Any Information Is Just Worth One Single Screenshot: Unifying Search With Visualized Information Retrieval

Zheng Liu[1*], **Ze Liu**[2,4*], **Zhengyang Liang**[3,4], **Junjie Zhou**[3,4], **Shitao Xiao**[4]
Chao Gao[4,5], **Chen Jason Zhang**[1], **Defu Lian**[2†]
[1]The Hong Kong Polytechnic University, [2]University of Science and Technology of China
[3]Beijing University of Posts and Telecommunications, [4]Beijing Academy of Artificial Intelligence,
[5]The Hong Kong University of Science and Technology
zhengliu1026@gmail.com, lz123@mail.ustc.edu.cn, liandefu@ustc.edu.cn

Abstract

With the popularity of multimodal techniques, it receives growing interests to acquire useful information in visual forms. In this work, we formulate an emerging IR paradigm called *Visualized Information Retrieval*, or **Vis-IR**, where multimodal information, such as texts, images, tables and charts, is jointly represented by a unified visual format called **Screenshots**, for various retrieval applications. We further make three key contributions for Vis-IR. First, we create **VIRA** (Vis-IR Aggregation), a large-scale dataset comprising a vast collection of screenshots from diverse sources, carefully curated into captioned and question-answer formats. Second, we develop **UniSE** (Universal Screenshot Embeddings), a family of retrieval models that enable screenshots to query or be queried across arbitrary data modalities. Finally, we construct **MVRB** (Massive Visualized IR Benchmark), a comprehensive benchmark covering a variety of task forms and application scenarios. Through extensive evaluations on MVRB, we highlight the deficiency from existing multimodal retrievers and the substantial improvements made by UniSE. Our data, model and benchmark have been made publicly available[1], which lays a solid foundation for this emerging field.

1 Introduction

Information retrieval has experienced tremendous progresses in recent years, driven by breakthroughs in foundation models. The growing capacity of language models has enabled precise and generalized text retrieval (Xiao et al., 2024; Wang et al., 2022, 2023), while the development in vision-language models has extended information retrieval to a broader range of data modalities (Wei et al., 2024; Zhang et al., 2024a; Zhou et al., 2024b). With

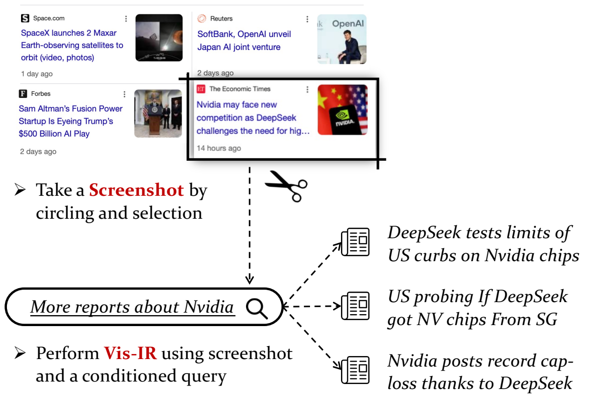


Figure 1: A use case of Vis-IR. Users take a screenshot of their interested news by circling and selection, and search for relevant news reports associated with "Nvidia" based on a query conditioned on the screenshot.

the recent popularity of multimodal techniques (Li et al., 2023; Liu et al., 2023; Team et al., 2023), there is growing interest in *uniformly representing multimodal data in visual forms and leveraging these visual representations for diverse information retrieval tasks* (shown in Figure 1) (Faysse et al., 2024; Ma et al., 2024; Zhang et al., 2024b). For example, researchers can circle and select a part of a paper, which may include texts, figures, equations, and figures, and search for relevant literature to address a specific question about the selected content. Similarly, people take a screenshot of an advertisement on their cellphones and retrieve descriptions of the corresponding product[2]. These innovative paradigms will greatly enhance the flexibility of information retrieval in real-world scenarios, leading to a unified paradigm for search engines. In this paper, we formally define these problems as *Visualized Information Retrieval*, or **Vis-IR** for brevity. We also define the visual representation for a mixture of multimodal data as a **screenshot**, which is treated as a unified entity in the retrieval process.

Despite that common multimodal retrievers can be applied for Vis-IR in a zero-shot manner, and

*Co-first authors of this paper.
†Corresponding author of this paper.
[1]https://huggingface.co/BAAI/BGE-VL-Screenshot
[2]https://blog.google/products/search/google-circle-to-search-android/

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19238–19261

there are preliminary works focused on improving representations of specific multimodal data, like screenshots of Wiki webpages (Ma et al., 2024), several fundamental challenges still persist for this emerging problem. Particularly, there are no tailored models which can offer unified, high-quality support for various Vis-IR applications. Besides, there is a lack of comprehensive benchmarks to evaluate a retriever's performance on Vis-IR tasks. Finally, no specialized datasets are curated and published for training competent Vis-IR models.

To address the above challenges, we present the following three resources for Vis-IR.

First, we create a large-scale dataset, namely **VIRA** (Vis-IR, Aggregation), which comprises a vast collection of 13 million screenshot data from diverse sources, like webpages from news websites, shopping platforms, and Wikipedia, homepages from GitHub, research papers from ArXiv, general PDF documents, and various forms of charts. Each screenshot is assigned with a *fine-grained caption* by either extracting from meta data or annotating with sophisticated OCR tools. We further produce two types of *question-answer data* for the well-captioned screenshots: 1) q2s tuples, which comprises a screenshot and a query about the screenshot, 2) sq2s triples, which contain source screenshots, conditioned queries about the source, and target screenshots. Altogether, over 20 million data samples are included by VIRA.

Second, we develop a family of embedding models, called **UniSE** (Universal Screenshot Embeddings), built on top of our created dataset. These models enable screenshots to either serve as queries or be queried by various data formats. For example, using a screenshot to search for an image or a document, or retrieving a screenshot based on a textual question. We adopt two alternative model architectures for UniSE: one based on CLIP (Radford et al., 2021), which is more efficient, the other one leveraging MLLMs (Liu et al., 2023), which is more expressive. This allows users to flexibly choose the model that best suits their own needs.

Third, we construct **MVRB** (Massive Visualized IR Benchmark) to evaluate multimodal retrievers' performance on general Vis-IR tasks. The benchmark includes various task types, such as screenshot-based multimodal retrieval (screenshot to anything, anything to screenshot) and screenshot-conditioned retrieval (e.g., searching for documents using queries conditioned on screenshots). It also covers a variety of important domains, including news, products, papers, and charts. We conduct a comprehensive experimental study using MVRB, which highlights significant deficiencies in existing methods and uncovers key factors influencing the development of Vis-IR retrievers. The results also demonstrate the effectiveness of UniSE in both in-domain and out-of-domain applications.

We've publicly released all resources at http://huggingface.co/BAAI/BGE-VL-Screenshot to facilitate future advancements in this critical field.

2 Dataset: VIRA

In this section, we present the VIRA (Vis-IR Aggregation) dataset, which offers diverse and large-scale training data for developing Vis-IR models. While creating the dataset, we emphasize two main factors: 1. **comprehensiveness** of screenshots, 2. **utility** of annotation. The creation's outline is introduced as follows. Due to the space limitation, the entire details are provided in Appendix A.

2.1 Screenshot collection

We perform massive collection of screenshot data spanning **seven categories**, which include *News*: from popular websites like BBC, CNN, and Fox, *Products*: from Amazon, *Research Papers*: from ArXiv, *General Documents*: from PDF Association, *Charts*: from ArXiv, *Common Knowledge*: from Wikipedia; *Project Homepages*: from GitHub. The data is evenly distributed across these categories, leading to **13 million screenshots** in total. The collected data encompasses a rich variety of formats, showing diverse combinations of text, figures, and structures presented in various layouts. These features substantially enhance the generalization for models trained on corresponding data.

2.2 Screenshot Annotation

We perform two types of annotations for the collected data. First, we assign each screenshot with a fine-grained **caption**, which paraphrases its detailed semantic. The caption is produced by in alternative ways. For those associated with complete meta data, e.g., project homepage from Github, we apply proper data extraction pipelines to acquire the captions. While for other free-form screenshots, we obtain captions based on OCR methods. Second, we create **question-answering** annotations for the well-captioned screenshots, which closely align with typical retrieval tasks. Specifically, we consider the following forms of question-answering

