Retrieval Evaluation for Long-Form and Knowledge-Intensive Image–Text Article Composition

Jheng-Hong Yang[1], **Carlos Lassance**[2], **Rafael Sampaio de Rezende**[3], **Krishna Srinivasan**[4], **Stéphane Clinchant**[3], **Jimmy Lin**[1]

[1]University of Waterloo, [2]Cohere, [3]Naver Labs Europe, [4]Google Research

Abstract

This paper examines the integration of images into Wikipedia articles by evaluating image–text retrieval tasks in multimedia content creation, focusing on developing retrieval-augmented tools to enhance the creation of high-quality multimedia articles. Despite ongoing research, the interplay between text and visuals, such as photos and diagrams, remains underexplored, limiting support for real-world applications. We introduce AToMiC, a dataset for long-form, knowledge-intensive image–text retrieval, detailing its task design, evaluation protocols, and relevance criteria. Our findings show that a hybrid approach combining a sparse retriever with a dense retriever achieves satisfactory effectiveness, with nDCG@10 scores around 0.4 for Image Suggestion and Image Promotion tasks, providing insights into the challenges of retrieval evaluation in an image–text interleaved article composition context. The AToMiC dataset is available at https://github.com/TREC-AToMiC/AToMiC.

1 Introduction

The ability to produce high-quality image–text content, like poetry and essays, is crucial, with diverse applications in education and entertainment domains. The creation of high-quality multimedia content is a complex task, particularly on platforms like Wikipedia, which hosts more than 6 million articles and serves as a primary reference for millions of users around the world. The integration of relevant images into textual content is critical for enhancing reader engagement, comprehension, and the overall quality of knowledge dissemination. However, despite the availability of over 100 million media files on Wikimedia Commons, selecting and aligning images with corresponding text remains a significant challenge. This is particularly evident in knowledge-intensive and long-form content, where the relevance of an image is not just a matter of keyword matching but requires deep contextual and semantic alignment with the text (Zhang et al., 2023; Dong et al., 2024; Zhang et al., 2024). Developing authoring tools to assist in multimedia content creation is therefore critical yet challenging for platforms such as Wikipedia.[1]

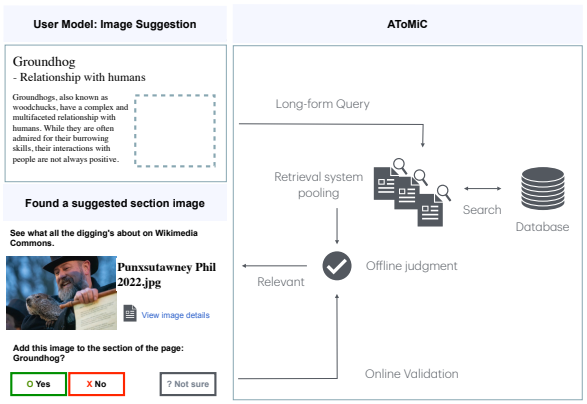


Figure 1: Conceptual plot illustrating the scope of AToMiC, featuring an image suggestion for the article *Groundhog - Relationship with humans*.

Recent advances in foundation models have significantly improved the ability to learn joint representations of images and text across diverse datasets (Radford et al., 2021; Li et al., 2022; Singh et al., 2022; Liu et al., 2024; Zhang et al., 2023; Beyer et al., 2024). These models leverage vast amounts of image–text data to align visual and textual inputs, achieving remarkable success in a variety of retrieval tasks. However, they are primarily designed to align structured, shorter texts, or alternative texts to perform effectively. More specifically, many models struggle to accurately recognize tailed entities represented in text (Hu et al., 2023; Chen et al., 2023). Taking the article in Figure1 as an example, recognizing entities like Punxsutawney Phil, a central figure in Groundhog Day celebration, can be challenging solely from an

[1]https://www.mediawiki.org/wiki/Structured_Data_Across_Wikimedia/Image_Suggestions

Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia, pages 36–45

image or text description.[2] This reliance poses difficulties when models are applied to more complex tasks, such as retrieving images opted for long-form texts, e.g., a section, where queries are implicit, long-form, and require a deep understanding of context, semantics, and world knowledge.

