LightSTAR: Efficient Visual Document Retrieval via Lightweight Selection with Vision-Adaptive Refinement

Tongkun Guan[1*], Haocheng Wang[1*], Wei Shen[1(✉)], and Xiaokang Yang[1]

MoE Key Lab of Artificial Intelligence, AI Institute, School of Computer Science, Shanghai Jiao Tong University, China
{gtk0615,wanghaocheng2023}@sjtu.edu.cn

Abstract. Visual document retrieval requires rapidly locating relevant pages from large multi-modal corpora in response to user queries. While recent methods powered by Multi-modal Large Language Models (MLLMs) show competitive accuracy, they suffer from prohibitive computational costs by applying intensive MLLM encoding to every single page. Meanwhile, we observe that user queries are typically keyword-anchored, containing semantically rich words that are expected to appear directly in the visible text of relevant pages, offering an efficient cue for quickly narrowing down candidate pages. Building on this insight, we propose LightSTAR, an efficient framework that decomposes visual document retrieval into: 1) LLM-free Visual Selection, which utilizes content-grounded query encoding to focus on informative words and employs LLM-free visual embeddings to produce a high-recall candidate set; and 2) Vision-adaptive Semantic Refinement, which further performs fine-grained semantic matching exclusively on these top candidates via adaptive region-wise feature fusion to effectively combine textual and layout cues, optimized through a hardness-aware contrastive objective. Experimental results demonstrate that LightSTAR achieves state-of-the-art retrieval accuracy while reducing end-to-end latency by several-fold, offering a highly practical solution to the accuracy-efficiency trade-off in visual document retrieval. Code is available at `https://github.com/bokufa/LightSTAR`.

Keywords: Visual Document Retrieval · Multi-modal Large Language Model · Visual Language Model

1 Introduction

The explosion of digital documents (from lengthy legal contracts to technical reports filled with diagrams and tables) is pushing retrieval systems beyond plain text. Modern applications such as retrieval-augmented generation (RAG) [30], enterprise search, and scientific assistance require processing hundreds or thousands of pages where crucial information is encoded not only in text but also in

*Equal contribution. ✉Corresponding author.

visual structure: layouts, figures, tables, and typography. This shift calls for retrieval systems capable of rapidly locating relevant pages from large multi-modal corpora. Such systems must not only understand complex visual semantics, but also operate efficiently at scale.

Traditional text-based retrieval systems struggle when crucial information is embedded in rich visual and structural elements, as they rely on imperfect text extraction and often lose critical layout and visual cues essential for understanding document semantics [51]. This gap has motivated the development of visual document retrieval [9], which operates directly on document images to preserve native layout and multi-modal cues. Recent breakthroughs in multi-modal large language models (MLLMs) [1, 31, 32] have enabled sophisticated approaches to this challenge. Methods such as ColPali [9] and VisRAG [49] leverage powerful MLLMs to embed document images and text queries into unified semantic spaces, employing late interaction mechanisms [26, 31, 37] to capture fine-grained cross-modal correspondences. While these MLLM-based approaches achieve competitive retrieval accuracy on challenging benchmarks [9, 20, 34], a major obstacle remains: they require MLLM-level computation for every page during indexing and retrieval. For long documents and frequently updated corpora, encoding all pages with a full MLLM is prohibitively expensive in latency, throughput, and cost [4, 7, 25]. This inefficiency stems from the fact that, for any given query, only a tiny fraction of pages are potentially relevant, yet existing systems apply the same heavy computation to the vast majority of obviously irrelevant pages. This calls for a selective computation in which inexpensive signals are used to prune large corpora before invoking costly multi-modal reasoning.

This paper is motivated by a key observation from real-world document retrieval scenarios: user queries are typically keyword-anchored [21, 43, 51]. They contain a compact set of semantically rich content words (specific entities, domain terminology, distinctive actions, or characteristic attributes) that are expected to appear on relevant pages, while most other pages lack such cues and can be filtered using much cheaper visual-text signals. This suggests that expensive MLLM computation can be reserved for refining a compact candidate set identified through lightweight filtering. Based on this principle, we propose **LightSTAR** (**Light**weight **S**elec**T**ion with vision-**A**daptive **R**efinement for efficient visual document retrieval), a novel framework that strategically allocates computation to where it matters. LightSTAR decomposes visual document retrieval into: **1) LLM-free Visual Selection.** We develop a lightweight, text-aware visual encoder, where query can directly interact with document images with a shared embedding space, to perform fast corpus-wide filtering and produce a small candidate set without introducing any large language model (LLM). To enhance retrieval effectiveness, we propose content-grounded query embedding, which applies linguistic analysis to filter semantically informative tokens while removing function words that are visually uninformative and ubiquitously distributed across documents. The filtered query tokens are then matched against document patch embeddings via a novel scale-adaptive late interaction scheme, enabling efficient, high-recall pruning at low cost. **2) Vision-adaptive Seman-**

tic Refinement. On the filtered candidates, we perform MLLM-based refinement for fine-grained semantic matching. We reuse the text-aware visual encoder to obtain semantically aligned representations. However, text-aware features alone may overlook complementary background and layout cues. To address this, we introduce an adaptive region-wise feature fusion strategy that preserves text-aware features on foreground/text regions while injecting layout-aware features on background/layout regions. Moreover, since candidates contain similar pages, we adopt a hardness-aware contrastive objective that emphasizes confusable negatives, encouraging more discriminative fine-grained representations.

Finally, our main contributions can be summarized as twofold:

- We introduce LightSTAR, which strategically decouples visual document retrieval into fast corpus-wide filtering via a lightweight LLM-free Visual Selection module and fine-grained candidate matching via Vision-adaptive Semantic Refinement.
- Experimental results demonstrate that LightSTAR achieves state-of-the-art retrieval quality while significantly improving latency efficiency compared to existing MLLM-based methods, making it practical for real-world deployments with large-scale document collections.

