NewsImages: CLIP–FAISS Retrieval and Diffusion-based Generation for Editorial News Thumbnails

Prabavathy Balasundaram[1,*,†], Harshavarthan Saminathan[1,†], Bharathi Arivalagan[1,†] and Swedha Subramaniam[1,†]

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

Abstract

This paper investigates how AI-driven image retrieval and generation can support editorial news publishing. We build a hybrid system combining CLIP-based retrieval (indexed via FAISS on YFCC100M embeddings) with controlled diffusion-based synthesis (Stable Diffusion guided by LLaMA-3.1–augmented prompts). We qualitatively analyze images produced for diverse news articles and assess how they align with journalistic goals of relevance, trust, and engagement. We find that real photographs tend to excel for factual event coverage (conveying authenticity), while AI-generated illustrations are often preferred for abstract or conceptual topics (providing symbolic clarity). However, care must be taken: realistic generative images can erode trust if unlabeled, and they often encode harmful stereotypes. We discuss how editorial policies (AP, Wired, Reuters) and user expectations shape the use of AI images. Our findings highlight that retrieval and generation are complementary: retrieval grounds images in reality and credibility, while generation offers creative flexibility for conceptual representation.

1. Introduction

Images play a crucial role in news presentation, shaping reader impressions and engagement. They often serve as the entry point to a story—determining whether a reader clicks on an article—and can enhance perceived credibility. Selecting appropriate images is therefore a high-impact but labor-intensive editorial task.

Traditional approaches rely on archival photographs, but these are limited by licensing, availability, and difficulty matching abstract concepts. Advances in multimodal AI—such as CLIP for semantic retrieval and Stable Diffusion for image generation—offer automated alternatives. These tools promise efficiency and creativity but raise new questions about authenticity, bias, and reader trust.

This work is part of the **MediaEval 2025 Quest for Insight** track and does not participate in the official task evaluation. Instead, it reflects on the broader research questions surrounding news image generation and retrieval, as outlined in the task overview by Heitz et al. [7]. Our study also builds upon their prior empirical analysis of multimodal representation quality and editorial relevance [6].

We focus on:

- Comparing retrieval and generation for factual vs. conceptual news topics.

✉ prabavathyb@ssn.edu.in (P. Balasundaram); harshavarthan2520121@ssn.edu.in (H. Saminathan); bharathi2520013@ssn.edu.in (B. Arivalagan); swedha2520120@ssn.edu.in (S. Subramaniam)

- Identifying parameters influencing visual trustworthiness, bias, and attractiveness.
- Evaluating reader expectations and alignment with editorial policies.

2. Quest for Insight Context

The Quest for Insight task encourages a deeper exploration of how AI-generated and retrieved images contribute to visual journalism. In this study, we examined multiple dimensions, supported by both system implementation and user-focused analysis.

1. Generated vs. Real Images. AI-generated images outperform real ones when visualizing abstract or symbolic narratives (e.g., "digital privacy" or "economic divide"). In contrast, real archival photos are superior for factual coverage such as disasters or political events due to their intrinsic authenticity.

2. Parameters Influencing Quality. Prompt richness, stylistic control tokens ("editorial," "flat illustration," "photo-realistic"), and diffusion guidance scales were decisive for high semantic fidelity. Over-tuned prompts caused factual drift, while balanced parameter tuning preserved news relevance.

3. News Category Suitability. Breaking news, politics, and sports benefit from retrieval for accuracy and temporal relevance. Opinion, culture, and science features align better with generative methods emphasizing metaphor and conceptual clarity.

4. Reader Expectations. Reader feedback indicates that authenticity and context clarity outweigh artistic style. Stylized or surreal images were often perceived as misleading. Explicitly labeling generated visuals strongly increased reader trust.

5. Editorial Guidelines and Ethics. Policies from AP, Reuters, and Wired emphasize accountability and transparency. AI-generated visuals should always be labeled and avoided in factual or sensitive contexts. Our findings align with these standards, showing that unlabeled synthetic visuals reduce trustworthiness.