To tackle the challenges, we initiated the AToMiC (Authoring Tools for Multimedia Content Creation) project, specifically designed for evaluating image–text retrieval within the context of multimedia content creation for Wikipedia articles. Unlike previous approaches, AToMiC focuses on the unique challenges posed by using entire, knowledge-intensive articles as implicit queries. This requires a sophisticated understanding of the article's content and its purpose, ensuring that the retrieved images not only match the text but also contribute meaningfully to the article's overall narrative and informational value to content creators.

We introduce two key retrieval tasks in assisting multimedia article composition:[3]

- Image Suggestion Task (T2M): This task focuses on text-to-image retrieval, where the goal is to retrieve images that best enhance specific sections of text.

- Image Promotion Task (M2T): This task involves image-to-text retrieval, where the objective is to identify the most suitable textual context for existing images.

To support these tasks, we worked with NIST to curate 24K and 14K graded relevance labels, respectively, using 16 different retrieval systems, ranging from widely used vision–language pretrained models (Radford et al., 2021; Li et al., 2022) to summarization-based systems (Long et al., 2024) and learned sparse retrieval systems (Nguyen et al., 2024). Our findings indicate that while many image–text retrieval models have been proposed in recent years, they still require strong signals from image captions to deliver relevant results. Additionally, we observed that integrating CLIP with a text-based learned sparse retrieval system (Formal et al., 2021, 2022) can enhance the overall effectiveness of a hybrid retrieval system, achieving approximately 0.4 in nDCG@10.

We further validated the relevance labels in a real-world context by attaching relevant images to Wikipedia articles and obtaining feedback from experienced Wikipedia editors. Specifically, in June 2024, we selected 14 vital articles and attached 18 relevant images based on the relevance labels we curated. During the past three months, the survival rate of these images has been approximately 94%.[4] This result highlights the effectiveness of our proposed evaluation framework in real-world applications, extending beyond laboratory settings.

The remainder of this paper is structured as follows: Section 3 provides a detailed overview of the evaluation process; Section 4 presents the task outcomes; Section 5 offers an analysis of the resources and labels generated; Section 6 discuss our findings in applying AToMiC in the wild; and Section 7 concludes our discussion.

2 Related Work

Existing works such as WebQA (Chang et al., 2022), CIRR (Liu et al., 2021), FashionIQ (Wu et al., 2021), ReMuQ (Luo et al., 2023), OVEN (Hu et al., 2023), and INFOSEEK (Chen et al., 2023) have made substantial contributions to various multimodal retrieval tasks. For instance, WebQA excels in visual question answering tasks, using multimodal input to answer complex open-domain questions. CIRR and FashionIQ are tailored for composed image retrieval and attribute-based searches, particularly within the fashion industry, where image modifications based on textual input are common. ReMuQ focuses on retrieving content to answer multimodal questions, while OVEN emphasizes object-centric and zero-shot retrieval, respectively, often within knowledge-rich domains like Wikipedia. INFOSEEK enhances retrieval through semantic navigation and knowledge exploration, but it is better suited for explicit, well-defined queries.

3 Evaluation Overview

This section offers a thorough overview of AToMiC evaluation process in TREC 2023. We begin by introducing our foundational test collections, AToMiC, which serve as the cornerstone for our assessment. Following this, we explore the intricacies of our task design, providing a detailed examination of the challenges and objectives that shape the evaluation process. We then outline our evaluation protocols, focusing on critical aspects

[2] https://en.wikipedia.org/wiki/Groundhog_Day

[3] In our context, "images" refer to both the pixel values and their associated captions, hence the task is aptly termed as Text-to-Media (T2M) and Media-to-Text (M2T), respectively.

