual anchors. *Sports* and *visual* articles favor CLIP: $(0.50, 0.35, 0.15)$ and $(0.55, 0.30, 0.15)$ respectively. The top-ranked candidate with a valid image file is selected.

3.4. Run 4 — RealVisXL Image Generation

The generative run uses `SG161222/RealVisXL_V4.0`, a photorealistic fine-tune of Stable Diffusion XL [6], to synthesize an image conditioned on the article. A prompt is constructed from three sources: (1) the article type's visual style cue (e.g., "sports action photography, athletes, stadium"), (2) the article title verbatim, and (3) up to 80 characters of the hybrid-matched candidate title for additional contextual grounding. Named entities longer than four characters are also appended as key subjects. A unique random seed per article ensures all images within the run are distinct, as required by the task rules. Images are generated at 460×260 pixels in landscape orientation and saved as PNG.

4. Results and Analysis

Table 1 presents the final crowdsourced Likert scores for all four submitted runs.

Table 1: Final evaluation scores (5-point Likert scale, higher is better).

Run	Mean Score
Basic Retrieval (MPNet)	2.205
Hybrid Retrieval (CLIP + BGE + TF-IDF)	2.162
Official Retrieval (BGE + CrossEncoder)	2.112
RealVisXL Generated	**3.209**

4.1. Retrieval Runs

All three retrieval runs score in a narrow band between 2.11 and 2.21, suggesting that ceiling effects limit pure retrieval on this dataset. The Basic MPNet run marginally outperforms both more sophisticated approaches, which is counterintuitive. This likely reflects the nature of the test set: articles drawn from MIND and historical archives may not have close semantic matches in the GDELT-based candidate pool. In such cases, the CrossEncoder in Run 2 may select contextually precise but visually poor candidates, while the TF-IDF component of the hybrid run may over-match on surface form without visual relevance.

The narrow gap among retrieval runs further suggests that scaling up model capacity alone does not resolve the fundamental distributional mismatch between training-domain GDELT images and the out-of-domain test article vocabulary.

4.2. Generative Run

The RealVisXL run achieves a mean score of 3.209, more than a full Likert point above the best retrieval baseline. This outcome demonstrates that, for novel or out-of-distribution news events, a well-prompted generative model can produce images rated as considerably more fitting than anything retrievable from an existing pool.

The style-conditioned prompting strategy appears central to this result. By mapping article type to a visual style descriptor before composing the prompt, the model is steered toward photorealistic news photography conventions rather than generic photo aesthetics. The hybrid-matched title used as secondary context provides topical grounding without requiring the model to hallucinate subject matter not present in the article itself.

One limitation of the generative approach is consistency: different seeds and prompt lengths produce large variance in image quality for similar articles. Future work should explore systematic prompt optimization and negative prompt tuning to suppress common failure modes such as distorted faces and illegible text.

5. Conclusion

We have presented the FAST-MS(DS) submission to the MediaEval 2026 News Image Recommendation task, comprising three retrieval-based and one generative approach. Our main finding is that type-aware generative image synthesis substantially outperforms text-based retrieval from an existing image pool, achieving a mean Likert score of 3.209 versus a best retrieval score of 2.205.

Among the retrieval runs, simpler MPNet cosine similarity matched or exceeded more complex two-stage and hybrid pipelines, pointing to a ceiling imposed by candidate pool coverage rather than ranking quality. Future work will investigate enriching the candidate pool with web-retrieved images, improving prompt engineering for the generative model, and exploring fine-tuned CLIP variants adapted to the news photography domain.

Declaration on Generative AI. During the preparation of this work, the authors used Claude (Anthropic) for grammar and structure review, and RealVisXL V4.0 for image generation in Run 4. After using these tools, the authors reviewed and edited all content as needed and take full responsibility for the publication's content.

References

[1] Task Organizers. MediaEval 2026 News Image Recommendation Task Overview. *MediaEval Benchmark*, 2026.

[2] F. Liu, Y. Wang, T. Wang, and V. Ordonez. Visual News: Benchmark and Challenges in News Image Captioning. In *Proceedings of EMNLP*, 2020.

[3] A. Biten, L. Gomez, M. Rusinol, and D. Karatzas. Good News, Everyone! Context Driven Entity-Aware Captioning for News Images. In *Proceedings of CVPR*, 2019.

[4] A. Radford, J. W. Kim, C. Hallacy, et al. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of ICML*, 2021.

[5] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of EMNLP*, 2019.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of CVPR*, 2022.

[7] MediaEval 2025 Participants. Working Notes of the MediaEval 2025 Workshop. *CEUR Workshop Proceedings*, 2025.