# TREC2023 AToMiC Overview

Jheng-Hong Yang
University of Waterloo,
Canada

Carlos Lassance
Naver Labs Europe, France

Rafael Sampaio de Rezende
Naver Labs Europe, France

Krishna Srinivasan
Google Research,
United States

Miriam Redi
Wikimedia Foundation,
United Kingdom

Stéphane Clinchant
Naver Labs Europe, France

Jimmy Lin
University of Waterloo,
Canada

## ABSTRACT

This paper presents an exploration of evaluating image–text retrieval tasks designed for multimedia content creation, with a particular focus on the dynamic interplay among various modalities, including text and images. The study highlights the pivotal role of visual-textual multimodality, where elements such as photos, graphics, and diagrams are not merely ornamental but significantly augment, complement, or even reshape the meaning conveyed by textual content. This integration of multiple modalities is central to crafting immersive and captivating multimedia experiences. In the context of detailing the TREC initiative's evaluation process for the year, the paper introduces the AToMiC test collection, which serves as the foundational framework for evaluation. The authors delve into the distinctive task design, elucidating the specific challenges and objectives that characterize this year's evaluation. The paper further outlines the evaluation protocols, encompassing methodologies such as pooling dependencies and the criteria employed for relevance judgments. This overview offers valuable insights into the intricate process of evaluating multimedia retrieval systems, underscoring the evolving complexity and interdisciplinary nature of this field.

## KEYWORDS

Authoring tools; Multimedia content creation; Image–Text Retrieval

## 1 INTRODUCTION

The creation of multimedia content involves understanding the connections between elements encoded in different modalities, including textual, visual, audio, and more. For instance, in the realm of visual-textual multimodality, visual elements such as photos, graphics, and diagrams are employed to enhance textual information, either by embellishing, complementing, or altering the content's meaning [4, 6]. Conversely, textual support in the form of articles, paragraphs, and captions can be used to enhance the presentation of an image. However, despite the myriad approaches to multimodal multimedia content creation, much of the recent research in multimedia information retrieval relies on datasets that predominantly focus on simple text-image relationships through proxy tasks.

Certainly, a commonly utilized surrogate task within the domain of multimedia content creation is the retrieval of images based on captions. However, the existing datasets crafted for image retrieval from text and text retrieval from images present several challenges. These issues encompass a misalignment with authentic user needs, concerns related to suboptimal labeling, and limitations in terms of corpus size. These drawbacks hinder the accurate assessment of the effectiveness of multimodal retrieval systems.

| Split | Training | Validation | Test | Other |
|---|---|---|---|---|
| # texts (T) | 3,002,458 | 17,173 | 9,873 | 7,105,240 |
| # images (M) | 3,386,183 | 16,131 | 8,605 | 7,608,283 |
| # qrels | 4,401,903 | 17,801 | 9,873 | - |
| # M/T | 1.47 (± 2.72) | 1.03 (± 0.43) | 1.00 (± 0.00) | - |
| # T/M | 1.30 (± 0.82) | 1.10 (± 0.43) | 1.15 (± 0.53) | - |

**Table 1: AToMiC dataset statistics. The number of texts/images is defined by the relevant labels. # M/T (T/M) stands for the number of relevant images per text (texts per image).**

To tackle these challenges, we launched the TREC 2023 AToMiC initiative, aimed at addressing these limitations and establishing a more resilient evaluation framework for contemporary systems. Although we had 11 teams registered, only three of them actively participated in the evaluation process. Despite the limited number of participating teams, we successfully generated new labels and additional resources, significantly improving the quality of evaluation in this domain.

In the subsequent sections, we will provide an overview of the evaluation procedure in Section 2, present the task results in Section 3, conduct an analysis of the generated resources and labels in Section 4, and conclude our discussion in Section 5.

## 2 EVALUATION OVERVIEW

This section provides a comprehensive overview of the evaluation process for the TREC initiative this year. We will begin by introducing our test collections, known as AToMiC, which form the foundation for our evaluation. Following this, we will delve into the intricacies of our task design, offering insights into the specific challenges and objectives that underpin this year's TREC evaluation. Additionally, we will outline our evaluation protocols, including key elements such as pooling depth and the criteria used to define relevance judgements. To provide context and benchmarks for our evaluation, we will also introduce the baseline systems that serve as reference points for system performance. Furthermore, we will showcase the runs we have received from participants, highlighting the approaches and strategies employed to tackle the tasks.

### 2.1 Dataset – AToMiC

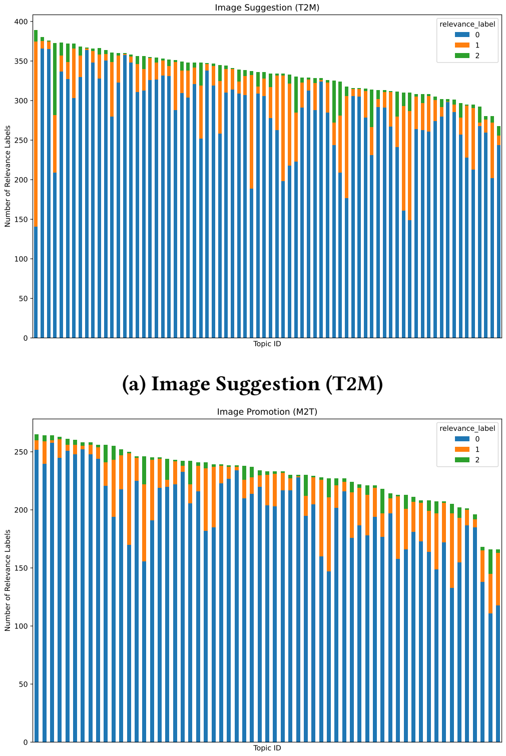
We rely on the AToMiC dataset [9] as the foundational resource for constructing our test collection, which forms the basis for the TREC evaluation. The AToMiC dataset is an extension of the Wikipedia-based Image Text (WIT) dataset [8] and encompasses two retrieval tasks designed for multimedia content creation: image suggestion and promotion (see subsection 2.2). Table 1 provides a comprehensive overview of the dataset's essential statistics. This extensive corpus, comprising approximately 10 million documents, encompasses both text and image collections across various partitions, including train, validation, test, and others. To facilitate system development and evaluation, we not only supply sparse labels (qrels) but also offer a set of development topics, complete with their dense labels [1]. For a more detailed understanding of the AToMiC dataset, we encourage interested readers to explore the previous work [9].

