ModernVBERT: Towards Smaller Visual Document Retrievers

Anonymous authors
Paper under double-blind review

Abstract

Retrieving specific information from a large corpus of documents is a widely prevalent industrial use case of modern AI, notably due to the popularity of Retrieval-Augmented Generation (RAG) systems. Although neural document retrieval models have historically operated exclusively in the text space, Visual Document Retrieval (VDR) models – large vision–language decoders repurposed as embedding models which directly work with page screenshots as inputs – are increasingly popular due to the performance and indexing latency gains they offer. In this work, we show that, while cost-efficient, this approach of repurposing generative models bottlenecks retrieval performance. Through controlled experiments, we revisit the entire training pipeline, and establish a principled recipe for improving visual document retrieval models. We notably measure the impact of attention masking, image resolution, modality alignment data regimes, and late interaction centered contrastive objectives which emerge as central performance factors. Building on these insights, we release *ModernVBERT*, a compact 250M-parameter vision–language encoder that outperforms recent models up to 10 times larger when fine-tuned on document retrieval tasks, enabling efficient inference on cheap CPU hardware and greatly reducing latency and costs while maintaining strong performance. Models, code and data are available in the public version of this work under an open license.

1 Introduction

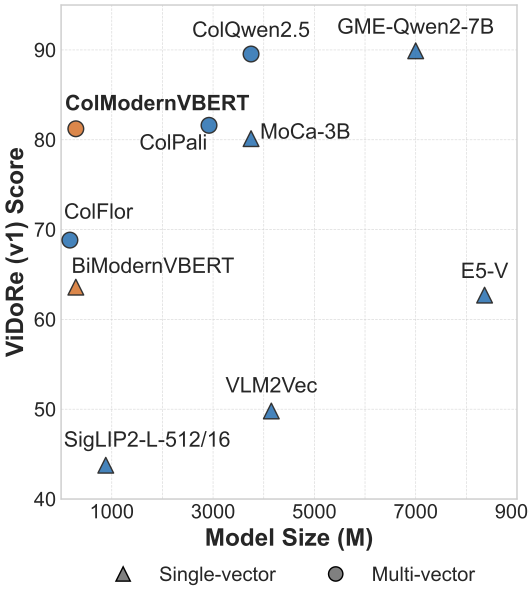


Figure 1: **Pareto efficiency.** *ColModernVBERT* outperforms models in its category on ViDoRe, achieving a leading performance-size tradeoff.

The ability to quickly locate specific information in vast document collections is a core building block of digital systems today, supporting use cases that range from web search and virtual assistants to enterprise knowledge management. Neural information retrieval (IR) models, and in particular dense retrievers, have become the de facto backbone of modern search systems thanks to their strong semantic matching capabilities and good scalability properties (Reimers & Gurevych, 2019; Karpukhin et al., 2020; Wang et al., 2022).

This trend is amplified by the widespread adoption of Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), where a retriever is used to select a small set of relevant documents that condition a downstream generator. In such systems, the first-stage retrieval module is a well-known bottleneck: its recall directly upper-bounds the quality of the generated answers, while its latency and indexing costs partially drive the overall system efficiency (Lin & Byrne, 2022). As a result, improving document retrieval, especially for long, complex files such as PDFs, scientific articles, and reports, is a key lever for making industrial RAG deployments more accurate and cost-effective.

Visual Document Retrieval. Historically, document retrieval in these settings has operated purely in the text space. To index PDFs or scans, practitioners first run heavy preprocessing pipelines that include Optical Character Recognition (OCR), layout analysis, and heuristic passage segmentation, before embedding the resulting text spans with a neural encoder. This approach suffers from several limitations: OCR and layout parsing can be brittle and slow, complex visual elements such as tables, figures, and typography are often poorly captured, and any error or bias introduced during preprocessing is propagated to the retriever.

Visual Document Retrieval (VDR) has emerged as a compelling alternative to such text-based systems. Rather than indexing pre-extracted textual content, VDR models directly operate on page screenshots: given a user query, they retrieve relevant document pages by matching the query against image-based representations of the pages (Faysse et al., 2025). By bypassing OCR and layout parsing, VDR yields simpler end-to-end pipelines, significantly reduces indexing latency, and better exploits visual cues such as layout, figures, and fonts, while achieving strong performance on visually rich benchmarks like ViDoRe.

Limits of Generative VLM Repurposing. Most current VDR systems are obtained by repurposing large generative vision–language decoders (Alayrac et al., 2022) as retrieval encoders via post-hoc contrastive fine-tuning (Ma et al., 2024; Faysse et al., 2025; Jiang et al., 2025). While cost-efficient, this design choice bottlenecks retrieval performance and efficiency: model sizes, attention patterns, image resolutions, and training objectives are designed for generative use cases rather than optimized for retrieval which has been shown in text models to be suboptimal (Lee et al., 2025; Gisserot-Boukhlef et al., 2025). Furthermore, scaling trends (Wei et al., 2022) are less pronounced for embedding models; while correlated with model size, strong retrieval performance remains attainable with small models (Clavié, 2024).

Recent papers and model releases in the visual retrieval space have claimed performance improvements by scaling the amount of contrastive data and the compute budget (Zhang et al., 2025a; Xu et al., 2025), modifying the attention mask (Chen et al., 2025), increasing image resolutions (Cohere, 2024) or by introducing more diverse tasks and data sources (Jiang et al., 2025).

In this work, we attempt to centralize these efforts and systematically disentangle the impact of core design decisions in visual retriever training. Through controlled experiments—ranging from language model pretraining to multi-stage, domain-specific fine-tuning, we aim to answer a central question:

Which design choices best boost performance in modern visual document retrievers?

Contribution 1. We revisit core assumptions in visual retriever design, showing that token-level training objectives benefit retrievers by strengthening image–text token alignment—rather than merely producing stronger image embeddings. Our results indicate that causal attention is suboptimal in document retrieval, with bidirectional masking offering clear improvements in multi-vector settings, and that other parameters such as image resolution data mixes should not be overlooked in the training pipeline.

Contribution 2: *ModernVBERT*. Building on these insights, we release *ModernVBERT*, a small 250M multimodal encoder that aligns a pretrained language encoder with a vision encoder through Masked Language Modeling (MLM) objective, and *ColModernVBERT* a variant fine-tuned for document retrieval. Despite its modest size and limited training budget, *ColModernVBERT* matches models 10x larger on standard visual document retrieval benchmarks, demonstrating the interest of designing a retrieval focused model from the ground up. We release the model, intermediate checkpoints, and the training code in the public version of the paper.

2 METHODOLOGY

Our analysis aims at quantifying the impact of design decisions made when training visual retrievers. In opposition to previous work, we begin our analysis as early as language model modality

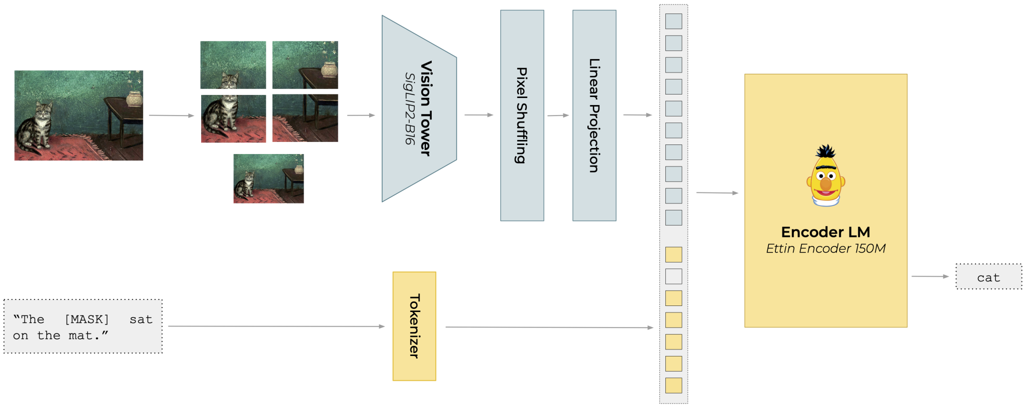


Figure 2: **MLM-based early fusion architecture.** The visual encoder produces patch representations, which are passed to a language model. Our end-to-end bidirectional attention fused architecture is trained with Masked Language Modeling objectives and is perfectly suited for sequence and token-level representation tasks.

alignment and iteratively study design choices by modifying design choices independently to reduce confounding factors as much as possible (Allen-Zhu & Li, 2025).

Controlled Experimental Setup. A central point of interest is the impact of causal and bidirectional attention masks. While recently studied for textual representation applications (Gisserot-Boukhlef et al., 2025; Weller et al., 2025), we extend the experiment to the vision modality. We use checkpoints released by Gisserot-Boukhlef et al. (2025) which consist in a series of identical 210M parameter transformer models based on the Llama architecture (Touvron et al., 2023) trained on 100B tokens that differ only in their attention masking strategy during language model training but that are perfectly identical in terms of training data seen, model size and architecture, learning rate scheduling, etc... The checkpoints we use are `enc` a bidirectional encoder trained with Masked Language Modeling (MLM), `dec`, a causal decoder trained with next token prediction, and `dec-enc` a causal decoder annealed over the end of its textual training by removing the causal mask and switching the training objective to MLM. For the vision tower, we employ the vision component of `siglip2-base-16b-512` (Tschannen et al., 2025), a 86M parameter vision transformer contrastively trained on billions of text-image pairs. All ablations thus stem from iso-data controlled setups, and as further described, are further trained on the same data sequence, with the same batch sizes, optimizers, schedulers and on the same hardware.

Model Architecture. Our analysis are not centered around model architectures and to draw broadly applicable insights, we design vision-language models following current standard training practices. In line with most recent work, we employ the early fusion architecture (Alayrac et al., 2022) illustrated in Figure 2, in which visual patch embeddings produced by the vision encoder are projected into the language model input embedding space and concatenated with text token embeddings to encourage joint processing (Li et al., 2022; Alayrac et al., 2022; Wang et al., 2024; Yang et al., 2025; Marafioti et al., 2025). As described in subsection 2.1, we generalize the training loss to function both with causal and masked language modeling objectives. To handle dynamic resolutions, we split large images into 512×512 pixel tiles as expected by the SigLIP encoder[1]. Following current standard practices, we further process a downscaled version of the full image to improve inter-tile consistency and global visual understanding (Lin et al., 2023; Ye et al., 2023). The vision tower produces 1024 pixel patch representations for each tile[2], which we compress to 64 tokens through *pixel shuffling* (Shi et al., 2016) with a compression ratio $r = 4$, following prior work on models

[1] Images are downscaled (or upscaled) so that the lengths and widths reach a multiple of 512 pixels to preserve the aspect ratio, padding is used on the smaller side when necessary (i.e. a 1024x1000 px image would be padded to 1024x1024 px).

