Unveil: Unified Visual-Textual Integration and Distillation for Multi-modal Document Retrieval

Anonymous ACL submission

Abstract

Document retrieval in real-world scenarios faces significant challenges due to diverse document formats and modalities. Traditional text-based approaches rely on tailored parsing techniques that disregard layout information and are prone to errors, while recent parsing-free visual methods often struggle to capture fine-grained textual semantics in text-rich scenarios. To address these limitations, we propose **Unveil**, a novel visual-textual embedding framework that effectively integrates textual and visual features for robust document representation. Through knowledge distillation, we transfer the semantic understanding capabilities from the visual-textual embedding model to a purely visual model, enabling efficient parsing-free retrieval while preserving semantic fidelity. Experimental results demonstrate that our visual-textual embedding method surpasses existing approaches, while knowledge distillation successfully bridges the performance gap between visual-textual and visual-only methods, improving both retrieval accuracy and efficiency.

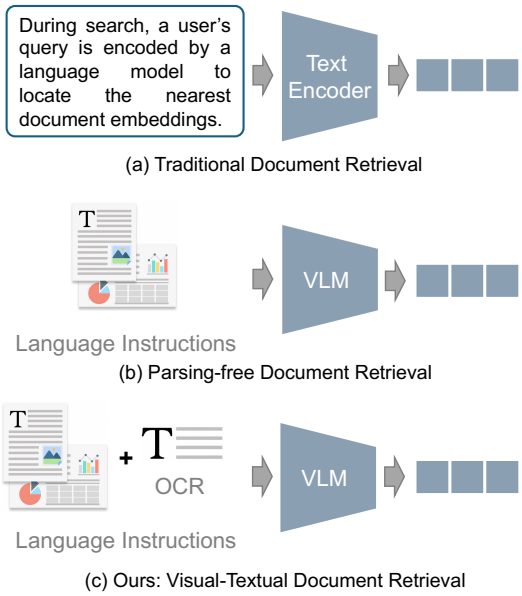


Figure 1: Comparison of document retrieval methods. Traditional approaches parse text and use text encoders for embeddings, while parsing-free methods directly process document screenshots with visual language models. Our visual-textual approach leverages both modalities, effectively addressing diverse scenarios.

1 Introduction

Document retrieval for real-world applications remains a challenging task due to the need to effectively handle diverse document formats, including text, images, charts, and complex visual layouts. As shown in Figure 1, traditional document retrieval predominantly relies on Optical Character Recognition (OCR) to convert scanned or image-based documents into machine-readable text. Subsequently, approaches such as the lexical-based BM25 (Robertson et al., 2009) and embedding-based techniques like Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) are utilized to model the semantic relevance between queries and documents. However, OCR-dependent pipelines come with significant limitations. They not only add computational overhead but also introduce potential recognition errors. Furthermore, these approaches often miss crucial visual contextual elements, which are essential for comprehending document content (Zhang et al., 2024; Faysse et al., 2024; Ma et al., 2024a).

Recent research has shifted toward parsing-free techniques that directly utilize visual inputs such as document screenshots (Faysse et al., 2024; Ma et al., 2024a; Zhou et al., 2024; Ni et al., 2021). These methods leverage Vision-Language Models (VLMs) and Multi-Modal Large Language Models (MLLMs) to process entire pages directly, preserving rich structural and graphical information (Cho et al., 2024; Yu et al., 2024; Yao et al., 2024a). While these approaches circumvent the computational overhead and complexity associated with OCR, our empirical analysis reveals significant limitations in their textual understanding capabil-

ities. As shown in Figure 2, our comparison of text-based and visual-based methods across both text-rich scenarios (web-page retrieval) and visual-rich scenarios (visual document retrieval) reveals distinct performance patterns. In visual-rich scenarios where layout and graphical elements are crucial, these visual-based methods outperform traditional text-based approaches, highlighting their superior ability to process spatial and structural information (Masry et al., 2022; Tanaka et al., 2023; Tito et al., 2023). However, when handling text-rich contexts, visual-based methods struggle to capture semantic details that text-based methods process effectively.

This observation underscores a fundamental challenge: text-based methods excel in modeling linguistic semantics but overlook crucial layout and graphical details, while purely visual methods preserve visual context but struggle with fine-grained language understanding (Faysse et al., 2024; Zhang et al., 2024; Ni et al., 2021; Ma et al., 2024a). To address this limitation, we propose **Unveil** (**U**nified **V**isual-**T**ext **I**ntegration and Distil**l**ation), a novel framework that bridges the gap between textual and visual document understanding. Our approach consists of two key components: First, we develop a visual-textual embedding approach that integrates both textual and visual inputs, leveraging the complementary strengths of both modalities for comprehensive document representations. Second, we conduct knowledge distillation to transfer semantic understanding from the teacher model (visual-textual embedding model) to the student model (purely visual model), enabling enhanced text comprehension without OCR dependency during inference. Specifically, we propose several techniques to facilitate this distillation process: (1) Representation Alignment: The student model is trained to replicate the teacher model's representations by minimizing the distance between their query and document representations. (2) Soft Label Distillation: We utilize the teacher model to provide a fine-grained label distribution for the student model. (3) Adaptive Re-Weighting: We dynamically identify instances where discrepancies exist between the teacher and student models, assigning higher weights to these instances.

