Beyond Bag-of-Patches: Learning Global Layout via Textual Supervision for Late-Interaction Visual Document Retrieval

Anonymous ACL submission

Abstract

Visual Document Retrieval (VDR) models mostly rely on late interaction architectures, in which documents are represented by a set of local patch embeddings and then matched against query tokens. While efficient, this architecture prioritizes local similarity over global layout structure of documents to estimate relevancy between documents and query. In practice, documents may contain locally relevant regions while remaining globally irrelevant. This leads to errors as relevance originates from layout structure of documents with heterogeneous layouts combining figures, tables, and text. We make document layout learnable without changing inference. We propose a visual encoder that augments local patch representations with a global layout embeddings. These embeddings are trained through descriptor-guided structural supervision that distills page-level layout priors into the retrieval space. Across four ViDoRe-v2 datasets, our model improves over the strongest architecturally comparable ColPali/ColQwen baseline by +2.4 nDCG@5 and +2.3 MAP@5, with statistically significant per-dataset gains over ColQwen.

1 Introduction

Document retrieval is a core component of retrieval-augmented generation (RAG), where retrieval quality directly determines factual grounding (Lewis et al., 2020; Mei et al., 2025; Wu et al., 2025). Documents in practical domains exhibit heterogeneous layouts combining text, tables, and figures (Dong et al., 2025; Peng et al., 2025). This incentivizes models to integrate semantically and visually diverse features. However, recent embedding models fail to capture page-level semantics in visually structured pages since they are primarily optimized for text-dominant corpora (Mei et al., 2025; Abootorabi et al., 2025).

A promising solution is *visual document retrieval (VDR)* models, which treat document pages as images to avoid expensive OCR (Faysse et al., 2025; Xu et al., 2025; Team, 2025; Günther et al., 2025). These VDR models often adopt *late interaction* (a.k.a *MaxSim*), representing each page as a set of patch embeddings and computing relevance via a sum of maximum similarities between query tokens and page patch tokens (Khattab and Zaharia, 2020). While being effective for capturing fine-grained local evidence such as table cells, figure regions, or typographic cues, their *MaxSim* formulation ignores spatial relationships between matched regions and biases retrieval toward isolated local matches.

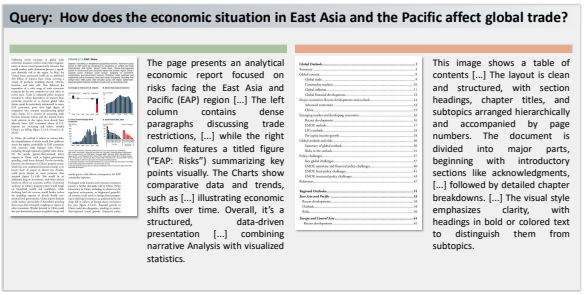

Figure 1: An example of a layout-rich query. The relevant page (left) requires cross-region reasoning over text, figures, and layout. The distractor (right) shares lexical overlap but is globally irrelevant because it is a table of contents. Holistic page descriptors train the model to encode page-level structure missed by patch-only MaxSim scoring.

In practical domains, relevance in visually rich documents depends also on *page-level structure*, i.e., how textual and visual elements interact across the page. As illustrated in Figure 1, answering the query requires linking narrative text about economic risks with corresponding figures, rather than matching either in isolation. This exposes a fundamental challenge as pages may contain locally relevant elements while remaining globally irrelevant. As a result, page-level structure signals, which may be partially encoded by vision encoders,

remain weakly expressed in the retrieval space after late-interaction scoring. We argue that this is a structural limitation of patch-based retrieval and empirically show that global relevance cannot be recovered from local similarity alone.

We propose a late-interaction encoder with a trainable `[CLS]`-style global token that is jointly optimized with local patches to capture page-level layout structure. The core idea lies in *descriptor-guided structural supervision*. We train the global token using *holistic page descriptors*, instead of standard image–caption objectives. These automatically-generated descriptors summarize page-level organization and are used only during training, encouraging the model to encode structural relationships that enable retrieval in layout-rich domains where relevance depends on cross-region interactions. At inference, descriptors are discarded since the model indexes only the page images plus one learned global vector, preserving the standard visual retrieval pipeline with negligible overhead.

We experiment on the four datasets of Visual Document Retrieval Benchmark V2 (ViDoRe-v2). We compare our model with `ColPali`/`ColQwen` and larger models such as `Nemo-3b`. To study the effect of descriptor-guided supervision, we introduce two `Cross-Context` methods as baselines. These methods inject textual descriptors at inference and act as an oracle for explicit access to global structure information. Our model improves over `ColPali`/`ColQwen` by **+2.4 nDCG@5** and **+2.3 MAP@5**, with gains concentrated on layout-rich pages. Our model also performs on par with or surpasses the Cross-Context oracle without requiring descriptors at inference. This indicates that our method can capture global structural signals not merely as exploitable features, but as parameterized retrieval prior.

Contributions. Our key contributions are **(1)** We identify a limitation of late-interaction retrieval for visual documents, where the MaxSim lacks a mechanism to model page-level structure; **(2)** We show that late-interaction retrieval fails to represent cross-region structure; so we introduce a learnable global bottleneck to encode it, **(3)** We show that this training-only structural signal improves retrieval without requiring descriptors at inference, yielding consistent gains over comparable `ColPali`/`ColQwen` baselines, especially on layout-rich pages.