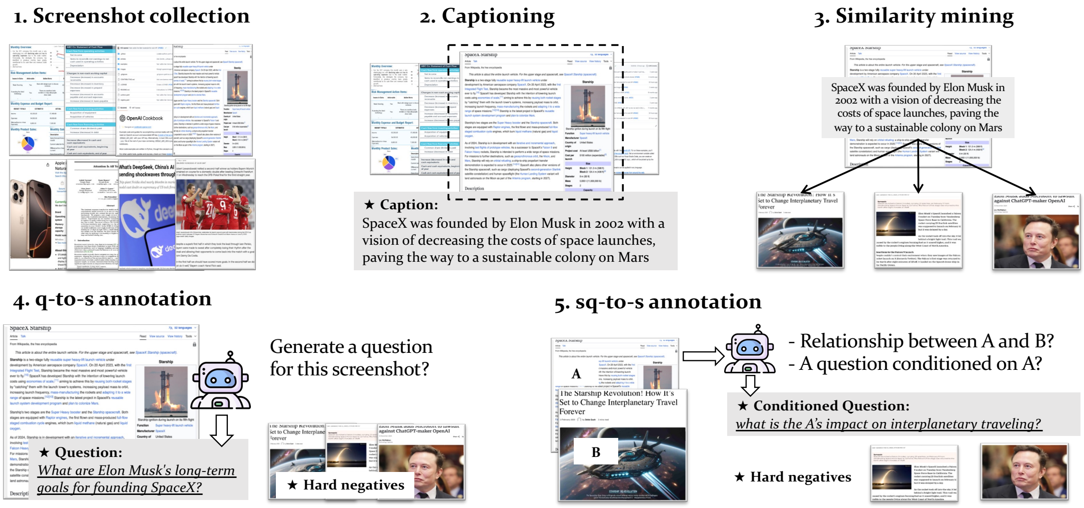


Figure 2: Creation process of VIRA dataset, including 1) comprehensive screenshot collection from various sources, 2) fine-grained screenshot captioning, 3) similar screenshots mining, 4) q2s annotation, and 5) sq2s annotation.

data. 1. **q2s tuples**, denoted as (q, s). For a screenshot s, we prompt a large language model (LLM) to generate a question q based on the screenshot's caption. For example, given a screenshot with the title "*Nvidia posts record cap-loss due to DeepSeek*", a question like "*What's the impact of DeepSeek on Nvidia?*" would be generated. 2. **sq2s triplets**, denoted as (s1, q, s2). Following MegaPairs methodology (Zhou et al., 2024a), for a pair of relevant screenshots (s1, s2) mined from the corpus, we prompt the LLM to analyze their relationship based on their captions and generate a relational question. For example, if s1 is a breaking news screenshot and s2 is a subsequent news report, a question like "*What's the follow-up news to it?*" is generated. Finally, we create **13 millions** screenshot-caption samples and **7 millions** question-answering samples from the above operations.

We supplement the above data with **hard negatives**, which are crucial for training retrieval models. For each $q2s$ tuple, we introduce hard negatives based on either textual or visual similarity (using off-the-shelf embedding models). For a $q2s$ tuple, a screenshot s' is selected if it meets any of these conditions: *1)* s' has a similar caption with s, *2)* s' is visually similar with s, *3)* s' is textually similar with q. While for a s_1q2s_2 triplet, we use s' as a hard negative, if it enjoys a strong textual or visual similarity with the retrieval target s_2.

3 Model: UniSE

We develop retrieval models based on VIRA dataset called Universal Screenshot Embeddings (UniSE), providing unified supports for general Vis-IR tasks.

3.1 Embedding

We adopt two structures for UniSE, allowing users to make flexible selection based on their own needs.

UniSE-CLIP. The first type adopts the *CLIP architecture* (Radford et al., 2021), which is more time-efficient. In this form, the screenshot s and text input t are encoded by the visual and textual encoder of CLIP, respectively: $\boldsymbol{e}_s \leftarrow \phi_v(s)$, $\boldsymbol{e}_t \leftarrow \phi_t(t)$. For a composed input of a screenshot s and conditioned textual query q, the joint representation is computed by linear combining the screenshot and query's embedding: $\boldsymbol{e}_{s,q} \leftarrow \boldsymbol{e}_s + \boldsymbol{e}_q$. Specifically, we utilize OpenAI's CLIP-Large[3], leveraging both its architecture and pre-trained weights as the foundation for UniSE-CLIP.

UniSE-MLLM. The second type is back-ended by *MLLMs (multimodal large language models)* (Liu et al., 2023; Wang et al., 2024), which is more expressive, especially in representing composed inputs. Without losing generality, a composed query is presented by the following template:

[Task]: \$task, [Query]: \$s-tok, \$q-tok, [EOS]

In this place, \$task indicates the task type, \$s-tok stands for the visual tokens from screenshot, while \$q-tok represents the textual tokens from query. The output embedding from the special token [EOS] is used to represent the composed input. For UniSE-MLLM, we adopt Qwen2-VL-2B[4] and initialize it with its pre-trained parameters.

[3]https://huggingface.co/openai/clip-vit-large-patch14
[4]https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct

3.2 Training

We propose a two-stage training workflow based on the composition of VIRA dataset. First, we perform pre-training using screenshots and their captions, denoted as $\{(s_i, c_i)\}_N$. In this stage, the model aims to capture the fine-grained semantic about the screenshot by learning to discriminate its relevant caption from irrelevant ones. Thus, we employ a bidirectional contrastive learning loss, ensuring that the model can correctly match screenshots to their captions and captions to their screenshots:

$$\mathcal{L}_{s_1} = \mathcal{L}_{con}(\boldsymbol{e}_s, \boldsymbol{e}_c) + \mathcal{L}_{con}(\boldsymbol{e}_c, \boldsymbol{e}_s) \tag{1}$$

where $\boldsymbol{e}_s$ and $\boldsymbol{e}_c$ represent the embeddings of screenshots and captions, respectively. The function $\mathcal{L}_{con}(\boldsymbol{u}, \boldsymbol{v})$ is formulated as:

$$\mathcal{L}_{con}(\boldsymbol{u}, \boldsymbol{v}) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log \frac{\exp(\boldsymbol{u}_i^T \boldsymbol{v}_i/\tau)}{\sum_{j\in\mathcal{B}} \exp(\boldsymbol{u}_i^T \boldsymbol{v}_j/\tau)} \tag{2}$$

where $\mathcal{B}$ represents the set of in-batch query samples, and τ is the temperature parameter that controls the strength of penalties on negative samples.

We continue to fine-tune the pre-trained model based on the question-answering data, denoted as $\{(q_i, s_i)\}_N$ (q_i can either be a textual query for a q2s tuple, or a combination of screenshot and its conditioned query for a sq2s triplet.) With this operation, the model is strengthened in handling retrieval-related tasks. In this stage, the following loss function is minimized:

$$\mathcal{L}_{s_2} = \mathcal{L}_{con}(\boldsymbol{e}_q, \boldsymbol{e}_s) \tag{3}$$

where $\boldsymbol{e}_q$ and $\boldsymbol{e}_s$ represent the embeddings of queries and screenshots, respectively. Additional training details and the hyper-parameter settings are provided in Appendix B.2.

4 Benchmark: MVRB

We create the Massive Vis-IR Benchmark (MVRB), which comprehensively evaluate a retriever's capability in handling general Vis-IR tasks. The outline of MVRB and its creation process is presented below, with more details provided in Appendix C.

4.1 Evaluation Tasks

Following the practice of popular retrieval benchmarks, such as MTEB (Muennighoff et al., 2022) and MMEB (Jiang et al., 2024b), we introduce four task categories, each encompassing multiple tasks relevant to critical application scenarios.

4.1.1 Screenshot Retrieval

Screenshot Retrieval (SR) consists of evaluation samples, each comprising a textual query q and its relevant screenshot s: (q, s). The retrieval model needs to precisely retrieve the relevant screenshot for a testing query from a given corpus S. Each evaluation sample is created in two steps: 1) sample a screenshot s, 2) prompt the LLM to generate a search query based on the caption of screenshot. We consider seven tasks under this category, including product retrieval, paper retrieval, repo retrieval, news retrieval, chart retrieval, document retrieval, and slide retrieval.

4.1.2 Composed Screenshot Retrieval

Composed Screenshot Retrieval (CSR) is made up of sq2s triplets. Given a screenshot $s1$ and a query q conditioned on $s1$, the retrieval model needs to retrieve the relevant screenshot $s2$ from the corpus S. We define four tasks for this category, including product discovery, news-to-Wiki, knowledge relation, and Wiki-to-product. All tasks in this category are created by human annotators. For each task, annotators are instructed to identify relevant screenshot pairs and write queries to retrieve s_2 based on s_1. Further details on task definitions and the annotation process are provided in Appendix C.2.

4.1.3 Screenshot Question Answering

Screenshot Question Answering (SQA) comprises sq2a triplets. Given a screenshot s and a question q conditioned on s, the retrieval model needs to retrieve the correct answer a from a candidate corpus A. Each evaluation sample is created in three steps: 1) sample a screenshot s, 2) prompt the MLLM to generate a question q, 3) prompt the MLLM to generate the answer a for q based on s. The following tasks are included in this category: product-QA, news-QA, Wiki-QA, paper-QA, repo-QA.

4.1.4 Open-Vocab Classification

Open-Vocab Classification (OVC) is performed using evaluation samples of screenshots and their textual class labels. Given a screenshot s and the label class C, the retrieval model needs to discriminate the correct label c from C based on the embedding similarity. We include the following tasks in this category: product classification, news-topic classification, academic-field classification, knowledge classification. For each task, we employ human labelers to create the label class and assign each screenshot with its correct label.