[2] The SigLIP tower takes 512x512 px images and process them by 16x16 px patches (Dosovitskiy et al., 2020). This results in $(512/16)^2 = 1024$ patches.

of comparable size (Marafioti et al., 2025). We highlight the impact of image resolution and this parameter on the number of visual tokens in Appendix C.6.1.

Training Procedure. Our experiments focus on retrieval performance. We employ a standard biphasic training procedure, in which we first run modality alignment to train a pretrained textual language model to understand visual inputs through language modeling objectives (Liu et al., 2023b) (subsection 2.1), then rely on a second text-image contrastive learning phase to learn efficient image representations (Radford et al., 2021) (subsection 2.2). We further describe the general setup, and detail specific modifications to the default training procedure in the experiment section.

2.1 Modality Alignment

We align the vision encoder tower with the language model by training the image embedding projection layer to map visual features into the language model embedding space. The pretrained language model is also fine-tuned with Low-Rank Adapters (LoRA) (Hu et al., 2021), allowing both image and text models to adapt jointly while reducing the risk of monomodal performance collapse (Alayrac et al., 2022; Liu et al., 2023b; Laurençon et al., 2024c; McKinzie et al., 2024; Marafioti et al., 2025).

Alignment Loss. For decoder-based models, we train with Causal Language Modeling (CLM) loss on the text tokens, as standardly done in VLM modality alignment:

$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{T} \log P_\theta\big(x_t \mid x_{<t}\big), \tag{1}$$

where $x_{<t}$ denotes all tokens preceding position t. We generalize this training scheme to bidirectional encoders models, by using the Masked Language Modeling (MLM) loss on the textual tokens:

$$\mathcal{L}_{\text{MLM}} = -\sum_{t\in\mathcal{M}} \log P_\theta\big(x_t \mid x_{\setminus\mathcal{M}}\big), \tag{2}$$

where $\mathcal{M}$ is the set of masked token positions and $x_{\setminus\mathcal{M}}$ is the input with those tokens masked out.

Modality Alignment Corpus. Models are modality aligned on a large corpus in large parts derived from The Cauldron 2 (Laurençon et al., 2024c) and Docmatix (Laurençon et al., 2024a). Our objective being to train document focused retrieval models, we use an adjusted training mixture that upsamples images containing text and documents with varying level of complexities. Our final training corpus consists of approximately 2B text tokens, and includes diverse sources such as web pages, books, and scientific papers. Mixture details are given in Appendix A.3.1. We note that controlling the exact data distribution during this phase enables the models we train to specialize early and achieve good document focused downstream performances which many large models struggle with (Liu et al., 2023a).

Parameters. All models are trained using a masking ratio of 0.5 and user-prompt masking to avoid overfitting on chat-template format (Huerta-Enochian & Ko, 2024; Shi et al., 2024; Allal et al., 2025). We employ WSD scheduler (Hu et al., 2024b) with the first 5% of the training as warmup, the last 20% as decay and a maximum learning rate of 1e-4. The ablation models are aligned on 3.5B tokens. We provide additional details on the training setup in Appendix A.1.

2.2 Contrastive Post-Training

Once the language model has learned to process image tokens jointly with text tokens, we specialize models through a contrastive post-training stage designed to enhance the semantic representation of the output embeddings produced by the model (Reimers & Gurevych, 2019).

Post-training Pairs. The post-training dataset used as starting point in our ablations comprises 118k document-query pairs from the ColPali corpus Faysse et al. (2025) as well as another 118k of natural image-description pairs from the MSCOCO train set (Lin et al., 2015).

Contrastive Loss. We employ the InfoNCE loss (van den Oord et al., 2019), defined as

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{q},\mathbf{d}^+) = -\log\frac{\Phi(\mathbf{q},\mathbf{d}^+)}{\Phi(\mathbf{q},\mathbf{d}^+) + \sum_{\mathbf{d}^-\in\mathcal{N}_q}\Phi(\mathbf{q},\mathbf{d}^-)}, \tag{3}$$

where $\mathbf{d}^+$ denotes the positive target for the query $\mathbf{q}$, $\mathcal{N}_{\mathbf{q}} = \mathcal{N}_{\mathbf{q}}^{\text{in}} \cup \mathcal{N}_{\mathbf{q}}^{\text{hard}}$ the set of negative targets (in-batch and hard negatives when mentioned), and $\Phi(\mathbf{q}, \mathbf{d})$ a similarity function between the token(s) of the query and the documents.[3]. For general-domain post-training we compute the loss symmetrically (Radford et al., 2021).

Batches Curation. In contrastive learning, batch diversity critically impacts retrieval entropy. Overly heterogeneous batches lead to trivial retrievals, while curated batches yield richer training signals. We employ task-aware batching (Li et al., 2023), grouping documents by source to ensure a homogeneous batch composition.

2.3 Ablation Evaluation Setup

The contrastively trained models are evaluated on retrieval and zero-shot classification tasks across multiple domains. Although the main focus remains document retrieval capabilities, evaluated by aggregating scores from the ViDoRe and ViDoRe v2[4] (Macé et al., 2025) benchmarks (nDCG@5), we also assess more generalist image retrieval capabilities by selecting tasks from MIEB (Xiao et al., 2025a). For natural image retrieval, we aggregate MSCOCO retrieval (Lin et al., 2015) and Flickr30k retrieval (nDCG@10) (Plummer et al., 2016) test sets. Finally, following practices in (Muennighoff et al., 2022), we assess both zero-shot and fine-tuning abilities of our models on general classification tasks. Specifically, we measure classification accuracy by fine-tuning a logistic regression head on top of our model's embedding on Stanford Cars (Krause et al., 2013) and Food101 (Bossard et al., 2014), and we evaluate zero-shot performance on FER2013 (Khaireddin & Chen, 2021) and EuroSAT (Helber et al., 2019) and aggregate the results.

3 What Makes a Great Visual Retriever?

Vision-language retrievers built upon existing generative VLMs often inherit design choices and weights that may not be well suited for all embedding tasks. Here, we analyze these critical design choices hoping to derive clear insights for developing efficient visual retrievers. Importantly, although we assess design decisions at different stages of the training pipelines, evaluation are always done end-to-end on the final evaluation signal.

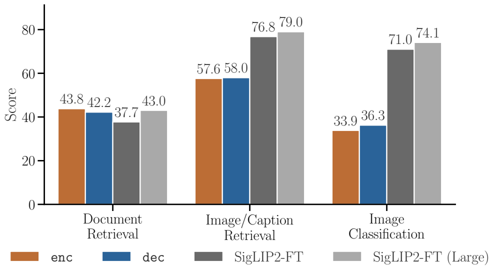


Figure 3: **Impact of Modality Alignment objective on downstream tasks.** Early Fusion of vision and text models boosts document retrieval tasks regardless of the LM objective, but degrades natural image and classification tasks w.r.t. the standalone *fine-tuned* vision model SigLIP. Reported scores are aggregated MIEB scores (nDCG, Accuracy.)

3.1 Modality Alignment Design

Language modeling Modality Alignment improves document understanding. According to benchmarks such as MIEB (Xiao et al., 2025a), dual encoder models explicitly trained on contrastive image-text tasks outperform repurposed VLMs in natural image classification tasks. To assess this, we train an encoder and a decoder vision-language model using the methodology described in section 2 on a mix of natural image and document data (alignment and contrastive training). We compare them with *SigLIP2-FT*, the 378M dual vision encoder model whose vision component is used by the vision tower of both VLMs, and with the larger *SigLIP2-FT Large* (881M parameters). Both SigLIP-FT models are finetuned in the same conditions as the VLMs, and initialized from pre-trained weights from scratch on billions of text-image pairs. [5]. As shown in Figure 3, the two early fusion VLM variants severely underperform the SigLIP2-FT dual

[3] We use the last (EOS) token for causal models, and mean pool all sequence tokens for bidirectional encoders for single-vector models. Alternatively, we use all document and query tokens without pooling for late interaction matching (Faysse et al., 2025). Details in Appendix A.2

[4] We report only the English splits of ViDoRe v2, as our base models are trained on English data only.

[5] We report the performance of the untrained *off-the-shelf* SigLIP in Appendix C.1

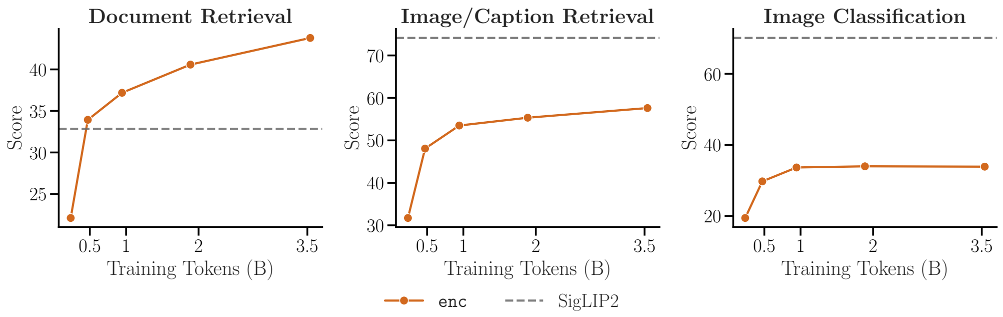


Figure 4: **Modality alignment scaling of early fusion encoders for up to 1 epoch (3.5B tokens) of data.** The dashed line indicates the vision encoder evaluated standalone without further training. Our findings show that retrieval tasks benefits from extended modality alignment phase, particularly in document retrieval, where performance quickly surpasses that of the standalone vision encoder.

encoders on natural image tasks. In contrast, they achieve significant gains on document retrieval tasks (+6.1 nDCG@5 on ViDoRe and ViDoRe v2 datasets w.r.t. base), even edging out *SigLIP2-FT Large* that contains 3.5x vision parameters more than both VLMs.

