Tevatron 2.0: Unified Document Retrieval Toolkit across Scale, Language, and Modality

Xueguang Ma*
x93ma@uwaterloo.ca
University of Waterloo
Waterloo, Canada

Luyu Gao*
luyug@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, United States

Shengyao Zhuang*
shengyao.zhuang@csiro.au
CSIRO
Queensland, Australia

Jiaqi Samantha Zhan
j35zhan@uwaterloo.ca
University of Waterloo
Waterloo, Canada

Jamie Callan
callan@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, United States

Jimmy Lin
jimmylin@uwaterloo.ca
University of Waterloo
Waterloo, Canada

Abstract

Recent advancements in large language models (LLMs) have driven interest in billion-scale retrieval models with strong generalization across retrieval tasks and languages. Additionally, progress in large vision-language models has created new opportunities for multimodal retrieval. In response, we have updated the Tevatron toolkit, introducing a unified pipeline that enables researchers to explore retriever models at different scales, across multiple languages, and with various modalities. This demo paper highlights the toolkit's key features, bridging academia and industry by supporting efficient training, inference, and evaluation of neural retrievers. We showcase a unified dense retriever achieving strong multilingual and multimodal effectiveness, and conduct a cross-modality zero-shot study to demonstrate its research potential. Alongside, we release OmniEmbed, to the best of our knowledge, the first embedding model that unifies text, image document, video, and audio retrieval, serving as a baseline for future research.

CCS Concepts

• **Information systems** → **Language models**; **Multimedia and multimodal retrieval**.

Keywords

Multimodal retrieval; Neural retrieval toolkit; Unified Retrieval Pipeline

ACM Reference Format:
Xueguang Ma, Luyu Gao, Shengyao Zhuang, Jiaqi Samantha Zhan, Jamie Callan, and Jimmy Lin. 2025. Tevatron 2.0: Unified Document Retrieval Toolkit across Scale, Language, and Modality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3726302.3730135

*The first three authors contributed equally.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGIR '25, Padua, Italy
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730135

1 Introduction

The Tevatron toolkit [5] is initially developed in 2021 during the BERT era, a period marked by the rise of neural dense retrievers, represented by models like Dense Passage Retrieval (DPR) [8], which became a focal point in information retrieval research. Tevatron is designed to provide an efficient and flexible framework, enabling researchers to easily modify or integrate new model architectures and training strategies for their specific needs. Since the initial release, dozens of previously published works have cited its usage. These works span a wide range of topics, including diverse text representation methods, multilingual and cross-lingual retrieval, and retrieval-augmented generation (RAG) pipelines.

Recent advancements in large language models (LLMs) have introduced new opportunities and challenges in the field of information retrieval. Billion-scale retrieval models have demonstrated superior generalizability across tasks and languages compared to their BERT-era counterparts [16, 25]. Additionally, the emergence of multimodal large language models [7, 23] has opened new possibilities for multimodal retrieval, leveraging large vision-language models to eliminate the need for document parsing and the cumbersome maintenance of disparate modalities [4, 14, 15]. These developments point toward a promising direction: the creation of unified retrieval systems capable of handling diverse tasks, languages, and modalities.

However, several challenges hinder the exploration of these advancements. First, training billion-scale retrieval models requires significantly more GPU memory, posing a barrier to researchers with limited computational resources. Second, while unified retrieval across modalities is an attractive goal, existing data formats designed for text retrieval are often ill-suited for organizing multimodal data. Third, as dense retrievers increasingly serve as industry solutions, there is a growing demand for more efficient inference methods to meet real-world deployment requirements.

In this work, we present updates to the Tevatron toolkit that address these challenges. Our contributions include a new data organization to support unified multimodal retriever training, as well as the integration of best practices from general LLM training and inference to enhance memory and computational efficiency. To facilitate further research, we have converted multiple open-source datasets into a unified format and hosted them in the Tevatron Hugging Face collection, which we will maintain and update continuously. We also demonstrate Tevatron's capability to support future research by training a unified dense retriever capable of multilingual and multimodal retrieval and providing analysis such as cross-modality generalization.

2 Tevatron-v2 Overview

2.1 Unified Data Management

Data management is a critical aspect often overlooked in individual research projects. As dense retrieval increasingly aims to handle generalizable tasks, it becomes essential to efficiently explore multiple sets of training data and their combinations.

