# Multimodal Fusion in NewsImages 2023: Evaluating Translators, Keyphrase Extraction, and CLIP Pre-Training

Tien-Huy Nguyen[1,4,*], Hoang-Long Nguyen-Huu[1,4,*], Thien-Doanh Le[2,4,*], Huu-Loc Tran[1,4,*], Quoc-Khanh Le-Tran[1,4,*], Hoang-Bach Ngo[3,4], Minh-Hung An[5] and Quang-Vinh Dinh[6,†]

[1]*University of Information Technology, Ho Chi Minh City, Vietnam*
[2]*International University, Ho Chi Minh City, Vietnam*
[3]*University of Science, VNU-HCM*
[4]*Vietnam National University, Ho Chi Minh City, Vietnam*
[5]*FPT Telecom, Ho Chi Minh City, Vietnam*
[6]*Vietnamese German University, Binh Duong, Vietnam*

Abstract

Matching the most appropriate image to its corresponding article poses a significant challenge in this landscape. This paper explores the intricate challenge of matching headline images to news articles, utilizing the zero-shot capability of CLIP to address the complex relationship between texts and both real and AI-generated images in the MediaEval 2023 News-Images Challenge. Additionally, analyzes the ramifications of diverse translation methodologies on the efficacy of CLIP performance. The innovative approach involving key phrase extraction for CLIP input demonstrates competitive results across various benchmarks in information extraction and matching.



## 1. Introduction

In today's Internet age, online news articles play a crucial role as fundamental sources of information on current events, employing compelling titles and content segments to engage and inform readers effectively. Journalists strategically integrate images to enhance content intuitiveness, enabling a comprehensive understanding of the presented information and captivating the reader's attention. The MediaEval Multimedia Evaluation benchmark, with a focus on the NewsImages task [1], explores the intricate relationship between textual narratives and visual elements in news articles, contributing significantly to understanding collaborative dynamics in news discourse. Recent advancements, exemplified by Contrastive Language-Image Pretraining (CLIP) [2], provide a robust foundation for research combining text and real images, comprehensively exploring their relationship in news articles. Taking advantage of CLIP's zero-shot capabilities to evaluate experiments on real images and AI-generated images.

*MediaEval'23: Multimedia Evaluation Workshop, February 1–2, 2024, Amsterdam, The Netherlands and Online*
*All authors contributed equally to this paper.
†Corresponding author.
22520567@gm.uit.edu.vn (T. Nguyen); 22520817@gm.uit.edu.vn (H. Nguyen-Huu); Itcsiu22237@student.hcmiu.edu.vn (T. Le); 22520796@gm.uit.edu.vn (H. Tran); 22520638@gm.uit.edu.vn (Q. Le-Tran); nhbach22@apcs.fitus.edu.vn (H. Ngo); hungam@fpt.com (M. An); vinh.dq2@vgu.edu.vn (Q. Dinh)
0009-0000-0196-6083 (T. Nguyen); 0009-0003-4473-0769 (H. Nguyen-Huu); 0009-0002-9261-5223 (T. Le); 0009-0006-0954-2713 (H. Tran); 0009-0000-3990-4232 (Q. Le-Tran); 0000-0002-5224-6323 (H. Ngo); 0009-0001-0394-4731 (M. An); 0000-0002-8025-2501 (Q. Dinh)

## 2. Related works

Understanding the interaction between text and images in news is crucial for grasping news content creation. Recent studies challenge the notion of a simple text-image connection, highlighting the limitations of traditional image captioning models. New dynamic attention-based models, like those by Messina et al. [3] and Qizhang et al. [4], offer adaptability but increase computational complexity. Nelleke Oostdijk's analysis in [5] highlighted the limitations of a simplistic correlation between modalities, demonstrating that images possess the capability to depict entities within text or unrelated visual elements. Research like Lidia Pivovarova's [6] in the MediaEval 2021 NewsImages task, which integrated knowledge distillation and a visual topic model, shows that images can represent entities from text or unrelated visuals, and alignment between text and visual topics is possible. HCMUS [7] achieved competitive results through advanced text preprocessing and the utilization of the CLIP pre-trained model; however, this approach also relied on the translator. Our method further investigates the effect of different translators on performance, using CLIP and key phrase extraction to predict relevant text for images, thus deepening the understanding of the complex text-image relationship.