This confirms large-scale contrastive training remains best for high-level image representation tasks (natural images), but sequentially combining a vision model with a pretrained language model facilitates document representation tasks, even with significantly less contrastive post-training. As the rest of this paper shows, steering away from the dual encoder architecture further enables improving performance through many avenues other than text to image contrastive training, for which supervised training samples can be hard to obtain.

Scaling the modality alignment phase for better token representations. Prior work shows that scaling the modality alignment phase of VLMs improves their generative abilities (Beyer et al., 2024; McKinzie et al., 2024; Wang et al., 2024). We test whether similar gains hold in retrieval by contrastively fine-tuning `enc` checkpoints during MLM modality alignment. Figure 4 illustrates the results of post-trained checkpoints on diverse tasks. Although document retrieval improves consistently with more modality alignment data – largely surpassing the vision tower evaluated in isolation and showing clear scaling benefits – natural image tasks plateau past 1B tokens, far from the standalone dual encoder baseline. This shows that document and natural image retrieval leverage different mechanisms and should not be optimized the same way. *Document Retrieval benefits from learning fine-grained interactions between image and text tokens through the language model, while the LM has limited utility for high level natural image tasks.*

Bidirectional attention fully unlocks Late Interaction. Inspired by the effectiveness of bidirectional attention in text-only retrieval (Gisserot-Boukhlef et al., 2025; Weller et al., 2025)[6], we investigate if it surpasses causal attention in *visual document retrieval*, particularly when using the multi-vector late interaction matching common in SOTA visual retrievers (Khattab & Zaharia, 2020; Faysse et al., 2025). Figure 5 reports single vector and late interaction results on the ViDoRe benchmark for various model variants. On top of the standard `enc` (MLM) and `dec` (CLM) models, we evaluate the `dec-enc` and the `dec` models modality aligned with MLM objectives to determine whether bidirectional attention capabilities can be obtained in later stages of training.

Single-vector embedding results are close between bidirectional and causal attention models for document retrieval, with `enc` slightly outperforming `dec` by +1.6 nDCG@5.

Intuitively however, bidirectional attention makes a huge difference when used in late interaction settings, substantially exceeding the causal counterpart by +10.6 nDCG@5. Causal decoders are incapable of correctly contextualizing image or text token representations seen at the beginning of the sequences. *This is a key result as almost all current visual retrievers, including late interaction variants, are causal models, clearly indicating some performance is left on the table.*

[6] Chen et al. (2025) investigate post-hoc removal of the attention mask during visual retrieval fine-tuning.

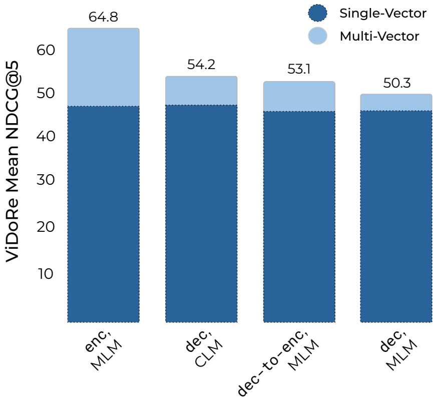


Figure 5: **Impact of attention masks and training objectives on document retrieval performances.** We report the average nDCG@5 on English splits of ViDoRe benchmarks for models post-trained on ColPali.

	HR Cooldown	Document Retrieval	Image/Caption Retrieval	Image Classification	**Average**
512px	✗	30.7	58.8	**41.4**	43.6
1024px	✗	42.2	**58.9**	37.2	**46.1**
2048px	✗	43.8	57.6	33.9	45.1
2048px	✓	**45.8**	57.8	33.7	45.8

Table 1: **Effect of image resolution on VL encoder abilities.** Document retrieval performance increases with higher image resolution. Further annealing the encoder on high-resolution images (HR Cooldown) at the end of modality alignment yields additional gains. By contrast, for non-document tasks, raising the resolution tends to degrade performance.

Removing the causal attention mask during training does not suffice to recover the `enc` late interaction performance at these data regimes. This indicates converting trained decoders as late interaction retrievers is highly non trivial, and confirms the insights from Weller et al. (2025); when possible, training encoder models from scratch remain better for retrieval tasks.

3.2 Contrastive Training Design

The previous subsection established bidirectional encoder models to often be the best option when training visual retrievers. In the following experiments, we assess contrastive training choices and only report results for the encoder model for simplicity.

Image resolution benefits are task-specific. Image resolution plays a critical role in VLM generative capabilities, notably in document-focused tasks, as higher-resolution inputs enables the model to capture finer visual cues (Hu et al., 2024a; Marafioti et al., 2025). Modality alignment is done at a fixed image resolution of 1024 pixels (longer side) and we report scores of contrastive training runs with varying settings in Table 1. To vary the resolution, images of the highest quality available are scaled to the desired size (often downscaled) before being fed to the image tokenizer. Our findings confirm that embedding tasks are strongly sensitive to image-resolution. In particular, *training with higher resolution inputs substantially improves the results on visual document retrieval benchmarks*, consistent with prior work in generative settings Beyer et al. (2024); McKinzie et al. (2024). Furthermore, adding a cool-down phase by showing higher-resolution images towards the end of the modality alignment phase yields additional gains. This suggests that models can adapt their attention mechanisms to finer details when exposed to increased resolution. Interestingly, these findings do not hold in natural image tasks, where high resolution can even degrade performance.

	Document Retrieval	Image/Caption Retrieval	Image Classification	**Average**
Baseline CL Mix	43.9	**57.2**	36.1	45.7
+ *Text→Text Pairs*	45.6	53.2	35.7	44.8
+ *Image→Caption Pairs*	**45.8**	54.4	**49.9**	**50.0**

Table 2: **Impact of contrastive training mixtures on downstream tasks.** Incorporating text-only pairs improves performance on document retrieval, but degrades other performances. Adding natural images-captions pairs substantially enhances performance on classification tasks.

Increasing the pool of contrastive pairs. A severe limitation that current visual retrievers face is the lack of large volumes of high quality (document image, query pairs). Previous work (Ma et al., 2024; Faysse et al., 2025; Jiang et al., 2025; Zhang et al., 2025a) has relied on a mix of repurposed existing visual question answering datasets and synthetically generated queries with external LLMs. Even put together however, the field is only a year old, and these datasets remain small in size and often of poor quality.

A central question in our study is whether the abundance of *text-only* query–document pairs can be exploited to improve *visual* retrieval via cross-modal capability transfer. To probe this, we run contrastive training under three regimes. Unlike prior work that "warms up" visual retrievers or trains exclusively with text-only pairs (Ma et al., 2024; Jiang et al., 2024), we *interleave* text-only pairs and text–image pairs throughout training at a 1:1 ratio. The dataset sources are detailed in Appendix A.3.3

As reported in Table 2, incorporating text-only pairs yields a sizeable improvement on visual document retrieval (+1.7 nDCG@5), indicating clear cross-modal transfer—likely facilitated by the backbone's jointly learned text–image embedding space. This result suggests that domain-specific training corpora can be assembled irrespective of native modality, reducing duplication of effort and lowering data-collection costs.

We further evaluate training with *NatCap*, a corpus of natural images paired with synthetic, highly detailed captions (see Appendix A.3.2). This scaling step improves downstream performance across the board—most notably on natural-image tasks, and with a smaller but consistent gain on document retrieval (+0.2 nDCG@5). Together, these findings underscore the importance of scaling contrastive learning with high-quality data, but which doesn't need to be exclusively image document focused.

4 Building a Small yet Mighty Visual Retriever.

4.1 Training.

Recipe. Putting together the results from our experiments, we devise a training recipe for a small visual document retriever *ModernVBERT*. It combines a state-of-the-art 150M text bidirectional encoder (Weller et al., 2025) with the ModernBERT architecture (Warner et al., 2024a) and a small vision encoder SigLIP2-16B-512 of 100M parameters (Tschannen et al., 2025). We modality align both models with a MLM objective for 10B tokens, 3 times longer than during our experiments. To boost document understanding, we augment the input image resolution from 1024px to 2048px during a modality alignment cooldown stage (2B tokens). We call the resulting model *ModernVBERT*. Following the findings of Section 3.2, we then scale the contrastive training mix from previous experiments to combine document–query pairs with text-only pairs, and use 1 hard negatives for each document-query pair and 2 for each text-only pairs. We opt for a 2/1 text-to-image ratio following our ablation results introduced in Appendix C.3.1. This results in *ColModernVBERT*, a compact late interaction model. For reference, we also train *BiModernVBERT*, a single vector variant. More training details are provided in Appendix A.1.

4.2 Results.

ColModernVBERT. The resulting model, *ColModernVBERT* showcases strong performances on visual document retrieval benchmarks, especially relative to its size category (Figure 1). Despite having over 10 times less parameters than models such as ColPali released only a year ago, it is only 0.6 nDCG@5 points below on the aggregated ViDoRe benchmark scores (Table 3). It also edges many larger single-vector repurposed VLM models released within the year (Chen et al., 2025; Jiang et al., 2024; 2025). It however falls short of top model performance on ViDoRe which are built on larger decoder VLMs pretrained and aligned on billions of tokens of text and image data.

