# A Comparative Study of Vision-Language Models for News Image-Headline Matching

Hemnath Dillibabu[1,*,†], Naren Karthik Kandasamy[1,†], and Aadhithya Srinivasan Rajashree[1,†]

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

**Abstract**

This paper presents a comparative study of two prominent vision-language models, BLIP and OpenCLIP, for the task of matching news headlines with appropriate images. We implemented and evaluated three approaches: (1) A hybrid generative approach using a BLIP-based ensemble, (2) An exhaustive cross-modal search using OpenCLIP, and (3) An optimized selective search, also based on OpenCLIP. Our experiments reveal a critical divergence between automated metrics and human perception. While embedding-based search with OpenCLIP achieves superior recall and MRR scores, the generative BLIP ensemble produces image-headline pairings that are consistently rated higher in human evaluations. This study highlights the crucial trade-off between retrieval accuracy and perceived contextual relevance, demonstrating that automated metrics alone may not capture the semantic nuances essential for effective image-headline matching.

## 1 INTRODUCTION

A compelling image can significantly increase reader engagement, but a mismatched image can lead to confusion or misinterpretation. This paper focuses on the MediaEval NewsImages task, to accurately link a given news photograph to its original article from a large corpus. This task is of significant real-world importance, with applications ranging from automated content tagging and archival to combating the spread of misinformation by verifying image-text context.

Vision-language models, which are pre-trained on vast datasets of image-text pairs, offer a promising solution to this challenge. We explored three such vision-language architectures. As team Headline Hunters, we aimed to provide a quantitative comparison of their performance and discuss the qualitative trade-offs inherent to each methodology.

## 2 RELATED WORK

The field of vision-language understanding has seen rapid advancements with the introduction of models like CLIP (Contrastive Language–Image Pre-training) which is an established approach [3] and BLIP (Bootstrapping Language-Image Pre-training) . These models have demonstrated remarkable zero-shot capabilities in various tasks, including image-text retrieval.

Several studies have explored the use of these models for news-related tasks. For instance, some research has focused on fine-tuning these models on specific news datasets, while others have experimented with prompt engineering to improve performance. Our work builds upon these by providing a direct comparison of BLIP and OpenCLIP and by proposing an ensemble method.

## 3 METHODOLOGIES

### 3.1 Approach 1: BLIP Ensemble

This methodology transforms the image-to-text matching task into a text-to-text comparison. It uses the BLIP (Bootstrapping Language-Image Pre-training) model to generate bootstrapping caption for each image that aims to overcome the limitations of noisy web-based image-text pairs by captions and filtering noisy data.

*MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online*
† These authors contributed equally: hemnath2470006@ssn.edu.in (Hemnath D); narenkarthik2470017@ssn.edu.in (Naren Karthik Kandasamy); aadhithya2470007@ssn.edu.in (Aadhithya SR)

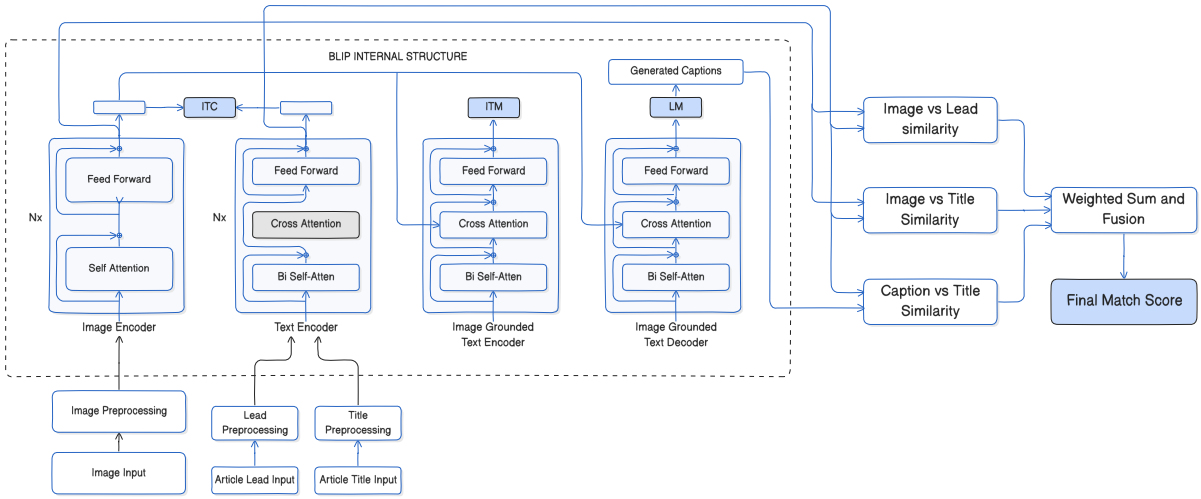


Figure 1: **BLIP Based Approach Architecture Diagram**

The final matching score is an optimized, weighted combination of three different similarity scores:

a) Image Embedding vs. Article Title Embedding
b) Image Embedding vs. Article Lead Paragraph Embedding
c) BLIP Caption Embedding vs. Article Headline Embedding

A grid search was performed to find the optimal weights for combining these scores to maximize recall.