Tevatron-v1 is specifically designed for text retrieval models, where training data for different text retrieval tasks could be unified into the following format for contrastive learning:

```
{
   "query_id": "<query id>",
   "query": "<query text>",
   "positive_passages": [
     {"docid": "<passage id>", "title": "<passage title>",
      "text": "<passage body>"}, ...],
   "negative_passages": [
     {"docid": "<passage id>", "title": "<passage title>",
      "text": "<passage body>"}, ...]
}
```

In this format, the raw document content for both positive and negative candidates is stored alongside each query instance. While this provides transparency for users to explore and debug the training data, it is not well-suited for multimodality data (where documents or queries can be images) or mixed-modality data (where queries and documents can contain both text and images).

As the number of queries expands, and each query is typically associated with multiple hard negative documents, the document lists across multiple queries often contain duplicate content. For instance, if each document in a corpus is associated with a query, and we aim to build training data with 20 hard negatives per query, the storage requirement for the training data becomes 20 times that of the corpus. This storage overhead becomes particularly prohibitive for image data, which is significantly more costly to store compared to text.

To address these challenges, we introduce a new unified data format for Tevatron-v2:

```
query:
{
   "query_id": "<query id>",
   "query_text": "<query text>",
   "query_image": "<query image>",
   "query_video": "<path to video>",
   "query_audio": "<path to audio>"
   "positive_document_ids": ["<document id>", ...],
   "negative_document_ids": ["<document id>", ...],
}
corpus:
{
   "docid": "<document id>",
   "document_text": "<document text>",
   "document_image": "<document image>",
   "document_video": "<document video path>",
   "document_audio": "<document audio path>",
}
```

Finetune	ZeRO	FlashAttn	GPU Memory	Training Time
Full-FT	ZeRO0	No	OOM	–
Full-FT	ZeRO3	No	63,146 MiB ×4	27 hours
Full-FT	ZeRO3	Yes	62,916 MiB ×4	26 hours
Full-FT	ZeRO3-off	Yes	21,764 MiB ×4	44 hours
LoRA	ZeRO0	No	28,554 MiB ×4	19 hours
LoRA	ZeRO0	Yes	28,172 MiB ×4	18 hours
LoRA	ZeRO3	Yes	25,778 MiB ×4	25 hours
LoRA	ZeRO3	Yes	69,324 MiB ×1	74 hours

Table 1: GPU memory and training time for fine-tuning Llama3.1-8B as a dense retriever using Tevatron-v2 with RepLlama [16] recipe on MS MARCO Passage.

In this new format, we decouple the training queries from the corpus. For document candidates associated with each query, we only store document IDs, dynamically loading the raw content within the dataloader. This approach significantly reduces storage requirements.

Additionally, this format efficiently organizes various data modality combinations without introducing complexity. It is compatible with text-only retrieval data, image-only retrieval data, and mixed-modality retrieval data. If a query or document contains more than one piece of text or image, we assume they can be merged into a single representation.

The dataloader and collator in Tevatron-v2 are designed to be modality-agnostic, enabling seamless training across different data combinations. For example, mixing text-only and image retrieval datasets allows training queries to encounter documents from varied modalities, helping to reduce modality bias. To support flexible data integration, Tevatron-v2 also introduces a Multi-Dataset class, allowing users to configure multiple training datasets.

2.2 GPU Memory Efficiency

The primary challenge in training billion-scale dense retrieval models lies in their substantial GPU memory requirements. Additionally, retriever training benefits from larger batch sizes and a large set of hard negative candidates for each update, further escalating these memory demands. Tevatron-v2 integrates several best practices from general large language model (LLM) training, including LoRA (Low-Rank Adaptation) [6], DeepSpeed ZeRO optimization [19], and FlashAttention [2], to address these challenges.

Table 1 demonstrates that each memory-efficient strategy significantly reduces GPU memory usage. For example, full fine-tuning without optimization results in out-of-memory errors. Enabling ZeRO stage 3 reduces memory usage to 63 GB per GPU on a 4-GPU machine, with a training time of 27 hours. With CPU offloading (zero3-off), memory usage drops to 22 GB per GPU, though training time increases to 44 hours due to offloading overhead.

For LoRA-based fine-tuning, the memory usage is much lower than full fine-tuning. Combining LoRA with ZeRO stage 3 and FlashAttention further reduces memory needs. Notably, training on a single GPU with LoRA, zero3, and FlashAttention is feasible, consuming 69,324 MiB of memory in charge of longer training time, but it makes things possible to explore on very limited compute resources. FlashAttention generally provides greater benefits in longer-context scenarios. However, in the MS MARCO Passage dataset, where documents are short (averaging 60 words), its impact on speed and memory cost is less pronounced.