2 Method

Late-interaction retrieval lacks a mechanism to encode global layout, as patch embeddings are optimized independently and matched locally. We address this by introducing a global structural bottleneck that enables page-level context to influence retrieval. This context includes layout, spatial organization, and distributed visual patterns that simple aggregations (e.g., mean pooling) cannot capture, as we show empirically.

Figure 2 shows an overview of our method. We augment the multi-vector visual encoder with a trainable global token and descriptor-guided structural supervision. The descriptor introduces a compressed latent summary of relational structure, forcing the global token to become a structure-preserving bottleneck. The global token acts as an aggregation variable over local patches through self-attention. By attending to all patches simultaneously, it encodes spatial relationships that are otherwise lost in independent local patch matching. This consolidates page-level layout information into a single representation rather than leaving it implicitly distributed across patches.

2.1 Training

Multi-Vector Encoder with Trainable Global Token. We build on a Vision Large Language Model (VLLM) that projects a document page into a sequence of patch embeddings $\mathbf{I} \in \mathbb{R}^{L_p \times d}$. While the transformer backbone allows for inter-patch attention during encoding, the late-interaction similarity function matched at retrieval time operates on patches independently, effectively treating the page as a "bag-of-patches." To preserve and aggregate global page-level semantics, we append a trainable global token $\mathbf{g} \in \mathbb{R}^d$ (randomly initialized) to the visual sequence $\mathbf{I}' = [\mathbf{I}; \mathbf{g}] \in \mathbb{R}^{(L_p+1) \times d}$. The same shared parameter $\mathbf{g}$ is also appended to the textual descriptor sequence, as described below. By passing $\mathbf{I}'$ through the self-attention layers of the language model, the output hidden state of $\mathbf{g}$ (denoted $\mathbf{g}_v$) acts as a structural bottleneck. It is trained to aggregate cross-region relationships that would otherwise be lost during the independent MaxSim matching phase. As a result, global layout information is explicitly represented in the multi-vector index and can influence retrieval alongside local patch embeddings.

Holistic Page Descriptors as Layout Supervision. To train the global token $\mathbf{g}$ to represent layout pri-

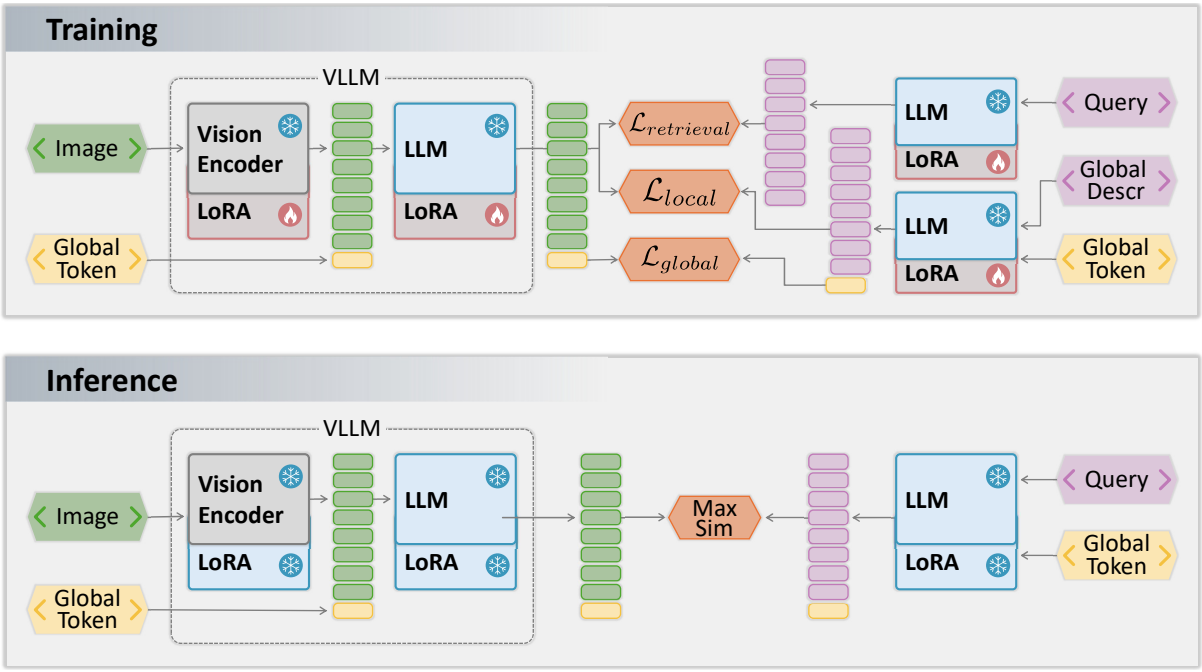

Figure 2: A high-level view of our approach. **Top (Training):** We append a trainable global token to the vision-encoder patch embeddings and contextualize it jointly with local features in VLLM. Training optimizes three objectives: a retrieval late-interaction max-similarity loss between query and document patches, a local max-similarity loss between holistic page descriptors and document patches, and a global alignment loss using cosine similarity logits aligning the image-derived global token with the descriptor-derived global representation. **Bottom (Inference):** Only image and query inputs are used. The global token is appended to both sequences and contextualized without requiring textual descriptors. Document relevance is scored using late-interaction similarity.

ors, we supervise it using holistic page descriptors. These descriptors (Figure 1) summarize a page's spatial organization (e.g., the positioning of columns, tables, and headers) while omitting fine-grained textual content. By abstracting local details, we force the global token to encode structural layout rather than a redundant representation of local patches. Further details about generation of these descriptors are in Appendix 7.2.

