NanoVDR: Distilling a 2B Vision-Language Retriever into a 70M Text-Only Encoder for Visual Document Retrieval

Anonymous ACL submission

Abstract

Vision-Language Model (VLM) based retrievers have advanced visual document retrieval (VDR) to impressive quality. They require the same multi-billion parameter encoder for both document indexing and query encoding, incurring high latency and GPU dependence even for plain-text queries. We observe that this design is unnecessarily symmetric: documents are visually complex and demand strong visual understanding, whereas queries are just short text strings. NanoVDR exploits this query–document asymmetry by decoupling the two encoding paths: a frozen 2B VLM teacher indexes documents offline, while a distilled text-only student as small as 69M parameters encodes queries at inference. The key design choice is the distillation objective. Through systematic comparison of six objectives across three backbones and 22 ViDoRe benchmark datasets, we find that pointwise cosine alignment on query text consistently outperforms ranking-based and contrastive alternatives, while requiring only pre-cached teacher query embeddings and no document processing during training. Furthermore, we identify cross-lingual transfer as the primary performance bottleneck, and resolve it cheaply by augmenting training data with machine-translated queries. The resulting NanoVDR-S-Multi (DistilBERT, 69M) retains 95.1% of teacher quality and outperforms DSE-Qwen2 (2B) on v2 and v3 with 32× fewer parameters and 50× lower CPU query latency, at a total training cost under 13 GPU-hours.

1 Introduction

Visual document retrieval (VDR) has achieved remarkable effectiveness in retrieving information from visually rich documents—financial reports with charts, scientific papers with figures, industrial manuals with diagrams—by treating each page as an image rather than relying on OCR-based text extraction (Faysse et al., 2024; Ma et al., 2024). State-of-the-art systems use Vision-Language Models (VLMs) to encode both queries and document pages into a shared embedding space (Faysse et al., 2024; Ma et al., 2024; Xin Huang, 2025; Nomic AI, 2025).

However, these systems apply the same heavyweight VLM encoder for both document indexing and query encoding. This results in high computational overhead at query time, requiring multi-billion parameter models and GPU inference even for plain-text queries, and leads to large index storage costs for high-dimensional representations.

A key observation is that this design is unnecessarily symmetric: documents are visually complex and genuinely require strong visual understanding, whereas queries are short text strings that carry no visual content. Using a multi-billion parameter VLM to encode text-only queries wastes the model's visual processing capacity entirely.

To exploit this query–document asymmetry, we propose NanoVDR, which decouples the two encoding paths through knowledge distillation (Figure 1a). A frozen VLM teacher indexes documents offline, producing single-vector visual embeddings; a lightweight text-only student (as small as 69M parameters) encodes queries at inference by mapping them into the teacher's embedding space via a learned projector. The student requires no vision module and runs on CPU in ∼50 ms, enabling single-vector cosine similarity retrieval.

The central design choice is the distillation objective—how to train the student to faithfully represent queries in the teacher's visual space. Through systematic comparison of six objectives across three backbones and the full 22-dataset ViDoRe benchmark, we make the following contributions:

- **Asymmetric distillation framework.** We propose a framework that distills a frozen 2B VLM teacher into text-only student encoders

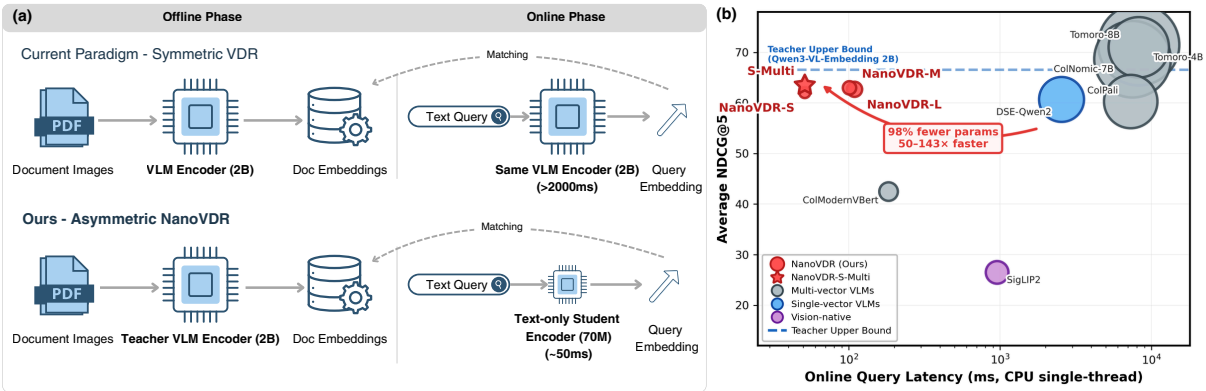

Figure 1: **Motivation and deployment advantage of NanoVDR.** (a) **Symmetric vs. asymmetric retrieval:** Current VDR systems (top) use the same heavy VLM encoder (2B) for both offline document indexing and online query encoding (>2,000 ms per query). NanoVDR (bottom) decouples the two: the frozen VLM teacher encodes documents offline, while a distilled text-only student (70M) encodes queries online in ∼50 ms on CPU. (b) **Performance vs. latency:** On the ViDoRe benchmark (mean NDCG@5 across v1/v2/v3), NanoVDR models achieve near-teacher accuracy (dashed line) at 50–143× lower CPU latency. Bubble size is proportional to model parameter count; the star (★) marks NanoVDR-S-Multi, the multilingual-augmented variant (§3.3).

(69–151M) for VDR. We show that pointwise cosine alignment which directly matches student and teacher query embeddings, outperforms all ranking-based and contrastive objectives. It only pre-cache teacher query embeddings and eliminating corpus-related processing from training entirely.

- **Extreme efficiency.** NanoVDR-S (DistilBERT, 69M) outperforms DSE-Qwen2 (2B) on v2 and v3 with 32× fewer parameters and 50× lower query latency (Figure 1b), at a total training cost under 13 GPU-hours. It also keeps the storage efficiency compared to multi-vector methods, which is inherited from the teacher model.

- **Cross-lingual augmentation.** We further identify in asymmetric encoders for VDR task, the cross-lingual transfer is the primary performance bottleneck, rather than cross-modal transfer. We resolve it via query-only multilingual augmentation, raising teacher retention from 92.4% (NanoVDR-S) to 95.1% (NanoVDR-S-Multi) at low cost.

2 Related Work

2.1 Visual Document Retrieval

Visual document retrieval treats document pages as images and uses VLMs for both query and document encoding. ColPali (Faysse et al., 2024) adapts PaliGemma into a ColBERT-style (Khattab and Zaharia, 2020) late interaction model, where each document page produces hundreds of token-level embeddings scored via MaxSim. The same work introduced the ViDoRe benchmark spanning diverse document types. DSE (Ma et al., 2024) takes a single-vector approach, producing one embedding per document screenshot using Qwen2-VL (Wang et al., 2024). VisRAG (Yu et al., 2024) demonstrates retrieval-augmented generation over visual documents. While retrieval quality has steadily improved, each generation of models has grown larger: more recent multi-vector systems based on 4–8B VLMs (Xin Huang, 2025; Nomic AI, 2025) achieve the highest quality but with query latency exceeding 7 seconds on CPU, further widening the efficiency gap. Vision-native encoders such as SigLIP2 (Tschannen et al., 2025) and Jina-CLIP (Koukounas et al., 2024) offer lighter alternatives but substantially lag behind VLM-based approaches on document retrieval tasks.

