Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking

Mingxin Li*   Yanzhao Zhang*   Dingkun Long*   Keqin Chen
Sibo Song   Shuai Bai   Zhibo Yang   Pengjun Xie
An Yang   Dayiheng Liu   Jingren Zhou   Junyang Lin

Tongyi Lab, Alibaba Group

🤗 https://huggingface.co/collections/Qwen
https://modelscope.cn/organization/qwen
https://github.com/QwenLM/Qwen3-VL-Embedding

Abstract

In this report, we introduce the Qwen3-VL-Embedding and Qwen3-VL-Reranker model series, the latest extensions of the Qwen family built on the Qwen3-VL foundation model. Together, they provide an end-to-end pipeline for high-precision multimodal search by mapping diverse modalities, including text, images, document images, and video, into a unified representation space. The Qwen3-VL-Embedding model employs a multi-stage training paradigm, progressing from large-scale contrastive pre-training to reranking model distillation, to generate semantically rich high-dimensional vectors. It supports Matryoshka Representation Learning, enabling flexible embedding dimensions, and handles inputs up to 32k tokens. Complementing this, Qwen3-VL-Reranker performs fine-grained relevance estimation for query-document pairs using a cross-encoder architecture with cross-attention mechanisms. Both model series inherit the multilingual capabilities of Qwen3-VL, supporting more than 30 languages, and are released in **2B** and **8B** parameter sizes to accommodate diverse deployment requirements. Empirical evaluations demonstrate that the Qwen3-VL-Embedding series achieves state-of-the-art results across diverse multimodal embedding evaluation benchmarks. Specifically, Qwen3-VL-Embedding-8B attains an overall score of **77.8** on MMEB-V2, ranking first among all models (as of January 8, 2025). This report presents the architecture, training methodology, and practical capabilities of the series, demonstrating their effectiveness on various multimodal retrieval tasks, including image-text retrieval, visual question answering, and video-text matching.
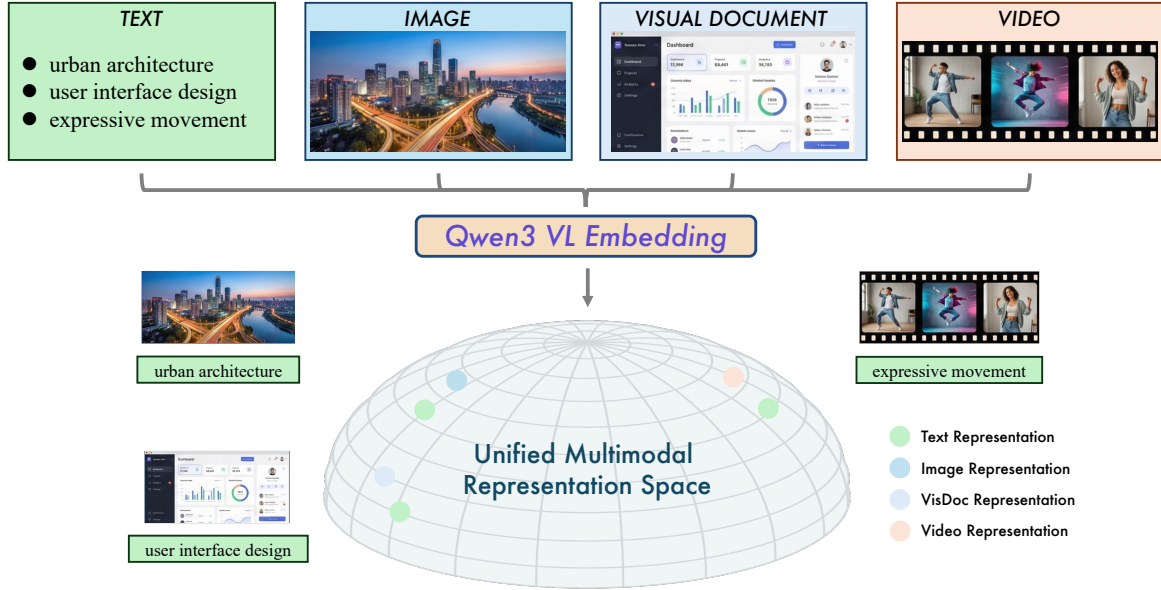
---
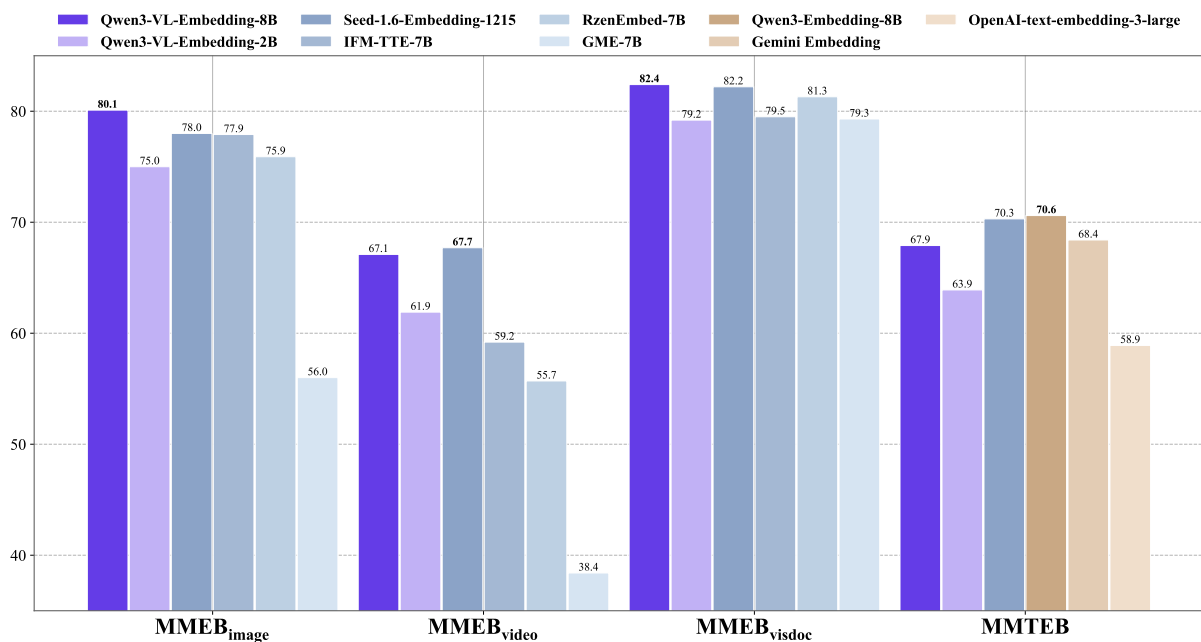
*Equal contribution.

Figure 1: Illustration of the Unified Multimodal Representation Space. Qwen3-VL-Embedding model series represent multi-source data (Text, Image, Visual Document, and Video) into a common manifold. By aligning semantic concepts across modalities (e.g., the text "urban architecture" and its corresponding image), the model achieves a holistic understanding of complex visual and textual information.

1 Introduction

The exponential growth of multimodal content on the internet has fundamentally transformed how information is created, shared, and consumed. Modern digital ecosystems are increasingly populated with diverse data modalities, including natural images, text documents, infographics, screenshots, and videos. This proliferation necessitates advanced retrieval systems capable of understanding and matching semantic concepts across different modalities, moving beyond traditional text-only search paradigms. Multimodal search, which aims to retrieve relevant content regardless of the query or document modality, has emerged as a critical capability for applications ranging from e-commerce product discovery to scientific literature exploration and social media navigation (Faysse et al., 2025; Fu et al., 2025).

Within contemporary multimodal retrieval architectures, embedding and reranking models constitute the two most critical modules. The field of multimodal representation learning has witnessed significant progress over the past decade (Manzoor et al., 2023; Mei et al., 2025). Among these pioneering works, CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021) has been particularly influential by demonstrating that large-scale contrastive learning on image-text pairs can produce powerful aligned representations. Its success has cemented the importance of learning shared embedding spaces where semantically similar content is positioned proximate in the representation space regardless of its modality.