2.3 Inference Efficiency

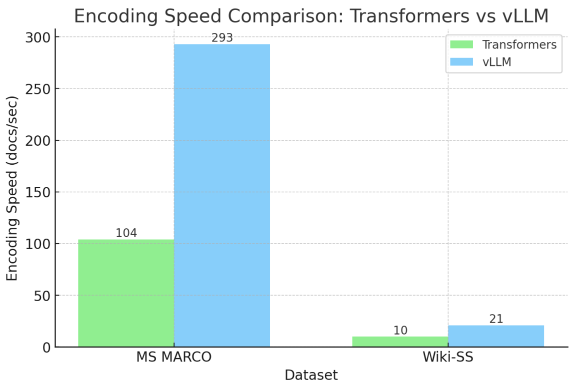

Figure 1: Encoding speed comparison between Transformers and vLLM implementation. For text, we used a retriever based on Llama3.1-8B retriever; for Wiki-SS, we used a multimodal retriever based on QWen2-VL-2B retriever with 784×784 image inputs.

vLLM [12] is an optimized serving and inference library for LLMs, designed to deliver high-throughput and low-latency inference. Integrating vLLM into Tevatron-v2 offers two key advantages. First, it significantly improves encoding speed, making the process more efficient. Second, it simplifies the deployment of models trained with the Tevatron codebase, enabling users to deploy them more easily and effectively. Additionally, vLLM majorly supports generation models, allowing retrievers to work closely with generation models in frameworks like retrieval augmented generation. This integration makes such collaborative workflows more seamless and efficient.

To quantitatively evaluate the benefits of vLLM, we compared the encoding speed of vLLM and the standard Transformers implementation in two settings: text encoding on the MS MARCO passage ranking corpus and document image encoding on the Wiki-SS (Wikipedia screenshot corpus). For text encoding, we used an 8B LLM retriever based on the Llama3.1-8B architecture. For Wiki-SS encoding, we employed a multimodal retriever based on the QWen2-VL-2B backbone, with images input at a resolution of 784x784. As shown in Figure 1, vLLM outperforms the Transformers implementation in both text and image encoding tasks by around 3 times of encoding speed.

Besides encoding latency, as the size of hidden states in large language models (LLMs) continues to grow, the increasing dimensionality of text embeddings raises concerns about storage costs for corpus indexes and search latency. To address this challenge, we integrate Matryoshka Representation Learning (MRL) [11] into Tevatron-v2. This approach enables the model to learn nested, scalable representations, where embeddings can be truncated to smaller dimensionalities without significant effectiveness degradation. During training, users can specify the target dimensionalities they wish to optimize for, allowing the model to adapt to diverse requirements. At inference time, this flexibility empowers users to dynamically adjust the text embedding dimensionality based on specific storage or computational constraints.

2.4 Comparison with Other Toolkits

The IR community has developed a number of outstanding toolkits that have greatly advanced research and experimentation. For example, Pyserini enables reproducible information retrieval experiments with robust integration of sparse and dense retrievers [13]. MTEB provides a comprehensive benchmark suite for evaluating embedding models across a wide range of retrieval and general embedding tasks [17]. Toolkits like sentence-transformers [20], BGE [1] and RAG-Retrieval [28] offer flexible and performant implementations of state-of-the-art retrievers and rerankers, making them widely adopted in both academia and industry. There are also specialized toolkits designed for LLM-based reranking [21, 22, 27].

Tevatron aims to offer an extensible framework that supports rapid prototyping and scalable training, with modular components for data processing, model training, and evaluation. It is designed with flexibility in mind, making it easy to incorporate new training paradigms, languages, and modalities. This enables researchers to explore new scenarios within a single, cohesive framework, paving the way for more generalizable and powerful retrieval systems.

3 Experiment: Unified Dense Retriever

To demonstrate the capabilities of our toolkit, we present the training of a dense retrieval model using a set of Tevatron self-contained data (available on Hugging Face) to develop a unified multimodal and multilingual dense retriever. We further explore in-modality zero-shot and cross-modality zero-shot settings to highlight the potential of Tevatron-v2 for future research.