Three concurrent directions attempt to bridge efficiency and visual understanding. VISTA (Zhou et al., 2024) augments a frozen text encoder (BGE-Base, 110M) with a ViT image tokenizer (∼196M total), enabling multi-modal retrieval without modifying the text backbone; however, the ViT remains required at inference, and the model has not been evaluated on document-level benchmarks. ModernVBERT (Teiletche et al., 2025) builds a purpose-designed 250M vision-language encoder by fusing a SigLIP2 vision encoder with a ModernBERT

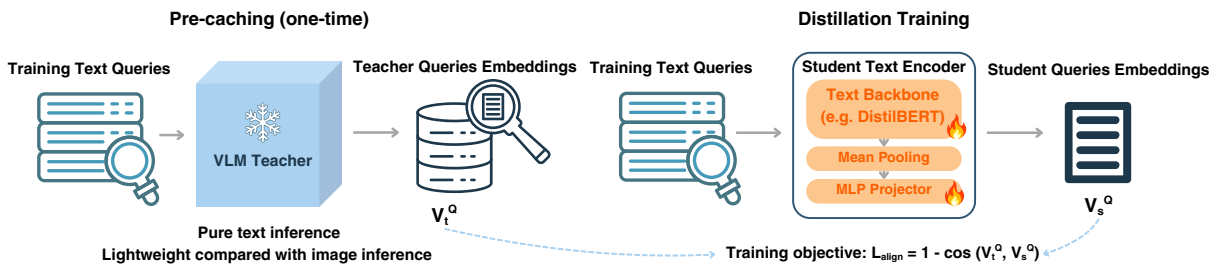

Figure 2: **Query-centric distillation training of NanoVDR. Left:** The frozen VLM teacher pre-caches training query embeddings via text-only inference. **Right:** The student text encoder is trained to minimize $\mathcal{L}_{\text{align}} = 1 - \cos(\mathbf{v}_t^Q, \mathbf{v}_s^Q)$, requiring no document images or negative sampling.

backbone via early fusion, matching ColPali-level quality at 12× fewer parameters; nevertheless, both query and document encoding still require the full vision-language model. SERVAL (Nguyen et al., 2025) takes a generate-then-encode approach: a VLM generates textual descriptions of document images, which are then indexed by a standard text encoder. While zero-shot and effective (63.4 NDCG@5 on ViDoRe v2 with a 72B VLM + 7B encoder), the pipeline requires massive VLM inference for every document at indexing time. Our approach differs fundamentally: we distill the VLM's embedding space directly into a tiny text-only encoder (69M), requiring neither a vision module at inference nor VLM-scale caption generation.

2.2 Knowledge Distillation in Dense Retrieval

Knowledge distillation (Hinton et al., 2015) has been extensively applied to dense text retrieval, where using separate encoders for queries and documents is well-established (DPR (Karpukhin et al., 2020), ColBERT (Khattab and Zaharia, 2020)). NanoVDR extends this asymmetry across modalities, pairing a VLM document encoder with a text-only query encoder. TAS-B (Hofstätter et al., 2021) uses topic-aware sampling with balanced training from a cross-encoder teacher. MarginMSE (Hofstätter et al., 2020) distills pairwise margin scores to train efficient bi-encoders. RankDistil (Reddi et al., 2021) applies listwise KL-divergence with curriculum learning. These approaches operate within a single modality (text-to-text) and rely on ranking-based objectives. In the vision-language domain, CLIP-KD (Yang et al., 2024) and TinyCLIP (Wu et al., 2023) compress CLIP models via combinations of feature alignment and affinity mimicking, but target image classification rather than document retrieval. Most closely related is Unveil (Sun et al., 2025), which distills an OCR-augmented VLM teacher (∼3B) into an image-only VLM student of the same size, combining representation alignment with soft-label KL-divergence. Our work takes a fundamentally different approach: we perform cross-modal distillation from a VLM teacher to a text-only student, and show that pure spatial alignment suffices—eliminating document representations during training entirely.

3 Methodology

3.1 System Overview

Figure 1a illustrates the overall architecture. Given a corpus of N document pages, each rendered as an image d_j, and a text query q, visual document retrieval aims to rank pages $\mathcal{D} = \{d_1, \ldots, d_N\}$ by relevance to q. NanoVDR decouples the two encoding paths entirely: a frozen VLM teacher g indexes each page image offline as $\mathbf{v}_j^D = g(d_j) \in \mathbb{R}^d$, while a lightweight text-only student f_θ encodes queries online as $\mathbf{v}_s^Q = f_\theta(q) \in \mathbb{R}^d$. Retrieval is performed via cosine similarity: $\text{score}(q, d_j) = \mathbf{v}_s^{Q\top} \mathbf{v}_j^D$.

Following Sentence-Bert (Reimers and Gurevych, 2019), the student text encoder consists of a pre-trained backbone h, mean pooling , and a two-layer MLP projector:

$$f_\theta(q) = \text{norm}\big(\text{MLP}(\text{pool}(h(q)))\big) \qquad (1)$$

where $\text{MLP}(\mathbf{x}) = W_2\,\sigma(W_1\mathbf{x} + b_1) + b_2$ with GELU activation. The teacher remains completely frozen throughout; specific model choices are detailed in §4.

3.2 Query-Centric Distillation

Figure 2 illustrates the training pipeline, which proceeds in two stages (left to right). First, the frozen VLM teacher encodes all training queries in text-only mode, producing target embeddings

$\mathbf{v}_t^Q = g(q) \in \mathbb{R}^d$. Second, the student text encoder is trained to produce query embeddings close to the teacher's. The alignment loss directly minimizes the angular distance:

$$\mathcal{L}_{\text{align}} = 1 - \frac{\mathbf{v}_s^Q \cdot \mathbf{v}_t^Q}{\|\mathbf{v}_s^Q\| \, \|\mathbf{v}_t^Q\|} \tag{2}$$

Because the teacher maps both queries and documents into the same embedding space, training the student to match teacher query embeddings simultaneously enables retrieval against teacher document embeddings, despite the student never seeing any images. This pointwise formulation requires no document embeddings, no negative sampling, and no corpus-level processing.

A key practical advantage of alignment-only distillation ($\mathcal{L}_{\text{align}}$) is that it requires only teacher query embeddings, which are text-encoded. Ranking-based objectives additionally require teacher document embeddings ($\mathcal{L}_{\text{rank}}$, Eq. 5) to construct in-batch similarity distributions, necessitating the teacher to process every training image—the dominant bottleneck in the pre-caching pipeline.

3.3 Multilingual Query Augmentation

Because alignment training is purely query-centric, extending the student to new languages requires only additional query text—not new document images or teacher re-encoding. We translate ∼489K English training queries into five target languages (Portuguese, Spanish, German, French, Italian) using Helsinki-NLP Opus-MT models (Tiedemann and Thottingal, 2020), balancing each language to ∼200K queries. Each translated query is re-encoded by the frozen teacher in text mode, producing a new target embedding. The augmented dataset combines these 778K translations with the original 711K pairs, yielding 1.49M training pairs (details in Appendix I).

4 Experimental Setup

We evaluate NanoVDR on the ViDoRe benchmark against 10 baselines spanning three model categories, followed by systematic ablation of the distillation objective (§6).

4.1 Datasets and Evaluation

We evaluate on the full public ViDoRe benchmark (Faysse et al., 2024; Macé et al., 2025; Loison et al., 2026), comprising 22 datasets across three versions (Appendix A): v1 (10 datasets: DocVQA, ArXivQA, InfoVQA, TabFQuAD, TatDQA, ShiftProject, and four SyntheticDocQA domains), v2 (4 datasets: ESG reports, biomedical lectures, economics reports, and human-labeled ESG reports), and v3 (8 datasets: finance with English and French corpora, HR, energy, industrial, pharmaceutical, physics, computer science). We report NDCG@5 (Järvelin and Kekäläinen, 2002) as the primary metric, averaged per benchmark version.