2 Related Work

Vision-language Models. Vision-language models have achieved remarkable progress in document understanding tasks [10, 12–19, 28, 40, 44–46]. Early works such as LayoutLM [48] and LayoutLMv2 [47] incorporate layout information by combining visual features with text embeddings extracted from OCR. More recent approaches adopt an OCR-free paradigm: Donut [27] employs an encoder-decoder architecture to directly generate text from document images, while Pix2Struct [29] pretrains on page screenshots for structured understanding. TokenFD [18] further advances OCR-free document understanding by introducing a token-level approach for fine-grained document feature extraction. The emergence of powerful generalist MLLMs—such as LLaVA [32], PaliGemma [3], Qwen-VL [2], and InternVL [6]—has further advanced document understanding capabilities. These models leverage large-scale pretraining to achieve strong vision-language alignment, enabling them to understand complex document layouts, extract textual content, and reason over visual elements.

Visual Document Retrieval. Visual document retrieval aims to retrieve relevant documents given text queries, where documents are represented as images containing mixed modalities—text, tables, figures, and complex layouts interleaved on a single page. Early approaches typically extract text via OCR engines and then apply text-based retrieval methods [39,41]. However, such pipelines suffer from OCR errors [51] and fail to capture visual layout information that is crucial for document understanding. Recent advances leverage multi-modal large language models to directly encode document images without relying on external OCR. ColPali [9] employs PaliGemma to generate multi-vector representations for document images and adopts late interaction [26] for query-document

matching, achieving strong retrieval performance. Follow-up works extend this paradigm with different MLLM backbones: ColQwen, a variant of ColPali, builds upon Qwen-VL [2] for enhanced multilingual support. VisRAG [49] investigates single-vector representations as an alternative to multi-vector approaches. Despite their effectiveness, these methods uniformly require LLM-based encoding for the entire corpus during indexing. Our work addresses this limitation via the combination of a lightweight LLM-free Visual Selection module (efficiently filters candidates) and a MLLM-based Vision-Adaptive Refinement module (distinguishes the subtle semantic differences of the candidates), achieving both scalability and accuracy.

3 Methodology

Problem Formulation. Visual document retrieval has emerged as a critical capability for processing real-world multi-modal documents (*e.g.*, technical reports, legal contracts, and scientific manuscripts) that often span hundreds or thousands of visually rich pages. For instance, in RAG pipelines, efficient first-stage evidence selection over large page collections is a prerequisite for any downstream reasoning, making the ability to rapidly and accurately identify relevant pages from large-scale document corpora directly determines the effectiveness and scalability of the entire system.

Specifically, we formalize this visual document retrieval task as a ranking problem over document images. Given a natural language query q and a collection of document pages $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, where each d_i represents a document image containing rich visual and textual information, the objective of visual document retrieval is to rank all pages $\mathcal{D}$ by their relevance to the query:

$$[d_{\pi_1}, d_{\pi_2}, \ldots, d_{\pi_N}] = \underset{d_i \in \mathcal{D}}{\operatorname{argsort}}\, \mathcal{S}(q, d_i), \tag{1}$$

where $\mathcal{S} : (q, d_i) \to \mathbb{R}$ is a similarity scoring function that quantifies the semantic correspondence between query q and document page d_i. π is a permutation such that $\mathcal{S}(q, d_{\pi_1}) \geq \mathcal{S}(q, d_{\pi_2}) \geq \cdots \geq \mathcal{S}(q, d_{\pi_N})$.

To address this problem, existing approaches predominantly rely on encoding each document page with powerful MLLMs [9, 49], leveraging their sophisticated capacity for fine-grained visual understanding to produce rich semantic representations. While this strategy effectively yields superior retrieval quality, it introduces a critical computational bottleneck: applying large-scale MLLMs to every page in a long document results in substantial computational overhead and prohibitive latency, limiting its practicality for large-scale or time-sensitive applications. This motivates our focus on retrieval methods that efficiently narrow the search space over long multi-modal documents without sacrificing the semantic fidelity required for accurate retrieval.

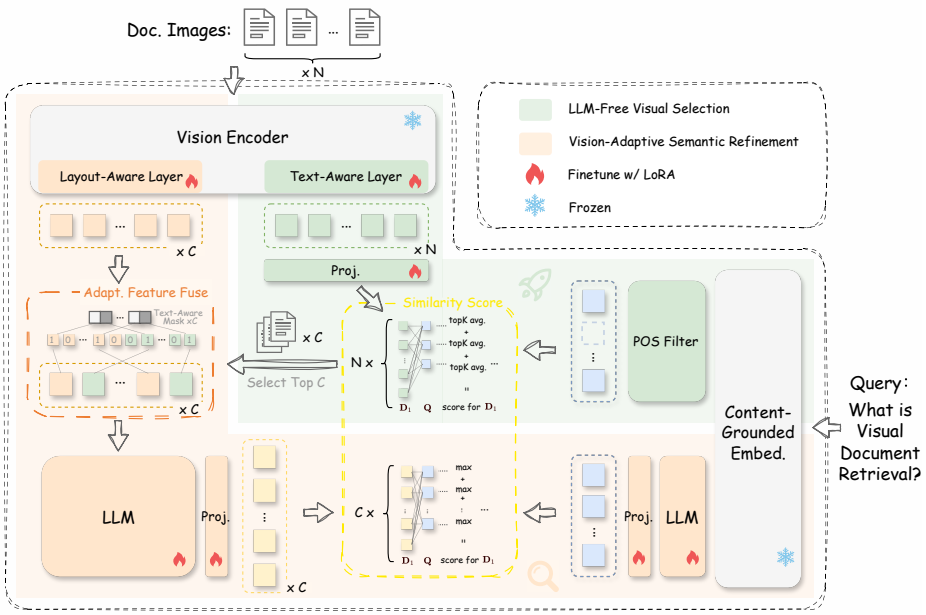

Fig. 1: Overview of the proposed method LightSTAR. Our method decomposes retrieval into LLM-Free Visual Selection and Vision-Adaptive Semantic Refinement, achieving state-of-the-art accuracy with lowest latency cost.

3.1 Overview

As formulated above, the core challenge lies in balancing retrieval accuracy against computational efficiency when processing long multi-modal documents. To address this, we begin with a key empirical observation: in practical scenarios, user queries typically correspond to only a small subset of pages within a document collection. Because queries are often keyword-anchored and contain distinctive textual cues such as domain-specific terms expected on relevant pages, pages lacking these fundamental cues can be efficiently filtered using accurate visual signals, obviating the need for expensive semantic modeling. This observation suggests that by first identifying a small set of likely relevant pages using inexpensive visual-text signals, expensive full semantic modeling can be selectively applied only where it is truly needed.