## 2.2 Task Design

In alignment with the AToMiC dataset's design principles, we have chosen evaluation topics that cater to the requirements of two distinct user models. Additionally, our selection of test topics takes into account the needs of both editors, who seek to enhance articles lacking images, and maintainers, who are responsible for monitoring the overall quality of all Wikipedia articles. Consequently, our emphasis lies on the selection of vital articles within Wikipedia to serve as evaluation topics for the tasks designed for these two user models: image suggestion (T2M) and image promotion (M2T).

*Image Suggestion (T2M).* The Image Suggestion (T2M) task revolves around the search scenario of identifying pertinent images to enhance textual content. To create this task, we meticulously selected 500 section topic candidates from articles listed in Wikipedia's vital articles at level 3.[2] Wikipedia's vital articles list consists of a meticulously curated collection of articles that are deemed pivotal in offering a comprehensive overview of human knowledge. These articles span a wide spectrum of topics and are often regarded as the foundation of Wikipedia's content, serving as indispensable reference points for readers seeking authoritative information. Our rationale for concentrating on these specific sections lies in their pivotal role within the Wikipedia ecosystem. By initially assessing them in the English language, we aim to pinpoint opportunities for enhancing the representation of vital content in other languages. This approach aligns seamlessly with our overarching objective of enhancing Wikipedia's accessibility and comprehensiveness for diverse linguistic communities.

*Image Promotion (M2T).* The Image Promotion (M2T) task is centered around a search scenario in which image providers aim to identify the most suitable attachment points: the text section within a article. Selecting the right images is a complex task, as predicting the images of greatest interest from the perspective of image providers can be challenging. To streamline the image selection process, we employ a multi-stage filtering approach to choose images from the image suggestion task. Initially, we utilize three fusion methods, namely, top-K, RRF, and RBP, to combine the image ranking lists generated by our baseline systems for 200 T2M topics, with the pooling depth set at 20. Subsequently, we merge the resulting image pools and eliminate duplicate images based on their IDs. Finally, we remove near duplicates among the images using the `fastdup` library and randomly chose 200 images as image topic candidates.[3]



**(a) Image Suggestion (T2M)**

**(b) Image Promotion (M2T)**

**Figure 1: Amount of labels generated for each topic, separated by relevance level and ordered by total amount of annotations.**

*Metrics.* In assessing the performance of our retrieval system, we anticipate dealing with ranked lists that prioritize the top positions as the most critical. Therefore, our primary metric of choice is the normalized Discounted Cumulative Gain (nDCG). This selection is particularly apt because we have access to graded annotation levels, allowing us to gauge the quality of our results with fine granularity. In addition to nDCG, we recognize the importance of understanding the interplay between other widely used metrics prevalent in different research communities. Metrics such as Mean Average Precision (mAP), Success, and Recall play vital roles in assessing retrieval effectiveness in various contexts. Investigating these metrics in conjunction with nDCG will provide a more comprehensive view of system performance and allow us to draw meaningful comparisons across different evaluation scenarios. By exploring these relationships, we aim to gain deeper insights into the strengths and limitations of our retrieval system, ultimately contributing to a more robust and nuanced evaluation framework.

## 2.3 Annotation Protocols

Our annotation process involves presenting annotators with candidates from participant runs, each with a specified pooling depth. Subsequently, after removing certain queries that do not meet evaluation criteria, the final assessment is conducted over 80 queries

[1] https://huggingface.co/datasets/TREC-AToMiC/AToMiC-Baselines/tree/main/dev_set

[2] https://en.wikipedia.org/wiki/Wikipedia:Vital_articles/Level/3

[3] https://github.com/visual-layer/fastdup

**Table 2: Image Suggestion (T2M) Results, ordered by nDCG@10.**

| Run ID | Team | Retrieval | Multimedia | mAP | nDCG@1K | nDCG@10 | Recall@1K | Success@1 | Success@10 |
|---|---|---|---|---|---|---|---|---|---|
| UvA-IRLab | IRLab-Amsterdam | Learned-Sparse | Image | 0.1526 | 0.4460 | 0.4060 | 0.6452 | 0.2973 | 0.6081 |
| b_splade_pp | baselines | Learned-Sparse | Caption | 0.1501 | 0.4461 | 0.4051 | 0.6452 | 0.2838 | 0.6081 |
| b_fsum_all | baselines | Hybrid | Image+Caption | 0.1183 | 0.5390 | 0.3109 | 0.8920 | 0.2297 | 0.5270 |
| b_bm25 | baselines | Sparse | Caption | 0.0761 | 0.3257 | 0.3036 | 0.4820 | 0.1351 | 0.5541 |
| UvA-IRLab-mlp-mlm-caption | UAmsterdam | Learned-Sparse | Caption | 0.0757 | 0.2741 | 0.2317 | 0.4273 | 0.1486 | 0.4865 |
| UvA-IRLab-mlp-mlm-img_cap | UAmsterdam | Learned-Sparse | Caption | 0.0760 | 0.2751 | 0.2315 | 0.4286 | 0.1486 | 0.4865 |
| finetune_large_t2i | uogTr | Dense | Image | 0.0857 | 0.2949 | 0.2206 | 0.4475 | 0.1351 | 0.3514 |
| b_clip_vith14_laion | baselines | Dense | Image | 0.0674 | 0.3011 | 0.2139 | 0.4699 | 0.1486 | 0.3784 |
| b_clip_vitg14_laion | baselines | Dense | Image | 0.0626 | 0.3039 | 0.2075 | 0.4596 | 0.1081 | 0.3514 |
| finetune_base | uogTr | Dense | Image | 0.0427 | 0.2365 | 0.1841 | 0.3352 | 0.0676 | 0.3243 |
| b_clip_vitl14_laion | baselines | Dense | Image | 0.0538 | 0.2790 | 0.1817 | 0.4700 | 0.1622 | 0.3378 |
| UvA-IRLab-mlp-mlm-cap1 | UAmsterdam | Learned-Sparse | Caption | 0.0234 | 0.1441 | 0.1426 | 0.2012 | 0.0811 | 0.2703 |
| b_clip_vitb32_laion | baselines | Dense | Image | 0.0248 | 0.1991 | 0.1396 | 0.2884 | 0.0135 | 0.2432 |
| b_flava | baselines | Dense | Image | 0.0031 | 0.0572 | 0.0752 | 0.0294 | 0.0000 | 0.0676 |
| UvA-IRLab-mlp-mlm-images | UAmsterdam | Learned-Sparse | Image | 0.0005 | 0.0179 | 0.0175 | 0.0286 | 0.0000 | 0.0405 |
| pretrain_base | uogTr | Dense | Image | 0.0000 | 0.0031 | 0.0050 | 0.0028 | 0.0000 | 0.0000 |