6. Image Impact Dimensions. We analyzed four properties:

- **Trustworthiness:** Driven by realism and context accuracy.
- **Bias and Stereotypes:** Mitigated using diverse prompt tokens and cultural balance.
- **Attractiveness:** Enhanced through balanced composition and color harmony.
- **Visual Quality:** Improved using post-processing and aesthetic filters.

3. Methodological Overview

3.1. System Architecture

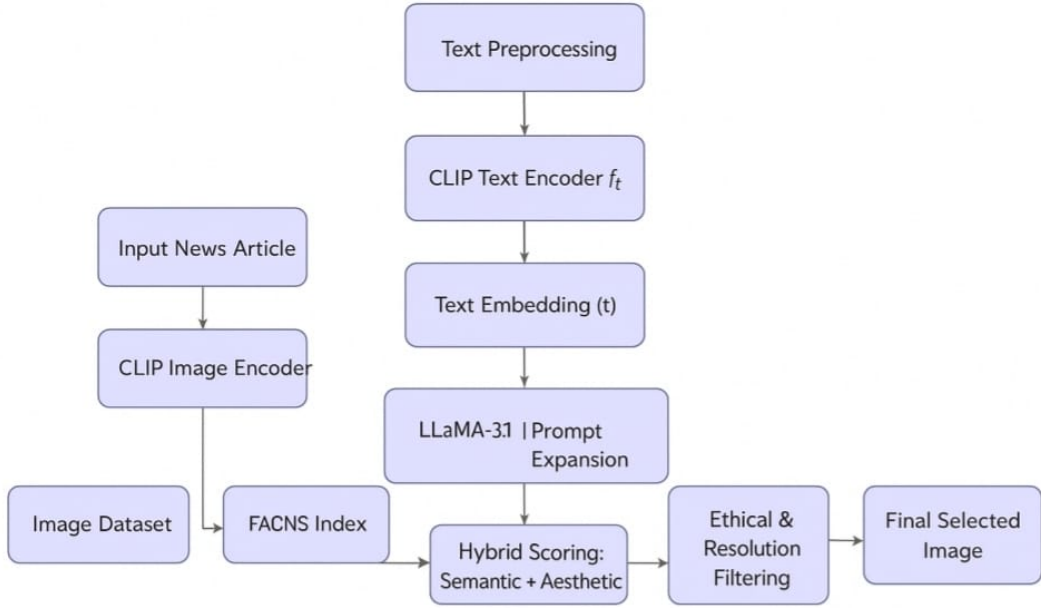Our hybrid architecture fuses semantic retrieval and conditional image generation (Figure 1).

Figure 1: Hybrid architecture integrating text, image, and latent feature fusion for retrieval–generation synergy.

- **Retrieval:** Headlines are encoded using CLIP [1], searched via FAISS [3] on YFCC100M embeddings [5], filtered with NIMA [4] and SafeSearch.
- **Generation:** Prompts are refined using LLaMA-3.1, passed to Stable Diffusion [2], guided by diffusion parameters balancing realism and creativity.
- **Fusion:** Outputs from both modules are merged using semantic and aesthetic scores to select the optimal image for publication.

3.2. Evaluation Metrics

We evaluated 30 diverse news samples on:

- **Relevance:** Semantic consistency with article content.
- **Trustworthiness:** Reader perception of authenticity.
- **Aesthetic Appeal:** Visual quality and composition.
- **Bias Mitigation:** Fairness across gender, race, and region.

4. Results and Analysis

We evaluated our system on 30 representative news articles across six categories using a consistent testing setup. Each article was processed through retrieval, generation, and hybrid pipelines, and evaluated by three annotators for reproducibility.