As the development of foundation models accelerates, multimodal pre-trained vision-language models (VLMs) such as Qwen-VL (Wang et al., 2024b; Bai et al., 2025) and GPT-4o (Hurst et al., 2024) have achieved unprecedented success in multimodal comprehension. Building on these breakthroughs, the multimodal retrieval community has increasingly explored training unified multimodal embedding models based on VLMs. Notable efforts in this space include E5-V (Jiang et al., 2024), GME (Zhang et al., 2025b), BGE-VL (Zhou et al., 2025), and VLM2Vec (Meng et al., 2025; Jiang et al., 2025). Training unified multimodal representations based on VLMs offers several compelling advantages. First, VLMs possess inherent cross-modal alignment through their pre-training on large-scale image-text datasets. Second, they leverage sophisticated attention mechanisms to capture fine-grained interactions between visual and textual elements. Third, they provide a natural pathway to handling complex multimodal documents such as infographics and presentation slides where visual and textual information are deeply intertwined. Furthermore, VLM-based approaches can inherit the extensive multilingual and multi-domain knowledge encoded in foundation models, enabling more robust generalization across diverse retrieval scenarios.
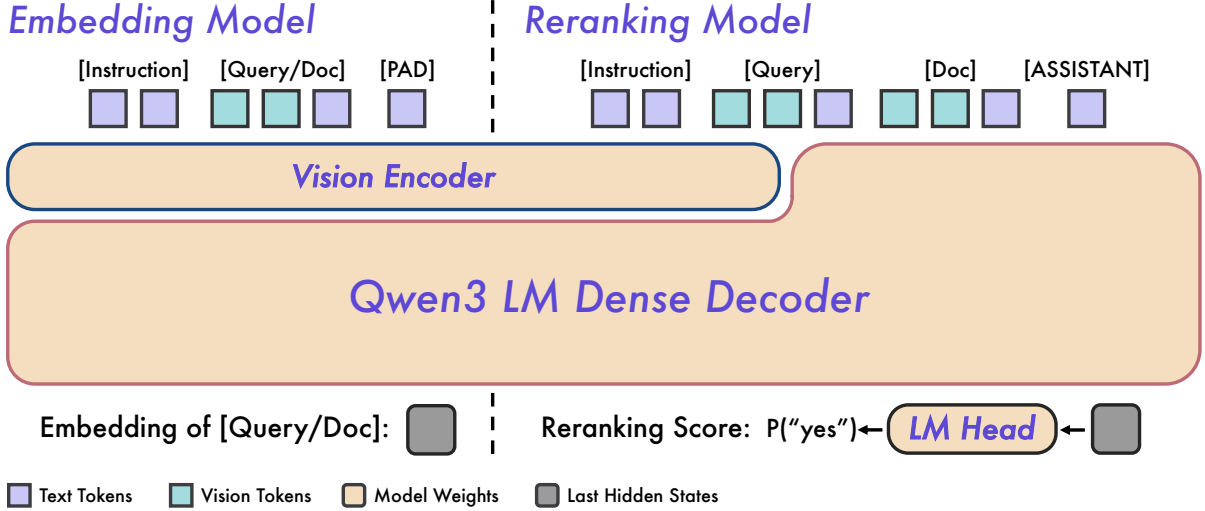
Figure 2: Overview of the Qwen3-VL-Embedding and Qwen3-VL-Reranker architecture.

In this work, we introduce the Qwen3-VL-Embedding and Qwen3-VL-Reranker model series, which are specifically designed for multimodal retrieval applications. Built upon the powerful Qwen3-VL (Bai et al., 2025) foundation model, these models bring together advanced vision-language understanding capabilities with specialized training methodologies tailored for retrieval tasks. The Qwen3-VL-Embedding series employs a sophisticated multi-stage training paradigm that progresses from contrastive pre-training on large-scale multimodal data to knowledge distillation from ranking models, ultimately producing semantically rich embeddings that capture nuanced relationships across modalities. These models support Matryoshka Representation Learning (Kusupati et al., 2022), allowing users to flexibly select embedding dimensions according to their storage and computational constraints without retraining. Additionally, we incorporate quantization-aware training strategies during the training process to ensure that the generated embeddings maintain robust performance after quantization. This capability significantly improves the storage efficiency and computational friendliness of downstream tasks. The models can process inputs containing up to 32,768 tokens, enabling comprehensive understanding of long documents and videos. Complementing the embedding models, the Qwen3-VL-Reranker series adopts a cross-encoder architecture that performs deep cross-attention between query and document representations, providing precise relevance scores for candidate retrieval results. Both model series inherit the impressive multilingual capabilities of the Qwen3-VL foundation model, supporting more than 30 languages with high proficiency, and are released in two sizes (2B, and 8B parameters) to accommodate diverse application scenarios.

We evaluate the Qwen3-VL-Embedding and Qwen3-VL-ReRanker model series across a comprehensive set of benchmarks spanning multiple tasks and domains. Experimental results demonstrate that our embedding and reranking models achieve state-of-the-art performance across multiple types of downstream tasks. For example, the flagship model Qwen3-VL-Embedding-8B attains a score of 77.8 on the MMEB-V2 benchmark (Meng et al., 2025), as evaluated in January 2026, surpassing all models currently on the leaderboard[1], including both open-source models and closed-source API services. Beyond multimodal evaluation, in pure text evaluation, the Qwen3-VL-Embedding-8B model achieves a mean task score of 67.9 on the MTEB Multilingual benchmark (Enevoldsen et al., 2025a), demonstrating highly competitive performance. Moreover, our reranking model delivers competitive results across a range of retrieval tasks. The Qwen3-VL-Reranker-2B model exceeds previously top-performing models in numerous retrieval tasks, while the larger Qwen3-VL-Reranker-8B model demonstrates even superior performance, improving ranking results by 4.1 points over the 2B model across multiple tasks. Furthermore, we include a constructive ablation study to elucidate the key factors contributing to the superior performance of the Qwen3-VL-Embedding series, providing insights into its effectiveness.

In the following sections, we present the architectural design of our model, elaborate on the training procedures, report comprehensive experimental results for both the embedding and reranking components, and conclude this technical report by synthesizing key findings and discussing promising avenues for future investigation.

---

[1] https://huggingface.co/spaces/TIGER-Lab/MMEB-Leaderboard

Table 1: Model specifications for the Qwen3-VL-Embedding and Qwen3-VL-Reranker. "Quantization Support" indicates the supported quantization formats for the embeddings. "MRL support" denotes whether the embedding model allows user-specified embedding dimensionalities. "Instruction-aware" indicates whether the models support task-specific customization of the input instruction.

Model Type	Size	Layers	Sequence Length	Embedding Dimension	Quantization Support	MRL Support	Instruction Aware
Qwen3-VL-Embedding	2B	28	32K	2048	Yes	Yes	Yes
	8B	36	32K	4096	Yes	Yes	Yes
Qwen3-VL-Reranker	2B	28	32K	-	-	-	Yes
	8B	36	32K	-	-	-	Yes

2 Model

Qwen3-VL-Embedding and Qwen3-VL-Reranker models are designed to make task-aware relevance judgments for multimodal instances. As shown in Figure 2, the embedding model follows a bi-encoder architecture to produce dense vector representations of instances and uses cosine similarity as the relevance measure. In contrast, the reranking model adopts a cross-encoder architecture to provide more fine-grained relevance estimates for each query–document pair.

Model Architecture   Both the embedding and reranking models are built on the Qwen3-VL backbone, using causal attention. After being trained on a large-scale collection of multimodal, multi-task relevance data, they retain the backbone's world knowledge, multimodal perception, and instruction-following capabilities, while additionally gaining the ability to estimate relevance. We train two model sizes—2B and 8B—and summarize their specifications in Table 1.

Embedding Method   The embedding model extracts task-aware dense vectors for multimodal inputs. The input format follows the Qwen3-VL context structure, where the instruction is passed as a system message, with the default instruction being "Represent the user's input." The multimodal instance to be represented is passed as a user message, and it can be in the form of text, images, videos, or any combination of these modalities. Finally, a "PAD" (<|endoftext|>) token is appended to the input, and the last hidden state corresponding to this token is used as the dense vector representation of the instance.

Input Template for Embedding

```
<|im_start|>system
{Instruction}
<|im_end|>
<|im_start|>user
{Instance}
<|im_end|><|endoftext|>
```

Reranking Method   The reranking model adopts a pointwise ranking approach, which evaluates the relevance between a pair of multimodal instances according to the relevance definition provided in the instruction. The input format follows the Qwen3-VL context structure, where both the relevance-defining instruction and the pair of multimodal instances to be evaluated are passed as user messages. These multimodal inputs can be text, images, videos, or any combination of these modalities. Finally, the relevance estimation for the pair is obtained by calculating the model's probability of predicting "yes" or "no" as the next output token.

Input Template for Reranking

```
<|im_start|>system
Judge whether the Document meets the requirements based on the Query and the Instruct
 ↪  provided. Note that the answer can only be "yes" or "no".
<|im_end|>
<|im_start|>user
<Instruct>: {Instruction}
```

```
<Query>: {Query}
<Document>: {Document}
<|im_end|>
<|im_start|>assistant
```

3 Data

To endow the model with universal representation capabilities across diverse modalities, tasks, and domains, we curated a massive-scale dataset. The distribution of different categories within the dataset is illustrated in Figure 3. However, both publicly available and proprietary in-house data exhibit significant imbalances and, in specific scenarios, notable scarcity across these dimensions. To address these challenges, we leverage data synthesis to construct a balanced training corpus that ensures robust coverage across all modalities, tasks, and domains.

