

RzenEmbed: Towards Comprehensive Multimodal Retrieval

Weijian Jian Yajun Zhang Dawei Liang Chunyu Xie Yixiao He
Dawei Leng[†] Yuhui Yin

360 AI Research

Abstract

The rapid advancement of Multimodal Large Language Models (MLLMs) has extended CLIP-based frameworks to produce powerful, universal embeddings for retrieval tasks. However, existing methods primarily focus on natural images, offering limited support for other crucial visual modalities such as videos and visual documents. To bridge this gap, we introduce RzenEmbed, a unified framework to learn embeddings across a diverse set of modalities, including text, images, videos, and visual documents. We employ a novel two-stage training strategy to learn discriminative representations. The first stage focuses on foundational text and multimodal retrieval. In the second stage, we introduce an improved InfoNCE loss, incorporating two key enhancements. Firstly, a hardness-weighted mechanism guides the model to prioritize challenging samples by assigning them higher weights within each batch. Secondly, we implement an approach to mitigate the impact of false negatives and alleviate data noise. This strategy not only enhances the model’s discriminative power but also improves its instruction-following capabilities. We further boost performance with learnable temperature parameter and model souping. RzenEmbed sets a new state-of-the-art on the MMEB benchmark. It not only achieves the best overall score but also outperforms all prior work on the challenging video and visual document retrieval tasks. Our models are available in <https://huggingface.co/qihoo360/RzenEmbed>.

1 Introduction

Multimodal retrieval, which aims to find semantically related information across heterogeneous data types like text, images, and video, is a fundamental task in artificial intelligence. Early approaches relied on hand-crafted features and shallow fusion mechanisms, which struggled to capture high-level semantic correspondences. The rise of deep contrastive learning has revolutionized this field, enabling models to learn rich, shared embedding spaces from massive image-text corpora.

Landmark models such as CLIP (Radford et al., 2021), Florence-2 (Xiao et al., 2024), and FG-CLIP (Xie et al., 2025) have demonstrated remarkable zero-shot transfer capabilities by aligning global image and text representations through contrastive objectives. More recently, Multimodal Large Language Models (MLLMs) like LLaVA (Liu et al., 2023) and Qwen2-VL (Wang et al., 2024b) have extended these frameworks by leveraging language modeling objectives to produce unified, semantically grounded embeddings. These advances have significantly improved performance on standard retrieval benchmarks, particularly in image-text settings.

However, these successes remain largely confined to natural images paired with descriptive text. As we move toward truly universal multimodal systems, there is a growing need to support more complex and structured visual modalities, such as videos with temporal

[†] Corresponding Author, E-mail: lengdawei@360.cn

dynamics and visual documents with layout-sensitive semantics. Unfortunately, most existing embedding models are not designed to handle such diversity. When applied to video or document retrieval, they suffer from degraded performance due to misaligned temporal segments, noisy captions, and structural ambiguities. This narrow generalization hinders the development of universal retrieval systems in real-world applications.

The Multimodal Embedding Benchmark (MMEB) (Jiang et al., 2025; Meng et al., 2025) has emerged to evaluate this broader vision of universal retrieval, requiring strong performance across a heterogeneous suite of tasks. Yet, current methods struggle on the more challenging sub-tasks of MMEB, especially video and visual document retrieval, which presents several technical challenges in their training paradigms. First, the standard contrastive learning objective can be compromised by the presence of false negatives (semantically similar samples incorrectly treated as negatives) and hard negatives (subtly different samples that the model struggles to distinguish), which impairs the final discriminative ability of the embeddings (Robinson et al., 2021). Second, the temperature parameter in InfoNCE is typically shared or fixed, despite differing optimal scales across tasks (e.g., fine-grained document retrieval may require sharper similarity distributions than coarse video retrieval) (Qiu et al., 2023). Third, the design of text prompts significantly influences embedding quality. Until now, systematic strategies for generating consistent and compact representations remain underexplored (Ju & Lee, 2025).

To address these challenges, we introduce RzenEmbed, a unified framework for learning universal embeddings across text, images, videos, and visual documents. Our approach uses a two stage training strategy. The first stage establishes broad cross-modal alignment using diverse multimodal datasets. The second stage refines the model with task-aware improvements, including a hardness-weighted mechanism to reduce the impact of false negatives and emphasize hard negatives, a learnable temperature module for per-task scaling, and a compact embedding prompt design to ensure discriminative representations. We also apply model souping to improve stability and final performance. On MMEB, RzenEmbed achieves new state-of-the-art results, outperforming all previous methods in overall score and especially in video and visual document retrieval tasks.

The main contributions of this work are summarized as follows:

- We propose RzenEmbed, a unified framework with a two-stage training strategy to learn highly discriminative and universal embeddings for text, images, videos, and visual documents.
- We introduce a method to identify and eliminate false negative samples, alongside a hardness-weighted mechanism that enhances the model’s ability to learn from challenging samples.
- We integrate a learnable temperature mechanism, a embedding prompt design, and model souping, further improving the model’s robustness and performance across diverse modalities.
- We achieve SOTA performance on MMEB, setting new benchmarks in challenging cross-modal retrieval tasks.