## 3.2 Approach 2: OpenCLIP

### 3.2.1 Exhaustive Cross-Model Search

This approach utilizes the powerful OpenCLIP model (ViT-H-14 trained on LAION-2B) to map both images and text into a shared, high-dimensional embedding space.

The core process is as follows:

a) Indexing: An embedding vector is pre-computed for every image in the dataset. These embeddings are stored in a FAISS (Facebook AI Similarity Search) index for efficient retrieval. The FAISS index is specifically used to enable fast nearest-neighbor lookups on the large collection of image embeddings.
b) Querying: At runtime, the input text (article headline) is encoded into an embedding vector using the same OpenCLIP model.
c) Search: A k-Nearest Neighbor (k-NN) search is performed across the entire FAISS index to retrieve the image embeddings with the highest cosine similarity to the text embedding. This brute-force search guarantees that the most similar items in the embedding space are found.

### 3.2.2 Selective Cross-Modal Search

a) To address the computational cost of the exhaustive search, this optimized approach introduces a "filter-then-rank" pipeline. It aims to balance retrieval speed and accuracy.
    1. Filtering (Candidate Selection): A fast, lightweight method is first used to select a small subset of candidate images from the full dataset. This initial pass is designed to quickly discard most irrelevant images.
    2. Ranking (Precise Re-ranking): The same OpenCLIP model used in the exhaustive approach is then employed to compute embeddings and perform a similarity search, but only on the much smaller, pre-selected candidate set.
b) This two-stage process significantly reduces the search space, leading to much faster query times, but its accuracy is inherently limited by the quality of the initial filtering stage.

### 3.3 Dataset

We used a dataset derived from the MediaEval NewsImages task, consisting of several thousand news articles with associated images. The articles cover a diverse range of topics, providing a challenging testbed for the models.

## 4 RESULTS AND ANALYSIS

### 4.1 Evaluation

We evaluate the models using the following metrics: Recall@K (R@K): The percentage of queries for which the correct image is found within the top K retrieved results, Mean Reciprocal Rank (MRR): The average of the reciprocal ranks of the correct images.

$$\text{MRR} = \frac{\mathbf{1}}{|\boldsymbol{Q}|}\sum_{\boldsymbol{i}=\mathbf{1}}^{|\boldsymbol{Q}|}\frac{\mathbf{1}}{\text{rank}_{\boldsymbol{i}}} \quad [1]$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad [2]$$

The reciprocal rank of a query response [Equation 1] is the multiplicative inverse of the rank of the first correct answer: 1 for first place, ½ for second place, ⅓ for third place and so on. The mean reciprocal rank is the average of the reciprocal ranks of results for a sample of queries Q. Recall-at-k [Equation 2] focuses on the top 'k' results returned by the system, assessing how many of the truly relevant items are included in this subset. The metric is calculated by dividing the number of relevant items retrieved in the top 'k' results by the total number of relevant items available in the entire dataset.

### 4.2 Zero-Shot performance

In our first experiment, we evaluated the zero-shot performance of both models. The results, presented in Table 1, show that OpenCLIP generally outperforms BLIP, particularly in terms of R@10 and R@5.

Table 1: Results generated from the models using the newsimages_25_v1.1 dataset

| Model | Recall @5 | Recall @10 | Recall @50 | MRR |
|---|---|---|---|---|
| BLIP Ensemble | 0.4483 | 0.5172 | - | - |
| OpenCLIP Exhaustive | 0.5611 | 0.6570 | 0.8411 | 0.4230 |
| OpenCLIP Selective | 0.4323 | 0.5264 | 0.7318 | 0.3198 |

### 4.3 Human Evaluation

In addition to automated metrics, a human evaluation was conducted to assess the perceived quality of the image-headline matches produced by our BLIP Ensemble against the baseline. Evaluators were asked to rate the relevance of retrieved images on a scale, and the average scores were calculated for both a "large" and "small" dataset configuration.