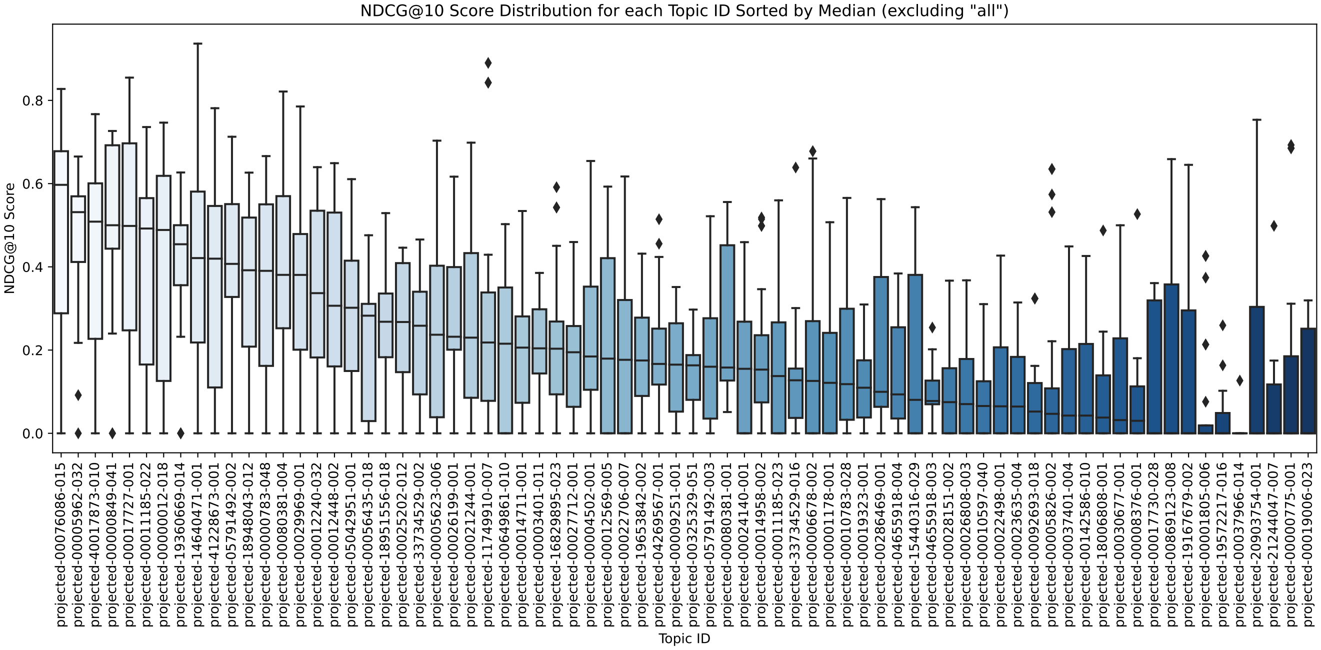


**Figure 2: nDCG@10 per Topic (T2M)**

for T2M and 70 queries for I2T. The overarching objective of our annotation guidelines is to identify the most suitable image that effectively complements the given section (or vice versa). However, it is important to note that we also accept instances where the selected image provides value by illustrating the entire article, even if it doesn't specifically correspond to the exact section under consideration. For a comprehensive overview of the annotation process, including the number of annotations and their associated relevance labels, please refer to Figure 1.

*Pooling depth.* In our evaluation process, NIST carefully considers the depth of pooling for different tasks. For the Image Suggestion (T2M) task, we annotate the top 25 candidates when assessing baseline runs, and we extend this to 30 candidates for the participant runs. In contrast, for the Image Promotion (I2M) task, we annotate the top 30 candidates from all runs. This tailored pooling depth strategy allows us to comprehensively evaluate the performance of different systems in diverse contexts.