For training, we aggregate 726K query-document image pairs from four public sources after quality filtering and case-insensitive deduplication: VisRAG-Synthetic (Yu et al., 2024) (234K, 32.9%), ColPali training set (Faysse et al., 2024) (109K, 15.3%), VisRAG-InDomain (Yu et al., 2024) (94K, 13.2%), and VDR-Multilingual (Cimolai and Markewich, 2025) (en/es/it/de/fr) (275K, 38.6%). We hold out 2% via stratified sampling for validation (14.5K pairs), yielding 711K training pairs (Appendix B). The validation set is used for model selection (best checkpoint by validation loss).

4.2 Baselines

We select 10 baselines that represent the full spectrum of current VDR approaches, from large multi-vector VLMs to lightweight vision-native encoders, to contextualize NanoVDR's efficiency–quality tradeoff: (1) **Multi-vector VLMs** with MaxSim scoring: Tomoro-8B/4B (Xin Huang, 2025), ColNomic-7B (Nomic AI, 2025), ColPali (Faysse et al., 2024), ColModernVBert (Teiletche et al., 2025); (2) **Single-vector VLMs**: our teacher Qwen3-VL-Embedding-2B (Li et al., 2026) and DSE-Qwen2 (Ma et al., 2024); (3) **Vision-native encoders**: SigLIP2 (Tschannen et al., 2025), Jina-CLIP (Koukounas et al., 2024), BiModernVBert (Teiletche et al., 2025). Models in categories (1)–(2) and BiModernVBert are fine-tuned for document retrieval; SigLIP2 and Jina-CLIP are general-purpose contrastive models used zero-shot. All baselines are evaluated under identical conditions.

4.3 Implementation Details

The VLM teacher is Qwen3-VL-Embedding-2B (Li et al., 2026) (built on Qwen3-VL (Bai et al., 2025)), producing $d = 2048$-dimensional embeddings. We train three student variants of increasing capacity: NanoVDR-S (DistilBERT (Sanh et al., 2019), 66M+2M projector = 69M), NanoVDR-M (BERT-base (Devlin et al., 2019), 110M+2M

Table 1: Main results (NDCG@5 ×100) on the ViDoRe benchmark. Each benchmark version is averaged over the indicated number of datasets. Params = total model parameters; for NanoVDR this includes the backbone + MLP projector. Scoring = retrieval scoring method: MaxSim (token-level late interaction) or Cosine (single-vector dot product). †: Qwen3-VL-Embedding-2B (Li et al., 2026), our frozen teacher used for offline document indexing. Best per group in **bold**. Subscripts on NanoVDR rows indicate teacher retention (%, student/teacher).

Model	Backbone	Params	Scoring	ViDoRe v1 (10)	ViDoRe v2 (4)	ViDoRe v3 (8)
Multi-vector VLMs (MaxSim scoring, multi-token document representation)						
Tomoro-8B	ColQwen3-8B	8.0B	MaxSim	**90.6**	65.0	**59.0**
Tomoro-4B	ColQwen3-4B	4.0B	MaxSim	90.2	**65.2**	57.6
ColNomic-7B	ColQwen2.5-7B	7.0B	MaxSim	89.8	60.4	55.9
ColPali	PaliGemma-3B	3.0B	MaxSim	84.2	54.7	42.0
ColModernVBert	ModernBERT-base	250M	MaxSim	76.7	33.4	17.4
Single-vector VLMs (Cosine scoring, one vector per document)						
DSE-Qwen2	Qwen2-VL-2B	2.0B	Cosine	**85.1**	55.7	41.3
Qwen3-VL-Emb†	Qwen3-VL-2B	2.0B	Cosine	84.3	**65.3**	**50.0**
Vision-native encoders (Cosine scoring, visual-text contrastive pre-training)						
JinaCLIP	CLIP-ViT-L	400M	Cosine	**53.7**	**26.7**	**20.7**
SigLIP2	So400m	400M	Cosine	44.6	20.1	14.8
BiModernVBert	ModernBERT-base	250M	Cosine	37.4	10.9	5.5
NanoVDR (text-only student, Cosine scoring, distilled from Qwen3-VL-Emb†)						
NanoVDR-L	ModernBERT-base	151M	Cosine	**82.4**$_{(97.7\%)}$	61.5$_{(94.2\%)}$	44.2$_{(88.4\%)}$
NanoVDR-M	BERT-base	112M	Cosine	82.1$_{(97.4\%)}$	**62.2**$_{(95.3\%)}$	44.7$_{(89.4\%)}$
NanoVDR-S	DistilBERT	69M	Cosine	82.2$_{(97.5\%)}$	60.5$_{(92.6\%)}$	43.5$_{(87.0\%)}$
NanoVDR-S-Multi	DistilBERT	69M	Cosine	82.2$_{(97.5\%)}$	61.9$_{(94.8\%)}$	**46.5**$_{(93.0\%)}$

= 112M), and NanoVDR-L (ModernBERT-base (Warner et al., 2025), 149M+2M = 151M). Each uses a two-layer MLP projector (768 → 768 → 2048) to match the teacher's embedding space.

All experiments are conducted on NVIDIA H200 GPUs (141 GB HBM3e); all GPU-hour figures in this paper refer to this hardware. Training uses OneCycleLR scheduling (peak lr=2e-4, 3% warmup), batch size 256 with gradient accumulation 4 (effective 1024), for 20 epochs (∼13.9K steps). Training takes 10–12 hours per model on a single GPU (10.1h for NanoVDR-S, 10.5h for NanoVDR-M, 11.7h for NanoVDR-L). Since the alignment objective is purely pointwise and query-centric, the total training cost for our best NanoVDR model including pre-caching teacher query embeddings (∼1 GPU-hour via text-mode inference), is under 13 GPU-hours. Ranking-based ablation variants additionally require cached document embeddings; a full cost comparison is in Appendix E.

5 Main Results

5.1 Retrieval Performance against Heavyweight VLMs

Table 1 presents NDCG@5 across the three ViDoRe benchmark versions (per-dataset breakdowns in Appendix C).

NanoVDR-S approaches VLM-level quality. Our smallest variant, NanoVDR-S (DistilBERT, 69M), achieves 82.2/60.5/43.5 on v1/v2/v3, retaining 92.4% of teacher performance with 29× fewer parameters. With multilingual query augmentation (NanoVDR-S-Multi; §3.3), retention rises to 95.1% (82.2/61.9/46.5), making the 69M text-only student competitive with the 2B VLM teacher. The larger backbones NanoVDR-M (112M) and NanoVDR-L (151M) offer only marginal improvements over NanoVDR-S, confirming that the query embedding task does not require extensive model capacity.

Text-only students outperform VLM baselines on harder benchmarks. On the more challenging v2 and v3, all NanoVDR variants surpass both ColPali (3B, multi-vector) and DSE-Qwen2 (2B, single-vector), despite being text-only models with 32× fewer parameters than DSE-Qwen2. NanoVDR-S-Multi achieves the highest NDCG@5 among all student variants on v3 (46.5, +4.5 over ColPali); NanoVDR-M leads on v2 (62.2, +6.5 over DSE-Qwen2).

5.2 Extreme Efficiency and Deployment Cost

Table 2 summarizes deployment costs. The efficiency gains of NanoVDR come from two sources: the distilled student itself (query encoding latency and model size), and the inherited single-vector

Table 2: Deployment cost comparison (query-time). All latency measured on a single CPU thread (batch size 1). Params = total parameters. Dim = embedding dimension per vector. Tok/Doc = tokens stored per document page (1 = single-vector). Size = query encoder checkpoint (float32); for NanoVDR this is the student only. Encode = single-query encoding latency. Ret/N = scoring latency against N candidate documents. Index/1M = index storage for 1M documents (float32 for single-vector, float16 for multi-vector). NanoVDR shares the teacher's (Qwen3-VL-Embedding-2B) pre-computed document index. The teacher itself is excluded as it is used only for offline GPU-based document indexing.