[4] 17 out of 18, as of August 2024

Task	Description	# Samples
T2M	Corpus (Images)	11,019,202
	Query (Train)	3,002,458
	Qrels (Train)	4,401,903
	Query (Eval)	74
	Qrels (Eval)	24,728
M2T	Corpus (Texts)	10,134,744
	Query (Train)	3,386,183
	Qrels (Train)	4,401,903
	Query (Eval)	61
	Qrels (Eval)	14,078

Table 1: Statistics of the AToMiC dataset. T2M: Image Suggestion; M2T: Image Promotion.

such as pooling depth and criteria for relevance judgments. To establish context and provide benchmarks, we introduce the baseline systems that serve as performance reference points. Additionally, we present participant reports, shedding light on the diverse approaches employed to address the tasks.

3.1 AToMiC Test Collection

AToMiC is an extension of the Wikipedia-based Image Text (WIT) dataset (Srinivasan et al., 2021), specifically designed to support two key retrieval tasks in multimedia content creation: image suggestion and image promotion (see subsection 3.2). Table 1 provides a summary of the key statistics. The corpus comprises approximately 10 million *documents*, integrating both text and image collections. To facilitate system development, we provide around 3 million *queries* and 4 million sparse *qrels* (relevance judgments) derived from image–text pairs extracted from Wikipedia.[5] Additionally, we offer 24K and 14K dense qrels for the 74 and 61 evaluation topics of the respective tasks.[6]

3.2 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).

[5]On average, there is only one image per section.

[6]Find more details in (Yang et al., 2023).

Image Suggestion (T2M). The Image Suggestion (T2M) task focuses on the scenario of identifying relevant images to enhance textual content. For this task, we selected 500 imageless sections from articles listed in Wikipedia's Level 3 Vital Articles.[7] The Vital Articles list is a carefully curated collection of articles considered essential for providing a comprehensive overview of human knowledge. These articles cover a wide range of topics and serve as a foundational reference point for readers seeking authoritative information.

Our focus on these specific sections stems from their critical importance within the Wikipedia ecosystem. By initially evaluating them in the English language, we aim to identify opportunities to improve the representation of vital content across other languages. Following the annotation process, we further refined the dataset by filtering out poorly performing and inappropriate sections, resulting in 74 test queries for this task, as shown in Table 1.

Image Promotion (M2T). The Image Promotion (M2T) task focuses on a search scenario where image providers aim to identify the most appropriate attachment points within an article's text sections. To simplify the image selection process, we employ a multi-stage filtering approach using images from the image suggestion task. Initially, we apply three fusion methods—top-K, RRF, and RBP—to combine the image ranking lists generated by our baseline systems for 200 T2M topics, with a pooling depth set at 20. We then merge the resulting image pools and remove duplicate images based on their IDs. Finally, we eliminate near-duplicate images using the `fastdup` library and randomly select 200 images as candidates for image topics.[8] After the annotation process, we further refine the dataset by filtering out poorly performing and inappropriate images, resulting in 61 test queries for this task, as detailed in Table 1.

Metrics. In assessing the effectiveness of retrieval systems, 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, which allows us to gauge the quality of our results with fine granularity. In addition to nDCG, we recognize the

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

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

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 provides a more comprehensive view of system performance across different evaluation scenarios. By exploring these relationships, we aim to gain insights into the strengths and limitations of the retrieval systems involved in AToMiC.

3.3 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 the evaluation criteria, the final evaluation is performed for 80 queries for T2M and 70 queries for M2T. The objective of our annotation guidelines is to identify the most suitable image that complements the given section (or vice versa). However, it is important to note that we accept instances where the selected image provides value by illustrating the entire article, even if it does not correspond to the exact section under consideration.