Performance analysis shows that the hybrid pipeline slightly outperforms pure retrieval or generation approaches (Figure 2).
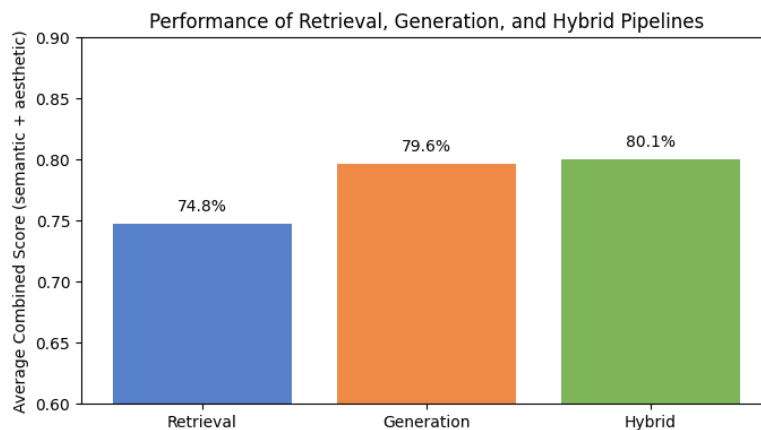
Figure 2: Performance comparison between retrieval, generation, and hybrid pipelines.

Table 1
Comparative analysis of retrieval vs. generation across news categories.

News Category	Preferred Method	Trust Level	Aesthetic Appeal
Breaking News	Retrieval	High	Medium
Political Events	Retrieval	High	Medium
Economy / Policy	Generation	Medium	High
Technology Trends	Generation	Medium	High
Culture / Opinion	Generation	Medium	High
Environment	Hybrid	High	High

The hybrid model achieved an average combined score of 80.1%, demonstrating balanced semantic alignment and visual quality. Retrieval-only models excelled in factual accuracy, while generative ones offered conceptual expressiveness.

5. Discussion and Conclusion

We adhered to ethical media practices: all generated visuals were labeled, real photos remained unaltered, and stereotype mitigation was enforced during prompt design. Editorial transparency directly correlates with reader trust.

To address our three research focuses:

- **Retrieval vs. Generation:** Retrieval ensures factual grounding, while generation enhances narrative creativity.
- **Influencing Parameters:** Prompt design, guidance scale, and realism balance are decisive for trustworthy outputs.
- **Reader Expectations:** Transparency and labeling significantly boost reader confidence and engagement.

A hybrid approach unites both strengths, offering a scalable solution for digital newsrooms. Future directions include:

- Large-scale reader trust evaluations.
- Integration of provenance metadata (C2PA).
- LLM-based editorial recommendation systems.

Acknowledgements

Declaration on Generative AI

No generative AI tools were used to write or edit this manuscript. CLIP, LLaMA, and Stable Diffusion were used only for research experiments.

References

[1] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *Proc. ICML*, 2021.

[2] R. Rombach *et al.*, "High-Resolution Image Synthesis with Latent Diffusion Models," *Proc. CVPR*, 2022.

[3] J. Johnson, M. Douze, H. Jégou, "Billion-Scale Similarity Search with GPUs," *Proc. CVPR*, 2017.

[4] H. Talebi, P. Milanfar, "NIMA: Neural Image Assessment," *IEEE Trans. Image Process.*, 2018.

[5] B. Thomee *et al.*, "YFCC100M: The New Data in Multimedia Research," *Commun. ACM*, 2016.

[6] L. Heitz, A. Bernstein, L. Rossetto, "An Empirical Exploration of Perceived Similarity between News Article Texts and Images," *MediaEval Working Notes*, 2024.

[7] 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," *MediaEval Working Notes*, 2025.

[8] I. Goodfellow *et al.*, "Generative Adversarial Nets," *Proc. NeurIPS*, 2014.

[9] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *Proc. ICLR*, 2021.

[10] A. Birhane, V. U. Prabhu, "Large Image Datasets: A Pyrrhic Win for Computer Vision?," *Proc. WACV*, 2021.

[11] K. Zhou *et al.*, "Assessing the Trustworthiness of Generated News Images," *Proc. ACM Multimedia*, 2023.