Building on this insight, we propose LightSTAR, an effective framework that strategically decouples retrieval into:

LLM-free Visual Selection. Without introducing any LLM, we develop a lightweight, text-aware visual encoder to rapidly filter the document collection and identify a promising candidate set $\mathcal{D}_c \subset \mathcal{D}$ where $|\mathcal{D}_c| \ll |\mathcal{D}|$, achieving computational complexity of $\mathcal{O}(N \cdot C_{\text{VE}})$ with $C_{\text{VE}} \ll C_{\text{MLLM}}$. C_{VE} and C_{MLLM} denote the complexity of the visual encoder and MLLM, respectively;

Vision-adaptive Semantic Refinement. The filtered candidate set $\mathcal{D}_c$ is further processed by a powerful MLLM to perform fine-grained semantic matching, reducing complexity to $\mathcal{O}(|\mathcal{D}_c| \cdot C_{\text{MLLM}})$ where $|\mathcal{D}_c| \ll N$. Additionally, it shares

the same visual encoder as the first-stage LLM-free Visual Selection module, eliminating redundant parameters and lowering the memory footprint, which further enhances computational efficiency.

This design achieves overall complexity of $\mathcal{O}(N \cdot C_{\text{VE}} + |\mathcal{D}_c| \cdot C_{\text{MLLM}})$, yielding substantial speedups when $|\mathcal{D}_c|/N$ is small, while retaining most of the semantic expressiveness compared to full MLLM-based retrieval with complexity $\mathcal{O}(N \cdot C_{\text{MLLM}})$. This architecture provides a practical approach to managing the efficiency–quality trade-off in large-scale retrieval within long multi-modal documents. By strategically allocating computation where it matters most, LightSTAR fully leverages the efficiency of lightweight encoders and the accuracy of MLLMs, making it practical for large-scale deployment. Figure 1 illustrates the complete pipeline, where the two modules synergistically combine efficiency and accuracy to enable practical large-scale visual document retrieval.

3.2 LLM-free Visual Selection

To reduce computational overhead in large-scale document image retrieval, this module performs rapid candidate selection over the entire document collection without introducing any large language model. Instead, it relies exclusively on lightweight text-aware visual representations that capture visual-text signal for relevance estimation.

Content-grounded Query Embedding. In real-world retrieval scenarios, user queries are typically formulated to explicitly specify the key information needed to locate target documents. Such queries naturally contain semantically rich content words expected to appear in the visual textual content of relevant pages [21, 38, 41, 43]. We also conduct an analysis on existing datasets which shows that the vast majority of queries contain visually observable content tokens in their ground-truth pages (Please see Appendix for details). However, natural language queries inevitably include function words (*e.g.*, *the*, *of*, *and*, *to*, *is*) that, while grammatically necessary, carry minimal semantic content and rarely contribute to visual document matching. Function words introduce noise in visual matching because they appear ubiquitously across documents regardless of relevance, creating spurious similarity signals. Treating function words equally with content words dilutes the discriminative power of query representations.

To obtain a content-grounded query representation, we apply part-of-speech (POS) tagging [33] to extract semantically informative tokens while filtering out function words. Specifically, we retain only semantically informative tokens while discarding function words unlikely to contribute to visual relevance:

$$q' = \{\, t_i \in q \mid \text{POS}(t_i) \notin \{\text{CC}, \text{IN}, \text{DT}\} \ \wedge \ t_i \notin \{\mathcal{B}, \mathcal{P}\}\}, \tag{2}$$

where $\text{POS}(\cdot)$ denotes the POS tagger output. The set $\{\text{CC}, \text{IN}, \text{DT}\}$ contains the excluded POS categories (coordinating conjunctions, prepositions/subordinating conjunctions, and determiners). $\mathcal{B}$ and $\mathcal{P}$ denote predefined sets of *be*-verbs and punctuation tokens, respectively. The filtered query q' is encoded via embedding lookup as $\mathbf{Q}$, each retained word t_i is mapped to its embedding vector $\mathbf{q}_i$:

$$\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{n_q}]^\top = f_{\text{norm}}(f_{\text{text}}(q')) \in \mathbb{R}^{n_q \times d_v}, \tag{3}$$

where $f_{\text{text}}(\cdot)$ denotes the text embedding function, including tokenization and embedding lookup, n_q is the number of query tokens, and d_v denotes the hidden dimension of the visual encoder.

This linguistically motivated filtering removes visually uninformative function words while preserving semantically meaningful content tokens. Since only a small set of high-frequency function words are discarded, the semantic structure of the query remains largely intact. The filtering rule is lightweight and can be adapted to other languages by adjusting the set of function-word categories.

Text-aware Visual Embedding. Given a document image d of size $H \times W$, we feed it into a visual encoder built on a Vision Transformer (ViT) [8] backbone. The image is partitioned into non-overlapping patches of size $p \times p$, producing $n_d = \frac{HW}{p^2}$ visual patch tokens. Each patch is processed through the ViT layers, followed by a lightweight projection layer that maps features into a shared embedding space aligned with textual representations, and then normalized for similarity calculation:

$$\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_{n_d}]^\top = f_{\text{norm}}(f_{\text{proj}}(f_{\text{ViT}}(d))) \in \mathbb{R}^{n_d \times d_v}, \tag{4}$$

where $\mathbf{D}$ denotes the output passage embedding, and $[\mathbf{d}_1, \ldots, \mathbf{d}_{n_d}]^\top$ contains the text-aware visual embedding for each patch.

To bootstrap our visual encoder with strong multi-modal priors, we initialize it from a pretrained multi-modal large language model. And to enable the encoder to extract text-rich representations directly from raw document images, we pretrain it following a visual-text alignment objective [18], which leverages token-mask pairs to establish fine-grained correspondence between image patches and textual embeddings. The encoder is subsequently fine-tuned for document retrieval while preserving learned visual-textual alignment.

Scale-adaptive Late Interaction. Given query embeddings $\mathbf{Q} \in \mathbb{R}^{n_q \times d_v}$ and document embeddings $\mathbf{D} \in \mathbb{R}^{n_d \times d_v}$, we compute a document-level similarity score via late interaction. A standard approach [26] applies max-pooling over document tokens for each query token. However, this is fragile: a single visually similar but semantically irrelevant patch can dominate the score, leading to spurious matches.