3.1 Dataset Format

The complete dataset comprises multiple sub-datasets, denoted as $\mathcal{D} = \{D_i\}_{i=1}^M$. Each sub-dataset D_i is defined by a quadruple $D_i = (I_i, Q_i, C_i, R_i)$, structured as follows:

- **Instruction** (I_i): A textual description defining the specific relevance criteria and task objectives for the sub-dataset.
- **Queries** (Q_i): A collection of N_q query objects, $Q_i = \{q_j\}_{j=1}^{N_q}$. Each q_j can consist of text, images, videos, or any multimodal combination thereof.
- **Corpus** (C_i): A repository of N_d document objects, $C_i = \{d_j\}_{j=1}^{N_d}$. Similar to queries, each d_j may be a single modality or a multimodal composite of text, images, and videos.
- **Relevance Labels** (R_i): This component identifies the relationships between queries and documents, denoted as $R_i = \{(q_j, \{d_{j,k}^+\}_{k=1}^{n^+}, \{d_{j,k}^-\}_{k=1}^{n^-})\}_{j=1}^{N_q}$. For each query q_j, $\{d_{j,k}^+\}_{k=1}^{n^+} \subset C_i$ represents the set of relevant documents (positive documents), while $\{d_{j,k}^-\}_{k=1}^{n^-} \subset C_i$ represents the set of irrelevant documents (negative documents).

Representative dataset examples are presented in Appendix A.

3.2 Data Synthesis

Figure 3: Distribution of different categories in the training data.
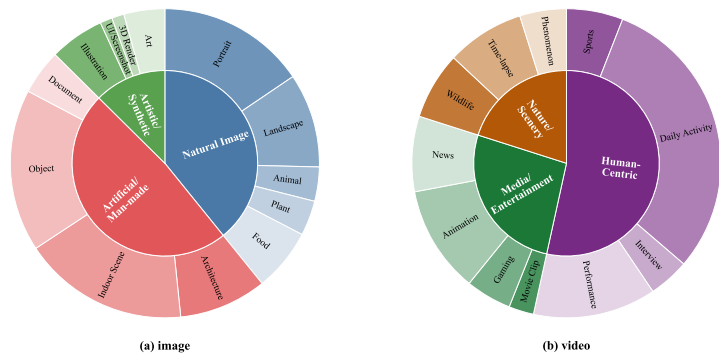

(a) image (b) video

Figure 4: Data distribution of the seed pool for data synthesis.

We employ data synthesis to construct various sub-datasets D_i. Specifically, we extend the methodology introduced in Qwen3 Embedding (Zhang et al., 2025c) to multimodal scenarios.

Seed Pool Construction Since the diversity of synthesized data depends on the underlying seed pool, we first aggregate an extensive collection of high-quality and diverse raw image and video datasets. To establish a high-quality foundation, we first apply coarse-grained quality filtering to prune assets with low resolutions or irregular aspect ratios. This is followed by structural refinement, specifically employing scene cut detection and removing static or corrupted segments, to preserve the integrity of temporal dynamics in video data. Subsequently, we leverage Qwen3-VL-32B (Bai et al., 2025) to generate fine-grained categorical labels for the remaining assets. To ensure cross-modal alignment, we implement a rigorous filtering mechanism that excludes samples with low-confidence annotations or poor visual-text correspondence, as measured by similarity scores from the GME (Zhang et al., 2025b) embedding model. Finally, we perform category-wise rebalancing on the refined dataset to construct the final seed pool. The resulting category distribution is illustrated in Figure 4.

Based on the seed pool, we leverage Qwen3-VL-32B (Bai et al., 2025) to perform multimodal and multi-task annotation.

Image Tasks Annotation We synthesize image datasets across three primary task paradigms:

1. **Image Classification**: The query q comprises an image and a classification instruction, while the document d is the specific category label. We synthesize datasets for a wide range of classification tasks, including object recognition, scene parsing, landmark identification, and action recognition. For each sample, the model designates a specific task type and annotates the image with its ground-truth category along with a semantically confusing negative label.

2. **Image Question Answering**: The query q consists of an image and a grounded question, and the document d is the corresponding answer. We generate diverse QA pairs covering factoid identification, visual reasoning, OCR-based data extraction, and domain-specific knowledge inquiry. Following a prescribed task orientation, the model formulates a question based on the visual content, providing a ground-truth response and a plausible but deceptive distractor.

3. **Image Retrieval**: The query q is a search text, and the document d is the candidate image. We synthesize retrieval queries across a hierarchy of semantic depths, spanning direct visual descriptions, abstract narrative scenarios, compositional logical constraints, and knowledge-centric textual localization. The model assigns a specific retrieval intent and generates a corresponding search query that captures either the salient visual features or the embedded textual logic within the image.

Video Tasks Annotation We synthesize video datasets across four primary task paradigms:

1. **Video Classification**: The query q combines a video with a classification task, and the document d is the resulting category. We synthesize datasets for diverse classification tasks, including activity recognition, scene parsing, event categorization, and sentiment/intent analysis. For each sample, the model identifies its category and generates a semantically related negative label.

2. **Video Question Answering**: The query q includes a video and a question, while the document d is the answer. We generate diverse QA pairs spanning factual identification, temporal grounding, thematic reasoning, and cinematic analysis. Guided by a specified task type, the model formulates a question and provides a correct response and a deceptive distractor.

3. **Video Retrieval**: The query q is a textual description, and the document d is the video. We synthesize retrieval queries across a spectrum of semantic granularities, ranging from entity and action-centric searches to temporal-event descriptions, thematic/emotional discovery, and instructional tutorial localization. The model produces a search query that captures the primary events and thematic content of the video.

4. **Moment Retrieval**: The query q is a textual query (optionally including a keyframe), and the document d is a specific video segment. The moment retrieval task aims at fine-grained temporal grounding. The model identifies a specific target—such as an action, object, or character—and localizes a relevant temporal segment. Simultaneously, it identifies an irrelevant segment with a clear temporal gap to serve as a negative contrast.

Prior to synthesizing task-specific annotations, we require the model to generate a descriptive caption for each image or video to provide necessary context. This two-step approach ensures higher quality and consistency in the subsequent annotation generation. Selected prompt examples for the synthesis of specific tasks are provided in Appendix B.
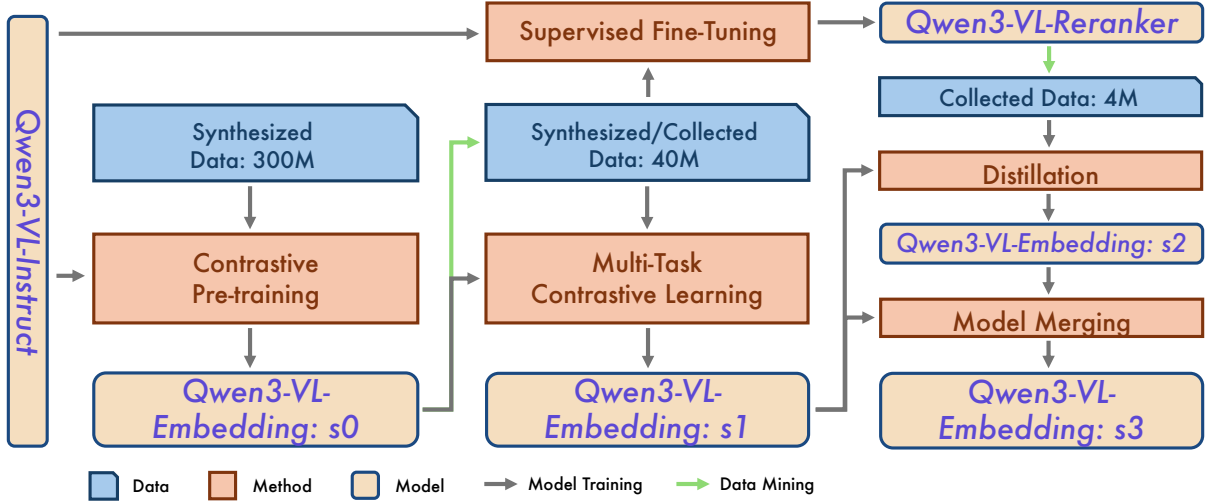
Figure 5: The multi-stage training pipeline of Qwen3-VL-Embedding and Qwen3-VL-Reranker.

3.3 Positive Refinement and Hard Negative Mining

Hard negative samples play a crucial role in contrastive representation learning (Robinson et al., 2021). To enhance the quality of positive pairs and identify effective hard negatives, we implement an automated two-stage mining pipeline: Recall and Relevance Filtering.

Recall For each sub-dataset D_i, we use an embedding model to extract representations for all queries $q_j \in Q_i$ and documents $d_k \in C_i$. For each query q_j, we retrieve the top-K most relevant candidates $\{d_k\}_{k=1}^{K}$ based on cosine similarity, denoted as relevance scores $S = \{s_{j,k}\}_{k=1}^{K}$.

Relevance Filtering Finally, we refine the relevance labels R_i based on the relevance scores S to eliminate noise:

- **Positive Refinement**: We retain q_j only if at least one positive document $d^+ \in \{d_k\}_{k=1}^{K}$ achieves a score $s > t^+$, where t^+ is a hyperparameter acting as the score threshold. If no such candidate exists, the query q_j is discarded.
- **Hard Negative Selection**: For a valid query q_j, we compute the average score of its refined positive samples, $\bar{s}^+$. Any non-positive document $d \in \{d_k\}_{k=1}^{K}$ is selected as a hard negative only if its score satisfies $s < \bar{s}^+ + \delta^-$, where δ^- is a small safety margin to prevent the inclusion of "false negatives".