2 Related Work

Embedding is an indispensable technique in modern information retrieval, drawing similar items closer while distancing dissimilar ones within a vector space. Traditional methods typically focus on unimodal queries and targets—such as text-to-text, image-to-image, or text-to-image retrieval—and are typically addressed separately using distinct methodologies. The emergence of CLIP (Radford et al., 2021) unifies these tasks within a single framework leveraging contrastive learning. Building upon this, Wei et al. (2024) introduce the unified instruction-guided retriever UniIR, capable of processing eight distinct tasks with mixed modalities within a single framework. However, UniIR adopts CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) as its base models, which exhibit limitations in following complex instructions.

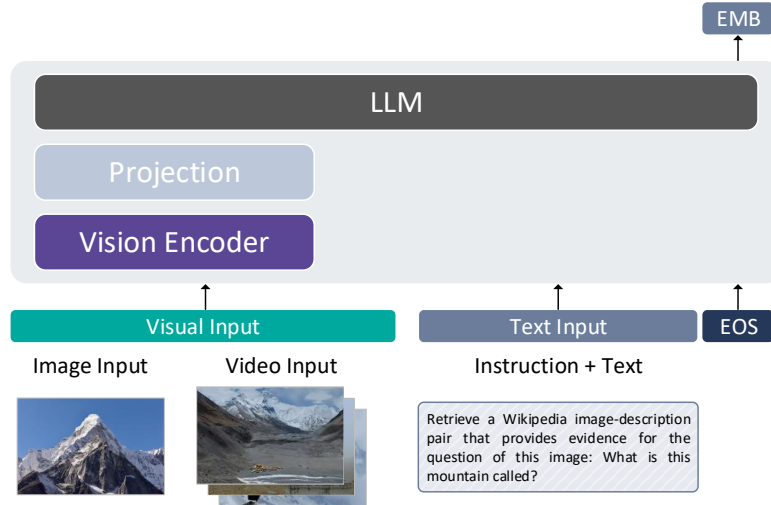


Figure 1: The architecture of RzenEmbed. The model takes both visual input (images or videos) and text input (instruction + text). Visual data is processed by a vision encoder and projection layer before being fed into the LLM. The LLM jointly encodes both modalities, and the final embedding (EMB) is extracted from the last token’s hidden state.

Recently, the advancement of Multimodal Large Language Models (MLLMs) has spurred numerous efforts (Jiang et al., 2025; Lin et al., 2025; Zhou et al., 2025; Zhang et al., 2025; Chen et al., 2025; Lan et al., 2025; Gu et al., 2025a; Thirukovalluru et al., 2025; Xue et al., 2025) to adapt them for multimodal embedding tasks. Jiang et al. (2025) introduce MMEB, a benchmark for training and evaluating multimodal embeddings, alongside VLM2Vec, a contrastive framework for converting any MLLM into an embedding model. Lin et al. (2025) propose MM-Embed, which enhances text retrieval capabilities through modality-aware hard negative mining and continuous fine-tuning. Zhou et al. (2025) present a data synthesis method, MegaPairs, leveraging public images and vision-language models. Similarly, Zhang et al. (2025) introduce a fused-modal data synthesis approach for training a general multimodal embedder. Chen et al. (2025) further develop a synthesis method covering diverse tasks, languages, and modality combinations, training a multimodal multilingual E5 model. Lan et al. (2025) introduce LLaVE, which improves multimodal embeddings by leveraging the discriminative difficulty of negative pairs. Gu et al. (2025a) propose UniME, a two-stage method: the first stage distills textual knowledge from an LLM teacher to enhance the MLLM’s language component, while the second stage employs instruction tuning augmented with hard negatives. Thirukovalluru et al. (2025) present a novel batch construction technique, B3, constructing a sparse dataset similarity graph and applying community detection to identify clusters of strong negatives. Finally, Xue et al. (2025) advocate amplifying gradients for hard negatives within the Info-NCE loss to learn more discriminative multimodal embeddings.

Previous research predominantly focuses on image and text modalities, largely ignoring other visual modalities like video and visual documents. This limitation restricts the practical applicability of these approaches in real-world scenarios. To address these shortcomings, Meng et al. (2025) present MMEB-V2, extending the original MMEB benchmark with five novel tasks encompassing video and visual documents. They concurrently propose VLM2Vec-V2, a unified framework designed to learn embeddings across images, videos, and visual documents. Most recently, ByteDance Seed introduce Seed-1.6-Embedding (ByteDance Seed, 2025), which employs a three-stage training strategy comprising text continual training, multimodal continual training, and fine-tuning. Seed-1.6-Embedding achieves the highest overall score on the MMEB-V2 benchmark.