Aligning Global Token to Textual Descriptions. We enforce structural consistency at the page level by aligning the visual global token with the descriptor's textual representation. The textual descriptor is passed through the same Large Language Model (LLM) backbone of the VLLM. It ensures that the text encoder and vision-language encoder share the same semantic vocabulary and places both visual and textual tokens in a shared embedding space. The embedding layer of the LLM encodes the textual descriptor into a sequence $\mathbf{E}_{\text{desc}} \in \mathbb{R}^{L_T \times d}$, where L_T is the number of textual tokens in desc. The same shared global token $\mathbf{g}$ is appended to $\mathbf{E}_{\text{desc}}$. After processing through the self-attention layers, its output hidden state serves as the textual structural token $\mathbf{g}_{\text{desc}} \in \mathbb{R}^d$. We apply the InfoNCE objective, which is a contrastive loss, averaged over the batch, to align $\mathbf{g}_v$ with $\mathbf{g}_{\text{desc}}$:

$$\mathcal{L}_{\text{global}} = -\frac{1}{B}\sum_{i=1}^{B}\left(s_{ii} - \log\sum_{j=1}^{B}\exp(s_{ij})\right), \tag{1}$$

where $s_{ij} = \frac{\cos(\mathbf{g}_v^i, \mathbf{g}_{\text{desc}}^j)}{\tau}$, $\mathbf{g}_v^i$ and $\mathbf{g}_{\text{desc}}^i$ denote the output hidden state of the shared global token $\mathbf{g}$ on the visual and descriptor sequences of the i-th sample, respectively. τ is a temperature hyperparameter and B is the batch size. By using in-batch negatives, the global token learns the unique layout semantics of a page relative to other pages. Unlike standard image-caption matching, this loss trains the visual global token to serve as a latent proxy for the high-level layout priors, effectively distilling the descriptor's structural semantics into the visual embedding space.

Aligning Textual Descriptors with Local Patch Embeddings. To complement the global token's macro-level awareness, we enforce fine-grained structural alignment at the patch level. We optimize local patch embeddings $\mathbf{I} \in \mathbb{R}^{L_p \times d}$ to align with

the textual descriptor embeddings $\mathbf{E}_{\text{desc}} \in \mathbb{R}^{L_g \times d}$:

$$\mathcal{L}_{\text{local}} = -\sum_{k=1}^{L_p} \max_{1 \le j \le L_g} \cos(\mathbf{i}_k, \mathbf{e}_j), \qquad (2)$$

which encourages each patch to match the most relevant descriptor token. By grounding local patches in the structural context provided by the descriptor, we ensure the model maintains structural coherence even when queried at a granular level.

Joint Objective for Global and Local Structured Retrieval. We combine our structural alignment objectives with the standard multi-vector retrieval loss $\mathcal{L}_{\text{retrieval}}$, which aligns patch embeddings $\mathbf{I}$ directly with query embeddings $\mathbf{Q}$ via late interaction. Thus, the final joint objective is:

$$\mathcal{L} = \mathcal{L}_{\text{global}} + \mathcal{L}_{\text{local}} + \mathcal{L}_{\text{retrieval}} \qquad (3)$$

This objective couples global structure with local evidence, enabling both to influence the learned representation space. To maintain computational efficiency, we restrict updates to LoRA adapters within the VLLM transformer layers, keeping the base vision encoder and LLM weights frozen. Both levels contribute to the final retrieval score, enabling the model to jointly reason over local visual evidence and global page structure. The auxiliary objectives introduce negligible overhead, as they operate on existing hidden states via cosine similarity.

2.2 Inference and Scoring

At inference, our model operates as a purely visual retriever and does not require any textual descriptors. The trained LoRA adapters remain part of the model (frozen after training) and no adapters are removed. This makes the method computationally viable for large-scale RAG settings where real-time descriptor generation would be prohibitively expensive. Both queries and documents are encoded as multi-vector representations augmented with global tokens, $\mathbf{Q} \in \mathbb{R}^{(L_q+1) \times d}$ and $\mathbf{D} \in \mathbb{R}^{(L_d+1) \times d}$. The same shared parameter $\mathbf{g}$ is appended to both sequences. Although $\mathbf{g}$ is trained only via document and descriptor pairs, the shared LLM backbone allows it to aggregate contextual information from query tokens in the same embedding space where document layout was aligned during training. The resulting output representations will differ based on their respective input contexts, analogously to how a `[CLS]` token produces different but meaningful representations for different inputs. We compute the retrieval score using the late-interaction function:

$$s(\mathbf{Q}, \mathbf{D}) = \sum_{i=1}^{L_q+1} \max_{1 \le j \le L_d+1} \mathbf{Q}_i \mathbf{D}_j^\top, \qquad (4)$$

where $\mathbf{Q}$ and $\mathbf{D}$ are the query and document embedding sequences (including global tokens).

Each query token, including the global token, can attend to all document tokens via MaxSim. Local patch similarities capture fine-grained evidence and the global token captures coarse layout structure. Together, they influence the final similarity score. This mechanism allows structural priors to directly influence the similarity score without increasing computational complexity beyond the addition of a single vector per document.

3 Experiments

3.1 Datasets

We evaluate `Global-Descr-Embed` (Section 2) on ViDoRe-v2 (Macé et al., 2025), the current standard for assessing visual retrieval models under realistic document layouts. Table 1 summarizes the dataset statistics. We evaluate on four diverse domains: economic reports, biomedical materials, and ESG industry disclosures (ESGH and ESGR). These datasets exhibit heterogeneous structures, including multi-column text, embedded charts, and complex tabular data.