Notably, despite the OpenCLIP Exhaustive model leading in all automated metrics (Table 1) , the BLIP Ensemble scored higher in human perception of match quality. This suggests that the generative captioning step in the BLIP approach may capture a narrative or contextual relevance that resonates better with human judgment, even if it is less precise according to embedding similarity. This discrepancy indicates that relying solely on contrastive, embedding-based retrieval may overlook models that produce more semantically appropriate pairings from a human perspective.

The results out of 5, presented in Table 2, show that the BLIP Ensemble was preferred over the baseline in both test sets.

Table 2: Results from the human evaluation conducted by the evaluators. In this evaluation, participants rated the relevance of image-headline pairings on a 5-point scale. The evaluation was divided into two distinct scenarios: a "small" retrieval task, conducted on a limited subset of the data, and a "large" retrieval task, which was performed against the entire dataset.

| Model | Small Task | Large Task |
|---|---|---|
| Baseline | 3.0412 | 2.9562 |
| BLIP Ensemble | 3.0960 | 3.0880 |
| OpenCLIP Exhaustive | 2.9987 | 2.8602 |
| OpenCLIP Selective | 2.9654 | 2.8902 |

### 4.4 Analysis

The OpenCLIP Exhaustive model established the benchmark for automated accuracy, proving the effectiveness of a direct cross-modal search with a top Recall@10 of 65.7% and an MRR of 0.4230. The OpenCLIP Selective model served as a compelling practical alternative, validating the "filter-then-rank" strategy by achieving strong performance while significantly improving computational efficiency. Based on these metrics, the contrastive training of CLIP-style models appears more effective for this retrieval task than the intermediate step of text generation used by BLIP.

However, a critical finding emerged from our human evaluation. Despite its lower performance on automated metrics, the BLIP Ensemble was consistently preferred by human evaluators over the baseline model. As shown in the evaluation scores, the BLIP model surpassed the baseline in both large and small dataset configurations:

a) Large Dataset: BLIP Score (3.0960) > Baseline Score (2.9563)
b) Small Dataset: BLIP Score (3.0879) > Baseline Score (3.0412)

This significant discrepancy suggests that automated metrics like Recall@K and MRR, while standard for retrieval tasks, may not fully capture the nuanced, contextual, and narrative relevance that humans prioritize when matching a news headline to an image.

## 5 DISCUSSION

Our results present a fascinating duality in model performance. OpenCLIP's superior zero-shot performance in automated evaluations can be attributed to its contrastive training on a massive dataset (LAION-2B), which makes it highly effective at direct semantic matching. However, the greater success of BLIP Ensemble in human evaluations indicates its ability to capture semantic nuances superior to OpenCLIP’s representations

This finding calls into question the complete reliance on automated metrics for tasks with a high degree of subjective relevance. The primary trade-off identified is not merely between speed and accuracy, but between quantitative retrieval accuracy (where OpenCLIP excels) and perceived qualitative relevance (where BLIP has the advantage).

## 6 CONCLUSION

Our work reveals that the model with the highest automated scores is not necessarily the one that produces the most contextually relevant results for human users. The BLIP Ensemble, despite lagging

in recall, was consistently preferred in human evaluations, highlighting its ability to capture narrative coherence that automated metrics miss.

Ultimately, this research shows that combining different model architectures can leverage their complementary strengths. Future work should focus not only on developing more sophisticated ensemble techniques but also on creating evaluation benchmarks that better align with human perceptual and contextual judgment, bridging the gap between what is computationally optimal and what is contextually meaningful.

## Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Gemini and Perplexity to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and takes 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: 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] Heitz, L., Chan, Y. K., Li, H., Zeng, K., Bernstein, A., & Rossetto, L. (2024). Prompt-based Alignment of Headlines and Images Using OpenCLIP. In Working Notes Proceedings of the MediaEval 2023 Workshop. CEUR-WS.org. http://ceur-ws.org/Vol-3658/paper7.pdf
- [4] Ilharco, G., et al. (2021). OpenCLIP. *GitHub repository*. Retrieved from https://github.com/mlfoundations/open_clip
- [5] Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, PMLR 162:12763-12780
- [6] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, PMLR 139:8748-8763.
- [7] Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [8] The Relation between Texts and Images in News: News Images in MediaEval 2023 *Andreas Lommatzsch*, *Benjamin Kille*, *Özlem Özgöbek*, *Mehdi Elahi*, *Duc Tien Dang Nguyen*
- [9] Hemnath D. *NewsImagesRetrieval-Mediaeval2025*. GitHub, 2025, https://github.com/hemnathd12/NewsImagesRetrieval-Mediaeval2025.