Pooling. Pooling is a classical method adopted in early TREC evaluations and used to select documents for human assessment. This approach merges the top-ranking results from multiple runs into a single pool, with only the documents within this pool being evaluated. Collaborating with NIST, we adjust the depth of pooling based on the specific task at hand. For the Image Suggestion (T2M) task, we annotate the top 25 candidates during baseline assessments and expand this to 30 candidates for participant runs. Conversely, in the Image Promotion (M2T) task, we consistently annotate the top 30 candidates across all runs.

Relevance Judgments. Our annotation process involves categorizing candidate results into three graded relevance levels to capture the nuances of their suitability. NIST annotators make relevance judgments based on the following criteria:

- 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 and serve the intended purpose effectively.

3.4 Baseline Systems

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 utilize different techniques to represent multimedia information, thereby offering a comprehensive range of methods for evaluation. The baseline methods include:

Dense Retrieval Models. We employ representative dense retrieval models with pretrained vision–language models, specifically OpenCLIP (Ilharco et al., 2021), BLIP (Li et al., 2022), and FLAVA (Singh et al., 2022). We apply these models in a zero-shot fashion and only encode the pixel values of images without accessing their captions.

Traditional Sparse Retrieval. We employ traditional sparse retrieval using BM25, utilizing captions as the sole representation of images. This approach serves as a text-only baseline, providing a benchmark to evaluate the performance of more advanced techniques that integrate texts and images.

Learned Sparse Retrieval. We also utilize SPLADE (Formal et al., 2021, 2022), a learned sparse retrieval approach, to encode and index image captions. For this purpose, we specifically employ the SPLADE++ (ED) model (Formal et al., 2022).

Here is a breakdown of the individual baseline systems: (a) `b_bm25`: Traditional sparse retrieval using `Anserini` with default parameters $(k1, b) = (0.9, 0.4)$; (b) `b_splade_pp`: Learned sparse retrieval with the SPLADE++ (ED) model (Formal et al., 2022); (c) `b_clip_vit{g14,h14,l14,b32}`: Dense retrievers in various sizes provided by OpenCLIP (Ilharco et al., 2021); (d) `b_flava`: Dense retrieval using FLAVA (Singh et al., 2022); (e) `b_fsum_all`: An ensemble model that combines scores from all baseline systems by summing min-max normalized relevance scores.

3.5 Systems from Participants

UAmsterdam. UAmsterdam submitted T2M runs using Learned Sparse Retrieval techniques (Nguyen et al., 2024). Their approach consistently employed a DistilBERT query encoder, with multimedia representation varying between captions or images depending on the model. Training took around 18 hours on an A100 GPU, while indexing required approximately 80 hours. Their Anserini-based system processed fewer than 100 queries per second (QPS) using 60 CPUs. Notably, only images with English captions were included in the indexing process.

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 by training a Multi-Layer Perceptron (MLP) and a Masked Language Modeling (MLM) head. The adaptation process took approximately 8 hours on an A6000 GPU, with indexing completed in just 30 minutes. Reported query latency was ≈ 3 seconds.

uogTr. The uogTr team submitted three runs using cascaded systems that combined a summarization model with CLIP (Long et al., 2024). Two runs utilized a pre-trained `base` model, while the third employed a fine-tuned `large` model. Pretraining took around 10 hours on four A6000 GPUs. Fine-tuning took 25 hours for the base and 75 hours for the large model.

4 Results

In this section, we present the results for two 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). In our analysis of Recall@1K, the hybrid model achieved the best results. This outcome was anticipated, likely due to its ability to leverage different information. However, since the hybrid model includes multiple evaluated models, this could contribute to result variability.

Interestingly, there was no clear advantage between models using either image or caption 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).