Model	Params	Dim	Tok/Doc	Size	Scoring	Encode	Ret/1K	Ret/10K	Index/1M
Multi-vector VLMs (MaxSim scoring)									
Tomoro-8B	8,768M	320	1,280	35.1 GB	MaxSim	8,225 ms	706 ms	7,124 ms	819 GB
ColNomic-7B	7,829M	128	1,000	31.3 GB	MaxSim	7,471 ms	234 ms	2,385 ms	256 GB
ColPali	2,964M	128	1,030	11.9 GB	MaxSim	7,284 ms	269 ms	2,786 ms	264 GB
ColModernVBert	259M	128	1,000	1.0 GB	MaxSim	183 ms	242 ms	2,553 ms	256 GB
Single-vector models (Cosine scoring)									
DSE-Qwen2	2,209M	1,536	1	8.8 GB	Cosine	2,539 ms	0.14 ms	2.3 ms	6.1 GB
SigLIP2	1,136M	1,152	1	4.5 GB	Cosine	952 ms	0.14 ms	2.0 ms	4.6 GB
NanoVDR (text-only student, Cosine scoring)									
NanoVDR-L	151M	2,048	1	605 MB	Cosine	109 ms	0.20 ms	2.6 ms	8.2 GB
NanoVDR-M	112M	2,048	1	447 MB	Cosine	101 ms	0.21 ms	2.5 ms	8.2 GB
NanoVDR-S	69M	2,048	1	274 MB	Cosine	51 ms	0.22 ms	2.5 ms	8.2 GB

representation from the teacher (retrieval scoring and index storage). All latency measurements are collected on a single CPU thread (AMD EPYC 9474F, batch size 1).

Query encoding latency and model size (distillation advantage). NanoVDR-S encodes a query in 51 ms on CPU—50× faster than DSE-Qwen2 (2.5 s) and 143× faster than ColPali (7.3 s); even NanoVDR-L completes in 109 ms. The NanoVDR-S checkpoint is 274 MB, compared to 11.9 GB for ColPali and 35.1 GB for Tomoro-8B, enabling deployment on edge devices without GPU memory.

Inherited efficiency from single-vector teacher. The remaining advantages—single-vector cosine scoring (2.5 ms per 10K documents vs. 7.1 s for MaxSim) and compact index storage (8.2 GB vs. 264–819 GB per 1M pages)—are inherited from the teacher's single-vector architecture; our contribution is making the query encoder small enough to exploit them in practice.

6 Ablation and Analysis

Our ablation spans 6 loss configurations × 3 student backbones × 3 benchmarks = 54 evaluation points, all trained to convergence under the identical settings described in §4 (711K pairs, 20 epochs). The only controlled variable is the loss function. Note that ranking-based and InfoNCE variants require pre-cached teacher document embeddings for computing in-batch similarity distributions ($B = 256$ in-batch negatives); the pure alignment objective does not.

6.1 The Monotonic Superiority of Spatial Alignment

We compare six distillation objectives spanning the full spectrum from pure ranking ($\lambda_a = 0, \lambda_r = 1$) to pure alignment ($\lambda_a = 1, \lambda_r = 0$), plus a hard-label InfoNCE baseline (Table 3). The combined loss is $\mathcal{L} = \lambda_a \mathcal{L}_{\text{align}} + \lambda_r \mathcal{L}_{\text{rank}}$, where $\mathcal{L}_{\text{rank}}$ is KL-divergence over in-batch similarity distributions:

$$\mathbf{p}_t = \text{softmax}\big(\mathbf{v}_t^Q \mathbf{V}^{D\top} / \tau_t\big), \tag{3}$$

$$\mathbf{p}_s = \text{softmax}\big(\mathbf{v}_s^Q \mathbf{V}^{D\top} / \tau_s\big), \tag{4}$$

$$\mathcal{L}_{\text{rank}} = D_{\text{KL}}(\mathbf{p}_t \| \mathbf{p}_s) \tag{5}$$

where $\mathbf{V}^D \in \mathbb{R}^{B \times d}$ is the matrix of in-batch teacher document embeddings, and $\tau_t = 0.07$, $\tau_s = 0.05$ are temperature parameters (the softer teacher distribution encourages the student to preserve relative ranking structure). The InfoNCE baseline replaces soft teacher distributions with hard one-hot labels, where $\mathbf{v}_+^D$ is the positive document embedding:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\mathbf{v}_s^Q \cdot \mathbf{v}_+^D / \tau_s)}{\sum_j \exp(\mathbf{v}_s^Q \cdot \mathbf{v}_j^D / \tau_s)} \tag{6}$$

The result is consistent: as the alignment weight increases relative to ranking, NDCG@5 improves monotonically in the 3-backbone average on all three benchmarks, with consistent trends per backbone (Appendix D). Pure alignment outperforms pure ranking by +1.1/+4.0/+2.5 on v1/v2/v3.

Table 3: Loss ablation (NDCG@5 $\times 100$, averaged over 3 backbones). λ_a, λ_r = weights for $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{rank}}$. All trained under identical settings (20 epochs, lr=2e-4).

λ_a	λ_r	v1 (10)	v2 (4)	v3 (8)
1	0	**82.2**	**61.4**	**44.1**
1	0.5	81.6	59.8	42.8
1	1	81.5	59.1	42.5
0.5	1	81.5	58.6	42.1
0	1	81.1	57.4	41.6
InfoNCE		71.5	39.8	30.0

Table 4: Distillation retention by query language (NanoVDR-S). Train% = proportion of 711K training pairs in each language. Teacher and Student are mean NDCG@5 ($\times 100$). Ret. = Student / Teacher. All 22 ViDoRe datasets pooled (19,537 queries).

Language	Train%	#Queries	Teacher	Student	Ret.
English	68.7	6,237	64.0	60.3	94.3%
French	7.6	3,074	56.2	51.7	92.1%
Italian	7.6	2,419	49.0	44.1	90.0%
Spanish	8.1	2,694	51.4	46.1	89.7%
German	8.0	2,694	49.3	42.3	85.7%
Portuguese	0.0	2,419	48.7	36.8	75.6%

This is notable because ranking-based losses (KL-divergence, MarginMSE) are the prevailing choice in retrieval distillation (Hofstätter et al., 2021, 2020). We conjecture that alignment's advantage stems from the high quality of our teacher's embedding space: when the teacher provides well-structured coordinates, direct spatial alignment exploits richer geometric signal than relative ranking alone. Supporting this, we find that teacher quality is the strongest predictor of distillation success ($r = +0.607$), while student–teacher cosine similarity shows near-zero correlation with retention ($r = +0.094$; Appendix H)—suggesting that the geometric structure of the space matters more than pointwise proximity.

6.2 The Necessity of Soft-Label Distillation

The InfoNCE (van den Oord et al., 2018) baseline uses hard one-hot labels—the student learns that query i matches document i and nothing else—rather than the teacher's soft ranking distribution. The degradation is severe: -10.7 on v1, -21.6 on v2, and -14.1 on v3 compared to alignment (Table 3, bottom row).

This 10–22 point gap shows that faithfully reproducing the teacher's continuous embedding geometry is substantially more informative than fitting binary relevance boundaries. The teacher's "dark knowledge" (Hinton et al., 2015) (the geometric relationships encoded in its embedding space) is the critical ingredient for cross-modal transfer.

6.3 Data Efficiency

Training on random subsets of the 711K pairs reveals strong diminishing returns (Appendix F): at 25% of training data (178K pairs), NanoVDR-S already achieves 93%/82%/70% retention on v1/v2/v3. Even at 10% (71K pairs), the model reaches 79% retention on v1 (66.7 NDCG@5). The v3 benchmark saturates slower than v1, likely because it demands broader cross-lingual coverage (§6.4).

6.4 Cross-Lingual Transfer: The Primary Bottleneck

A natural concern with text-only distillation is whether the modality gap (text-only student vs. vision-language teacher) limits performance on visual documents. We conduct a per-query analysis on NanoVDR-S across all 22 ViDoRe datasets (19,537 queries), grouping every query by language to disentangle language effects from document content.