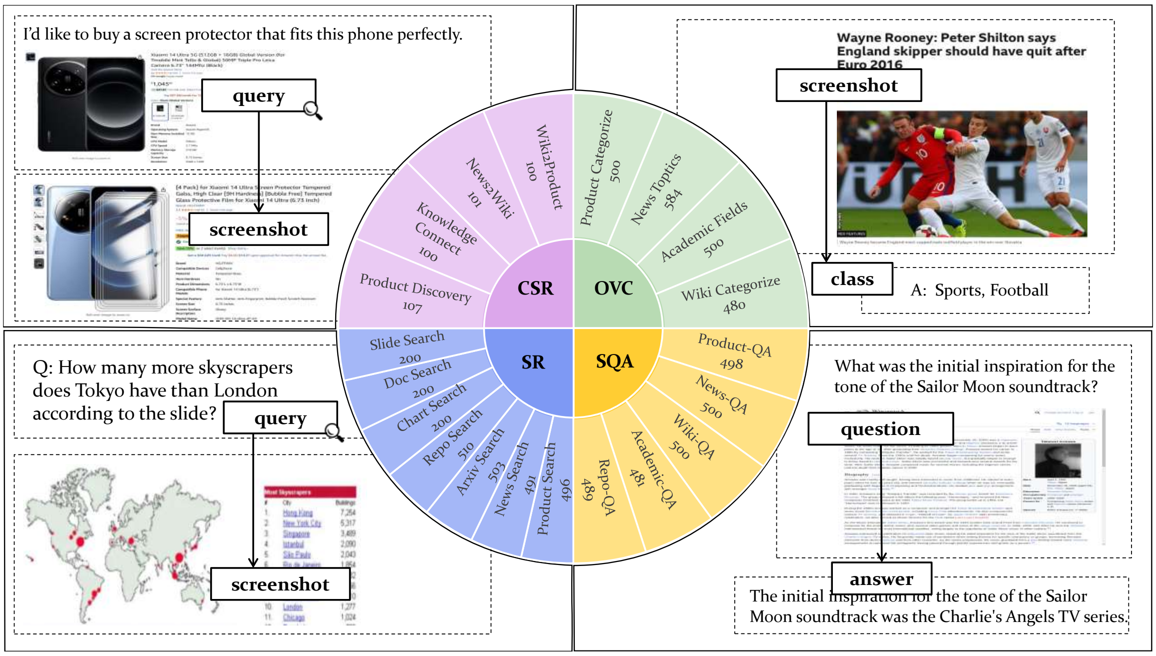


Figure 3: MVRB benchmark. There are four task categories: screenshot retrieval, composed screenshot retrieval, screenshot question answering, and open-vocab classification. Each category covers multiple concrete task scenarios.

4.2 Optimizations

We conduct the following operations to optimize the quality and usability of the benchmark.

We first introduce **quality-control** while creating the evaluation samples. Our quality-control framework comprises two main operations: **automatic assessment** and **human verification**. We employ multiple MLLMs to constitute a quality-control committee, which make assessment based on three criteria. 1) Clarify, whether the query conveys a concrete information need, 2) Reasonability, whether the query is appropriate in practice, 3) Correctness, whether the query can be addressed by the retrieval target (a screenshot or an answer). Each evaluation sample is independently reviewed by the MLLMs, and if it fails to meet any criterion, it is removed. The remaining samples are then verified by human labelers using the same principles. An evaluation sample is successfully created only if it passes both quality-control stages.

We then **construct the corpus** for each retrieval task. On the one hand, a corpus needs to maintain sufficient difficulty to effectively distinguish the capabilities of different retrievers. On the other hand, its scale should be properly controlled to prevent excessive evaluation costs. Considering these factors, the retrieval corpus is created in two steps. First, a moderate number of retrieval candidates (around 5,000) are randomly sampled for each task[5]. Second, a group of hard negative samples are introduced for each sample. For (composed) screenshot retrieval, hard negatives are selected based on their similarity to the query. For screenshot question answering, hard negatives are generated by modifying the correct answer using an LLM (see Appendix C for details). The hard negatives also undergo quality-control before being added to the corpus.

5 Experiments

Our experiments focus on three main perspectives. **1**. A comprehensive study for existing methods' abilities in Vis-IR. **2**. The value brought by VIRA dataset. **3**. The improvement achieved by UniSE.

5.1 Baselines

Our baseline methods fall into three main categories. (1) *OCR + Text Retrievers*, which leverage advanced OCR tools[6] to extract text from screenshots, followed by textual retrieval. (2) *General Multimodal Retrievers*, which are vision-language embedding models fine-tuned on diverse multimodal retrieval tasks, such as image retrieval, composed image retrieval, and visual question answering. In our experiments, these models process screenshots in a zero-shot manner. (3) *Screenshot Document Retrievers*. These models are built upon MLLMs architectures and fine-tuned by screenshot-based document retrieval data. Although these

[5] This allows users to complete their evaluation in a few hours using one A800-80G GPU, e.g., 3.5h for UniSE-MLLM.

[6] https://github.com/PaddlePaddle/PaddleOCR.

Models	Backbone	#params	Per Meta-Task Score				Overall
			SR	CSR	SQA	OVC	
number of datasets	–	–	7	4	5	4	20
OCR + Text Retrievers							
BM25 (Robertson et al., 2009)	–	–	40.71	28.51	38.46	6.23	30.81
DPR (Karpukhin et al., 2020)	BERT-Base	109M	24.98	18.73	27.35	22.08	23.74
BGE (Xiao et al., 2024)	BERT-Base	109M	45.67	36.73	36.29	47.91	41.99
E5-Mistral (Wang et al., 2023)	Mistral-7B	7.11B	38.68	<u>51.44</u>	43.00	45.08	45.51
General Multimodal Retrievers							
VISTA (Zhou et al., 2024b)	BERT-Base	196M	5.21	11.29	25.78	16.61	13.85
Uni-IR (Wei et al., 2024)	CLIP-Large	428M	12.35	35.92	29.68	20.06	19.63
CLIP (Radford et al., 2021)	CLIP-Large	428M	18.89	25.39	23.90	30.40	23.75
SigLIP (Zhai et al., 2023)	SOViT-400m	878M	38.33	34.48	19.60	40.64	33.34
E5-V (Jiang et al., 2024a)	LLaVA-1.6	8.35B	34.11	26.59	5.23	32.85	25.13
VLM2Vec (Jiang et al., 2024b)	Phi-3.5-V	4.15B	15.93	48.05	**49.42**	23.24	32.19
MM-Embed (Lin et al., 2024)	LLaVA-1.6	7.57B	25.86	40.93	42.83	32.67	34.48
Screenshot Document Retrievers							
ColPali (Faysse et al., 2024)	Paligemma	2.92B	<u>61.73</u>	35.00	35.32	31.04	43.64
DSE (Ma et al., 2024)	Phi-3-V	4.15B	61.54	37.78	39.24	31.51	45.21
GME (Zhang et al., 2024b)	Qwen2-VL	2.21B	61.62	37.68	37.78	<u>47.98</u>	<u>48.14</u>
UniSE-CLIP (ours)	CLIP-Large	428M	35.95	43.38	28.13	40.62	36.41
UniSE-MLLM (ours)	Qwen2-VL	2.21B	**69.63**	**54.49**	<u>43.20</u>	**48.26**	**55.72**

Table 1: Overall performance on MVRB (measure by Recall@1). The aggregation result for each task category and the average score is reported for each method. Top scores are marked in **bold** while runner-ups are <u>underlined</u>.

models are trained on datasets involving screenshots, their effectiveness is severely constrained by limited data and task diversity, making them insufficient for general Vis-IR tasks. We adopt multiple popular methods for each category to ensure a thorough comparison. Detailed specifications about the baseline methods are provided in Appendix D.

5.2 Main Results

The overall performance of baseline retrievers and our UniSE models are presented in Table 1. We report the aggregation result for each task category, leaving the detailed performance for each task presented in Appendix E. We have identified the following key observations from this table:

(1) Our UniSE models achieve the leading performance across all task categories on MVRB. Specifically, UniSE-MLLM surpasses the previous best screenshot document retriever, GME (Zhang et al., 2024b), by 7.6% in average score. Additionally, the UniSE-CLIP model, with only 428 million parameters, outperforms all baselines of no greater sizes while delivering comparable performance to MM-Embed (Lin et al., 2024), the strongest general multimodal retriever, which has 7.57 billion parameters. The superiority of UniSE models verifies the effectiveness of our VIRA dataset.

(2) The Vis-IR paradigm offers significant advantages over traditional text-based retrieval.. As shown in Table 1, even with a complex retrieval workflow powered by advanced OCR tools and powerful text retrievers (Xiao et al., 2024), traditional methods still fall significantly behind UniSE. This suboptimal performance is likely due to the loss of crucial layout and visual semantics, even though OCR can accurately extract textual information. This highlights the great potential of Vis-IR as a unified approach for document retrieval.

(3) Zero-shot application of general multimodal retrievers results in suboptimal performance for Vis-IR. While these methods can be directly applied to screenshot data, most of them perform worse than other baselines. For instance, the average score of MM-Embed and MLV2Vec falls significantly behind that of GME and DSE, despite using similar MLLM backbones and being fine-tuned on massive datasets. This highlights a key distinction between screenshots and common multimodal data, as the latter lacks a rich combination of data elements presented in visual forms.