4 Training Strategy

To train our Qwen3-VL-Embedding and Qwen3-VL-Reranker, we employ a multi-stage training pipeline, as shown in Figure 5. This approach is designed to mitigate the data imbalance between abundant weakly-supervised data and scarce high-quality samples (Wang et al., 2022; Li et al., 2023; Chen et al., 2024; Zhang et al., 2025c). The model is first pre-trained on vast amounts of weakly supervised, noisy data to establish a baseline for relevance understanding and to boost generalization. We then perform fine-tuning on high-quality, task-specific datasets to steer the model toward more precise relevance scoring and fine-grained interaction. In addition to the aforementioned reasons, another objective of the multi-stage training strategy is to bootstrap both data quality and model performance. As the training progresses through successive stages, the model's capabilities are continuously enhanced. This improvement, in turn, facilitates more effective data mining, thereby refining the quality of the training data. This iterative cycle ultimately leads to a substantial boost in the model's overall performance.

4.1 Multi-stage Training

We implement a three-stage training strategy as follows:

Stage 1: Contrastive Pre-training   In the pre-training stage, we first perform contrastive learning on the embedding model using large-scale, multimodal, and multi-task synthetic data. The synthetic data utilized in this stage is mined using the methodology described in Section 3.3, utilizing an existing open-source model (Zhang et al., 2025b) as the embedding model.

The optimization objective employed during training is defined in Equation 1. Upon completion of this stage, we obtain the initial model version, Qwen3-VL-Embedding: s0.

Stage 2: Multi-Task Contrastive Learning and Supervised Fine-Tuning   In this stage, we primarily utilize a combination of curated public datasets and proprietary in-house data, augmented with sampled synthetic data to address the task imbalance inherent in existing datasets. Benefiting from the improved multi-task performance of Qwen3-VL-Embedding: s0, we employ this model to perform data mining, thereby ensuring high data quality across various tasks. We then train our embedding model using multi-task contrastive learning, implementing tailored contrastive objectives for different task types (see Section 5.1 for details). This results in Qwen3-VL-Embedding: s1.

Simultaneously, we train a new reranking model, Qwen3-VL-Reranker, by training on the retrieval-specific subset of the newly mined data, using Equation 4 as the optimization objective. This subset encompasses diverse tasks, including image retrieval, video retrieval, moment retrieval, and visual document retrieval. The resulting model demonstrates superior performance across these retrieval-centric tasks.

Stage 3: Distillation and Model Merging   In this final stage, we further enhance the embedding model by distilling the relevance discrimination expertise from the previously trained Qwen3-VL-Reranker. To achieve this, we curate a compact sub-dataset from both public and proprietary sources, ensuring a balanced distribution across multiple retrieval categories. We then employ Qwen3-VL-Reranker to generate fine-grained relevance scores for this subset, which serves as the supervision signal for training the embedding model under the objective defined in Equation 3. This distillation process yields Qwen3-VL-Embedding: s2.

While Qwen3-VL-Embedding: s2 exhibits significant gains in retrieval-centric tasks, it suffers a slight performance degradation in classification and QA tasks. To address this, we merge Qwen3-VL-Embedding: s2 with Qwen3-VL-Embedding: s1 using the methodology proposed by Li et al. (2024). This process results in our final model, Qwen3-VL-Embedding: s3, which achieves optimal and balanced performance across all evaluated tasks.

4.2   Implementation

We employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) for model training, with the model parameters initialized from Qwen3-VL-Instruct. This approach offers several key advantages: 1) reduced memory footprint, allowing for larger effective batch sizes; 2) enhanced generalization performance; and 3) significantly more efficient hyperparameter search for model merging (Li et al., 2024). Additionally, we adopt dynamic resolution and frame rates. For the image modality, we preserve the original aspect ratio while capping the maximum token consumption at 1,280 (approximately 1.3×10^6 pixels). For video, we first sample at 1 FPS with a maximum of 64 frames. For each frame, the aspect ratio is maintained, and the total token budget for all frames is constrained to 4,500 (approximately 9.2×10^6 pixels).

5   Training Objective

This section outlines the training objectives for the Qwen3-VL-Embedding and Qwen3-VL-Reranker. For Qwen3-VL Embedding model, We extend the loss function from the Qwen3 Embedding model (Zhang et al., 2025c) to handle a wider variety of data types. We also integrate two key techniques: Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) to produce variable-dimension embeddings, and Quantization-Aware Training (QAT) (Esser et al., 2020) to support multiple numerical precisions. Together, these methods reduce storage and compute costs, improving inference efficiency. The Qwen3-VL-Reranker adopts the same objective function as Qwen3 Reranker (Zhang et al., 2025c). The specific loss functions for each model are detailed below.

5.1   Loss Functions for the Embedding Model

The training of Qwen3-VL-Embedding involves diverse data types across multiple stages. To accommodate this, we employ distinct loss function tailored to the specific characteristics of each data category.

Loss for Retrieval Data This category includes data from various multimodal and cross-modal retrieval tasks, such as Text-to-Text (T2T), Text-to-Image (T2I), and Image+Text-to-Image+Text (IT2IT) retrieval. In Stage 1, we use the same InfoNCE loss (Oord et al., 2018) formulation as in the Qwen3-Embedding:

$$\mathcal{L}_{\text{retrieval}} = -\frac{1}{N} \sum_{i}^{N} \log \frac{e^{(s(q_i, d_i^+)/\tau)}}{Z_i}, \tag{1}$$

where $s(\cdot, \cdot)$ is a similarity function (we use cosine similarity), τ is a temperature parameter, and Z_i aggregates scores from the positive pair and various types of negative pairs:

$$Z_i = e^{(s(q_i, d_i^+)/\tau)} + \sum_{k}^{K} m_{ik} e^{(s(q_i, d_{i,k}^-)/\tau)} + \sum_{j \neq i} m_{ij} e^{(s(q_i, q_j)/\tau)} + \sum_{j \neq i} m_{ij} e^{(s(d_i^+, d_j)/\tau)} + \sum_{j \neq i} m_{ij} e^{(s(q_i, d_j)/\tau)}$$

corresponding to similarities with (1) the positive document d_i^+, (2) K hard negatives $\{d_{i,k}^-\}_{k=1}^K$, (3) other in-batch queries $\{q_j\}_{j \neq i}$, (4) other in-batch documents $\{d_j\}_{j \neq i}$ contrasted with d_i^+, and (5) other in-batch documents $\{d_j\}_{j \neq i}$ contrasted with q_i. m_{ij} is a masking factor to mitigate the impact of false negatives:

$$m_{ij} = \begin{cases} 0, & \text{if } s_{ij} > s(q_i, d_i^+) + 0.1 \text{ or } d_j = d_i^+, \\ 1, & \text{otherwise}, \end{cases}$$

where s_{ij} denotes the corresponding similarity score (e.g., $s(q_i, d_j)$ or $s(q_i, q_j)$).

In Stage 2, we further modify the objective by removing the query–query and document–document terms from Z_i. Empirically, this adjustment yields better performance on high-quality multimodal retrieval data.

Loss for Classification Data For text or image classification tasks, we likewise formulate training as contrastive learning. Specifically, the instance to be classified is treated as a query q, and its class label is treated as the corresponding document d^+. In contrast to retrieval, negative samples are restricted to explicitly incorrect labels for the same query, while other labels in the batch are ignored to avoid introducing false negatives.

Semantic Textual Similarity (STS) Data STS datasets are symmetric and thus do not admit a natural query–document asymmetry. Moreover, supervision is typically provided as real-valued similarity scores. To exploit this fine-grained signal, we optimize the model with the CoSent loss (Huang et al., 2024), which encourages cosine similarities between paired embeddings to preserve the ordering induced by ground-truth similarity scores:

$$\mathcal{L}_{\text{sts}} = \log \left(1 + \sum_{\hat{s}(q_i, d_j) > \hat{s}(q_m, d_n)} \exp \left(\frac{\cos(q_m, d_n) - \cos(q_i, d_j)}{\tau} \right) \right), \tag{2}$$

where $\hat{s}(q_i, d_j)$ denotes the ground-truth score for the pair (q_i, d_j).

Distillation Data In the final training stage, we further improve the embedding model via knowledge distillation. We sample a high-quality subset from the union of all training data and use a strong reranker to provide supervision. Concretely, for each query q, we pre-compute (offline) reranker relevance logits for its positive document and k negatives. During training, we compute embedding-based scores online using cosine similarity and minimize a distribution-matching objective (cross-entropy) to align the embedding model's score distribution with that of the reranker:

$$\mathcal{L}_{\text{distill}} = -\sum_{i=1}^{k+1} P_{\text{reranker}}(d_i \mid q) \log P_{\text{embedding}}(d_i \mid q), \tag{3}$$

where $P(d_i \mid q)$ is the softmax distribution over the $(k+1)$ candidate documents (one positive and k negatives) for query q.