Language determines retention. Table 4 aggregates retention (student NDCG@5/ teacher NDCG@5 $\times$ 100) by the six evaluation languages. The hierarchy broadly tracks training data coverage: English (68.7% of training data) achieves 94.3% retention, French/Italian/Spanish (7–8% each) 90–92%, German (8.0%) 85.7%, and Portuguese (entirely absent from training) only 75.6%. The Pearson correlation between training data proportion and retention is $r = +0.563$ ($n = 6$; suggestive but not statistically significant at this sample size).

Within-dataset evidence isolates the language effect. On the eight ViDoRe v3 multilingual subsets—where the same document corpus is queried in all six languages—English queries average 92.8% retention versus 75.4% for Portuguese, a 17.4 pp gap on identical corpora. Per-dataset breakdowns (Appendix G) show the gap is largest for dense technical terminology and smallest for notation-heavy scientific content.

The modality gap is not the bottleneck. For English queries, NanoVDR-S retains 94.3% of teacher quality across 20 datasets spanning diverse visual document types—infographics, financial tables, scientific diagrams, and dense text—demonstrating

Table 5: Effect of multilingual query augmentation by language (NDCG@5 ×100). Original = NanoVDR-S trained on 711K pairs; +Aug = retrained on 1.49M pairs with translated queries. Δ = +Aug − Original. Ret. = +Aug / Teacher. Same 19,537 queries as Table 4.

Language	#Queries	Teacher	Original	+Aug	Δ	Ret.
English	6,237	64.0	60.3	60.3	+0.0	94.3%
French	3,074	56.2	51.7	53.2	+1.5	94.7%
Italian	2,419	49.0	44.1	45.7	+1.6	93.3%
Spanish	2,694	51.4	46.1	47.8	+1.8	93.1%
German	2,694	49.3	42.3	45.4	+3.1	92.0%
Portuguese	2,419	48.7	36.8	46.1	+9.3	94.6%

that a text-only student can effectively retrieve visual documents through spatial alignment alone. The dominant bottleneck is cross-lingual transfer: NanoVDR-S uses DistilBERT, a primarily English-trained backbone, which limits its ability to align non-English query embeddings with the teacher's multilingual space. Since alignment training is query-centric and requires no document images, this gap can be addressed cheaply via query-only data augmentation (§3.3).

Beyond language, we find that teacher quality ($r = +0.607$) and corpus size ($r = -0.566$) predict the remaining retention variation on English-only queries, while direct student–teacher cosine similarity shows near-zero correlation ($r = +0.094$); full analysis is in Appendix H.

6.5 Validating Multilingual Augmentation

Table 5 validates the multilingual augmentation described in §3.3 by breaking down NanoVDR-S-Multi's improvement by query language. The gains concentrate precisely where the cross-lingual analysis above predicted: Portuguese (previously absent from training) gains +9.3 NDCG, while English shows zero regression. After augmentation, all six languages achieve >92% retention—the maximum cross-lingual gap narrows from 18.6 pp to 2.7 pp. On the benchmark level (Table 1, NanoVDR-S-Multi), ViDoRe v1 performance is preserved (82.2 → 82.2); the multilingual v2 and v3 gain +1.4 and +3.0 respectively. These results confirm that the primary bottleneck for NanoVDR is cross-lingual transfer, not cross-modal transfer, and that it can be resolved cheaply through query-only data augmentation.

7 Conclusion

We presented NanoVDR, a framework for asymmetric cross-modal distillation that enables efficient visual document retrieval. Given a high-quality VLM teacher, the simplest possible objective—pointwise cosine alignment—consistently outperforms ranking-based and contrastive alternatives, while eliminating document representations and negative sampling from training. The query-centric nature of alignment makes the primary bottleneck cross-lingual rather than cross-modal, which we resolve cheaply via translated query augmentation, closing a 19 pp language gap without additional image processing. With this augmentation, NanoVDR-S-Multi (DistilBERT, 69M) retains 95.1% of teacher quality and outperforms DSE-Qwen2 (2B) on v2 and v3 with 32× fewer parameters, 50× lower query latency, and 32–100× less index storage—all at a training cost under 13 GPU-hours.

Limitations

NanoVDR inherits the teacher's document embedding quality as its performance ceiling, therefore the student cannot surpass the teacher. This work does not explore reducing the offline indexing cost, which still requires the full 2B VLM teacher to encode every document image; techniques such as teacher compression or progressive indexing remain directions for future work. We evaluate exclusively on visual document retrieval (first-stage ranking) with text-only queries; whether the framework generalizes to other retrieval settings is left to future investigation. Additionally, our multilingual query augmentation relies on lightweight machine translation models , which may introduce semantic shifts or struggle with domain-specific terminology (e.g., in finance or physics); future work could leverage more advanced LLMs for higher-fidelity translations.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Marco Cimolai and Logan Markewich. 2025. Visual document retrieval goes multilingual. Hugging Face Blog. https://huggingface.co/blog/vdr-2b-multilingual.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, pages 4171–4186.

Manuel Faysse, Hugues Fernandez, Nicolas Musiconi, Gautier Music, Hervé Déjean, Sébastien Keraron, and Stephane Moens. 2024. ColPali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. *Proceedings of SIGIR*, pages 113–122.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of SIGIR*, pages 39–48.

Andreas Koukounas, Georgios Mastrapas, Michael Günther, Bo Wang, Nan Sturua, Isabelle Mohr, Saba Krimmel, Jie Bai, Florian Facciolo, Scott Seidel, and 1 others. 2024. Jina CLIP: Your CLIP model is also your text retriever. *arXiv preprint arXiv:2405.20204*.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, and 1 others. 2026. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*.

Antonio Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel Moreira, Bo Liu, Manuel Faysse, Celine Hudelot, and Gautier Viaud. 2026. ViDoRe v3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios. *arXiv preprint arXiv:2601.08620*.

Xueguang Ma, Sheng Fang, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251*.

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

Thong Nguyen, Yibin Lei, Jia-Huei Ju, and Andrew Yates. 2025. SERVAL: Surprisingly effective zero-shot visual document retrieval powered by large vision and language models. In *Proceedings of EMNLP*, pages 30807–30822.

Nomic AI. 2025. Nomic embed multimodal: Open source multimodal embedding models for text, images, pdfs, and charts. Hugging Face: nomic-ai/colnomic-embed-multimodal-7b.

Sashank J. Reddi, Rama Kumar Pasumarthi, Aditya Krishna Menon, Ankit Singh Rawat, Felix X. Yu, Seungyeon Kim, Andreas Veit, and Sanjiv Kumar. 2021. RankDistil: Knowledge distillation for ranking. In *Proceedings of AISTATS*, pages 2368–2376.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Hao Sun, Yingyan Hou, Jiayan Guo, Bo Wang, Chunyu Yang, Jinsong Ni, and Yan Zhang. 2025. Unveil: Unified visual-textual integration and distillation for multi-modal document retrieval. In *Proceedings of ACL*, pages 23935–23945.

Paul Teiletche, Quentin Macé, Max Conti, Antonio Loison, Gautier Viaud, Pierre Colombo, and Manuel Faysse. 2025. Modernvbert: Towards smaller visual document retrievers. *arXiv preprint arXiv:2510.01149*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, Hongyang Chao, and Han Hu. 2023. TinyCLIP: CLIP distillation via affinity mimicking and weight inheritance. In *Proceedings of ICCV*, pages 2858–2868.

Kye Min Tan Xin Huang. 2025. Beyond text: Unlocking true multimodal, end-to-end rag with tomoro colqwen3. Hugging Face: TomoroAI/tomoro-colqwen3-embed-8b.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. 2024. CLIP-KD: An empirical study of CLIP model distillation. In *Proceedings of CVPR*, pages 26145–26154.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Shi, Xu Bai, Zhiyuan Liu, and Maosong Sun. 2024. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. VISTA: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of ACL*, pages 3185–3200.