Our framework offers a flexible retrieval system. For text-rich scenarios that require precise semantic nuances, the visual-textual model—which incorporates both textual and visual inputs—can be employed. Alternatively, in scenarios where efficiency or OCR-free processing is preferred, the distilled visual-only model serves as a practical alternative while maintaining comparable semantic understanding. We validate our approach on 12 datasets encompassing both text-rich and visual-rich scenarios. Experimental results indicate that the visual-textual embedding model consistently outperforms both text-based and visual-based methods. Furthermore, the distillation process effectively reduces the gap between visual-textual and visual-only approaches, enhancing retrieval accuracy and efficiency.

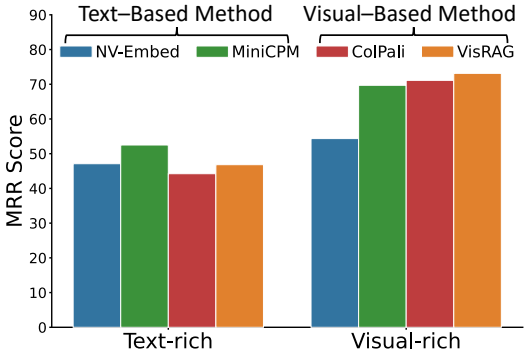


Figure 2: Empirical analysis on retrieval performance under text-rich and visual-rich scenarios.

In summary, our contributions are as follows:

- We identify that existing text-based and visual-based methods struggle to adapt across different scenarios. To address this, we introduce Unveil, a visual-textual embedding approach that integrates textual and visual features for comprehensive document understanding.
- We propose several knowledge distillation strategies to transfer the visual-textual model's robust textual understanding to a purely visual model, enabling parsing-free retrieval without compromising accuracy.
- Extensive experiments demonstrate that our visual-textual embedding method outperforms existing text-based and visual-based methods. Additionally, the knowledge distillation effectively reduces the gap between visual-textual and visual-only approaches, enhancing retrieval accuracy and efficiency.

2 Methodology

Our proposed method seeks to bridge the gap between visual-textual and visual-only approaches.

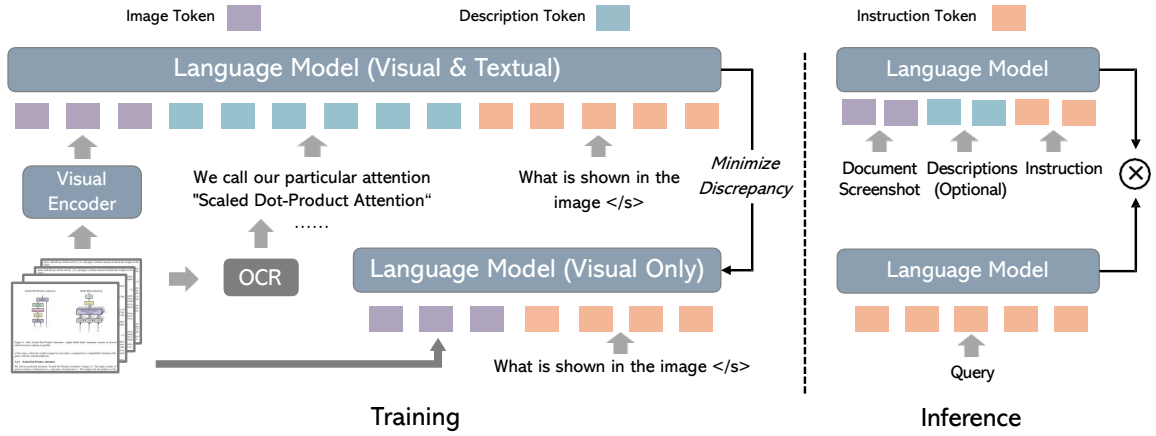


Figure 3: Unveil consists of: (a) a visual-textual embedding model that jointly processes document images and OCR text, and (b) a purely visual model that operates on document images only. During training, knowledge distillation is employed to transfer semantic understanding from the teacher (visual-textual) to the student (visual-only) model. At inference time, the framework offers flexibility to choose between the two models based on efficiency requirements.

Initially, Unveil learns a textual-visual embedding model that leverages both OCR-derived text and visual inputs. Subsequently, it distills the strong capacity of the visual-textual embedding model into a purely visual model. The distilled visual model thus retains the semantic richness characteristic of visual-textual embeddings while achieving high efficiency without the need for textual input.

2.1 Task Definition

Given a query q and a corpus $\mathcal{C}$ comprising visual documents $\{d_1, d_2, \ldots, d_n\}$, the task of multimodal document retrieval is to identify the k visual documents most relevant to the query q. Relevance is assessed using a similarity metric to measure the similarity between the query and document embeddings. Here, a visual document represents a complete information snippet (e.g., a web article or a PPT page), while the query is purely textual.

2.2 Unveil Framework

As illustrated in Figure 3, Unveil comprises two components: the teacher model (Visual-Textual model) and the student model (Visual-Only model). We initially train these models independently, followed by knowledge distillation to produce a visual-only model capable of robust document retrieval without OCR reliance during inference.