*Relevance judgements.* Our annotation process involves categorizing candidate results into three distinct relevance levels to capture the nuances of their suitability. Annotators make relevance judgements based on the following criteria:

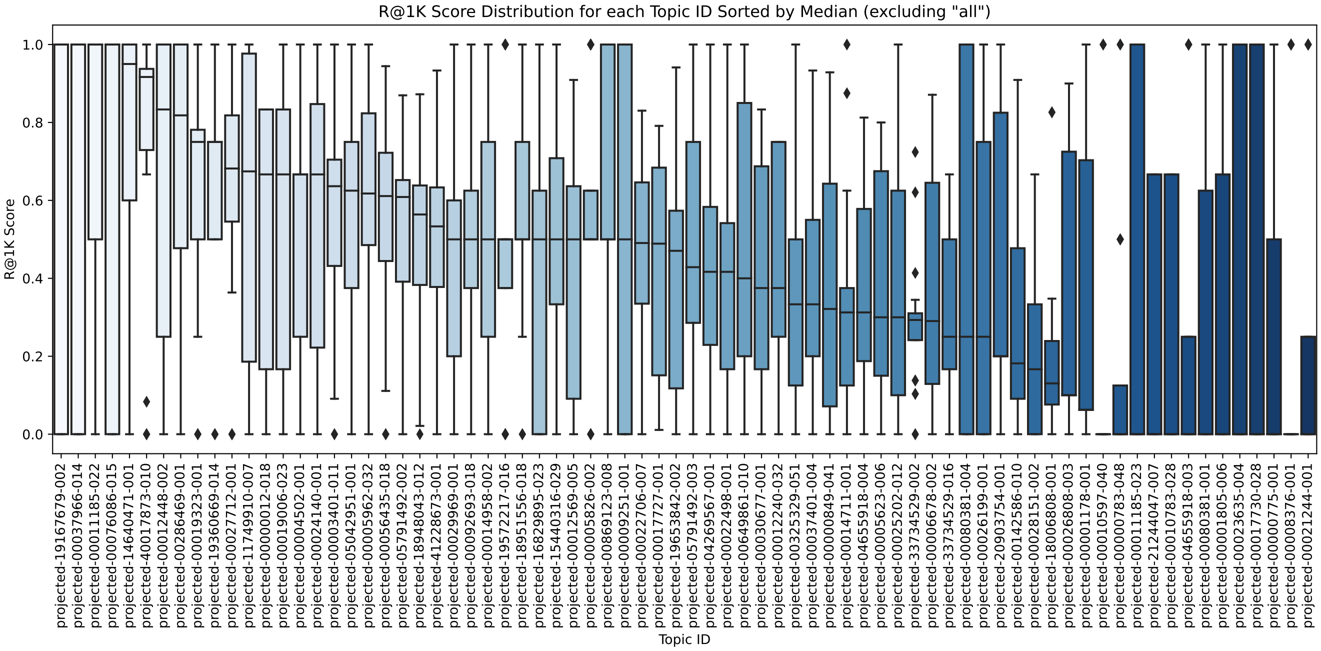


**Figure 3: R@1K per Topic (T2M)**

**Table 3: Image Promotion (M2T) Results, ordered by nDCG@10**

| Run ID | Team | Retrieval | Multimedia | mAP | nDCG@1K | nDCG@10 | Recall@1K | Success@1 | Success@10 |
|---|---|---|---|---|---|---|---|---|---|
| b_fsum_ all_i2t | baselines | Hybrid | Image+Caption | 0.2100 | 0.6308 | 0.4029 | 0.9776 | 0.2131 | 0.6066 |
| b_splade_pp_i2t | baselines | Learned-Sparse | Caption | 0.2408 | 0.4687 | 0.3691 | 0.7821 | 0.1967 | 0.5574 |
| b_clip_vitg14_laion_i2t | baselines | Dense | Image | 0.0776 | 0.4243 | 0.2790 | 0.6849 | 0.0656 | 0.3279 |
| b_bm25_i2t | baselines | Sparse | Caption | 0.1992 | 0.3163 | 0.2784 | 0.4314 | 0.2295 | 0.4098 |
| b_clip_vith14_laion_i2t | baselines | Dense | Image | 0.0751 | 0.3996 | 0.2403 | 0.6634 | 0.0656 | 0.3934 |
| b_clip_vitl14_laion_i2t | baselines | Dense | Image | 0.0650 | 0.3703 | 0.2103 | 0.5996 | 0.0656 | 0.2623 |
| finetune_base_i2t | uogTr | Dense | Image | 0.0588 | 0.2695 | 0.1864 | 0.4828 | 0.1148 | 0.2295 |
| b_clip_vitb32_laion_i2t | baselines | Dense | Image | 0.0565 | 0.2755 | 0.1597 | 0.4761 | 0.0820 | 0.1967 |
| finetune_large_i2t | uogTr | Dense | Image | 0.0362 | 0.2516 | 0.1213 | 0.5403 | 0.0492 | 0.2131 |
| b_flava_i2t | baselines | Dense | Image | 0.0155 | 0.0916 | 0.0595 | 0.1644 | 0.0164 | 0.0492 |
| pretrain_base_i2t | uogTr | Dense | Image | 0.0018 | 0.0148 | 0.0110 | 0.0184 | 0.0000 | 0.0328 |

- Non-Relevant (0): Candidates that are deemed not relevant to the task at hand fall into this category. They do not contribute meaningfully to the intended purpose.
- Relevant but Not Ideal (1): Candidates that possess some degree of relevance to the task but are not considered the best or most fitting options are categorized as relevant but not ideal. They provide value but may have room for improvement.
- Good Match (2): The highest level of relevance is assigned to candidates that are an excellent match for the task. These candidates align exceptionally well with the criteria and serve the intended purpose effectively.

*Feedback from the annotation period.* During the annotation period, valuable insights were gathered from our annotators, shedding light on specific challenges and unique aspects of the dataset. The feedback highlights the following key observations:

- Difficulty in Annotating Images Without English Captions: Annotators encountered challenges when tasked with annotating images that lacked English captions. The absence of textual context made it challenging to identify and understand certain concepts or individuals solely based on visual content. This highlights the importance of textual information in facilitating image understanding and relevance assessment.
- Uniqueness of Image Collection: The image collection presented in the dataset was noted to have certain characteristics that set it apart from conventional image datasets. This distinctiveness is attributed to the source of the images, which is WikiMedia. As a result, the dataset exhibits unique characteristics that may differ from more standardized image collections. Understanding these idiosyncrasies is essential for accurate evaluation and interpretation of results.