Most sub-1B parameter models evaluated on document retrieval benchmarks are dual encoder models, since early fusion generative models that perform well are not common at this scale. The most related model is a 176M late interaction model, ColFlor (Masry & Hoque, 2024), trained from the Florence2 model (Xiao et al., 2023). ColFlor is 12.7 nDCG@5 points under *ColModernVBERT*. *ColModernVBERT* also largely outperforms off-the-shelf dual encoders, even when those have substantially larger parameter counts. These results highlights the benefits of multi-phase training and

	Late Interaction	Model Size (B)	ViDoRe(v1)	ViDoRe(v2,eng)	**Average**	**Latency (ms)**
≥ 1B Parameters						
MoCa-3B (Chen et al., 2025)		3.75	80.1	53.8	66.9	158
VLM2Vec (Jiang et al., 2025)		4.15	49.8	36.5	43.1	211
GME-Qwen2 (Zhang et al., 2025a)		8.29	89.9	61.8	75.8	412
E5-V (Jiang et al., 2024)		8.36	62.7	49.4	56.1	434
ColPali (Faysse et al., 2025)	✓	2.92	81.6	56.8	69.2	175
ColQwen2.5 (Faysse et al., 2025)	✓	3.75	89.5	61.5	75.5	158
Jina-v4 (Günther et al., 2025)	✓	3.75	90.4	60.1	75.2	158
NemoRetriever-3B (Xu et al., 2025)	✓	4.40	91.0	66.3	78.7	155
≤ 1B Parameters						
Jina CLIP* (Koukounas et al., 2024)		0.22	17.6	14.0	15.8	**14**
BGE Visualized M3* (Zhou et al., 2024)		0.87	12.4	10.2	11.3	38
SigLIP2-L-512/16* (Tschannen et al., 2025)		0.88	43.8	27.0	35.4	25
ColFlor (Masry & Hoque, 2024)	✓	0.17	68.8	43.0	55.9	17
BiModernVBERT (ours)		0.25	63.6	35.7	49.7	20
***ColModernVBERT* (ours)**	✓	0.25	**81.2**	**56.0**	**68.6**	20

Table 3: **Performance on ViDoRe.** Our model *ColModernVBERT* offers the best performance-size tradeoff, significantly outperforming existing sub-1B models and matching the performance of models up to 10x larger with substantially lower inference CPU latency Details and GPU latencies in Appendix C.6.2. Models marked with * are not specifically trained for VDR. Bold values indicate the best performance amongst sub-1B models.

early fusion architectures for multi-modal document related tasks, even at smaller parameter counts. We also attribute the strong performance of *ColModernVBERT* at smaller model sizes to the symbiosis of native bidirectional attention and Late Interaction matching, which largely boosts performance relative to comparable decoder models (Section 3.1).

Speed. As noted by Xiao et al. (2025b), multi-vector visual retrievers are not bottlenecked in their inference speed by the late interaction matching operation, but rather by the latency required to encode queries with the text model. Our model demonstrates that strong performance is not incompatible with speed, even when running inference on consumer CPUs, which is the standard setting in most industrial local deployments of text embedding models. Latencies are computed by averaging query encoding times of all NanoBEIR queries, which are 23.4 word and 147.5 character long on average, and are run with batch size 1 to replicate online use cases. To prevent RAM bottlenecks, we benchmark on very high RAM (2TB) CPU cloud environments, but note models larger than 3B parameter require more than 12 GB RAM to run optimally.[7] (Table 3). *ModernVBERT* achieves more than a 7x speedup on CPU over models with similar performances on ViDoRe. We further report model latency results on GPU hardware in Appendix C.6.2. We notably demonstrate that with batched inference, *ModernVBERT* based query encoders are able to encode 5000 queries per second on Nvidia H100 GPUs. ModernVBert's small model size also enables efficient batching when encoding documents.

5 Related Work

Repurposing VLMs for Representation Learning. Motivated by the zero-shot performances of generative VLMs (Alayrac et al., 2022; Lucas Beyer* et al., 2024; Bai et al., 2023), recent studies have explored repurposing these for multimodal embedding tasks (Ma et al., 2024; Faysse et al., 2025; Jiang et al., 2025; Zhang et al., 2025a). As backbone generative models improved, retriever

[7] With more standard CPU RAM settings such as those found in low-end servers or Google Colab (12GB RAM), models above 3B parameters must rely on memory offloading to run, which adds up to dozens of seconds of latency per query.

performance improved as well showcasing the central impact of language model pretraining and modality alignment (Xu et al., 2025; Nussbaum et al., 2025). These model remain inherently constrained by their causal attention mechanisms which has been shown in text settings to limits represational efficiency (Gisserot-Boukhlef et al., 2025; Weller et al., 2025). Recent work attempts to address this issue by modifying VLM attention during continual pretraining (Chen et al., 2025) or contrastive tuning (Jiang et al., 2025; Xu et al., 2025), but no recent work attempts to align natively bidirectional language encoder models with vision encoders. The recent release of long sequence text encoders (Warner et al., 2024a; Boizard et al., 2025) makes this possible.

Late Interaction in Visual Document Retrieval To further boost performance, visual document retrievers leverage the late interaction mechanism (Khattab & Zaharia, 2020) which matches multiple query embeddings with multiple document embeddings through the MaxSim operation (Faysse et al., 2025; Günther et al., 2025; Xu et al., 2025). This enables more granular interactions between image and query tokens, at the cost of additional storage and a slight compute overhead during the matching operation. Efficiency gains have come from improving the storage costs through quantization (Bergum, 2025), token pruning (Faysse et al., 2024) and more recently the use of Matrioshka losses to compact multi-token representations (Xiao et al., 2025b). Ultimately, the performance bottleneck when running visual retrieval inference with such models now resides mostly in the necessity to rely on costly GPU hardware to encode queries, which sets apart text from vision retrieval. This paper fills this gap by using encoders that run on CPU, of parameter sizes comparable to commonly used local text embedding models (Chen et al., 2024; Enevoldsen et al., 2025).

6 Conclusion

In this paper we question design decisions of current VLM-based retriever models, providing crucial insights into what matters when training early-fusion vision encoders. Our study notably shows that these models generally do not improve retrieval on natural-image tasks compared to dual encoders, yet strong vision-language alignment is essential for document-centric retrieval. We uncover a tight synergy between bidirectional attention and late-interaction retrieval, which underscores a fundamental limitation of repurposing decoder-style generative VLMs for retrieval. To mitigate data scarcity in contrastive learning, we propose augmenting limited image-document/text-query pairs with larger, lower-cost corpora from other modalities. Guided by these insights, we trained *ColModernVBERT*, a compact yet powerful 250M-parameter multimodal encoder that matches the performance of models up to 10× larger on visual retrieval benchmarks. We release models and training code to help practitioners reduce cost and latency when deploying visual retrievers in real-world applications, and to encourage research on efficient multimodal embedding models.

Future Work & Limitations. By design, our analysis targets relatively small models. An important next step is to test whether the observed patterns persist at larger scales—for example, to more rigorously probe the interplay between late interaction and bidirectional attention. Our study also focuses exclusively on English. While we expect the broad trends to generalize and see clear value in releasing multilingual variants, it remains unclear how allocating parameters to additional languages trades off against the understanding of the vision modality, and to what extent this penalizes English retrieval performance as the number of languages are scaled (Fernandes et al., 2023). Finally, although we center on retrieval and sequence-level zero-shot classification, the modality-aligned encoder can be fine-tuned for a range of token-level tasks, including OCR error detection, token-level classification, visual named entity recognition, visually grounded token-level object detection, contextual embeddings (Conti et al., 2025). We release our base model to encourage exploration of these directions.

Ethics Statement

Environmental Costs. Training *ColModernVBERT*required approximately 2,000 H100 GPU-hours in total, which we estimate corresponds to 41 kg of CO_2[8], based on standard assumptions of GPU power draw, datacenter efficiency, and grid carbon intensity. This estimate follows methodologies

[8] Carbon footprint estimated with *Green Algorithms* (Lannelongue et al., 2021): $E = t \times P \times \mathrm{PUE}$, $\mathrm{CO_2e} = E \times \mathrm{CI}$. With $t = 2000$ GPUh, $P = 0.35$ kW (H100 PCIe), PUE $= 1.3$, and CI $= 45$ gCO$_2$/kWh, this gives $E \approx 910$ kWh and $\mathrm{CO_2e} \approx 41$ kg.

such as Green Algorithms (Lannelongue et al., 2021) and related analyses of the carbon footprint of machine learning (Strubell et al., 2019; Patterson et al., 2021). Across the entire project, all combined experiments totaled about 18k H100-hours. To mitigate costs and promote sustainable research practices, we release all model checkpoints and training artifacts to facilitate reuse, extension, and reproducibility without necessitating retraining. Additionally, this work shows efficiency gains with smaller models to aim to limit the inference costs of visual retrieval, and consequently reduce the environmental footprint. Our model performs query encoding efficiently on CPUs, keeping inference costs low and reducing barriers to adoption.

Safety and Bias. From a safety perspective, our encoder-only retriever poses less risk than generative models: it produces fixed-length embeddings rather than free-form content, reducing avenues for harmful content generation, hallucination, or deceptive outputs; nonetheless, retrieval systems can still propagate biases present in the underlying data, which we address through dataset curation open release.

AI Assistance. Parts of this paper were prepared with the assistance of an AI-based writing tool used for copy editing and stylistic refinement. All generated text was carefully reviewed, verified, and revised by the authors, who take full responsibility for the accuracy and originality of the final manuscript.

Reproducibility Statement

For transparency and to foster future work, we release our training data, model checkpoints (base models and adapters), and the complete codebase under the MIT License, as detailed in the main paper and repository. The supplementary material specifies training configurations for all models (also provided in the corresponding HuggingFace repositories), describes our synthetic data generation process, and reports expanded evaluation results to support exact replication.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, 2024. URL `https://arxiv.org/abs/2404.14219`. Version Number: 2.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL `https://arxiv.org/abs/2204.14198`.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model, 2025. URL `https://arxiv.org/abs/2502.02737`.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 1, learning hierarchical language structures, 2025. URL https://arxiv.org/abs/2305.13673.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. 2023. doi: 10.48550/ARXIV.2308.12966. URL https://arxiv.org/abs/2308.12966. Publisher: arXiv Version Number: 3.

Jo Bergum. Scaling ColPali to billions of PDFs with Vespa — blog.vespa.ai. https://blog.vespa.ai/scaling-colpali-to-billions/, 2025.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024. URL https://arxiv.org/abs/2407.07726.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. Eurobert: Scaling multilingual encoders for european languages, 2025. URL https://arxiv.org/abs/2503.05500.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 446–461, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10599-4.

Antoine Chaffin. Gte-moderncolbert, 2025. URL https://huggingface.co/lightonai/GTE-ModernColBERT-v1.

Haonan Chen, Hong Liu, Yuping Luo, Liang Wang, Nan Yang, Furu Wei, and Zhicheng Dou. Moca: Modality-aware continual pre-training makes better bidirectional multimodal embeddings, 2025. URL https://arxiv.org/abs/2506.23115.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, 2024. URL https://arxiv.org/abs/2402.03216. Version Number: 3.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost, 2016. URL https://arxiv.org/abs/1604.06174.

Benjamin Clavié. Towards better monolingual japanese retrievers with multi-vector models, 2024. URL https://arxiv.org/abs/2312.16144.