Unified View of Retrieval Models Both the visual-textual and visual-only models employ a dual encoder architecture to model the similarity between queries and documents. The key distinction lies in the input to the document encoder.

For the visual-textual model's document encoder, we begin by employing OCR on each document image d_i to derive a textual description t_i. The document image is then processed by the vision encoder of the vision-language model to yield visual tokens. The encoded visual latent embeddings are concatenated with a text prompt for input to the subsequent language model: *"<s><img> <description> What is shown in this image?</s>"*. In contrast, for the visual-only model's document encoder, the input to the subsequent language model is: *"<s><img> What is shown in this image?</s>"*. For both models, the input to the query encoder is the query text. To aggregate sequence information using a language model with uni-directional attention, following prior work (Ma et al., 2024a), we use the embedding of the end-of-sequence token </s> from the last hidden state as the representation. The representation of the queries and documents are calculated as follows:

$$\begin{aligned} \boldsymbol{q} &= \text{VLM}(q)[-1] \\ \boldsymbol{d}_s &= \text{VLM}(\text{<img>}, \text{<inst>})[-1] \\ \boldsymbol{d}_t &= \text{VLM}(\text{<img>}, \text{<desc>}, \text{<inst>})[-1] \end{aligned} \tag{1}$$

where $\boldsymbol{d}_s, \boldsymbol{d}_t$ denote the document representation from the student and teacher, respectively.

The query-document similarity is measured using cosine similarity between their embeddings:

$$\text{Sim}(\boldsymbol{q}, \boldsymbol{d}) = \frac{\boldsymbol{q}^T \cdot \boldsymbol{d}}{\|\boldsymbol{q}\| \cdot \|\boldsymbol{d}\|} \tag{2}$$

During training, our embedding model is opti-

mized using the InfoNCE loss (Oord et al., 2018):

$$\mathcal{L}_{\text{hard}} = -\sum_{d^+ \in \mathbb{D}^+} \log \frac{\exp(\text{Sim}(\boldsymbol{q}, \boldsymbol{d}^+))}{\sum_{d \in \mathbb{D}} \exp(\text{Sim}(\boldsymbol{q}, \boldsymbol{d}))} \quad (3)$$

After independently training both models, we freeze the teacher model and leverage it to guide the student model during knowledge distillation.

Representation Alignment To align the representation of the student and teacher model, we define a representation alignment loss:

$$\mathcal{L}_{\text{align}} = \frac{1}{n}\sum_{i=1}^{n}(\|\boldsymbol{d}_t - \boldsymbol{d}_s\|_2^2 + \|\boldsymbol{q}_t - \boldsymbol{q}_s\|_2^2) \quad (4)$$

where n is the number of query-doc pairs.

Minimizing $\mathcal{L}_{\text{align}}$ encourages $\boldsymbol{d}_s$ and $\boldsymbol{q}_s$ to inherit the teacher's textual representation abilities. As training progresses, the student model learns to encode in a manner reflecting both textual semantics and visual features, despite never explicitly encountering textual data during inference.

Soft Label Distillation The teacher model's score distribution conveys fine-grained similarity information, unlike hard one-hot labels. We leverage this by aligning the student's distribution with the teacher's: The label distribution of the student model and the teacher model are defined as follows:

$$\boldsymbol{t} = \underset{d \in \mathbb{D}}{\forall} \frac{\exp(\text{Sim}(\boldsymbol{d}, \boldsymbol{d}_t))}{\sum_{d' \in \mathbb{D}} \exp(\text{Sim}(\boldsymbol{d}, \boldsymbol{d}'_t))} \quad (5)$$

$$\boldsymbol{s} = \underset{d \in \mathbb{D}}{\forall} \frac{\exp(\text{Sim}(\boldsymbol{d}, \boldsymbol{d}_s))}{\sum_{d' \in \mathbb{D}} \exp(\text{Sim}(\boldsymbol{d}, \boldsymbol{d}'_s))} \quad (6)$$

The soft label distillation loss is calculated as:

$$\mathcal{L}_{\text{soft}} = D_{\text{KL}}(\boldsymbol{t}/\tau, \boldsymbol{s}/\tau) \quad (7)$$

where τ is the temperature parameter.

Adaptive Re-Weighting Discrepancies between student and teacher models on certain documents can reveal student misinterpretations. We propose focusing on these discrepancies by giving them higher weights using KL Divergence:

$$w_i = \frac{\exp(-D_{\text{KL}}(\boldsymbol{t}_i, \boldsymbol{s}_i)/\tau)}{\sum_{\boldsymbol{j}=1}^{K} \exp(-D_{\text{KL}}(\boldsymbol{t}_j, \boldsymbol{s}_j)/\tau)} \quad (8)$$

where w_i denotes the importance of document d_i.

Finally, the total loss combines both the alignment loss and soft label distillation loss:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{n} w_i \times (\mathcal{L}^i_{\text{align}} + \mathcal{L}^i_{\text{soft}}) \quad (9)$$

Inference During inference, Unveil offers two inference modes to cater to different needs regarding performance and computational efficiency.