A ViDoRe Benchmark Details

The Visual Document Retrieval (ViDoRe) benchmark comprises three progressively challenging versions. Table 6 provides a cross-version summary; Table 7 lists all 22 evaluation datasets.

v1 (Faysse et al., 2024). Introduced alongside ColPali, v1 contains 10 datasets in English and French. Five are sourced from established visual QA benchmarks (DocVQA, ArXivQA, InfoVQA, TatDQA, TabFQuAD) with human-authored queries; five use queries generated by Claude-3 Sonnet over curated document collections. With state-of-the-art models exceeding 90 NDCG@5, v1 is approaching saturation.

v2 (Macé et al., 2025). Designed to address v1's saturation, v2 introduces 4 datasets with queries in up to 4 languages (English, French, Spanish, German). Queries are generated through a blind contextual process—annotators receive only document metadata, not full pages—reducing extractive bias and requiring cross-document reasoning. The fourth dataset (ESG Reports Human-Labeled) provides expert human annotations over the same ESG corpus. Best models score 59–65 NDCG@5.

Table 6: Cross-version comparison of the ViDoRe benchmark.

	v1	v2	v3
Datasets	10	4	8
Query languages	2	4	6
Query source	VQA + LLM	LLM + human	LLM + expert
SOTA NDCG@5	>90	59–65	57–59

v3 (Loison et al., 2026). The most comprehensive version, v3 provides 8 public datasets across 6 languages (adding Italian and Portuguese). Annotation involved approximately 12,000 person-hours, combining LLM-synthesized queries with extensive expert review. Each query includes page-level relevancy rankings, bounding box annotations, and multilingual translations. The 8 domains span enterprise scenarios from finance to physics, with document corpora in English or French.

B Training Data Details

We aggregate training data from four publicly available visual document retrieval datasets, totalling 726K unique query–document image pairs after quality filtering and case-insensitive query deduplication (from 761K raw pairs). A 2% stratified split yields 711K training and 14.5K validation pairs. Table 8 provides the breakdown.

VisRAG-Synthetic (Yu et al., 2024). This dataset contains 234K synthetically generated query–page image pairs. Document pages are sourced from four domains: OpenStax college-level textbooks (10K pages), ICML 2023 papers (5K pages), NeurIPS 2023 papers (5K pages), and Manuallib product manuals (20K pages). Queries are generated by prompting GPT-4o with each page image. Training batches are organized so that all samples within a batch of 128 originate from the same source domain.

ColPali Training Set (Faysse et al., 2024). The ColPali training set comprises 127K query–page pairs, of which 63% come from existing academic VQA benchmarks (DocVQA, ArXiv QA, TatDQA, Infographic-VQA) and 37% are synthetically gen-

Table 7: All 22 ViDoRe evaluation datasets used in this work. Doc = document corpus language. Query = languages in which queries are provided. Source: H = human-authored (from VQA benchmarks or expert annotation); L = LLM-generated; L+H = LLM-generated with human review. Six languages = EN/FR/ES/DE/IT/PT.

Dataset	Ver.	Document Domain	Doc	Query	Source
DocVQA	v1	Industrial documents	EN	EN	H
ArXivQA	v1	Scientific papers	EN	EN	H
InfoVQA	v1	Infographics	EN	EN	H
TatDQA	v1	Financial tables	EN	EN	H
TabFQuAD	v1	Tables in French PDFs	FR	FR	H
SyntheticDocQA-AI	v1	AI-related documents	EN	EN	L
SyntheticDocQA-Energy	v1	Energy sector reports	EN	EN	L
SyntheticDocQA-Gov.	v1	Government reports	EN	EN	L
SyntheticDocQA-Hlt.	v1	Healthcare documents	EN	EN	L
ShiftProject	v1	Environmental reports	FR	FR	L
ESG Reports	v2	ESG / sustainability	EN	EN/FR/ES/DE	L+H
Biomedical Lectures	v2	Biomedical slides	EN	EN/FR/ES/DE	L+H
Economics Reports	v2	Economics reports	EN	EN/FR/ES/DE	L+H
ESG Reports (Human)	v2	ESG / sustainability	EN	EN	H
Finance-EN	v3	US annual reports	EN	6 langs	L+H
Finance-FR	v3	French annual reports	FR	6 langs	L+H
Computer Science	v3	CS textbooks	EN	6 langs	L+H
HR	v3	EU HR reports	EN	6 langs	L+H
Energy	v3	French energy reports	FR	6 langs	L+H
Industrial	v3	USAF technical orders	EN	6 langs	L+H
Pharmaceutical	v3	FDA reports	EN	6 langs	L+H
Physics	v3	French physics lectures	FR	6 langs	L+H

Table 8: Training data composition after deduplication. VDR-Multilingual is split by language.

Source	Subset	Train	%
VisRAG-Synthetic	—	233,817	32.9
ColPali	—	109,044	15.3
VisRAG-InDomain	—	94,016	13.2
VDR-Multilingual	Spanish	57,491	8.1
	German	56,994	8.0
	French	54,079	7.6
	Italian	53,787	7.6
	English	52,375	7.4
Total		**711,603**	**100.0**

erated from web-crawled PDF pages across energy, healthcare, government, AI, and business domains using Claude-3 Sonnet. The dataset is English-only by design, enabling study of zero-shot cross-lingual transfer. After our deduplication pipeline, 109K pairs remain.

VisRAG-InDomain (Yu et al., 2024). This dataset provides 123K query–page pairs drawn from six established visual QA benchmarks: PlotQA (56K, scientific plots), ArXivQA (26K, academic papers), Infographic-VQA (18K), MP-DocVQA (11K), SlideVQA (8K, presentation slides), and ChartQA (4K). Since these benchmarks overlap with ViDoRe evaluation sets, we rely on the official train splits to avoid contamination. After deduplication, 94K pairs remain.

VDR-Multilingual (Cimolai and Markewich, 2025). This multilingual dataset contains approximately 500K query–image pairs across five European languages (English, German, French, Spanish, Italian), collected from public internet PDFs. Queries are synthetically generated using Gemini-1.5-Pro and Qwen2-VL-72B. Each record also includes 16 hard negative document IDs (unused in our align-only training). After deduplication, 275K pairs remain with a balanced distribution across languages (52–57K per language). This is the only multilingual source in our training mixture, providing cross-lingual diversity. The dataset is released under the Apache 2.0 license.

C Per-Dataset Results

Tables 9–11 provide the full per-dataset breakdown for all three ViDoRe benchmark versions, covering all baselines and NanoVDR variants (including NanoVDR-S-Multi with multilingual augmentation) evaluated under identical conditions.

D Per-Backbone Loss Ablation

Table 12 shows loss ablation broken down by student backbone.

Table 9: Per-dataset NDCG@5 (×100) on ViDoRe v1. †: Qwen3-VL-Embedding-2B (our teacher). Best overall in **bold**, best among NanoVDR variants underlined.