Cohere. Introducing Rerank 3: A New Foundation Model for Efficient Enterprise Search & Retrieval, April 2024. URL https://cohere.com/blog/rerank-3.

Max Conti, Manuel Faysse, Gautier Viaud, Antoine Bosselut, Céline Hudelot, and Pierre Colombo. Context is gold to find the gold passage: Evaluating and training contextual document embeddings. *arXiv preprint arXiv:2505.24782*, 2025.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL https://arxiv.org/abs/2307.08691.

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. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2020. doi: 10.48550/ARXIV.2010.11929. URL https://arxiv.org/abs/2010.11929. Publisher: arXiv Version Number: 2.

Sebastian Dziadzio, Vishaal Udandarao, Karsten Roth, Ameya Prabhu, Zeynep Akata, Samuel Albanie, and Matthias Bethge. How to merge your multimodal models over time?, 2024. URL https://arxiv.org/abs/2412.06712.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. Mmteb: Massive multilingual text embedding benchmark, 2025. URL https://arxiv.org/abs/2502.13595.

Manuel Faysse, Patrick Fernandes, Nuno M. Guerreiro, António Loison, Duarte M. Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro H. Martins, Antoni Bigata Casademunt, François Yvon, André F. T. Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. CroissantLLM: A Truly Bilingual French-English Language Model. 2024. doi: 10.48550/ARXIV.2402.00786. URL https://arxiv.org/abs/2402.00786. Publisher: arXiv Version Number: 3.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models, 2025. URL https://arxiv.org/abs/2407.01449.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. Scaling laws for multilingual neural machine translation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10053–10071. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/fernandes23a.html.

Hippolyte Gisserot-Boukhlef, Nicolas Boizard, Manuel Faysse, Duarte M. Alves, Emmanuel Malherbe, André F. T. Martins, Céline Hudelot, and Pierre Colombo. Should we still pretrain encoders with masked language modeling?, 2025. URL https://arxiv.org/abs/2507.00994.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and Han Xiao. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval, 2025. URL https://arxiv.org/abs/2506.18902.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2019. URL https://arxiv.org/abs/1709.00029.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding, 2024a. URL https://arxiv.org/abs/2409.03420.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024b. URL https://arxiv.org/abs/2404.06395.

Mathew Huerta-Enochian and Seung Yong Ko. Instruction fine-tuning: Does prompt loss matter?, 2024. URL https://arxiv.org/abs/2401.13586.

Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights, 2022. URL https://arxiv.org/abs/2208.05592.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models, 2024. URL https://arxiv.org/abs/2407.12580.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks, 2025. URL https://arxiv.org/abs/2410.05160.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering, 2020. URL https://arxiv.org/abs/2004.04906. Version Number: 3.

Yousif Khaireddin and Zhuofa Chen. Facial emotion recognition: State of the art performance on fer2013, 2021. URL https://arxiv.org/abs/2105.03588.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. 2020. doi: 10.48550/ARXIV.2004.12832. URL https://arxiv.org/abs/2004.12832.

Andreas Koukounas, Georgios Mastrapas, Michael Günther, Bo Wang, Scott Martens, Isabelle Mohr, Saba Sturua, Mohammad Kalim Akram, Joan Fontanals Martínez, Saahil Ognawala, Susana Guzman, Maximilian Werk, Nan Wang, and Han Xiao. Jina CLIP: Your CLIP Model Is Also Your Text Retriever, 2024. URL https://arxiv.org/abs/2405.20204. Version Number: 1.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

Loïc Lannelongue, Joe Grealey, and Michael Inouye. Green algorithms: Quantifying the carbon footprint of computation. *Advances in Science*, 7(34):eabf3899, 2021. doi: 10.1126/sciadv.abf3899.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions., 2024a.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, May 2024b. URL http://arxiv.org/abs/2405.02246. arXiv:2405.02246 [cs].

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024c. URL https://arxiv.org/abs/2405.02246.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models, 2025. URL https://arxiv.org/abs/2405.17428.

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. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2020. URL https://arxiv.org/abs/2005.11401. Version Number: 4.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL https://arxiv.org/abs/2201.12086.

Mingxin Li, Zhijie Nie, Yanzhao Zhang, Dingkun Long, Richong Zhang, and Pengjun Xie. Improving general text embedding model: Tackling task conflict and data imbalance through model merging, 2024. URL https://arxiv.org/abs/2410.15035.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning, 2023. URL https://arxiv.org/abs/2308.03281.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, 2014. URL https://arxiv.org/abs/1405.0312. Version Number: 3.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11238–11254, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.772. URL https://aclanthology.org/2022.emnlp-main.772/.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models, 2023. URL https://arxiv.org/abs/2311.07575.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning, 2023a. URL https://arxiv.org/abs/2310.03744. Version Number: 2.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b. URL https://arxiv.org/abs/2304.08485.

Lucas Beyer*, Andreas Steiner*, André Susano Pinto*, Alexander Kolesnikov*, Xiao Wang*, Xiaohua Zhai*, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Jeremiah Harmsen, Daniel Keysers, Neil Houlsby, Xi Chen, Emanuele Bugliarello, Thomas Unterthiner, Keran Rong, Matthias Minderer, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Julian Eisenschlos, Manoj Kumar, Matko Bošnjak, Matthias Bauer, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Paul Voigtlaender, Pinelopi Papalampidi, Olivier Henaff, Skanda Koppula, Xi Xiong, Radu Soricut, Model release contributors and general support, Tris Warkentin, Kat Black, Luiz Gustavo Martins, Glenn Cameron, Raj Gundluru, Manvinder Singh, Meg Risdal, Nilay Chauhan, Nate Keating, Nesh Devanathan, Elisa Bandy, Joe Fernandez, Antonia Paterson, Jenny Brennan, Tom Eccles, Pankil Botadra, Ben Bariach, Lav Rai, Minwoo Park, Dustin Luong, Daniel Vlasic, Bo Wu, Wenming Ye, Divyashree Sreepathihalli, Kiranbir Sodhia, Alek Andreev, Armand Joulin, Surya Bhupatiraju, Minh Giang, Joelle Barral, and Zoubin Ghahramani. PaliGemma, 2024. URL https://www.kaggle.com/m/23393.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding, 2024. URL https://arxiv.org/abs/2406.11251.

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

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models, 2025. URL https://arxiv.org/abs/2504.05299.

Ahmed Masry and Enamul Hoque. Colflor: Towards bert-size vision-language document retrieval models. 2024.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods, analysis & insights from multimodal llm pre-training, 2024. URL https://arxiv.org/abs/2403.09611.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive Text Embedding Benchmark, 2022. URL https://arxiv.org/abs/2210.07316. Version Number: 3.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder, 2025. URL https://arxiv.org/abs/2402.01613.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Haisam Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. The carbon footprint of machine learning training will plateau, then shrink. *IEEE Computer*, 54(12):18–28, 2021. doi: 10.1109/MC.2021.3120015.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016. URL https://arxiv.org/abs/1505.04870.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models, 2020. URL https://arxiv.org/abs/1910.02054.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019. URL https://arxiv.org/abs/1908.10084. Version Number: 1.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, 2016. URL https://arxiv.org/abs/1609.05158.

Zhengyan Shi, Adam X. Yang, Bin Wu, Laurence Aitchison, Emine Yilmaz, and Aldo Lipani. Instruction tuning with loss over instructions, 2024. URL https://arxiv.org/abs/2405.14394.

Ken Shoemake. Animating rotation with quaternion curves. 19(3):245–254, July 1985. ISSN 0097-8930. doi: 10.1145/325165.325242. URL https://doi.org/10.1145/325165.325242.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1355.

Yi-Lin Sung, Linjie Li, Kevin Lin, Zhe Gan, Mohit Bansal, and Lijuan Wang. An empirical study of multimodal model merging, 2023. URL https://arxiv.org/abs/2304.14933.

Nandan Thakur, Crystina Zhang, Xueguang Ma, and Jimmy Lin. Fixing data that hurts performance: Cascading llms to relabel hard negatives for robust information retrieval, 2025. URL https://arxiv.org/abs/2505.16967.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023. doi: 10.48550/ARXIV.2307.09288. URL https://arxiv.org/abs/2307.09288. Publisher: arXiv Version Number: 2.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text Embeddings by Weakly-Supervised Contrastive Pre-training, 2022. URL https://arxiv.org/abs/2212.03533. Version Number: 2.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024a. URL https://arxiv.org/abs/2412.13663.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024b. URL https://arxiv.org/abs/2412.13663.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. URL https://arxiv.org/abs/2206.07682.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. Seq vs seq: An open suite of paired encoders and decoders, 2025. URL https://arxiv.org/abs/2507.11412.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks, 2023. URL https://arxiv.org/abs/2311.06242.

Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth Enevoldsen, and Niklas Muennighoff. Mieb: Massive image embedding benchmark, 2025a. URL https://arxiv.org/abs/2504.10471.

Zilin Xiao, Qi Ma, Mengting Gu, Chun cheng Jason Chen, Xintao Chen, Vicente Ordonez, and Vijai Mohan. Metaembed: Scaling multimodal retrieval at test-time with flexible late interaction, 2025b. URL https://arxiv.org/abs/2509.18095.

Mengyao Xu, Gabriel Moreira, Ronay Ak, Radek Osmulski, Yauhen Babakhin, Zhiding Yu, Benedikt Schifferer, and Even Oldridge. Llama nemoretriever colembed: Top-performing text-image retrieval model, 2025. URL https://arxiv.org/abs/2507.05513.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. Qwen2.5-1m technical report, 2025. URL https://arxiv.org/abs/2501.15383.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Alex Lin, and Fei Huang. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model, 2023. URL https://arxiv.org/abs/2310.05126.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms, 2025a. URL https://arxiv.org/abs/2412.16855.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025b. URL https://arxiv.org/abs/2506.05176.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. Vista: Visualized text embedding for universal multi-modal retrieval, 2024. URL https://arxiv.org/abs/2406.04292.

A Training

A.1 Implementation and Resources

Model	Batch Size	Learning Rate	Training Steps	Training GPU Hours
Modality Alignment				
ModernVBERT-base (Table 5)	4096	1e-4	5500	1920h
Contrastive Learning				
Generalist contrastive training (Table 7)	256	2e-4	3917	80h
Document Specialization				
Document-focused contrastive training w/ hard negatives (Table 7)	64	2e-4	19602	160h

Table 4: Training details of our final models at each training stage. GPU Hours are on 80GB H100 GPUs.

