

 jinaai / **jina-embeddings-v4** 

 like 458

Follow  Jina AI 1.6k

 Visual Document Retrieval

 Transformers

 Safetensors

ColPali

 sentence-transformers

 multilingual

feature-extraction

vidore

multimodal-embedding

multilingual-embedding

Text-to-Visual Document (T → VD) retrieval

sentence-similarity

mteb

vllm

custom_code

 arxiv:2506.18902

 Region: EU



 Deploy ▾

Use this model ▾

 **Model card**

 Files

 xet

 Community 81



The embedding model trained by [Jina AI](#).

Jina Embeddings v4: Universal Embeddings for Multimodal Multilingual Retrieval

[GGUF](#) | [Blog](#) | [Technical Report](#) | [API](#)

Intended Usage & Model Info

jina-embeddings-v4 is a universal embedding model for multimodal and multilingual retrieval. The model is specially designed for complex document retrieval, including visually rich documents with charts, tables, and illustrations.

Built on [Qwen/Qwen2.5-VL-3B-Instruct](#), jina-embeddings-v4 features:

- **Unified embeddings** for text, images, and visual documents, supporting both dense (single-vector) and late-interaction (multi-vector) retrieval.
- **Multilingual support** (30+ languages) and compatibility with a wide range of domains, including technical and visually complex documents.
- **Task-specific adapters** for retrieval, text matching, and code-related tasks, which can be selected at inference time.
- **Flexible embedding size:** dense embeddings are 2048 dimensions by default but can be truncated to as low as 128 with minimal performance loss.

Summary of features:

Feature	Jina Embeddings V4
Base Model	Qwen2.5-VL-3B-Instruct
Supported Tasks	retrieval, text-matching, code
Model DType	BF16
Max Sequence Length	32768
Single-Vector Dimension	2048
Multi-Vector Dimension	128
Matryoshka dimensions	128, 256, 512, 1024, 2048
Pooling Strategy	Mean pooling
Attention Mechanism	FlashAttention2

Training & Evaluation

Please refer to our [technical report of jina-embeddings-v4](#) for training details and benchmarks.

Usage

- Requirements
- via [Jina AI Embeddings API](#)
- via [transformers](#)
- via [sentence-transformers](#)
- via [vLLM](#)

Jina-VDR

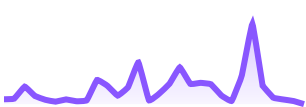
Alongside jina-embeddings-v4, we’re releasing [Jina VDR](#), a multilingual, multi-domain benchmark for visual document retrieval. The task collection can be viewed [here](#), and evaluation instructions can be found [here](#).

License

This model was initially released under cc-by-nc-4.0 due to an error. The correct license is the Qwen

Downloads last month

103,327



 **Safetensors** ⓘ

Model size

4B params

Tensor type

BF16

 Chat template

 Files info

 **Inference Providers** NEW

 Visual Document Retrieval

This model isn't deployed by any Inference Provider.

 17 Ask for provider support

 **Model tree for jinaai/jina-embed...**

Quantizations

3 models

 **Spaces using jinaai/jina-embe...** 6

 [jinaai/jina-embeddings-v4-retrieva...](#)

 [Navid-AI/The-Arabic-RAG-Leaderboard](#)

 [jadenhoch/jina-embeddings-v4](#)

 [Gopal2002/jina-embadding-v4-late-c...](#)

 [srinivasbilla/jina-embeddings-v4-retr...](#)

 [sanchitshaleen/chat-with-your-data](#)

 **Collection including jinaai/jina-e...**

jina-embeddings-v4

 Collection

Universal Em... • 10 items • Up... • △ 2

 **Paper for jinaai/jina-embedding...**

jina-embeddings-v4: Universal Emb...

 Paper • 2506.18902 • Publi... • △ 12

Research License, as this model is derived from Qwen-2.5-VL-3B which is governed by that license.

Contact

Join our [Discord community](#) and chat with other community members about ideas.

Citation

If you find jina-embeddings-v4 useful in your research, please cite the following paper:

```
@misc{günther2025jinaembeddingsv4universal,
  title={jina-embeddings-v4: Universal},
  author={Michael Günther and Saba Stuker},
  year={2025},
  eprint={2506.18902},
  archivePrefix={arXiv},
  primaryClass={cs.AI},
  url={https://arxiv.org/abs/2506.18902}
}
```