For fine-tuning, we use the training set of `ColPali` (Faysse et al., 2025), consisting of 118k training and 500 development samples. Each instance pairs an English text query with a relevant document page. We augment each instance with a holistic page descriptor via one-time offline preprocessing (Appendix 7.2), enabling controlled evaluation of retrieval under diverse layout conditions.

3.2 Evaluation Metrics

We report Normalized Discounted Cumulative Gain (`nDCG`) and Mean Average Precision (`MAP`).

Dataset	Docs	Queries	Pages	Pages / Query
Economics	5	232	452	15.6
Biomedical	27	640	1,016	3.2
ESGH	27	52	1,538	2.5
ESGR	30	228	1,538	3.9

Table 1: The dataset statistics of ViDoRe-v2.

nDCG@k measures ranking quality by penalizing relevant documents that appear lower in the results list. This is critical for RAG, where top-ranked pages dominate generation quality. MAP@k measures retrieval consistency by averaging precision across all relevant pages for a given query. It captures a model's ability to retrieve all relevant pages even in use cases with long-tail queries. For all experiments, we report metrics at $k = 5$, following the default setting established by the ViDoRe benchmark.

3.3 Methods Compared

To study the impact of structural supervision, we compare Global-Descr-Embed with two architectural variants and state-of-the-art visual retrievers.

Cross-Context-Frozen (CrossContext-ZS). This training-free variant concatenates visual patches with pre-generated textual descriptors and passes them through the VLLM (Figure 3), enabling joint attention over local and global semantics. Because it uses textual descriptors during retrieval, it acts as an oracle baseline for descriptor informativeness. While computationally expensive, it reveals to what extent holistic page descriptors benefit retrieval.

Cross-Context-FineTuned (CrossContext-FT). To examine whether global structure can be encoded by the model parameters when holistic page descriptors are available, we fine-tune the cross-context variant on the same vidore/colpali_train_set used for Global-Descr-Embed training. The descriptor tokens are concatenated with the visual patch tokens and passed through the VLLM, enabling joint contextualization of local visual evidence and page-level descriptor information. Unlike Global-Descr-Embed, this model uses no separate trainable global token; the descriptor tokens serve as the source of global structural context. We update only Low-Rank Adapters (LoRA) parameters and optimize with the standard late-interaction retrieval objective using in-batch negatives, with early stopping on the development split.

State-of-the-art (SOTA) Methods. We compare Global-Descr-Embed with leading open-source VDR systems on ViDoRe-v2, including our primary baseline ColQwen2.5 (Faysse et al., 2025) (same architecture and pretraining), as well as llama-nemoretriever-colembed (Xu et al., 2025), colnomic-embed-multimodal (Team, 2025), and ColPali (Faysse et al., 2025), all evaluated without additional fine-tuning for comparability.

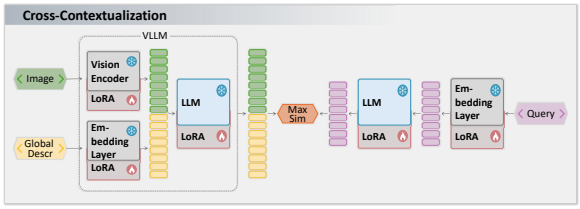

Figure 3: The cross-context baseline for frozen and fine-tuned variants. Patch-level image embeddings are concatenated with embeddings of tokens of the holistic page descriptor.

3.4 Implementation Details

The vision and text encoders are initialized from ColQwen2.5 pretrained LoRA adapters and weights. The global token embedding is initialized randomly and added to the model's vocabulary. Base encoder weights remain frozen; only LoRA parameters and the retrieval projection head are updated. We fine-tune Global-Descr-Embed for 6 epochs with a batch size of 128 using AdamW with a learning rate of 5×10^{-5}, weight decay 10^{-4}, and a linear warm-up over the first 100 steps. The full implementation details are in Appendix 7.3.

4 Results

Main Findings. Overall, our evaluation reveals four primary trends that characterize the impact of holistic page descriptors on retrieval performance: **(1)** Global-Descr-Embed consistently outperforms all architecturally comparable baselines. We observe a +2.4 point improvement in average nDCG@5 and a +2.3 point gain in average MAP@5 over our primary baseline, ColQwen2.5. **(2)** These gains concentrate on layout-rich datasets (e.g., ESGH), demonstrating the importance of page-level layout signals. **(3)** Ablations reveal that global and local cues are individually insufficient and so their synergy drives the gains. **(4)** Improvements scale with spatial heterogeneity and multi-region reasoning demands.

4.1 Document Retrieval

Table 2 reports retrieval effectiveness across the four ViDoRe-v2 datasets.

Global-Descr-Embed achieves the strongest

Method	Size	Economics		Biomedical		ESGH		ESGR		Avg	
		nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP
Architecturally and size comparable											
ColPali (Faysse et al., 2025)	2B	47.5	17.7	55.2	46.1	52.2	46.1	48.5	36.1	50.9	36.5
Colnomic-3b (Team, 2025)	3B	54.3	22.5	**62.6**	**53.2**	64.3	57.9	56.4	43.5	59.4	44.3
ColQwen2.5-T	3B	53.0	21.0	58.4	49.2	62.0	54.8	51.9	39.6	56.3	41.2
ColQwen2.5 (Faysse et al., 2025)	3B	55.9	23.4	61.2	52.1	68.1	60.4	55.9	43.8	60.3	45.0
Global-Descr-Embed (ours)	3B	**58.2**†	**25.2**†	61.7†	52.4†	**73.4**†	**67.1**†	**57.3**†	**44.6**†	**62.7**	**47.3**
Architecturally different OR significantly larger											
Colnomic-7b (Team, 2025)	7B	55.5	23.7	62.9	54.3	69.8	62.4	54.1	43.1	60.6	45.9
Nemo-3b (Xu et al., 2025)	4.4B	57.8	24.4	62.7	53.9	75.4	68.3	57.6	45.5	63.5	48.0