	English								French		
Model	ArXivQA	TatDQA	InfoVQA	DocVQA	Syn-AI	Syn-En.	Syn-Gov.	Syn-Hlt.	Shift.	TabFQ.	Avg
Tomoro-8B	**91.3**	**81.5**	**94.6**	**66.2**	98.9	96.2	**96.9**	98.7	87.6	94.3	**90.6**
Tomoro-4B	90.9	80.8	94.2	66.0	98.5	95.9	95.7	98.9	87.4	94.0	90.2
ColNomic-7B	87.9	80.5	91.9	60.9	**100.0**	**96.8**	95.8	**99.1**	**88.7**	**96.0**	89.8
ColPali	82.7	70.3	85.4	58.9	96.7	94.2	96.0	96.9	75.2	85.9	84.2
ColModernVBert	72.7	71.3	83.5	47.2	96.7	89.0	93.0	94.7	50.1	68.9	76.7
DSE-Qwen2	85.6	67.5	87.2	56.3	96.6	92.3	96.0	96.5	80.3	93.0	85.1
Teacher†	81.8	63.0	88.4	49.3	97.4	90.9	95.3	97.0	83.4	**96.3**	84.3
JinaCLIP	68.2	31.8	60.9	27.6	67.7	64.6	67.5	68.6	33.6	46.9	53.7
SigLIP2	35.0	18.8	56.2	22.0	55.5	57.3	42.0	62.6	29.1	67.5	44.6
BiModernVBert	40.4	27.5	43.0	15.2	45.2	45.2	41.2	49.5	18.7	48.0	37.4
NanoVDR-L	79.1	59.6	87.4	46.2	96.1	89.3	93.9	96.4	79.9	95.8	82.4
NanoVDR-M	78.2	59.5	86.0	46.6	95.5	89.6	94.3	94.9	80.8	96.0	82.1
NanoVDR-S	77.6	58.7	86.6	45.8	95.3	90.3	95.0	95.4	81.4	96.0	82.2
NanoVDR-S-Multi	78.7	58.4	85.9	46.3	96.5	91.0	93.7	94.9	80.9	95.6	82.2

Table 10: Per-dataset NDCG@5 (×100) on ViDoRe v2. †: Qwen3-VL-Embedding-2B (our teacher). Best overall in **bold**, best among NanoVDR variants underlined.

Model	ESG	BioMed	Econ	ESG-H	Avg
Tomoro-8B	60.4	65.4	59.5	74.7	65.0
Tomoro-4B	62.6	65.6	56.0	**76.8**	**65.2**
ColNomic-7B	53.6	63.2	55.6	69.4	60.4
ColPali	54.9	55.9	49.4	58.7	54.7
ColModernVBert	30.5	26.6	27.2	49.3	33.4
DSE-Qwen2	54.9	55.9	52.0	60.0	55.7
Teacher†	**65.4**	**68.4**	**63.4**	63.7	**65.3**
JinaCLIP	16.1	31.3	35.4	24.1	26.7
SigLIP2	9.4	33.4	28.3	9.4	20.1
BiModernVBert	6.1	7.6	15.8	14.0	10.9
NanoVDR-L	63.5	62.0	59.3	61.2	61.5
NanoVDR-M	62.7	63.6	59.8	62.8	62.2
NanoVDR-S	61.1	61.0	57.8	62.1	60.5
NanoVDR-S-Multi	62.0	62.7	60.5	62.5	61.9

E Training Cost: Alignment vs. Ranking

Table 13 quantifies this difference. For our 711K-pair training set, encoding document images with the 2B-parameter VLM teacher costs 24 GPU-hours (8 shards on H200), while text query encoding completes in under 1 GPU-hour—a 30× speedup. Alignment-only training eliminates the image encoding step entirely, reducing the total pre-caching cost from ∼25 to ∼1 GPU-hour and halving the embedding storage from 5.8 GB to 2.9 GB.

Combined with the accuracy advantage demonstrated in Table 3, alignment-only distillation is dominant in our setting: it is both more accurate and substantially cheaper.

F Data Efficiency

We investigate how much training data is needed by training NanoVDR-S (DistilBERT, align-only) on random subsets of {1, 5, 10, 25, 50, 100}% of the 711K training pairs, all with identical hyperparameters (20 epochs, lr=2e-4). Figure 3 shows clear diminishing returns: at just 25% of training data (178K pairs), the model already achieves 93%/82%/70% retention on v1/v2/v3, covering most of the gap to the full-data model (98%/93%/87%). The marginal gain from quadrupling data (25%→100%) adds only +4.2/+7.0/+8.6 NDCG points, compared to +25.2/+29.4/+22.9 from the 5%→25% jump.

The three benchmarks saturate at markedly different rates: v1 (English-dominant, smaller cor-

Table 11: Per-dataset NDCG@5 (×100) on ViDoRe v3. †: Qwen3-VL-Embedding-2B (our teacher). Best overall in **bold**, best among NanoVDR variants underlined.

Model	Fin-EN	Fin-FR	CS	HR	Energy	Indust.	Pharma	Physics	Avg
Tomoro-8B	**62.8**	**45.9**	72.4	**61.5**	**65.8**	52.5	**64.3**	**46.7**	**59.0**
Tomoro-4B	61.3	42.9	72.5	57.3	63.3	**52.7**	64.0	46.6	57.6
ColNomic-7B	55.2	42.8	**74.1**	57.0	62.4	49.4	61.1	45.0	55.9
ColPali	34.9	21.8	63.6	43.4	45.7	34.6	52.7	39.5	42.0
ColModernVBert	18.3	11.6	25.9	12.9	22.9	11.4	19.4	17.0	17.4
DSE-Qwen2	40.1	22.4	60.8	40.8	45.0	30.4	51.8	39.0	41.3
Teacher†	53.5	32.5	69.3	48.9	54.8	38.2	60.2	42.3	50.0
JinaCLIP	13.4	9.4	37.9	21.7	19.2	10.6	29.5	23.8	20.7
SigLIP2	6.4	4.8	28.2	13.6	14.8	12.8	20.7	17.0	14.8
BiModernVBert	4.5	2.0	11.6	4.3	4.8	3.7	8.2	5.1	5.5
NanoVDR-L	46.5	29.2	59.1	43.2	51.7	28.7	55.1	40.1	44.2
NanoVDR-M	47.0	30.2	60.3	43.1	51.8	30.2	55.1	39.9	44.7
NanoVDR-S	45.6	30.2	57.8	42.1	50.9	27.9	54.0	39.3	43.5
NanoVDR-S-Multi	48.9	32.2	62.5	45.3	52.7	32.2	57.4	40.5	46.5

Table 12: Loss ablation (NDCG@5 ×100) per variant. S = NanoVDR-S, M = NanoVDR-M, L = NanoVDR-L.

			ViDoRe v1			ViDoRe v2			ViDoRe v3		
Config	λ_a	λ_r	S	M	L	S	M	L	S	M	L
Align	1	0	82.2	82.1	82.4	60.5	62.2	61.5	43.5	44.7	44.2
Align-dom	1	0.5	81.4	81.6	81.8	59.5	59.7	60.2	41.9	43.5	43.1
Combined	1	1	81.2	81.4	82.0	58.6	59.4	59.3	41.4	42.9	43.0
Rank-dom	0.5	1	81.1	81.6	81.8	57.3	58.6	59.9	40.9	42.6	42.8
Rank	0	1	80.3	81.1	81.9	56.5	57.0	58.9	40.0	41.8	42.9
InfoNCE	hard		71.4	71.5	71.7	39.2	38.9	41.4	29.8	30.3	30.0

Table 13: Pre-caching and training cost comparison for align-only ($\lambda_r = 0$) vs. ranking-based ($\lambda_r > 0$) distillation on 711K training pairs with a 2B VLM teacher.