The first mode, **Visual-Textual Mode**, uses both OCR text and document images to achieve optimal retrieval performance. By combining rich visual features with extracted textual information, it maximizes semantic understanding for scenarios requiring high precision. The second mode, **Visual-Only Mode**, relies solely on the distilled visual model without OCR dependency, maintaining competitive accuracy through advanced visual representations. This mode significantly reduces computational overhead, making it ideal for efficiency-critical applications.

3 Experiment Setup

We evaluate Unveil on two distinct multi-modal retrieval scenarios: visual document retrieval, which emphasizes visual content, and web page retrieval, which focuses on textual content.

3.1 Visual Document Retrieval

Dataset We employ question-document pairs from various VQA datasets, each targeting distinct document types: MP-DocVQA (Tito et al., 2023) for industrial documents, ArXivQA (Li et al., 2024), ChartQA (Masry et al., 2022), InfographicsVQA (Mathew et al., 2022), and PlotQA (Methani et al., 2020) for different types of figures, as well as SlideVQA (Tanaka et al., 2023) for presentation slides. We adhere to the datasets' original train-test splits, except for MP-DocVQA and InfographicsVQA, where the validation split is utilized as our evaluation set. We construct the retrieval corpus by collecting the positive documents linked to each query from the training and evaluation sets.

Evaluation Following conventional assessment approaches for VQA datasets, we apply Recall@10 and MRR@10 as evaluation metrics.

3.2 Web-Page Retrieval

Dataset Following (Ma et al., 2024a), we employ the Wiki-SS-corpus[1] as our retrieval corpus. This dataset is compiled from English Wikipedia pages via URLs, with screenshots captured automatically over four days, from May 20 to May 23, 2024. The corpus comprises 1,267,874 Wikipedia

[1] https://huggingface.co/datasets/Tevatron/wiki-ss-corpus

Model	ArxivQA		ChartQA		DocVQA		InfoVQA		PlotQA		SlideVQA		AVG	
	Rec	MRR	Rec	MRR	Rec	MRR	Rec	MRR	Rec	MRR	Rec	MRR	Rec	MRR
Text-Based Models														
BM25	42.30	32.48	56.69	43.66	86.38	73.56	83.19	69.03	50.48	33.18	91.16	76.65	68.37	54.76
GTR	40.82	31.17	56.55	43.50	74.19	57.09	84.95	67.82	44.87	28.83	90.43	74.82	65.30	50.54
BGE-Large	38.40	29.78	53.76	42.47	77.54	59.63	87.98	70.86	47.60	32.06	92.26	75.77	66.26	51.76
NV-Embed	44.92	35.13	52.09	43.04	80.26	60.50	91.84	77.05	47.67	31.55	93.90	78.83	68.45	54.35
MiniCPM	69.53	57.02	73.96	60.91	93.24	80.56	94.67	82.34	63.86	45.07	96.93	92.30	82.03	69.70
Visual-Based Models														
SigLIP	50.50	35.57	66.16	47.62	54.55	34.86	68.08	47.40	52.82	25.89	87.22	78.06	63.22	44.90
ColPali	81.11	69.85	77.16	62.68	94.78	83.64	94.82	81.92	60.66	40.84	97.32	86.83	84.31	70.96
DSE	85.41	72.11	78.13	63.42	94.20	80.41	97.07	84.96	63.82	43.82	97.01	93.08	85.94	72.96
VisRAG	84.93	71.41	78.83	64.54	94.73	80.12	96.33	85.53	64.30	44.31	97.38	92.94	86.08	73.14
Hybrid Models														
DSE	77.57	63.76	74.51	62.68	93.88	81.55	96.58	85.39	64.22	45.70	97.20	93.84	83.99	72.16
VisRAG	83.58	69.48	77.58	64.35	95.64	83.27	96.63	85.70	64.61	46.00	97.71	93.61	85.96	73.74
Unveil (Ours)														
Visual-Textual	**86.24**	**73.67**	79.53	**66.75**	**96.06**	**83.88**	97.26	86.19	64.63	45.91	**97.87**	**94.37**	86.93	**75.13**
Visual-Only	86.23	73.27	**80.36**	66.40	95.74	82.53	**97.61**	**86.39**	**64.82**	**46.16**	97.61	93.75	**87.06**	74.75

Table 1: Overall performance on Visual Document Retrieval. The best retrieval performance is marked in **bold**.

screenshots. To reduce inference time, we sample 112,888 screenshots to serve our retrieval corpus. For training, we use the Wiki-SS-NQ dataset[2], which is constructed by performing a BM25 search for each question to retrieve positive documents, thus forming query-document pairs.

Given the extensive use of the Wikipedia corpus in open-domain QA tasks, we make evaluation using several widely utilized QA datasets. These include open-domain QA datasets such as NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and WebQ (Berant et al., 2013), multihop datasets like Wikihop (Yang et al., 2018) and HotpotQA (Ho et al., 2020), as well as the ambiguous dataset ASQA (Stelmakh et al., 2022).

Evaluation Consistent with previous practices for evaluating the effectiveness of retrieval in QA datasets, we use Recall@10 and MRR@10 as evaluation metrics. Specifically, a question is considered correctly answered if its retrieved documents contain at least one answer from the answer list.