3.1 Setup

3.1.1 Training data.

- BGE-Training Data [1]: A set of text retrieval training data, such as web search (MS MARCO Passage), Wikipedia-based QA and Fact verification (NQ, FEVER), scientific document retrieval (SCIDOCS), etc. 1.84 million in total derived from bge-full-training data.
- WikiSS [15]: This dataset, released as part of the DSE work, is based on natural question query sets and the Wikipedia screenshot corpus. It includes 29.3k training samples.
- PixMo-docs: PixMo-Docs [3] is originally a collection of synthetic question-answer pairs about computer-generated images (e.g., charts, tables, diagrams, and documents), we converted PixMo-Docs into retrieval training data through filtering and hard negative mining. We removed questions that were too specific and unsuitable for open-domain question answering (e.g., "What is the summary of the text?"). This was done by leveraging DSE as a retriever to search for each query. If the corresponding paired document did not appear in the top-100 retrieval results, the query was deemed unsuitable and removed. We identified challenging negative examples for each question by randomly sampling non-positive documents from the top-100 retrieval results. After processing, we obtained 1.75 million training samples.
- ColPali-training-data [4]: This dataset is the training data for the ColPali model, a multi-vector document screenshot retriever. It includes 127k query-image pairs. The original

Method	Base Model	BEIR	ViDoRe	MIRACL
BGE-M3 [1]	X-RoBERTa	50.0	66.1*	69.2
MistralE5 [25]	Mistral0.1-7B	59.0	-	62.2
DSE-QWen2 [15]	Qwen2vl-2B	-	85.8	-
GME-2B [30]	Qwen2vl-2B	55.4	87.8	-
Tevatron-WikiSS	Qwen2.5vl-3B	40.8	73.3	30.6
Tevatron-BGE	Qwen2.5vl-3B	57.0	76.4	67.5
Tevatron-VL	Qwen2.5vl-3B	54.3	85.3	66.5
Tevatron-Omni	Qwen2.5Omni	58.2	85.8	69.1

Table 2: Multimodal and multilingual retrieval results. *The result of BGE-M3 on ViDoRe is based on OCR.

release did not contain hard negatives, so we followed a similar approach as above to mine hard negative candidates.
- MSR-VTT [26] and AudioCaps [9] are video and audio retrieval datasets, respectively. We include them to support training the Omni variants of our model. See additional details below.

3.1.2 Hyperparameters. The Tevatron-VL model, based on Qwen2.5-VL-3B-Instruct [23], is trained on the combined text and visual datasets for one epoch. The Tevatron-Omni model, based on Qwen2.5-Omni-7B [7], additionally incorporates video and audio data. Training is performed on a machine with 8×NVIDIA H100 GPUs using a batch size of 128 queries. Each query is paired with one positive and three hard negative documents. We adopt LoRA for memory-efficient training.

3.1.3 Evaluation. To assess the model's effectiveness on retrieval tasks across languages and modalities, we evaluate it on the following widely used retrieval evaluation datasets:

- BEIR [24]: A diverse collection of English retrieval tasks. We evaluate on its 13 publicly available datasets, with effectiveness reported in nDCG@10.
- MIRACL [29]: A multilingual retrieval evaluation benchmark covering 18 languages, with effectiveness reported in nDCG@10.
- ViDoRe [4]: A multimodal evaluation benchmark that includes table, chart, and document screenshot retrieval tasks. Effectiveness is reported in nDCG@5.

Note that Tevatron-v2 also self-contains evaluation scripts for above evaluation.

3.1.4 Model Variants. To further investigate the effectiveness of in-modality and cross-modality zero-shot generalization, we trained two additional variants:

- Tevatron-WikiSS, which is trained exclusively on Wikipedia screenshot-based document image retrieval data. This variant is evaluated using the ViDoRe benchmark for in-modality zero-shot effectiveness.
- Tevatron-BGE, which is trained solely on text retrieval data to adapt the multimodal backbone as a retriever. This variant is evaluated using the ViDoRe benchmark for cross-modality zero-shot effectiveness.

Method	MSR-VTT	AudioCaps
CLIP [18]	31.2	-
CE [10]	-	23.1
Tevatron-Omni	51.3	34.0

Table 3: Recall@1 of video and audio retrieval tasks of Tevatron-Omni compared to baselines.

3.2 Results

Table 2 presents a comparison of retrieval effectiveness across various models, highlighting the effectiveness of our unified retriever model, Tevatron-VL and Tevatron-Omni, against other existing representative models. The results demonstrate that our unified retriever achieves competitive effectiveness across multiple dimensions of retrieval tasks, including English multi-task retrieval, multilingual retrieval, and multimodal retrieval. In addition, our Tevatron-Omni also effectively supports video and audio embedding as illustrated in Table 3. Existing retrieval models are typically optimized for one or two specific dimensions. For instance, BGE-M3 [1] and MistralE5 [25] are primarily focused on text retrieval tasks. In contrast, Tevatron is designed to easily handle all these optimizations across retrieval tasks, languages, and modalities. Besides, our training data and training pipeline are fully open-sourced.