3 Method

3.1 Architecture

Our primary objective is to learn a unified embedding space capable of supporting a diverse range of modalities and tasks. This requires a model backbone that can flexibly encode text, images, and videos, while efficiently processing long-context inputs. Recent advances in Multimodal Large Language Models (MLLMs) (Liu et al., 2024; Abdin et al., 2024; Gu et al., 2024; Beyer et al., 2024; Hong et al., 2024; Li et al., 2025; Wang et al., 2024b; Bai et al., 2025; Wang et al., 2025) have demonstrated remarkable performance across various benchmarks and serve as a strong foundation for multimodal embedding systems (Jiang et al., 2025; Faysse et al., 2025; Zhang et al., 2025; Chen et al., 2025; Lan et al., 2025; Meng et al., 2025; Gu et al., 2025b).

In light of these requirements, we adopt Qwen2-VL (Wang et al., 2024b) as the backbone. As illustrated in Figure 1, the architecture is designed to process both visual and textual data seamlessly. The choice of Qwen2-VL is motivated by its key features that align with our goals: (1) **Native Dynamic Resolution**, which efficiently handles visual inputs of varying resolutions; (2) **Multimodal Rotational Position Embeddings (M-RoPE)**, enabling robust modeling of static images and temporal features in videos; and (3) **Strong Generalization**, particularly for instruction-following tasks. These capabilities make it an ideal choice for scalable and generalizable encoding of heterogeneous multimodal data.

The model accepts two primary types of input: **visual input** and **text input**.

- **Visual Input** can be either static images or videos. For video tasks, we represent the video as a sequence of frames sampled at a fixed interval to ensure consistent temporal coverage.
- **Text Input** is structured as a combination of an Instruction and associated Text. This instruction-based format guides the model to perform specific tasks, such as retrieval or question answering.

The Large Language Model (LLM) jointly processes the sequence of projected visual tokens and text tokens. Inspired by established practices in text embedding (Li et al., 2023b; Wang et al., 2024a), we extract the embedding (EMB) from the final hidden state of the last token (EOS token) of the entire input sequence. This single vector serves as a comprehensive and unified representation for the given multimodal input.

3.2 Training

3.3 Training Objective

We train Rzenembed using a contrastive learning framework. Our approach is fundamentally based on the InfoNCE loss (Rusak et al., 2025), a cornerstone objective in self-supervised and contrastive learning. The core principle of InfoNCE is to train a model that pulls the representation of a query (or "anchor") and its corresponding "positive" sample closer together in the embedding space, while simultaneously pushing it apart from a set of "negative" samples.

Formally, given a query vector q , a corresponding positive sample vector k^+ , and a set of N negative sample vectors $\{k_i^-\}_{i=1}^N$, the InfoNCE loss is defined as:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(q, k^+)/\tau)}{\exp(\text{sim}(q, k^+)/\tau) + \sum_{i=1}^N \exp(\text{sim}(q, k_i^-)/\tau)} \quad (1)$$

Here, $\text{sim}(\cdot)$ denotes a similarity function, typically cosine similarity, and τ is a temperature hyperparameter that controls the sharpness of the distribution.

Despite its widespread success, the standard InfoNCE loss suffers from two notable limitations in practice:

- **False Negatives:** A training batch may contain samples that are semantically similar to the query but are inadvertently treated as negatives. This penalizes the model for recognizing valid similarities, which can hinder convergence.
- **Dominance of Easy Negatives:** A typical batch is often dominated by easy negatives (samples that are apparently different from the query). This causes the model to allocate most of its learning capacity to trivial distinctions while neglecting the more informative hard negatives, which are crucial for learning a fine-grained representation space.

To address these challenges, we introduce two key modifications to the standard InfoNCE framework, as detailed below.

3.3.1 False Negative Mitigation

False negatives are instances within a training batch that, despite being sampled as negatives, are semantically similar or even equivalent to the query. For instance, in a batch of text passages, two documents discussing the same topic might be incorrectly contrasted, misleading the model.

To mitigate the impact of false negatives, we adopt a straightforward yet effective filtering strategy. During training, for each query-positive pair (q, k^+) , we identify potential false negatives from the set of negative samples $\{k_i^-\}$. A negative sample k_i^- is considered as a false negative if its similarity to the positive sample k^+ exceeds a predefined threshold δ :

$$\text{sim}(k_i^-, k^+) > \delta \quad (2)$$

These identified false negatives are then excluded from the denominator of the InfoNCE loss calculation (Equation 1) for that specific query. This simple mechanism prevents the model from being penalized for clustering semantically similar samples together, thereby achieving more stable and meaningful learning.

3.3.2 Hardness-Weighted Strategy

Hard negatives are samples that are semantically distinct from the query but lie close to it in the embedding space, making them difficult for the model to distinguish. For example, an image of a Husky might be a hard negative for a query image of a Samoyed, since they are visually similar but belong to different classes. Effectively learning from these challenging examples is critical for developing a robust and discriminative model, as easy negatives offer little learning signal.