3.3 Implementation Details

Our framework involves initially training both a teacher and a student model independently, followed by knowledge distillation. Throughout both stages, models are fine-tuned using in-batch negatives for two epochs, with a batch size of 16 and a learning rate of 2e-5 on 8 NVIDIA A100 80GB GPUs. We initialize the models with MiniCPM-V 2.0 (OpenBMB, 2024; Yao et al., 2024a). Additional details regarding the training and document parsing are provided in Appendices B and C.

3.4 Baselines

We compare our method with the following retrieval approaches:

- **Text-Based Models**: This category encompasses BM25, a well-known lexical model, as well as advanced text embedding models such as BGE-Large-en-v1.5[3](Xiao et al., 2023), GTR-T5-Large[4](Ni et al., 2021), NV-Embed-v1[5](Lee et al., 2024a), and MiniCPM[6](Yao et al., 2024b), which has been fine-tuned for dense text retrieval.

- **Visual-Based Models**: This includes SigLIP[7](Zhai et al., 2023), a model in the CLIP style for vision tasks; ColPali(Faysse et al., 2024), a multi-vector retrieval model; as well as DSE (Ma et al., 2024a) and VisRAG (Yu et al., 2024), which are state-of-the-art visual embedding models.

- **Hybrid Models**: We also create hybrid models by interpolating similarity scores from the retrieval results of visual-based retriever like

[2] https://huggingface.co/datasets/Tevatron/wiki-ss-nq

[3] https://huggingface.co/BAAI/bge-large-en-v1.5

[4] https://huggingface.co/sentence-transformers/gtr-t5-large

[5] https://huggingface.co/nvidia/NV-Embed-v1

[6] openbmb/MiniCPM-V-2

[7] https://huggingface.co/HuggingFaceM4/siglip-so400m-14-980-flash-attn2-navit

Model	NQ		TriviaQA		WebQ		Wikihop		HotpotQA		ASQA		AVG	
	Rec	MRR	Rec	MRR	Rec	MRR	Rec	MRR	Rec	MRR	Rec	MRR	Rec	MRR
Text-Based Models														
BM25	61.33	40.01	72.67	56.35	64.07	41.73	36.02	23.98	48.93	33.64	70.50	48.07	58.92	40.63
GTR	66.84	52.22	57.41	40.70	73.13	55.96	25.14	15.11	38.23	25.04	79.66	64.28	56.74	42.22
BGE-Large	68.39	54.22	61.03	44.36	73.97	56.99	26.65	16.74	42.55	29.05	80.67	66.46	58.88	44.64
NV-Embed	69.97	54.61	68.84	52.04	75.10	55.77	31.62	19.48	45.63	31.78	82.12	68.96	62.21	47.11
MiniCPM	75.23	60.04	77.44	63.56	75.76	59.26	39.10	25.44	50.93	36.11	83.24	69.47	66.95	52.31
Visual-Based Models														
SigLIP	59.57	41.45	53.25	34.08	58.33	39.09	22.30	13.93	31.63	18.89	67.82	47.11	48.82	32.42
ColPali	68.78	53.18	60.70	44.82	73.57	56.50	27.59	16.97	40.77	27.43	81.68	66.47	58.85	44.23
DSE	71.70	55.90	73.07	55.61	71.67	54.30	35.03	21.70	45.53	30.43	78.66	62.21	62.61	46.69
VisRAG	72.17	56.20	72.37	55.55	71.38	53.96	34.13	20.46	45.60	31.06	79.78	63.72	62.57	46.83
Hybrid Models														
DSE	73.80	59.56	76.18	61.74	73.79	57.98	37.17	23.88	48.20	34.25	81.34	67.03	65.08	50.74
VisRAG	73.80	59.73	75.84	61.22	74.14	57.84	36.13	23.16	48.67	34.81	81.79	68.36	65.06	50.85
Unveil (Ours)														
Visual-Textual	**75.80**	**61.86**	**78.75**	**64.68**	**75.81**	**60.02**	**40.57**	**26.23**	**52.40**	**36.62**	**84.13**	**70.40**	**67.91**	**53.30**
Visual-Only	72.20	57.30	74.64	59.07	73.84	55.84	35.50	22.23	48.13	32.95	80.11	65.31	64.07	48.78

Table 2: Overall performance on Web-Page Retrieval. The best retrieval performance is marked in **bold**.

DSE and VisRAG and with text-based retrievers MiniCPM (Ma et al., 2024b).

4 Experimental Results

4.1 Main Result

In this section, we present experiments in both visual document retrieval and web page retrieval scenarios. Based on the results shown in Tables 1 and 2, several observations can be made:

First, text-based and visual-based models each exhibit unique advantages in different scenarios. For example, in web page retrieval, the text-based method MiniCPM significantly outperforms visual-based models. Conversely, in visual document retrieval, visual-based approaches excel. This highlights that these models cannot achieve superior performance across both scenarios. Interestingly, the simple lexical method BM25 outperforms more powerful dense retrieval models like BGE-Large. This can be attributed to the fact that text within these visual documents is often fragmented and semantically incoherent. In such cases, string matching might be a more effective solution.

Second, hybrid models yield intermediate results, which is understandable given that, in web-page retrieval, the scores generated by text-based models might be adversely affected by the less accurate scores from visual-based models, which leads to performance inferior to that of text-based models alone. This demonstrates that merely merging the outputs of the two models does not inherently enhance performance. Additionally, these models necessitate inference from both models, which increases the inference cost.