Zero-Shot Generalization. An exciting finding from the Tevatron-BGE model variants, which trains a large vision-language model backbone using only text retrieval data, demonstrates strong cross-modal zero-shot effectiveness on ViDoRe. The effectiveness score is even higher than that of in-modality zero-shot when using WikiSS as training data. Since the BGE data encompasses a more diverse range of retrieval tasks, this suggests that the backbone model excels in aligning textual and visual inputs. Furthermore, it highlights that fine-tuning the model to learn relevancy is more critical than focusing on modality alignment. This implies that even in the absence of multimodal retrieval training data, training on diverse text-only retrieval tasks can potentially yield effective multimodal retrieval. This approach may prove more effective than relying on text-based retrieval combined with OCR, as seen in models like BGE-M3 on ViDoRe. This showcases the flexibility and research-friendly design of Tevatron-v2, empowering researchers to experiment with diverse training data and tasks while exploring new research questions.

4 Conclusion

In this paper, we introduce Tevatron-v2, a unified and efficient toolkit for advancing large-scale, multimodal, and multilingual retrieval research. By demonstrating the training of a unified dense retriever capable of handling both multilingual text and image document retrieval, Tevatron-v2 showcases its versatility and potential to drive future research in retrieval, making it a valuable tool for both academia and industry. Tevatron-v2 is fully open-sourced at https://github.com/texttron/tevatron.

Acknowledgments

We sincerely thank the open-source community, as well as the contributors and users who have supported the development. This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

[1] Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 2318–2335.

[2] Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Re. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.).

[3] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. *arXiv:2409.17146* (2024).

[4] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. ColPali: Efficient Document Retrieval with Vision Language Models. *arXiv:2407.01449* (2024).

[5] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Tevatron: An Efficient and Flexible Toolkit for Neural Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 3120–3124.

[6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

[7] Jinzheng He Hangrui Hu Ting He Shuai Bai Keqin Chen Jialin Wang Yang Fan Kai Dang Bin Zhang Xiong Wang Yunfei Chu Junyang Lin Jin Xu, Zhifang Guo. 2025. Qwen2.5-Omni Technical Report. *arXiv:2503.20215* (2025).

[8] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 6769–6781.

[9] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating Captions for Audios in The Wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 119–132. doi:10.18653/v1/N19-1011

[10] A. Sophia Koepke, Andreea-Maria Oncescu, João F. Henriques, Zeynep Akata, and Samuel Albanie. 2023. Audio Retrieval With Natural Language Queries: A Benchmark Study. *IEEE Transactions on Multimedia* 25 (2023), 2675–2685.

[11] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. 2022. Matryoshka Representation Learning. In *Advances in Neural Information Processing Systems*.

[12] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles* (Koblenz, Germany) *(SOSP '23)*. Association for Computing Machinery, New York, NY, USA, 611–626.

[13] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2356–2362. doi:10.1145/3404835.3463238

[14] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. Universal Multimodal Retrieval with Multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*.

[15] Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying Multimodal Retrieval via Document Screenshot Embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 6492–6505.

[16] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2421–2425.

[17] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 2014–2037.

[18] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. 2021. A Straightforward Framework for Video Retrieval Using CLIP. In *Pattern Recognition: 13th Mexican Conference, MCPR 2021, Mexico City, Mexico, June 23–26, 2021, Proceedings* (Mexico City, Mexico). Springer-Verlag, Berlin, Heidelberg, 3–12.

[19] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 3505–3506.

[20] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[21] Sahel Sharifymoghaddam, Ronak Pradeep, Andre Slavescu, Ryan Nguyen, Andrew Xu, Zijian Chen, Yilin Zhang, Yidi Chen, Jasper Xian, and Jimmy Lin. 2025. Rank-LLM: A Python Package for Reranking with LLMs. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pandova, Italy) *(SIGIR '25)*. Association for Computing Machinery, Pandova, Italy.

[22] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14918–14937.

[23] Qwen Team. 2025. Qwen2.5-VL.

[24] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[25] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving Text Embeddings with Large Language Models. *arXiv:2401.00368* (2024).

[26] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5288–5296.

[27] Crystina Zhang, Sebastian Hofstätter, Patrick Lewis, Raphael Tang, and Jimmy Lin. 2025. Rank-Without-GPT: Building GPT-Independent Listwise Rerankers on Open-Source Large Language Models. In *Advances in Information Retrieval*, Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto (Eds.). Springer Nature Switzerland, Cham, 233–247.

[28] Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. Jasper and Stella: distillation of SOTA embedding models. *arXiv:2412.19048* (2025).

[29] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics* 11 (09 2023), 1114–1131.

[30] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025. GME: Improving Universal Multimodal Retrieval by Multimodal LLMs. *arXiv:2412.16855* (2025).