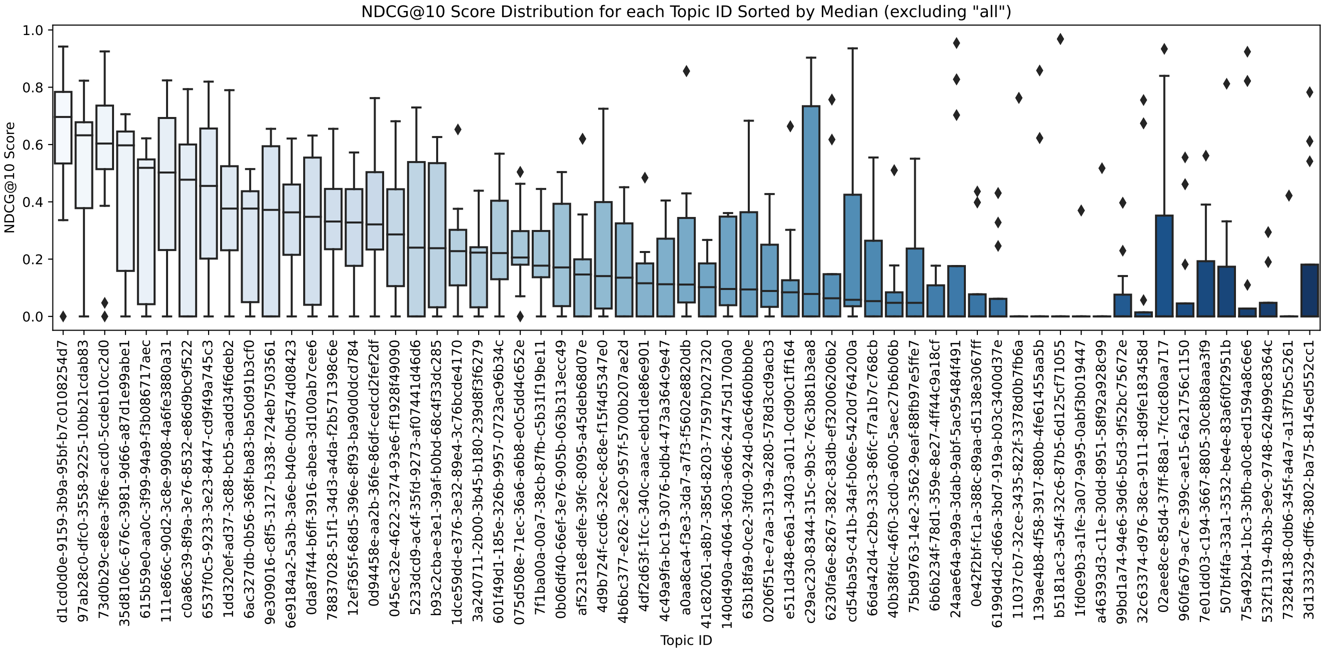


**Figure 4: nDCG@10 per Topic (M2T)**

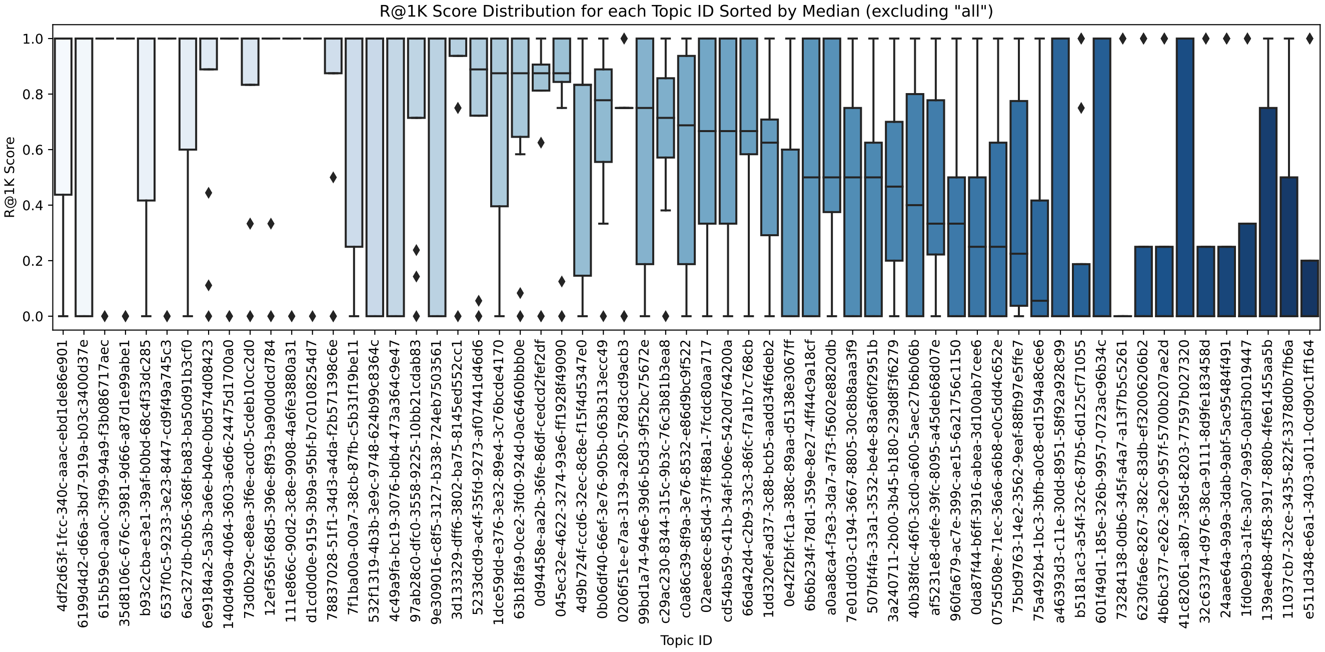


**Figure 5: R@1K per Topic (M2T)**

This feedback underscores the importance of considering the nuances of the dataset and its content when designing evaluation tasks and interpreting the outcomes. It also highlights the need for strategies to address challenges related to image understanding and relevance assessment, especially in scenarios where textual context may be limited.

