

Towards Mixed-Modal Retrieval for Universal Retrieval-Augmented Generation

Chenghao Zhang
Renmin University of China
Beijing, China
davidzhang@ruc.edu.cn

Guanting Dong
Xinyu Yang
Renmin University of China
Beijing, China

Zhicheng Dou*
Renmin University of China
Beijing, China
dou@ruc.edu.cn

Abstract

Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm for enhancing Large Language Models (LLMs) by retrieving relevant documents from an external corpus. However, existing RAG systems primarily focus on unimodal text documents, and often fall short in real-world scenarios where both queries and documents may contain mixed modalities (such as text and images). In this paper, we address the challenge of Universal Retrieval-Augmented Generation (URAG), which involves retrieving and reasoning over mixed-modal information to improve vision-language generation. To this end, we propose **Nyx**, a unified mixed-modal-to-mixed-modal retriever tailored for URAG scenarios. To mitigate the scarcity of realistic mixed-modal data, we introduce a four-stage automated pipeline for data generation and filtering, leveraging web documents to construct **NyxQA**, a dataset comprising diverse mixed-modal question-answer pairs that better reflect real-world information needs. Building on this high-quality dataset, we adopt a two-stage training framework for **Nyx**: we first perform pre-training on **NyxQA** along with a variety of open-source retrieval datasets, followed by supervised fine-tuning using feedback from downstream Vision-Language Models (VLMs) to align retrieval outputs with generative preferences. Experimental results demonstrate that **Nyx** not only performs competitively on standard text-only RAG benchmarks, but also excels in the more general and realistic URAG setting, significantly improving generation quality in vision-language tasks. Our code is released at <https://github.com/SnowNation101/Nyx>

CCS Concepts

• **Information systems** → **Web search engines**; **Multimedia and multimodal retrieval**; **Web mining**; **Question answering**; **Retrieval models and ranking**.

Keywords

Multimodal Retrieval-Augmented Generation, Vision-Language Model, Multimodal Embedding, Contrastive Learning

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License. *SIGIR '26, Melbourne, VIC, Australia*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2599-9/2026/07
<https://doi.org/10.1145/3805712.3809716>



Figure 1: An illustration of the input patterns of “mixed-modal” content in the URAG scenario.

ACM Reference Format:

Chenghao Zhang, Guanting Dong, Xinyu Yang, and Zhicheng Dou. 2026. Towards Mixed-Modal Retrieval for Universal Retrieval-Augmented Generation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26)*, July 20–24, 2026, Melbourne, VIC, Australia. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3805712.3809716>

1 Introduction

Large Language Models (LLMs) have shown remarkable capabilities in text comprehension and generation [8, 11, 43]. To extend their capabilities to multimodal understanding, Vision-Language Models (VLMs) incorporate visual encoders to process text and image inputs [1, 52]. However, like LLMs, VLMs often struggle with queries needing up-to-date or external knowledge. Retrieval-Augmented Generation (RAG) addresses this by retrieving documents from an external corpus to complement internal knowledge [10, 27]. Building on this, Multimodal RAG (MRAG) extends the paradigm to settings where both queries and documents may contain text, images, or both [5, 45].

Current MRAG methods fall broadly into two categories: (1) **The divide-and-conquer approach**, which utilizes text queries for text documents and visual queries for images; (2) **The cross-modal retrieval**, which uses text queries to retrieve visual content. However, both paradigms suffer from notable limitations. They often overlook the spatial and logical relationships between images and text within a document, making it difficult to capture fine-grained interactions crucial for downstream reasoning.

Furthermore, web documents in the real world are far more complex and diverse. As illustrated in Figure 1, they may include pure text, individual images, paired image-text content, or arbitrarily interleaved sequences of text and images. We refer to this broad

spectrum of formats as *mixed-modal* content, where the interplay between modalities plays a critical role in conveying meaning.

While recent efforts, such as VLM2Vec [20], have introduced unified multimodal embedding models, these approaches mainly focus on embedding pure text, individual images, or neatly aligned text-image pairs. Consequently, they face challenges in handling complex multimodal structures, such as interwoven or densely intertwined text and images. Furthermore, their application within the MRAG framework is still largely unexplored.

This gap becomes even more pronounced in the context of Universal Retrieval-Augmented Generation (URAG), as both queries and documents can be of arbitrary mixed modalities. Unlike traditional settings with purely textual queries, URAG introduces dual challenges: understanding heterogeneous inputs and retrieving from equally diverse corpora. An effective retriever needs to be capable of encoding various content types and matching them with complex document structures. This imposes new technical requirements on representation learning, matching precision, and alignment with downstream VLMs, highlighting the need for a truly universal, flexible, and vision-language-aware retrieval paradigm.

To address these challenges, we propose **Nyx**, a unified retriever designed for mixed-modal-to-mixed-modal retrieval in URAG scenarios. To mitigate data scarcity of realistic URAG data, we first introduce a four-stage fully automatic pipeline to build **NyxQA**, a novel dataset tailored for URAG. **NyxQA** consists of three components: (1) a large-scale mixed-modal multiple-choice question answering (QA) dataset, (2) a corresponding mixed-modal document corpus, and (3) a pretraining dataset for contrastive learning.

Our construction process begins with sampling naturally interleaved image-text documents from the web to form the corpus. We then employ a powerful VLM to generate QA pairs conditioned on these documents. To ensure high data quality, we apply a multi-step post-processing procedure, yielding a clean and diverse QA set. Finally, based on these QA pairs, we mine hard negatives from the corpus to form the pretraining triplets used for contrastive learning. Unlike existing multimodal datasets limited to specific modality combinations, **NyxQA** supports retrieval and generation involving arbitrarily structured text, images, and their interleaved formats.

Building upon this dataset, we adopt a two-stage training framework to develop **Nyx** from a VLM. In the first stage, we pretrain the retriever on **NyxQA** and several public datasets to establish general-purpose mixed-modal retrieval capabilities. To balance retrieval effectiveness and efficiency, we incorporate Matryoshka Representation Learning (MRL) [24], resulting in a compact yet expressive embedding model, termed **Nyx**-pretrained. In the second stage, we conduct feedback-driven fine-tuning by harvesting preference signals from the downstream VLM. We then use the resulting preference dataset to further fine-tune the retriever, encouraging it to retrieve contexts that are both semantically relevant and more conducive to the generator’s output quality, yielding the final retriever, **Nyx**.

Extensive experiments demonstrate that **Nyx** consistently enhances retrieval accuracy and downstream reasoning performance in challenging mixed-modal scenarios, showcasing its strong suitability for URAG tasks.

In conclusion, our contributions are as follows:

(1) We pioneer the exploration of the Universal Retrieval-Augmented Generation (URAG) problem, addressing scenarios where both queries and documents consist of arbitrarily interleaved image-text content.

(2) We introduce **NyxQA**, a dataset specifically designed for real-world URAG applications, created through a comprehensive four-step web-based multimodal data synthesis pipeline. This dataset offers a wide variety of interleaved content formats, serving as an effective benchmark for practical multimodal retrieval tasks.

(3) We propose a two-stage training framework for **Nyx**, combining MRL-based contrastive pretraining with downstream VLM preference fine-tuning to better align retrieval with generation.

2 Related Work

Multimodal Retrieval-Augmented Generation (MRAG) extends the traditional RAG framework to multimodal settings by retrieving text, images, or image-text pairs from an external corpus to support Vision-Language Models (VLMs) in generating textual responses [5]. Current methods use various retrieval strategies: dual-path strategies retrieve text with text queries and images with image queries [9, 35]; cross-modal retrieval techniques [6, 45]; and treating multimodal documents as images for retrieval [38].