We list hyperparameters and resource details in Table 4 for the various training stages of our final models. We employ ZeRO stage 1 optimizer (Rajbhandari et al., 2020) for our modality alignment runs. All ablation models are contrastively trained with gradient checkpointing (Chen et al., 2016) to reduce memory usage. All training runs are performed with FlashAttention 2.0 (Dao, 2023). For LoRA configurations, we consistently use a rank `r` of 32, `lora_alpha` of 32, and a dropout of 0.1. For the implementation, we start from m4[9] and ColPali[10] codebases for training, and use the MTEB[11] repository for evaluation.[12]

A.2 Similarity Functions

Single-Vector Similarity. For single-vector models, we apply mean pooling for MLM-aligned encoders and end-of-sequence (EOS) pooling for CLM-based models and compute the cosine similarity of a query q and a document d as

$$\Phi_{\text{CosSim}}(\mathbf{q}, \mathbf{d}) = \exp(\cos(\mathbf{E}_q, \mathbf{E}_d)/\tau) \tag{4}$$

Multi-Vector Similarity. For multi-vector models, we adopt the standard late-interaction scoring function defined as:

$$\Phi_{\text{LI}}(q, d) = \sum_{i \in [\![1, N_q]\!]} \max_{j \in [\![1, N_d]\!]} \left\langle \mathbf{E}_q^{(i)}, \mathbf{E}_d^{(j)} \right\rangle, \tag{5}$$

where $\mathbf{E}_q^{(i)}$ and $\mathbf{E}_d^{(j)}$ denote token-level embeddings for the query and document, respectively.

A.3 Data

A.3.1 Modality Alignment Mixture

For our modality alignment trainings, we rely on The Cauldron dataset (Laurençon et al., 2024b) and its Docmatix extension (Laurençon et al., 2024a). Table 5 provides further details on the constitution of this dataset.

A.3.2 *NatCap*

To enrich our contrastive learning data mixture, we construct *NatCap* (Natural Captions), a large-scale dataset containing around 333000 contextualized image–caption pairs. This dataset is created by generating synthetic captions, along with cross-class and in-class discriminative tags, from existing image classification datasets (see Table 6). For this purpose, we leverage `Gemini-flash-2.5`[13] which produces captions conditioned on both the image content and the accompanying dataset metadata, as illustrated in Figure 6. We detail the prompt below.

[9] SmolVLM trainer, https://github.com/huggingface/smollm

[10] https://github.com/illuin-tech/colpali

[11] https://github.com/embeddings-benchmark/mteb

[12] We will release our training codebases in the public version of this paper

[13] https://ai.google.dev/gemini-api/docs/models?hl=fr#gemini-2.5-flash

Dataset Subsection	# Images	# QA Pairs	# Tokens	% Mix
Captioning	609,843	612,768	62,906,011	3.13
Real-world VQA	457,360	2,125,615	23,318,335	1.16
OCR, Document Understanding	2,499,258	11,415,478	426,806,479	21.21
Chart/Figure Understanding	539,743	24,444,120	30,315,784	1.51
Table Understanding	163,568	229,077	21,371,931	1.06
Reasoning, Logic, Maths	490,870	2,212,629	32,450,213	1.61
Screenshot to Code	547,974	548,296	336,299,551	16.71
Text-only Instructions	0	21,482,682	1,079,001,075	53.61
Total	**5308616**	**63070665**	**2012469379**	**100.00**

Table 5: Aggregated statistics of modality alignment datasets from The Cauldron 2 (Laurençon et al., 2024c) and Docmatix (Laurençon et al., 2024a), showing image counts, QA pairs, token counts, and the proportional contribution of each subsection to the overall mixture.

Dataset	Description	# Items
Caltech101	General objects.	3.000
Caltech256	General objects.	30.000
Cars	Car model classification.	8.000
Country211	Country where the picture is taken.	28.000
DTD	Describable textures (texture attributes).	4.000
EuroSat	Land use / area zone type.	16.000
FER2013	Facial emotion recognition.	28.000
FGCVAircraft	Aircraft model recognition.	3.000
Food101	Food categories.	75.000
OxfordPets	Dog/cat species.	3.000
RESISC45	Aerial scene / area zone type.	18.000
SUN397	General scenes.	109.000
VOC2007	General objects.	8.000
TOTAL		**333000**

Table 6: ***NatCap* Dataset Composition.** *NatCap* spans 13 different sources covering various images types. The total dataset is composed of 333k pairs

A.3.3 Contrastive Training Mix

In this subsection, we describe the composition of our data mixes used in the contrastive training stages. Table 7 outlines the datasets included in each mix, including the Document-Focused variant employed for *ColModernVBERT*.

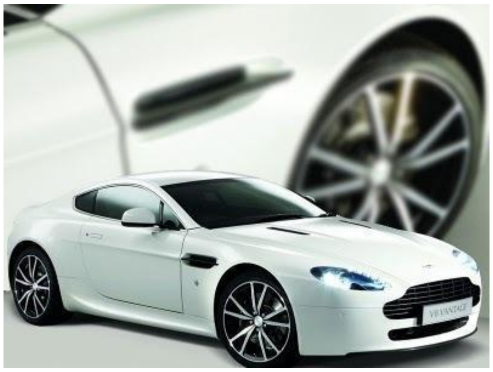
B Baselines Details

In this section, we describe the models evaluated in as comparison to our document retriever model.

MoCa-3B (Chen et al., 2025). A modality-aware continual pretraining model that transforms a causal vision-language model into a bidirectional multimodal embedding model, using interleaved image-text reconstruction and contrastive alignment to support cross-modal retrieval.

GME-Qwen2 (Zhang et al., 2025a). A unified multimodal embedder built on Qwen2-VL (Wang et al., 2024), which produces shared embedding representations across text, image, and fused input modalities, enabling universal multimodal retrieval.

VLM2Vec (Jiang et al., 2025). A method that trains a vision-language encoder by converting a VLM through extensive contrastive post-training. Flagship model is based on the model Phi-3.5 (Abdin et al., 2024).



Class label: “Aston Martin V8 Vantage Coupe 2012”

NatCap caption: “A white 2012 Aston Martin V8 Vantage Coupe is showcased against a blurred background featuring a large car wheel.”

Figure 6: Example from the NatCap dataset

Source	Description	Pairs	Epochs
Generalist Mix			
ColPali (Faysse et al., 2025)	Query–Document images for visual retrieval	118k	1
MSCOCO (Lin et al., 2014)	Natural images with human-written captions	118k	1
NatCap (ours, subsampled)	Diverse images with synthetic captions	118k	1
RLHN (Thakur et al., 2025)	Text–text pairs for complex retrieval	680k	1
TOTAL		**1030k**	
Document-Focused Mix			
ColPali (Faysse et al., 2025)	Query–Document images for visual retrieval	118k	3
RLHN (Thakur et al., 2025)	Text–text pairs for complex retrieval	300k	3
TOTAL		**1254k**	

Table 7: **Data mixes for contrastive trainings.** The *Generalist Mix* spans over 1M diverse pairs, while the *Document-Focused Mix* emphasizes document retrieval with extra ColPali epochs.

E5-V (Jiang et al., 2024). An adaptation of the E5 embedding approach to multimodal models: it trains only on text pairs yet bridges the modality gap to handle image inputs, reducing cost while achieving universal embeddings.

ColPali (Faysse et al., 2025). A vision-based document retrieval model that processes document pages as images (no OCR) and produces multi-vector embeddings via a late-interaction mechanism over PaliGemma (Beyer et al., 2024), enabling efficient and accurate retrieval.

ColQwen2.5 (Faysse et al., 2025). An extension of ColPali (Faysse et al., 2025) using Qwen2-VL (Wang et al., 2024) as the backbone, carrying forward the late interaction retrieval paradigm over page image embeddings, capturing layout and textual context without OCR.

Jina-v4 (Günther et al., 2025). A multimodal embedding model combining visual and textual inputs with support for multi-vector (late interaction) embeddings, using adapters over a unified backbone to excel on visually rich document retrieval.

NemoRetriever (Xu et al., 2025). An LI retriever that combines vision-language embeddings with a ColEmbed design, enabling high performance on visual document retrieval with structured patch matching and efficient similarity.

Jina CLIP (Koukounas et al., 2024). A smaller scale vision-language model using CLIP embeddings, applied to document retrieval tasks; although not LI, it offers a lightweight multimodal baseline.

BGE Visualized M3 (Zhou et al., 2024). A vision-enhanced version of BGE M3 (Chen et al., 2024) that supports visual inputs and extends embedding models into multimodal domains.

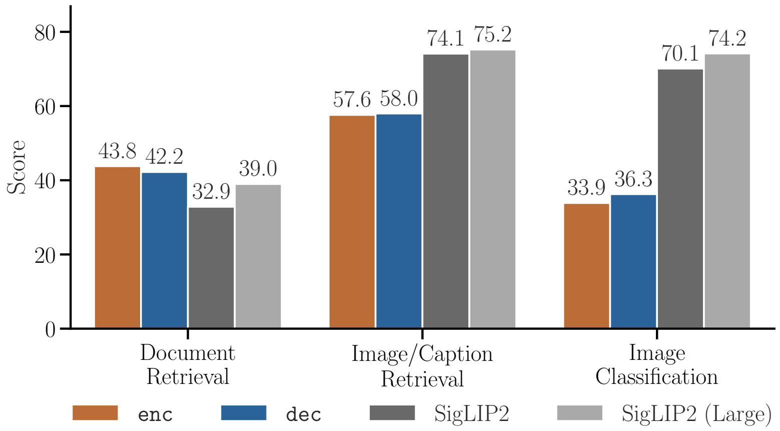


Figure 7: **Impact of Modality Alignment objective on downstream tasks.** Early Fusion of vision and text models boosts document retrieval tasks regardless of the LM objective, but degrades natural image and classification tasks w.r.t. the standalone *off-the-shelf* vision model SigLIP. Reported scores are aggregated MIEB scores (nDCG, Accuracy.)

SigLIP2-L-512/16 (Tschannen et al., 2025). A multilingual vision-language bi-encoder model, which combines image and text modalities to yield unified embeddings across languages. This configuration handles images of 512x512 pixels and create subpatches of 16x16 pixels.