## 2.4 Baselines

In our effort to enrich the diversity of annotations and submissions, we incorporate baseline runs based on three primary approaches for multimedia retrieval. These approaches leverage different techniques to represent multimedia information, thereby providing a broad spectrum of methods for evaluation. The baseline methods include:

- **Dense Multimodal Models**: We employ state-of-the-art dense multimodal models, specifically OpenCLIP [3], BLIP [5], and FLAVA [7]. These models utilize images to represent multimedia information, offering a robust approach to retrieval.
- **Traditional Sparse Retrieval with BM25**: We apply traditional sparse retrieval using BM25, employing captions as the primary source of multimedia information. This method serves as a baseline to assess the performance of more advanced techniques.
- **Learned Sparse Retrieval with SPLADE**: We utilize SPLADE [1, 2], a learned sparse retrieval approach, also using captions to represent multimedia information. SPLADE++ ED model [1] is specifically employed for this purpose.

Here is a breakdown of the individual baseline systems:

- `b_bm25`: Traditional sparse retrieval using `Anserini` default parameters.
- `b_splade_pp`: Learned sparse retrieval employing the SPLADE++ ED model [1].
- `b_clip_vit{g14,h14,l14,b32}_laion`: Different sizes of dense OpenCLIP [3] models are used for representation.
- `b_flava`: A dense retrieval model based on FLAVA [7].
- `b_fsum_all`: An ensemble of all other baseline systems, aggregating their scores through the sum of normalized scores, with values ranging from 0 to 1.

## 2.5 Participants

For this task, we had three participating teams, in addition to the baseline submissions from organizers:

*UAmsterdam:* UAmsterdam submitted T2M runs using Learned Sparse Retrieval techniques. In their approach, the query encoder consistently utilized a DistilBERT model, while the multimedia representation could either be the caption or the image, depending on the specific model. The training process consumed approximately 18 hours on an A100, and indexing took around 80 hours. Their Anserini-based system operated at less than 100 queries per second (QPS) on 60 CPUs. It's important to note that their indexing process only included images with English captions.

*IRLab-Amsterdam:* IRLab-Amsterdam submitted a single run that involved adapting a pre-existing multi-modal model (CLIP) into a Learned Sparse method. This adaptation was achieved through the training of a Multi-Layer Perceptron (MLP) and a Masked Language Modeling (MLM) head component. Adapting the model took approximately 8 hours on an A6000 GPU, while indexing was completed in just half an hour. Their reported query latency stands at around 3 seconds.

*uogTr:* The uogTr team submitted three runs using dense retrievers. Two of these runs were based on a model they pre-trained with a size referred to as "base" (according to CLIP nomenclature), and the third run was based on a “large” fine-tuned model. The pretraining phase took around 10 hours with the assistance of four A6000 GPUs, while fine-tuning the base model required approximately 25 hours, and the large model demanded 75 hours. The precise indexing and retrieval times for their systems were not specified.

# 3 RESULTS

In this section, we present the results for two distinct tasks: the Image Suggestion Task (T2M) and the Image Promotion Task (M2T) as shown in Table 2 and Table 3, respectively.

*Image Suggestion Task (T2M).* Firstly, in our analysis of R@1k, we observed that the hybrid model achieved the best results. This outcome was anticipated due to the hybrid model's capacity to harness a broader range of signals by utilizing both image and caption information. However, it's essential to acknowledge that the hybrid model comprises multiple evaluated models, which could contribute to result variability. Secondly, it is somewhat disheartening to note that there isn't a significant advantage observed between models that utilize either the image or the caption for representation. We suspect that this lack of distinction may stem from potential biases in the annotation process, which may have favored images with English captions due to the annotation's inherent difficulty (further analysis is provided in the subsequent section). Thirdly, we also observed that while the hybrid model faced challenges in terms of nDCG@10, it exhibited improvement in nDCG@1K. This positive development offers some optimism for the viability of the hybrid strategy, incorporating both captions and images to convey multimedia information effectively. In conclusion, it appears that there is substantial room for progress in this task. This assertion is supported by the notable difference in nDCG@10 scores observed here compared to the benchmarks commonly seen in TREC tasks.

*Image Promotion Task (M2T).* In our attempt to apply a similar analysis as in the T2M task, we initially note that this task exhibits less diversity in positive outcomes. The top two methods in terms of nDCG@10 also display notably high R@1k (up to 97%). This result was expected, considering that only one team participated in this task, supplemented by baseline methods. Once again, akin to the T2M task, we observe limited advantages in employing the image alone for representation. The nDCG@10 scores in this task are comparatively low when compared to other tasks, signifying significant room for improvement. However, a notable distinction from the T2M task is that in the M2T task, the hybrid approach yielded the most successful results. In summary, while the M2T task shows promise, it also highlights areas for improvement, particularly in enhancing the utilization of images for promoting content. Notably, the success of the hybrid approach in this task sets it apart from the T2M task.

# 4 ANALYSIS

## 4.1 Annotations per Topic

Figure 1 compares the performance of two tasks: Image Suggestion (T2M) and Image Promotion (M2T). Panel (a) of Figure 1 showcases the output of the Image Suggestion method (T2M). We observe a high frequency of blue bars across all topic IDs, signifying a large number of images with a relevance level of ’0’. There is a noticeable pattern where the number of highly relevant images (green) is consistently lower than those of moderate relevance (orange), which in turn is lower than the least relevant images (blue). Both **T2M (a)** and **M2T (b)** generate a larger number of low-relevance labels, with high-relevance labels being the least frequent. This observation suggests that it is still a challenging task for most systems, highlighting a potential area for algorithm refinement. It is worth noting that, M2T appears to have more portion of relevant labels compared to T2M, we believe this comes from the density of English Wikipedia, with a large number of articles that expand what sometimes would be a single section (leading to the image being relevant to the section and to the expanding articles).