To force our model to focus on these informative samples, we incorporate an exponentially hardness-weighted strategy (Robinson et al., 2021). Instead of treating all negatives equally, this method assigns a higher weight to harder negatives in the loss computation. Specifically, we re-weight each negative sample k_i^- based on its similarity to the query q . The weight w_i is defined as:

$$w_i = \exp(\alpha \cdot \text{sim}(q, k_i^-)) \quad (3)$$

where $\alpha > 0$ is a hyperparameter that controls the strength of the weighting. The modified loss function then becomes:

$$\mathcal{L}_{\text{WHNM}} = -\log \frac{\exp(\text{sim}(q, k^+)/\tau)}{\exp(\text{sim}(q, k^+)/\tau) + \sum_{i=1}^N w_i \cdot \exp(\text{sim}(q, k_i^-)/\tau)} \quad (4)$$

This mechanism ensures that negatives with higher similarity to the query (i.e., harder negatives) receive a larger weight w_i , thereby amplifying their contribution to the loss gradient. This effectively directs the model’s attention towards learning the fine-grained distinctions necessary to resolve these challenging cases.

By combining false negative elimination and hardness-weighted strategy, our training objective enables Rzenembed to learn a more robust and discriminative embedding space.

3.4 Recipe

We train RzenEmbed on a diverse mixture of embedding tasks that span multiple modalities and task types. The training process is divided into two distinct stages: multimodal continual pre-training and fine-tuning. Furthermore, we incorporate a learnable temperature mechanism, a embedding prompt design, and model souping, which further enhance the model’s robustness and performance across a wide range of tasks.

3.4.1 Multimodal Continual Training

The primary goal of this stage is to equip our model with fundamental embedding capabilities. This involves learning to align representations across text, image, and video modalities into a unified semantic space.

In this stage, we deliberately avoid instruction-based fine-tuning. The sole focus is on developing the model’s capacity to generate high-quality and well-aligned embeddings. To this end, we utilize a diverse mixture of training data, categorized into three types:

- **Unimodal Data:** Text-to-Text ($T \rightarrow T$) pairs for improving textual understanding.
- **Cross-modal Data:** Text-to-Image ($T \rightarrow I$) and Text-to-Video Description ($T \rightarrow VD$) pairs for learning cross-modal alignment.
- **Fused-modal Data:** Image-Text-to-Image ($IT \rightarrow I$) pairs, where the model uses a source image and a differential text description to retrieve a target image.

For unimodal ($T \rightarrow T$) training, we leverage established datasets such as MS-MARCO (Nguyen et al., 2016), NQ (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), and TriviaQA (Joshi et al., 2017). Our cross-modal data is sourced from $T \rightarrow I$ pairs in the LAION (Schuhmann et al., 2022) dataset and $T \rightarrow VD$ pairs from ShareGPT4V (Chen et al., 2024). For fused-modal training ($IT \rightarrow I$), we use the Megapairs dataset (Zhou et al., 2025), which is specifically designed for this kind of differential image retrieval. To ensure a balanced data distribution during training, we sample from these datasets with the following proportions: 0.3 million $T \rightarrow T$ pairs, 0.25 million $T \rightarrow VD$ pairs, 2 million $T \rightarrow I$ pairs, and 2.5 million fused-modal pairs.

Enhancing Data with Detailed Recaptioning To improve the model’s comprehension of long and detailed text, we enhance our $T \rightarrow I$ training data. We use a powerful large multimodal model, CogVLM-19B (Hong et al., 2024), to recaption images from the LAION-2B dataset (Schuhmann et al., 2022). This strategy fosters a tighter semantic alignment between visual and textual modalities by training on high-quality image-description pairs. This process simultaneously serves as a data-denoising step, yielding representations more robust to the inherent noise of web-crawled captions. Moreover, by replacing generic labels (e.g., “a cat”) with fine-grained descriptions (e.g., “an orange tabby cat basking in the sun”), our approach enables the model to capture subtle semantic nuances and produce more discriminative embeddings.

Finally, all public datasets undergo a rigorous cleaning process. We employ sophisticated filtering algorithms to remove noise, duplicates, and irrelevant content. We also systematically discard blurry, corrupted, or low-resolution images to ensure the high quality of our training corpus.

3.4.2 Fine-Tuning

The objective of this stage is to comprehensively improve the model’s ability to handle a wide range of specialized scenarios and complex tasks. We achieve this by introducing a diverse mixture of instruction-formatted data.

We systematically construct a high-quality fine-tuning dataset structured around three key dimensions: **task type**, **input modality**, and **task scenario**. This dataset includes the training set from MMEB-v2 (Meng et al., 2025), supplemented by a wide array of public multimodal retrieval and question-answering (QA) datasets.

Similar to the pre-training stage, all data undergoes a strict cleaning process to ensure high quality. A key aspect of our strategy is that **each training batch is sampled from single dataset** (except classification dataset). This approach concentrates hard negative samples within each batch, making the contrastive learning objective more effective. The instruction data in this stage is highly diverse and covers a broad spectrum of tasks:

- **For Images:** Tasks include classification, QA (both multiple-choice and open-ended), retrieval, and grounding.
- **For Visual Documents (VisDoc):** The primary task is Visual Document Retrieval.
- **For Videos:** Tasks encompass Video Retrieval, Moment Retrieval, Video Classification, and Video Question Answering.

To maintain a balanced task distribution and prevent the model from overfitting to any single task, we cut the number of samples from each individual dataset at 100,000.