In real-world applications, queries and documents often contain *mixed-modal* inputs—combinations of text and images. New multimodal deep search paradigms introduce iterative retrieval [14], where intermediate queries may also be mixed-modal. However, a unified retrieval framework for URAG scenarios remains undeveloped.

Multimodal Embedding Retrievers focus on retrieving relevant multimodal documents by encoding both queries and documents into a shared embedding space. In text-only contexts, embedding-based retrievers have shown strong performance across various tasks and languages [3, 17, 22, 41]. Extending this to multimodal scenarios, cross-modal retrievers like CLIP [18, 23, 32, 46] and vision-language models such as BLIP-2 [13, 28] encode text and images into a unified space, enabling retrieval tasks such as image-to-text, text-to-image, and image-to-image.

Recent advancements [19, 20, 42, 50] have leveraged VLMs as general-purpose encoders for text, images, and image-text pairs. Other studies have focused on using synthetic data [2, 49, 51] and improving contrastive learning objectives [25, 39] to enhance embedding quality. Building on this, MME [48] utilized synthetic data to improve performance on interleaved text-image retrieval in the wikiHow task.

However, these methods lack support for text-to-text and general interleaved text-image retrieval in URAG scenarios. Moreover, most retrievers are trained independently of downstream VLMs, resulting in suboptimal alignment. Therefore, this paper proposes a unified retriever, **Nyx**, which builds a bridge for mixed-modal to mixed-modal retrieval, leading to better alignment with VLM generation.

3 Methodology

3.1 Problem Formulation: URAG

This work addresses the task of **Universal Retrieval-Augmented Generation (URAG)**, which aims to generate high-quality textual

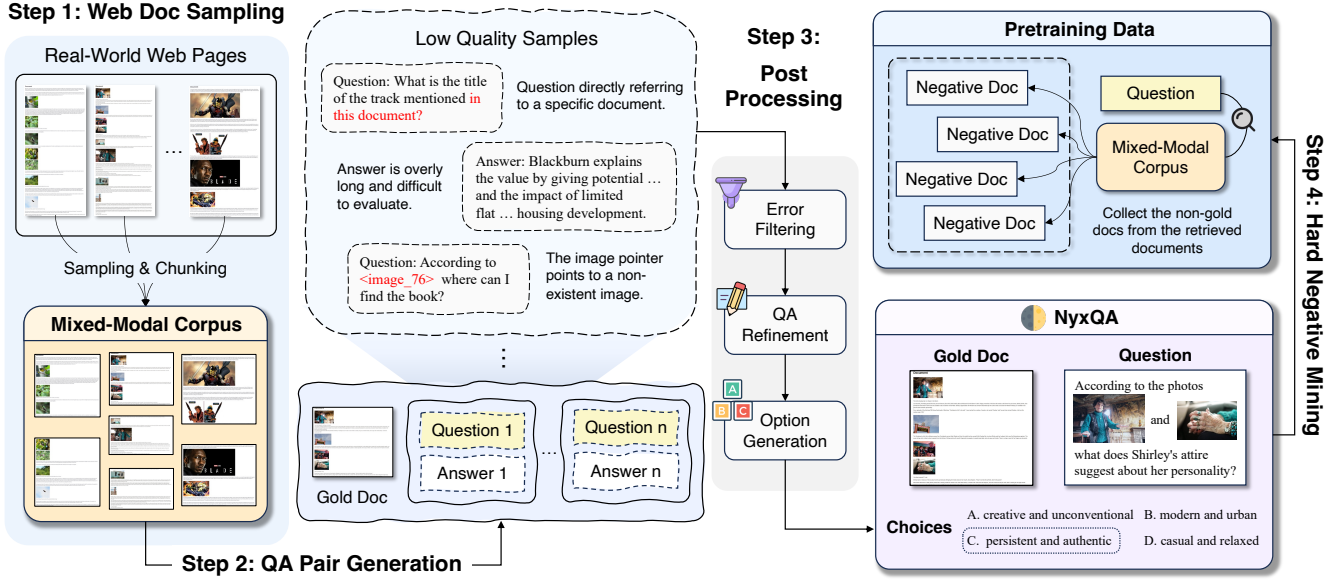


Figure 2: The proposed four-step automated NyxQA construction pipeline.

responses to mixed-modal queries by retrieving and leveraging relevant information from a mixed-modal corpus. A **mixed-modal content** x is defined as an ordered sequence of elements, where each element can be either a textual segment or an image. Formally, it can be represented as: $x \in \{a_1 a_2 \dots a_n \mid a_i \in \{\mathcal{T}, \mathcal{I}\}\}$.

To effectively accomplish this task, we presuppose access to a **mixed-modal corpus** C . A retriever is required to retrieve a relevant subset of **documents** $\{d_i\} \subseteq C$, conditioned on a mixed-modal **question** q . The retrieved content then serves to guide the generation of the ultimate textual response. Formally, the objective is to learn a retrieval function $\mathcal{R}$ and a generation model $\mathcal{G}$ such that: $y = \mathcal{G}(p, q, \mathcal{R}(q))$, where p denotes a textual prompt, q is the mixed-modal query, and $\mathcal{R}(q)$ represents the set of retrieved documents. The output y is a **natural-language textual response**. The retrieval function $\mathcal{R}$ selects the top- K most relevant entries from the corpus C based on their relevance to the query q :

$$\mathcal{R}(q) = \text{Top-}K_{d \in C} \text{sim}(q, d),$$

where d denotes a document in the mixed-modal corpus and $\text{sim}(\cdot, \cdot)$ is a similarity-based relevance function defined within the joint vision-language embedding space.

The generation model $\mathcal{G}$, typically instantiated as a VLM, processes both the query q and the retrieved documents $\mathcal{R}(q)$ to yield a coherent and factually grounded textual output y .

3.2 NyxQA: A Dataset for URAG

To simulate a realistic web environment, we introduce **NyxQA**, a large-scale mixed-modal dataset designed for the URAG setting. Our dataset comprises three components: (1) a high-quality multiple-choice QA dataset $\mathcal{D}_{\text{NyxQA}}$ with mixed-modal questions, (2) a corresponding mixed-modal document corpus C_{mix} , and (3) a contrastive pretraining set $\mathcal{D}_{\text{pretrain}}$ containing positive and hard negative examples for retriever training.

The construction of **NyxQA** follows a four-stage pipeline (Figure 2). First, we sample and segment web documents to create a diverse mixed-modal corpus. Next, we utilize a VLM to generate QA pairs from these document segments. This is followed by a post-processing pipeline to filter errors, refine answers, and format multiple-choice options. Finally, we employ hard negative mining using an existing retriever to produce high-quality contrastive training triplets for pretraining the mixed-modal retriever.

Step 1: Web Document Sampling. To obtain naturally occurring mixed-modal documents, we sample from OBELICS [26], a large-scale dataset of web pages featuring interleaved text and images that reflect real-world multimodal distributions. Following standard practices in text-only RAG [21], each document is segmented into smaller *chunks* $\{d_i\}_{i=1}^N$, where each d_i contains up to 200 textual tokens (excluding image tokens from the count). This segmentation maintains semantic coherence and prevents length imbalance caused by densely illustrated documents. The resulting set of chunks forms our mixed-modal corpus $C_{\text{mix}} = \{d_i\}_{i=1}^N$, comprising 49,723 segments in total (35,178 text-only, 13,266 single-image, and 1,279 multi-image). We then perform stratified sampling to select 10,000 chunks from C_{mix} for QA pair generation, while preserving the original modality ratios.