(4) Existing screenshot document retrievers cannot comprehensively address Vis-IR tasks. Although baselines like ColPali, DSE, and GME achieve substantial improvements in average score,

Models	Domains					
	Prod.	News	Wiki	Paper	Repo	Others
#Datasets	4	3	3	3	2	5
BM25	28.65	28.23	21.36	39.72	45.67	34.65
DPR	17.28	33.64	25.96	21.47	33.08	27.07
BGE	38.08	48.44	34.77	41.11	51.81	47.66
VISTA	14.49	19.76	17.48	11.53	20.39	6.39
Uni-IR	19.34	24.26	23.62	8.87	18.54	21.56
CLIP	25.96	35.57	27.70	12.14	21.75	20.27
SIGLIP	41.46	38.73	30.19	21.09	36.10	31.76
E5-V	18.09	28.28	20.81	21.32	21.11	35.37
VLM2Vec	28.41	38.56	39.60	20.74	37.97	31.50
MM-Embed	31.72	40.33	36.84	23.52	42.29	35.22
ColPali	36.24	41.83	26.38	47.96	55.28	53.76
DSE	37.05	48.60	28.52	46.83	60.47	52.23
GME	34.73	56.53	37.17	50.08	62.64	53.46
UniSE-CLIP	37.92	43.37	36.10	26.45	38.28	36.44
Uni-MLLM	**49.33**	**64.30**	**44.50**	**54.12**	**70.17**	**57.61**

Table 2: Average performance (SR, CSR, SQA, OVC) on different domains. Recall@1 is the evaluation metric.

they still underperform other methods in CSR and SQA tasks. Additionally, these methods exhibit inconsistent performance across different domains, as shown in Table 2. In contrast, UniSE-MLLM maintains the leading performance across both task categories and application domains, which demonstrates its well-rounded Vis-IR capabilities.

5.3 Data Ablation

We perform extensive ablation studies to analyze the value from VIRA dataset. First, we examine the impact from captions and question-answering data, which are used for pre-training and fine-tuning, respectively. Next, we look into the roles of different question-answering data, including the s2q tuples and sq2s triplets. Finally, we assess the performance gain from using hard negatives.

5.3.1 Two Annotations

We first explore the impact from the two types of annotations within our VIRA dataset, as shown in Table 3. We compare three approaches: the first one relies solely on screenshot captions (first row), the second one leverages only question-answering data (second-row), the last one uses both captions and question-answering data (last row). To ensure a fair comparison, these methods are fine-tuned for the same number of training steps.

We derive the following key observations from the experiment results. First, UniSE demonstrates strong Vis-IR capabilities solely with pre-training on screenshot-caption data, already surpassing ColPaLI (Faysse et al., 2024) by 2.5% in overall score. While the alignment between screenshots and captions differs significantly from downstream tasks, it enables the model to capture fine-grained screenshot semantics, laying a solid foundation for further training. Second, question-answering data alone yields even stronger performance, surpassing the caption-only method by 7.8%. Third, the combined use of both captions and question-answering data provides additional improvements, achieving performance gains of 9.6% and 1.8% over methods using only a single type of annotation data.

5.3.2 Composite QA Data

We further investigate the benefits of using composite question-answering data, specifically q2s tuples and sq2q triplets, as shown in Table 3. To assess their individual contributions, we conduct two ablation studies: one using only q2s tuples (3rd row, w/o sq2q) and the other using only sq2q triplets (4th row, w/o q2s). The results indicate that both q2s tuples and sq2q triplets significantly enhance UniSE's overall performance, improving the average score by 6.8% and 8.6%, respectively, compared to the screenshot-caption pre-trained model. Besides, sq2q triplets provide an extra 1.8% gain over q2s tuples alone. Finally, the effects of both data types are complementary, as their joint usage leads to the highest performance (6th row, use all).

5.3.3 Domain Diversity

We examine the detailed impact of using diverse data from multiple domains. To this end, we introduce an ablation approach that includes question-answering data solely from Wikipedia while retaining diverse caption data (5th row, w/o. diversity). The experimental results indicate a clear advantage of using diverse training data, as the default method (6th row, use all) outperforms the ablation approach by 2.2% in overall performance. This underscores the value of domain diversity in providing uniform support across various Vis-IR tasks.

5.3.4 Hard Negatives

Finally, we explore the impact of incorporating hard negatives into the training data. In our experiment, we use three ablation methods, as shown in Table 4: 1) without hard negatives (1st row), 2) with only sq2s hard negatives (2nd row), and 3) with only s2q hard negatives (3rd row). Compared to the method without hard negatives, introducing either type of hard negatives leads to improved overall performance. Furthermore, combining both types of hard negatives results in the best performance.

Caption	Instructions			Overall Score
	q2s	sq2s	Diversity	
✓	✗	✗	–	46.14
✗	✓	✓	✓	53.95
✓	✓	✗	✓	52.94
✓	✗	✓	✓	<u>54.72</u>
✓	✓	✓	✗	53.53
✓	✓	✓	✓	**55.72**

Table 3: UniSE-MLLM's performance from different data, including 1. caption-only, 2. question-answer only, i.e., q2s and sq2s, 3. w/o. sq2s, 4. w/o. q2s, 5. w/o. diversity, using Wiki as the only source, 6. using all.

Hard Negatives		Overall Score	
q2s	sq2s	UniSE-MLLM	UniSE-CLIP
✗	✗	54.26	34.99
✗	✓	<u>54.68</u>	<u>36.19</u>
✓	✗	54.50	36.08
✓	✓	**55.72**	**36.41**

Table 4: Impact from different hard negative configurations. **q2s**: incorporate hard negatives for q2s tuples; **sq2s**: incorporate hard negatives for sq2s triplets.

While the quality of hard negatives could be further improved with more costly processing, the current results offer valuable support for enhancing Vis-IR models developed from the VIRA dataset.

6 Related Work

Neural Document Retrieval. Document retrieval is crucial for a wide range of problems, like search engines, open-domain question answering, and retrieval-augmented generation (Karpukhin et al., 2020; Lewis et al., 2020; Yuan et al., 2023). Traditionally, document retrieval primarily rely on text-based methods, with significant advancements propelled by pre-trained language and large language models. These models have facilitated the development of effective document retrieval systems (Ni et al., 2022; Wang et al., 2022; Xiao et al., 2024; Li et al., 2024a). However, text-based retrieval methods require prior document parsing and are incapable of handling unstructured information within documents, such as complex layouts, charts, and visual elements (Ma et al., 2024). This limitation results in information loss and highlights the need for new retrieval methods that go beyond text.

Multimodal Retrieval. The development of vision-language models (VLMs) (Radford et al., 2021; Jia et al., 2021; Liu et al., 2023) has significantly advanced the capabilities of general-purpose multimodal retrievers (Wei et al., 2024; Zhou et al., 2024b; Jiang et al., 2024b; Lin et al., 2024), enabling them to perform massive multimodal retrieval tasks, such as text-to-image retrieval (Chen et al., 2015), composed image retrieval (Wu et al., 2021; Baldrati et al., 2023), and visual grounding (Zhu et al., 2016), etc. However, these methods assume that multimodal documents are well-preprocessed into interleaved image-text sequences, which is unsuitable for the diverse and unstructured nature of many real-world documents.

Recent studies (Ma et al., 2024; Faysse et al., 2024; Zhang et al., 2024b) have explored the use of document screenshots as unified input representations, i.e., visualized information retrieval (Vis-IR). This approach eliminates the need for additional content extraction, preserves all information within the document, and shows potential for effectively handling the complexity of real-world documents. Despite these advancements, existing Vis-IR datasets remain limited to specific domains (e.g., Wiki-SS (Ma et al., 2024)) and primarily evaluate performance by adapting existing document visual question answering benchmarks to text-to-screenshot retrieval tasks (Tanaka et al., 2023; Mathew et al., 2021a, 2022; Zhu et al., 2022; Li et al., 2024b). This narrow scope restricts the generalizability and applicability of current methods. Therefore, there is a clear need for diverse datasets and comprehensive benchmarks to advance this field and enable more robust and scalable solutions.

7 Conclusion

In this work, we introduce Vis-IR, a novel paradigm that enables multimodal search in a highly unified manner. Our work makes three fundamental contributions to this emerging field: (1) the creation of VIRA, a large-scale, diverse, and well-annotated dataset for developing Vis-IR models; (2) the development of UniSE, a family of embedding models designed for general Vis-IR applications; and (3) the construction of MVRB, a comprehensive benchmark for evaluating Vis-IR performance. Our experiments uncover the limitations of existing methods in performing Vis-IR tasks and demonstrate the significant improvements brought by VIRA and UniSE. To drive further advancements, we will publicly release all resources and continue expanding in key directions, such as increasing data diversity and incorporating multilingual annotations.

Limitations

While this work makes substantial progress over existing methods, several limitations remain to be addressed in the future. First, incorporating multilingual annotations would enhance the development and application of Vis-IR across different cultures. Second, the diversity of data sources can be further improved. We plan to expand the dataset with more critical sources and explore the use of synthetic data. Third, we aim to train the model on both screenshot and general multimodal data, enabling it to serve broader applications.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. U24A20253). We appreciate the anonymous reviewers' constructive suggestions to this paper.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. 2023. Getting vit in shape: Scaling laws for compute-optimal model design. *Advances in Neural Information Processing Systems*, 36:16406–16425.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15338–15347.