Merging Image Classification dataset The MMEB-v2 training set can be divided into three categories: images, visual documents, and videos. Most image classification datasets have very few categories. For example, HatefulMemes dataset has only 2 categories, VOC2007 dataset has 20 categories, and N24News dataset has 24 categories. During training under contrastive learning, it is necessary to construct an image-text similarity matrix. This results in a large number of false negative samples. Therefore, we merge all image classification datasets into a new dataset. This significantly reduces the number of false negative samples when a batch is sourced from this new dataset.

Enhancing Video Data We observed that existing video training sets, such as those in MMEB-v2, primarily consist of short videos (under 30 seconds) with high frame-to-frame similarity, making the tasks relatively simple. To address this, we reduce our reliance on this data and supplement our training with a broad collection of public video datasets, processed with the following strategies to increase task difficulty:

- **Segmenting Long Videos:** We divide long videos into multiple short clips, each with a corresponding description. Since these clips originate from the same source video, they serve as natural hard negatives for one another during training.
- **Incorporating Long-Form Videos:** We add long videos (1–3 minutes) paired with holistic descriptions of their overall content. This encourages the model to develop an understanding of long-range temporal dependencies and global context.

3.4.3 Task-Specific Learnable Temperature

In contrast to the standard InfoNCE loss, where the temperature τ is a manually-tuned hyperparameter, we adopt a learnable temperature, following recent work by Li et al. (2023c). This allows the model to dynamically control the sharpness of the softmax probability distribution during training. The temperature $\tau > 0$ governs this sharpness: a smaller τ creates a sharper distribution, compelling the model to focus on the hardest negative samples, whereas a larger τ yields a smoother distribution, encouraging the model to consider all negative samples more uniformly.

Our work extends this concept to a large-scale, multi-task setting. Designed for broad multimodal understanding, our training set is organized into seven distinct tasks, including image classification, image question answering, image retrieval, image grounding, document retrieval, video retrieval, and video question answering. Instead of using a single global temperature, we introduce a dedicated, learnable temperature parameter τ_t for each task t . This allows the model to learn an optimal, task-specific sharpness, accommodating the varying difficulty and sample distributions across different tasks.

To ensure positivity ($\tau_t > 0$) and stable optimization, we employ a re-parameterization strategy. For each task-specific temperature τ_t , we introduce a corresponding learnable scalar θ_t and define the temperature as:

$$\tau_t = \exp(\theta_t). \quad (5)$$

This formulation inherently constrains τ_t to be positive and allows for the unconstrained optimization of θ_t via standard backpropagation, which is updated jointly with other model parameters.

3.4.4 Embedding Prompt

We leverage Qwen2-VL (Wang et al., 2024b) in our contrastive learning architecture, which is primarily trained in a generative manner, but this can pose a significant challenge for discriminative representation learning.

To overcome this, we strategically employ a combination of system prompts and representation prompts (Ju & Lee, 2025), which forces the model to generate representations suitable for discriminative learning.

we use “Given an image, summarize the provided image in one word. Given only text, describe the text in one word.” as the system prompt. And for plain text queries, the representation prompt is “Represent the given text in one word.”, for multimodal queries, the representation prompt is “Represent the given image in one word.”

During model training, the input query is structured as “<system prompt> <query> <representation prompt>”. In inference mode, the query is modified accordingly.

3.4.5 Model Souping

We further enhance the model’s performance by employing the model souping technique specifically for LoRA adapters (Hu et al., 2022; Vera et al., 2025). Instead of deploying multiple specialized LoRA adapters individually, we first consolidate their learned low-rank weight matrices into a single, generalized adapter through a weighted aggregation or other fusion strategies. This “souped” LoRA adapter then captures the complementary knowledge from the individual adapters. Subsequently, this consolidated LoRA adapter is seamlessly merged with the pre-trained base model, creating a unified and more versatile retrieval model. This approach effectively distills the collective expertise of multiple specialized adapters into a single, efficient entity, significantly reducing computational overhead and memory footprint while preserving or enhancing retrieval performance.

4 Experiments

4.1 Train Data

Our training methodology is structured in two sequential stages to instill robust textual, cross-modal retrieval, and instruction-following capabilities. In the first stage, we utilized 5 million data entries to develop foundational embedding skills. For text retrieval, this involved incorporating datasets such as MS-MARCO (Nguyen et al., 2016), NQ (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), TriviaQA (Joshi et al., 2017), SQuAD (Rajpurkar et al., 2016), FEVER (Thorne et al., 2018), and AllNLI for SimCSE (Gao et al., 2021), totaling approximately 300,000 entries. To enable cross-modal retrieval, we randomly sample 2 million entries from LAION-2B (Schuhmann et al., 2022), including both original and CogVLM-19B-generated captions, and supplemented this with 2.5 million randomly sampled MegaPairs dataset entries (Zhang et al., 2024) to ensure basic multimodal retrieval proficiency. The second stage primarily focused on training with the MMEB v2 training set, augmented by mME5-synthetic data and 400,000 video clips sampled from the VideoChat-Flash dataset. This advanced stage aimed to cultivate strong instruction-following retrieval capabilities by exposing the model to diverse multimodal instruction scenarios.