Step 2: QA Pair Generation. For each sampled chunk d_i , whether text-only or containing images, we use a VLM to generate up to five context-independent raw QA pairs $(q_{ij}^{\text{raw}}, a_{ij}^{\text{raw}})$, ensuring that each question can be answered solely based on its associated chunk. For chunks with images, we specifically prompt the VLM to create questions that reference the visual content. Since the model outputs text only, we use special tags such as `<image k>` to denote the k -th image within the chunk. This process produces the raw QA dataset $\mathcal{D}_{\text{raw}} = \{(d_i, q_{ij}^{\text{raw}}, a_{ij}^{\text{raw}})\}$, which contains diverse samples

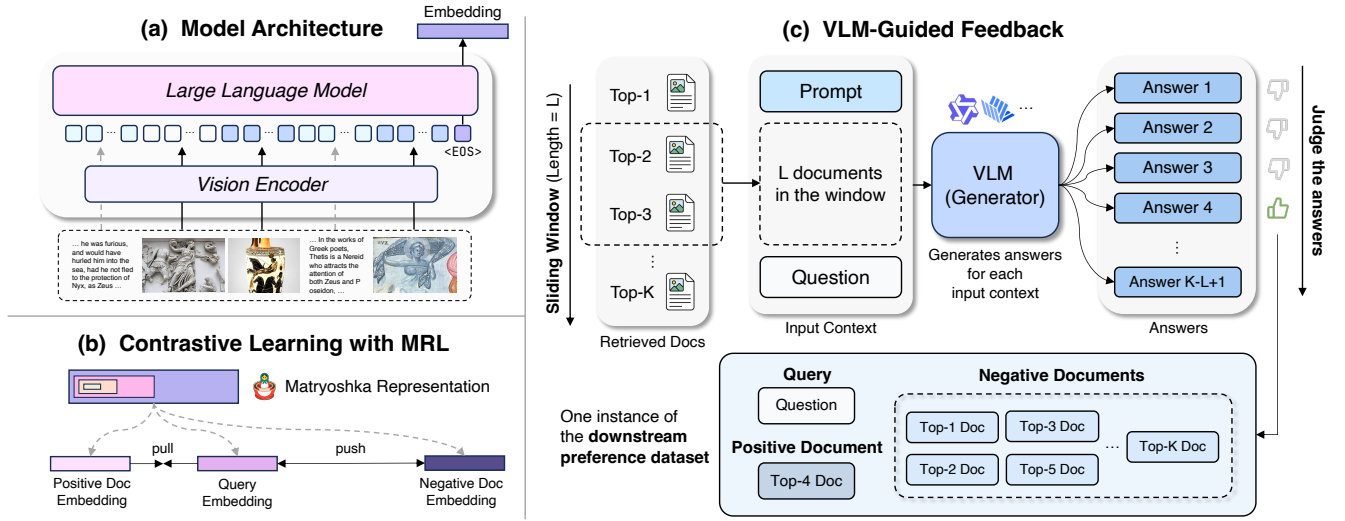


Figure 3: Overview of the Nyx architecture and its training paradigm.

ranging from pure text to multi-image questions, thereby enriching the modality diversity of **NyxQA**.

Step 3: Post Processing. The initial set $\mathcal{D}_{\text{raw}}$ of generated QA pairs is of suboptimal quality, containing various errors that could adversely affect subsequent training and evaluation. Therefore, we perform a three-stage post-processing procedure on the raw data to produce the final **NyxQA** dataset.

- **Error Filtering.** Questions with explicit contextual references (e.g., phrases like “in this document”) are removed using rule-based filters. In addition, we ensure image–text consistency by verifying that the image tags mentioned in the generated question correspond to actual images present in the chunk d_i .
- **QA Refinement.** We further refine the filtered QA pairs using a VLM to enhance clarity and completeness. Each retained pair $(q_{ij}^{\text{raw}}, a_{ij}^{\text{raw}})$ is compressed to its essential content, eliminating redundancy while preserving factual accuracy. This process yields concise, self-contained questions and answers that align closely with the corresponding gold document d_i^+ , resulting in the refined set $(d_i^+, q_{ij}, a_{ij}^+)$.
- **Option Generation.** For each refined QA, an LLM generates three semantically plausible distractors $\{a_{ij}^-\}$ for the question q_{ij} . After shuffling the distractors with the correct answer, we finalize each sample with the question, four options, and the gold document, forming our multiple-choice dataset $\mathcal{D}_{\text{NyxQA}} = \{(q_{ij}, \{a_{ij}^+, a_{ij}^-\}, d_i^+)\}$.

Step 4: Hard Negative Mining. To enhance retriever pretraining, we construct contrastive triplets based on $\mathcal{D}_{\text{NyxQA}}$. Each question q_{ij} serves as a query, with its corresponding gold document d_i^+ (the same chunk d_i that was used to generate the corresponding QA instance) designated as the positive sample. We then employ mmE5 [2] to retrieve the top-10 relevant documents from the mixed-modal corpus $\mathcal{C}_{\text{mix}}$. From these, we select five documents that differ from d_i^+ as hard negatives $\{d_{ij}^-\}$, prioritizing

the highest-ranked candidates. This yields the pretraining dataset $\mathcal{D}_{\text{pretrain}} = \{(q_{ij}, d_i^+, \{d_{ij}^-\})\}$, a contrastive training set specifically designed for mixed-modal retrieval.

3.3 Nyx: Training Paradigm

Our goal is to build a unified retriever capable of handling mixed-modal queries and documents across diverse real-world scenarios. To this end, we begin by pretraining **Nyx** on a large-scale corpus that includes both public and synthetic datasets spanning various modality configurations, resulting in **Nyx**-pretrained. This initialization equips the retriever with general-purpose retrieval capabilities across text-only, image-only, and multimodal pairs.

However, generic pretraining may not fully align with the specific information needs of downstream VLMs during generation. Therefore, in the second stage, we fine-tune **Nyx**-pretrained through a feedback-driven learning process, leveraging VLM responses to construct high-quality examples that reflect the actual relevance signals needed for multimodal generation.

Throughout both stages, we employ contrastive learning with Matryoshka Representation Learning (MRL) [24], as illustrated in Figure 3(b), to enable scalable and efficient retrieval while maintaining high embedding quality. We detail our training objective and the two-stage training procedure below.

Training Objective. As shown in Figure 3(a), our retriever is built upon Qwen-2.5-VL-3B-Instruct [1] pretrained VLM, which serves as our backbone encoder. For a given input sequence, we leverage the last hidden state of the final $\langle \text{EOS} \rangle$ token as the global embedding for retrieval.

Following established practices in embedding model training, we construct each training instance as a triplet $\{q, d^+, \{d_n^-\}_{n=1}^N\}$, where q is a query, d^+ is a positive document, and $\{d_n^-\}$ are N negative documents. An instruction string is prepended to each query before encoding. Both queries and documents may come

from mixed modalities (e.g., text, image, or interleaved image-text), allowing **Nyx** to operate in a unified embedding space.

To learn discriminative representations, we adopt the InfoNCE loss for contrastive learning. Let $\mathbf{h}_q \in \mathbb{R}^d$, $\mathbf{h}^+ \in \mathbb{R}^d$, and $\{\mathbf{h}_n^-\}_{n=1}^N \subset \mathbb{R}^d$ denote the embeddings of the query, the positive document, and the negative documents, respectively. We apply MRL, which encourages the full embedding $\mathbf{h} \in \mathbb{R}^d$ to remain informative even when truncated to lower-dimensional subspaces. This enables flexible trade-offs between retrieval performance and memory efficiency.

Specifically, for a set of target dimensions $\{\text{dim}_1, \dots, \text{dim}_K\}$, where each dim_k is smaller than the model’s original embedding dimension, we truncate each embedding to its first dim_k dimensions, denoted as $\mathbf{h}^{(\text{dim}_k)} \in \mathbb{R}^{\text{dim}_k}$. For each dim_k , we compute an InfoNCE loss as:

$$\mathcal{L}_{\text{Info}}^{(\text{dim}_k)} = -\log \frac{\phi(\mathbf{h}_q^{(\text{dim}_k)}, \mathbf{h}^{+(\text{dim}_k)})}{\phi(\mathbf{h}_q^{(\text{dim}_k)}, \mathbf{h}^{+(\text{dim}_k)}) + \sum_{n=1}^N \phi(\mathbf{h}_q^{(\text{dim}_k)}, \mathbf{h}_n^{-(\text{dim}_k)})}, \quad (1)$$

where $\phi(\mathbf{a}, \mathbf{b}) = \exp(\text{sim}(\mathbf{a}, \mathbf{b})/\tau)$, and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity with temperature hyperparameter $\tau > 0$. Here, $\mathbf{h}^{+(\text{dim}_k)}$ and $\mathbf{h}_n^{-(\text{dim}_k)}$ correspond to the positive and the n -th negative sample document embeddings, respectively.

The final training objective aggregates the InfoNCE losses over all truncated dimensions as a weighted sum:

$$\mathcal{L}_{\text{MRL}} = \sum_{k=1}^K w_k \cdot \mathcal{L}_{\text{Info}}^{(\text{dim}_k)}, \quad \text{where } \sum_{k=1}^K w_k = 1, \quad (2)$$

with w_k denoting the weight for the k -th dimension. This objective encourages each embedding prefix to preserve semantic integrity under varying retrieval constraints.

Stage 1: Pretraining with Mixed-Modal Data. In the first stage, we pretrain **Nyx** as a general-purpose retriever using a large-scale corpus constructed from both public and synthetic data sources. Following mmE5 [2], we include MMEB [20] and synthetic triplets from the mmE5 pipeline. To ensure the model’s ability to handle genuinely mixed-modal scenarios, we further integrate our proposed **NyxQA** dataset, which supports retrieval across diverse modality combinations.

Since real-world retrieval tasks still predominantly involve textual inputs, we further enhance the retriever’s text understanding ability by introducing additional text-only datasets. Specifically, we use the training sets from HotpotQA [44], 2WikiMultiHopQA [15], and MuSiQue [40]. For each query, we retrieve the top- K documents from the full Wikipedia corpus using E5-v2 [41], and treat the top-1 document as the positive sample, while selecting negative samples from documents ranked beyond top-10.

All the datasets described above are combined and jointly used to train the initial retriever, resulting in the **Nyx**-pretrained model.

Stage 2: Supervised Fine-tuning with VLM-Guided Feedback. While **Nyx**-pretrained demonstrates strong retrieval performance, it is not explicitly optimized for supporting downstream generation by VLMs. To bridge this gap, we introduce a fine-tuning stage that leverages feedback from a VLM to align the retriever with the actual information needs during generation.

Given a dataset $D = \{(q_i, a_i)\}$ of queries and their corresponding answers, along with a retrieval corpus C , we proceed as follows.

For each query q_i , we first use **Nyx**-pretrained to retrieve the top- K candidate documents $\{d_1, d_2, \dots, d_K\}$. Then, using a sliding window of length L , we construct a sequence of candidate contexts by grouping contiguous subsets of the retrieved documents. Each context window is concatenated with the query and fed into the VLM to generate an answer.

We select the first context window that either yields a correct answer or exceeds a pre-defined generation-quality threshold (e.g., EM, F1). Prior work suggests that LLMs exhibit biased document usage in long-context RAG, often relying more on higher-ranked and boundary-positioned documents [29, 36]. Therefore, the first document in this window is treated as the positive sample d^+ , and the remaining $K - 1$ documents are used as negative samples $\{d_n^-\}$. If no window meets the quality threshold, the entire instance is discarded from the feedback dataset.

By applying this procedure to all queries in D , we construct a **downstream preference dataset** from the feedback $D_{\text{pref}} = \{(q_i, d_i^+, \{d_{i,n}^-\})\}$, which reflects the actual preferences of the VLM in real generation scenarios. We then fine-tune **Nyx**-pretrained on this dataset using the same contrastive learning framework described earlier, thus obtaining the final retriever **Nyx**.

4 Experimental Settings

4.1 Datasets

HotPotQA [44] is a widely used multi-hop QA dataset requiring reasoning across multiple Wikipedia articles. It contains 90,447 training samples, and we evaluate on a held-out validation set of 250 examples. **2WikiMultiHopQA** [15] is a large-scale multi-hop reasoning dataset built from Wikidata and Wikipedia, with diverse questions and annotated reasoning chains. It includes 15,000 training samples, and we evaluate on the test set of 250 examples. **Bamboogle** [31] contains manually curated multi-hop questions for compositional reasoning, with up to four inference steps. It provides only a test set of 125 examples, which we use for evaluation. **MuSiQue** [40] targets sequential multi-hop inference where each step depends on the previous one. It comprises 19,938 training examples, and we evaluate on its development set of 250 held-out samples.

For the four text-only datasets above, we build two Wikipedia-derived corpora for training and evaluation. For each corpus, we retrieve top-20 relevant Wikipedia passages for all questions using the E5 retriever over the full Wikipedia dump, then deduplicate the results. The training corpus is used in feedback collection, where **Nyx**-pretrained retrieves passages to construct the “downstream VLM preference dataset” for fine-tuning, while the evaluation corpus is used for final benchmarking on the four text-only datasets.

MultimodalQA [37] requires joint reasoning over text, tables, and images. It contains 23,817 training and 2,411 testing examples. We merge these modalities to build a mixed-modal corpus of 285,370 instances for MMQA. **ScienceQA** [30] is a large-scale multimodal science QA dataset with lecture-style explanations and context in text or images. We construct two corpora: a deduplicated lecture corpus and a training QA-pair example corpus. During testing, we retrieve one lecture and two example QAs as external support. The dataset includes 12,726 training and 4,241 testing examples.

Table 1: The overall results on the six RAG datasets. To ensure consistent evaluation, the top document retrieved by each retriever was combined with the corresponding question, then input into Qwen2.5-VL-7B for answer generation. The exception is SciQA, where the retrieval content consists of one lecture and two example-based retrieval results to suit the dataset’s structure. This setup isolates the effect of the retrievers, facilitating a controlled comparison of retrieval performance. The best results are highlighted in bold, and the second-best results are underlined.

Method	HotpotQA		Bamboogle		MuSiQue		SciQA	MMQA		NyxQA	Avg.
	EM	F1	EM	F1	EM	F1	Acc	EM	F1	Acc	
<i>Direct Answer</i>											
InternVL3 (8B) [52]	16.40	22.88	9.60	15.49	3.60	8.16	78.87	20.07	23.99	53.33	25.24
Qwen2.5-VL (7B) [1]	12.40	18.36	6.40	11.50	3.29	7.32	77.98	20.73	24.39	50.17	23.25
<i>Text RAG</i>											
E5-v2 (0.1B) [41]	14.40	19.18	7.20	12.80	2.40	6.79	–	–	–	–	–
<i>Vision-Language RAG</i>											
CLIP (0.4B) [32]	14.00	21.12	6.40	11.64	3.20	6.74	73.07	18.03	20.67	61.50	23.64
VLM2Vec (4B) [20]	14.40	22.08	10.40	16.95	3.60	10.12	79.56	19.91	23.34	56.50	25.69
VisRAG-Ret (3B) [45]	12.08	19.84	8.80	16.05	3.60	8.29	80.45	18.84	21.55	64.33	25.38
mmE5 (11B) [2]	17.60	24.30	13.60	18.69	5.20	9.70	<u>81.40</u>	<u>34.00</u>	<u>38.50</u>	66.83	30.71
<i>Ours</i>											
Nyx-pretrained (3B)	<u>22.00</u>	<u>31.38</u>	<u>16.00</u>	<u>22.87</u>	<u>5.60</u>	<u>11.00</u>	81.33	31.75	35.97	<u>74.83</u>	<u>33.27</u>
Nyx (3B)	24.40	33.19	16.80	25.93	7.20	12.80	81.75	39.66	44.50	81.83	36.81

4.2 Training Settings

We train **Nyx** based on the Qwen2.5-VL-3B model using a single node equipped with $8 \times$ NVIDIA A800-SXM4-80GB GPUs. To enable efficient fine-tuning, we apply Low-Rank Adaptation (LoRA) [16] with a rank of 8. Each GPU processes a batch of 20 samples, and we accumulate gradients over 4 steps, resulting in an effective batch size of 640. The temperature parameter τ in the InfoNCE loss is fixed at 0.02. To prevent memory overflow when processing multi-image inputs, we cap the visual input resolution at $400 \times 28 \times 28$ pixels.

We use DeepSpeed [34] with bf16 mixed-precision training and enable gradient checkpointing for memory efficiency. The optimizer is AdamW, combined with a linear learning rate scheduler. The base learning rate is set to $1e-5$, with a warmup ratio of 0.05, and a maximum gradient norm of 5.0. The contrastive loss function uses a temperature of 0.02 and a negative sampling ratio of 1.

For efficient memory optimization, we employ DeepSpeed’s ZeRO optimization [33] at stage 2, which partitions model states across devices to significantly reduce memory consumption. This allows us to train larger models without overflow. ZeRO stage 2 also enables communication optimizations such as allgather and reduce-scatter, improving parallelism and computational efficiency. These configurations, combined with gradient checkpointing and mixed-precision training, work together to maximize both performance and memory efficiency during training.

5 Main Experiments

Our main experiments consist of two parts, where we first evaluate the generation performance in URAG scenarios and then examine the embedding performance, given that the model is inherently an embedding model.

5.1 Evaluation Methodology for Generation

Datasets and Metrics. We evaluate RAG pipelines incorporating our retriever across two categories: (1) text-only datasets and (2) multimodal datasets, including MRAG and URAG.

For text-only RAG, following RESEARCH [4], we evaluate on **HotpotQA** [44], **MuSiQue** [40], and **Bamboogle** [31]. For each question, we use E5-v2 to retrieve the top 20 documents from Wikipedia, merge and deduplicate them to create a task-specific corpus, and randomly sample up to 250 questions per dataset.

In the case of MRAG and URAG, we utilize **MultimodalQA** (MMQA) [37], **ScienceQA** (SciQA) [30], and **NyxQA**. MMQA requires retrieval from a corpus containing text, tables, and images. Since the tables in the MMQA corpus are originally stored in JSON format, we employ the Python tabulate library to convert them into a more comprehensible text table format. SciQA presents image-text questions, and we construct its corpus using associated lectures and QA examples. **NyxQA**, constructed from authentic web pages, covers a broader spectrum of input and document modalities, thereby facilitating more realistic and comprehensive evaluation of URAG in real-world web environments.

During pretraining, we use the training sets of 2WikiMulti-HopQA, HotpotQA, MuSiQue, and **NyxQA**, treating Bamboogle, MMQA, and SciQA as out-of-domain (OOD) evaluation sets. For feedback-based fine-tuning, feedback is collected from HotpotQA, MuSiQue, MMQA, SciQA, and **NyxQA**, with Bamboogle serving as the OOD benchmark.

For multiple-choice datasets, we report **accuracy** (Acc), while for open-ended QA, we adopt **exact match** (EM) and **F1 score** (F1), reflecting both strict correctness and token-level overlap between predictions and references.

Table 2: Performance comparison on the MMEB benchmark, which includes 36 tasks spanning four categories: classification (Cls.), visual question answering (VQA), retrieval (Retr.), and visual grounding (Grd.). Bold values denote the best results.

Models	Per Meta-Task Score				Overall
	Cls.	VQA	Retr.	Grd.	
<i>Zero-Shot Setting Models</i>					
CLIP (0.4B) [32]	42.8	9.1	53.0	51.8	37.8
BLIP2 (4B) [28]	27.0	4.2	33.9	47.0	25.2
OpenCLIP (0.4B) [7]	47.8	10.9	52.3	53.3	39.7
E5-V (8B) [19]	21.8	4.9	11.5	19.0	13.3
MagicLens (0.4B) [47]	38.8	8.3	35.4	26.0	27.8
MMRet (7B) [51]	47.2	18.4	56.5	62.2	44.0
VLM2Vec (4B) [20]	52.8	50.3	57.8	72.3	55.9
<i>Supervised Fine-tuning Models</i>					
CLIP (0.4B) [32]	55.2	19.7	53.2	62.2	45.4
OpenCLIP (0.4B) [7]	56.0	21.9	55.4	64.1	47.2
VLM2Vec (4B) [20]	61.2	49.9	67.4	86.1	62.9
MMRet (7B) [51]	56.0	57.4	69.9	83.6	64.1
mmE5 (11B) [2]	67.6	62.8	70.9	89.7	69.8
mmE5-Qwen-3B (3B)	56.6	56.0	59.4	71.5	59.0
<i>Ours</i>					
Nyx -pretrained (3B)	55.2	53.7	58.4	70.5	57.5
Nyx (3B)	57.9	57.5	61.8	75.7	61.1

Baseline Models. For the text-only retriever, we use E5-v2 [41] as the unimodal RAG baseline, since it serves as the backbone model for constructing our text-only retrieval datasets and is also one of the most widely used retrievers in text-based RAG systems [21]. For multimodal retrievers, we use well-supervised fine-tuned embedding models CLIP [32], VLM2Vec [20] and mmE5 [2], as well as a retriever for visual document retrieval for RAG, VisRAG-Ret [45]. We also report the direct answering results of InternVL3-8B [52] and Qwen2.5-VL-7B as baselines for comparison.

5.2 Results on Generation Performance

Our generation performance results are presented in Table 1. Overall, **Nyx** consistently outperforms all baselines, clearly demonstrating its superiority. We further highlight the following insights:

Performance in Text-Only RAG. Despite the powerful 11 billion parameter VLM backbone of mmE5, our 3 billion parameter **Nyx**-pretrained model still outperforms mmE5 on HotpotQA, Bamboogle, and MuSiQue, with performance gains of 9% and 6% on HotpotQA and Bamboogle, respectively. This result shows the strength of targeted training. Moreover, **Nyx** substantially surpasses the text-only retriever E5 that is commonly used in RAG frameworks, further demonstrating its effectiveness in unimodal retrieval.

Multimodal RAG Performance. In multimodal tasks, **Nyx**-pretrained performs competitively on MMQA and **NyxQA**, though it trails mmE5 slightly on SciQA. This may be attributed to **Nyx**'s smaller parameter count and its broader training coverage, which

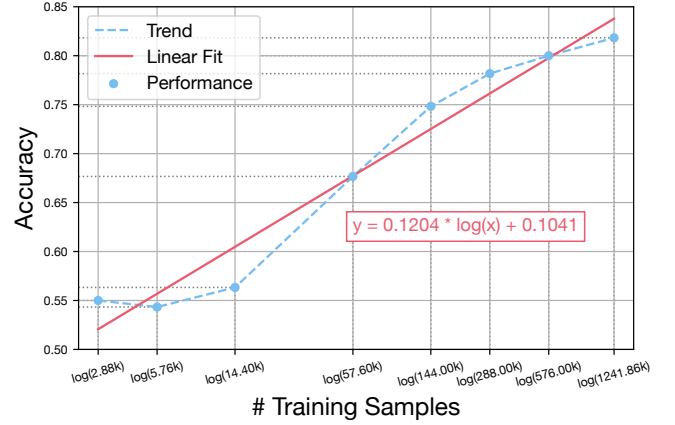


Figure 4: Impact of training data scale on NyxQA accuracy when training Nyx with varying sample sizes.

includes interleaved and text-only examples. Nevertheless, its robust performance across different input types highlights the benefit of mixed-modal training. After incorporating feedback from downstream VLMs, **Nyx** achieves the best performance across all multimodal benchmarks, with great results on MMQA (F1: 35.97% → 44.50%) and **NyxQA** (Accuracy: 74.83% → 81.83%). On SciQA, the gain is modest, possibly due to the limited informativeness of the provided lecture corpus. Nonetheless, fine-tuning with feedback still leads to alignment with the VLM's preferences.

A McNemar's test was conducted on **NyxQA** to assess the performance differences among mmE5, **Nyx**-pretrained, and **Nyx** as retrievers. The comparison between mmE5 and **Nyx**-pretrained yielded a test statistic of 19.0631 (χ^2 , 1 degree of freedom) with a p-value of 0.0000. Furthermore, the comparison between **Nyx**-pretrained and **Nyx** resulted in a test statistic of 15.7538 and a p-value of 0.0001. These results provide strong evidence that the retrieval performance differs significantly across the methods.

Beyond Gold Documents: Learning from Preference. An interesting observation arises from **NyxQA**, where each question is originally paired with a generation-originated "golden" document. Although semantically relevant, these gold documents do not always lead to correct answers during inference. Our feedback analysis shows that documents preferred by the VLM may differ from the labelled positives. Incorporating this preference signal during fine-tuning leads to a 7-point accuracy gain on **NyxQA**. This suggests the importance of further aligning retrieval models with downstream generative utility in URAG systems.

5.3 Embedding Capability Analysis

Although aligning mixed-modal retriever with downstream models can enhance the generative quality of the final VLM, the capability of the retriever itself is also of central importance. We evaluate the embedding ability of our models on the MMEB benchmark [20], and the results are reported in Table 2. In particular, *mmE5-Qwen2.5-3B* is trained on Qwen-2.5-VL-3B-Instruct, with its training data consisting of MMEB-labelled data together with the retrieval and VQA

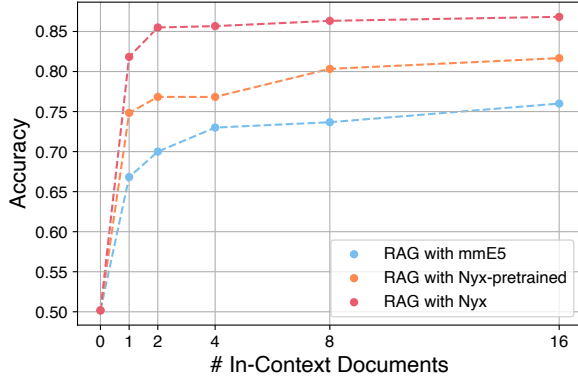


Figure 5: Impact of the number of in-context documents on NyxQA.

subsets of the mmE5 synthetic data (excluding classification subset to maintain alignment with Nyx-pretrained and Nyx settings), serving as the ablation setting.

Compared to mmE5, mmE5-Qwen2.5-3B performs worse across all capabilities, which can be attributed to its smaller backbone size and the exclusion of the classification subset from the mmE5 synthetic data. Nevertheless, it still surpasses most of the models. When OOD pure text data and NyxQA mixed-modal data are included, the mismatch with the MMEB evaluation pattern results in a 1.5% overall performance drop. However, after fine-tuning with feedback from a VLM on OOD datasets with entirely different tasks, Nyx outperforms mmE5-Qwen2.5-3B across all capabilities, achieving a 2.1% overall improvement. These findings further demonstrate that incorporating VLM feedback not only improves the performance of URAG systems but also enhances the embedding capability of dense retrievers themselves.

6 Quantitative Analysis

6.1 Impact of Data Scale on URAG Performance

The scalability of training data is crucial for building effective retrievers. Prior studies have shown an approximately logarithmic-linear relationship between the volume of training data and the quality of retrieval model embeddings [2, 12]. In this section, we further examine how the scale of training data affects Nyx’s performance in the URAG setting.

As illustrated in Figure 4, the performance trend closely follows a logarithmic-linear curve, consistent with previous findings. The steady improvement of URAG performance with increasing data scale further confirms the high quality and diversity of our training data. This indicates that enhancements in the retriever’s independent capabilities translate proportionally into gains in end-to-end URAG performance. Thus, increasing training data is expected to predictably enhance URAG scenario generalization.

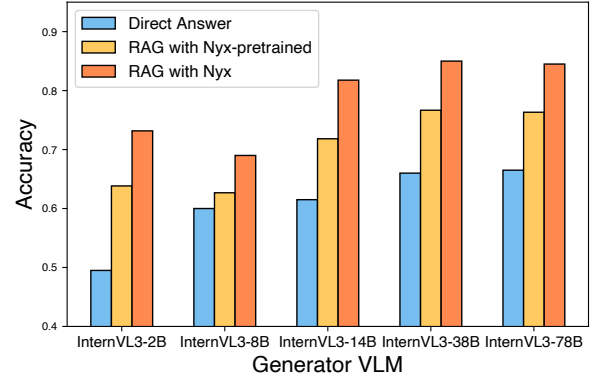


Figure 6: Performance of Nyx when paired with InternVL3 generators of different parameter scales on NyxQA.

6.2 Effect of Retrieved Document Count

To examine how the number of retrieved documents influences generation quality, we vary the number of documents fed into Qwen2.5-VL-7B from 0 to 16, evaluating the URAG results of mmE5, Nyx-pretrained, and Nyx. As shown in Figure 5, adding more documents consistently improves all retrievers, though the gains diminish as the count increases. Nyx consistently outperforms both Nyx-pretrained and mmE5, demonstrating robust performance even with fewer documents and confirming the effectiveness of feedback-based fine-tuning in producing informative retrievals. Overall, the results highlight the critical importance of high-quality top-ranked retrieval for efficient generation.

6.3 Generalization Across Generators

While Nyx is fine-tuned with supervision from Qwen2.5-VL-7B, we also examine whether such supervision generalizes to other VLMs. To assess this, we evaluate its performance across InternVL3 models of varying sizes used as generators. As shown in Figure 6, Nyx consistently outperforms the direct-answer baseline across all InternVL3 variants, indicating that supervision from Qwen2.5-VL-7B transfers effectively across different generator architectures. Integrating Nyx yields further improvements, particularly for InternVL3-2B and InternVL3-14B, with absolute gains exceeding 0.2 points.

However, the degree of improvement varies with generator size, indicating different alignment preferences among models. Additionally, since performance does not increase monotonically with model size, this suggests that generator size is not a reliable predictor of RAG pipeline performance. Effective alignment is crucial in bridging semantic gaps across VLMs.

6.4 Effect of MRL

In real-world retrieval systems, reducing embedding dimensions can significantly decrease memory usage and speed up retrieval. To adapt to varying resource budgets, we incorporate MRL into our contrastive training framework. MRL ensures the model maintains meaningful representations across reduced dimensions. We train

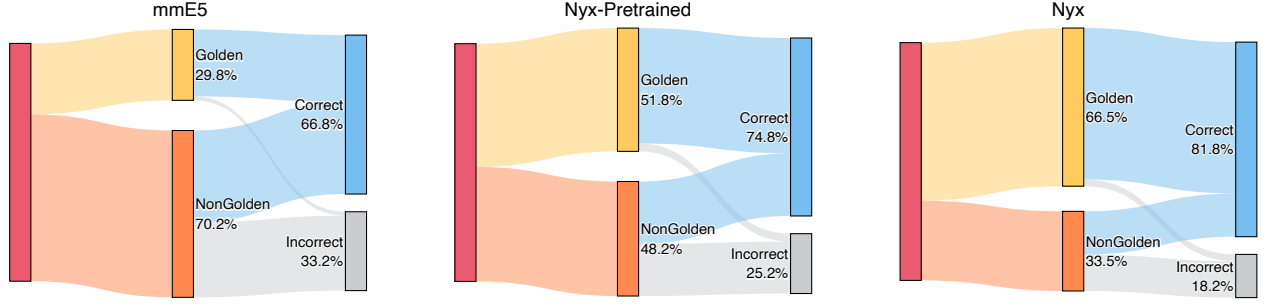


Figure 7: Comparison of retrieval and answer correctness distributions on NyxQA for mmE5, Nyx-pretrained, and Nyx.

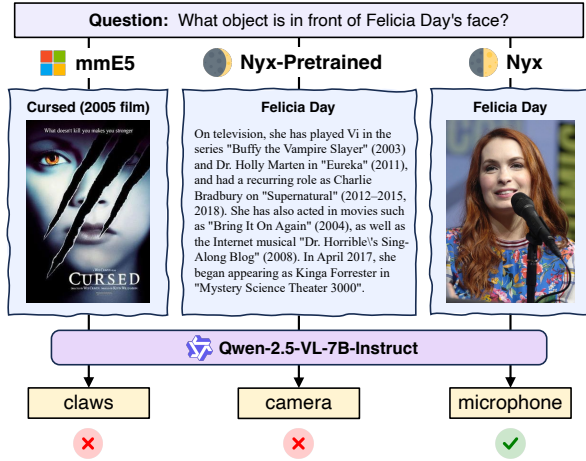


Figure 8: Case study on MMQA. The top-1 retrieved documents by mmE5, Nyx-pretrained, and Nyx are shown together with the corresponding answers produced by VLM.

the model to generate effective embeddings at four target dimensions: 2048, 1024, 512, and 256, with weights of [1.0, 1.0, 0.2, 0.2] (before normalization), where 2048 is the default VLM output.

As shown in Table 3, the 1024-dimensional variant achieves accuracy comparable to the 2048 one while halving storage, and even the 512- and 256-dimensional versions maintain strong performance. These results highlight MRL’s ability to provide efficient, resource-aware retrieval with graceful performance degradation under limited memory or latency budgets.

6.5 Impact of Retrieved Docs on Generation

To investigate the impact of retrieved documents on generation, we begin with a case study on MMQA. As shown in Figure 8, we examine how retrieval influences the produced answers. Unlike MMEB, which assumes modality-specific similarity computation, real-world retrieval entails cross-modal relevance estimation across multimodal documents. In this example, mmE5 retrieves an image-text pair focused solely on “face”, missing the query subject; **Nyx**-pretrained correctly identifies “Felicia Day” but provides the textual evidence that fails to support the answer; in contrast, **Nyx** retrieves

Table 3: Performance of Nyx on NyxQA under different output dimensions.

Setting	Output Embedding Dimension			
	2048-dim	1024-dim	512-dim	256-dim
Weight	1.0000	1.0000	0.2000	0.2000
Accuracy	0.8183	0.8100	0.7800	0.7467

Table 4: Impact of different retriever backbones on performance. Results are averaged over the six datasets reported in the main results.

Backbone VLM	Pretrained	w/ Feedback
InternVL3-2B	31.84	35.62
Qwen-2.5-VL-3B	33.27	36.46

the correct entity along with both the proper title and visual information, directly grounding the generated response.

Building on these qualitative observations, we further perform a quantitative analysis on the **NyxQA** dataset to study the relationship between retrieval correctness and answer correctness. The retrieval quality is visualized in Figure 7 using Sankey diagrams. The results reveal two key trends: (1) higher proportions of golden documents lead to higher answer accuracy; and (2) even with non-golden documents, nearly half of the answers remain correct, demonstrating the robustness of VLMs. These findings suggest that improving retrievers is crucial not only for ensuring faithful grounding but also for mitigating noise from irrelevant evidence. Future improvements may arise from modelling VLM preferences on non-golden evidence, which can sometimes diverge from human intuition.

6.6 Ablation on Backbone Model

To evaluate the architectural robustness of our framework, we replace the VLM backbone Qwen-2.5-VL-3B with InternVL3-2B while keeping the generator as Qwen-2.5-VL-7B constant. As illustrated in Table 4, InternVL3-2B achieves consistent performance gains of 3.78 following feedback-driven fine-tuning. However, it is observed that InternVL3-2B exhibits lower absolute performance compared to Qwen-2.5-VL-3B in both pretrained and fine-tuned settings. This

disparity can be primarily attributed to the smaller parameter scale of InternVL3-2B, alongside fundamental differences in their base VLM architectures and original pre-training data distributions. Nevertheless, these results underscore the robustness of our two-stage approach across varying model architectures.

7 Conclusion

To enable Universal Retrieval-Augmented Generation (URAG) over arbitrarily mixed-modal questions and corpora, we constructed **NyxQA**, the first large-scale and comprehensive dataset that faithfully reflected real-world URAG scenarios, where text, images, and their interleaved combinations naturally coexisted. Building on this foundation, we introduced **Nyx**, a unified multimodal retriever explicitly optimized for such settings. **Nyx** was initially pretrained via contrastive learning with Matryoshka Representation Learning on a diverse mixture of public and synthetic data, and was subsequently fine-tuned using feedback from a downstream vision-language generator, thereby better aligning retrieval relevance with generation utility. Extensive experiments demonstrated that this simple yet effective pipeline achieved consistent and substantial improvements over both unimodal and multimodal baselines across all modality combinations, underscoring the promise of unified mixed-modal retrieval for next-generation URAG systems.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Nos. 62272467 and 625B2178). The work was partially done at the Beijing Key Laboratory of Research on Large Models and Intelligent Governance and the Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. *CoRR* abs/2502.13923 (2025). arXiv:2502.13923 doi:10.48550/ARXIV.2502.13923
- [2] Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. 2025. mmE5: Improving Multimodal Multilingual Embeddings via High-quality Synthetic Data. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025 (Findings of ACL)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 8254–8275. <https://aclanthology.org/2025.findings-acl.433/>
- [3] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *CoRR* abs/2402.03216 (2024). arXiv:2402.03216 doi:10.48550/ARXIV.2402.03216
- [4] Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. 2025. ReSearch: Learning to Reason with Search for LLMs via Reinforcement Learning. *CoRR* abs/2503.19470 (2025). arXiv:2503.19470 doi:10.48550/ARXIV.2503.19470
- [5] Wenhui Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. 2022. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7–11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 5558–5570. doi:10.18653/V1/2022.EMNLP-MAIN.375
- [6] Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. 2024. MLLM Is a Strong Reranker: Advancing Multimodal Retrieval-augmented Generation via Knowledge-enhanced Reranking and Noise-injected Training. *CoRR* abs/2407.21439 (2024). arXiv:2407.21439 doi:10.48550/ARXIV.2407.21439
- [7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17–24, 2023*. IEEE, 2818–2829. doi:10.1109/CVPR52729.2023.00276
- [8] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fugong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiali Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiusi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhua Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shutong Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, and Wangding Zeng. 2024. DeepSeek-V3 Technical Report. *CoRR* abs/2412.19437 (2024). arXiv:2412.19437 doi:10.48550/ARXIV.2412.19437
- [9] Guanting Dong, Chenghao Zhang, Mengjie Deng, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. 2025. Progressive Multimodal Reasoning via Active Retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 3579–3602. <https://aclanthology.org/2025.acl-long.180/>
- [10] Guanting Dong, Yutao Zhu, Chenghao Zhang, Zechen Wang, Ji-Rong Wen, and Zhicheng Dou. 2025. Understand What LLM Needs: Dual Preference Alignment for Retrieval-Augmented Generation. In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, Guodong Long, Michale Blumstein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (Eds.). ACM, 4206–4225. doi:10.1145/3696410.3714717
- [11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Baptiste Rozière, Bethany Bion, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiohu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelfer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The Llama 3 Herd of Models. *CoRR* abs/2407.21783 (2024). arXiv:2407.21783 doi:10.48550/ARXIV.2407.21783
- [12] Yan Fang, Jingtao Zhan, Qingyao Ai, Jiaxin Mao, Weihang Su, Jia Chen, and Yiqun Liu. 2024. Scaling Laws For Dense Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (Washington DC, USA) (SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1339–1349. doi:10.1145/3626772.3657743
- [13] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. ColPali: Efficient Document Retrieval with Vision Language Models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025*. OpenReview.net. <https://openreview.net/forum?id=ogjBpZ8uSi>
- [14] Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025. WebWatcher: Breaking New Frontier of Vision-Language Deep Research Agent. *CoRR* abs/2508.05748 (2025). arXiv:2508.05748 doi:10.48550/ARXIV.2508.05748
- [15] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *Proceedings of the 28th International Conference on Computational*

- Linguistics, COLING 2020, Barcelona, Spain (Online), December 8–13, 2020, Doria Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 6609–6625. doi:10.18653/V1/2020.COLING-MAIN.580
- [16] 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. <https://openreview.net/forum?id=nZeVKeeFYf9>
- [17] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Trans. Mach. Learn. Res.* 2022 (2022). <https://openreview.net/forum?id=jKNpXi7b0>
- [18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139), Marina Meila and Tong Zhang (Eds.)*. PMLR, 4904–4916. <http://proceedings.mlr.press/v139/jia21b.html>
- [19] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. E5-V: Universal Embeddings with Multimodal Large Language Models. *CoRR* abs/2407.12580 (2024). arXiv:2407.12580 doi:10.48550/ARXIV.2407.12580
- [20] Ziyang Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. 2025. VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025*. OpenReview.net. <https://openreview.net/forum?id=TE0KozWYAF>
- [21] Jiajie Jin, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. 2025. FlashRAG: A Modular Toolkit for Efficient Retrieval-Augmented Generation Research. In *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025 - 2 May 2025*, Guodong Long, Michale Blumstein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (Eds.). ACM, 737–740. doi:10.1145/3701716.3715313
- [22] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick 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*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 6769–6781. doi:10.18653/V1/2020.EMNLP-MAIN.550
- [23] Andreas Koukounas, Georgios Mastrapas, Michael Günther, Bo Wang, Scott Martens, Isabelle Mohr, Saba Sturua, Mohammad Kalim Akram, Joan Fontanals Martínez, Saahil Ognawala, Susana Guzman, Maximilian Werk, Nan Wang, and Han Xiao. 2024. Jina CLIP: Your CLIP Model Is Also Your Text Retriever. *CoRR* abs/2405.20204 (2024). arXiv:2405.20204 doi:10.48550/ARXIV.2405.20204
- [24] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham M. Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka Representation Learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/c32319f4868da7613d78af999310e42-Abstract-Conference.html
- [25] Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. 2025. LLaVE: Large Language and Vision Embedding Models with Hardness-Weighted Contrastive Learning. *CoRR* abs/2503.04812 (2025). arXiv:2503.04812 doi:10.48550/ARXIV.2503.04812
- [26] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10–16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/e2cfb719f58585f779d0a4f907bd618-Abstract-Datasets_and_Benchmarks.html
- [27] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- [28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICML 2023, 23–29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 19730–19742. <https://proceedings.mlr.press/v202/li23q.html>
- [29] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguistics* 12 (2024), 157–173. doi:10.1162/TACL_A_00638
- [30] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html
- [31] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6–10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 5687–5711. doi:10.18653/V1/2023.FINDINGS-EMNLP.378
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139), Marina Meila and Tong Zhang (Eds.)*. PMLR, 8748–8763. <http://proceedings.mlr.press/v139/radford21a.html>
- [33] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. ZeRO: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (Atlanta, Georgia) (SC '20)*. IEEE Press, Article 20, 16 pages.
- [34] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (Virtual Event, CA, USA) (KDD '20)*. Association for Computing Machinery, New York, NY, USA, 3505–3506. doi:10.1145/3394486.3406703
- [35] Monica Riedler and Stefan Langer. 2024. Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications. *CoRR* abs/2410.21943 (2024). arXiv:2410.21943 doi:10.48550/ARXIV.2410.21943
- [36] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-Augmented Black-Box Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16–21, 2024*, Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (Eds.). Association for Computational Linguistics, 8371–8384. doi:10.18653/V1/2024.NAACL-LONG.463
- [37] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. MultiModalQA: complex question answering over text, tables and images. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*. OpenReview.net. <https://openreview.net/forum?id=ee6W5UgQLa>
- [38] Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kiyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2025. VDocRAG: Retrieval-Augmented Generation over Visually-Rich Documents. In *CVPR*.
- [39] Raghuveer Thirukovalluru, Rui Meng, Ye Liu, Karthikeyan K, Mingyi Su, Ping Nie, Semih Yavuz, Yingbo Zhou, Wenhui Chen, and Bhuwan Dhingra. 2025. Breaking the Batch Barrier (B3) of Contrastive Learning via Smart Batch Mining. *CoRR* abs/2505.11293 (2025). arXiv:2505.11293 doi:10.48550/ARXIV.2505.11293
- [40] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Trans. Assoc. Comput. Linguistics* 10 (2022), 539–554. doi:10.1162/TACL_A_00475
- [41] 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. *CoRR* abs/2212.03533 (2022). arXiv:2212.03533 doi:10.48550/ARXIV.2212.03533
- [42] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhui Chen. 2024. UniLR: 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 (Lecture Notes in Computer Science, Vol. 15145)*, Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer, 387–404. doi:10.1007/978-3-031-73021-4_23

- [43] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *CoRR* abs/2505.09388 (2025). arXiv:2505.09388 doi:10.48550/ARXIV.2505.09388
- [44] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 2369–2380. doi:10.18653/V1/D18-1259
- [45] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net. <https://openreview.net/forum?id=zG459X3Xge>
- [46] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. 2021. Florence: A New Foundation Model for Computer Vision. *CoRR* abs/2111.11432 (2021). arXiv:2111.11432 <https://arxiv.org/abs/2111.11432>
- [47] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhui Chen, Yu Su, and Ming-Wei Chang. 2024. 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.
- [48] Xin Zhang, Ziqi Dai, Yongqi Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, Jun Yu, Wenjie Li, and Min Zhang. 2025. Towards Text-Image Interleaved Retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 4254–4269. <https://aclanthology.org/2025.acl-long.214/>
- [49] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. GME: Improving Universal Multimodal Retrieval by Multimodal LLMs. *CoRR* abs/2412.16855 (2024). arXiv:2412.16855 doi:10.48550/ARXIV.2412.16855
- [50] Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, 3185–3200. doi:10.18653/V1/2024.ACL-LONG.175
- [51] Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yuezhe Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. 2025. MegaPairs: Massive Data Synthesis for Universal Multimodal Retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 19076–19095. <https://aclanthology.org/2025.acl-long.935/>
- [52] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingting Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhui Wang. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *CoRR* abs/2504.10479 (2025). arXiv:2504.10479 <https://doi.org/10.48550/arXiv.2504.10479>