## 4.2 Metrics per Topic

In this section, we analyze the results in terms of nDCG@10 and R@1K for two distinct tasks denoted as T2M and M2T. The evaluation encompasses all the systems mentioned earlier, and we present the findings using box plots in Figure 2, Figure 3 (for T2M), and Figure 4, Figure 5 (for M2T). Upon closer examination of these figures, it becomes evident that both tasks exhibit similar trends. The systems tend to perform sub-optimally in terms of nDCG@10 while maintaining relatively high R@1K scores for most topics. This suggests that there is substantial room for improvement in terms of early precision for both tasks.

Notably, M2T demonstrates superior performance in terms of R@1K compared to T2M. This observation aligns with the insights gained from Figure 1: M2T has more portion of relevant labels compared to T2M. We speculate that this observation may be attributed to annotators' tendencies to overlook images lacking English captions when performing the T2M task, resulting in more non-relevant labels. In contrast, for the M2T task, all candidate selections involve well-structured English Wikipedia articles.

## 4.3 Example annotation

*T2M topic* `projected-19572217-016`*: Diabetes - Diagnosis.* One example of topic on the T2M was the diagnosis section of the diabetes page. We depict 3 examples of good matches (rel=2) in Figure 6 note how even without an English caption there might be images that are relevant to it. We also noticed that some images without captions (or without English captions) got selected, which is a positive, but may have hindered teams that were not able to use images without English caption. Not surprisingly, this topic is also one with the worst median nDCG@10 and largest variation on R@1k (some models 100%, some 0% and an average of around 50%). Looking at the images the one without the caption looks like the perfect candidate for illustrating the section, while the other two are good matches.

*M2T topic* `1dd320ef-ad37-3c88-bcb5-aadd34f6deb2` *- Map of Kenya.* In Figure 7, we present an image depicting a map of Kenya. We have chosen this particular image for analysis because it offers a distinct departure from traditional image caption datasets; it is not a typical "natural" image, but rather a map. Additionally, this image was assigned the highest number of positive sections. In total, we identified 90 sections related to this topic, out of which 24 were deemed to be particularly relevant. It is noteworthy that these relevant sections predominantly originate from the same set of pages, owing to the substantial volume of information available on English Wikipedia. For instance, we observed references to Geography, Demography, Politics, and the Outline of Kenya, which exist in English but may not have equivalents in other languages. This observation hints at the potential for discovering intriguing insights by exploring less densely populated languages on Wikipedia, as they may offer a more diverse range of multimedia content with fewer overlapping or redundant pages.

# 5 CONCLUSION

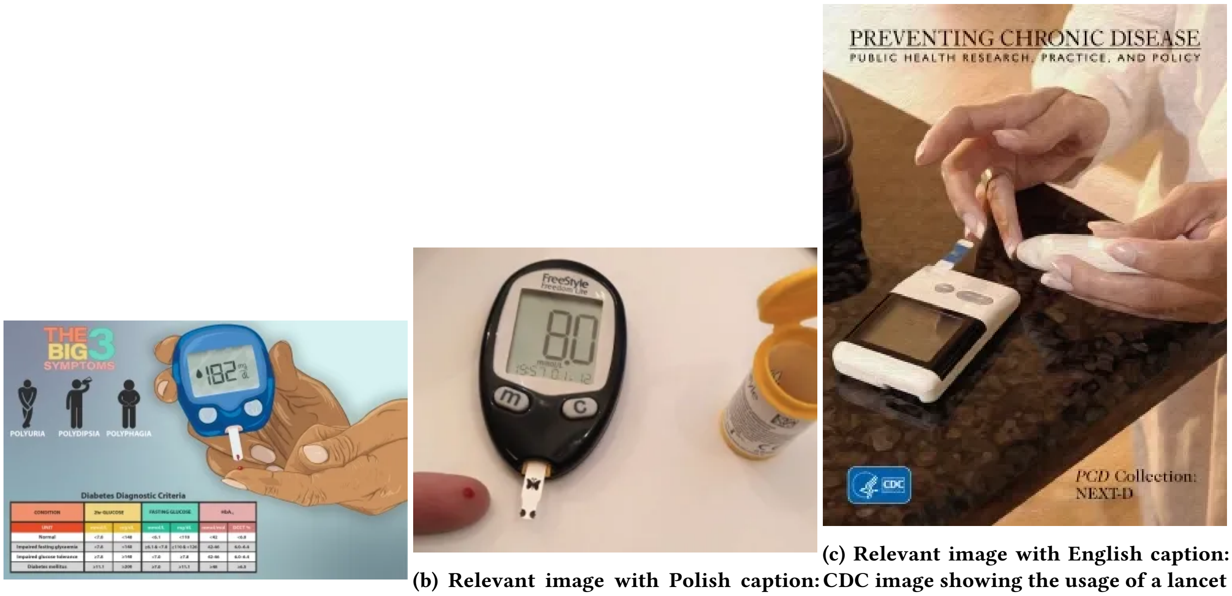
The findings of the TREC2023 AToMiC presented in this study highlight the significant advancements in multimedia retrieval systems nowadays, particularly in the tasks of Image Suggestion (T2M) and Image Promotion (M2T). The results underscore the efficacy of the hybrid model, which, by integrating both image and textual caption information, outperforms models that rely on single modalities. This success can be attributed to the model's ability to capture a more comprehensive understanding of the content, tapping into the nuanced interplay between visual and textual elements.

## 5.1 Discussion

The hybrid model's performance in the R@1k metric indicates a substantial step forward in aligning multimedia content with complex user queries. By leveraging a broader range of signals, the model not only enhances the relevance of suggested images but also provides insights into the dynamic nature of how multimedia content is consumed and understood. These insights are crucial for developing more sophisticated algorithms that can cater to diverse user needs in an increasingly multimedia-rich online environment.