4.2 Training Configuration

We fine-tune the model for a single epoch using the AdamW optimizer. For parameter-efficient tuning, we apply Low-Rank Adaptation (LoRA) to all linear layers of both the vision encoder and the Large Language Model (LLM), with a uniform rank of 64. The

Table 1: Results on the MMEB-V1 benchmark (Jiang et al., 2025). The results in **bold** and underlined represent the best and second-best performances of different model sizes, respectively. IND: in-distribution, OOD: out-of-distribution. [†]: link to the model’s homepage.

Model	Backbone	Model Size	Per Meta-Task Score				Average Score		
			Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# of datasets →			10	10	12	4	20	16	36
<i>Encoder-Only Models</i>									
CLIP (Radford et al., 2021)	-	0.428B	42.8	9.1	53.0	51.8	37.1	38.7	37.8
BLIP-2 (Li et al., 2023a)	-	3.74B	27.0	4.2	33.9	47.0	25.3	25.1	25.2
SigLIP (Zhai et al., 2023)	-	0.203B	40.3	8.4	31.6	59.5	32.3	38.0	34.8
OpenCLIP (Cherti et al., 2023)	-	0.428B	47.8	10.9	52.3	53.3	39.3	40.2	39.7
UniIR (BLIP_FF) (Wei et al., 2024)	-	0.247B	42.1	15.0	60.1	62.2	44.7	40.4	42.8
UniIR (CLIP_SF) (Wei et al., 2024)	-	0.428B	44.3	16.2	61.8	65.3	47.1	41.7	44.7
MagiClens (Zhang et al., 2024)	-	0.428B	38.8	8.3	35.4	26.0	31.0	23.7	27.8
<i>Closed-source Models</i>									
Seed-1.6-embedding [†]	Seed1.6-flash	unknown	76.1	74.0	77.9	91.3	-	-	77.8
<i>~ 2B Models</i>									
VLM2Vec (Jiang et al., 2025)	Phi-3.5-V	4.15B	54.8	54.9	62.3	79.5	66.5	52.0	60.1
VLM2Vec (Jiang et al., 2025)	Qwen2-VL	2.21B	59.0	49.4	65.4	73.4	66.0	52.6	59.3
VLM2Vec-V2 (Meng et al., 2025)	Qwen2-VL	2.21B	62.9	56.3	69.5	77.3	-	-	64.9
UniME-V2 (Gu et al., 2025b)	Qwen2-VL	2.21B	62.1	56.3	68.0	72.7	67.4	58.9	63.6
GME (Zhang et al., 2025)	Qwen2-VL	2.21B	54.4	29.9	66.9	55.5	-	-	51.9
LLaVE (Lan et al., 2025)	Aquila-VL	1.95B	62.1	60.2	65.2	84.9	69.4	59.8	65.2
B3 (Thirukovalluru et al., 2025)	Qwen2-VL	2.21B	67.0	61.2	70.9	79.9	72.1	63.1	68.1
UNITE (Kong et al., 2025)	Qwen2-VL	2.21B	63.2	55.9	65.4	75.6	65.8	60.1	63.3
ColPali-v1.3 (Faysse et al., 2025)	PaliGemma	2.92B	40.3	11.5	48.1	40.3	-	-	34.9
CAFe (Yu et al., 2025)	LLaVA-OV	0.894B	59.1	49.1	61.0	83.0	64.3	53.7	59.6
Ops-MM-embedding-v1 [†]	Qwen2-VL	2.21B	<u>68.1</u>	<u>65.1</u>	69.2	80.9	-	-	<u>69.0</u>
RzenEmbed (ours)	Qwen2-VL	2.21B	68.5	66.3	74.5	90.3	76.1	67.4	72.3
<i>~ 7B Models</i>									
VLM2Vec (Jiang et al., 2025)	LLaVA-1.6	7.57B	61.2	49.9	67.4	86.1	67.5	57.1	62.9
VLM2Vec (Jiang et al., 2025)	Qwen2-VL	8.29B	62.6	57.8	69.9	81.7	65.2	56.3	65.8
UniME-V2 (Gu et al., 2025b)	LLaVA-OV	8.03B	65.3	67.6	72.9	90.2	74.8	66.7	71.2
UniME-V2 (Gu et al., 2025b)	Qwen2-VL	8.29B	64.0	60.1	73.1	82.8	72.0	63.0	68.0
GME (Zhang et al., 2025)	Qwen2-VL	8.29B	57.7	34.7	71.2	59.3	-	-	56.0
LLaVE (Lan et al., 2025)	LLaVA-OV	8.03B	65.7	65.4	70.9	91.9	75.0	64.4	70.3
B3 (Thirukovalluru et al., 2025)	Qwen2-VL	8.29B	<u>70.0</u>	66.5	<u>74.1</u>	84.6	<u>75.9</u>	<u>67.1</u>	72.0
UNITE (Kong et al., 2025)	Qwen2-VL	8.29B	68.3	65.1	71.6	84.8	73.6	66.3	70.3
QQMM-embed (Xue et al., 2025)	LLaVA-OV	8.297B	66.8	66.8	70.5	90.4	74.7	65.6	70.7
CAFe (Yu et al., 2025)	LLaVA-OV	8.03B	65.2	65.6	70.0	91.2	75.8	62.4	69.8
Ops-MM-embedding-v1 [†]	Qwen2-VL	8.29B	69.7	<u>69.6</u>	73.1	87.2	-	-	<u>72.7</u>
LamRA [†]	Qwen2-VL	8.29B	59.2	26.5	70.0	62.7	-	-	54.1
LamRA [†]	Qwen2.5-VL	8.29B	51.7	34.1	66.9	56.7	-	-	52.4
RzenEmbed (ours)	Qwen2-VL	8.29B	70.6	71.7	78.5	92.1	78.5	72.7	75.9