Table 2: We report performance on ViDoRe-v2 using nDCG@5, MAP@5, and their means. The first block contains architecturally comparable multi-vector retrievers with similar parameter scale (2–3B), representing our primary baselines. Global-Descr-Embed improves consistently over the strongest 3B ColPali-family baseline (ColQwen2.5). The second block includes larger or architecturally different models (4–7B) trained with different objectives and is shown only for reference. ESGH and ESGR are the human-labeled and Restaurant version of ESG dataset. ColQwen2.5-T is ColQwen2.5 where supplied with OCR-based textual representations of documents. Results marked with † are statistically significant (Wilcoxon Signed-Rank test, $p < 0.05$) compared to the ColQwen2.5.

Method	Economics		Biomedical		ESGH		ESGR		Avg	
	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP
Global-Descr-Embed	58.2	25.2	61.7	52.4	73.4	67.1	57.3	44.6	**62.7**	**47.3**
w/o local patch embeddings	49.2	18.0	43.8	35.8	28.1	22.8	19.5	12.9	35.2	22.4
w/o query global token	58.1	24.9	62.2	52.9	71.6	65.0	57.7	45.3	62.4	46.9
w/o image global token	58.4	25.1	61.9	52.6	72.5	66.2	57.4	44.9	62.6	47.2
Fine-tuned w/o $\mathcal{L}_{\text{global}}$	55.0	23.5	60.2	51.4	69.7	62.2	57.0	45.1	60.5	45.6
Fine-tuned w/o $\mathcal{L}_{\text{local}}$	57.8	24.9	60.9	52.0	71.0	64.1	57.5	45.1	61.8	46.5
Mean-Pooling	38.7	28.7	17.1	17.6	25.5	14.0	22.6	13.1	11.7	15.3
Max-Pooling	16.3	13.6	6.6	10.2	11.7	4.0	9.9	4.6	7.7	6.5
Median-Pooling	38.5	26.1	16.9	16.7	24.6	13.9	20.5	12.5	10.9	14.4

Table 3: Ablation study on ViDoRe-v2. The Mean, Max, and Median poolings are deterministic ways to induce global representations from local patch embeddings.

average performance among all ColPali- and ColQwen-based systems, which are closest counterparts in terms of architecture and training setup. It gains +2.4 nDCG@5 and +2.3 MAP@5 points compared with ColQwen2.5, showing that the proposed page-level structure modeling encodes a complementary structural prior not captured by patch-level similarity.

When comparing with large multimodal retrievers such as Nemo-3b (4.4B) and ColNomic-7B; our 3B-parameter model delivers competitive average performance, and in some cases exceeds ColNomic-7B, in particular on layout-rich datasets ESGH and ESGR. This suggests improved parameter efficiency as our 3B model outperforms a 7B model by leveraging structural inductive biases. These patterns indicate that descriptor-guided structural alignment contributes robustness that is not recovered by scaling parameters and training data alone. The weak results of the OCR-augmented ColQwen2.5-T confirm that visual and layout signals drive retrieval accuracy on multimodal pages more than textual transcription alone. These findings show that a compact model with explicit structural bias can match or surpass substantially larger encoders on visually heterogeneous documents.

4.2 Ablation Study

Table 3 contrasts the contribution of the global token and local patch embeddings as well as alternative pooling strategies for aggregating local patch embeddings to obtain global embeddings. while we report per-dataset scores for completeness, we focus our analysis on the mean (Avg) scores, which are more reliable than dataset-specific fluctuations.

Removing local patch embeddings sharply degrades performance (nDCG: -27.5, MAP: -24.9). In this configuration, the model relies solely on

the global token embeddings, which encode page-level layout. This lacks the granularity to precisely match localized elements such as individual words, numbers, or icons. This confirms that local patches remain the primary evidence for fine-grained matching. Ablating the global tokens yields small but consistent declines (nDCG: -0.3, MAP: -0.4). The small magnitude of this drop suggests that fine-tuned local patches partially learn to capture structural signals of the document. The explicit token still provides the optimal anchor and adds information that improves performance on layout-rich pages. Ablating either the query- or image-side global token yields similar drops (nDCG≤ -0.4). This symmetry suggests that the benefit of global structure comes from aligning query intent with document layout, not from either side alone.

Effects of the Loss Functions Removing the global-token alignment loss (w/o L_{global}) drops performance (nDCG: -2.2, MAP: -1.7), showing that explicit supervision of the global representation is the larger contributor in this ablation. Dropping the local-descriptor alignment loss (w/o L_{local}) also lowers performance (nDCG: -0.9, MAP: -0.8), indicating that patch-level structural grounding is essential for retrieval as well. The larger impact of $\mathcal{L}_{\text{global}}$ on training compared with removing the global token at inference (-0.1 to -0.3 nDCG) indicates that the global token functions primarily as a training-time structural bottleneck, shaping patch representations through layout supervision rather than contributing solely via an additional inference-time vector.

Pooling is an Inadequate Substitute. Mean, Max, and Median pooling variants of Global-Descr-Embed show a substantial performance drop, confirming that naively collapsing local patches into a single vector removes crucial structure signals and cannot replace trained global representations.