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*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024a. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2024b. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chaofan Li, Zheng Liu, Shitao Xiao, Yingxia Shao, and Defu Lian. 2024a. Llama2vec: Unsupervised adaptation of large language models for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3490–3500.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024b. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2024. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021a. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021b. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9844–9855. Association for Computational Linguistics.

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 *International conference on machine learning*, pages 8748–8763. PMLR.

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

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXVII*, volume 15145 of *Lecture Notes in Computer Science*, pages 387–404. Springer.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. 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.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

Huaying Yuan, Zhicheng Dou, Yujia Zhou, Yu Guo, and Ji-Rong Wen. 2023. Vile: Block-aware visual enhanced document retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3104–3113.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. 2024a. Magiclens: Self-supervised image retrieval with open-ended instructions. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024b. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. 2024a. Megapairs: Massive data synthesis for universal multimodal retrieval. *arXiv preprint arXiv:2412.14475*.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024b. VISTA: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3185–3200, Bangkok, Thailand. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

Appendix

A Details of VIRA Dataset Construction

A.1 Data Collection

We collect massive screenshots spanning seven categories. Based on the acquisition method, these data primarily fall into two groups: those crawled by us and those curated from publicly available open-source datasets. For the crawled data, we use the automated platform `Playwright`[7] to scrape target websites and capture screenshots of the main content pages. To ensure the completeness of the captured images, we perform a verification process using the built-in function of `Playwright` and discard any incompletely rendered images. Our entire crawling process adheres to the `robots.txt` regulations of the respective websites. For the other curated, we have carefully curated the files to ensure they can be converted into complete screenshots, containing both visual and textual content.

Self-Crawled Datasets. The screenshots obtained through crawling fall into four categories: News, Products, Research Papers, and Project Homepages. The details of the source and processing steps for each category are as follows:

- **News**: We scrap news webpages from five mainstream news websites, including BBC, CNN, Fox, CGTN, and Global Times. For each news page, we use `Playwright` to capture full-page scrolling screenshots and extract text content from the corresponding HTML.

- **Products**: We primarily focus on product detail pages from the Amazon e-commerce website. Utilizing `Playwright`, we capture screenshots of these pages, which typically feature product images, titles, prices, categories, and other relevant details of the product. Additionally, we extract textual information from the HTML of each product webpage.

- **Research Papers**: We collect research papers from arXiv using the arXiv API[8] , spanning from January 2018 to November 2024. For each paper, we retain only the most recent version. The original format of these papers is PDF. We convert each page of every paper

[7] https://playwright.dev/python/

[8] https://info.arxiv.org/help/api/basics.html

into an image and extract the text from each page using the PyMuPDF[9] library.

- **Project Homepage**: We collect project homepages from GitHub repositories. Initially, we filter out repositories that contain a README.md file. Using Playwright, we capture screenshots of the README region from the homepages of selected repositories and extract the relevant text from corresponding HTML. Since some README files contain minimal information (e.g., only the repository name), we discard screenshots with a height of less than 300 pixels.

Curated Datasets. We curate open-source datasets to obtain screenshots for three additional categories: General Documents, Charts, and Common Knowledge. The details of the source and collection process for each category are as follows:

- **General Documents**: We utilize the publicly available PDFA[10] dataset, a document dataset filtered from SafeDocs. To construct our dataset, we first filter out documents in PDF format and convert each page into an image using PyMuPDF. The source dataset provides OCR data for each document, which we process by stitching together the extracted text from top-left to bottom-right, forming a coherent caption corresponding to each image.
- **Charts**: We utilize the publicly accessible dataset ArxivCap (Li et al., 2024b), which comprises image-caption pairs derived from 572,000 Arxiv papers. Each entry contains an image, like a table or figure, and its caption, initially provided as separate text-image pairs rather than a cohesive screenshot. Therefore, we render each image alongside its caption into a single, unified screenshot.
- **Common Knowledge**: We used the publicly available dataset Wiki-SS-Corpus (Ma et al., 2024). This dataset is derived from screenshots of Wikipedia entry webpages and includes caption data for the first 500 words of each page.

After collecting the data, we apply a filtering process to ensure content quality and appropriateness. Specifically, we perform keyword matching on the caption text to filter out NSFW content. Additionally, we remove low-quality entries based on aspect ratio and caption length, discarding screenshots with an aspect ratio exceeding 9 and captions with fewer than 100 characters.

[9] https://pypi.org/project/PyMuPDF/

[10] https://hf-mirror.com/datasets/pixparse/pdfa-eng-wds

A.2 Data Annotation

The annotation of VIRA data is categorized into two main types: caption and question-answering. Caption is the text within the collected screenshots, which is annotated and generated during the screenshot collection process, as detailed in Appendix A.1. As for the question-answering, we have developed two types of question-answering annotations: q2s tuples and sq2s triplets. The detailed process of creating these question-answering annotations is as follows:

q2s tuples. A q2s tuple consists of a question q and a corresponding screenshot s that can be used to answer q. Given that we have access to caption for each screenshot, we leverage a large language model (LLM) to generate question-answer pairs closely related to the screenshot. Our approach is similar to the methodology used in constructing Docmatix (Laurençon et al., 2024). To enhance annotation quality, we design domain-specific prompts tailored to different categories of screenshots, with the corresponding prompts illustrated in Figure 4. For the whole annotation process, we employ the open-source model Qwen2.5-72B-Instruction (Team, 2024), which requires approximately 2,078 A800 GPU hours in total.

To enhance the effectiveness of q2s tuples, we augment each tuple with hard negatives. This augmentation process leverages both the text embedding model BGE (Xiao et al., 2024) and the visual embedding model EVA-CLIP (Sun et al., 2023). Specifically, we first encode all captions to construct a corpus embedding and separately encode both the question and the caption of the target screenshot to obtain their respective embeddings. Based on these embeddings, we retrieve the top 15 and top 10 candidates from the corpus for the question and the target screenshot, respectively. To mitigate false negatives, we exclude the top-1 candidate for the question and the top-3 candidates for the target screenshot.

For screenshots in the domains of of News, Products, and Common Knowledge—where natural images frequently appear—we further employ EVA-CLIP to encode the target screenshot and retrieve the top-10 candidates from the corpus, excluding the top-2 to reduce potential false negatives. Finally, we merge all remaining candidate sets and randomly sample 8 candidates to serve as hard negatives for the q2s tuple.

sq2s triplets. An sq2s triplet consists of two relevant screenshots—a query screenshot and a target screenshot—along with a conditional query. The target screenshot is used to respond the query conditioned on the query screenshot. To construct an sq2s triplet, we first mine a pair of relevant screenshots from the corpus. Following a similar approach used for augmenting q2s tuples with hard negatives, we employ BGE and EVA-CLIP to retrieve the top-10 candidates and randomly select one as the relevant screenshot for each given screenshot. For each pair of relevant screenshots, we prompt LLM to analyze their relationship based on their captions and generate a relational query. To enhance annotation quality, we design domain-specific prompts tailored to different categories of screenshot pairs, with the corresponding prompts are illustrated in Figure 5 . We utilize the open-source model Qwen2.5-72B-Instruction, and the entire annotation process requires approximately 1002 A800 GPU hours.

Additionally, we augment each sq2s triplet with hard negatives. The conditional query, query screenshot, and target screenshot are used to retrieve candidates from the corpus, following the same methodology applied in augmenting q2s tuples with hard negatives. From the retrieved candidates, 8 candidates are randomly sampled to serve as hard negatives for each sq2s triplet.

A.3 Statistics of VIRA

The VIRA is an extensive dataset, comprising a total of 20 million data entries. Of these, approximately 12.96 million are caption data, and around 7.11 million are question-answering data, which includes 5.97 million q2s tuples and 1.14 million sq2s triplets. A detailed breakdown of the data for each type in different domains is provided in Table 5.

Type	Domain	Number
Caption	News	1.85M
	Products	1.24M
	Research Papers	2.39M
	Project Homepage	2.38M
	General Documents	1.82M
	Charts	2.01M
	Knowledge	1.27M
s2q	News	1.24M
	Products	787.5K
	Research Papers	997.4K
	Project Homepage	1.0M
	General Documents	1.09M
	Knowledge	850.2K
qs2s	News	248.0K
	Products	496.1K
	Research Papers	89.8K
	Project Homepage	69.7K
	General Documents	87.5K
	Knowledge	155.4K

Table 5: Detailed data counts for each domain in VIRA

B More Details of UniSE Models

B.1 Preprocessing Strategies for Screenshot Images

For the UniSE-CLIP model, all input screenshot images are resized to 224×224 during both training and evaluating, following the default settings of the original CLIP model. In contrast, the UniSE-MLLM model employs a **smart resize** strategy for both training and evaluation, which preserves the original aspect ratio of the screenshot images while maximizing their resolution to retain visual quality. Specifically, the maximum image size is set as $M \times 28 \times 28$, where M denotes the maximum number of image tokens. For an image with height H and width W exceeding this limit, its dimensions are adjusted to $H' = \lfloor \frac{H}{\beta} \rfloor \times 28$ and $W' = \lfloor \frac{W}{\beta} \rfloor \times 28$, where $\beta = \sqrt{\frac{W \times H}{M \times 28 \times 28}}$. In our UniSE-MLLM, M is set to 2500.