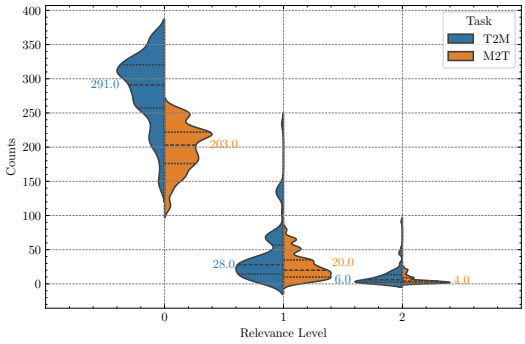


Figure 2: Violin plots of label counts for all topics, categorized by relevance level (0, 1, 2) and task types (T2M, M2T). The annotated values representing the median for each case.

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). For M2T task, our first note is that this task exhibits less diversity in positive outcomes since teams from Amsterdam did not participate this task. The top two methods in terms of nDCG@10 also display notably high Recall@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.

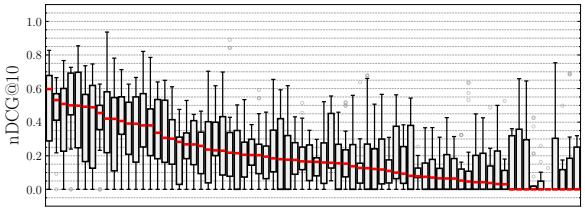
In summary, while the M2T task shows promise, it also highlights areas for improvement, particularly in enhancing the utilization of images for promoting content. Notably, the success of the hybrid approach in this task sets it apart from the T2M task.

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

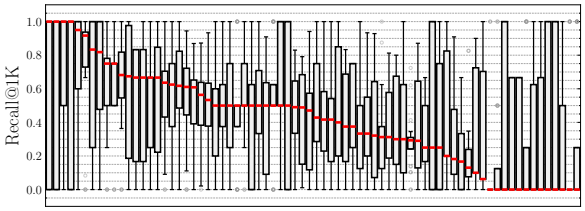
Run ID	Team	Retrieval	Multimedia	mAP	nDCG@1K	nDCG@10	Recall@1K	Success@1	Success@10
UvA-IRLab	IRLab-Amsterdam	Learned-Sparse	Image	0.1526	0.4460	0.4060	0.6452	0.2973	0.6081
b_splade_pp	baselines	Learned-Sparse	Caption	0.1501	0.4461	0.4051	0.6452	0.2838	0.6081
b_fsum_all	baselines	Hybrid	Image+Caption	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

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



(a) nDCG@10 for each topic

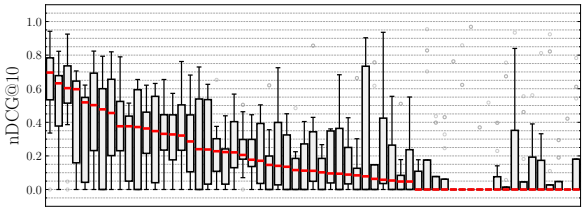


(b) Recall@1K for each topic

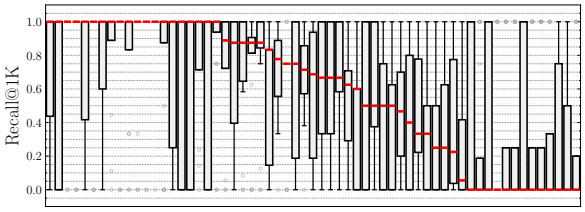
Figure 3: Image Suggestion (T2M) evaluation results. The box plots present the evaluation metrics by topic, with (a) nDCG@10 and (b) Recall@1K.

