MIRACLE: Multimodal Image-text Retrieval and Analysis for Contextual Long-form Evaluation

Md Messal Monem Miah
Texas A&M University
messal.monem@tamu.edu

Agent Chatterjee
Arizona State University
achatt39@asu.edu

Arindam Mitra
Microsoft
arindammitra2@gmail.com

Ruihong Huang
Texas A&M University
huangrh@cse.tamu.edu

Man Luo
Intel Lab
mluo26@asu.edu

Abstract

Multimodal retrieval is becoming increasingly vital as media platforms often feature content combining text and images. This is especially prevalent in long-form, information-rich articles. Recognizing the evolving complexity of such content, we introduce a novel benchmark, MIRACLE: Multimodal Image-text Retrieval and Analysis for Contextual Long-form Evaluation. MIRACLE distinguishes itself significantly from existing datasets by presenting unique challenges: 1) It extends the retrieval context with an average length of 402 words, surpassing the scope of prior benchmarks. 2) It intricately weaves multiple images within the textual narrative, demanding sophisticated interpretative analyses from retrieval systems. Upon evaluating state-of-the-art models using MIRACLE, we observe that the benchmark poses a considerable challenge in terms of both long context and complex interleaved image-text structures, indicating a need for more advanced models tailored to its demands. Our findings underscore MIRACLE's potential to drive progress in the field by pushing the boundaries of current multimodal retrieval systems.

1. Introduction

Information Retrieval (IR) is an crucial task with aim of sourcing relevant data from vast repositories in response to user queries. Traditionally focused on textual content [21], IR has evolved with the advent of Multimodal Information Retrieval (MIR) [2, 4, 14, 26], which enhances the retrieval process by incorporating various data types, including text, images, audio, and video. This advancement in IR illustrates the significant role of multimodal content in improving cognitive comprehension and retrieval accuracy [23]. In MIR, textual elements providing conceptual explanations are synergized with visual elements for a richer, more comprehensive understanding of the data.

The need for MIR becomes particularly evident in the analysis of long-context multimodal content. In scenarios such as academic research, legal document analysis, or multimedia storytelling, where meaning and relevance are embedded in extended narratives and detailed expositions, traditional short-context retrieval systems are inadequate. These systems often miss the nuanced connections present in longer sequences of text and multimedia. Therefore, understanding and extracting value from such complex data necessitates the retrieval of long-context and multiple images interleaved with text content.

Recognizing the ubiquitous nature of multimodal media in articles, we identify a gap in current multimodal retrieval datasets, which primarily focus on either text-based or image-based targets, overlooking the intricate challenge of retrieving targets composed of long text and multiple images. To bridge this gap, our work introduces a benchmark named MIRACLE: **M**ultimodal **I**mage-text **R**etrieval and **A**nalysis for **C**ontextual **L**ong-form **E**valuation. The source of MIRACLE is from WikiHow [1], where the query is a text and the target is a long text context interleaved with multiple images. An example form MIRACLE is given in Figure 1. MIRACLE challenges existing models, primarily limited by their ability to encode a single modality or handle extended encoding lengths. This new benchmark not only serves as an evaluative tool but also as a catalyst for advancing multimodal retrieval technologies, encouraging the development of models capable of adeptly interpreting the complexities of multimodal data.

Using the MIRACLE benchmark, we evaluate several leading retrieval models, including text-based retrieval models such as BM25 [20] and Contriever [6], and