B.2 Training Details

For both UniSE-CLIP and UniSE-MLLM models, training is conducted in two stages: pre-training and instruction fine-tuning. In the pre-training stage, both models are trained on VIRA's screenshot-caption subset, which contains approximately 13 million screenshot-caption pairs. In the instruction fine-tuning stage, the pre-trained models are further refined using VIRA's question-answering subset, comprising approximately 6 million items. For all models and training stages, the initial learning rate is set to 5×10^{-6}, with a linear decay schedule applied during training. Other training details are shown below.

UniSE-CLIP Model. In the pre-training stage, the UniSE-CLIP model is trained with a batch size of 8192 for a single epoch. During the instruction fine-tuning stage, the model uses a batch size of 4096 for one epoch, where each query is paired with one positive screenshot and one hard negative sample. All model parameters are updated during both stages of training.

UniSE-MLLM Model. In the pre-training stage, the UniSE-MLLM model is trained with a batch size of 2048 for one epoch. During the instruction fine-tuning stage, the model uses a batch size of 1024 for one epoch, with each query paired with one hard negative sample. Unlike UniSE-CLIP, UniSE-MLLM employs LoRA (Hu et al., 2022) to fine-tune the language model component of Qwen, while all other layers remain frozen throughout the training process. The LoRA rank is set to 32.

C Details of MVRB Benchmark Creation

We construct our benchmark through machine annotation, human annotation, and filtering and reorganizing other publicly available datasets. Examples of tasks in our benchmark are presented in Figure 9 and Figure 10. The specific number of queries and the corpus for each subtask of our benchmark are listed in Table 6. Below, we provide detailed information on our construction process for each meta task.

C.1 Screenshot Retrieval

Screenshot Retrieval (SR) tasks are constructed through machine annotation and filtering/restructuring of publicly available datasets. Tasks constructed by machine annotation include *product retrieval*, *paper retrieval*, *repo retrieval*, and *news retrieval*. For these tasks, we use captions of screenshots to generate question-answer pairs using LLMs with the questions serving as queries. To ensure quality, we implement a two-stage quality control process involving **automatic assessment** and **human verification**. First, each sample is assessed by three MLLMs (LLaVA1.6-34B[11], Molmo-72B[12], Llama-3.2-90B-Vision-Instruct[13]) across three dimensions: 1) *Clarity* : whether the query expresses a concrete information need, e.g., filtering out vague queries like "*a sports car*", 2) *Reasonableness*: whether the query is realistic, e.g., filtering out "*a horse that can swim*"), and 3) *Correctness*: whether the query can be answered by the retrieved screenshot, e.g., filtering out "*what is the citation count of this paper*" since it cannot be answered solely from the screenshot, with the corresponding prompts detailed in Figure 6. Each evaluation sample is independently reviewed by the MLLMs, and any sample failing any criterion is discarded. The remaining samples are further verified by human labelers using the same principles. An evaluation sample is successfully created only if it passes both stages of quality control.

Tasks derived from filtering and restructuring publicly available datasets include *chart-retrieval*, *document-retrieval*, and *slide-retrieval*. We selected test sets from ChartVQA (Masry et al., 2022), DocVQA (Mathew et al., 2021b), and SlideVQA (Tanaka et al., 2023), using their questions as queries and corresponding images as target screenshots. To refine the evaluation dataset, we remove cases where a single query matches multiple images, such as in SlideVQA. We also filter out context-dependent questions, like "*What is the content of Section 3.4.5?*", ensuring that each query is tailored for retrieving a single screenshot.

To increase task difficulty, we introduce hard negatives based on the query and target screenshot, following a method similar to the hard negatives augmentation approach for q2s tuples and sq2s triplets (Appendix A.2). To prevent false negatives, we use MLLM filtering to ensure no relevant screenshots are mistakenly included in the hard negatives, with the corresponding prompt shown in Figure 7.

C.2 Composed Screenshot Retrieval

We construct these tasks through human annotation, encompassing four subtasks: *product discovery*, *news-to-Wiki*, *knowledge relation*, and *Wiki-to-product*. These tasks are not only challenging but also hold significant real-world relevance. Anno-

[11]https://huggingface.co/liuhaotian/llava-v1.6-34b
[12]https://huggingface.co/allenai/Molmo-72B-0924
[13]https://huggingface.co/meta-llama/Llama-3.2-90B-Vision-Instruct

tators are guided to gather query screenshots from appropriate platforms, generate realistic and high-quality queries, and acquire the matching target screenshots. To enhance both the complexity and the discriminative aspect of the task, each annotator also identifies 8∼10 similar screenshots as difficult negatives for each original screenshot. Below, we outline the construction process for each specific CSR task:

- **Product Discovery**: This task simulates a real-world shopping scenario where users search for a related product based on specific needs such as price, color, brand, or accessories. Annotators browse Amazon, search for products from a predefined category set, capture screenshots, and then generate queries based on the information in the screenshots. They then retrieve another product screenshot that satisfies the query.

- **News-to-Wiki**: This task mirrors a scenario where users read news articles and seek additional details on people, events, locations, or causes, available in Wikipedia entries. Annotators explore news platforms like BBC, Fox, and CNN, develop queries from the content, and find relevant Wikipedia entries, capturing screenshots that answer the query.

- **Knowledge Relation**: This task simulates a scenario where users browsing Wikipedia are interested in exploring related topics. When looking at one entry, they often seek additional information, which is available in another Wikipedia article. Annotators search Wikipedia, capture relevant screenshots, formulate queries based on the screenshot content, and find another entry screenshot that fulfills the query.

- **Wiki-to-Product**: This task captures the scenario where users reading Wikipedia articles about products may wish to purchase them. Annotators are given Wikipedia entries related to products and retrieve corresponding screenshots from Amazon and Wikipedia to create query-target pairs.

C.3 Screenshot Question Answering

Screenshot Question Answering (SQA) tasks are constructed through machine annotation and include *product-QA*, *news-QA*, *Wiki-QA*, *paper-QA*, and *repo-QA*. For these tasks, we utilize collected screenshots and generate question-answer pairs using MLLMs, with the corresponding prompts provided in Figure 8. To ensure quality of each evaluation example, we follow the same **automatic assessment** and **human verification** process used in creating SR tasks, assessing the evaluation example based on clarity, reasonableness, and correctness. To further increase task difficulty, we leverage large language models (LLMs) to generate hard negative answers—plausible yet incorrect answers that closely resemble the correct one—thereby making the task more challenging.

C.4 Open-Vocab Classification

For Open-Vocab Classification (OVC) tasks, we start by constructing a mutually orthogonal and well-organized category corpus. Labels are then manually annotated on screenshots. The details of the taxonomy creation for each task are listed as follows:

- **Product Classification**: Categories are sourced from Amazon product pages and use hierarchies like *"Arts, Crafts & Sewing - Knitting & Crochet - Crochet Hooks"*. A word tree is used to refine these categories to ensure exclusivity.

- **News Topics**: Categories are collected from CNN, BBC, and Fox News and are manually consolidated to form a mutually exclusive and comprehensive set that covers a wide range of real-world scenarios.

- **Academic Fields**: Predefined categories are obtained from the Arxiv homepage[14].

- **Knowledge Classification**: Categories are defined based on the Wikipedia category page[15] and are manually combined to create an exclusive and comprehensive classification set for Wikipedia entries.

D Baselines and Evaluation Setup

We select retrieval models that are widely adopted in practice and academic research, which can be divided into three main categories: OCR + Text Retrievers, General Multimodal Retrievers, and Visualized Document Retrievers. In the OCR +

[14] https://arxiv.org/

[15] https://en.wikipedia.org/wiki/Wikipedia:Contents/Categories

Task	SubTask	#Query	#Corpus
SR	Product Retrieval	496	5436
	Paper Retrieval	503	5021
	Repo Retrieval	510	5024
	News Retrieval	491	5401
	Chart Retrieval	200	5000
	Doc Retrieval	200	5000
	Slide Retrieval	200	5000
CSR	Product Discovery	107	1012
	News-to-Wiki	101	1010
	Knowledge Relation	100	1011
	Wiki-to-product	100	969
SQA	Product-QA	498	2988
	News-QA	500	3000
	Wiki-QA	500	3000
	Academic-QA	481	2886
	Repo-QA	489	2934
OVC	Product Classification	500	5000
	News Toptics	584	74
	Academic Fields	500	142
	Knowledge Classification	480	46

Table 6: The specific number of queries and the corpus for each subtask

Text Retrievers category, we used Paddle OCR to extract text from the image part and performed text concatenation in the order from the top-left to the bottom-right. For evaluation metrics, we used the Recall metric and recorded the Recall@1 for each task.