5 Analysis

Label Distribution by Topic. Figure 2 presents the distribution of labels across different relevance levels for two tasks: Text-to-Media (T2M) and Media-to-Text (M2T). The plot depicts the label counts at each relevance level, with separate distributions for each task. Annotations indicate the median number of labels within each category, where blue represents T2M and orange represents M2T, across relevance levels 0, 1, and 2. Both tasks show a similar trend: the majority of labels fall into the lowest relevance level (`rel` = 0), with medians of 291.0 for T2M and 203.0 for M2T, while the number of highly relevant labels (`rel` = 2) is substantially lower, with medians of 6.0 for T2M and 4.0 for M2T. T2M generally has a higher median count at the lowest relevance level compared to M2T, whereas M2T displays a slightly higher median at the moderate relevance level (`rel` = 1), with medians of 28.0 for T2M and 20.0 for M2T. This distribution underscores the ongoing challenge of assigning higher relevance labels, especially for systems processing the nuanced content typical of English Wikipedia articles. The findings suggest a need for further algorithmic improvements to effectively identify highly relevant pairs.



(a) nDCG@10 for each topic



(b) Recall@1K for each topic

Figure 4: Image Promotion (M2T) evaluation results. The box plots present the evaluation metrics by topic, with (a) nDCG@10 and(b) Recall@1K.

Evaluation Metrics by Topic. This section analyzes the results of nDCG@10 and Recall@1K for the tasks T2M and M2T across all evaluated systems. To understand system effectivness across different test topics, we present results using box plots in Figure 3a, Figure 3b (for T2M), and Figure 4a, Figure 4b (for M2T). Upon closer examination of these figures, it becomes evident that both tasks exhibit similar trends. The systems tend to perform suboptimally in terms of nDCG@10 while maintaining relatively high Recall@1K scores. This suggests that there is substantial room for improvement in terms of early precision.

In particular, M2T demonstrates superior performance in terms of Recall@1K compared to T2M. This observation aligns with the insights gained from Figure 2: M2T has a higher proportion 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 candidates are well-structured English Wikipedia articles.

6 AToMiC in the Wild

To assess the model's performance in more challenging tasks and its applicability to real-world scenarios, we deployed the *offline* relevance labels generated by AToMiC onto *online* Wikipedia articles. By attaching relevant images to selected Wikipedia sections, we aimed to evaluate the longevity and impact of these images in an authentic editorial environment. Acknowledging Wikipedia's *not a lab* policy,[9] all uploaded images were vetted by humans as part of the standard Wikipedia editing process.


Drawing [edit]

Main article: Drawing

Drawing is a means of making an image, illustration or graphic using any of a wide variety of tools and techniques available online and offline. It generally involves making marks on a surface by applying pressure from a tool, or moving a tool across a surface using dry media such as graphite pencils, pen and ink, inked brushes, wax color pencils, crayons, charcoals, pastels, and markers. Digital tools, including pens, stylus, that simulate the effects of these are also used. The main techniques used in drawing are: line drawing, hatching, crosshatching, random hatching, shading, scribbling, stippling, and blending. An artist who excels at drawing is referred to as a *draftsman* or *draughtsman*.[11]

Drawing and painting go back tens of thousands of years. Art of the Upper Paleolithic includes figurative art beginning between about 40,000 to 35,000 years ago. Non-figurative cave paintings consisting of hand stencils and simple geometric shapes are even older. Paleolithic cave representations of animals are found in areas such as Lascaux, France and Altamira, Spain in Europe, Maros, Sulawesi in Asia, and Gabarnmung, Australia.

In ancient Egypt, ink drawings on papyrus, often depicting people, were used as models for painting or sculpture. Drawings on Greek vases, initially geometric, later developed into the human form with black-figure pottery during the 7th century BC.[12]

With paper becoming common in Europe by the 15th century, drawing was adopted by masters such as Sandro Botticelli, Raphael, Michelangelo, and Leonardo da Vinci, who sometimes treated drawing as an art in its own right rather than a preparatory stage for painting or sculpture.

Christiaan Tonnis - Female Warrior #14 'Extinction', pencil and colored pencil on paper, 1981

Figure 5: Example of an image attached to a Wikipedia article, Visual Arts - Drawing. The image was selected as a "Good match" (`rel` = 2) annotation from the AToMiC dataset.[11]

We selected 14 level-3 vital articles on Wikipedia and manually attached 18 images chosen from highly relevant (`rel` = 2) image–text pairs (see more details in Appendix B). This experiment was conducted in June 2024, and only one image was subsequently removed by Wikipedia editors. Several key insights emerged from this end-to-end experiment:

Real-World Applicability. We achieved a high retention rate of 94% (17 out of 18 images at the time of submission) for the selected relevant images. On one hand, Figure 5 illustrates a survival test sample from our experiment on the Wikipedia page for Visual Arts—Drawing. Originally, the *Drawing* section had no attached image. We selected this image from the highly relevant (`rel` = 2) annotations due to its strong relevance and comprehensive coverage of the content of the section. On the other hand, the only image was removed by Wikipedia editors because the article already contained a sufficient number of images.[10] This result highlights additional challenges, such as the need for page-level relevance optimization and the nuanced judgment required for precise annotation.

[9] https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not

[10] https://en.wikipedia.org/w/index.php?title=Aircraft&diff=1227096926&oldid=1227092091

[11] Source: https://en.wikipedia.org/wiki/Visual_arts#Drawing; screenshot captured on August 28, 2024.

Challenges. To ensure the feasibility of real-world experiments, we introduced an additional filtering process to identify the *golden* labels from the NIST annotations. This process involved manually refining the initial 832 (`rel = 2`) down to 18 images according to our judgment. After the filtering process, we found that the focus shifts towards selecting the most impactful image, the one that truly enhances the article's content, similar to optimizing for the NDCG@1 metric. This requires applying additional criteria to ensure that the chosen image not only meets relevance standards but also significantly elevates the *overall quality* of the article. The selected images should be visually compelling and convey *key ideas* or *added value* relevant to the *entire article*, rather than merely aligning with specific sentences or words, as demonstrated in Figure 5.

7 Conclusion

This research highlights significant advancements in multimedia content creation, particularly through the integration of diverse content modalities. The success of hybrid models in Image Suggestion and Image Promotion tasks underscores the value of combining multiple information sources to enhance content quality and address complex user queries. The strong performance in Recall@1K indicates a substantial leap forward in developing algorithms suited to a multimedia-rich online environment.

However, challenges remain in interpreting multimedia content, especially due to the complexity of visual and textual interrelations. Addressing these challenges requires careful consideration of context, cultural nuances, and potential biases. Expanding beyond English-language content is crucial to make the model more applicable to the multilingual and multicultural landscape.

Collaboration with platforms like Wikimedia underscores the importance of aligning AI research with real-world content needs. Practical, user-centered research is essential for the continued development of effective multimedia content creation systems. Looking ahead, key areas for future work include reducing English-centric bias through multilingual expansion, establishing a year-round evaluation event or continuous (Chiang et al., 2024), and enhancing collaboration with content platforms. Implementing preference-based evaluations will also offer better insights into user satisfaction and content relevance.

In sum, we curated and studied a new benchmark dataset for multimedia content creation and opens avenues for further refinement, particularly in expanding multilingual capabilities and ensuring alignment with diverse user expectations and ethical standards.

Acknowledgements

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. PaliGemma: A versatile 3b VLM for transfer. *arXiv preprint arXiv:2407.07726*.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. WebQA: Multihop and multimodal QA. In *Proc. of IEEE/CVF CVPR*, pages 16495–16504.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? In *Proc. of EMNLP*, pages 14948–14968.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot Arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. InternLM-XComposer-2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural ir models more effective. In *Proc. of SIGIR*, pages 2353–2359.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proc. of SIGIR*, pages 2288–2292.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of

wikipedia entities. In *Proc. of ICCV*, pages 12065–12075.

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.

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*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Proc. of NeurIPS*, 36.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proc. of IEEE/CVF ICCV*, pages 2125–2134.

Zijun Long, Xuri Ge, Richard McCreadie, and Joemon M. Jose. 2024. CFIR: Fast and effective long-text to image retrieval for large corpora. In *Proc. of SIGIR*, page 2188–2198.

Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2023. End-to-end knowledge retrieval with multi-modal queries. In *Proc. of ACL*, pages 8573–8589.

Thong Nguyen, Mariya Hendriksen, Andrew Yates, and Maarten de Rijke. 2024. Multimodal learned sparse retrieval with probabilistic expansion control. In *Proc. of ECIR*, pages 448–464.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. of ICML*, pages 8748–8763.

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*, pages 15638–15650.

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*, page 2443–2449.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. FashionIQ: A new dataset towards retrieving images by natural language feedback. In *Proc. of IEEE/CVF CVPR*, pages 11307–11317.

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 *Proc. of SIGIR*, pages 2975–2984.

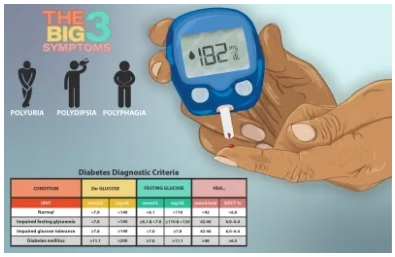
Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. 2024. InternLM-XComposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*.

Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. 2023. InternLM-XComposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

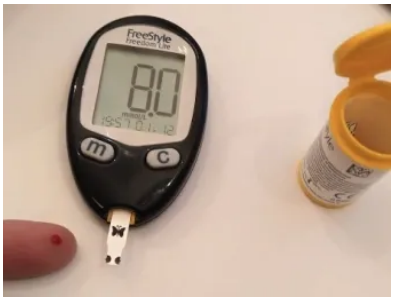
A Case Study

T2M topic: 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 Recall@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.

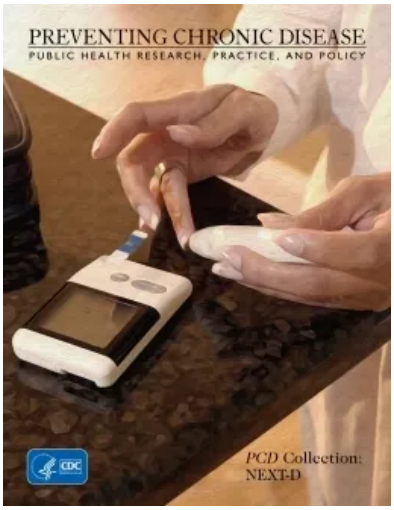
M2T topic: Map of Kenya. In Figure 7, we present an image depicting a map of Kenya. We have chosen this particular image for analysis because it offers a distinct departure from traditional image caption datasets; it is not a typical "natural" image, but rather a map. Additionally, this image was assigned 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



(a) Relevant image without caption



(b) Polish caption: Průběžné měření hladiny cukru v krvi



(c) 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

Wikipedia, as they may offer a more diverse range of multimedia content with fewer overlapping or redundant pages.

B In-the-Wild Evaluation

The following is the list of test topics (article - section) and the corresponding images that were uploaded to Wikipedia as part of the evaluation process:

- **Afterlife - Reincarnation**: (Uploaded 2 images)
- **Aircraft - History** (Uploaded 2 images)
- **Biotechnology - Definition** (Uploaded 1 image)
- **Grammar - Education** (Uploaded 2 images)
- **History of film - 1980S** (Uploaded 1 image)
- **Internal combustion engine - History** (Uploaded 1 image)
- **Iron age - History of the concept** (Uploaded 1 image)
- **Latin - Grammar** (Uploaded 1 image)
- **Mediterranean sea - Biogeochemistry** (Uploaded 1 image)
- **Orbit - History** (Uploaded 2 images)
- **Realism (arts) - Theatre** (Uploaded 1 image)
- **Roald Amundsen - Early life** (Uploaded 1 image)
- **Visual arts - Drawing** (Uploaded 1 image)
- **Wind - On other planets** (Uploaded 1 image)