[1] https://www.wikihow.com/Main-Page

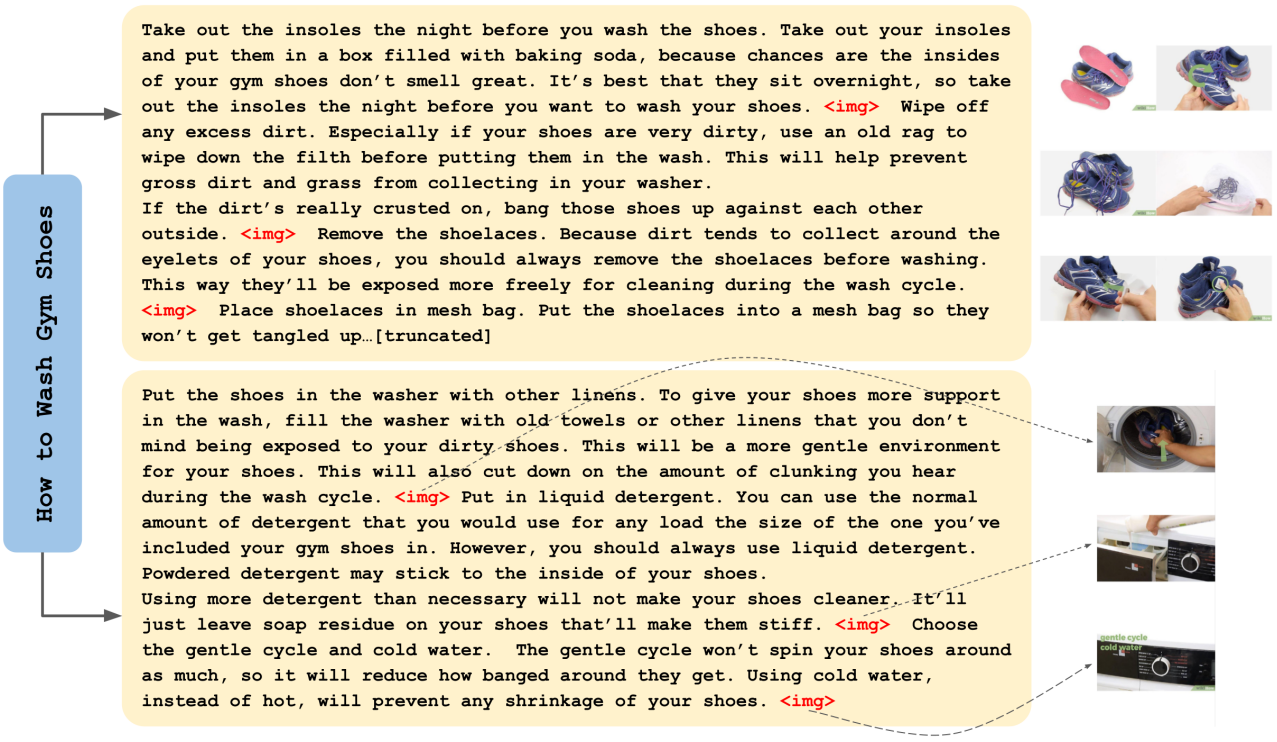

Figure 1. An example of MIRACLE, the input is a text query denoted by the blue box and the target(s) are long-form articles with multiple images as shown by yellow boxes. Each article contains multiple images and the position of the images are denoted by <*img*> identifiers in the text.

multimodal-based models such as CLIP [18] and BLIP [9]. To address the challenges posed by articles with long textual contexts and multiple images, we comprehensively analyze three aggregation functions: summation, mean, and maximum, each serving a unique role in processing multimodal content. Additionally, we explore hybrid models that utilize the strengths of the best text-based and image-based retrievers. This approach aims to leverage the distinct advantages of each modality. Our experimental findings are insightful and highlight several key aspects. 1) When comparing the performance of BM25 and CLIP on MIRACLE against previous benchmarks [2, 10], we observe significant performance drop on MIRACLE. This discrepancy underscores the unique challenges our dataset presents, establishing MIRACLE as a complementary benchmark in the field. 2) Our analysis reveals that text-only based retrievers generally outperform image-only based ones on our dataset. This indicates that the text content provides more information compare to the images. 3) We discover that integrating text and image scores results in slightly better performance than text-only based retrieval system. This outcome indicates the importance of considering information from both modalities in multimodal content retrieval, highlighting the synergistic potential of combining textual and visual data. These findings not only contribute to our understanding of multimodal retrieval dynamics but also guide future advancements in the field.

2. Related Work

2.1. Multimodal Retrieval Datasets

Multimodal retrieval tasks, defined as those involving a source or target composed of multiple modalities, are increasingly prevalent. A notable example is WebQA [2], a multimodal and multihop QA dataset, where queries are text-based questions and the context is provided through either images or text. Building upon this, ReMuQ [14] augments WebQA into a benchmark for multimodal retrieval, introducing queries that combine text and images, each holding mutually exclusive information, and targets comprised of text-based knowledge capable of answering such queries. In the domain of image-text modality, efforts to compile a retrieval corpus for the OkVQA dataset include sourcing text output from Wikipedia [4], Google search text snippets [12], or knowledge graphs [15]. FashionIQ [26] represents another approach, where queries consist of a text description alongside an image, with the target being a similar image that matches the features described in the text. Similarly, CIRR [11] adopts this approach but expands beyond the fashion domain, using text