D.1 OCR + Text Retrievers

BM25 (Robertson et al., 2009) is a classical information retrieval algorithm improved from TF-IDF, which optimizes relevance scoring through document length normalization (parameter b) and term frequency saturation mechanisms (parameter k_1). For evaluation, we use the code of dorian-brown/rank_bm25.
DPR (Karpukhin et al., 2020) employs dual BERT encoders for dense retrieval, mapping queries and documents into a shared vector space via contrastive learning. Trained with an "in-batch negatives" strategy for efficiency. For evaluation, we use the weight of facebook/dpr-question_encoder-single-nq-base.
BGE (Xiao et al., 2024) is a widely used BERT-based text embedding model. It is trained on a large text corpus and optimized using contrastive learning. For evaluation, we utilize the base version of BGE with weights from BAAI/bge-base-en-v1.5.
E5-Mistral (Wang et al., 2023) is a large language model-based embedding model derived from the Mistral-7B architecture. It uses the last hidden state of the [EOS] token as the embedding for the input text. For evaluation, we utilize the weights from intfloat/e5-mistral-7b-instruct.

D.2 General Multimodal Retrievers

VISTA (Zhou et al., 2024b) is a universal multi-modal retriever based on general text embedding models, such as pre-trained BERT models (Devlin et al., 2018; Xiao et al., 2024). It processes input images by converting them into sequences of patches, which are then fed into the pre-trained text embedding model alongside text tokens in an inter-leaved manner. Although VISTA was trained on natural images, we evaluated its performance on the MVRB by directly inputting document screen-shot images.
Uni-IR (Wei et al., 2024) is a single retriever to accomplish any multimodal retrieval task. UniIR can follow the instructions to take a heterogeneous query to retrival from a heterogeneous candidate pool with millions of candidates in diverse modalities. We use the weight of TIGER-Lab/UniIR.
CLIP (Radford et al., 2021) is a multimodal contrastive learning model comprising a ViT image encoder and Transformer text encoder. Pretrained on 400 million image-text pairs, CLIP maps cross-modal content into a shared vector space, enabling zero-shot image classification and cross-modal retrieval. It processes an image as 14*14 patches and outputs the 1+196 (ClS token and 196 patch tokens) sequence to represent the image. We use openai/clip-vit-large-patch14 for evaluation.
SigLIP (Zhai et al., 2023) is a multimodal model based on SoViT-400m (Alabdulmohsin et al., 2023) architecture with a better contrastive loss fuction. The sigmoid loss operates solely on image-text pairs and does not require a global view of the pairwise similarities for normalization. This allows further scaling up the batch size, while also performing better at smaller batch sizes. It pre-trained on WebLi (Chen et al., 2022) at resolution 384*384. We use the weight of google/siglip-so400m-patch14-384.
E5-V (Jiang et al., 2024a) is a unified multimodal embedding framework that maps text/image inputs

into a shared semantic space via prompt mechanisms. Innovatively trained with text-only pairwise data through single-modality contrastive learning. We use the weight of royokong/e5-v for evaluation.
VLM2Vec (Jiang et al., 2024b) builds upon the multimodal large language model Phi-3.5-V (Abdin et al., 2024). For any given input, it leverages the last hidden state of the [EOS] token to generate embeddings. VLM2Vec was trained on 20 diverse multimodal datasets, encompassing a range of tasks such as classification, visual question answering, retrieval, and visual grounding. Although the primary focus of its training corpus is on natural images, it also includes several document question answering datasets, such as DocVQA (Mathew et al., 2021a). This extensive training enables VLM2Vec to exhibit strong performance in the Screenshot Question Answering (SQA) tasks within our MVRB benchmark, outperforming other models, including our own UniSE, which are zero-shot for these tasks. Despite its competitive performance in SQA tasks due to this specialized training, VLM2Vec is classified as a general-purpose retriever in our taxonomy because of its limited exposure to document data, both in terms of quantity and the narrow focus on question answering tasks. For evaluation, we use the weight of TIGER-Lab/VLM2Vec-Full.
MM-Embed (Lin et al., 2024) is a multimodal dual-encoder model addressing vision-text bias through modality-aware negative sample mining. It supports hybrid-modality queries (e.g., text+image combinations) and enhances retrieval precision in complex scenarios via large language model reranking. We use the weight of nvidia/MM-Embed.

D.3 Screenshot Document Retrievers

DSE (Ma et al., 2024) is a bi-encoder model designed to encode document screenshots into dense vectors for document retrieval. It trained on Wiki pages and documents. We use the weight of Tevatron/dse-phi3-docmatix-v2.
ColPali (Faysse et al., 2024) is an advanced document retrieval model designed to leverage vision-language models (VLMs) to generate high-quality contextual embeddings from document page images. The model is based on PaliGemma-3B (Beyer et al., 2024) and employs the ColBERT (Khattab and Zaharia, 2020) strategy to create multi-vector representations of document page, thereby enhancing the accuracy and efficiency of retrieval. We use the weight of vidore/colpali-v1.3-hf for evaluation.
GME (Zhang et al., 2024b) is based on the Qwen2-VL model, and utilizes a contrastive learning approach to integrate diverse data into a unified semantic space. The GME has been trained on the visual document dataset Docmatix (Laurençon et al., 2024), which makes it have a great power to handle the visual document. For evaluation, we use the weight of Alibaba-NLP/gme-Qwen2-VL-2B-Instruct.

E Detailed Results for Each Task

The detailed results of different retrievers for each task on MVRB are presented in Table 7 (A) and Table 7 (B).

	BM25	DPR	BGE	E5-Mistral	CLIP	SIGLIP	VISTA	Uni-IR
SR								
Product	59.27	29.44	62.70	69.75	34.88	55.24	10.48	24.19
Paper	70.38	23.66	50.89	69.18	6.96	26.24	0.60	1.19
Repo	44.71	30.78	53.92	60.20	15.69	47.65	3.14	7.84
News	38.08	26.48	43.18	47.45	24.24	33.20	4.28	11.20
Chart	9.00	8.50	21.50	33.00	4.50	24.00	0.50	4.50
Document	19.00	19.00	30.50	29.00	7.50	21.00	2.50	6.00
Slide	44.50	37.00	57.00	51.50	38.50	61.00	15.00	31.50
Avg	40.71	24.98	45.67	51.44	18.89	38.33	5.21	12.35
CSR								
Knowledge Relation	19.00	15.00	18.00	23.00	18.00	30.00	6.00	23.00
News2Wiki	20.79	10.89	17.82	31.68	16.83	21.78	6.93	18.81
Product Discovery	25.23	28.04	27.10	33.64	32.71	55.14	25.23	29.91
Wiki2Product	49.00	21.00	84.00	92.00	34.00	31.00	7.00	47.00
Avg	28.51	18.73	36.73	45.08	25.39	34.48	11.29	29.68
SQA								
Repo-QA	46.63	35.38	49.69	33.33	27.81	24.54	37.63	29.24
News-QA	38.40	35.40	38.80	57.60	31.80	19.80	24.00	22.20
Product-QA	22.49	10.04	26.31	27.71	15.26	16.06	15.46	12.85
Paper-QA	46.78	30.15	32.85	52.18	22.45	21.62	31.60	21.62
Wiki-QA	38.00	25.80	33.80	44.20	22.20	16.00	20.20	21.20
Avg	38.46	27.25	36.29	43.00	23.90	19.60	25.78	21.42
OVC								
Product Classification	7.6	1.60	36.20	41.00	21.00	39.40	6.80	10.40
News Topics	8.22	39.04	63.35	59.42	50.68	63.18	30.99	39.38
Knowledge Classification	7.08	37.08	52.50	42.08	42.91	44.58	26.25	26.67
Academic Fileds	2.00	10.60	39.60	12.20	7.00	15.40	2.40	3.80
Avg	6.23	22.08	47.91	38.68	30.40	40.64	16.61	20.06
Overall Score								
All	30.81	23.74	41.99	45.51	23.75	33.34	13.85	19.63

Table 7 (A): Detailed performance results for each task of different retrievers on MVRB (Recall@1).