5.1.1 Additional Techniques for Efficient Inference

In practical retrieval systems, index construction requires storing a large number of embeddings offline. To reduce storage overhead and improve retrieval efficiency, we incorporate the following auxiliary training objectives.

Matryoshka Representation Learning (MRL)   When optimizing the objectives described above, we compute each loss not only on the full-dimensional embedding, but also on truncated lower-dimensional prefixes of the same representation (Kusupati et al., 2022). Empirically, training over a sufficiently dense set of MRL dimensions yields strong generalization, enabling competitive performance at intermediate dimensions that are not explicitly included during training.

Quantization-Aware Training (QAT)   Storing embeddings with lower numerical precision (int8 or binary) can further reduce both storage and compute overhead. To preserve embedding quality under low-precision representations, we adopt a quantization-aware training (QAT) strategy. Concretely, during training we compute the optimization objective using both full-precision embeddings and their low-precision (quantized) counterparts, so that the model learns to produce embeddings that are robust to quantization. This allows the learned representations to better adapt to low-bit embedding formats, mitigating the performance degradation that may otherwise occur at deployment time. We instantiate QAT with Learned Step Size Quantization (LSQ) (Esser et al., 2020). LSQ treats the quantization scale (step size) as a learnable parameter and optimizes it jointly with the model weights via backpropagation. In addition, it uses a Straight-Through Estimator (STE) (Bengio et al., 2013) to propagate gradients through the non-differentiable rounding operation, enabling end-to-end training under simulated quantization.

5.2   Loss Function for the Reranking Model

We frame reranking as a binary classification problem: given a query–document pair, the model predicts either a special yes token (relevant) or no token (irrelevant).

$$\mathcal{L}_{\text{reranking}} = -\log p(l|I, q, d), \tag{4}$$

where $p(\cdot|*)$ denotes the probability assigned by the VLM. The label l is "yes" for positive pairs and "no" for negatives. This loss function encourages the model to assign higher probabilities to correct labels, thereby improving the ranking performance (Dai et al., 2025).

During inference, the final relevance score is computed by applying the sigmoid function to the difference between the logits of the 'yes' and 'no' tokens:

$$s = \text{sigmoid}(\text{logit}(\text{yes}) - \text{logit}(\text{no})). \tag{5}$$

6   Evaluation

6.1   Multimodal Benchmarks

To evaluate the overall performance of Qwen3-VL-Embedding in multimodal and multi-task representation learning, we report its results on the MMEB-v2 benchmark (Meng et al., 2025). MMEB-v2 provides a comprehensive assessment spanning three primary domains—Image, Video, and Visual Document—comprising nine task categories and 78 datasets in total. We compared our model against several prominent open-source and proprietary baselines. During evaluation, the context length is constrained to 16,384 tokens. For image-based tasks, the maximum token consumption is set at 1,800, while for video-based tasks, we cap the total tokens at 15,000 and the frame count at 64. As summarized in Table 2, the results demonstrate that our model achieves state-of-the-art (SOTA) average performance and exhibits exceptional proficiency across all three domains. Specifically, Qwen3-VL-Embedding-8B achieves an average score of 77.8 on MMEB-v2, representing a 6.7% improvement over the previous best open-source model.

6.2   Visual Document Benchmarks

In addition to the evaluation datasets in MMEB-V2, we conducted further tests on the latest JinaVDR (Günther et al., 2025) and Vidore-v3 [3] benchmarks for visual document retrieval tasks. We compared our models with current state-of-the-art ColPali-style models, with the results illustrated in Table 3. As shown, our embedding model achieves performance comparable to ColPali-style models that require significantly higher computational costs. Furthermore, our reranker model substantially outperforms ColPali models of a similar parameter size.

---

[2]https://huggingface.co/datasets/VLM2Vec/MMLongBench-page-fixed,
https://huggingface.co/datasets/VLM2Vec/ViDoSeek-page-fixed
[3]https://huggingface.co/collections/vidore/vidore-benchmark-v3

Table 2: Results on the MMEB-V2 benchmark (Meng et al., 2025). CLS: classification, QA: question answering, RET: retrieval, GD: grounding, MRET: moment retrieval, VDR: ViDoRe, VR: VisRAG, OOD: out-of-distribution. [†]: link to the model's homepage. All models except IFM-TTE have been re-evaluated on the updated VisDoc OOD[2] split.

Model	Size	Image					Video					VisDoc					All
		CLS	QA	RET	GD	Overall	CLS	QA	RET	MRET	Overall	VDRv1	VDRv2	VR	OOD	Overall	
# of Datasets →		10	10	12	4	36	5	5	5	3	18	10	4	6	4	24	78
Open-Source Models																	
VLM2Vec (Jiang et al., 2025)	2B	58.7	49.3	65	72.9	59.7	33.4	30.5	20.6	30.7	28.6	49.8	13.5	51.8	48.2	44	47.7
VLM2Vec-V2 (Meng et al., 2025)	2B	62.9	56.3	69.5	77.3	64.9	39.3	34.3	28.8	36.8	34.6	75.5	44.9	79.4	62.2	69.2	59.2
GME (Zhang et al., 2025b)	2B	54.4	29.9	66.9	55.5	51.9	34.9	42.0	25.6	31.1	33.6	86.1	54.0	82.5	67.5	76.8	55.3
Ops-MM-embedding-v1[†]	2B	68.1	65.1	69.2	80.9	69.0	53.6	55.6	41.8	33.7	47.6	76.4	53.2	77.6	64.2	70.8	64.6
RzenEmbed (Jian et al., 2025)	2B	68.5	66.3	74.5	90.3	72.3	50.4	49.7	46.6	38.9	47.3	87.1	55.1	87.2	43.4	74.5	67.2
VLM2Vec (Jiang et al., 2025)	8B	62.7	56.9	69.4	82.2	65.5	39.1	30.0	29.0	38.9	33.7	56.9	9.4	59.1	54.0	49.1	53.1
GME (Zhang et al., 2025b)	8B	57.7	34.7	71.2	59.3	56.0	37.4	50.4	28.4	37.0	38.4	89.4	55.6	85.0	68.3	79.3	59.1
Ops-MM-embedding-v1[†]	8B	69.7	69.6	73.1	87.2	72.7	59.7	62.2	45.7	43.2	53.8	80.1	59.6	79.3	67.8	74.4	68.9
RzenEmbed (Jian et al., 2025)	8B	70.6	71.7	78.5	92.1	75.9	58.8	63.5	51.0	45.5	55.7	89.7	60.7	88.7	69.9	81.3	72.9
Closed-Source Models																	
IFM-TTE[†]	8B	76.7	78.5	74.6	89.3	77.9	60.5	67.9	51.7	54.9	59.2	85.2	71.5	92.7	53.3	79.5	74.1
Seed-1.6-embedding-0615[†]	-	76.1	74.0	77.9	91.3	77.8	55.0	60.8	51.3	53.5	55.3	85.3	56.6	84.7	68.6	77.7	72.6
Seed-1.6-embedding-1215[†]	-	75.0	74.9	79.3	89.0	78.0	85.2	66.7	59.1	54.8	67.7	90.0	60.3	90.0	70.7	82.2	76.9
Qwen3 VL Embedding Models																	
Qwen3-VL-Embedding-2B	2B	70.3	74.3	74.8	88.5	75.0	71.9	64.9	53.9	53.3	61.9	84.4	65.3	86.4	69.4	79.2	73.2
Qwen3-VL-Embedding-8B	8B	74.2	81.1	80.2	92.3	80.1	78.4	71.0	58.7	56.1	67.1	87.2	69.9	88.7	73.3	82.4	77.8

Table 3: Results on visual document retrieval benchmarks. All results are obtained from our experimental runs.

Model	Size	VisRAG	VisDocOOD	Vidore-v1	Vidore-v2	Vidore-v3	JinaVDR	Avg
llama-nemoretriever-colembed-1b-v1 (Xu et al., 2025)	1B	82.4	65.6	90.5	62.1	55.5	66.4	70.4
llama-nemoretriever-colembed-3b-v1 (Xu et al., 2025)	3B	85.5	69.7	91.0	55.5	57.1	67.8	71.1
colnomic-embed-multimodal-3b (Team, 2025)	3B	86.8	71.0	89.7	63.5	56.4	77.6	74.2
colqwen2.5-v0.2 (Faysse et al., 2025)	3B	86.6	70.9	89.5	59.3	52.4	75.6	72.4
tomoro-colqwen3-embed-4b (Huang &Tan, 2025)	4B	89.0	75.9	90.6	66.0	60.2	76.2	76.5
colnomic-embed-multimodal-7b (Team, 2025)	7B	88.7	75.6	90.0	62.0	57.6	78.9	75.5
tomoro-colqwen3-embed-8b (Huang &Tan, 2025)	8B	90.2	76.8	90.8	67.7	61.6	79.2	77.7
Qwen3 VL Embedding Models								
Qwen3-VL-Embedding-2B	2B	86.3	74.3	84.4	65.3	52.9	71.0	72.2
Qwen3-VL-Embedding-8B	8B	88.8	78.3	87.2	69.9	59.0	76.9	76.7
Qwen3 VL Reranking Models								
Qwen3-VL-Ranker-2B	2B	89.7	77.5	90.3	62.5	60.8	80.9	77.0
Qwen3-VL-Ranker-8B	8B	91.1	80.4	91.7	71.3	66.7	83.6	80.8

6.3 Text Benchmarks

Table 4 compares our Qwen3-VL-Embedding models with standard text-only embedding models on the MMTEB (Enevoldsen et al., 2025b) benchmark. Compared to text-only Qwen3 embedding models of similar sizes, the Qwen3-VL-Embedding model series show slightly lower performance. Nevertheless, Qwen3-VL-Embedding maintains competitive performance on pure text tasks. Specifically, Qwen3-VL-Embedding-8B achieves a mean task score of 67.9 on MMTEB, performing on par with other similarly sized text-only embedding models.