4.3 Impact of Global Descriptors

Table 4 compares Global-Descr-Embed to its two Cross-Context counterparts (CrossContext-ZS and CrossContext-FT). These models require holistic page descriptors at inference. Despite not using descriptors at inference, our Global-Descr-Embed outperforms CrossContext-ZS and performs on par with CrossContext-FT, indicating that page layout information extraction is learned via encoder parameters.

Method	nDCG	MAP
CrossContext-ZS	61.7	46.4
CrossContext-FT	62.5	47.1
Global-Descr-Embed	**62.7**	**47.3**

Table 4: Comparing Global-Descr-Embed with CrossContext-ZS and CrossContext-FT.

Feature	Failures	Improvements
Text	11.56	14.14
Image	2.91	3.31
List	1.47	2.80
N Words	289.27	380.53
N VE	2.24	3.24

Table 5: Summary of the mean of certain document page features of failure cases and improvements of our method compared to its baseline ColQwen2.5 on the ESGH dataset

4.4 Case Study: Layout Complexity in ESGH

We use ESGH as a focused case study because it contains the most structurally complex layouts in ViDoRe-v2, including dense text regions, visual elements, lists, and spatially dispersed content, as illustrated in Figure 4 and summarized in Section 4.4.

As shown in Figure 4, simple pages with contiguous text boxes offer limited visual or spatial variation, whereas complex layouts combine images, multiple text regions, and large tables, requiring integration of visual semantics, textual cues, and layout relationships. Section 4.4 shows that improvement cases consistently have more text regions, visual elements, lists, and higher content density than failure cases, reflecting the same complexity signals—high spatial entropy and element diversity—linked to larger gains in earlier analysis. These patterns align with our quantitative findings: Global-Descr-Embed excels when spatial-semantic reasoning across heterogeneous, multi-region layouts is needed, but offers smaller gains on simpler, text-only pages where local matching suffices. This case study thus confirms that descriptor-guided structural supervision delivers its strongest benefits in visually rich, structurally complex documents.

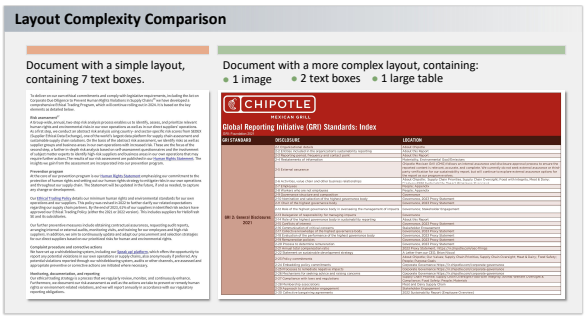

Figure 4: Examples of ESGH page layouts: a simple, text-only page (top) and a complex, multi-region layout with mixed modalities (bottom).

5 Related Work

We discuss related work in two groups: OCR-based and visual document retrieval.

OCR-based Retrieval Recent OCR-based methods (Wang et al., 2024; Wei et al., 2025, 2024; Niu et al., 2025) typically follow two stages. First they convert pages into textual and structural tokens, such as words, lines, tables, and detected layout regions, using an OCR engine. Retrieval is then performed in the text space with methods such as dense text encoders (Reimers and Gurevych, 2019), BM25 (Robertson et al., 1994), and layout-aware models (Xu et al., 2020, 2021; Huang et al., 2022). While effective, these pipelines inherit OCR errors such as incorrect reading order, misaligned bounding boxes, missing captions, or failure to parse figures and charts. As a result, their performance degrades on visually rich pages, where crucial semantic cues are encoded in the raw layout, not in transcribed text. In contrast, our approach operates directly on the visual document image and learns a global structural representation that does not depend on OCR or textual content.

Visual Document Retrieval Visual document retrieval bypasses OCR by embedding an entire document page as an image and learning similarity directly in a multimodal space. Early vision-language document models such as LayoutLM (Xu et al., 2020; Huang et al., 2022) combine textual, layout, and visual cues for document understanding, but they are primarily designed for single-vector prediction settings rather than fine-grained retrieval.

Recent systems adopt a multi-vector late-interaction paradigm (Faysse et al., 2025; Xu et al., 2025; Team, 2025; Günther et al., 2025), in which a page is represented as a *set* of patch embeddings and relevance is computed by matching query tokens against the strongest-scoring page patches. This formulation is effective for local cues such as words, table cells, or icons, but it biases scoring toward isolated matches. In multimodal pages, however, relevance often depends on page-level structure, namely how textual and visual elements are arranged and related across the page. Although some of this information may be present implicitly in visual encoders (Maniparambil et al., 2024), it is not explicitly supervised and may be underemphasized by late-interaction scoring. Our work addresses this limitation by introducing descriptor-guided structural supervision for late-interaction retrieval, implemented through an explicit global token trained jointly with patch embeddings.

Unlike global-pooling or CLS-token variants (Radford et al., 2021; Huang et al., 2022; Dosovitskiy et al., 2021; Zhai et al., 2022), our contribution is not the use of a single additional vector by itself. Instead, we train a retrieval-aware global representation with holistic page descriptors that summarize layout and organization during training only. This supervision encourages the encoder to represent cross-region structure while preserving the fine-grained local matching behavior of multi-vector retrieval, and it removes the need for textual descriptors at inference.