learning rate is initialized to $2e-4$ and decayed following a cosine schedule. We use a global batch size of 768 and a weight decay of $5e-2$. For our proposed training strategies, we set the reweighting factor $\alpha = 9$ for the hardness-weighted strategy and the threshold $\delta = 0.95$ for the false negative elimination strategy. The maximum number of input tokens for both images and videos is 1280. To enhance memory efficiency, we employ bfloat16 (bf16) mixed-precision training and enable gradient checkpointing. All experiments are conducted on 16 NVIDIA A800 (80GB) GPUs.

4.3 Main Results

We evaluate RzenEmbed on a comprehensive suite of benchmarks that span diverse task types and modalities. Specifically, our evaluation results on the MMEB-V1 (Jiang et al., 2025) and MMEB-V2 (Meng et al., 2025) benchmarks are reported in Table 1 and Table 2, respectively.

Results on MMEB-V1 We report RzenEmbed’s performance on MMEB-V1 in Table 1, alongside a comparison with recent related works. The results reveal that RzenEmbed achieves the best performance in both 2B and 7B model scales. Moreover, RzenEmbed attains optimal performance across Per Meta-Task and in both OOD and IND task divisions, which

Table 2: Results on the MMEB-V2 benchmark (Jiang et al., 2025). The results in **bold** and underlined represent the best and second-best performances of different model sizes, respectively. 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.

Model	Backbone	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
Closed-source Models																		
Seed-1.6-embedding†	Seed1.6-flash	unknown	76.1	74.0	77.9	91.3	77.8	55.0	60.9	51.3	53.5	55.3	85.5	56.6	84.7	43.1	73.4	71.3
~ 2B Models																		
VLM2Vec (Jiang et al., 2025)	Qwen2-VL	2.21B	58.7	49.3	65.0	72.9	59.7	33.4	30.5	20.6	33.0	29.0	49.8	13.5	51.8	33.5	41.6	47.0
VLM2Vec-V2 (Meng et al., 2025)	Qwen2-VL	2.21B	62.9	56.3	<u>69.5</u>	77.3	64.9	39.3	34.3	28.8	38.5	34.9	75.5	44.9	79.4	39.4	65.4	58.0
GME (Zhang et al., 2025)	Qwen2-VL	2.21B	54.4	29.9	66.9	55.5	51.9	34.9	42.0	25.6	32.4	33.9	86.1	54.0	82.5	43.1	72.7	54.1
ColPali-v1.3 (Faysse et al., 2025)	PaliGemma	2.92B	40.3	11.5	48.1	40.3	34.9	26.7	37.8	21.6	25.5	28.2	83.6	52.0	81.1	43.1	71.0	44.4
CAFe (Yu et al., 2025)	LLaVA-OV	0.894B	56.4	45.3	57.6	72.0	55.4	33.9	41.7	29.7	39.7	35.9	56.9	32.6	68.6	30.7	51.4	49.7
Ops-MM-embedding-v1†	Qwen2-VL	2.21B	<u>68.1</u>	<u>65.1</u>	69.2	<u>80.9</u>	<u>69.0</u>	53.6	55.7	<u>41.8</u>	33.7	47.6	<u>87.0</u>	57.6	<u>85.4</u>	<u>43.3</u>	<u>74.4</u>	<u>63.4</u>
RzenEmbed (ours)	Qwen2-VL	2.21B	68.5	66.3	74.5	90.3	72.3	<u>50.4</u>	<u>49.7</u>	46.6	<u>38.9</u>	<u>47.3</u>	87.1	<u>55.1</u>	87.2	43.4	74.5	67.2
~ 7B Models																		
VLM2Vec (Jiang et al., 2025)	Qwen2-VL	8.29B	62.7	56.9	69.4	82.2	65.5	39.1	30.0	29.0	40.6	34.0	56.9	9.4	59.1	38.1	46.4	52.3
GME (Zhang et al., 2025)	Qwen2-VL	8.29B	57.7	34.7	71.2	59.3	56.0	37.4	50.4	28.4	38.2	38.6	<u>89.4</u>	55.6	<u>85.0</u>	44.4	<u>75.2</u>	57.8
CAFe (Yu et al., 2025)	LLaVA-OV	8.03B	63.6	61.7	69.1	<u>82.6</u>	67.6	35.8	58.7	34.4	39.5	42.4	70.7	49.6	79.5	38.1	63.9	60.6
Ops-MM-embedding-v1†	Qwen2-VL	8.29B	<u>69.7</u>	<u>69.6</u>	<u>73.1</u>	87.2	<u>72.7</u>	59.7	62.2	<u>45.7</u>	<u>43.2</u>	<u>53.8</u>	80.1	<u>59.6</u>	79.3	43.3	70.3	67.6
LamRA†	Qwen2-VL	8.29B	59.2	26.5	70.0	62.7	54.1	39.3	42.6	24.3	34.6	35.2	22.0	11.5	37.4	21.0	23.9	40.4
LamRA†	Qwen2.5-VL	8.29B	51.7	34.1	66.9	56.7	52.4	32.9	42.6	23.2	37.6	33.7	56.3	33.3	58.2	40.1	50.2	47.4
RzenEmbed (ours)	Qwen2-VL	8.29B	70.6	71.7	78.5	92.1	75.9	<u>58.8</u>	63.5	51.0	45.5	55.7	89.7	60.7	88.7	44.4	77.1	71.6