6.4 Evaluation for Reranking Model

Table 5 presents the evaluation results across various reranking tasks. For multimodal retrieval, we utilize the MMEB-v2 suite, covering image, video (including moment retrieval), and visual document tasks. Text retrieval is evaluated using MMTEB, while visual document retrieval is further assessed on MMEB-v2, JinaVDR, and ViDoRe v3. To ensure a fair comparison, we use Qwen3-VL-Embedding-2B to retrieve the top 100 candidates before applying the reranking models for refinement. Our results demonstrate that all three Qwen3-VL-Reranker models consistently outperform the base embedding model and baseline rerankers, with the 8B variant achieving the best performance across most tasks.

Table 4: Performance on MTEB Multilingual (Enevoldsen et al., 2025a). For compared models, the scores are retrieved from MTEB online leaderboard on December 25th, 2025.

Model	Size	Mean (Task)	Mean (Type)	Bitext Mining	Class-ification	Clus-tering	Inst. Retrieval	Multilabel Class.	Pair Class.	Rerank	Retrieval	STS
Open-Source Models												
KaLM-Embedding-Gemma3-12B-2511 (Zhao et al., 2025)	12B	72.3	62.5	83.8	77.9	55.8	5.5	33.0	84.7	67.3	75.7	79.0
llama-embed-nemotron-8b (Babakhin et al., 2025)	8B	69.5	61.1	81.7	73.2	54.4	10.8	29.9	84.0	67.8	68.7	79.4
NV-Embed-v2 Lee et al. (2024)	7B	56.3	49.6	57.8	57.3	40.8	1.0	18.6	78.9	63.8	56.7	71.1
GritLM-7B (Muennighoff et al., 2024)	7B	60.9	53.7	70.5	61.8	49.8	3.5	22.8	80.9	63.8	58.3	73.3
BGE-M3 (Chen et al., 2024)	0.6B	59.6	52.2	79.1	60.4	40.9	-3.1	20.1	80.8	62.8	54.6	74.1
multilingual-e5-large-instruct (Wang et al., 2024a)	0.6B	63.2	55.1	80.1	64.9	50.8	-0.4	22.9	80.9	62.6	57.1	76.8
gte-Qwen2-1.5B-instruct (Li et al., 2023)	1.5B	59.5	52.7	62.5	58.3	52.1	0.74	24.0	81.6	62.6	60.8	71.6
gte-Qwen2-7b-instruct (Li et al., 2023)	7B	62.5	55.9	73.9	61.6	52.8	4.9	25.5	85.1	65.6	60.1	74.0
Qwen3-Embedding-0.6B (Zhang et al., 2025c)	0.6B	64.3	56.0	72.2	66.8	52.3	5.1	24.6	80.8	61.4	64.6	76.2
Qwen3-Embedding-4B (Zhang et al., 2025c)	4B	69.5	60.9	79.4	72.3	57.2	11.6	26.8	85.1	65.1	69.6	80.9
Qwen3-Embedding-8B (Zhang et al., 2025c)	8B	70.6	61.7	80.9	74.0	57.7	10.1	28.7	86.4	65.6	70.9	81.1
Closed-Source Models												
text-embedding-3-large [†]	-	58.9	51.4	62.2	60.3	46.9	-2.7	22.0	79.2	63.9	59.3	71.7
Cohere-embed-multilingual-v3.0 [†]	-	61.1	53.2	70.5	63.0	46.9	-1.9	22.7	79.9	64.1	59.2	74.8
Gemini Embedding (Lee et al., 2025)	-	68.4	59.6	79.3	71.8	54.6	5.2	29.2	83.6	65.6	67.7	79.4
Seed-1.6-embedding-1215[†]	-	70.3	61.3	78.7	76.8	56.8	-0.0	46.2	85.5	66.2	66.1	75.9
Qwen3 VL Embedding Models												
Qwen3-VL-Embeddnig-2B	2B	63.9	55.8	69.5	65.9	52.5	3.9	26.1	78.5	64.8	67.1	74.3
Qwen3-VL-Embeddnig-8B	8B	67.9	58.9	77.5	72.0	55.8	4.5	28.6	81.1	65.7	69.4	75.4

Table 5: Evaluation results for reranking models and baselines. All scores are obtained from our experimental runs.

Model	Size	MMEB-v2(Retrieval)				MMTEB(Retrieval)	JinaVDR	ViDoRe(v3)
		Avg	Image	Video	VisDoc			
Qwen3-VL-Embedding-2B	2B	73.4	74.8	53.6	79.2	68.1	71.0	52.9
jina-reranker-m0 [†]	2B	-	68.2	-	85.2	-	82.2	57.8
Qwen3-VL-Reranker-2B	2B	75.1	73.8	52.1	83.4	70.0	80.9	60.8
Qwen3-VL-Reranker-8B	8B	79.2	80.7	55.8	86.3	74.9	83.6	66.7

7 Analysis

7.1 Efficacy of Matryoshka Representation Learning and Embedding Quantization
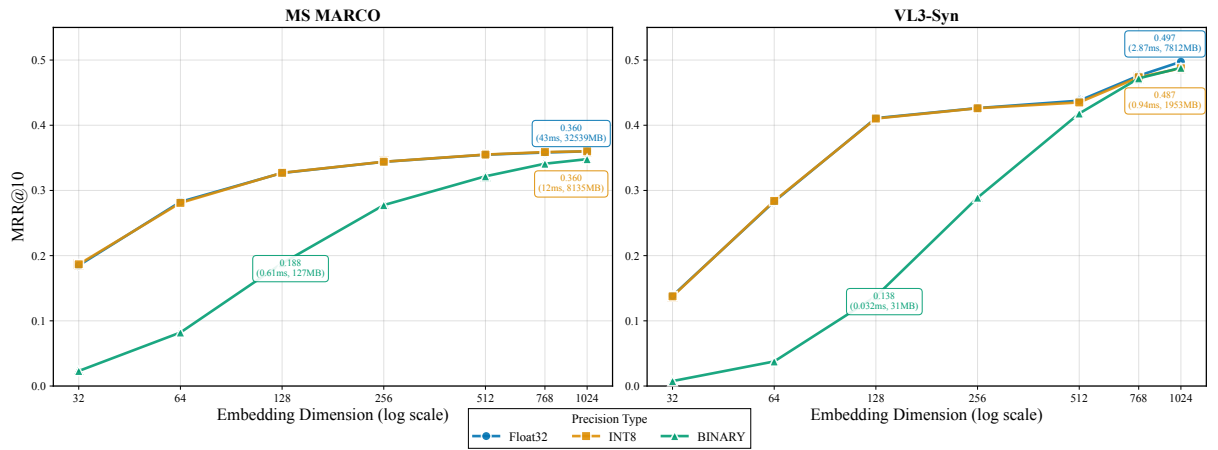

Figure 6: Performance analysis of different embedding dimensions and embedding quantization on MSMARCO Passage Dataset (text to text retrieval) and VL3-Syn Dataset (text to image retrieval).

Embedding models are foundational to modern retrieval systems, spanning both unimodal tasks (e.g., text retrieval) and cross-modal scenarios (e.g., text to image retrieval). In large-scale production environments, the corpus size often reaches millions or even billions of entries. Consequently, optimizing storage requirements for the corpus while enhancing computational efficiency by reducing retrieval latency is a critical challenge. The Qwen3-VL-Embedding series addresses these needs by integrating Matryoshka Representation Learning (MRL) and Quantization-Aware Training (QAT) into its training pipeline.