Methods	DocVQA		InfoVQA	
	Rec	MRR	Rec	MRR
Ours	95.74	82.53	97.61	86.39
-w/o Adaptive Re-Weighting	95.64	82.45	97.41	86.20
-w/o Representation Alignment	95.26	81.49	97.21	85.89
-w/o Soft Label Distillation	94.94	80.89	96.68	85.06
-w/o Distillation	94.20	80.41	97.07	84.96

Table 3: Ablation Study. We experiment by gradually removing all components and observing the performance.

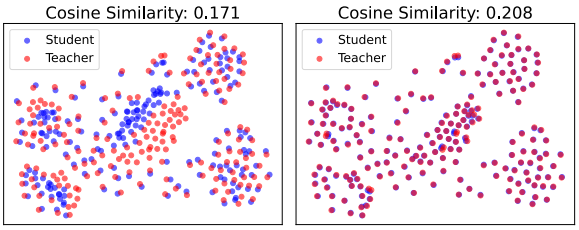
Third, our method Unveil, specifically the visual-textual variant, consistently achieves the highest performance across all retrieval scenarios, confirming its effectiveness in integrating information from both modalities for improved outcomes. Furthermore, the distilled visual-only version exhibits superior performance compared to both text-based and visual-based models and can even achieve performance comparable to the teacher model while requiring no text input. This is mainly because our distillation framework can effectively transfer comprehensive knowledge to the student model.

4.2 Ablation Study

In this section, we evaluate the effectiveness of each component by incrementally removing them and observing the changes in performance. "w/o Representation Alignment" and "w/o Soft Label Distillation " refer to the removal of representation alignment and soft label distribution loss, respec-

Length	NQ		TriviaQA		WebQ	
	Rec	MRR	Rec	MRR	Rec	MRR
0	71.70	55.90	73.07	55.61	71.67	54.30
512	72.73	57.53	75.94	60.93	73.55	57.00
1024	73.70	57.73	76.81	61.41	72.81	56.79
2048	74.43	59.68	78.11	63.81	74.33	58.46
3096	75.27	60.83	78.71	63.87	76.11	59.09

Table 4: Performance of teacher model using different input text lengths.



(a) Before Distillation (b) After Distillation

Figure 4: The visualization of document embeddings.

tively, following the removal of the adaptive reweighting. "w/o distillation" represents the visual model before distillation.

As shown in Table 3, removing each component results in performance degradation, confirming the effectiveness of each component. Specifically, we find that removing the representation alignment loss leads to significant degradation in model performance. This is because token representation contains the most valuable information about the query and document, and forcing the visual model to produce representations similar to the visual-textual model is the most direct way to learn from it. Additionally, removing the soft label distillation also results in performance degradation, primarily because the teacher provides a soft label that helps the student model discern fine-grained differences between documents within the same batch.

4.3 Analysis

Impact of Text Length In this section, we analyze the impact of text length on the performance of the teacher model. Specifically, we gradually increase the text length from 0 to 3096 and observe the changes in performance.

As shown in Table 4, the model performance improves as the context length increases. This is expected because longer texts can provide more useful information about the image. However, longer contexts also incur additional inference costs, highlighting the importance of distilling the strong capabilities of the visual-textual teacher model into a visual-only student model. Additionally, we observe a saturation phenomenon in performance. Specifically, there is a significant performance increase when the text length grows from 0 to 512, but the improvement becomes less pronounced as the length increases from 2048 to 3096. Therefore, selecting an intermediate text length might offer a good balance between effectiveness and efficiency.

Visualization of Embeddings In this section, we analyze the effects of the distillation process by visualizing the document representations of the student and teacher models before and after distillation. Specifically, we sample 200 documents from the ChartQA dataset and apply t-SNE to these document representations.

As shown in Figure 4, the representations of the student model become much more aligned with those of the teacher model after distillation, confirming the effectiveness of the representation alignment technique. Additionally, the cosine similarity between the student and teacher models also increases. Consequently, after distillation, the student model is able to achieve performance similar to that of the teacher model without incurring additional computational costs from input text.