We address this with a scale-adaptive aggregation mechanism. For each query token $\mathbf{q}_i$, we select the top-k most similar document patches and average their similarities:

$$\mathcal{S}(q, d) = \sum_{i=1}^{n_q} \frac{1}{k} \sum \{topK(\{\langle \mathbf{q}_i, \mathbf{d}_1 \rangle, \ldots, \langle \mathbf{q}_i, \mathbf{d}_{n_d} \rangle\}, k)\}, \tag{5}$$

where $topK(\mathbf{x}, k)$ is a function that returns top-k elements in $\mathbf{x}$, and

$$k = \max\left(1, \lfloor \gamma \cdot n_d \rfloor\right), \quad \gamma \in (0, 1). \tag{6}$$

This design serves two main purposes. (1) It stabilizes the matching score by aggregating the top-k similarities rather than relying on a single maximum response. This reduces sensitivity to noisy activations or accidental local matches, mitigating score fluctuation and producing a more robust matching signal. (2) It

allows k to scale with document complexity. For high-resolution or visually dense documents, a larger k aggregates evidence from multiple relevant regions and further suppresses noise. For simpler documents with fewer patches, a smaller k preserves fine-grained precision and avoids over-smoothing discriminative signals.

Optimization Objective. To train the visual encoder to produce embeddings that maximize $\mathcal{S}(q, d)$ for relevant query-document pairs, we employ in-batch contrastive learning. Given a training batch of B query–document pairs $\{(q_i, d_i^+)\}_{i=1}^B$, where d_i^+ denotes the ground-truth relevant document for query q_i, we adopt in-batch contrastive learning. All other documents in the batch serve as negatives for each query. The loss function is formulated as:

$$\mathcal{L}_{sel} = \frac{1}{B}\sum_{i=1}^{B} \log\left(1 + \exp\left(\max_{j\neq i}\mathcal{S}(q_i, d_j) - \mathcal{S}(q_i, d_i^+)\right)\right), \tag{7}$$

This formulation encourages the model to rank the positive document above all in-batch negatives. During training, we apply Low-Rank Adaptation (LoRA) [22] to the final ViT layer and the projector, keeping all other parameters frozen. This minimizes training cost while achieving strong retrieval performance.

3.3 Vision-adaptive Semantic Refinement

While the LLM-free Visual Selection module efficiently narrows the search space from N pages to a manageable candidate set $\mathcal{D}_c$ ($|\mathcal{D}_c| \ll N$), we further introduce MLLM-based refinement to model the semantic nuance. This stage builds upon our proposed visual encoder to inject richer semantic representations into the filtered candidates, circumventing the computational bottleneck of processing the entire corpus. Specifically, we encode the query q with an embedding layer and an LLM. For document d, visual features are first extracted via a visual extraction module and adapted via an MLP before being fed into the LLM. Both modalities are ultimately projected into a shared normalized embedding space for similarity computation:

$$\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_{n_q}] = \mathcal{F}_{\text{query}}(q) \quad \in \mathbb{R}^{n_q \times d_p}, \tag{8}$$

$$\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_{n_d}] = \mathcal{F}_{\text{doc}}(d) \quad \in \mathbb{R}^{n_d \times d_p}, \tag{9}$$

where d_p represents the dimension of the projected vector space. The encoding pipelines are defined as $\mathcal{F}_{\text{query}} = f_{\text{norm}} \circ f_{\text{proj}} \circ f_{\text{LLM}} \circ f_{\text{text}}$ and $\mathcal{F}_{\text{doc}} = f_{\text{norm}} \circ f_{\text{proj}} \circ f_{\text{LLM}} \circ f_{\text{MLP}} \circ f_{\text{fuse}}$.

Then, we sum the maximum similarity of each query token across all document tokens to compute the similarity score via $\mathcal{S}(q, d) = \sum_{i=1}^{n_q} \max_{1\leq j\leq n_d} \langle \mathbf{q}_i, \mathbf{d}_j \rangle$.

To prevent excessive parameter growth and redundant computation, the refinement module reuses the visual encoder from the Visual Selection module mentioned in Sec. 3.2 via weight sharing. Furthermore, we introduce an adaptive feature fusion mechanism (f_{fuse}) to integrate visual representations from both stages, fostering effective cross-stage synergy. The detailed design is as follows.

Adaptive Visual Feature Fusion. Our architecture shares the lower 1 to $\ell-1$ ViT layers between the two modules. Given a document image d, we first extract the shared intermediate visual features $\mathbf{H}_{<\ell} = \mathcal{T}_{1:\ell-1}(d)$. For the final ℓ-th layer, we introduce a dual-branch design. Alongside the text-aware layer $\mathcal{T}_\ell^{\text{text}}$ inherited from the Visual Selection module, we instantiate a parallel ViT layer $\mathcal{T}_\ell^{\text{layout}}$ initialized from the original pretrained MLLM weights (see Figure 1). This produces two complementary visual feature streams:

1) Text-Aware Features ($\mathbf{F}_{\text{text}}$): Inherited from the LLM-free Visual Selection module, these features are optimized for visual-textual alignment and encode rich textual semantics:

$$f_{\text{ViT}} : \mathbf{F}_{\text{text}} = \mathcal{T}_\ell^{\text{text}}(\mathbf{H}_{<\ell}) \in \mathbb{R}^{n_d \times d_v}. \tag{10}$$

2) Layout-Aware Features ($\mathbf{F}_{\text{layout}}$): Retaining the general-purpose visual modeling capacity of the original MLLM, these features better capture non-textual structural semantics such as layouts, tables, figures, and whitespace:

$$\mathbf{F}_{\text{layout}} = \mathcal{T}_\ell^{\text{layout}}(\mathbf{H}_{<\ell}) \in \mathbb{R}^{n_d \times d_v}. \tag{11}$$

Since the text-aware layer is optimized to match textual content, $\mathbf{F}_{\text{text}}$ is highly responsive to text-dense regions. For accurate retrieval, both textual content and document structure (*e.g.*, table layouts, figure positions) are also critical. To reconcile this, we propose an adaptive fusion mechanism that spatially multiplexes the features based on regional characteristics.

The key insight is that patches in $\mathbf{F}_{\text{text}}$ with low textual alignment naturally exhibit high similarity to "space" semantics. We exploit this property to separate text-dominant regions from background-dominant regions. Specifically, we compute the similarity between $\mathbf{F}_{\text{text}}$ and a predefined whitespace token embedding $\mathbf{q}_{\text{space}} \in \mathbb{R}^{d_v}$ (obtained from the MLLM's vocabulary):

$$\mathbf{s} = \mathbf{F}_{\text{text}} \mathbf{q}_{\text{space}}^\top \in \mathbb{R}^{n_d}. \tag{12}$$

We normalize the similarity scores to $[0, 1]$ via min-max scaling to yield $\hat{\mathbf{s}}$, and apply an empirical threshold $\tau \in [0, 1]$ to generate a binary spatial mask $\mathbf{m}$:

$$\mathbf{m} = \mathbb{I}\left[\hat{\mathbf{s}} > \tau\right] \in \{0, 1\}^{n_d}, \tag{13}$$

where $\mathbb{I}[\cdot]$ is the indicator function. Here, $m_j = 1$ indicates a background-dominant patch, while $m_j = 0$ corresponds to a text-dominant patch.

Finally, the fused feature $\mathbf{F}_{\text{fused}}$, which is also the output embedding representation of function f_{fuse}, is computed as:

$$f_{\text{fuse}} : \mathbf{F}_{\text{fused}} = (\mathbf{1} - \mathbf{m}) \odot \mathbf{F}_{\text{text}} + \mathbf{m} \odot \mathbf{F}_{\text{layout}}, \tag{14}$$

where $\odot$ denotes element-wise multiplication with broadcasting along the feature dimension, and $\mathbf{1}$ is a vector of ones. As defined in Eq. (9), $\mathbf{F}_{\text{fused}}$ is subsequently fed into the MLP connector and the LLM backbone to output the contextualized document embeddings $\mathbf{D} = f_{\text{norm}} \circ f_{\text{proj}} \circ f_{\text{LLM}} \circ f_{\text{MLP}}(\mathbf{F}_{\text{fused}})$ for the final similarity computation.

This adaptive fusion strategy provides three distinct advantages: (1) it reuses representations via shared lower layers, avoiding redundant computational overhead; (2) it preserves the sharp text-sensitivity of the Visual Selection module for content-rich regions; and (3) it enriches the overall representation with essential structural and layout clues, yielding a comprehensive document understanding for accurate retrieval.

Hardness-aware Optimization Objective. After LLM-free Visual Selection, the candidate set $\mathcal{D}_c$ is enriched with plausibly relevant documents, making the retrieval task significantly more challenging than the initial filtering. Many candidates share similar textual content or visual layouts, requiring the retrieval to learn fine-grained distinctions. Standard contrastive learning treats all negatives equally, which is suboptimal when some negatives are far more confusable than others.

To train MLLM-based Vision-adaptive Refinement to focus on resolving these challenging cases, we adopt a hardness-weighted variant of InfoNCE loss [36]. Given a batch of B query–document pairs $\{(q_i, d_i^+)\}_{i=1}^B$, we compute all pairwise similarities within the batch. The loss is formulated as:

$$\mathcal{L}_{ref} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\mathcal{S}(q_i, d_i^+))}{\exp(\mathcal{S}(q_i, d_i^+)) + \sum_{j\neq i} w_{ij}\cdot\exp(\mathcal{S}(q_i, d_j))}, \tag{15}$$

where the hardness weight w_{ij} for each negative pair (q_i, d_j), $j \neq i$, is defined as:

$$w_{ij} = \mathrm{sg}(\exp(\mathcal{S}(q_i, d_j))), \tag{16}$$

where $\mathrm{sg}(\cdot)$ means computing with stop-gradient to prevent gradient flow.

This formulation adaptively emphasizes hard negatives: documents with high similarity to the query (i.e., confusing cases) receive exponentially larger weights, forcing the model to learn more discriminative representations for challenging distinctions. Easy negatives with low similarity contribute minimally to the loss, allowing efficient training focus on the decision boundary.

During refinement training, we apply LoRA to the transformer layers of the language model backbone and the output projection layer, following the same parameter-efficient strategy as LLM-free Visual Selection part. We freeze the shared ViT layers to maintain the visual-textual alignment learned in LLM-free Visual Selection, and only update the newly added parallel ViT branch, the MLP connector, the language model, and the projection layer via LoRA. This ensures stable training while adapting the model for fine-grained retrieval.

4 Experiments

Implementation Details. For LLM-free Visual Selection, we adopt InternViT-300M [11] as the visual encoder backbone. The model is fine-tuned with a batch size of 32 for 10 epochs on the training dataset [9]. The base learning rate is set to 5×10^{-4}. We apply LoRA with rank $r = 32$ and scaling factor $\alpha = 32$. For Vision-adaptive Semantic Refinement, we build upon InternVL3 [52] as the

vision-language backbone. The model is trained with a batch size of 80 for 10 epochs on the same training dataset. The base learning rate is also set to 5×10^{-4}. LoRA is applied with the same rank configuration. All training experiments are conducted on 4 NVIDIA A800 80GB GPUs. Additional implementation details and hyperparameter analyses are provided in Appendix.

Datasets. We conduct experiments on the ViDoRe benchmark [9], a large-scale visual document retrieval dataset consisting of diverse document types. ViDoRe provides documents as pure images, which requires joint modeling of textual content, layout structure, and visual cues. For scalability and latency analysis, we further construct multiple subsets by sampling from the full ViDoRe dataset, with sizes ranging from 500 to 7000 document pages. These subsets are used exclusively for measuring end-to-end retrieval latency and are not involved in accuracy evaluation.

Metrics. We evaluate retrieval performance using Recall@K and NDCG@K, following prior work on visual document retrieval. Recall@K measures candidate coverage by checking whether at least one relevant document appears within the top-K results, while NDCG@K evaluates ranking quality by considering both relevance and position. We primarily report Recall@100 for the Visual Selection module, as its objective is to preserve as many relevant documents as possible within a manageable candidate pool for subsequent refinement. For the final retrieval performance, we report NDCG@5 to focus on ranking quality.

Latency Evaluation. To evaluate retrieval efficiency, we measure end-to-end query latency including query/document encoding, similarity computation, and candidate refinement during retrieval. All methods are evaluated on a single NVIDIA A800 80GB GPU with a batch size of 16 under the same dataset and resolution settings to ensure fair comparison.

4.1 Main Results

Overall Accuracy. Table 1 presents the main retrieval results of LightSTAR across ViDoRe sub-datasets. Among the compared methods, LightSTAR achieves the highest overall average NDCG@5 score of 89.1. Compared with conventional text-based OCR pipelines and vision-language contrastive models, LightSTAR achieves substantial gains across all sub-datasets. Among MLLM-based methods, LightSTAR achieves state-of-the-art performance on 5 out of 8 sub-datasets, outperforming the previous best model ColQwen2.5 with an average NDCG@5 of 88.8. Crucially, LightSTAR achieves this top-tier performance with only 2B parameters, offering substantially better inference efficiency than existing MLLM-based retrievers.

Latency and Scalability. As shown in Figure 2, we compare end-to-end retrieval latency and scalability of LightSTAR with three representative MLLM-based methods (VisRAG-Ret, ColQwen2.5, and ColPali), which show competitive retrieval accuracy, to provide a comprehensive scalability analysis. While all methods exhibit an approximately linear increase in latency with corpus size, LightSTAR maintains a much flatter growth than all the baselines, demonstrating superior efficiency and scalability for large corpora. Its latency increases

Table 1: Performance comparison of LightSTAR with existing methods on the ViDoRe benchmark. Results are presented using NDCG@5(%). Latency (s) denotes end-to-end retrieval time measured on a corpus of 5,000 document pages as reported in Figure 2. The best results are in **bold** and the second best are underlined.

Method	#Params	ArxivQ	DocQ	InfoQ	TATQ	AI	Energy	Gov.	Health.	Avg.(↑)	Latency(↓)
Text-based Methods											
BM25 [39]	–	40.1	38.4	70.0	61.5	88.0	84.7	82.7	89.2	69.3	–
BGE-M3 [5]	569M	35.7	32.9	71.9	43.8	88.8	83.3	80.4	91.3	66.0	–
Vision-Language Contrastive Models											
CLIP [37]	151M	26.5	14.6	51.7	4.7	22.9	32.4	39.8	37.5	28.8	–
SigLIP [50]	883M	50.2	31.3	69.7	27.5	67.8	73.5	75.3	83.1	59.8	–
Nomic-Embed-Vision [35]	92M	17.1	10.7	30.1	2.7	12.9	10.9	11.4	15.7	13.9	–
MLLM-based Models											
VLM2Vec [24]	4B	42.8	26.7	66.7	21.4	53.5	63.5	64.0	70.7	51.2	–
E5-V [23]	8B	48.3	34.7	69.2	29.3	78.9	78.1	82.2	82.3	62.9	–
VisRAG-Ret [49]	3B	80.6	42.7	85.0	49.3	94.4	92.0	87.7	92.7	78.0	1219.9
ColPali [9]	3B	79.1	54.4	81.8	65.8	96.2	91.0	92.7	94.4	81.9	257.3
Nomic-Embed-Multimodal [42]	3B	**87.3**	59.6	89.8	69.4	97.5	93.3	95.7	97.9	86.3	–
ColQwen2.5	2B	87.2	63.3	91.3	79.5	98.6	**96.3**	**96.3**	98.0	88.8	466.6
Ours											
LightSTAR	2B	86.2	**63.5**	**91.7**	**81.0**	**99.5**	95.4	95.8	**99.8**	**89.1**	**123.9**

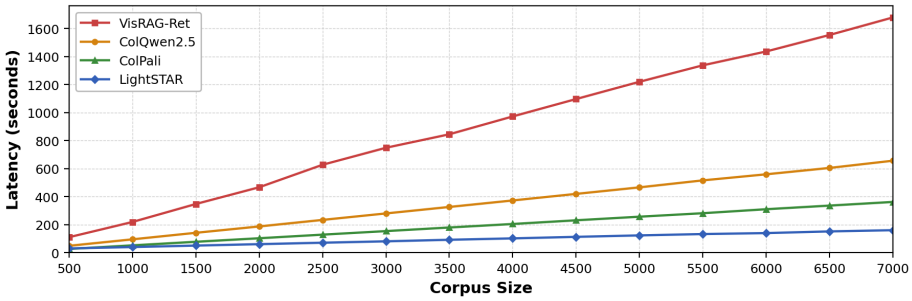

Fig. 2: End-to-end retrieval latency comparison between LightSTAR with three competitive MLLM-based retrievers (VisRAG-Ret, ColQwen2.5, and ColPali) as corpus size increases from 500 to 7,000 document pages. At 7,000 pages, LightSTAR is 10.4× faster than VisRAG-Ret, 4.1× faster than ColQwen2.5, and 2.3× faster than ColPali, demonstrating its computational efficiency for large-scale visual document retrieval.

from 31.1 s at 500 pages to only 161.1 s at 7000 pages, yielding the smallest slope among all methods. At 7,000 pages, LightSTAR is 10.4× faster than VisRAG-Ret (1678.9/161.1), 4.1× faster than ColQwen2.5 (656.7/161.1), and 2.3× faster than ColPali (362.6/161.1). These results confirm that our architecture effectively controls computational cost by restricting expensive reasoning to a small top candidate set.

Recall Analysis of the LLM-free Visual Selection Module. To highlight the fundamental role of our LLM-free Visual Selection in synergy with the refinement module, we examine its candidate coverage against existing LLM-free methods. As shown in Table 2, our selection module achieves an average Recall@100 of 97.8, substantially outperforming conventional text-based retrievers and vision-language contrastive models, while remaining fully LLM-free. These results show that our 542M lightweight visual encoder attains the highest candi-

Table 2: Performance comparison of the LLM-Free Visual Selection module with existing methods on the ViDoRe benchmark. Results are presented using Recall@100(%) metrics to show the recall coverage. The best results are in **bold** and the second best are underlined.

Method	#Params	ArxivQ	DocQ	InfoQ	TATQ	AI	Energy	Gov.	Health.	Avg.(↑)
Text-based Methods										
BM25 [39]	–	71.1	70.9	90.3	96.1	98.0	99.0	95.0	**100**	90.1
BGE-M3 [5]	569M	75.1	72.2	94.9	93.4	**100**	98.0	98.0	**100**	91.5
Vision-Language Contrastive Models										
CLIP [37]	151M	71.6	60.8	92.2	50.5	71.0	83.0	87.0	83.0	74.9
SigLIP [50]	883M	89.6	72.6	95.4	84.1	89.0	96.0	98.0	99.0	91.2
Nomic-Embed-Vision [35]	92M	66.8	53.9	71.7	41.3	49.0	52.0	46.0	50.0	53.8
Ours										
LLM-Free Visual Selection Module	542M	**96.8**	**88.9**	**98.6**	**98.9**	99.0	**100**	**100**	**100**	**97.8**

date coverage among LLM-free alternatives, preserving nearly all relevant pages within a fixed candidate budget and thus providing an indispensable foundation for subsequent refinement.

4.2 Ablation Studies

Ablation on the Overall Retrieval Architecture. Table 3 shows a comprehensive ablation results with three configurations: (1) selection-only without refinement, (2) refinement-only over the full corpus, and (3) the complete pipeline. The refinement-only setting achieves the highest NDCG@5 of 89.3, but incurs a latency increase of nearly three times, since it processes the entire corpus. In contrast, the selection-only configuration is significantly faster but suffers from a large performance drop (80.8 NDCG@5), indicating that visual similarity alone is insufficient for precise ranking. In contrast, the full LightSTAR pipeline achieves 89.1 NDCG@5, only 0.2 lower than refinement-only, while dramatically reducing inference cost. This demonstrates that the LLM-free Visual Selection module effectively preserves relevant candidates, allowing the refinement module to operate on a compact subset with negligible impact on ranking quality and achieving a superior trade-off between effectiveness and efficiency.

Ablation on Module Components. Table 4 analyzes the impact of different components in different modules.

1) Components of the LLM-free Visual Selection Module. Removing the content-grounded embedding leads to a 0.8 drop in Recall@100. This confirms that filtering function words and grounding embeddings in content-bearing tokens effectively reduces spurious visual-text alignments and stabilizes visual matching. Replacing the proposed scale-adaptive late interaction with a standard late interaction approach [26] which uses max-pooling aggregation causes a 1.0 decrease. This suggests that simple maximum similarity is highly sensitive to accidental high-scoring patches, resulting in unstable document-level matching. In contrast, adaptive top-k aggregation mitigates spurious matches and produces more robust similarity estimation.

2) Components of the Vision-adaptive Semantic Refinement Module. Removing the proposed region-adaptive fusion mechanism without introducing layout-

Table 3: Ablation study of LightSTAR modules. Performance is measured by average NDCG@5 on ViDoRe and end-to-end retrieval latency(s) measured on a corpus of 5,000 document pages. "Selection" and "Refinement" denote the LLM-free Visual Selection Module and Vision-adaptive Semantic Refinement Module, respectively.

Selection	Refinement	ViDoRe	Latency
✓	×	80.8	86.8
×	✓	89.3	465.1
✓	✓	89.1	123.9

Table 4: Ablation study of LightSTAR module components. Performance is measured by average Recall@100 and NDCG@5 on ViDoRe. "CG-Embed" and "SA-LateInt" denote Content-Grounded Embedding and Scale-Adaptive Late Interaction, respectively; "Feat-Fusion" denotes Adaptive Feature Fusion, × means using only text aware features; "HA-Obj" denotes Hardness-Aware Objective, × means using Eq.(7) as objective.

(a) Visual Selection Module

CG-Embed	SA-LateInt	ViDoRe (Recall@100 %)
✓	×	96.8
×	✓	97.0
✓	✓	97.8

(b) Semantic Refinement Module

Feat-Fusion	HA-Obj	ViDoRe (NDCG@5 %)
✓	×	87.9
×	✓	88.3
✓	✓	89.1

aware features reduces NDCG@5 from 89.1 to 88.3, indicating that the features generated by layout-aware ViT layer serve as complementary signals for text-optimized features. Leveraging the layout/background information captured by layout-aware features with adaptive fusion enables more accurate discrimination between subtle structural differences. Eliminating the hardness-aware objective reduces performance from 89.1 to 87.9. This demonstrates that emphasizing hard negatives during training enhances the model's ability to distinguish highly similar candidate documents.

5 Conclusion

In this work, we presented LightSTAR, a practical and efficient framework for visual document retrieval that strategically addresses the fundamental accuracy-efficiency trade-off faced by existing MLLM-based approaches. By leveraging the key insight that user queries are typically keyword-anchored, we decompose the retrieval process into two synergistic modules: a lightweight LLM-free Visual Selection module, followed by Vision-adaptive Semantic Refinement. Our extensive experiments demonstrate that LightSTAR achieves state-of-the-art retrieval accuracy while reducing end-to-end latency by several-fold compared to prior MLLM-based methods. Beyond immediate practical impact, LightSTAR establishes a new paradigm for visual document retrieval: rather than applying uniform heavy computation across all documents, intelligent computational allocation based on query characteristics and document properties offers a more sustainable path forward.

Acknowledgements

This work was supported by NSFC 62322604 and NSFC 62576207.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Bińkowski, M.a., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 23716–23736. Curran Associates, Inc. (2022)
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
3. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., Unterthiner, T., Keysers, D., Koppula, S., Liu, F., Grycner, A., Gritsenko, A., Houlsby, N., Kumar, M., Rong, K., Eisenschlos, J., Kabra, R., Bauer, M., Bošnjak, M., Chen, X., Minderer, M., Voigtlaender, P., Bica, I., Balazevic, I., Puigcerver, J., Papalampidi, P., Henaff, O., Xiong, X., Soricut, R., Harmsen, J., Zhai, X.: Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726 (2024)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
5. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: M3-embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 2318–2335. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024)
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
7. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 16344–16359. Curran Associates, Inc. (2022)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)

9. Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., HUDELOT, C., Colombo, P.: Colpali: Efficient document retrieval with vision language models. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) International Conference on Learning Representations. vol. 2025, pp. 61424–61449 (2025)
10. Fu, P., Guan, T., Wang, Z., Guo, Z., Duan, C., Sun, H., Chen, B., Jiang, Q., Ma, J., Zhou, K., Luo, J.: Multimodal large language models for text-rich image understanding: A comprehensive review. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025. pp. 19941–19958 (2025)
11. Gao, Z., Chen, Z., Cui, E., Ren, Y., Wang, W., Zhu, J., Tian, H., Ye, S., He, J., Zhu, X., et al.: Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. arXiv preprint arXiv:2410.16261 (2024)
12. Guan, T., Gu, C., Lu, C., Tu, J., Feng, Q., Wu, K., Guan, X.: Industrial scene text detection with refined feature-attentive network. IEEE Transactions on Circuits and Systems for Video Technology **32**(9), 6073–6085 (2022)
13. Guan, T., Gu, C., Tu, J., Yang, X., Feng, Q., Zhao, Y., Shen, W.: Self-supervised implicit glyph attention for text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15285–15294 (June 2023)
14. Guan, T., Lin, C., Shen, W., Yang, X.: Posformer: recognizing complex handwritten mathematical expression with position forest transformer. In: European Conference on Computer Vision. pp. 130–147. Springer (2025)
15. Guan, T., Shen, W., Yang, X.: CCDPlus: Towards accurate character to character distillation for text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
16. Guan, T., Shen, W., Yang, X., Feng, Q., Jiang, Z., Yang, X.: Self-supervised character-to-character distillation for text recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19473–19484 (2023)
17. Guan, T., Shen, W., Yang, X., Wang, X., Yang, X.: Bridging synthetic and real worlds for pre-training scene text detectors. In: European Conference on Computer Vision. pp. 428–446. Springer (2024)
18. Guan, T., Wang, Z., Fu, P., Guo, Z., Shen, W., Zhou, K., Yue, T., Duan, C., Sun, H., Jiang, Q., et al.: A token-level text image foundation model for document understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23210–23220 (2025)
19. Guan, T., Yang, Z., Wan, J., Yang, M., Guo, Z., Hu, Z., Luo, R., Chen, R., Jiang, S., Wang, P., et al.: Codepercept: Code-grounded visual stem perception for mllms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 33542–33552 (2026)
20. Günther, M., Sturua, S., Akram, M.K., Mohr, I., Ungureanu, A., Wang, B., Eslami, S., Martens, S., Werk, M., Wang, N., Xiao, H.: jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In: Adelani, D.I., Arnett, C., Ataman, D., Chang, T.A., Gonen, H., Raja, R., Schmidt, F., Stap, D., Wang, J. (eds.) Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025). pp. 531–550. Association for Computational Linguistics, Suzhuo, China (Nov 2025)
21. Guo, J., Xu, G., Cheng, X., Li, H.: Named entity recognition in query. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 267–274. SIGIR '09, Association for Computing Machinery, New York, NY, USA (2009)

22. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. ICLR (2022)
23. Jiang, T., Song, M., Zhang, Z., Huang, H., Deng, W., Sun, F., Zhang, Q., Wang, D., Zhuang, F.: E5-v: Universal embeddings with multimodal large language models. arXiv preprint arXiv:2407.12580 (2024)
24. Jiang, Z., Meng, R., Yang, X., Yavuz, S., Zhou, Y., Chen, W.: VLM2vec: Training vision-language models for massive multimodal embedding tasks. ICLR (2025)
25. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
26. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In: Huang, J.X., Chang, Y., Cheng, X., Kamps, J., Murdock, V., Wen, J., Liu, Y. (eds.) Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020. pp. 39–48. ACM (2020)
27. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII. p. 498–517. Springer-Verlag, Berlin, Heidelberg (2022)
28. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
29. Lee, K., Joshi, M., Turc, I.R., Hu, H., Liu, F., Eisenschlos, J.M., Khandelwal, U., Shaw, P., Chang, M.W., Toutanova, K.: Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 18893–18912. PMLR (23–29 Jul 2023)
30. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 9459–9474. Curran Associates, Inc. (2020)
31. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900. PMLR (17–23 Jul 2022)
32. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916. Curran Associates, Inc. (2023)
33. Loper, E., Bird, S.: Nltk: The natural language toolkit. arXiv preprint arXiv:cs/0205028 (2002)
34. Muennighoff, N., Tazi, N., Magne, L., Reimers, N.: MTEB: Massive text embedding benchmark. In: Vlachos, A., Augenstein, I. (eds.) Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 2014–2037. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023)

35. Nussbaum, Z., Duderstadt, B., Mulyar, A.: Nomic embed vision: Expanding the latent space. arXiv preprint arXiv:2406.18587 (2024)
36. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2019)
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021)
38. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (Apr 2009)
39. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M., et al.: Okapi at trec (1994)
40. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
41. SPARCK JONES, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation **28**(1), 11–21 (01 1972)
42. Team, N.: Nomic embed multimodal: Interleaved text, image, and screenshots for visual document retrieval (2025)
43. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)
44. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
45. Wang, Z., Guan, T., Fu, P., Duan, C., Jiang, Q., Guo, Z., Guo, S., Luo, J., Shen, W., Yang, X.: Marten: Visual question answering with mask generation for multi-modal document understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14460–14471 (2025)
46. Wen, C., Peng, Z., Huang, Y., Shen, W.: Efficient segmentation with multimodal large language model via token routing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 10593–10602 (2026)
47. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., Zhou, L.: LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 2579–2591. Association for Computational Linguistics, Online (Aug 2021)
48. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 1192–1200. KDD '20, ACM (Aug 2020)
49. Yu, S., Tang, C., Xu, B., Cui, J., Ran, J., Yan, Y., Liu, Z., Wang, S., Han, X., Liu, Z., Sun, M.: Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) In-

ternational Conference on Learning Representations. vol. 2025, pp. 21074–21098 (2025)
50. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11975–11986 (October 2023)
51. Zhang, J., Zhang, Q., Wang, B., Ouyang, L., Wen, Z., Li, Y., Chow, K.H., He, C., Zhang, W.: Ocr hinders rag: Evaluating the cascading impact of ocr on retrieval-augmented generation. arXiv preprint arXiv:2412.02592 (2025)
52. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., Gao, Z., Cui, E., Wang, X., Cao, Y., Liu, Y., Wei, X., Zhang, H., Wang, H., Xu, W., Li, H., Wang, J., Deng, N., Li, S., He, Y., Jiang, T., Luo, J., Wang, Y., He, C., Shi, B., Zhang, X., Shao, W., He, J., Xiong, Y., Qu, W., Sun, P., Jiao, P., Lv, H., Wu, L., Zhang, K., Deng, H., Ge, J., Chen, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)