6 Conclusions

We proposed `Global-Descr-Embed`, a visual retrieval model that is trained to capture retrieval-aware structural priors that alter the geometry of late-interaction retrieval. We augment patch-based encoders with a learned global token trained via descriptor-guided structural supervision. The contribution is not the global vector itself, but the descriptor-guided training objective that reshapes both the global token and local patch representations to encode structural priors. On ViDoRe-v2, our method achieves significant gains over strong baselines with fewer parameters. We observe the largest improvements on visually rich and multi-region documents. Ablations confirm that global context complements local features. Overall, our findings show that structural signals can be learned as part of the retrieval representation itself, enabling more robust and scalable document retrieval across diverse layouts. Future work may extend our method beyond single-page retrieval to multi-page reasoning.

Limitations

Our method assumes that holistic page descriptors faithfully capture the structural layout of each document page. These descriptors are generated offline by an external vision-language model. If the generating model is unreliable or systematically biased toward certain layout patterns, the training signal for the global token degrades. We do not study how sensitive retrieval performance is to descriptor quality or to the choice of generation model.

We evaluate exclusively on ViDoRe-v2, which comprises four datasets from academic and corporate reporting domains. Our training data consists of English document pages only. We make no claims about generalization to specialized domains such as legal or clinical text, or to retrieval benchmarks with different query distributions. Empirical gains may not transfer to document corpora with substantially different layout characteristics.

Performance gains are not uniform across layout types. Our analysis shows that the benefit of the global token scales with the spatial heterogeneity of document pages. On collections with homogeneous or simple layouts, improvements are smaller. Practitioners should expect reduced gains when the document corpus does not exhibit diverse multi-region structure.

Our method addresses single-page retrieval. Multi-page reasoning, document-level retrieval, and end-to-end retrieval-augmented generation are out of scope. Extending the approach to these settings would require architectural choices beyond those studied here.

Risks

This work presents a retrieval model intended for document search and retrieval-augmented generation. We do not foresee direct misuse specific to our method, but note the following considerations.

Efficient visual retrieval lowers the cost of searching large document corpora at scale. While the intended use is enterprise search and RAG pipelines, the same capability could be applied to sensitive or private document collections. We encourage practitioners to apply appropriate access controls when deploying retrieval systems over confidential data.

Training our model requires only LoRA fine-tuning on a single GPU, which is lightweight relative to large-scale generative model training. The environmental footprint is therefore modest.

References

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16776–16809, Vienna, Austria. Association for Computational Linguistics.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, page 929–947, New York, NY, USA. Association for Computing Machinery.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. 2025. Mmdocir: Benchmarking multi-modal retrieval for long documents. *arXiv preprint arXiv:2501.08828*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, et al. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. *arXiv preprint arXiv:2506.18902*.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Sedigheh Eslami, Scott Martens, Bo Wang, Nan Wang, and Han Xiao. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. *Preprint*, arXiv:2506.18902.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4083–4091.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Quentin Macé, António Loison, and Manuel Faysse. 2025. Vidore benchmark v2: Raising the bar for visual retrieval. *arXiv preprint arXiv:2505.17166*.

Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Mohamed EI Amine Seddik, Sanath Narayan, Karttikeya Mangalam, and Noel E. O'Connor. 2024. Do vision and language encoders represent the world similarly? In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14334–14343.

Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. 2025. A survey of multimodal retrieval-augmented generation. *arXiv preprint arXiv:2504.08748*.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, et al. 2025. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing. *Preprint*, arXiv:2509.22186.

Xiangyu Peng, Cab Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, and Chien-Sheng Wu. 2025. Unidocbench: A unified benchmark for document-centric multimodal rag. *arXiv preprint arXiv:2510.03663*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. pages 0–.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Nomic Team. 2025. Nomic embed multimodal: Interleaved text, image, and screenshots for visual document retrieval.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. 2024. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *Preprint*, arXiv:2409.01704.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Junda Wu, Yu Xia, Tong Yu, Xiang Chen, Sai Sree Harsha, Akash V Maharaj, Ruiyi Zhang, Victor Bursztyn, Sungchul Kim, Ryan A. Rossi, Julian McAuley, Yunyao Li, and Ritwik Sinha. 2025. Doc-react: Multi-page heterogeneous document question-answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 67–78, Vienna, Austria. Association for Computational Linguistics.

Mengyao Xu, Gabriel Moreira, Ronay Ak, Radek Osmulski, Yauhen Babakhin, Zhiding Yu, Benedikt Schifferer, and Even Oldridge. 2025. Llama nemoretriever colembed: Top-performing text-image retrieval model. *arXiv preprint arXiv:2507.05513*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *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)*, pages 2579–2591, Online. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1192–1200.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. *Preprint*, arXiv:2111.07991.

7 Appendix

7.1 Prompt for Textual Baseline

The instructions provided to GPT-5 for generating textual summaries from images of document pages is displayed in the following. The prompt guides the model to produce plain-text descriptions that capture the main ideas and relevant details, while avoiding verbatim transcription or formatting artifacts:

`You are given an image representing a single page from a PDF document. Your task is to produce a concise plain-text summary of the content in the image. Do not include any formatting tokens such as bold text, bullet points, or other markup; the output should be plain text only. Read the visible text and summarize its main ideas, key points, and relevant details instead of transcribing it verbatim. If any text is unclear or partially visible, note that it is unreadable rather than guessing. For non-text elements such as images, charts, or diagrams, briefly describe their content and placement.`

7.2 Holistic Page Descriptors

The instructions provided to GPT-5 for generating high-level textual descriptions of images are shown below. The prompt directs the model to focus on overall scene composition, spatial arrangement, and thematic content, avoiding fine-grained details unless they dominate the image:

`You are given an image and must produce a concise yet informative textual description that captures its overall scene, structure, and high-level visual concepts. Focus on the broad composition, spatial arrangement, and thematic content that can be perceived without examining small details, describing the type of scene or setting, the general layout, dominant visual patterns, and the main purpose or activity depicted if relevant. Avoid mentioning small objects, text fragments, or fine-grained elements unless they dominate the entire image, and use clear, natural language that reflects the image's global visual meaning.`

7.3 Implementation Details

The vision and text encoders are initialized from `ColQwen2.5` pretrained LoRA adapters and weights. The new `[global]` token embedding is initialized randomly and added to the model's vocabulary. Base encoder weights remain frozen; only LoRA parameters and the retrieval projection head are updated. We fine-tune `Global-Descr-Embed` for 6 epochs with a batch size of 128 using AdamW with a learning rate of 5×10^{-5}, weight decay 10^{-4}, and a linear warm-up over the first 100 steps. `CrossContext-FT` is tuned with early stopping on the development set, converging after approximately 0.25 epochs, compared to 6 epochs for `Global-Descr-Embed`. This reflects task difficulty rather than overfitting: `CrossContext-FT` has direct access to text descriptors at inference time, while `Global-Descr-Embed` must internalize the same structural priors into the visual global token without textual input, a harder objective that requires more training to converge. We train in `bfloat16` precision and enable gradient checkpointing (Chen et al., 2016) to maximize memory efficiency on the GPU. We conduct all training and evaluation on a single NVIDIA H200 GPU (141 GB memory) using PyTorch (Ansel et al., 2024) and the HuggingFace Transformers library (Wolf et al., 2020). Our `Global-Descr-Embed` implementation extends the open-source `ColPali` framework (Faysse et al., 2025), adapting its multimodal retrieval pipeline to incorporate holistic page descriptors alongside local patch embeddings. All baselines (`llama-nemoretriever-colembed`, `colnomic-embed-multimodal`) are evaluated using their official HuggingFace implementations. We set the temperature $\tau = 0.02$. We generate the holistic page descriptors using GPT-5 (Singh et al., 2025).

7.4 Performance by Query Type: Global layout vs. Local Detail Oriented

We label each query in ViDoRe-v2 as either *detail-oriented* or *layout-oriented* using `GPT-5`.

Figure 5 shows the performance of `Global-Descr-Embed` on layout oriented queries in comparison with `Nemo-3b`. The largest improvements belongs to the layout-rich ESGH and ESGR datasets, showing the soundness of the trained layout encoder. These two datasets contain multiple visual regions per page (tables,

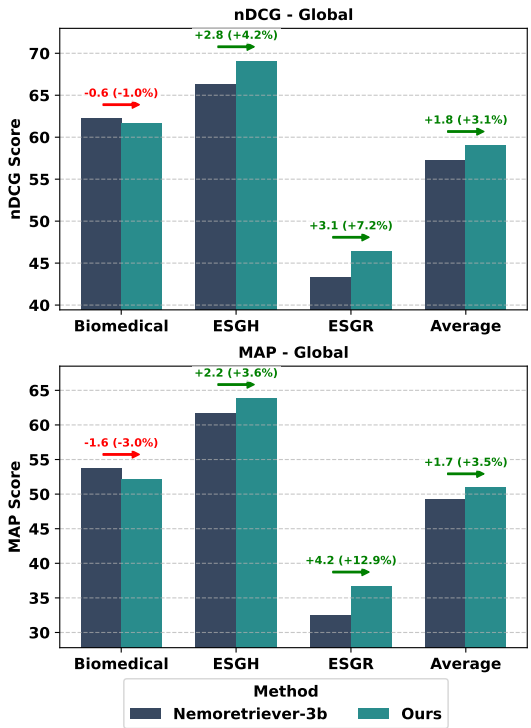

Figure 5: Performance on holistic-context queries: Comparison of our `Global-Descr-Embed` model with `Nemo-3b` across Biomedical, ESGH, and ESGR datasets, and averaged results. Bars show `nDCG@5` (top) and `MAP@5` (bottom). Green arrows show absolute and relative increases of performance, red arrows show decreases.

Dataset	Local	Global
Economics	2	230
Biomedical	195	445
ESG Human-Labeled	41	11
ESG Restaurant	128	100

Table 6: The number of detail-oriented (*local*) and layout-oriented (*global*) queries per ViDoRe-v2 dataset.

charts, and multi-column text), and relevance frequently depends on how these elements are spatially organized. In contrast, the Biomedical dataset is largely text dominant and contains many detail-oriented queries. Therefore, global layout information is less predictive for this dataset. In such cases, `Nemo-3b`, which is optimized for localized visual cues, performs comparatively well.

7.5 Spatial Shannon Entropy

We quantify the spatial dispersion of layout elements by computing the Shannon entropy of their center-point distribution over a $G \times G$ grid. Let $\mathcal{B} = \{b_1, \ldots, b_N\}$ be the set of detected blocks, and $\mathbf{c}_k = \left(\frac{x_{\min}^{(k)}+x_{\max}^{(k)}}{2}, \frac{y_{\min}^{(k)}+y_{\max}^{(k)}}{2}\right)$ their normalized center coordinates. Each $\mathbf{c}_k$ is assigned to a grid cell (i, j), with counts n_{ij} and cell probabilities $p_{ij} = n_{ij}/N$. The spatial Shannon entropy is then $H_{\text{spatial}} = -\sum_{i=1}^{G}\sum_{j=1}^{G} p_{ij} \log p_{ij}$, where $p_{ij} = 0$ terms are omitted. Higher H_{spatial} values indicate that elements are more uniformly distributed across the page, whereas lower values reflect spatial concentration.