## 3. Proposed Methods

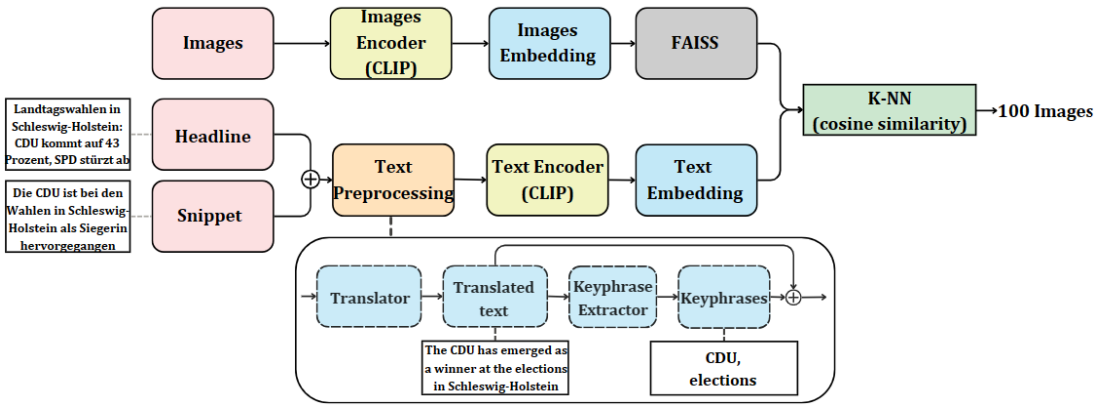


**Figure 1:** Our text-based image retrieval system architecture.

The fundamental concept of this architecture is to integrate both text and image as inputs by embedding them into a shared space. For the image input, each image undergoes vector embedding by the CLIP image encoder. These embeddings are then indexed using the Faiss library (as you can see in Figure 1). With the text input, we process the headline and snippet of the news (including translating text and optionally extracting keyphrases) before encoding into an embedding using the CLIP text encoder. In the end, we identify the 100 most relevant images by leveraging the K-NN (with cosine similarity) algorithm.

### 3.1. Translator

The dataset consists of three components, each derived from news content sourced from news portals, including GDELT1 and GDLT2, and a news feed RT dataset. The articles from RT News are written in German and paired with their corresponding English texts in the dataset. In [7], using Google Translate as a translator tool, in addition, we use another translator tool, mBART (multilingual Bidirectional and Auto-Regressive Transformers) [8], to experiment and

evaluate the impact of different translation methods on overall performance. Through this experimentation, we could have a better insight into each translator tool's advantages and disadvantages, allowing us to explore the relationship between the features of images and news.

### 3.2. Keyphrase

The section aims to analyze and extract relevant keyphrases from given inputs. The CLIP without keyphrase approach shows suboptimal accuracy, attributed to lengthy and noisy headlines and snippet sentences. The observation that images closely match the content in the headline and snippet, so to enrich key information for the image query, suggests the need for the keyphrase approach to extract more crucial entities. To address this problem, our approach uses KBIR (Keyphrase Boundary Infilling with Replacement) [9] pre-trained model, designed for NLP tasks, for effective keyphrase extraction and generation from text, crucial for the CLIP model.