	VLM2Vec	E5-V	MM-Embed	DSE	Colpali	GME	UniSE-CLIP	UniSE-MLLM
SR								
Product	23.59	33.46	28.83	71.17	72.50	57.80	53.02	75.40
Paper	7.75	39.36	18.09	78.02	80.31	74.55	34.19	90.26
Repo	17.65	36.08	30.59	74.01	68.23	76.47	40.98	83.33
News	17.52	26.88	24.03	50.61	41.10	57.00	39.92	75.96
Chart	11.50	23.00	11.00	29.00	24.00	39.00	11.00	36.50
Document	12.00	26.00	18.50	47.50	59.50	51.00	17.00	50.50
Slide	21.50	54.00	50.00	80.50	86.50	75.50	55.50	75.50
Avg	15.93	34.11	25.86	61.54	61.73	61.62	35.95	69.64
CSR								
Knowledge Relation	47.00	18.00	26.00	22.92	16.00	20.00	29.00	41.00
News2Wiki	49.50	37.62	41.58	21.88	16.80	19.80	27.72	44.55
Product Discovery	32.71	34.58	41.12	24.04	25.20	28.90	45.79	51.40
Wiki2Product	63.00	30.00	55.00	82.29	82.00	82.00	71.00	81.00
Avg	48.05	30.05	40.93	37.78	35.00	37.68	43.38	54.49
SQA								
Repo-QA	58.28	36.81	53.99	46.93	42.33	48.80	35.58	57.00
News-QA	51.40	36.20	50.20	40.52	40.40	43.00	33.00	47.60
Product-QA	36.55	19.68	30.12	22.98	17.67	18.80	19.68	25.10
Paper-QA	43.78	34.10	38.46	46.67	46.98	45.30	25.36	46.70
Wiki-QA	57.00	30.20	41.40	39.11	29.20	33.00	27.00	39.60
Avg	49.42	31.40	42.83	39.24	35.32	37.78	28.13	43.20
OVC								
Product Classification	20.80	14.20	26.80	30.00	29.60	33.40	33.20	45.40
News Topics	46.75	52.57	46.75	56.68	44.00	69.60	57.19	69.34
Knowledge Classification	14.79	45.42	43.13	23.54	33.95	58.50	52.29	52.91
Academic Fileds	10.60	19.20	14.00	15.80	16.60	30.40	19.80	25.40
Avg	23.24	32.85	32.67	31.51	31.04	47.98	40.62	48.26
Overall Score								
All	32.19	32.37	34.48	48.14	43.64	13.3	34.99	55.72

Table 7 (B): Detailed performance results for each task of different retrievers on MVRB (Recall@1).

You are an advanced assistant specializing in generating **question-answer pairs** according to news articles. Based on the provided **Passage**, generate a question and answer.

Follow these guidelines:
1. The question should be:
   Specific: Focus on a key detail, event, or fact from the **Passage**.
   Relevant: The query must focus directly on the content of the news articles and should not introduce any irrelevant or external information.
   Clear: The query should be specific and clear, avoiding vague or overly broad expressions.
2. The answer should be:
   Concise: Provide a direct answer to the query without unnecessary information.
   Accurate: Ensure the answer is based solely on the provided **Passage**.
3.The question and answer should be returned in the following format:
   {"question": <Generated question>, "answer": <Generated answer>}

Note: If no relevant query can be generated from the text, return an empty dictionary.

The provided **Passage** is as follows:
<Passage>

Figure 4: The prompt used for q2s annotation. This prompt is designed for the news domain. For other domains, the word in blue can be substituted with the appropriate term for that domain.

You are an expert in creating technical **question-answer pairs** designed for effective arXiv paper retrieval.

You will be given a **Query Text** (a passage from one section or page of a paper) and a **Candidate Text** (a passage from another related section or page of a paper). Your task is to generate QA pairs where:
- The question must combine contextually with the query text to retrieve the candidate text.
- The answer should be derived exclusively from the candidate text, with no external information added.

Before generating the QA pairs, you should summarize the relation between the two passages from the perspectives of methodology, experiments, insights, and so on.

Your output must adhere to the following JSON format:
   {"question": <Generated question>, "answer": <Generated answer>}

Note: If no relevant question can be generated from the text, return an empty dictionary.

The provided query text and candidate text are as follows:

Query Text: <Query Text>

Candidate Text: <Candidate Text>

Figure 5: The prompt used for sq2s annotation. This prompt is intended for the paper domain. For other domains, the word in blue can be substituted with the appropriate term for that domain.

Instruction: Evaluate the given sample, consisting of **a QA pairs and an image**, using the following three dimensions. Output **acceptance decisions** with detailed rejection reasons in JSON format.

Evaluation Dimensions
1. Clarity
   - Requirement: Question clearly and specifically expresses an information need.
   - Reject if: The question uses vague, overly general, or ambiguous expressions.
2. Plausibility
   - Requirement: Question is realistic and makes sense in the context of the provided image.
   - Reject if: The question contains impossible scenarios/actions.
3. Validity
   - Requirement: The answer must be derivable from the image itself.
   - Reject if: The question requires external knowledge or information not present in the image.

Output Format
```
{
    "accept": <boolean>,
    "reject_reasons": [
      {"dimension": "clarity", "detail": "specific explanation"},
      {"dimension": "plausibility", "detail": "..."},
      {"dimension": "validity", "detail": "..."}
    ]
}
```

Input
Question: <Question>
Answer: <Answer>
Image: <Image>

Output

Figure 6: The prompt used for qulatiy control. We employ MLLMs to evaluate the clarity, plausibility, and validity of the evaluation examples.

Instruction: You are an expert in Visual Question Answering (VQA). Given **a question and an image**, determine if the question can be answered based **solely** on the information in the image.

Input:
Question: <Question>
Image: <Image>

Output:
If the question can be answered using only the information from the image, output True.
If the question cannot be answered using only the information from the image, output False.

Figure 7: The prompt used for filtering out false negatives.

Instruction: You are a professional data annotator tasked with generating a high-quality question-answer pair based on the input text-rich document screenshot image. The generated question must align strictly with the input content, and all outputs must meet the defined requirements.

- A single, natural language question that is closely related to the content of the input image.
- A corresponding answer that directly and concisely addresses the generated question based on the input image.

Requirements :
The question must be:
- **Relevance**: Be directly based on the input image and avoid introducing irrelevant or external information.
- **Singularity**: Contain only a single question, not a compound or multi-part question.
- **Hard**: The generated question should have a certain level of difficulty and not be simple extraction-based questions.
- **Vision Centric**: The question should reference the visual content of the image, such as charts, diagrams, tables, or figures. If no such visual references exist, base the question on the text of the image.

Input and output example: <Demonstration>

Your task: The input image is <Input Image>

Figure 8: The prompt used for creating screenshot question answering task.

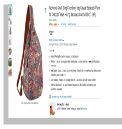
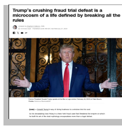
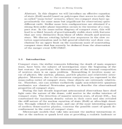
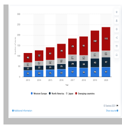
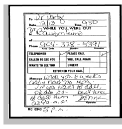
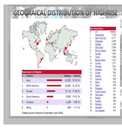
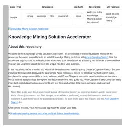
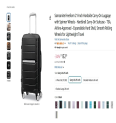
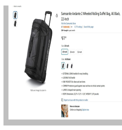
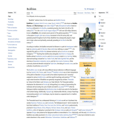
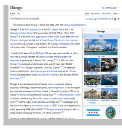
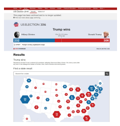
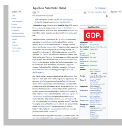
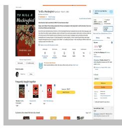
Task	Subtask	Query Text	Query Image	Target
Screenshot Retrieval	Product Retrieval	What are the dimensions of the Baosha Women's Small Sling Cross-body bag?	–	
	News Retrieval	What are the four foundational codes that explain Trump's tumultuous path through his life, according to the judge's ruling in the New York fraud case?	–	
	Paper Retrieval	What is the 'mass twins' scenario in the context of compact stars?	–	
	Chart Retrieval	What percentage of the prize pool did eSports team devils.one receive?	–	
	Document Retrieval	Who is the staff representative mentioned in the document?	–	
	Slide Retrieval	How many more skyscrapers does Tokyo have than London according to the slide?	–	
	Repo Retrieval	What are the prerequisites for successfully completing the Knowledge Mining Solution Accelerator?	–	
Composite Screenshot Retrieval	Product Discovery	I need a larger size of the same brand.		
	Knowledge Relation	Which temple is the Buddha statue in the image located?		
	News-to-Wiki	I need information about the political party of this man in this event.		
	Wiki-to-Product	Search the most relevant product page.		

Figure 9: Examples of tasks in Screenshot Retrieval and Composite Screenshot Retrieval in MVRB.

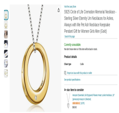
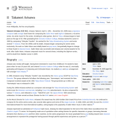
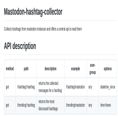
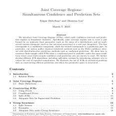
Task	Subtask	Query Text	Query Image	Target
Screenshot Question Answering	Product-QA	What type of chain is used in the S925 Circle of Life Cremation Memorial Necklace?		Cable.
	News-QA	What did Jeffrey Garten accidentally send to Ina Garten's publicist?		A text that said, "You're gonna be delicious tonight."
	Wiki-QA	What was the initial inspiration for the tone of the Sailor Moon soundtrack?		The initial inspiration for the tone of the Sailor Moon soundtrack was the Charlie's Angels TV series.
	Paper-QA	What is the key feature of the optimal contingent debt contract according to Proposition 5?		A contingent debt contract with at most one non-defaultable face value.
	Repo-QA	What does the Mastodon-hashtag-collector API method /hashtag/:hashtag return?		It returns the collected messages for a specific hashtag.
Open-Vocab Classification	Product Classification	–		Clothing, Shoes & Jewelry - Women - Uniforms, Work & Safety - Clothing - Medical - Scrub Tops
	News Topics	–		Sports, Football
	Arxiv Fileds	–		Artificial Intelligence
	Knowledge Classification	–		Society and social sciences, Government and Politics

Figure 10: Examples of tasks in Screenshot Question Answering and Open-Vocab Classification in MVRB.