To evaluate the practical impact of these strategies on retrieval performance, we conduct benchmarks

Table 6: Performance of Qwen3-VL-Embedding-2B across different training stages on the MMEB-V2.

| Model | Image | | | | | Video | | | | | VisDoc | | | | | All |
Stage	CLS	QA	RET	GD	Overall	CLS	QA	RET	MRET	Overall	VDRv1	VDRv2	VR	OOD	Overall	
s0	62.2	63.7	65.9	80.0	65.8	60.8	65.9	51.1	48.4	57.5	76.7	59.8	79.5	64.3	74.8	66.6
s1	71.2	75.8	72.4	88.3	74.8	73.0	67.7	51.3	41.6	60.3	83.5	58.8	84.9	66.4	77.1	72.1
s2	61.8	69.8	78.8	76.3	71.3	63.9	60.0	55.6	57.8	59.5	84.2	72.4	87.9	70.6	80.9	71.5
s3	70.3	74.3	74.8	88.5	75.0	71.9	64.9	53.9	53.3	61.9	84.4	65.3	86.4	69.4	79.2	73.2

across two representative tasks. The first is a text retrieval task utilizing the MSMARCO Passage Ranking dataset (Bajaj et al., 2016), where we sample 10,000 queries and use all passages from the training dataset as our test corpus. The second is a cross-modal text to image retrieval task based on the VL3-Syn (Zhang et al., 2025a) dataset, featuring 10,000 captions as queries and a corpus of 2,000,000 images. We adopt the Qwen3-VL-Embedding-2B model for experimentation and utilize MRR@10 as our primary evaluation metric. Furthermore, we provide a comprehensive analysis of index storage overhead and retrieval latency across varying embedding dimensions and quantization schemes to demonstrate the tradeoffs between accuracy and efficiency.

As illustrated in Figure 6, we observe consistent patterns in both text retrieval and text to image cross-modal retrieval. Regarding embedding dimensionality, retrieval performance degrades as dimensions decrease; however, within a reasonable range, this degradation is acceptable given the substantial savings in storage and retrieval latency. For instance, in text retrieval tasks, reducing the embedding dimension from 1024 to 512 results in only a 1.4% decrease in retrieval performance while achieving 50% storage reduction and doubling retrieval speed. Regarding embedding quantization, we find that int8 quantization preserves retrieval performance with negligible degradation, whereas binary quantization significantly impairs retrieval effectiveness. Moreover, this performance loss becomes increasingly pronounced as embedding dimensionality decreases.

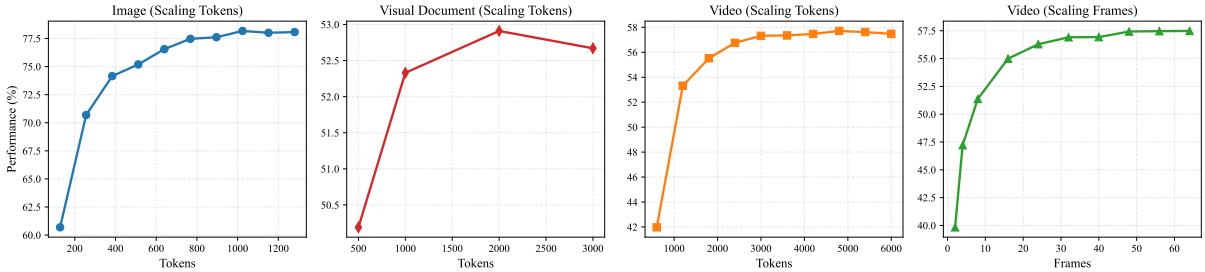## 7.2 Impact of Spatial and Temporal Granularity

Figure 7: Impact of visual granularity on model performance across different domains.

In this section, we investigate how model performance scales with visual granularity across different dimensions. Specifically, for the image modality, we examine the impact of spatial resolution as measured by the number of visual tokens. For video, we decouple our analysis into two axes: (i) temporal granularity, measured by the number of frames, and (ii) spatial resolution, quantified by the aggregate token budget across all frames.

We first analyzed the distribution of image/video resolutions and frame counts across the MMEB-v2 benchmarks, selecting several high-resolution tasks from the Image, Video, and Visual Document domains for our experiments. The results are illustrated in Figure 7. Our findings indicate a consistent trend where performance improves with increased resource consumption across all task categories. However, we observe a pronounced diminishing return as resource allocation grows, with a slight performance regression occurring at the highest levels of consumption. A potential explanation for this decline is the inherent performance degradation that the model encounters when processing excessively long contexts.

7.3 Performance Across Training Stages

In our multi-stage training pipeline, a total of four embedding models were produced. Table 6 details the performance of these four models at the 2B size. The results indicate that by distilling from a reranking model, the embedding model achieves a substantial performance boost in retrieval-oriented tasks.

Although a slight decline is observed in other task categories during this process, the final model merging stage successfully reconciles these trade-offs, leading to a robust and superior overall performance across all benchmarks.

8 Conclusion

In this work, we present Qwen3-VL-Embedding and Qwen3-VL-Reranker, a state-of-the-art model series for multimodal retrieval. By integrating a multi-stage training pipeline with high-quality multimodal data while maximally leveraging the multimodal knowledge and general understanding capabilities of Qwen3-VL Foundation models, Qwen3-VL-Embedding and Qwen3-VL-Reranker model series achieve unprecedented performance across a broad spectrum of multimodal retrieval benchmarks while maintaining strong pure-text capabilities. Furthermore, through matryoshka representation learning and quantization-aware training, the Qwen3-VL-Embedding series offers excellent practical deployment characteristics, significantly reducing computational costs for downstream tasks while preserving superior performance. Looking forward, promising directions include extending support to additional modalities, developing more efficient training paradigms, enhancing compositional reasoning capabilities, and establishing more comprehensive evaluation protocols. We believe these models represent a significant advancement in multimodal retrieval technology and hope they will facilitate further innovation in this rapidly evolving field.

References

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks, 2025. URL https://arxiv.org/abs/2511.07025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Ziqi Dai, Xin Zhang, Mingxin Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Supervised fine-tuning or contrastive learning? towards better multimodal llm reranking. *arXiv preprint arXiv:2510.14824*, 2025.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzeminski, Genta Indra Winata, et al. Mmteb: Massive multilingual text embedding benchmark. In *International Conference on Learning Representations*. International Conference on Learning Representations, 2025a.

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, 2025b. URL https://arxiv.org/abs/2502.13595.

Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. In *International Conference on Learning Representations*, 2020.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *ICLR*, 2025.

Chenghan Fu, Daoze Zhang, Yukang Lin, Zhanheng Nie, Xiang Zhang, Jianyu Liu, Yueran Liu, Wanxian Guan, Pengjie Wang, Jian Xu, et al. Moon embedding: Multimodal representation learning for e-commerce search advertising. *arXiv preprint arXiv:2511.11305*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

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

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. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xiang Huang, Hao Peng, Dongcheng Zou, Zhiwei Liu, Jianxin Li, Kay Liu, Jia Wu, Jianlin Su, and Philip S. Yu. Cosent: Consistent sentence embedding via similarity ranking. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:2800–2813, May 2024. ISSN 2329-9290. doi: 10.1109/TASLP.2024.3402087. URL https://doi.org/10.1109/TASLP.2024.3402087.

Xin Huang and Kye Min Tan. Beyond text: Unlocking true multimodal, end-to-end rag with tomoro colqwen3, 2025. URL https://tomoro.ai/insights/beyond-text-unlocking-true-multimodal-end-to-end-rag-with-tomoro-colqwen3.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. Rzenembed: Towards comprehensive multimodal retrieval. *arXiv preprint arXiv:2510.27350*, 2025.

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. *arXiv preprint arXiv:2407.12580*, 2024.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *ICLR*, 2025.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

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. *arXiv preprint arXiv:2405.17428*, 2024.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*, 2025.

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. *arXiv preprint arXiv:2410.15035*, 2024.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Muhammad Arslan Manzoor, Sarah Albarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–34, 2023.

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

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning, 2024.

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

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations (ICLR)*, 2021.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Nomic Team. Nomic embed multimodal: Interleaved text, image, and screenshots for visual document retrieval, 2025. URL https://nomic.ai/blog/posts/nomic-embed-multimodal.

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. *arXiv preprint arXiv:2212.03533*, 2022.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024a.

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

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

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.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9274–9285, 2025b.

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. *arXiv preprint arXiv:2506.05176*, 2025c.

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, Youcheng Pan, Yang Xiang, Meishan Zhang, Haofen Wang, Jun Yu, Baotian Hu, and Min Zhang. Kalm-embedding-v2: Superior training techniques and data inspire a versatile embedding model, 2025. URL https://arxiv.org/abs/2506.20923.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. Megapairs: Massive data synthesis for universal multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19076–19095, 2025.

A   Dataset Examples

Table 7: Dataset format examples: Docmatix [†] and MS-COCO (Lin et al., 2014).