### 3.3. Using CLIP as a Zero-shot retriever

Automatic image captioning and text-image matching have advanced significantly, typically requiring labelled data and specialized training. CLIP, a pre-trained neural network model, takes a unique approach by learning joint image and text representations without task-specific optimization. Its ability to transfer knowledge to other tasks without prior training, along with a large and varied pre-training dataset, especially with news articles collected on the internet, makes it a suitable and attractive option for tasks such as NewsImage. In addition, this allows the model to achieve state-of-the-art performance on tasks it hasn't been explicitly trained on before, offering a promising baseline for further research [10, 11, 12, 13]. This research investigates the zero-shot performance of CLIP, and the advanced ViT-L/14@336p model, the most potent CLIP variant, is employed for optimal results.

## 4. Experimental Results

The competition task requires participants to predict a sequentially organized list of images that closely aligns with the accompanying textual article. Evaluation employs the Mean Reciprocal Rank (MRR) metric and MeanRecall@K scores (K in 5, 10, 50, 100). Our research undergoes assessment on three datasets provided by the competition organizers, leading to distinct experimental methodologies and variations in textual input for CLIP due to inherent dissimilarities in each dataset. Consequently, our innovative approaches differ for each dataset under consideration.

**Table 1**
**Experimental results to assess how different translation methods impact the overall performance on the RT test set.**

| Translation Method \ Metric | MRR | MeanRecall@5 | MeanRecall@10 | MeanRecall@50 | MeanRecall@100 |
|---|---|---|---|---|---|
| Baseline | 0,22247 | **0,30767** | **0,38233** | 0,54367 | 0,61967 |
| mBART | 0,22451 | 0,30200 | 0,38133 | 0,55667 | 0,63533 |
| Google translate | **0,22527** | 0,30567 | 0,37633 | **0,56033** | **0,64433** |

The experimental findings on various lexical translation methodologies show relatively consistent results compared to utilizing the organizers' translated text. Comparisons between translation models indicate that using Googletrans is better mBART for the RT dataset, this led to the decision to leverage Googletrans for ongoing enhancements. However, based on

experimental reports, among many translation methodologies, we conclude that translator modules do not greatly affect CLIP's performance, so future methods should consider the option of removing the translator module to reduce pipeline complexity.

**Table 2**
**Experimental results of using the keyphrase approach on the RT test set.**

| Approach \ Metric | MRR | MeanRecall@5 | MeanRecall@10 | MeanRecall@50 | MeanRecall@100 |
|---|---|---|---|---|---|
| Non-keyphrases | **0,22527** | 0,30567 | 0,37633 | **0,56033** | **0,64433** |
| Keyphrases | 0,22304 | **0,30700** | **0,38367** | 0,54600 | 0,62033 |

In the experiment, incorporating keyphrases into the textual content that previously consisted of the headline and snippet as input for CLIP yields overall performance improvements, particularly in the MeanRecall@5 and MeanRecall@10 metrics (increases of 0.00133 and 0.00734, respectively), as shown in Table 2. The keyphrase's ability to encapsulate main ideas helps the model focus more on crucial information and clarify images query, resulting in commendable outcomes, particularly in retrieving images within the 5th and 10th ranks.

**Table 3**
**Experimental results of CLIP's zero-shot capability with both real and synthetic images on the GDELT1 and GDELT2 test set.**

| Datasets \ Metric | MRR | MeanRecall@5 | MeanRecall@10 | MeanRecall@50 | MeanRecall@100 |
|---|---|---|---|---|---|
| GDELT1 | 0,62243 | 0,77600 | 0,85200 | 0,94267 | 0,96533 |
| GDELT2 | 0,50842 | 0,62867 | 0,71800 | 0,86733 | 0,91533 |

Table 3 displays our evaluation outcomes on GDELT1 and GDELT2 test sets, which mainly include AI-generated images by Stable Diffusion [14], diverging from the RT dataset by using only headlines as input for CLIP. The experiment aims to leverage the pre-trained CLIP model's zero-shot capability, demonstrating a remarkably robust correlation between news text content and images, more specifically, CLIP produces well results on synthetic images, compared to experiments on real-image-exclusive RT dataset, as the nature of Stable Diffusion using the text encoder CLIP, the results are significantly better, showcasing state-of-the-art performance, being capable of querying images in almost any situation, and effectively addressing the challenge of establishing meaningful associations between articles and images, which contributes to helping journalists in recommendation systems for locating article-relevant images.