4.4 Case Study

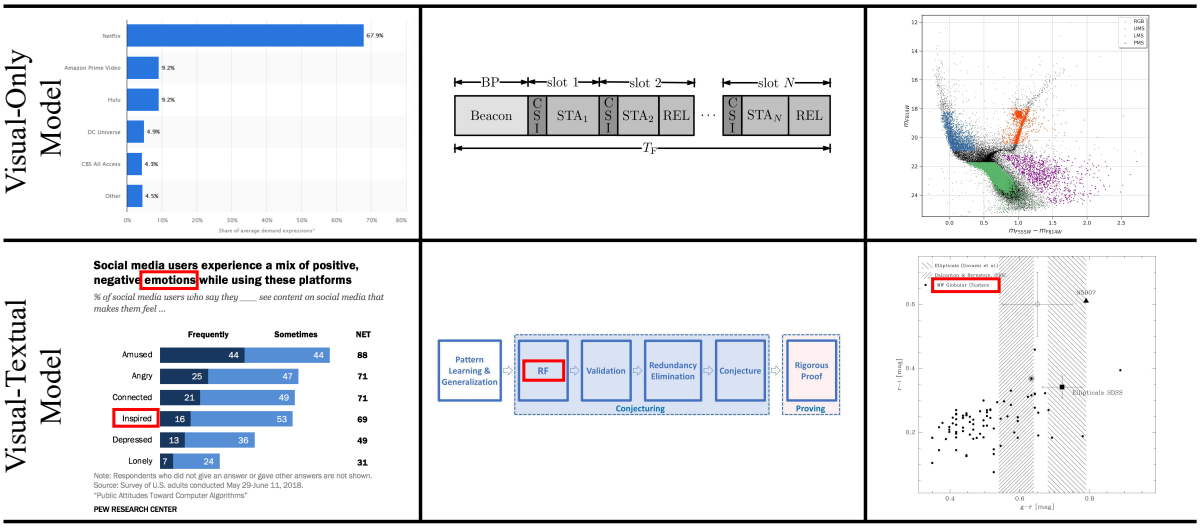
In this section, we analyze the effectiveness of the visual-textual model versus the visual-only model through several cases from the DocVQA and PlotQA datasets.

As illustrated in Figure 5, the visual-textual model demonstrates superior retrieval accuracy compared to its visual-only counterpart. The key distinction lies in their ability to capture semantic information: while the visual-only model primarily relies on visual patterns, potentially missing crucial textual context, the visual-textual model leverages both modalities to form a comprehensive understanding of the document content. This enhanced semantic comprehension directly translates to more accurate retrieval results.

5 Related Work

5.1 Multi-modal Document Retrieval

In multi-modal document understanding, models integrate visual and textual data to enhance information extraction. MPLUG-DocOwl (Ye et al.,



Question: What was the 4th most popular **emotion**?
Answer: Inspired

Question: What is the purpose of the "**RF**" block in the…
Answer: To provide a hypothesis

Question: Based on the points for **Globular Clusters**…
Answer: C

Figure 5: Case Study. We sample several cases from the DocVQA and PlotQA datasets to compare the performance of the visual-only method and the visual-textual method. The keywords that match between the questions and the content of the images are highlighted in red.

2023) introduces a modular multimodal large language model for OCR-free document understanding, leveraging both visual and textual content. MPLUG-DocOwl2 (Hu et al., 2024) extends this approach by focusing on high-resolution compression for multi-page documents. VISTA (Zhou et al., 2024) offers a method for visualized text embedding, enabling efficient multi-modal retrieval across document types. Unified multi-modal representations, such as in (Lee et al., 2024b), combine text and image features for improved retrieval and understanding. Document parsing challenges are also addressed by recent work on structured information extraction (Zhang et al., 2024), which focuses on methods for extracting and understanding document structures.

5.2 Multi-modal RAG

Multi-modal retrieval-augmented generation (RAG) models combine retrieval and generative techniques, leveraging both textual and visual information to enhance document processing tasks. VisRAG (Yu et al., 2024) uses vision-based retrieval to improve generative tasks such as document summarization by combining visual and textual content. M3DocRAG (Cho et al., 2024) extends RAG to multi-page, multi-document settings, improving the generation of summaries and answers by incorporating information from multiple document sources. M-Longdoc (Chia et al., 2024) introduces a retrieval-aware tuning framework that enhances the understanding of super-long documents by selecting relevant document segments for generation. Colpali (Faysse et al., 2024) applies vision-language models with retrieval for more efficient document retrieval, thereby boosting the quality of generation tasks. MM-Embed (Lin et al., 2024) proposes a unified framework for multimodal retrieval with LLMs, optimizing retrieval and generation for multi-modal documents.

6 Conclusion

In this paper, we identify that current text-based and visual-based methods lack adaptability across different scenarios. To overcome this, we introduce Unveil, a visual-textual embedding approach that integrates text and visual document features for enhanced semantic grounding. Our knowledge distillation technique transfers robust textual understanding from the visual-textual model to a purely visual model, allowing for parsing-free retrieval without sacrificing accuracy. Empirical results show that our visual-textual embedding method surpasses existing text-based and visual-based approaches. Additionally, the knowledge distillation bridges the gap between visual-textual and visual-only models, improving retrieval accuracy and efficiency.

Limitations

In this paper, we propose a multi-modal document retrieval framework that leverages both visual and textual information. We acknowledge a limitation in our approach: the visual-textual embedding model relies on textual inputs, necessitating OCR parsing of documents. This requirement can introduce additional computational overhead and may affect processing time, especially when dealing with large volumes of documents or when OCR accuracy is variable.

Ethics Statement

This research was conducted in full compliance with the ACL Ethics Policy. All datasets and large language models (LLMs) used for evaluation purposes are publicly available, ensuring transparency and reproducibility of our results. Our work is aimed at advancing multi-modal embedding techniques to improve document retrieval capabilities. We have carefully considered ethical implications and do not foresee any negative ethical impacts arising from our research.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP 2013*, pages 1533–1544.

Yew Ken Chia, Liying Cheng, Hou Pong Chan, Chaoqun Liu, Maojia Song, Sharifah Mahani Aljunied, Soujanya Poria, and Lidong Bing. 2024. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. *arXiv preprint arXiv:2411.06176*.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. 2020. Pp-OCR: A Practical Ultra Lightweight OCR System. *arXiv*, abs/2009.09941.