ColFlor (Masry & Hoque, 2024). A lightweight OCR-free visual document retriever with only 174M parameters built over Florence-2 and DaViT, delivering strong performance near ColPali with much lower computational cost and much faster encoding.

C Additional Ablations

C.1 Performance Against Off-the-Shelf Dual Encoder

We study whether using *off-the-shelf* performances of the standalone vision tower are not outweighing the burden of adding language parameters and re-training through language modeling, as proposed in our work. Figure 7 shows the results of the various models on the tasks described in Section 2. Similarly to Section 3.1, we observe that the early fusion model trained with LM objective significantly outperform the standalone vision tower on document retrieval tasks (+10.9 nDCG@5). It even surpass the larger dual encoder (+4.8 nDCG@5) on these latest tasks. We note that the standalone vision tower largely outperform the early fusion models on the other natural images tasks, supporting for the use of the SigLIP model for these tasks as found in various general benchmarks (Xiao et al., 2025a).

C.2 Scaling Dynamics of Attention Masks

We study the different training dynamics of the different training objectives. We compare the `enc` (MLM) approach with a traditional `dec` (CLM) objective. Figure 8 presents the performance of the two training objectives across a diverse set of tasks. While starting `dec` offers an advantage in low-data regimes, `enc` seems to catches up. In document retrieval tasks, it eventually surpasses `dec` and scales better.

C.3 Bridging the Gap with Longer Contrastive Training

We study the impact of additional in-distribution training pairs on embedding tasks by scaling the contrastive training stage. Starting from the final checkpoint of our encoder-based ablation model,

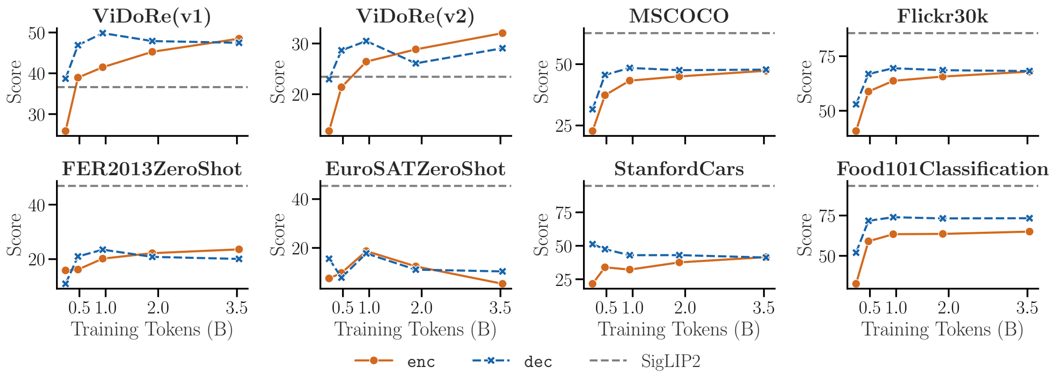


Figure 8: **Attention masks impact on modality alignment phase scaling.** The dashed line marks the vision tower baseline. The orange curve shows the model initialized from a decoder LM with a *CLM* objective, and the blue curve shows the model trained with an *MLM* objective from an encoder LM. CLM performs better in low-data regimes, but MLM scales more effectively, surpassing CLM in document retrieval, while captioning and classification remain below the CLIP baseline.

we double the contrastive dataset size at each step and train until convergence[14]. This setup tests whether scaling continues to improve performance. Figure 9 shows the scaling behavior. Performance improves overall with more in-distribution data. The vision-tower baseline is quickly surpassed on visual document benchmarks, and scaling narrows the gap on other tasks[15]. We note a plateau in captioning and classification, pointing to the need for more diverse data.

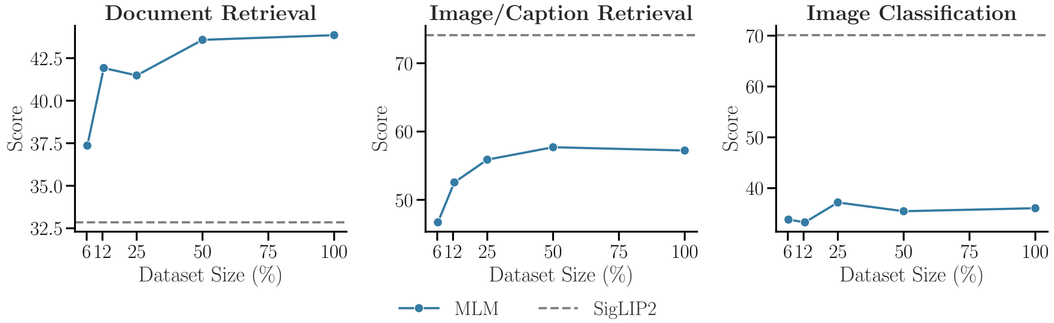


Figure 9: Contrastive training scaling. Each dot on the blue curve represents one fraction of the baseline contrastive training mix (ColPali + MSCOCO). Performance improves with more in-distribution data, surpassing the baseline on document benchmarks and narrowing the gap on image captioning. There is no clear improvement in image classification, highlighting the need for more diverse pairs.

C.3.1 Optimal Text-To-Image Ratio for Document Retrieval

Our findings in subsection 3.2 indicate that incorporating additional text-only pairs boosts document retrieval performance. While our initial experiment employed a 1:1 text-to-image ratio, we further investigate how varying this ratio impacts our broad set of tasks. We start from the best contrastive mix in Table 2, and vary the text-to-image ratio. As shown in Figure 10, increasing the number of text-only pairs *for a fixed amount of image pairs* consistently enhances retrieval performance. However, for natural image classification tasks, adding more text does not appear to provide benefits.

[14] To avoid overfitting, we set an early stopping on an eval set. We limit the number of step to one epoch on the full dataset.

[15] Note that the models probably won't fully recover baseline vision-tower performance. This highlights the need to choose models according to use case (e.g., lightweight CLIP-like models for image classification).

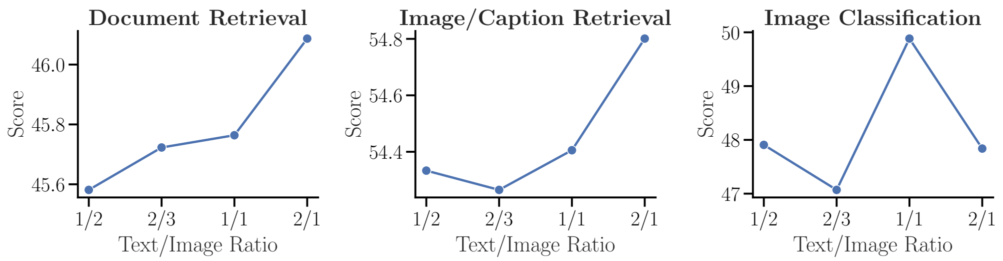


Figure 10: **Optimal text-to-image ratio in contrastive training mix.** Increasing the ratio in retrieval tasks consistently improves the performances.

C.4 Late Interaction for Non-Documental Retrieval

		Document Retrieval		Image/Caption Retrieval		
	Model Size	ViDoRe(v1)	ViDoRe(v2)	MSCOCO (T→I)	Flickr30k (T→I)	**Average**
CLIP Encoders						
siglip2-base-patch16-512	376M	36.6	23.4	66.2	86.9	53.3
siglip2-large-patch16-512	882M	43.8	27.0	67.1	88.9	56.7
clip-vit-base-patch16	151M	25.5	20.4	50.3	76.8	43.3
clip-vit-large-patch14	428M	38.0	28.6	52.7	79.3	49.6
VLM-based Encoders						
VLM2Vec-Full	4150M	49.8	36.5	59.5	81.8	56.9
e5-v	8360M	62.7	49.4	68.1	89.8	67.5
Early Fusion Encoders						
bge-visualized-base	196M	10.3	9.0	50.0	74.1	35.9
bge-visualized-m3	873M	12.4	10.2	39.6	69.0	32.8
***ModernVBERT*-embed**	252M	58.4	36.9	56.5	76.0	56.9
***ModernVBERT*-embed (multi-vector)**	252M	76.5	53.9	61.8	81.4	**68.4**

Table 8: **Generalist retrieval performances.** Late interaction benefits extend to non-documental retrieval tasks. Our multi-vector model increases its single-vector counterpart across all tasks, surpassing larger VLM-based retrievers.

We want to study if the multi-vector gains transfer to non-documental retrieval. To do so, we contrastively post-train our base model on our generalist post-training mix presented in Table 7. The late interaction generalist exhibits superior performance in retrieval setting, improving its single-vector performance by +20.2% (11.5 points), matching the performance of substantially larger VLM-based retrievers like E5-V (8.3B parameters, 67.5 points) and surpassing dual encoders like SigLIP (882M parameters, 56.7 points). This matches the capabilities observed in Section 3.1 for documental settings for models with native bidirectional attention, extending it to natural image tasks. This result extends the prevailing understanding from the document retrieval community, where the superiority of late-interaction is well-documented (Khattab & Zaharia (2020), Chaffin (2025), Faysse et al. (2025)). While this performance gap is widely accepted for document retrieval, its applicability to caption matching tasks has not really been addressed. Our findings provide strong evidence that the fine-grained matching capabilities of late-interaction models are a key driver of performance in this domain too.

C.4.1 Model Merging

Our contrastive learning stage provides direct performance trade-offs on different tasks. Following recent trends, we evaluate how model merging techniques allow to mitigate performance degradation on specific tasks, while maintaining the performance enabled by the contrastive training (Sung et al., 2023; Dziadzio et al., 2024; Li et al., 2024; Zhang et al., 2025b). We merge our ablation model after modality alignment with the checkpoint after the full contrastive learning with two methods: SLERP (Ilharco et al., 2022) and average merging (Shoemake, 1985). For SLERP, we compare three values for the λ coefficient (0.25, 0.5, 0.75). Figure 11 displays the the trends with the best method (SLERP, $\lambda = 0.75$). As we can see, the merged model mitigates the performance drop in Image/Caption Retrieval tasks, while maintaining significant gains on Image Classification tasks. However, merging strongly degrades performance on Document Retrieval, showing that benefits of merging embedding models are task-dependent.