descriptions to capture user needs. Dialogue-based image search is represented by ChatSearch [1], where multimodal dialogues serve as queries, with corresponding images as targets. OVEN-Wiki [5] compiles seven datasets with the goal of predicting a wiki-entity from a multimodal query, requiring the retrieval system to locate relevant wiki page information that may include both text and images. The InfoSeek [3] dataset introduces a QA approach with a sub-task focused on retrieving wiki-entities, where multimodal queries lead to outputs featuring Wikipedia page titles and images. EDIS [10] presents a multimodal web-content retrieval task centered on text queries from the news domain, aiming to understand entities and events. Finally, UniIR [25] amalgamates 8 existing datasets to create a benchmark for flexible multimodal searches, including multimodal-to-multimodal, image-to-multimodal, and text-to-multimodal searches. Our dataset distinguishes itself from these existing benchmarks in several aspects. It features significantly longer contexts and a higher number of images. Whereas most previous work focuses on entity-centric queries, our dataset shifts the emphasis to "How"questions, addressing a different spectrum of multimodal retrieval challenges.

2.2. Multimodal Retriever

A straightforward approach in multimodal retrieval involves converting images into corresponding text descriptions for use with a text retriever [4, 12]. However, this method may overlook the fine-grained details of the images. Vision and language (VL) models, such as those proposed in [7, 9, 19], have significantly enhanced multimodal retrieval performance. Although most VL models are not pre-trained specifically for retrieval tasks, they utilize text, visual, or cross embeddings to create dense semantic vectors representing images and text, followed by applying a scoring function for relevance assessment. Dual encoder architectures like CLIP and BLIP encode text and images separately. These models undergo pre-training with contrastive training objectives, closely mirroring retrieval tasks where the goal is to maximize the score of a relevant object from a set of candidates. Notably, the size of these candidate sets during training is often much smaller than the corpus size in retrieval tasks. Despite this, studies have shown that CLIP performs well in zero-shot cross-modal retrieval tasks. There are several methods to leverage CLIP for multimodal retrieval. One approach involves combining image and text scores to evaluate multimodal content [10, 14, 25]. Another strategy fuses the image and text embeddings, requiring an additional fusion module like a weighted vector for image and text scores [5], a simple multi-layer perception [5], or a more sophisticated cross-attention layer [10, 25]. Notably, Liu et al. [10] achieved significant improvements by fine-tuning CLIP and BLIP models with instruction and multitasking, a technique also effective in the NLP domain as demonstrated in[16, 17, 24]. Yasunaga et al. [27] approached multimodal content retrieval using CLIP, averaging the scores of each modality with L2 normalization. Luo et al. [14] developed a multimodal encoder for queries and a text knowledge encoder to form a comprehensive multimodal retriever. In our work, we focus on the zero-shot performance of existing multimodal retrievers and their efficacy in long-context multimodal retrieval tasks.

3. MIRACLE Benchmark

In MIRACLE benchmark, queries are text-based, while targets consist of multimodal content, featuring both long textual contexts and multiple images. This design sets MIRACLE apart from previous benchmarks, which typically do not combine extensive text with serials of images in their target content. Next, we will outline the procedure used to compile the MIRACLE dataset from WikiHow.

Dataset Collection We crawl data through the official wikiHow website[2], scraping text and images in an interleaved manner. We follow a two-step approach of getting all links, followed by parsing each link. In the first step, we initiate our crawlers through the official category pages (total count = 19) and recursively callback subsequent pages for wide coverage. To achieve comprehensive data collection, we employ pagination techniques to systematically retrieve articles, fetch newly indexed pages, and explore all sub-categories. This approach ensures a thorough and up-to-date coverage of all articles within a category. To ensure data quality, we only collect non-stubbed article links, i.e. only articles that are verified by an expert.

In the next step, the text and images are downloaded in the same order as they appeared in the original article. We assign a unique ID for each unique link and fetch the article title, article description and the corresponding *how-to* steps. Each *how-to* steps consist of their own unique heading, step description and an associated image. Each downloaded image is assigned a unique image ID (implemented as UUID) to avoid duplication of images. For a small number of articles that have videos instead, we only download the corresponding image thumbnail.

Data Statistic Table 1 presents the statistics of the MIRACLE dataset, comprising over 1 million images and approximately 240K multimodal articles. Each article in the dataset averages roughly 400 words and includes 4 images, signifying a scale substantially greater than that of other retrieval benchmarks.

[2] https://www.wikihow.com/Main-Page

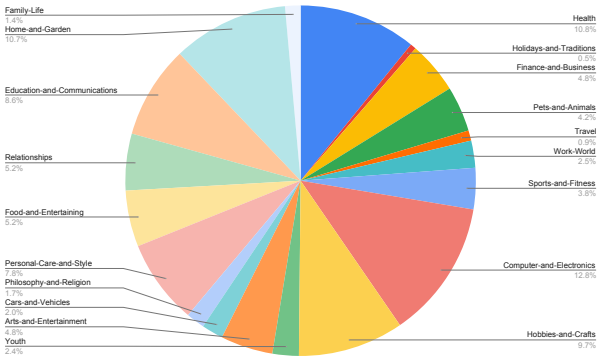

Figure 2

Statistic	Value
Total number of images	1,135,174
Total number of multimodal article	237,471
Total number of queries	26,221
Average length of text per article	402.5
Average number of images per article	4.2
Average length of queries	7.0

Table 1. Summary of the dataset statistics

4. Multimodal-Knowledge Retrieval Pipeline

We design multiple methods using the existing available retrieval models.

4.1. Retrieval Pipeline Using Single Source

Single source refers to only using either text or image as the targets.

BM25 BM25 [20] is a popular ranking function, designed to evaluate the relevance of articles to a specific search query. This method is particularly effective due to its incorporation of term frequency (TF) and inverse document frequency (IDF) while also accounting for document length normalization. Since, the articles in the dataset are multimodal, we disregard the images associated with the articles and retrieve the top articles based on the text using BM25.

Contriever Contriever [6], a dense retriever which is shown to be effective for zero-shot retrieval, is trained with contrastive learning objective in an unsupervised manner. Positive pairs are sourced from the same document using independent cropping, and negative pairs are sampled from different documents. Similar to the BM25, while using Contriever for retrieval of the articles, we only use the text from the articles and disregard the image information. We encode the text of the articles and the queries using the Contriever encoder and then calculate the cosine similarity to rank the relevant articles to any query. In our experiments, we use two model checkpoints provided by the authors, one is a pretrained model, and the other is a model that has been fine-tuned on the MSMARCO dataset.

CLIP Text Encoder (CLIP-T) CLIP [18] has become a popular choice for zero-shot multimodal retrieval due to its unique ability to jointly understand and represent both text and images within a shared embedding space. In our approach, we utilize the text encoder component of CLIP to encode both search queries and the textual content associated with articles. However, the maximum input token length of the text encoder in CLIP is 77 tokens, while the length of the text in the articles is usually longer. Hence, we segment the articles text into smaller chunks based on the $<img>$ identifiers present within the articles. Then, we utilize the CLIP text encoder to obtain the embeddings of each chunk and store all the chunk embeddings.

During inference, the CLIP text encoder is applied to get the embedding of the query, and cosine similarity is used to retrieve the top-100 chunks for each query. Since there might be multiple retrieved chunks from the same article, to get the final score for each article, we implement three pooling mechanisms - Sum, Mean and Maximum - to aggregate the scores of chunks from a distinct article. This aggregated score is then used to re-rank the articles, thus enabling a refined and more accurate retrieval process based on the pooled scoring mechanism.

An alternative implementation to this pooling based re-ranking involves averaging all the chunk embeddings from a distinct article to compute the article embedding. Retrieval is then executed based on the cosine similarity between these article embeddings and the query embedding. This method presents a streamlined approach, focusing on the overall representation of an article rather than individual text chunks, potentially offering a more holistic assessment of article relevance in relation to the query.

CLIP Image Encoder (CLIP-I) The methodologies discussed thus far only consider the textual information, leaving the visual components of the articles unused. In contrast to the previous approaches, in this method we exclusively consider the image content of the articles, aiming to evaluate whether images contain more relevant information that could potentially lead to enhanced retrieval performance. The procedure is almost identical to the previous method except that we compute the image embeddings of all images associated with the articles using the CLIP image encoder and store them instead of the chunk embeddings. During the inference, we retrieve the top-100 images corresponding to the given query based on the cosine similarity. Since there might be multiple retrieved images from the same ar-

ticle, we employ the aforementioned pooling mechanisms to calculate the aggregated article score. This aggregated score is then used to re-rank the articles. We also implement the alternative retrieval method, where we compute the article embedding based on the average of all the image embeddings corresponding to that article. Then we rank the articles based on the cosine similarity between these article embeddings and the query embedding.

BLIP Text Encoder (BLIP-T) Utilizing the advancements in multimodal learning, the BLIP model presents a significant leap in processing and understanding the interplay between text and visual content. BLIP has achieved state-of-the-art results on a wide range of vision-language tasks, such as image-text retrieval, image captioning, and visual question answering (VQA). In our methodology, we harness the text encoding capabilities of the BLIP model to efficiently encode search queries along with the textual content associated with the articles. We use a similar segmentation strategy as used for CLIP to break down the article text into chunks. For each segmented chunk, the BLIP text encoder is employed to generate embeddings, which are then stored for subsequent retrieval processes. During the retrieval phase, we aggregate of all chunk embeddings from an article to create a singular article embedding. Then we rank the articles based on the cosine similarity between query embedding and article embedding.

BLIP Image Encoder (BLIP-I) In this approach, we encode the images associated with the articles using image encoder of BLIP. During inference, we aggregate all the image embeddings associated with an article to calculate an article embedding. This consolidated embedding is then used to rank articles, offering a streamlined method that emphasizes the aggregate visual relevance of an article to the search query.

4.2. Retrieval Pipeline Using Multiple Sources

Multiple sources refer to using both image and textual content associated with the articles.

CLIP Image Encoder & Text Encoder (CLIP-I + CLIP-T) In this method, We leverage the multimodal processing capabilities of the CLIP and retrieve the top articles based on two strategies.

1. Score Aggregation: The textual query is encoded using the CLIP text encoder. Textual content of each article is segmented into multiple chunks, as previously outlined, and each chunk is encoded with the CLIP text encoder. These chunk embeddings are stored for further processing. Similarly, images associated with the articles are encoded using the CLIP image encoder, and their embeddings are stored. The next step involves retrieving the top-100 chunks and top-100 images based on the cosine similarity between the chunk/image embeddings and the query embedding. For each article, scores from multiple sources (either segments or images among the top-100) are combined using the three pooling mechanisms. These pooled scores are then used to compute an overall score for each article, leading to a re-ranking based on these aggregated scores.

2. Embedding Aggregation: Another way to utilize the multimodal information associated with the articles is to aggregate the chunk and the image embeddings corrsponding to an article to calculate the article embedding. We explore three techniques for this aggregation:

- All the chunk and image embeddings are considered independently and averaged to calculate the article's overall embedding. In this approach, equal weight is given to each chunk and image embedding.
- A combined image embedding is computed by averaging all the image embeddings corresponding to one article, while each chunk embedding is considered individually. We compute the article embedding by averaging the combined image embedding and chunk embeddings, putting more weight on the chunks and treating the images as auxiliary source.
- We compute one combined embedding for images and another for chunks by averaging within each modality. The final article embedding is then calculated by averaging these two modality-specific embeddings.

After extracting article embeddings in these three methods, articles are retrieved based on the cosine similarity between the query embedding and the calculated article embedding.

Combining BM25 & CLIP-I This retrieval method relies on the textual retrieval based on BM25 and image-based retrieval based on CLIP. As discussed in the BM25 retrieval, we retrieve top-100 articles based on the BM25 score for each query. Also, we retrieve the top-100 images based on the cosine similarity scores of the image embeddings extracted using CLIP image encoder and query embedding extracted using CLIP text encoder. After this retrieval stage we aggregate the modality specific scores using the three pooling mechanisms (Sum, Mean and Max). Finally, the articles are re-ranked based on the aggregated scores.

Combining Contriever & CLIP-I Similar to BM25 and CLIP, for this method we combine Contriever based textual retrieval and CLIP based image retrieval and re-rank the articles based on the aggregated scores.

ViLT Vision-and-Language Transformer (ViLT) [8] stands out as an incredibly effective method for multimodal

Method	Aggregation Strategy	Recall@5	Recall@10	NDCG@10
BM25	-	**42.4**	52.0	45.2
Contriever	-	26.2	35.4	29.5
Contriever-MSMARCO	-	41.7	**52.7**	**45.7**
CLIP-I	Sum of scores	5.7	8.4	6.3
	Mean of scores	4.5	6.8	5.2
	Max of scores	5.4	7.7	6.2
	Mean of embeddings	6.4	9.2	7.2
CLIP-T	Sum of scores	18.7	26.4	20.5
	Mean of scores	15.3	22.6	17.4
	Max of scores	19.2	25.8	21.1
	Mean of embeddings	24.8	32.9	27.3
BLIP-I	Mean of embeddings	0.1	0.3	1.2
BLIP-T	Mean of embeddings	20.6	28.0	22.8

Table 2. Comparison of the performances of the retrieval pipelines using single source on the MIRACLE dataset.

Method	Aggregation Strategy	Images	Chunks	Recall@5	Recall@10	NDCG@10
CLIP-I + CLIP-T	Sum of scores	-	-	18.6	26.4	20.3
	Mean of scores	-	-	11.4	16.8	13.2
	Max of scores	-	-	19.2	25.8	21.1
CLIP-I + CLIP-T	Mean of embeddings	all	all	9.8	13.7	11.4
		mean	all	25.2	33.5	27.5
		mean	mean	10.6	15.4	11.7
CLIP-I + BM25	Sum of scores	-	-	**42.7**	52.4	45.5
	Mean of scores	-	-	32.1	39.4	35.1
	Max of scores	-	-	42.4	52.0	45.2
CLIP-I + Contriever	Sum of scores	-	-	23.2	34.1	27.1
	Mean of scores	-	-	20.8	28.7	24.0
	Max of scores	-	-	26.2	35.4	29.5
CLIP-I + Contriever-MSMARCO	Sum of scores	-	-	26.3	39.8	31.7
	Mean of scores	-	-	30.7	38.6	34.6
	Max of scores	-	-	41.7	**52.7**	**45.7**
ViLT	Mean of embeddings	-	-	0.3	0.4	0.4
BLIP-M	Mean of embeddings	-	-	4.8	7.1	9.1
BLIP-M	Sum of scores	-	-	6.0	9.4	6.9
	Mean of scores	-	-	5.8	9.1	6.8
	Max of scores	-	-	6.9	10.0	7.9

Table 3. Comparison of the performances of the retrieval pipelines using multiple sources on the MIRACLE dataset.

information retrieval due to its unique capability of seamlessly merging text and visual information. We first create (chunk, image) pairs from the article by associating one image to one chunk of the article. Then we use ViLT to encode the pairs into multimodal embeddings. All of the pairwise multimodal embeddings from one article are averaged to generate the overall multimodal embedding of the article. On the contrary, the queries are all textual but ViLT encoding requires multimodal information. Hence, we incorporate a dummy blank images with the textual query to extract pseudo-multimodal embedding of the queries. Cosine similarity score is calculated between a query embedding and the article embeddings to retrieve top articles for any given query.

BLIP Multimodal Encoder(BLIP-M) We explore a multimodal setup that synergizes text and image data associated with the articles through the BLIP Multimodal Encoder. Similar to ViLT based approach, we start with segmentation of articles into discrete chunks, guided by the presence of image identifiers. The text chunks are paired with corresponding images, forming text-image pairs that encapsulate the multimodal chunks of an article. These pairs are then encoded using the BLIP Multimodal Encoder, which is specifically designed to understand and integrate the nuances of both textual and visual data within a unified embedding space. Retrieval is then performed based on the cosine similarity between the query embeddings and the multimodal embeddings of the text-image chunk pairs, identifying the top-ranking pairs for each query. To address the challenge of multiple pairs originating from the same article, we adopt pooling mechanisms—Sum, Mean, and Maximum—to aggregate the relevance scores of pairs from individual articles. This aggregated score forms the basis for a re-ranking process, allowing us to prioritize articles that are most relevant to the query across both their textual and visual components. Alternatively, to simplify the retrieval mechanism and focus on a more holistic representation of articles, we explore the consolidation of all text-image pair embeddings for a given article into a single, comprehensive article embedding. We then rank the articles based on the similarity of the query embeddings with the overall multimodal article embeddings.

5. Experimental Results

5.1. Evaluation Metrics

To assess the performance of retrieval models, we utilize the Recall@k metric, which measures the recall rate of the top-k retrieved items. In this study, we set k to 5, and 10. Mathematically, Recall@k is defined as follows:

$$\text{Recall@k} = \frac{1}{|Q|} \sum_{m=1}^{|Q|} \frac{\sum_{n=1}^{k} \text{rel}(m,n)}{\sum_{n} \text{rel}(m,n)} \tag{1}$$

where $|Q|$ is the total number of queries, and $\text{rel}(m,n)$ indicates the relevance score of the n-th article with the m-th query.

We also employ Normalized Discounted Cumulative Gain (NDCG). NDCG is particularly valuable in scenarios where the position of an item in the result list is significant. The computation of NDCG for a set of queries is as follows:

$$\text{NDCG} = \frac{1}{|Q|} \sum_{m=1}^{|Q|} \frac{\text{DCG}(m)}{\text{IDCG}(m)} \tag{2}$$

where $|Q|$ represents the total number of queries, and for each query q_m, the Discounted Cumulative Gain (DCG) is calculated by:

$$\text{DCG}(m) = \sum_{n=1}^{P} \frac{2^{\text{rel}(m,n)} - 1}{\log_2(1+n)} \tag{3}$$

Here, $\text{rel}(m,n)$ denotes the relevance score of the n-th item in the result list for query q_m, and P is the number of positions to consider.

Method	MIRACLE (ours)	EDIS	WebQA
BM25	42.4	18.0	-
CLIP	6.4	36.0	32.1

Table 4. Compare BM25 and CLIP baselines on our datasets and others using metric Recall@5. CLIP achieves the worse performance indicates its limited zero-shot performance on image style-shifting datasets.

5.2. Effectiveness of Popular Retrieval Methods on MIRACLE

Table 2, presents a quantitative comparison among various retrieval pipelines that use either text or image to retrieve the top articles. Notably, the BM25 method outperforms other baselines achieving Recall@5 of 42.4, Recall@10 of 52.0 and NDCG@10 of 45.2. While Contriever performs slightly lower than BM25, it still demonstrates reasonable recall and NDCG. On the other hand, CLIP and BLIP based methods, including CLIP-I, CLIP-T, BLIP-I and BLIP-T, exhibit lower scores than BM25 and Contriever. Particularly, CLIP-I and BLIP-I performs significantly worse than BM25, Contriever, CLIP-T and BLIP-T, indicating that relying solely on image-based retrieval on this task doesn't performs worse than text-based retrievals. One potential reason for this is that, the images in the articles are not always directly relevant to the query rather they complement the chunks of texts in the articles. In 3, we present two representative cases from the dataset. In the first example, the query talks about speaker and the images do not directly contain speaker but some tools that can be used to build one. In the second example, we observe a few generic image that do not correspond with the query without additional context. Additionally, we found that using embedding aggregation method performs significantly better than all three score aggregation methods for CLIP-based retrievals.

Table 3 presents results for retrieval pipelines that leverage both textual and visual information. ViLT demonstrates very poor recall and NDCG scores compared to other methods. We observe that a particular set of articles are always retrieved as top articles for all the queries which results in a lower score. Since we have used dummy blank image to make the queries multimodal, the retrieval method always retrieves articles with no or less number of images

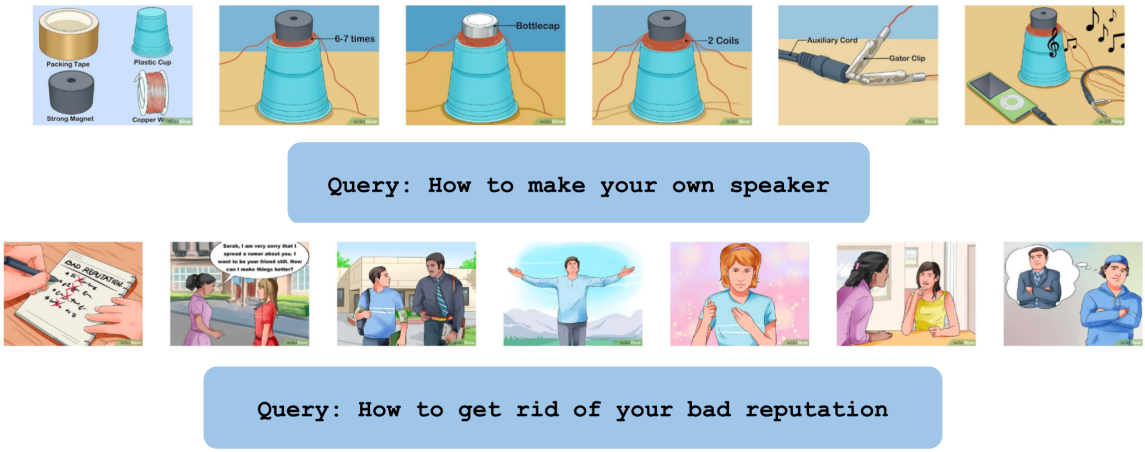

Figure 3. The images associated with articles do not always contain entities corresponding to the query.

regardless of the query. When we explore multimodal retrieval by combining CLIP-I and CLIP-T, intriguing results emerge. Sum and mean score-aggregation based retrieval methods perform worse compared to corresponding score-aggregation based CLIP-T alone and max score-aggregation retrieval performance is same for CLIP-T alone and combination of CLIP-I and CLIP-T. The reason behind this is that the retrieval score for top chunks are always higher than the retrieval scores of top images. Hence the maximum score for any article always comes from a chunk. In contrast, when employing embedding-based aggregation (averaging combined image embeddings with all chunk embeddings), we observe a performance boost in all three metrics compared to embedding-aggregation based CLIP-T retrieval. This indicates that images indeed contain valuable complementary information to enhance performance, but the method of utilization and weighting is crucial.

When we combine CLIP-I with BM25, we see a performance boost with a summation-based scoring aggregation strategy, while other aggregation methods result in decreased performance. This also suggests that image information can be useful when utilized appropriately. However, when combining CLIP-I with Contriever, we do not observe any performance gain. All the BLIP-M retrieval variants perform worse than CLIP variants.

In summary, our experiments indicate that images contain valuable supplementary information that can enhance retrieval, especially in multi-modal contexts, compared to text-only retrieval methods. The key lies in the proper utilization of images, with embedding-based aggregation proving to be particularly effective in improving performance.

5.3. Comparing Baselines on MIRACLE with other datasets

Since EDIS is an entity-centric benchmark where the entity is not mention in the question, it is not surprising that BM25 performs poorly on such kind of information seeking settings. On the other hand, BM25 can be usually good when the entity is mentioned in the question [13, 22]. This showcase that our dataset is a complementary with the existing benchmark. When using CLIP, the performance on our dataset is much worse compared to the other two datasets. The possible reason behind this discrepancy is that the style of the images in our dataset is very different compared to the pre-training datasets of the CLIP models, while the other two datsets images are from the same domain as the pre-training domain. This indicates that while many work have showcased that the CLIP generalize well on many tasks, such great performance only hold in the same style of images. Our experiments show that CLIP is not sufficient when the style of image is different.

6. Conclusion

The paper presents MIRACLE, a novel benchmark for multimodal image-text retrieval and analysis, focusing on long-form evaluation. This benchmark is unique in its integration of multiple images within long textual narratives, posing significant challenges for retrieval systems. The study evaluates zero-shot retrieval performance of the state-of-the-art models using MIRACLE, revealing the need for more advanced models and finetuning tailored to the demands of long-context and domain shifting. Findings underscore MIRACLE's potential to drive progress in the field by pushing the boundaries of current multimodal retrieval systems.

Limitations

The study's methodology and dataset, primarily based on the MIRACLE benchmark from a single platform and English-language content, establish a solid base for multimodal retrieval. However, they also indicate the potential benefits of including diverse languages and platforms for broader applicability. The zero-shot approach used provides a baseline for model performance, highlighting the untapped potential of methodologies like fine-tuning for enhanced effectiveness. These insights guide future research towards enriching the field of multimodal retrieval in more varied and complex scenarios.

References

[1] Anonymous. Chatsearch: a dataset and a generative retrieval model for general conversational image retrieval, 2023. 3

[2] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16495–16504, 2022. 1, 2

[3] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*, 2023. 3

[4] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5067–5077, 2022. 1, 2, 3

[5] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. *arXiv preprint arXiv:2302.11154*, 2023. 3

[6] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2022. 1, 4

[7] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 3

[8] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 5

[9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2, 3

[10] Siqi Liu, Weixi Feng, Wenhu Chen, and William Yang Wang. Edis: Entity-driven image search over multimodal web content. *arXiv preprint arXiv:2305.13631*, 2023. 2, 3

[11] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021. 2

[12] Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6417–6431, 2021. 2, 3

[13] Man Luo, Arindam Mitra, Tejas Gokhale, and Chitta Baral. Improving biomedical information retrieval with neural retrievers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11038–11046, 2022. 8

[14] Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. End-to-end knowledge retrieval with multi-modal queries. *arXiv preprint arXiv:2306.00424*, 2023. 1, 2, 3

[15] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121, 2021. 2

[16] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, 2022. 3

[17] Mihir Parmar, Swaroop Mishra, Mirali Purohit, Man Luo, Murad Mohammad, and Chitta Baral. Inboxbart: Get instructions into biomedical multi-task learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 112–128, 2022. 3

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sas-

try, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 4

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[20] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3 (4):333–389, 2009. 1, 4

[21] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008. 1

[22] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, 2021. 8

[23] Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. MIMOQA: Multimodal input multimodal output question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5317–5332, Online, 2021. Association for Computational Linguistics. 1

[24] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, 2022. 3

[25] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*, 2023. 3

[26] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. 1, 2

[27] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*, 2022. 3