However, the discussion would be incomplete without acknowledging the challenges that accompany the interpretation of multimedia content. The complexity of visual and textual interrelations poses a substantial obstacle, one that necessitates careful consideration of context, cultural nuances, and the potential for biases. The hybrid model's interpretability and its decision-making process must be examined to ensure fairness and equity in content curation. Moreover, the study's focus on English-language content raises questions about the model's applicability to the multilingual and multicultural landscape of the internet. Future iterations of this research must endeavor to incorporate a broader linguistic spectrum to ensure inclusivity and relevance across different demographics.

The engagement with WikiMedia Foundation also suggests a growing need for collaboration between AI research and content platforms. The goal of aligning model outputs with the content needs of real-world applications, like those of WikiMedia, underscores the importance of practical, user-centered research in the domain of multimedia content curation. In conclusion, the research presented provides a promising direction for multimedia content creation, with the hybrid model setting a new benchmark for performance. It opens avenues for further research, particularly in refining the model's capabilities, expanding its multilingual proficiency, and ensuring its alignment with diverse user expectations and ethical standards.


THE BIG 3 SYMPTOMS
POLYURIA
POLYDIPSIA
POLYPHAGIA
Diabetes Diagnostic Criteria
PREVENTING CHRONIC DISEASE
PUBLIC HEALTH RESEARCH, PRACTICE, AND POLICY
CDC
PCD Collection:
NEXT-D

**(a) Relevant image without caption**

**(b) Relevant image with Polish caption: Průběžné měření hladiny cukru v krvi**

**(c) Relevant image with English caption: CDC image showing the usage of a lancet and a blood glucose meter**

**Figure 6: Examples of relevant images for topic `projected-19572217-016`, Diabetes - Diagnosis.**

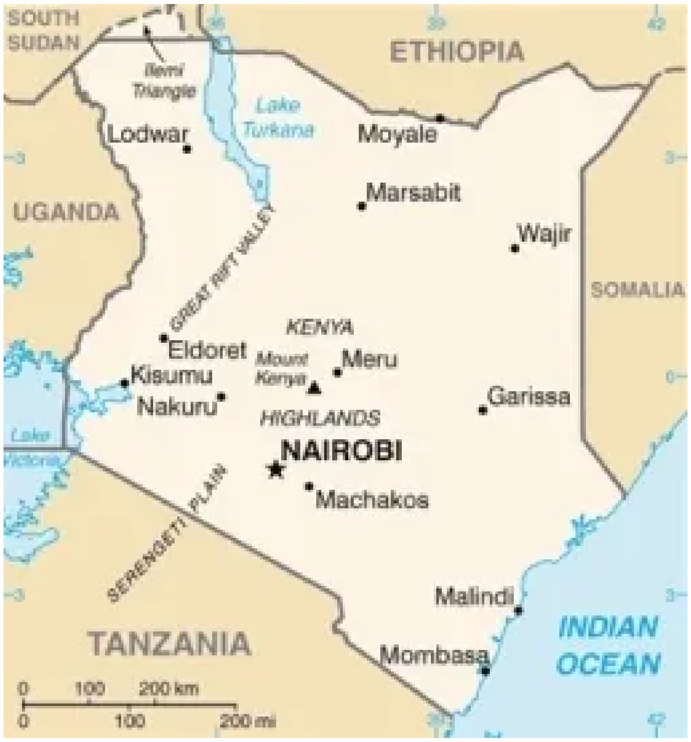


**Figure 7: Example of M2T topic `1dd320ef-ad37-3c88-bcb5-aadd34f6deb2` - Map of Kenya**

## 5.2 Future Work

The TREC2023 AToMiC has provided a comprehensive overview of the state-of-the-art in multimedia content creation, with a special emphasis on the interplay between various content modalities. The success of the hybrid model in the Image Suggestion and Image Promotion tasks demonstrates the value of integrating multiple sources of information to improve the relevance and quality of multimedia content suggestions.

The research conducted has significant implications for the development of AI-driven multimedia platforms, particularly in the context of content curation and recommendation systems. The hybrid model's adeptness at interpreting complex queries and returning relevant content sets a new standard for how such systems can be developed and refined.

Looking forward to TREC2024 and beyond, the study points towards several key areas of future work:

(1) **Multilingual and Multicultural Expansion.** To diminish the bias towards English-centric content and to embrace Wikipedia's multilingualism, there is a pressing need to introduce multilingual topics and recruit multilingual annotators.
(2) **Continuous Improvement and Evaluation.** The integration of a year-round leaderboard, leveraging the labels from TREC2023, could facilitate ongoing improvement and benchmarking of models.
(3) **Collaborative Labeling.** Working alongside content platforms like WikiMedia to validate the labeling process ensures that the research remains aligned with the practical needs of such organizations.
(4) **User-Centric Evaluation.** Adopting preference-based evaluations could further refine the understanding of user satisfaction and content relevance.

## ACKNOWLEDGMENTS

We would like to thank Catherine Zhou, Jack Lee and Jasper Xian for their invaluable help in experimenting and developing tools to work with AToMiC.

## REFERENCES

[1] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural ir models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2353–2359.

[2] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2288–2292.

[3] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP.* https://doi.org/10.5281/zenodo.5143773 If you use this software, please cite it as below..

[4] Joel R. Levin and Alan M. Lesgold. 1978. On Pictures in Prose. *ECTJ* 26, 3 (1978), 233–243.

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proc. of ICML.*

[6] Emily E. Marsh and Marilyn Domas White. 2003. A Taxonomy of Relationships between Images and Text. *Journal of documentation* 59, 6 (2003), 647–672.

[7] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A Foundational Language and Vision Alignment Model. In *Proc. of IEEE/CVF CVPR.* 15638–15650.

[8] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning. In *Proc. of SIGIR.* 2443–2449.

[9] Jheng-Hong Yang, Carlos Lassance, Rafael Sampaio De Rezende, Krishna Srinivasan, Miriam Redi, Stéphane Clinchant, and Jimmy Lin. 2023. AToMiC: An Image/Text Retrieval Test Collection to Support Multimedia Content Creation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2975–2984.