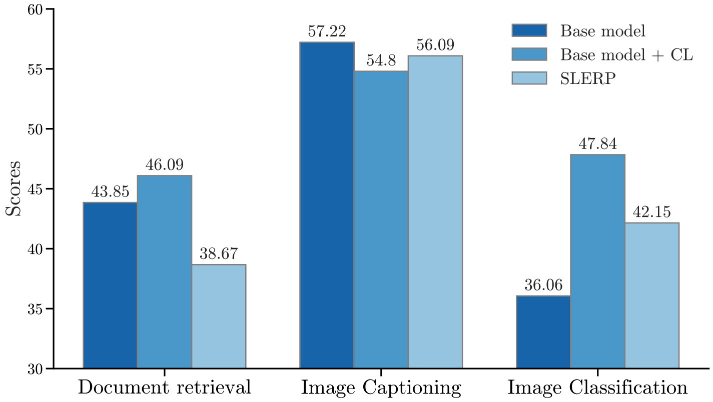


Figure 11: Merging model results across tasks. Benefits are task-dependent, with performance degradation w.r.t. both original models in Document Retrieval.

C.4.2 Curriculum For Document Retriever Contrastive Post-Training

We conduct an ablation study to determine the optimal contrastive training curriculum for specializing *ModernVBERT* in document retrieval. Specifically, we investigate whether a preliminary generalist contrastive training phase, intended to leverage a larger dataset, improves downstream performance. As shown in Table 9, our results demonstrate that this initial generalist phase is detrimental to final performance (-0.5%). The optimal strategy is to specialize the model on the target task directly after its initial Masked Language Modeling (MLM) alignment.

	ViDoRe(v1)	ViDoRe(v2)	**Average**
Document retrieval contrastive training starting checkpoint			
ModernVBERT-base	**81.2**	**56.0**	**68.6**
+ multi-vector generalist CL	80.7	55.4	68.1
+ single-vector generalist CL	80.6	54.0	67.3

Table 9: **Performance of *ModernVBERT* Document Specialisation Curriculums.** This table presents the performance of various contrastive training curriculums starting from *ModernVBERT*-base, on the ViDoRe(v1) and ViDoRe(v2) benchmarks. The generalist contrastive learning mix used in the last two models is detailed in Table 7. We see that a preliminary stage of generalist contrastive learning harms the final document retrieval performance, regardless of whether a multi-vector approach is used.

C.5 Text-Only Retrieval

The results in Table 10 detail the performance of *ColModernVBERT* and other baselines on the NanoBEIR text retrieval benchmark. It achieves an average NDCG@5 score competitive with single and multi vector models specialized for text, even without explicit optimization for this modality. This performance

Model	Params (M)	NDCG@5
Statistical		
BM25s	—	0.559
Single Vector		
Jina Embeddings v4	3577*	0.623
E5-large-v2	335	0.605
bge-m3 (Bi Encoder)	567	0.590
Qwen3-Embedding-0.6B	600	0.567
Multi Vector		
LightOn GTE-ModernColBERT v1	149	0.669
Jina ColBERT v2	137	0.642
bge-m3 (Late Interaction)	567	0.606
ColBERT v2	110	0.593
Colqwen2-v1.0	1580*	0.593
ColModernVBERT	150*	0.589
Colqwen2.5-v0.2	3145*	0.589

Table 10: Average NDCG@5 of *ColModernVBERT* on NanoBEIR, a text retrieval benchmark with multiple sub domains. *For multimodal models, we only consider parameters of the text encoder

is encouraging and indicates a promising direction for training a unified model for both text and image retrieval.

C.6 Model Latency

C.6.1 Image Resolution Tradeoffs

Figure 12 presents the pixel shuffling trade-off. Processing larger images creates more visual tokens, leading to very long sequences (around 17′500 tokens for a 2048x2048 px image with no pixel shuffling). Pixel shuffling allow to compress these sequence by concatenating the embeddings of spatially close patches. This diminishes the number of tokens for longer visual token embeddings. Table 11 presents the latency to process one image of various resolutions on one L4 GPU and CPU.

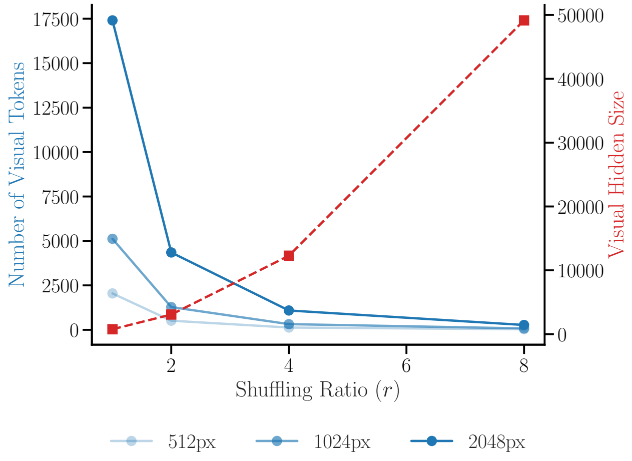


Figure 12: **Image processing parameters impact on visual tokens.** Here we assume a square image for simplicity. Scaling the image size introduces naturally more tokens, but having a large enough pixel shuffling ratio ($r \geq 4$) allows to counterbalance by concatenating spatially close patch representations.

	Num. Visual Tokens	CPU Latency (ms)	GPU Latency (ms)
512px	128	$287.2_{(\pm 7.8)}$	$43.6_{(\pm 1.4)}$
1024px	320	$1015.8_{(\pm 58.1)}$	$150.3_{(\pm 2.5)}$
2048px	1088	$2572.0_{(\pm 63.9)}$	$363.4_{(\pm 4.6)}$

Table 11: ***ModernVBERT* image processing latency**. Computing the average time to process a single image on GPU and CPU. The average is computed on 100 images. The values represent the mean latency in milliseconds, with the standard deviation included in parenthesis.

	Late Interaction	Model Size (B)	CPU Latency (ms)	GPU Latency (ms)	GPU Batching (ms)
≥ 1B Parameters					
MoCa-3B		3.75	$158_{(\pm 147)}$	$26_{(\pm 3)}$	4.54
VLM2Vec		4.15	$211_{(\pm 253)}$	$21_{(\pm 3)}$	2.82
GME-Qwen2-7B		8.29	$412_{(\pm 411)}$	$25_{(\pm 1)}$	9.07
E5-V		8.36	$434_{(\pm 379)}$	$22_{(\pm 2)}$	9.55
ColPali	✓	2.92	$175_{(\pm 113)}$	$14_{(\pm 1)}$	3.07
ColQwen2.5	✓	3.75	$158_{(\pm 147)}$	$26_{(\pm 2)}$	26
Jina-v4	✓	3.75	$158_{(\pm 147)}$	$26_{(\pm 2)}$	4.54
NemoRetriever-3B	✓	4.40	$155_{(\pm 118)}$	$20_{(\pm 2)}$	4.59
≤ 1B Parameters					
Jina CLIP		.22	$14_{(\pm 7)}$	$6_{(\pm 2)}$	.69
BGE Visualized M3		.87	$38_{(\pm 42)}$	$10_{(\pm 2)}$	.77
SigLIP2-L-512/16		.88	$25_{(\pm 8)}$	$6_{(\pm 1)}$	.10
ColFlor	✓	.17	$17_{(\pm 9)}$	$8_{(\pm 2)}$	.31
BiModernVBERT (ours)		.25	$20_{(\pm 11)}$	$14_{(\pm 2)}$	.20
***ColModernVBERT* (ours)**	✓	.25	$20_{(\pm 11)}$	$14_{(\pm 2)}$	.20

Table 12: **Text query encoding latency.** The latency is computed both on high-end CPUs (1TB RAM, 128 cores) and GPU (Nvidia H100, 80GB) (mean ± std). Since only 649 queries are used, standard deviations are not reported in GPU batching mode (batches of 512 queries by default), for which we report the inverse throughput (average latency per batch divided by the batch size).

C.6.2 Online Query Encoding Latency

We evaluate the query embedding speed of our model on GPU. We use a single Nvidia H100 with 80GB of VRAM. As for Section 4.2, latencies are computed in batch size 1 to simulate online situations, and are averaged over all NanoBEIR queries. Only the text parameters are loaded and run, to minimize memory usage. Parameters are cast to bfloat16 and Flash Attention 2 is used. The resulting speeds are often much faster than those obtained by running inference through each model's reference implementation. Results are shown in Table 12). Interestingly in this setup where memory is not a bottleneck, model depth seems to be a large performance driver, sometimes more the parameter count. We finally evaluate batched GPU throughput. We use batches of size 512 by default and iteratively half it when memory is insufficient. We observe that *ModernVBERT* based models are extremely fast and can process 5000 queries per second. In the table, the reported figures correspond to the inverted throughput (latency per batch divided by the number of queries per batch). These speed and throughput gains are made possible due to a combination of size, and efficient hardware-informed design as well as the support of flash attention and sequence packing other models of the size often lack (Warner et al., 2024b).

NatCap Annotation Prompt

You are an image annotator expert.

You will receive an image along with its classification label and the classification task scope, and your task is to provide contextualized metadata about it.

The output should be a JSON object with the following metadata fields:

- **caption**: A descriptive caption of the image accounting for its label. This should be a **unique** and concise sentence that describes the image in detail.
- **class_tags**: A list of tags that represents the image and can help identify the class. (e.g., for a car image with its model as a class, this could be some specific attribute of the car)
- **other_tags**: A list of tags that represents the image but can help identify the image among others of the same class. (e.g., for a car image with its model as a class, this could be its color or the background of the image)
- **is_image_class_explicit**: Boolean, could the class be inferred from the image alone? (e.g., the class is a country and you cannot necessarily infer it from the image alone, so this would be `false`)

Please ensure that the output is in valid JSON format.

Example:
You receive an image of what is clearly a car with its model as a class (here Audi TTS coupe 2012) for a car model classification task.
The output could be a JSON object like this:

```
{
  "caption": "A red Audi TTS coupe 2012 car parked on a sunny street
               in front of a sport shop.",
  "class_tags": ["sport coupe","four door coupe","17’’ alloy wheels"],
  "other_tags": ["sunny street","parked","red","sport shop"],
  "is_image_class_explicit": true
}
```

Classification scope: {task_info}
Image label: {label}
Answer: