

Towards Text-Image Interleaved Retrieval

Xin Zhang^{1,2*}, Ziqi Dai^{1*}, Yongqi Li², Yanzhao Zhang, Dingkun Long
Pengjun Xie, Meishan Zhang^{1†}, Jun Yu¹, Wenjie Li², Min Zhang¹

¹Harbin Institute of Technology, Shenzhen ²The Hong Kong Polytechnic University
{zhangxin2023, ziqi.dai}@stu.hit.edu.cn zhangmeishan@hit.edu.cn

Release at <https://github.com/vec-ai/wikiHow-TIIR>

Abstract

Current multimodal information retrieval studies mainly focus on single-image inputs, which limits real-world applications involving multiple images and text-image interleaved content. In this work, we introduce the text-image interleaved retrieval (TIIR) task, where the query and document are interleaved text-image sequences, and the model is required to understand the semantics from the interleaved context for effective retrieval. We construct a TIIR benchmark based on naturally interleaved wikiHow tutorials, where a specific pipeline is designed to generate interleaved queries. To explore the task, we adapt several off-the-shelf retrievers and build a dense baseline by interleaved multimodal large language model (MLLM). We then propose a novel Matryoshka Multimodal Embedder (MME), which compresses the number of visual tokens at different granularity, to address the challenge of excessive visual tokens in MLLM-based TIIR models. Experiments demonstrate that simple adaptation of existing models does not consistently yield effective results. Our MME achieves significant improvements over the baseline by substantially fewer visual tokens. We provide extensive analysis and will release the dataset and code to facilitate future research.

1 Introduction

Multimodal information retrieval (MIR) aims to retrieve relevant information involving multiple modalities (Wei et al., 2024), which plays a crucial role in various applications such as e-commerce search (Wu et al., 2021) and retrieval augmented generation (RAG) systems (Chen et al., 2022; Yasunaga et al., 2023). Current advanced multimodal retrievers (Zhou et al., 2024a; Lin et al., 2025) typically adopt the dense retrieval paradigm, where queries or documents are encoded into embeddings for vector similarity calculation. These models

*Equal contribution. †Corresponding author.

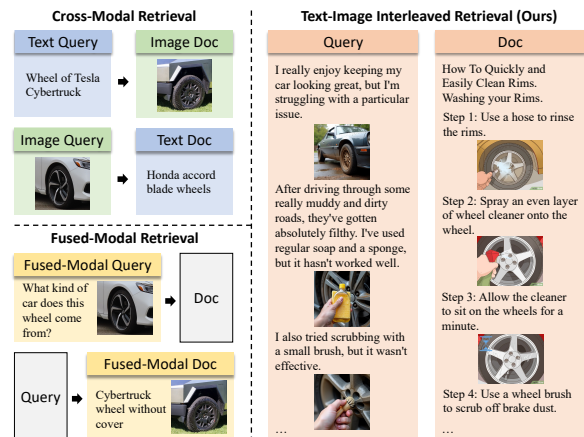


Figure 1: Comparison of our Text-Image Interleaved Retrieval task to previous settings. Blocks with black borders represent data in text, image or fused-modal.

have demonstrated promising results in scenarios involving cross-modal and fused-modal retrieval (Figure 1 left illustrates the settings).

Despite their effectiveness, most existing MIR studies permit only a single image in the query or document (Zhou et al., 2024a; Wei et al., 2024). This would largely limit users to clearly present their information needs and requirements, while also restricting the system from leveraging useful documents containing multiple images and interleaved text-image contents. For example, a tutorial for everyday skills, such as furniture assembly or cooking recipes, typically requires multiple illustrations to describe sequential steps (Figure 1 right). Similarly, users may need more than one photo to effectively describe their current problems or situations. Such cases would be inevitable in real-world RAG systems, demonstrating the necessity of interleaved-modal inputs in retrieval.

To explore the above scenarios, we introduce the text-image interleaved retrieval (TIIR) task, where both the query and document contain interleaved text and images (Figure 1 right). In TIIR, multiple text chunks and images are sequentially positioned

in a semantic manner, allowing for a more accurate expression of user intent and document information. However, this also presents challenges in understanding interleaved-modal content.

To advance the progress of TIIR, we first construct a new benchmark based on wikiHow¹, a large-scale collection of human-curated how-to guides with text and images (Yang et al., 2021). We convert the tutorial articles into a retrieval corpus of 150K interleaved documents. To obtain interleaved contextual queries, we design a novel and efficient pipeline that leverages powerful large language models (LLMs) (Laurençon et al., 2024; Yang et al., 2024) and a text-to-image generator (Labs, 2023) to automatically generate interleaved queries (§2.2) based on documents. We then employ human experts to annotate and filter out generation artifacts, resulting in 7,654 high-quality query-document pairs for testing, while the remaining generated queries are allocated to the training set. We dub this dataset as wikiHow-TIIR.

Beyond the data, building an effective TIIR model is complex due to the challenges in modeling interleaved-modal content. First, existing retrievers struggle to handle this task effectively due to their single-image constraints. Second, while fine-tuning multimodal LLMs (MLLMs) with interleaved inputs support (Lu et al., 2024) as dense TIIR models seems promising, the hundreds of visual tokens required per image (Yin et al., 2024) leads to prohibitively long sequences, resulting in both computational inefficiency and disproportionate visual dominance in the embedding space (§4.4). To address these issues, we propose a novel retriever, *i.e.*, Matryoshka Multimodal Embedder (MME), that compresses the number of visual tokens at different granularity (§3), thereby generating more effective embeddings for TIIR.

We conduct extensive experiments to explore our dataset and evaluate different retrievers (§4). Results show that the interleaved context is the key of TIIR modeling. Even with specialized adaption strategies, existing retrievers (non-interleaved) perform worse than the native-interleaved baseline, indicating the necessity of developing dedicated TIIR retrievers. In contrast, our suggested MME outperform the baseline by a large margin, demonstrating the effectiveness of our approach. We further conduct comprehensive analyses (§4.4) to understand the TIIR task and models.

¹<https://www.wikihow.com>.

Our contributions are as follows:

- We identify the text-image interleaved retrieval (TIIR) task and construct the wikiHow-TIIR benchmark. To the best of our knowledge, it is the first dataset for TIIR.
- We propose a novel TIIR model that compresses the number of visual tokens at different granularity, which successfully addresses the challenge in modeling interleaved content.
- We present extensive experiments and analyses on our dataset, including strategies for adapting existing retrievers. This provides insights for future work and applications.

2 WikiHow-TIIR Benchmark

2.1 Task Definition

We first define the text-image interleaved data instance X as a sequence of text and images, $X = [x_i, \dots, x_n]$, where x_i can be either a text chunk or an image, all of which are ordered contextually. Given a query X^Q and a corpus C consisting of documents $\{X_1^D, \dots, X_m^D\}$, the TIIR task is to retrieve the most relevant document X^D from C for X^Q . The relevance is determined by a similarity function $f(X^Q, X^D)$, which measures the semantic similarity at the image-text sequence level. The model is required to understand the semantics from contextually interleaved text and images for effective retrieval, which could be challenging to existing multimodal retrievers.

2.2 Data Construction

One of the most common scenarios involving interleaved text and images in everyday life is found in tutorials for daily skills or product manuals, where images are necessary to provide clearer and more vivid descriptions. WikiHow¹ is a widely used tutorial website that contains a large number of high-quality text-image tutorials that meet these criteria. Therefore, we choose wikiHow articles crawled by Yang et al. (2021) as our *retrieval corpus*. For each tutorial, we select the goal, step titles and corresponding images to build an interleaved document. We then generate and annotate queries.

Query Generation We design a query generation pipeline to mimic real-world scenarios where users may provide multiple images and text to describe their problems or situations. Given that current interleaved MLLMs are not yet sufficiently capable of handling complex text and image generation,

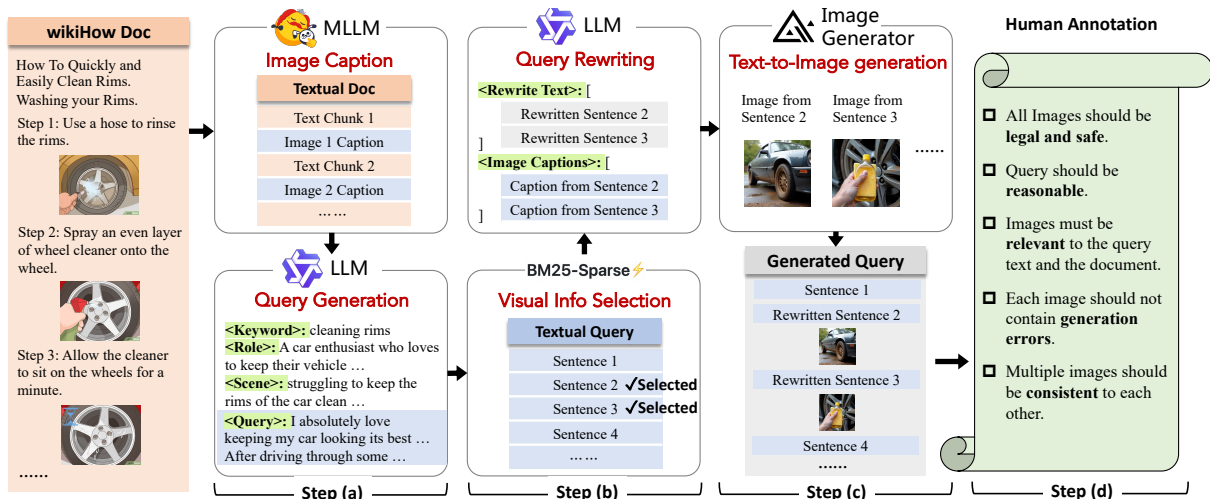


Figure 2: Our data construction workflow (§2.2), where step (a), (b) and (c) comprise the generation pipeline, and (d) shows the brief annotation guideline. Technical details and principles are provided in Appendix A.2 and A.3.

Part	#Examples	Avg./Min/Max #Images	Avg. Text #Tokens	#Pos.
Corpus	155,262	4.97 / 2 / 64	85.62	-
Train Query	73,084	2.88 / 2 / 4	105.15	1
Test Query	7,654	2.81 / 2 / 4	105.59	1

Table 1: Statistics of our constructed wikiHow-TIIR dataset, where Pos. denotes positive document. We count tokens by LLaMA tokenizer.

our pipeline centers on the text modality. It leverages image captioning and text-to-image generation for modality conversion, while utilizing more advanced LLMs to drive query text generation. As shown in Figure 2, it consists of three stages:

(a) *Query text generation*. Given an interleaved document X^D , we first generate captions for each image by a strong and efficient MLLM² (Laurençon et al., 2024). Then, based on the tutorial text and image captions, we instruct a powerful LLM³ (Yang et al., 2024) to write a text query T^Q target to one specific step of the document. We design a detailed in-context learning style prompt (Appendix Figure 9) to guide the query generation and minimize potential hallucinations.

(b) *Text-image information reorganization*. We split the query text into sentences and employ BM25 (Robertson et al., 2009) to identify the most informative ones⁴ $S_{\text{top-k}}$ for transforming the textual information into images. Next, we use the

²hf.co/HuggingFaceM4/Idfics3-8B-Llama3

³hf.co/Qwen/Qwen2.5-72B-Instruct

⁴We compute BM25 scores between query sentences and the positive document and take the top ones.

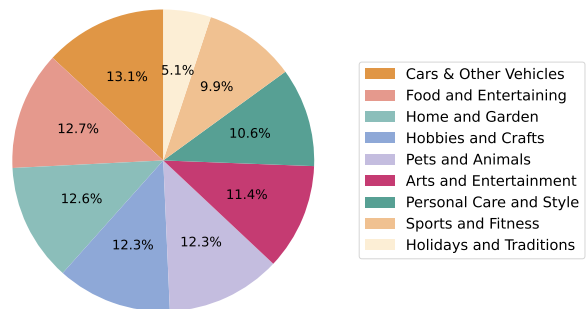


Figure 3: Categories of wikiHow-TIIR testset queries.

LLM to select entities or actions from $S_{\text{top-k}}$ to generate captions C^Q for images in query and rewrite the query text into T_r^Q to remove selected information. The prompt is in Appendix Figure 10. We instruct the LLM to generate scene descriptions using objective, third-person narratives and ensure the consistency to the query text.

(c) *Image generation*. We use a text-to-image generator⁵ (Labs, 2023) to generate images from image captions C^Q and merge with the rewritten query T_r^Q to form the final query X^Q .

Detailed design principles and implementations of our pipeline are provided in Appendix A.2.

We select around 80.7k tutorials from the corpus and generate one query for each tutorial. As the generated query is designed to be relevant to the corresponding tutorial, we take the tutorial as the *positive* document for the query.

Testset Annotation To build a high-quality testset for fair and reasonable evaluation, we further conduct a human annotation process to filter

⁵hf.co/black-forest-labs/FLUX.1-dev

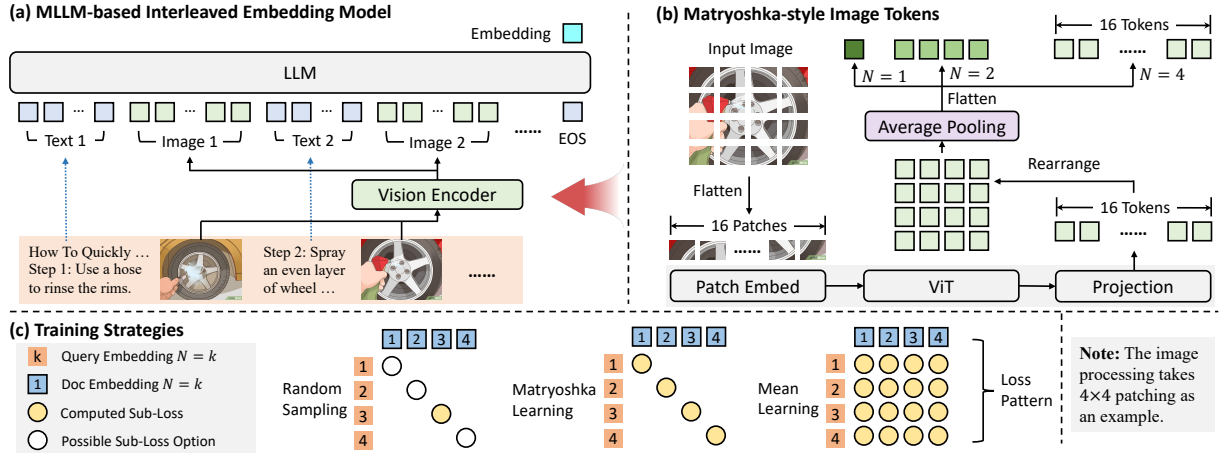


Figure 4: Our TIIR model overview, where (a) is the DPR baseline (§3.1), (b) illustrates the computation of visual tokens in different granularities, and (c) shows the training strategies of our MME.

out generation artifacts and ensure the generated queries are reasonable and contextually interleaved. We employ 10 graduate students in computer science as annotators. Our annotation guidelines primarily focus on five types of issues: **(1)** Images must not involve illegal content, sensitive topics, or contain offensive material such as pornography. **(2)** The overall content of the query should be reasonable and consistent with common sense. **(3)** Images must be relevant to the query text and the document. **(4)** Each image in the query should not contain obvious structural or textual errors. **(5)** If multiple images in the query depict the same subject or scene, they should not exhibit significant variations. The details refer to §A.3. We select around 10,000 query-document pairs with diverse wikiHow topic labels for annotation, resulting in 7,654 high-quality pairs as the final testset.

2.3 Data Statistics

Table 1 shows the statistics of the wikiHow-TIIR dataset. From all generated queries, we annotate 7,654 query-positive pairs as the testset, and the remaining 73,084 pairs are used as the trainset. Figure 3 illustrates the category distribution in our test set, which covers nine real-life domains: Vehicles, Food, Home Improvement, Crafts, Animals, Arts, Personal Care, Fitness, and Traditions.

3 Approach

3.1 Baseline Model

Our baseline is in the dense retrieval paradigm, where inputs are encoded by a backbone and a pooling operator is applied to obtain the sequence-level embeddings. To effectively model the seman-

tics of interleaved context, the interleaved MLLM is a natural backbone choice as the order of text chunks and images are kept in the input sequence and thus attention operations can better capture the multimodal interactions. In practice, we use the DeepSeek-VL (Lu et al., 2024) as the backbone and take [EOS] output state as the embedding.

We train it by InfoNCE (Oord et al., 2018) loss:

$$\mathcal{L} = -\log \frac{\exp(s(X^Q, X_+^D)/\tau)}{\sum_{i=1}^N \exp(s(X^Q, X_i^D)/\tau)}, \quad (1)$$

where τ denotes the temperature parameter. The X_+^D is the relevant document (positive) to X^Q , while others are irrelevant documents (negatives), which could be either hard negatives or in-batch negatives. $s(X^Q, X^D)$ is the relevance score between X^Q and X^D , measured by the cosine similarity between their respective embeddings.

3.2 Matryoshka Multimodal Embedder

Current MLLMs utilize Vision Transformers (ViTs) to encode images and a linear projection to convert into visual tokens, which are then concatenated with text tokens to form the input of the LLM backbone. As most models use a large number of visual tokens (e.g., 576) for each image, a substantial number of images from interleaved data could take excessive visual tokens, leading to great inefficiency and allowing visual information to disproportionately dominate the embedding space. Moreover, the token sequence will be truncated if it's too long to exceed the model's max context length, which may lose critical semantics for retrieval. Inspired by Cai et al. (2024), we introduce a novel Matryoshka Multimodal Embedder (MME) to address

No.	Setting	Model	#Param	Recall@5	MRR@5	MRR@10	nDCG@5	nDCG@10
<i>Non-Interleaved Models</i>								
1	Text w/ Merged Image	VISTA	0.21B	45.06	31.95	33.14	35.22	38.09
2		GME _{Qwen2-VL-2B}	2.21B	65.85	50.12	51.18	54.06	56.64
3		E5-V	8.36B	62.66	46.47	47.64	50.52	53.32
4		MM-Embed	8.18B	68.73	52.24	53.25	56.37	58.79
5		GME _{Qwen2-VL-7B}	8.29B	70.05	54.24	55.21	58.21	60.52
6	Text w/ Caption	BGE-v1.5 _{large}	0.34B	39.66	27.54	28.58	30.56	33.08
7		GTE-v1.5 _{large}	0.43B	41.44	27.74	28.94	31.14	34.06
8		GTE-Qwen2-1.5B	1.78B	47.24	33.40	34.63	36.85	39.80
9	Vector-Fusion	Jina-CLIP-v2	0.87B	58.80	43.29	44.44	47.17	49.97
10		CLIP _{large} Fine-tuned	0.43B	69.41	53.06	54.25	57.15	60.01
<i>Native Interleaved Models (Fine-tuned)</i>								
11	TIIR	DPR _{DeepSeek-VL}	1.98B	74.79	59.43	60.49	63.28	65.80
12		MME (Ours) _{<i>N=3</i>}		77.40	62.07	63.01	65.91	68.16

Table 2: Evaluation results on our wikiHow-TIIR. Text w/ Merged Image denotes the interleaved sequence is concatenated into a text-image pair. The description of Vector-Fusion is in §4.1.

these issues. MME produces a nested set of visual tokens for each image, which is a Matryoshka doll-like sequence across multiple coarse-to-fine granularities (Figure 4). At the inference time, we could set a certain token size to meet the requirement, which would be more flexible and efficient.

Technically, we introduce an average pooling after the visual projection of MLLM to compress the visual tokens into different lengths by different-sized pooling kernels. We take DeepSeek-VL-1.3B as an example. Its vision encoder⁶ divides an image into 24×24 patches (i.e., 576 in total) and outputs 576 visual features, which are then projected into visual tokens. We rearrange the visual tokens into a 24×24 grid and apply average pooling with kernel size $24/N$ to compress into $N \times N$ grid, resulting in flattened N^2 visual tokens. $N \in \{1, 2, 3, 4, 6, 8, 12, 24\}$.

In training, we propose three strategies to learn the nested visual tokens: **(1) Random sampling (Rand)**: We randomly sample a grid width N for each micro-batch, which is a simple and efficient way for the model to adapt inputs at different levels of granularity. **(2) Matryoshka learning (MRL)**: We train the model with all M kernel sizes simultaneously, where the model is trained with a weighted sum of M losses from different grid sizes. **(3) Mean learning (Mean)**: Similar to MRL, but we additionally compute losses between query and document embeddings of different sizes, the final loss is the mean of all $M \times M$ possible combinations.

⁶hf.co/timm/ViT-L-16-SigLIP-384

4 Experiments

4.1 Evaluated Models

Besides the DPR_{DeepSeek-VL} baseline (§3.1), we also adapt non-interleaved retrievers for evaluation:

- Single-image multimodal models, *i.e.*, VISTA (Zhou et al., 2024a), E5-V (Jiang et al., 2024), MM-Embed (Lin et al., 2025) and GME (Zhang et al., 2025b), where we concatenate multiple images into one (Appendix Figure 13 shows an example).

- Text models, *i.e.*, BGE (Xiao et al., 2024) and GTE (Zhang et al., 2024). We evaluate them by replacing images with text captions from a MLLM⁷ (details refer to Appendix §C.2).

- CLIP-style two-stream models, we evaluate the well-trained Jina-CLIP⁸ (Koukounas et al., 2024) and fine-tuned original CLIP (Radford et al., 2021) by a simple vector-fusion strategy. Given an input sequence, we concatenate all text chunks and encode as one text embedding \mathbf{t} , while all images are separately encoded as image embeddings $\{i_1, \dots, i_n\}$. The final embedding \mathbf{e} is the normalized sum of the normalized average embedding of images and the text embedding, *i.e.*, $\mathbf{e} = \text{Norm}(\text{Norm}(\text{Mean}(i_1, \dots, i_n)) + \mathbf{t})$.

4.2 Settings

Metrics We compute Recall@ k (the rate that positives are successfully retrieved within the top- k ranked results), MRR@ k (Mean Reciprocal Rank,

⁷hf.co/Qwen/Qwen2-VL-2B-Instruct

⁸hf.co/jinaai/jina-clip-v2

the average of reciprocal ranks of the first positive in the top- k) and $nDCG@k$ (normalized Discounted Cumulative Gain, the quality of ranking by considering both the relevance and position of positives within top- k) on our testset for evaluation.

Implementation We fine-tune OpenAI CLIP⁹ and DeepSeek-VL-1.3B¹⁰ ($DPR_{DeepSeek-VL}$ and our MME). We use a batch size of 32 and a learning rate of 5×10^{-5} with a linear warm-up scheduler to train the models for 3 epochs. The contrastive learning temperature τ is set to 0.05. We use in-batch negatives and 1 randomly selected hard negative. Other details are provided in Appendix §B.

4.3 Main Results

Table 2 presents the results on our wikiHow-TIIR benchmark. First, we focus on the evaluation of adapted non-interleaved models. For the single-image multimodal retrievers (setting Text w/ Merge Image in Table 2), by combining multiple images into one image, they could achieve reasonable performance. From VISTA to GME and then to MM-Embed, The scaling of the model size could bring consistent improvements. While E5-V appears to be an outlier with suboptimal performance, this is understandable given that it was trained solely on textual natural language inference data (Jiang et al., 2024), without exposure to either retrieval or multimodal data. It is remarkable to observe that foundation MLLMs can demonstrate such comparable performance. By replacing images with captions (setting Text w/ Caption), the text retrievers at different sizes perform worse than their similar-sized multimodal models, *e.g.*, BGE is worse than VISTA. This is because captions may not fully preserve the visual semantics (as we will analyze in Table 3). Regarding two-stream models, the vector-fusion strategy allows well-finetuned Jina-CLIP (Koukounas et al., 2024) to be directly adapted, achieving promising performance.

For native interleaved models, we can observe that: (1) The DPR baseline (row 11) performs better than fine-tuned CLIP (row 10), demonstrating the interleaved modeling provides a more accurate context understanding for TIIR; (2) Our proposed MME (row 12) further improves the performance by a large margin, indicating the effectiveness of our Matryoshka-style visual token learning.

⁹hf.co/openai/clip-vit-large-patch14

¹⁰hf.co/deepseek-ai/deepseek-vl-1.3b-base

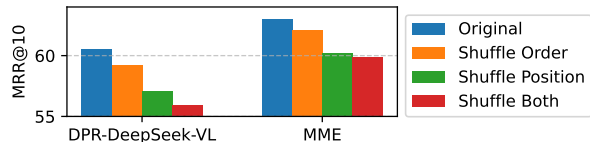


Figure 5: Results of interleaved models evaluated on settings of original data, shuffled image ordering, shuffled image position, and shuffled image ordering & position.

In summary, all adapted models are underperformed by the native interleaved models, which calls for developing TIIR support in future multimodal retrievers. It is also worth noting that, to ensure fair comparison to a reasonable extent, we do not fine-tune any off-the-shelf retrievers, and the fine-tuned models are initialized from weak checkpoints (models that have not been trained on any high-quality retrieval data).

4.4 Analysis

This subsection presents several in-depth analyses to understand the TIIR task and models. We address the following five research questions.

RQ1: Can the interleaved context be effectively modeled? Fig. 5 Given that text-image interleaved context lies at the core of our task, a natural question arises regarding its importance for retrieval. We examine this by manipulating the images in several ways: (1) shuffling the image ordering, (2) shuffling the image position, and (3) shuffling both image ordering and position. To ensure rigorous evaluation of these settings and isolate other potential confounding factors, we only evaluate the native interleaved models. Figure 5 demonstrates that shuffling both image ordering and position leads to significant performance degradation, indicating that both the order among images and the alignment between images and text affect the context semantics. Combining both settings further decreases the result. In summary, the performance drop empirically demonstrates that the interleaved context is effectively modeled and crucial for accurate retrieval.

RQ2: Are the off-the-shelf models adaptation strategies (§4.1) effective? Tab. 3 After recognizing the importance of interleaved context, we further evaluate the effectiveness of the adaptation strategies (§4.1) for off-the-shelf models. A direct probing to this question is hard to achieve, as they are not designed for the TIIR task. Fortunately, an elegant solution emerges: since all these models are

No.	Original Setting	Model	MRR@10	
			Original	Text
1		VISTA	33.14	40.86
2	Text w/ Merged Image	GME _{Qwen2-VL-2B}	51.18	42.80
3		E5-V	47.64	43.23
4		MM-Embed	53.25	53.11
5		BGE-v1.5 _{large}	28.58	44.10
6	Text w/ Caption	GTE-v1.5 _{large}	28.94	43.89
7		GTE-Qwen2-1.5B	34.63	46.17
8	Vector-Fusion	Jina-CLIP-v2	44.44	39.28
9	Visual Doc	GME _{Qwen2-VL-2B}	50.79	42.80

Table 3: Comparison of performance between original adaption and text-only evaluation (ignoring images). The adaption strategy could be considered as useful if text results are lower than the original.

proven to be powerful text retrievers, we could investigate this question by comparing their adapted performance against their text-only retrieval scores. Table 3 presents the results. We observe that for single-image multimodal retrievers, the adaption of merging multiple images into one does not always succeed. We suppose that the merged image (as the example in Figure 13) not only loses the interleaved context but also introduces noise in content understanding. The image caption strategy for text retrievers actually decreases the performance, which could be due to the fact that the generated captions are not as informative as the original images. Notably, the vector-fusion strategy improves the performance, which could be attributed to the feature-level fusion of text and images. Nonetheless, we suppose that these failures stem from the loss of interleaved data structure. Effectively preserving this interleaved context is crucial for enabling existing models to support TIIR.

RQ3: Can we model the interleaved context in the vision modality? Tab. 3 All adaptations in §4.1 preserve the original text information. For vision modality, a promising recent paradigm in multimodal retrieval is based on visual documents (Ma et al., 2024; Faysse et al., 2025), which takes screenshots of multimodal documents as input. Among evaluated models, GME (Zhang et al., 2025b) supports this mode. To explore its potential, we convert interleaved sequences into visual docs (as shown in Appendix Figure 14) for evaluation. The last row of Table 3 shows the results. Interestingly, this adaptation is also effective (*i.e.*, the adapted scores are higher than that of text-only) as it maintains the interleaved information structure.

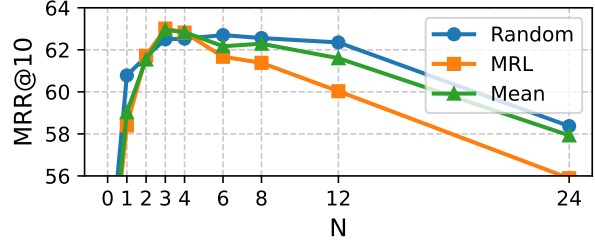


Figure 6: Performance curve of different settings of Matryoshka-style visual token, where all three different training strategies (§3.2) are presented. The best one (mean) is selected as the final model.

RQ4: Understanding the Matryoshka-style visual token. Fig. 6 & 7 Now we focus on the proposed MME model. In Table 2, for brevity, we only report the results of $N = 3$ of the best training strategy. To better understand the behavior, we display the performance curve of different visual token settings in Figure 6. We can see that, for all three training strategies, retrieval performance exhibits an inverted U-shaped relationship with the number of visual tokens, initially improving before declining. The observed pattern aligns well with the intuition: an insufficient number of visual tokens fails to capture the rich semantics of each image, while excessive tokens dominate the input sequence, leading to semantic bias in the embeddings as well as inaccurate retrieval results. This highlights the importance of compressing visual tokens for multiple images and interleaved retrieval models. In addition, all strategies reach the peak performance at $N = 3$, which implies the best visual token size is dataset/domain dependent. We further investigate the visual information dominance by calculating the normalization between distances of an embedding and both text-only embeddings (d_t) and full image tokens embeddings (d_i), as $(d_i - d_t)/(d_i + d_t)$, as plotted in Figure 7. The distribution aligns with the performance curve, where the optimal $N = 3$ yields a more balanced distribution, indicating a more effective balance between text information and visual influence.

RQ5: Encoding efficiency of MME. Tab. 4 The Matryoshka-style visual token also brings an enhancement in encoding efficiency, reducing the computational overhead of the large LLM backbone (Cai et al., 2024). To quantify the gain, we randomly select 1000 query-document pairs from the testset and measure the average sequence length, encoding time, and maximum batch size for different settings. Table 4 shows the results. In our

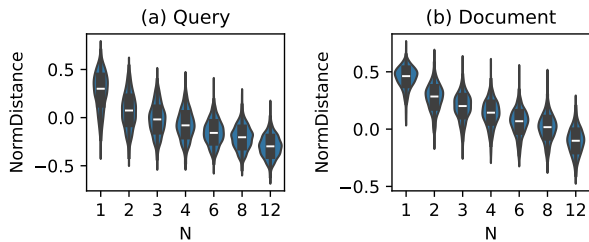


Figure 7: The distribution of the normalization between distances of an embedding with setting N and both text-only embeddings (d_t) and full image tokens embeddings (d_i), calculated as $(d_i - d_t)/(d_i + d_t)$. Higher values indicate text information dominance, while lower values suggest stronger visual influence. The distribution aligns with the performance curve, where the optimal $N = 3$ yields a more balanced distribution.

Setting	Avg. Seq. Len. Query/Doc	Encoding Time	Max Batch Size
$N = 1$	152.43/141.90	68.45	128
$N = 2$	160.86/156.76	69.51	128
$N = 3$	174.91/181.53	70.61	128
$N = 4$	194.58/216.21	71.94	128
$N = 6$	250.78/315.29	76.20	128
$N = 8$	329.46/454.00	85.27	64
$N = 12$	554.26/850.32	105.39	32
$N = 24$	1768.18/2770.74	187.03	16

Table 4: Inference efficiency of different token compression settings, measured by 1000 randomly selected testset pairs. Models are accelerated by FlashAttention-2 in float16. $N = 24$ is equivalent to the DPR baseline.

MME (§3.2), the visual token size of each image is controlled by the grid width N . As expected, decreasing N leads to reduced visual token numbers (sequence length), which translates into both accelerated encoding speeds (shorter time) and enhanced batch processing capabilities (larger batch size). In practice, the optimal N is determined by the trade-off between encoding efficiency and retrieval performance (Figure 6), which allows for flexible and efficient model deployment.

4.5 Discussion

We discuss about two potential scenarios beyond the tutorial-style data: (1) PDF Documents: They naturally contain interleaved text and visual content (tables, figures, diagrams). The retrieval of multimodal documents can be framed as a text-image interleaved retrieval task (Lee et al., 2024). (2) Social Media Contents: Users frequently incorporate visual elements (emojis, memes, photos, videos) within their textual conversations in non-fixed patterns. This is particularly relevant for multimodal

RAG tasks where retrieval of conversation history can be viewed as an interleaved retrieval task. We suppose, beyond these examples, there are likely many other scenarios where text and images naturally interleave, presenting opportunities for further exploration and application of our methodology.

5 Related Work

5.1 Multimodal Information Retrieval

Early Multimodal Information Retrieval tasks focused on cross-modal retrieval of text and image (Cao et al., 2022), where the goal is simply to retrieve captions of everyday images (Lin et al., 2014; Young et al., 2014). The scope has been extended to more complex scenarios, such as composed image retrieval (Liu et al., 2021), scientific contents (Wu et al., 2024), and visual documents (Ma et al., 2024; Faysse et al., 2025). Recent studies have been progressively exploring unified MIR settings (Zhou et al., 2024b). For instance, M-BEIR (Wei et al., 2024) integrates various image and text-related retrieval tasks, while UMRB (Zhang et al., 2025b) further extends the evaluation to encompass more textual datasets and visual document retrieval (Faysse et al., 2025). However, these benchmarks are constrained by their limitation to single-image queries or texts (Zhang et al., 2025b), lacking support for multi-image and interleaved contents. We construct a new text-image interleaved retrieval benchmark to meet the demands of complex multimodal RAG scenarios.

Current strong multimodal retrievers predominantly adopt the dense retrieval paradigm, which can be categorized into two main approaches: CLIP-style dual-stream models (Liu et al., 2023; Koukounas et al., 2024; Nussbaum et al., 2024) and language model-centric architectures (Lin et al., 2024; Zhou et al., 2024a; Jiang et al., 2024). Wang et al. (2024) proposed to compute unified multimodal embeddings from frozen LLM, which is not specifically designed for TIIR but shows potential in the multimodal context to image search task. A concurrent work (Lee et al., 2024) also explores interleaved embeddings for multimodal document retrieval, where a task-specific hierarchical encoder is suggested to retrieve interleaved documents parsed from Wikipedia webpage. In this work, we introduce the more generalized MLLM-based embedding model and propose a novel Matryoshka Multimodal Embedder to address the challenge of excessive visual tokens, which is crucial for TIIR.

5.2 Multimodal Interleaved Modeling

The modeling of interleaved text and image has been explored in various aspects, such as pre-training models (Alayrac et al., 2022; Laurençon et al., 2024) and corpus (Laurençon et al., 2023; Zhu et al., 2023). Notably, there exists a parallel line of research focusing on unified models that simultaneously handle both interleaved representation and generation tasks (Koh et al., 2023; Li et al., 2024; Zou et al., 2024). Their experimental datasets are converted from existing multimodal generation datasets with interleaved context, *e.g.*, Visual Storytelling (Huang et al., 2016), and less retrieval-oriented. Additionally, general interleaved corpus typically suffers from low knowledge density and logical coherence in image sequence (Zhang et al., 2025a), which might not be suitable for constructing an interleaved retrieval benchmark. In contrast, we build the TIIR dataset from human-curated high-quality tutorials (from wikiHow) for everyday skills, which are naturally interleaved and more informative for retrieval.

6 Conclusion

In this work, we introduce a new Text-Image Interleaved Retrieval (TIIR) task where the query and document are interleaved sequences of text and images, requiring the multimodal retriever to understand the semantics from interleaved context. We construct the wikiHow-TIIR benchmark based on the high-quality tutorial corpus from wikiHow, and present an efficient pipeline to generate text-image interleaved queries. We adapt several non-interleaved off-the-shelf multimodal and text retrievers to evaluate on our benchmark, showing that keeping interleaved structure is crucial for TIIR modeling. To explore native interleaved retrievers, we train interleaved MLLM-based DPR baseline and propose a novel Matryoshka Multimodal Embedder (MME) to address the challenge of excessive visual tokens. Evaluation results demonstrate the visual token compression strategy of MME achieves better performance and efficiency. We also present extensive analyses to understand the TIIR task and models, providing insights for future research in multimodal retrieval.

Limitations

Our work has two main limitations: First, during query generation, we rely on text-only LLMs to distribute multimodal information and construct

image-text interleaved sequences. This limitation stems from the unavailability of sufficiently capable interleaved MLLMs during the time of our data construction. We plan to explore the integration of newer models as they become available. Second, in our native interleaved model experiments, we exclusively employed the DeepSeek-VL-1.3B model without investigating larger 7B-scale models. This constraint is primarily due to computational resource limitations, as our training instances frequently generate input sequences with lengths ranging from 2K to 4K tokens, resulting in substantial memory consumption. With access to expanded computational resources in the future, we intend to scale up our model experiments.

Ethical Statement

We acknowledge that without proper training, the system may produce inaccurate retrieval results, which could negatively impact user experience. Our work is built upon existing resources (Yang et al., 2021) under the MIT license. All artifacts and data created through this research will maintain the same licensing terms to ensure proper usage and distribution rights. We welcome further research and applications based on our data. During data construction, we implemented strict controls and screening processes to ensure the absence of illegal or offensive content. As our data was created for purely academic purposes, we cannot be held responsible for any adverse effects resulting from commercial applications. Our data annotation was conducted with the help of computer science graduate students from our university. All annotators were fairly compensated for their contributions at appropriate market rates. In the preparation of this manuscript, we utilized LLM as a writing assistant, such as Claude and Qwen. However, all research design, implementation, and content creation were solely conducted and overseen by the authors. Through this statement, we aim to maintain transparency regarding the ethical considerations and methodological choices in our research.

Acknowledgments

This work receives support from the Natural Science Foundation of China (under Grants 62336008, 624B2048, 62125201, U24B20174), Research Grant Council of Hong Kong (PolyU/15209724), and the Shenzhen Science and Technology Program (under Grant ZDSYS20230626091203008).

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. **Flamingo: a visual language model for few-shot learning**. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736.
- Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. 2024. **Matryoshka multimodal models**. In *Workshop on Video-Language Models @ NeurIPS 2024*.
- Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. 2022. **Image-text retrieval: A survey on recent research and development**. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 5410–5417.
- Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William 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*, pages 5558–5570.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. **The faiss library**. *arXiv preprint arXiv:2401.08281*.
- Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELLOT, and Pierre Colombo. 2025. **Colpali: Efficient document retrieval with vision language models**. In *The Thirteenth International Conference on Learning Representations*.
- Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. **Visual storytelling**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.
- 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**. *arXiv preprint arXiv:2407.12580*.
- Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. **Grounding language models to images for multimodal inputs and outputs**. In *Proceedings of the 40th International Conference on Machine Learning*, pages 17283–17300.
- Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Günther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, and Han Xiao. 2024. **jina-clip-v2: Multilingual multimodal embeddings for text and images**. *Preprint*, arXiv:2412.08802.
- Black Forest Labs. 2023. Flux. <https://github.com/black-forest-labs/flux>.
- 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*.
- Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. 2023. **Obelics: An open web-scale filtered dataset of interleaved image-text documents**. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Jaewoo Lee, Joonho Ko, Jinheon Baek, Soyeong Jeong, and Sung Ju Hwang. 2024. **Unified multi-modal interleaved document representation for information retrieval**. *arXiv preprint arXiv:2410.02729*.
- Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan Kankanhalli. 2024. **Improving context understanding in multimodal large language models via multimodal composition learning**. In *Forty-first International Conference on Machine Learning*.
- Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. **MM-EMBED: Universal multimodal retrieval with multimodal LLMs**. In *The Thirteenth International Conference on Learning Representations*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. **Microsoft coco: Common objects in context**. In *Proceedings of the 13th European Conference on Computer Vision*, pages 740–755.
- Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. 2024. **PreFLMR: Scaling up fine-grained late-interaction multi-modal retrievers**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5294–5316.
- Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2023. **Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval**. In *The Eleventh International Conference on Learning Representations*.
- Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. **Image retrieval on real-life images with pre-trained vision-and-language models**. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134.
- Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. **Deepseek-vl: towards real-world vision-language understanding**. *arXiv preprint arXiv:2403.05525*.

- Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. [Unifying multimodal retrieval via document screenshot embedding](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505.
- Zach Nussbaum, Brandon Duderstadt, and Andriy Mulyar. 2024. [Nomic embed vision: Expanding the latent space](#). *arXiv preprint arXiv:2406.18587*.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. [Representation learning with contrastive predictive coding](#). *arXiv preprint arXiv:1807.03748*.
- 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.
- Chameleon Team. 2024. [Chameleon: Mixed-modal early-fusion foundation models](#). *arXiv preprint arXiv:2405.09818*.
- Ziyang Wang, Heba Elfardy, Markus Dreyer, Kevin Small, and Mohit Bansal. 2024. [Unified embeddings for multimodal retrieval via frozen LLMs](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1537–1547.
- 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 *Proceedings of 18th European Conference on Computer Vision*, volume 15145, pages 387–404.
- 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 (CVPR)*, pages 11307–11317.
- Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, et al. 2024. [SciMMIR: Benchmarking scientific multi-modal information retrieval](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12560–12574.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, 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*, page 641–649.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. 2021. [Visual goal-step inference using wikiHow](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2167–2179.
- Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. 2023. [Retrieval-augmented multimodal language modeling](#). In *Proceedings of the 40th International Conference on Machine Learning*, pages 39755–39769.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. [A survey on multimodal large language models](#). *National Science Review*, 11(12).
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. [From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions](#). *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Wenqi Zhang, Hang Zhang, Xin Li, Jiashuo Sun, Yongliang Shen, Weiming Lu, Deli Zhao, Yueting Zhuang, and Lidong Bing. 2025a. [2.5 years in class: A multimodal textbook for vision-language pretraining](#). *arXiv preprint arXiv:2501.00958*.
- Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. [mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.
- Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025b. [Bridging modalities: Improving universal multimodal retrieval by multimodal large language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024a. [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.
- Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. 2024b. [MARVEL: Unlocking the multi-modal capability of dense retrieval via visual module plugin](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14608–14624.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. [Multimodal c4: An open, billion-scale corpus of images interleaved with text](#). In *Thirty-seventh Conference on Neural Information Processing Systems: Datasets and Benchmarks Track*.

Xueyan Zou, Linjie Li, Jianfeng Wang, Jianwei Yang, Mingyu Ding, Junyi Wei, Zhengyuan Yang, Feng Li, Hao Zhang, Shilong Liu, et al. 2024. [Interfacing foundation models' embeddings](#). In *Advances in Neural Information Processing Systems*, volume 37.

Appendix

A WikiHow-TIIR

A.1 Data Collection

Our corpus construction adopts the wikiHow articles collected by [Yang et al. \(2021\)](#), systematically curated for Visual Goal-Step Inference (VGSI) research. This dataset comprises approximately 53,000 instructional articles. Structurally, each article decomposes a procedural objective (e.g., "hanging an ironing board") into multiple implementation methods (each article contains an average of 3 methods), with every method further annotated as stepwise components containing: (1) step headlines, (2) detailed descriptions, and (3) corresponding image demonstrations. We convert them into 155,262 self-contained, text-image interleaved documents, each structured as <Goal, Method Name, [(Step-Headline, Step-Image), ...]>.

Our multimodal query generation pipeline employs three state-of-the-art open-source architectures: Idefics3-8B-Llama3 ([Laurençon et al., 2024](#)), Qwen2.5-72B-Instruct ([Yang et al., 2024](#)), and FLUX.1-dev ([Labs, 2023](#)). The workflow initiates with systematic extraction of categorical metadata from wikiHow, successfully curating annotations for around 29,000 articles. Through stratified random sampling constrained by category distribution, we constructed: (1) A human-annotated test corpus comprising 7,654 queries and (2) A sample training partition containing 25,000 articles (pairs=73,084).


A.2 Query Generation

A.2.1 Query Text Generation

The reason why we select LLM to generate textual queries instead of MLLM is that: (1) At the time we conduct the study, MLLMs are not powerful enough to accept text-image interleaved data to perform complex task generation. (2) Considering that


[Prompt]
The above content is a step in a tutorial on <Goal>. Please describe the content in the picture and guess the hidden message the image conveys based on the text of step description.

[Example-Input]
<Goal> How to Hang an Ironing Board
<Description> An ideal door would be one located in your laundry room or near your laundry facilities. Hanging your ironing board in this area of your home helps you stay organized by keeping your ironing board close to other laundry items. If you don't have a laundry facility in your home, consider a place where you keep your laundry items, such as a closet that stores your detergent, laundry basket, etc. When hanging an ironing board over a door, the board can hang on either side. However, hanging the ironing board behind a door will present a more organized look as the board remains out of sight. While the over-the-door hanging method is one of the best ways to store your ironing board, there are many creative options out there.

<Image>


[Example-Output]
This picture is a step in a tutorial on how to hang an ironing board. There is a washing machine on the left and a glass door on the right. The glass door is surrounded by a green frame and a check mark. According to the text description and the content of the picture, the location of this picture is a bathroom or laundry room. At the same time, the green frame surrounding the glass door indicates that the tutorial recommends hanging the ironing board on the glass door of the laundry room.

[Input]
<Goal> How to Deal with Talkative Students
<Description> This signals to the student that the issue is closed and not open for debate. Don't wait for the student to give a response, as that wastes more time and opens the door for the student to create a distraction. You are thanking them for their positive behavior, which shows that you expect it to happen. If the student tries to argue with you anyway, keep your focus on their behavior. For example, the student may say, "I was only asking him if you said we had homework." You could respond, "I understand, but right now it's important that everyone listens to me." If you acknowledge their position and continue the redirection, they should stop talking.

<Image>


[Output]
The image is a step in a tutorial on how to deal with talkative students. In the image, there is a man in a white shirt and gray pants, and there is a speech bubble above his head with the text "Thank you...". The background is blurred. According to the text description and the content of the image, this image suggests that the teacher is thanking the student for their positive behavior, which shows that the teacher expects it to happen.

Figure 8: The example of the prompt, input and output of image caption.

we add design examples to the data generation process, if we use MLLM, we need to input more than ten images at a time, or even more, which brings great challenges to machine performance, runtime, and model capability. (3) Describe the image in the document through MLLM first and then use the textual document to generate data through LLM can effectively use the powerful performance of the current LLM, and can get better data generation effect in less resources and shorter running time.

Image Caption Therefore, we convert images to textual descriptions using Idefics3 by in-context learning style prompting. We chose this model considering that we fill in a well-designed example and the need to strike a balance between interleaved cross-modal alignment accuracy and computational efficiency. Specifically, we decompose each method into discrete steps and sequentially input stepwise data into the model to generate image captions that extract latent visual semantics. The implementation example is illustrated in Figure 8.


```

[Prompt]
The above is a tutorial on <Goal>, which contains text and pictures. The pictures are provided in the form of picture captions, with the <image> tag before the picture caption and the </image> tag after the picture caption. Please create a help post for this tutorial, including the keywords of tutorial, the identity of the help seeker, the difficulties encountered, and the content of the post, and generate at least 80 words, at most 110 words. In the post, please first describe the character's current situation and difficulties in detail based on the character's identity and situation with around 30 words, then give the character's own attempts to solve the problem, which may not be related to the tutorial, and finally ask for help. The content of the generated post must be consistent with the role identity and scenario.

[Example-Input]
<Goal> How to Hang an Ironing Board
<Headline> Hanging an Ironing Board Over a Door
<Steps>
Step 1: Find a door to hang your ironing board.
<image> This picture is a step in a tutorial on how to hang an ironing board. There is a washing machine on the left and a glass door on the right. The glass door is surrounded by a green frame and a check mark. According to the text description and the content of the picture, the location of this picture is a bathroom or laundry room. At the same time, the green frame surrounding the glass door indicates that the tutorial recommends hanging the ironing board on the glass door of the laundry room.</image>
Step 2: Ensure the board will fit on the back of the chosen door.
<image> This picture is a step in a tutorial on how to hang an ironing board. There is a washing machine on the left and a glass door on the right. The glass door is surrounded by a green frame and a check mark. According to the text description and the content of the picture, the location of this picture is a bathroom or laundry room. At the same time, the green frame surrounding the glass door indicates that the tutorial recommends hanging the ironing board on the glass door of the laundry room.</image>
Step 3: Buy a hanger with 2 hooks for a board with a T-shaped base.
<image> The picture is a step in a tutorial on how to hang an ironing board. In the picture, there is a hanger with two hooks and a green payment icon. According to the text description and the content of the picture, this picture suggests that you need to buy a hanger of the right shape to hang an ironing board.</image>

[Example-Output]
<Keywords> Ironing Board,Door,hanger
<Role> A housewife who just moved into a new home
<Scene> is worried about having too much furniture and didn't know where to put the ironing board.
<Query> God, I have so much furniture, I really don't know where to put it after moving! I packed too many things, including an ironing board, a hairdryer, a small sofa and so on. Our family recently moved downtown, but the reality is that our new house is much smaller than our previous one, so I really want to rearrange my furniture. Maybe I need to start with something that's easier to organize? What should I do with my ironing board? I think my laundry room should have some room for it, but I'm not quite sure how to make the most of the space. I'm even thinking about hanging it on a glass door - does that sound feasible?

[Input]
<Goal> How to Email Large Video Files
<Headline> Using Google Drive (Gmail)
<Steps>
Step 1: Open the Gmail website.
<image> The image is a step in a tutorial on how to email large video files. In the image, there is a web browser with a URL bar displaying "https://mail.google.com/mail". The URL bar is highlighted in green, and the "https" is bolded. The image also shows a back button and a home button. According to the text description and the content of the image, this image suggests that the first step in the tutorial is to open a web browser and navigate to the Gmail website.</image>
Step 2: Click Compose.
<image> The image is a screenshot of a Gmail inbox. The primary focus is on the "Compose" button, which is highlighted in red. The "Compose" button is located in the top left corner of the screen. Below the "Compose" button, there is a search bar where you can search for emails. The search bar is labeled "Google Search." Below the search bar, there is a list of folders, including "Primary," "Social," "Starred," "Drafts," "Notes," and "More." The "Primary" folder is selected, and there are two emails in this folder. The first email is from "Google Maps," and the second email is from "Facebook." Below the list of folders, there is a section labeled "Labels," which includes "Gmail" and "Social." The "Gmail" label is selected. The "More" option is also visible, which likely contains additional settings and options. The image also includes a "Sent Mail" section, which shows two emails. The first email is from "Gmail," and the second email is from "Yelp." Below the "Sent Mail" section, there is a "Twitter" label, which is selected. Finally, there is a "Pinterest" label, which is not selected. The image is a screenshot of a Gmail inbox, and the highlighted "Compose" button suggests that the user is about to compose a new email.</image>

[Output]
<Keywords> Gmail, Google Drive, large video files
<Role> A college student who needs to submit a project video
<Scene> is having trouble sending a large video file to a professor via email
<Query> Hey everyone, I'm struggling to send a large video file for my final project to my professor. I've tried compressing the file, but it's still too big for regular email. I've also attempted to use WeTransfer, but the file is too large for that as well. Does anyone have any suggestions on how I can use Google Drive to send this file through Gmail? I'm really stuck and would appreciate any help!

```

Figure 9: The example of the prompt, input and output of Query Text Generation.

Query Text Generation Following the text-only conversion of interleaved multimodal documents, we implement a two-stage query generation pipeline using a LLM. Current MLLMs (e.g., Chameleon (Team, 2024)) with joint text-image generation capabilities lack accessible image generation modules, necessitating sequential construction of image-text interleaved queries through: (1) Primary textual query synthesis using Qwen2.5-72B-Instruct, and (2) Subsequent multimodal composition. The Qwen2.5-72B-Instruct architecture is configured with a multi-perspective prompting framework across four semantic axes: keywords, character, scene, and query, simulating real-world problem-solving scenarios. The implementation

```

[Prompt]
Please rewrite part of a query. In the query given, it is required to replace the original part of the content through simple text and detailed image description without changing the semantics. The rewrite_text should not contain any specific items or scenes, and should simply link sentences while maintaining consistency. The image_caption should depict the situation taken by the phone or camera of the person who sent the query. Avoid showing what a person is doing in image_caption, and if necessary, replace it with local features, such as one hand is holding on something. Each description in the image_caption is related before and after, and the rewrite_text and image_caption must be the same length and correspond to the same position. The image_caption begins with 'a photo of'. Please reformat the sentences according to the content of the tutorial and query, and return the reformatted content in order.

[Example-Input]
<Doc>
Hanging an Ironing Board Over a Door
Step 1: Find a door to hang your ironing board. An ideal door would be one located in your laundry room or near your laundry facilities. Hanging your ironing board in this area of your home helps you stay organized by keeping your ironing board close to other laundry items. If you don't have a laundry facility in your home, consider a place where you keep your laundry items...
<Text>
[
'I have packed too many things, including an ironing board, a hairdryer, a small sofa and so on.',
'What should I do with my ironing board?',
'I'm even thinking about hanging it on a glass door - does that sound feasible?'
]
[Example-Output]
<Rewrite_text>
[
'Look I have packed so many things.',
'How can I deal with this one?',
'I even think about putting it here.'
]
<Image_caption>
[
'A picture of a cluttered, small living room filled with various pieces of furniture. On one side, a solid red V-shaped ironing board leans precariously against a wall. In the background, a small sofa sits amidst a pile of boxes, with a hairdryer peeking out from one of them. Sunlight filters through a window, casting warm light over the scene.'
'A photo of a solid red, V-shaped ironing board, which has been propped against a pristine white wall. The edges of the board are smooth, hinting at quality craftsmanship, and it appears stable in its leaning position. The white wall acts as a blank canvas that enhances the vividness of the red board, making it stand out even more. The absence of any decorations or textures on the wall draws the eye directly to the ironing board, allowing its color and shape to take center stage.'
'A photo of a tidy and functional laundry room, defined by its bright and inviting atmosphere. Central to the image is a sleek glass door, positioned prominently in the middle of the photo. Flanking the glass door are the room's essential appliances and furniture. To one side, there is a washing machine, and its metallic surface gleams under the illumination, suggesting it's well-maintained. Opposite the washing machine stands a striking red, V-shaped ironing board.'
]
[Input]
<Doc>
How to Quickly and Easily Clean Rims Washing your Rims
Step 1: Use a hose to rinse the rims. The first step to washing your rims is simply to rinse them thoroughly. Use a hose with a spray nozzle to remove any loose debris, dirt or grime from the wheels. Most wheel cleaners need to be applied to a wet surface. A power washer or professional car wash hose can remove a good deal of stuck on grime. A regular hose ...
<Text>
[
'I absolutely love keeping my car looking its best, but I've been having so much trouble with my rims.',
'I've tried using regular car soap and a sponge, but it doesn't seem to do the trick.',
'I even tried scrubbing them with a toothbrush, but it's just not effective enough.',
'Does anyone have any tips on how to quickly and easily clean rims?'
]
[Output]
<Rewrite_text>
[
'I really enjoy keeping my car looking great, but I'm struggling with a particular issue.',
'I've used regular soap and a sponge, but it hasn't worked well.',
'I also tried scrubbing with a small brush, but it wasn't effective.',
'Does anyone have any advice on how to clean this part quickly and easily?'
]
<Image_caption>
[
'A photo of a car parked in a driveway, with the focus on the wheels. The wheels are visibly dirty and covered in mud and grime, contrasting with the clean body of the car. The background is blurred, drawing attention to the dirty wheels and the car's sleek lines.'
'A photo of a hand holding a sponge and a bottle of regular car soap. The sponge is damp and the soap bottle is half-empty, with droplets of water visible on the label. The hand is about to apply the soap to a dirty wheel, but the wheel still looks grimy and unclear.'
'A photo of a hand holding a small, bristled brush, with the bristles touching the surface of a dirty wheel. The brush is worn, and the wheel still shows signs of dirt and grime, indicating that the cleaning attempt has not been successful.'
'A photo of a close-up of a dirty wheel, with a question mark drawn in the dirt using a finger. The wheel is in the foreground, and the background is blurred, emphasizing the need for effective cleaning tips.'
]

```

Figure 10: The example of the prompt, input and output of Text-image Information Reorganization.

example is demonstrated in Figure 9.

A.2.2 Text-image Information Reorganization

The construction of text-image interleaved queries presents dual modality coordination challenges during partial textual substitution: First, naive text-to-image conversion without original text retention induces inter-modal incoherence, where visual outputs fail to maintain linguistic continuity. Concurrently, directly processing non-objective textual queries through image generation models leads to visual semantic ambiguity due to conceptual abstraction. Second, preserving original textual components risks semantic redundancy, where visual representations become subsumed by textual semantics, negating their informational value. To solve these problems, we identify substitutable textual segments through semantic saliency analysis.

We implement a two-phase optimization method: Phase 1: Visual Info Selection. we segment query texts into constituent sentences and perform relevance ranking against source documents using BM25 to isolate the top- k ($k = 2, 3, 4$) maximally informative sentences. Phase 2: Query Rewriting. The selected sentences undergo semantic transformation via Qwen2.5-72B-Instruct, which: (1) Simulates human multimodal communication patterns by substituting text narratives with visual representations. (2) Synthesizes contextual bridging statements to maintain discourse continuity. This dual phase approach ensures the preservation of informational fidelity while achieving a human-aligned modality distribution, as demonstrated in Figure 10.

A.2.3 Image Generation

The image generation phase employs FLUX.1-dev, a state-of-the-art open-source image generation model, to generate images from captions. We configure the model with photorealistic constraints through the prompt ["photorealistic", "realistic", "photograph"] and set the output resolution to 512x512 pixels to ensure spatial consistency. The generated images are illustrated in Figure 11.

A.3 Data Annotation

We deploy a web-based annotation interface using Label Studio¹¹, hosting around 10,000 test instances requiring labeling, and engage 10 computer science graduate annotators via the university's information platform. After annotation, we imple-

¹¹<https://github.com/HumanSignal/label-studio>



Figure 11: Examples of generated images.



Figure 12: The example of our WikiHow-TIIR document and query.

ment dual verification mechanisms that include random sampling and statistical consistency checks. Annotators received performance-based remuneration calculated with hourly compensation rates averaging 12\$, exceeding local academic compensation standards.

On the whole, we establish strict guidelines that prioritize ethical and safety considerations, requiring all queries to: (1) adhere to legal standards, (2) exclude content involving pornography, violence or illegal activities, and (3) demonstrate rational and contextually appropriate requests.

We design an annotation methodology for image annotation comprising three key assessment dimensions: (1) Structural Integrity Evaluation: Annotators identify morphological anomalies in character and object generation. (2) Textual Content Classification: A three-tier text quality assessment. Level 1: No text. Level 2: Legible and comprehensible text. Level 3: Obvious textual errors (3) Semantic Relevance Verification. Annotators determine the image's contextual meaningfulness, excluding instances unrelated to the query or document.

Moreover, we set a comprehensive coherence

evaluation methodology to address potential inconsistencies arising from independent image generation: Level 1: Consistent subject/scene representation. Level 2: Minimal variations in subject/scene characteristics. Level 3: Significant divergences in subject/scene depiction. Annotators holistically analyze all images within a single query, systematically assessing visual consistency and identifying potential generative model limitations in maintaining semantic and visual coherence.

A.4 Data Statistics

Table 1 presents the dataset statistics. We calculate average text token lengths by concatenating text chunks and encoding them using LlamaTokenizer. Following the query generation methodology in §A.2, we create one positive query per document while utilizing same-article documents as hard negative samples (as stated in §A.1, each article contains an average of three documents).

B Implementation Details

We fine-tune OpenAI CLIP and DeepSeek-VL-1.3B. During training, we use a batch size of 32 and set a learning rate of $5 \times 10^{-5}/2 \times 10^{-5}$ with a linear warm-up scheduler for DeepSeek-VL-1.3B/CLIP. In our contrastive learning configuration, the temperature coefficient τ is empirically set to 0.05. Documents derived from identical source articles are designated as in-batch negatives. Specifically, we implement randomized selection of a single hard negative instance per mini-batch. The entire process undergoes three complete training epochs.

We select DeepSeek-VL-1.3B-base to train in four ways. (1) *Baseline*(DPR): We set the image token number as the model default, 576, to train. (2) *Random sampling* (Rand): We randomly sample a grid width N for each micro-batch. (3) *Matryoshka learning* (MRL): We train the model with all M kernel sizes simultaneously. (4) *Mean learning* (Mean): We additionally compute losses between query and document embeddings of different sizes, the final loss is the mean of all $M \times M$ possible combinations. All models are trained with the max token length of 4096, and test with the same.

Table 6 demonstrates Jina-CLIP-v2’s superior performance through normalized image-text embedding fusion approach (summation of averaged modality embeddings). This methodology was subsequently adopted for training clip-vit-large-patch14, with detailed performance metrics pro-

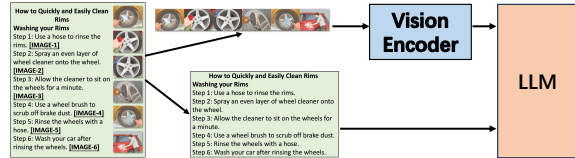


Figure 13: The example of the way that encodes text-image interleaved content with single-image multimodal retrievers.

Format	Prompt(query/doc)
only-text	<sent>\nSummary above query/tutorial in one word:
image+text	<image>\n<sent>\nSummary above query/tutorial in one word:

Table 5: The instructions of the E5-V model.

vided in the same table.

C Experiments Details

All experiments are conducted on a NVIDIA A100-80G 8-GPU server. All retrieval results were implemented using Faiss (Douze et al., 2024).

C.1 Single-image Multimodal Retrievers

Given architectural constraints in single-image multimodal retrievers that process only single image-text pairs per instance, we disentangle image-text interleaved data into images and text to encode. The implementation pipeline (Figure 13) demonstrates this separation process.

E5-V introduces unimodal training through text-only pairwise optimization. The architecture employs specialized markup templates for modality-specific encoding. The constructed prompts what we set are formally specified in Table 5 following standard template formatting conventions.

MM-Embed and GME_{Qwen2-VL-2B} require task-specific instructions appended to each query. We implement standardized prompts for both architectures: "Retrieve a wikiHow tutorial that provides an answer to the given query" for MM-Embed and "Find a wikiHow tutorial that matches the given query" for GME_{Qwen2-VL-2B}.

C.2 Text Models

For text models, we implement two encoding strategies for text-image interleaved data: (1) remove the images and keep only the text, and (2) replace the images with image captions. The latter employs the standardized prompt "Describe the image" for real-time inference simulation, replacing image with generated captions through the processing of Qwen2-VL-2B-Instruct.

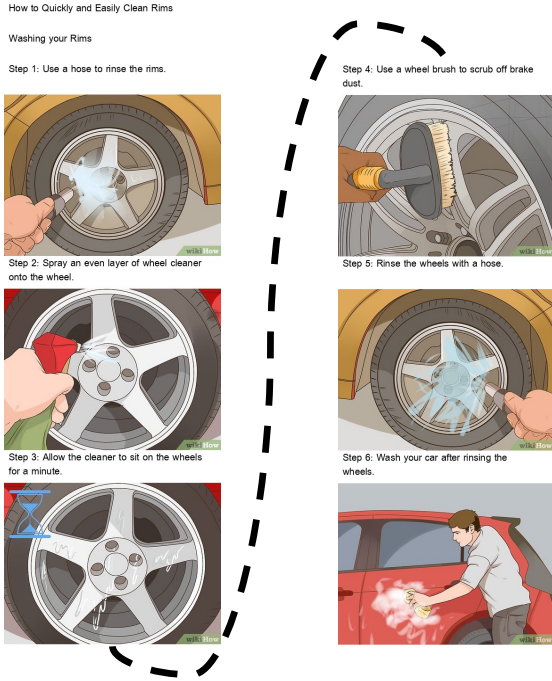


Figure 14: The example of visual document (image). The left and right images in the picture are joined up and down, but for the sake of the layout of the paper, we cut them and arrange them left and right.

We implement standardized prompt "Given a query, retrieve relevant wikiHow document that answer the query" for GTE-Qwen2-7B and "Represent this query for searching relevant wikiHow passages:" for BGE-v1.5_{large}.

C.3 Two-stream Models

For two-stream models, we employ separate text-image encoding pipelines. Text embeddings derive from concatenated document chunks, while visual encoding explores: (1) image concatenation, and (2) normalized mean pooling of individual image embeddings. Following established multimodal fusion methods (Liu et al., 2023), we evaluate three combination strategies: vector summation, feature concatenation, and element-wise multiplication, reporting optimal results in Table 2.

C.4 Visual Document (Image) Retrievers

For visual document (image) retrievers, we convert the whole query/document into one image. The example is shown in Figure 14.

C.5 Ablation Study

Finally, we conduct an ablation study to investigate the hyper-parameters in our model training. Due to

Model	Text&Image	Image	Recall@5	MRR@10	nDCG@10
Jina-CLIP-v2	Sum	mean	58.80	44.44	49.97
		concat	51.13	37.77	43.05
	Concatenate	mean	55.91	42.31	47.48
		concat	50.10	36.85	41.87
	Dot product	mean	30.36	21.58	25.18
		concat	24.61	17.35	20.39
CLIP _{large} Fine-tuned	Sum	mean	69.41	54.25	60.01
		concat	55.55	41.67	47.25
	Concatenate	mean	61.18	46.84	52.36
		concat	49.33	36.71	41.94
	Dot product	mean	16.19	11.78	13.82
		concat	10.5	7.43	8.79

Table 6: Evaluation results on our WikiHow TIIR of the two-stream models, Text&Image denotes the way we combine the text and image embedding, and Image denotes the way we get the image embedding.

Model	LoRA Rank	Learning Rate	MRR@10 ($N = 3$)
MME	16	5e-5	62.50
MME	16	1e-4	57.73
Rand	8	5e-5	62.17
Rand	32	5e-5	61.99

Table 7: Ablation study of different hyper-parameters in our MLLM-base model training. We perform hyper-parameter search on MME align since it's the fastest to train. The results of the best setting $N = 3$ are shown. As GPU resources are limited, we run all experiments with the same batch size of 32.

computational constraints¹², our hyper-parameter search is based-on the most training-friendly Rand strategy of MME. We vary the rank of LoRA (8, 16, 32) and learning rate (1e-4, 2e-5), where the LoRA rank controls the size of new learnable parameters in training. Although batch size substantially influences model performance (with larger batch sizes generally yielding better results in contrastive learning), we opt to maintain a fixed batch size, *i.e.*, the maximum allowable within GPU constraints, across all models to ensure fair comparison. Therefore, the impact of batch size is not discussed in this analysis. As shown in Table 7, the best setting is achieved with a rank of 16 and a learning rate of 5e-5.

¹²The training instances of our dataset frequently generate input sequences with lengths in the order of 4,000 tokens, resulting in substantial memory consumption.