## 5. Conclusion and future work

This study tackles the demanding text-image matching task in the MediaEval 2023 NewsImages challenge, achieving notable success using the pre-trained CLIP model's zero-shot capability. Our experiments underscore the efficacy of the model architecture and the benefits of employing a pre-trained model. We studied to experiment with the CLIP's ability on both real and synthetic images, yielding promising outcomes for real images and proficient performance on AI-generated images. In addition, we proved that adding a translator did not improve performance, so we consider not using it in the pipeline in the future. Conversely, using key phrases showed positive signs of slightly increasing the accuracy of top-5 and top-10 image queries.

Future efforts will concentrate on implementing a more extensive approach, exploring additional techniques, such as re-ranking strategies [15] or face recognition systems [16] enhance to enrich crucial information for the image query and to further improve overall performance.

## References


[1] A. Lommatzsch, B. Kille, Ö. Özgöbek, M. Elahi, D.-T. Dang-Nguyen, News images in mediaeval 2023 (2023).

[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] N. Messina, F. Falchi, A. Esuli, G. Amato, Transformer reasoning network for image-text matching and retrieval, CoRR abs/2004.09144 (2020). URL: https://arxiv.org/abs/2004.09144. arXiv:2004.09144.

[4] Q. Zhang, Z. Lei, Z. Zhang, S. Z. Li, Context-aware attention network for image-text retrieval, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3536–3545.

[5] N. Oostdijk, H. v. Halteren, E. Basar, M. A. Larson, The connection between the text and images of news articles: New insights for multimedia analysis (2020).

[6] L. Pivovarova, E. Zosa, Visual topic modelling for newsimage task at mediaeval 2021, in: Working Notes Proceedings of the MediaEval 2021 Workshop, MediaEval, 2021.

[7] T. Cao, N. Ngô, T.-D. Le, T. Huynh, N.-T. Nguyen, H. Nguyen, M. Tran, Hcmus at mediaeval 2021: Fine-tuning clip for automatic news-images re-matching 3181 (2021).

[8] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, CoRR abs/2001.08210 (2020). URL: https://arxiv.org/abs/2001.08210. arXiv:2001.08210.

[9] M. Kulkarni, D. Mahata, R. Arora, R. Bhowmik, Learning rich representation of keyphrases from text, CoRR abs/2112.08547 (2021). URL: https://arxiv.org/abs/2112.08547. arXiv:2112.08547.

[10] H.-N. Vu, H.-D. Nguyen, M.-T. Tran, Re-matching images and news using clip pretrained model (2022).

[11] Y. Zhang, Y. Shao, X. Zhang, W. Wan, J. Li, J. Sun, Clip pre-trained models for cross-modal retrieval in newsimages 2022 (2022).

[12] D. Galanopoulos, V. Mezaris, Cross-modal networks and dual softmax operation for mediaeval newsimages 2022 (2022).

[13] M.-D. Le-Quynh, A.-T. Nguyen, A.-T. Quang-Hoang, V.-H. Dinh, T.-H. Nguyen, H.-B. Ngo, M.-H. An, Enhancing video retrieval with robust clip-based multimodal system, in: Proceedings of the 12th International Symposium on Information and Communication Technology, SOICT '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 972–979. URL: https://doi.org/10.1145/3628797.3629011. doi:10.1145/3628797.3629011.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, 2022. arXiv:2112.10752.

[15] W. Liu, Y. Xi, J. Qin, F. Sun, B. Chen, W. Zhang, R. Zhang, R. Tang, Neural re-ranking in multi-stage recommender systems: A review, 2022. arXiv:2202.06602.

[16] J. Dalvi, S. Bafna, D. Bagaria, S. Virnodkar, A survey on face recognition systems, 2022. arXiv:2201.02991.