Dataset	Docmatix
Instruction	Find a screenshot that relevant to the user's question.
Queries (Q_i)	
q_01	What type of research project was announced by the Danish Cancer Society on 01/02/21?
Corpus (C_i)	
d_01	
d_02	
d_03	
Relevance (R_i)	`{q_01: pos: [d_01], neg: [d_02, d_03]};`

Dataset	MS-COCO
Instruction	Find an image caption describing the following everyday image.
Queries (Q_i)	
q_01	
Corpus (C_i)	
d_01	A man swinging a baseball bat on a baseball field.
d_02	The man is walking on the field to play a game of baseball.
d_03	A boy playing baseball waiting for a pitch.
Relevance (R_i)	`{q_01: pos: [d_01], neg: [d_02, d_03]};`

B   Examples of Data Synthesis Prompts

Image Question Answering Prompt

```
You are given an image. Your job is to create ONE high-quality multimodal training
↪  example for an Image Question Answering (IQA) dataset.
The final answer MUST be a single JSON object and nothing else.
```

```
STEP 1 - Visual Description (less or equal than 500 {language} words)
- General scene summary and object-level details (attributes, positions, relations).
- Contextual features (environment, lighting, actions).
- Brainstorm the types of reasoning enabled (e.g., spatial, comparative, predictive).

STEP 2 - Task Selection
Choose ONE task type from the list below that best fits the image content:
- Factoid Identification: Questions about specific entities, brands, or basic facts
↪ (e.g., "What brand is the watch?").
- Visual Reasoning: Questions requiring logical inference or analysis (e.g., "How many
↪ rats were fed the control diet?").
- OCR-based Data Extraction: Questions targeting text, tables, or document info (e.g.,
↪ "Who is the author of the book?").
- Domain-specific Knowledge Inquiry: Questions requiring specialized background
↪ knowledge (e.g., "What style of architecture is this?").

STEP 3 - Populate the Example
Fill every key below using double quotes. Do not add extra keys.
{
 "description": "<STEP 1 output>",
 "task_type": "<Task selected in STEP 2>",
 "question": "<A visually grounded question in {language}>",
 "positive_answer": "<Concise, correct answer in {language}>",
 "hard_negative_answer": "<A plausible but deceptive incorrect answer in {language}>"
}

Hard Constraints:
- "task_type" must be exactly chosen from the list in STEP 2.
- Ensure the question is directly answerable from the visual or embedded textual
↪ content.
- Output ONLY the JSON object.
```

Video Classification Prompt

```
You are given a video. Your job is to create ONE high-quality multimodal training
↪ example for a video classification dataset.
The final answer MUST be a single JSON object and nothing else.

STEP 1 - Visual Analysis (less or equal than 300 {language} words)
- General overview of the video content.
- Identify primary actions, environmental settings, and the overall event type.
- Brainstorm potential ways this video could support the classification tasks listed in
↪ STEP 2.

STEP 2 - Task Selection
Choose ONE task type from the list below that best fits the video:
- Activity Recognition: Identifying the main activity or action being performed.
- Scene Parsing: Determining the primary environment or setting of the video.
- Event Categorization: Classifying the video into a specific event type or intended
↪ purpose.
- Sentiment/Intent Analysis: Recognizing the dominant emotional tone or the sentiment
↪ expressed.

STEP 3 - Populate the Example
Fill every key below using double quotes. Do not add extra keys.
{
  "description": "<STEP 1 output>",
  "task_type": "<Task Selected in STEP 2>",
  "label": "<Correct label in {language}>",
  "misleading_label": "<Plausible but incorrect label in {language} for hard negative
    ↪ mining>"
```

```
    }

    Hard Constraints:
    - "task_type" must be exactly chosen from the list in STEP 2.
    - "description", "label", and "misleading_label" must be in {language}.
    - Output ONLY the JSON object—no extra text or explanations.
```

C   Model Applications and Examples

In this section, we present several real-world application scenarios to demonstrate the practical utility of Qwen3-VL-Embedding. The showcases in Table Tables 8 to 11 illustrate how the model handles diverse queries and complex visual data, providing a clearer understanding of its integration into downstream tasks.

Table 8: Similarity scores evaluated by Qwen3-VL-Embedding (text tasks).

Task	AG News (Zhang et al., 2015)		
Instruction	Classify the news article.		

Ex.	Query	Document	Sim.
1	**Text:** Fears for T N pension after talks Unions representing workers at Turner Newall say they are 'disappointed' after talks with stricken parent firm Federal Mogul.	**Text:** Business	0.55
2	**Text:** US fighter squadron to be deployed in South Korea next month (AFP) AFP - A squadron of US Air Force F-15E fighters based in Alaska will fly to South Korea next month for temporary deployment aimed at enhancing US firepower on the Korean peninsula...	**Text:** World	0.57

Task	SQuAD (Rajpurkar et al., 2016)		
Instruction	Retrieve passages that answer this question.		

Ex.	Query	Document	Sim.
1	**Text:** Which NFL team represented the AFC at Super Bowl 50?	**Text:** Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated...	0.81
2	**Text:** Who headlined the halftime show for Super Bowl 50?	**Text:** CBS broadcast Super Bowl 50 in the U.S., and charged an average of $5 million for a 30-second commercial during the game. The Super Bowl 50 halftime show was headlined by the British rock group Coldpl...	0.75

Task	MS MARCO (Lin et al., 2014)		
Instruction	Retrieve relevant passages.		

Ex.	Query	Document	Sim.
1	**Text:** walgreens store sales average	**Text:** The average Walgreens salary ranges from approximately $15,000 per year for Customer Service Associate / Cashier to $179,900 per year for District Manager. Average Walgreens hourly pay ranges from app...	0.77
2	**Text:** how much do bartenders make	**Text:** According to the Bureau of Labor Statistics, the average hourly wage for a bartender is 10.36, $and the average yearly take-home is$ $21,550$. Bartending can be a lot of things. For some it is exciting,...	0.81

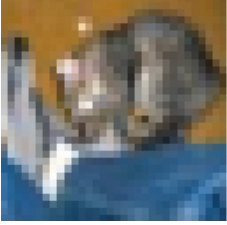Table 9: Similarity scores evaluated by Qwen3-VL-Embedding (image tasks).

Task	CIFAR-10 (Krizhevsky et al., 2009)		
Instruction	Classify the object in this image.		

Ex.	Query	Document	Sim.
1	**Image:**	**Text:** cat	0.67
2	**Image:**	**Text:** truck	0.69

Task	VQAv2 (Goyal et al., 2017)		
Instruction	Find the answer to this question about the image.		

Ex.	Query	Document	Sim.
1	**Text:** Where is he looking? **Image:**	**Text:** down	0.54
2	**Text:** What are the people in the background doing? **Image:**	**Text:** watching	0.67

Task	MS COCO (Lin et al., 2014)		
Instruction	Find images matching this description.		

Ex.	Query	Document	Sim.
1	**Text:** A man with a red helmet on a small moped on a dirt road.	**Image:**	0.52
2	**Text:** The bathroom is clean and ready to be used.	**Image:**	0.46

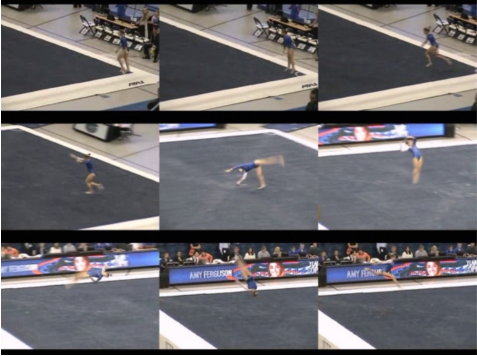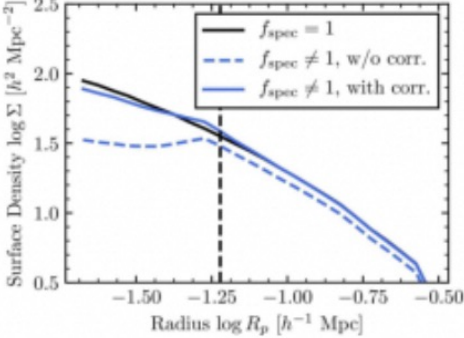Table 10: Similarity scores evaluated by Qwen3-VL-Embedding (video tasks).
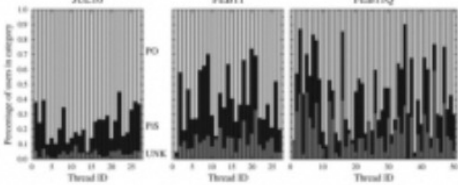
Task	**UCF101 (Soomro et al., 2012)**
Instruction	Classify the action in this video.

Ex.	Query	Document	Sim.
1	**Video:**	**Text:** FloorGymnastics	0.66

Task	**NExTQA (Xiao et al., 2021)**
Instruction	Find the answer to this question about the video.

Ex.	Query	Document	Sim.
1	**Text:** Why did the girl have painted nail polish on her nails... **Video:**	**Text:** (E) look nice	0.64

Task	**MST-VTT (Xu et al., 2016)**
Instruction	Find videos matching this description.

Ex.	Query	Document	Sim.
1	**Text:** baseball player hits ball	**Video:**	0.80

Table 11: Similarity scores evaluated by Qwen3-VL-Embedding (visual document tasks).

Task	ViDoRe_ArxivQA (Faysse et al., 2025)		
Instruction	Find documents that answer this question.		

Ex.	Query	Document	Sim.
1	**Text:** Based on the graph, what is the impact of correcting for fspec not equal to 1 on the surface density trend?	**Image:**	0.63
2	**Text:** Based on the progression from JUL10 to FEB11Q, what trend can be observed in the thread participation?	**Image:**	0.55