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

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*, pages 6609–6625. International Committee on Computational Linguistics.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: A benchmark for question answering research. *TACL 2019*, pages 452–466.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024a. Nv-embed: Improved techniques for training llms as generalist embedding models.

Jaewoo Lee, Joonho Ko, Jinheon Baek, Soyeong Jeong, and Sung Ju Hwang. 2024b. Unified multi-modal interleaved document representation for information retrieval. *arXiv preprint arXiv:2410.02729*.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal ArXiv: A Dataset for Improving Scientific Comprehension of Large Vision-Language Models. In *Proceedings of ACL*, pages 14369–14387.

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

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

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024b. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2425.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. Chartqa: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Proceedings of ACL*, pages 2263–2279.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2582–2591.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. 2021. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenBMB. 2024. openbmb/minicpm-v-2.

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

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8273–8288. Association for Computational Linguistics.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A Dataset for Document Visual Question Answering on Multiple Images. In *Proceedings of AAAI*, pages 13636–13645.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for Multipage DocVQA. *Pattern Recognition*, 144:109834.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

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. *arXiv preprint arXiv:1809.09600*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024a. Minicpm-V: A GPT-4v Level MLLM on Your Phone. *arXiv*, abs/2408.01800.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024b. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. 2023. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of ICCV*, pages 11941–11952.

Qintong Zhang, Victor Shea-Jay Huang, Bin Wang, Junyuan Zhang, Zhengren Wang, Hao Liang, Shawn Wang, Matthieu Lin, Wentao Zhang, and Conghui He. 2024. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *arXiv preprint arXiv:2410.21169*.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. Vista: Visualized text embedding for universal multi-modal retrieval. *arXiv preprint arXiv:2406.04292*.

A Dataset Statistics

Settings	NQ (Kwiatkowski et al., 2019)	TriviaQA (Joshi et al., 2017)	WebQ (Berant et al., 2013)	HotpotQA (Yang et al., 2018)	2WikiMultihopQA (Ho et al., 2020)	ASQA (Stelmakh et al., 2022)
Task	Open-domain QA	Open-domain QA	Open-domain QA	Multi-hop QA	Multi-hop QA	Ambiguous QA
Test Data	3,610	11,313	2,032	7,405	12,576	895
Metrics	Recall@10, MRR@10					

Table 5: Statistics and experimental settings of different tasks/datasets.

Settings	ArXivQA (Li et al., 2024)	ChartQA (Masry et al., 2022)	MP-DocVQA (Tito et al., 2023)	InfoVQA (Mathew et al., 2022)	PlotQA (Methani et al., 2020)	SlideVQA (Tanaka et al., 2023)
Task	Arxiv Figures	Charts	Industrial Documents	Infographics	Scientific Plots	Slide Decks
Test Data	8,640	718	1,879	2,046	11,307	1,640
Metrics	Recall@10, MRR@10					

Table 6: Statistics and experimental settings of different tasks/datasets.

B Training Details

Training Data In web page retrieval, we utilize 49,095 training pairs of query and positive documents. In visual document retrieval, we utilize 122,752 training pairs of query and positive documents.

Training Process We conducted full parameter fine-tuning during both stages. In the first stage, both student model and teacher were fine-tuned for 2 epochs with a learning rate of 2e-5 and a batch size of 16. In the second stage, the teacher model was frozen and the student was fine-tuned for 2 epochs with a learning rate of 2e-5 and a batch size of 16.

Model Inference After fine-tuning on the web page retrieval dataset, we tested the model on all the open-domain datasets, including open-domain QA datasets such as NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and WebQ (Berant et al., 2013), multi-hop datasets like Wikihop (Yang et al., 2018) and HotpotQA (Ho et al., 2020), as well as the ambiguous dataset ASQA (Stelmakh et al., 2022).

After fine-tuning on the visual document retrieval dataset, we tested the model on all the visual document datasets, including MP-DocVQA (Tito et al., 2023) for industrial documents, ArXivQA (Li et al., 2024), ChartQA (Masry et al., 2022), InfographicsVQA (Mathew et al., 2022), and PlotQA (Methani et al., 2020) for different types of figures, as well as SlideVQA (Tanaka et al., 2023).

C Document Parsing

Following (Yu et al., 2024), we use PaddlePaddle OCR (PPOCR) (Du et al., 2020) for document parsing. The process involves several stages:

1. **Text Detection**: A text detection model identifies text regions within the document and generates bounding boxes around them.
2. **Orientation Classification**: These detected regions are processed by a classification model to correct any orientation issues, such as rotation or flipping.
3. **Text Recognition**: A recognition model extracts the textual content from the corrected bounding boxes, returning the recognized text along with confidence scores. Only results with confidence scores above 0.6 are retained, and the bounding box coordinates, along with the recognized text, are stored for further processing.

Throughout this process, we apply a Layout Preserving policy. This approach maintains the original document structure by ordering the text boxes based on their spatial positions. Spaces and line breaks are dynamically inserted to reflect horizontal and vertical gaps between text regions. This ensures that the extracted text mirrors the original document layout, preserving its formatting in the final output.