Requirement	Align-only	With ranking
Teacher query encoding (text)	✓	✓
Teacher doc encoding (images)	—	✓
Query encoding time	0.8 GPU-h	0.8 GPU-h
Doc encoding time	—	24 GPU-h
Total pre-cache cost	**0.8 GPU-h**	**24.8 GPU-h**
Embedding storage	2.9 GB	5.8 GB
Temperature tuning (τ_t, τ_s)	not needed	required
Per-step complexity	$O(B)$	$O(B^2)$

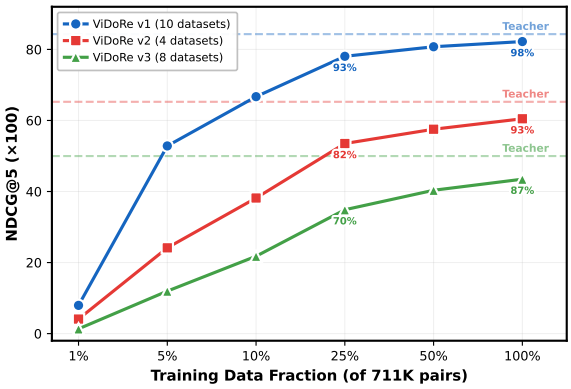

Figure 3: **Data efficiency of NanoVDR-S.** NDCG@5 vs. fraction of 711K training pairs, with teacher upper bounds (dashed). Percentages indicate retention (student/teacher). Diminishing returns are pronounced after 25% of training data.

pora) reaches 93% retention with 25% data, while v3 (6-language, larger corpora) requires 100% data to reach 87%. This mirrors the cross-lingual finding (§6.4): multilingual evaluation demands greater data diversity. Even at 10% (71K pairs), NanoVDR-S achieves 66.7 on v1 (79% retention), demonstrating that the alignment objective extracts useful cross-modal knowledge from remarkably few examples.

G Per-Language Retention Matrix

Table 14 presents the full per-language retention breakdown across all 22 ViDoRe datasets. The 10 ViDoRe v1 datasets are single-language (8 English, 2 French); ViDoRe v2 includes four evaluation languages (EN/FR/ES/DE); ViDoRe v3 includes

all six. Portuguese, entirely absent from training data, is consistently the lowest-retention language across all eight v3 datasets.

H Residual Factors Beyond Language

Controlling for language, we analyze English-only queries to identify what corpus-level properties predict the remaining variation in retention. We compute Pearson correlations over 19 datasets with distinct document corpora (Table 15), excluding ESG Reports (Human) which shares its corpus with ESG Reports.

Teacher quality is the strongest positive predictor. The teacher's own NDCG@5 correlates most strongly with student retention ($r = +0.607$): datasets where the teacher achieves higher absolute performance also exhibit higher student retention. A well-structured teacher embedding space—where relevant documents are clearly separated from irrelevant ones—facilitates distillation, while noisy or ambiguous teacher rankings propagate errors.

Larger corpora are harder to distill. The number of documents in the evaluation corpus shows a moderate negative correlation with retention ($r = -0.566$). Larger corpora present more confusable candidates, demanding finer-grained query discrimination to preserve top-k rankings.

Student–teacher alignment does not predict retention. The mean cosine similarity between student and teacher query embeddings—the direct optimization target of the alignment loss—shows near-zero correlation with dataset-level retention ($r = +0.094$). This reveals that retrieval success depends not on how closely the student replicates the teacher's query coordinate, but on whether the relevant document remains among the student's nearest neighbors. When documents are well-separated (high teacher quality, small corpus), even approximate alignment preserves the correct ranking; when documents cluster densely, even high-fidelity alignment may be insufficient.

I Multilingual Augmentation Pipeline

This section details the query-only data augmentation procedure summarized in §3.3.

I.1 Translation Procedure

We first extract all English queries from the four English-language training sources (VisRAG-Synthetic, VisRAG-InDomain, ColPali, VDR-English), yielding 489K source queries. For each target language, we compute the gap between its current count in the 711K training set and a target of 200K, then randomly sample that many English queries (seed = 42) for translation. Table 16 summarizes the per-language breakdown.

Translation uses Helsinki-NLP Opus-MT models (Tiedemann and Thottingal, 2020), a family of neural machine translation models trained on OPUS parallel corpora. For Portuguese, we use the multilingual `opus-mt-en-ROMANCE` model with a »pt« target language token; other languages use dedicated bilingual models (`opus-mt-en-{es,de,fr,it}`). All translations use batch size 64 with maximum sequence length 512.

I.2 Teacher Embedding Caching

Each translated query is encoded by the frozen Qwen3-VL-Embedding-2B teacher in text mode, using the same query instruction (*"Find a document image that matches the given query."*) applied during the original embedding caching. This produces a 2048-dimensional float16 embedding per translated query. No document image encoding is involved—the translated queries inherit the original query's positive document pairing, but the alignment-only objective does not use document embeddings during training. The total teacher encoding cost for 777,649 translated queries is under 15 minutes on a single H200 (915 queries/s in text-only mode).

I.3 Combined Training Set

The 777,649 translated queries are appended to the original 711,603 training pairs, yielding 1,489,252 combined training pairs. The resulting language distribution is approximately: English 385K (25.9%), Portuguese 200K (13.4%), Italian 200K (13.4%), French 200K (13.4%), German 200K (13.4%), Spanish 200K (13.4%), with the remainder from mixed-language sources.

I.4 Training Adjustments

To keep total training steps comparable despite the doubled dataset, we halve the epoch count (10 vs. 20) and slightly increase the learning rate (3e-4 vs. 2e-4). All other variables remain identical to the original NanoVDR-S configuration: DistilBERT backbone, align-only loss ($\lambda_a = 1, \lambda_r = 0$), batch

Table 14: Per-language retention (NanoVDR-S NDCG@5/ Teacher NDCG@5 $\times$ 100) across all 22 ViDoRe datasets. "—" = language not present in that dataset. V1 datasets are single-language (8 English, 2 French).

Dataset	Ver.	EN	FR	ES	DE	IT	PT
ArXivQA	v1	94.9	—	—	—	—	—
TatDQA	v1	93.1	—	—	—	—	—
InfoVQA	v1	97.9	—	—	—	—	—
DocVQA	v1	92.8	—	—	—	—	—
SyntheticDocQA-AI	v1	97.8	—	—	—	—	—
SyntheticDocQA-En.	v1	99.3	—	—	—	—	—
SyntheticDocQA-Gov.	v1	99.7	—	—	—	—	—
SyntheticDocQA-Hlt.	v1	98.3	—	—	—	—	—
ShiftProject	v1	—	97.6	—	—	—	—
TabFQuAD	v1	—	99.7	—	—	—	—
ESG Reports	v2	95.7	96.4	93.8	87.7	—	—
Biomedical Lectures	v2	89.8	90.9	91.6	84.2	—	—
Economics Reports	v2	95.5	91.5	95.3	82.0	—	—
ESG Reports (Human)	v2	97.4	—	—	—	—	—
Finance (EN corpus)	v3	89.7	89.9	85.5	84.5	89.0	71.3
Finance (FR corpus)	v3	94.0	94.2	97.9	90.6	98.1	81.8
CS	v3	92.5	84.5	83.8	82.4	84.4	72.6
HR	v3	92.4	90.4	87.6	88.5	89.0	67.9
Energy	v3	97.1	94.9	95.1	89.8	96.3	83.3
Industrial	v3	86.5	74.3	77.9	63.3	73.5	58.7
Pharma	v3	94.9	92.3	90.1	90.7	93.4	75.9
Physics	v3	95.3	93.5	95.4	90.1	92.0	91.4

Table 15: Pearson correlations between corpus-level features and English-only retention (NanoVDR-S, 19 datasets). Teacher = mean teacher NDCG@5; N = corpus size; $\bar{c}$ = mean student–teacher query cosine similarity.

	Teacher	N	$\bar{c}$
r with Ret.	+0.607	-0.566	+0.094

Table 16: Translation statistics per target language. Existing = count already in the 711K training set. Translated = newly machine-translated queries. Target = 200K per language.

Language	Existing	Translated	Total	Model
Portuguese	0	200,000	200,000	en-ROMANCE
Italian	53,787	146,213	200,000	en-it
French	54,079	145,921	200,000	en-fr
German	56,994	143,006	200,000	en-de
Spanish	57,491	142,509	200,000	en-es
Total		**777,649**		

size 256 with gradient accumulation 4, MLP projector ($768 \rightarrow 768 \rightarrow 2048$), OneCycleLR with 3% warmup, and seed 42.