Table 3: Ablations of strategies. The results in **bold** represent the best performances of different strategies.

Strategies	Merging classification dataset	Learnable temperature	System prompt	Dataset Resample	Overall	Image	Video	Visdoc
Baseline	✗	✗	✗	✗	65.7	71.4	43.5	73.8
Exp1	✓	✗	✗	✗	66.3	71.0	45.3	75.0
Exp2	✗	✓	✗	✗	66.4	71.5	44.0	75.3
Exp3	✗	✗	✓	✗	66.4	71.3	45.0	75.0
Exp4	✓	✓	✓	✗	66.7	71.6	45.8	75.0
Exp5	✓	✓	✓	✓	67.2	72.3	47.3	74.5

illustrates RzenEmbed’s exceptional adaptability to numerous tasks and its generalization power over data from different domains.

Results on MMEB-V2 The results in Table 2 demonstrate that RzenEmbed achieves excellent performance across tasks involving different input modalities, including images, videos, and visual documents on MMEB-V2. Overall, compared to the next best models of same scale, RzenEmbed 2B and 7B models show improvements of 3.4% and 4.0%, respectively. Notably, RzenEmbed’s 7B model outperforms the closed-sourced Seed-1.6-embedding on both the Video and VisDoc subtasks, as well as achieving a higher overall score on MMEB-V2. Analyzing individual tasks, RzenEmbed consistently achieves top performance in 9 and 11 tasks for the 2B and 7B versions, respectively, with a slight underperformance on a few video meta tasks. This further highlights RzenEmbed’s comprehensive multimodal representation capabilities.

4.4 Ablations of Strategies

In this section, we conduct a series of ablation studies to validate the effectiveness of each proposed strategy. The results are summarized in Table 3. Our baseline model is trained with a standard InfoNCE loss and achieves an overall score of 65.7.

Effect of Merging Classification Datasets We hypothesize that the limited label space of individual image classification datasets may restrict the model’s semantic understanding. To address this, we merge multiple classification datasets into a single, larger one with a richer label set. As shown in Table 3 (Exp1), this strategy improves the overall performance

Table 4: Results using different training mixtures. The best results are shown in **bold**.

Pooling	Overall	Image-Overall	Video-Overall	Visdoc-Overall
Mix 1	71.11	75.78	54.16	76.83
Mix 2	71.16	75.64	54.59	76.86
Mix 3	71.18	75.43	55.09	76.88
Souped	71.61	75.92	55.73	77.06

to 66.3. Notably, it enhances performance on video and VisDoc retrieval, suggesting that a broader semantic foundation for images benefits cross-modal learning.

Effect of Learnable Temperature Given the significant heterogeneity in the data distributions and task formats of our training datasets, a single, fixed temperature for the InfoNCE loss is suboptimal. We introduce task-specific learnable temperatures, grouping datasets into seven distinct tasks (e.g., image classification, VQA, retrieval). This allows the model to dynamically balance the penalties for negative samples across different tasks. Table 3 (Exp2) shows this mechanism lifts the overall score to 66.4 and achieves the best VisDoc performance (75.3), confirming the benefits of adaptive temperature scaling.

Effect of System Prompt Our model, Rzenembed, utilizes Qwen2-VL, a backbone pre-trained on generative tasks. To better adapt it for discriminative retrieval tasks, we employ an instruction-tuning approach inspired by prior work [Ju & Lee \(2025\)](#). Specifically, we prepend a system prompt, "summarize the user's intent in one word," to the query. This simple instruction guides the model to produce more discriminative embeddings. As seen in Table 3 (Exp3), this strategy alone improves the performance to 66.4, demonstrating its effectiveness in bridging the gap between generative pre-training and discriminative fine-tuning.

Effect of Dataset Resampling During training, we observed that the loss on video datasets converged much faster than on image datasets, indicating an imbalance in learning dynamics. To mitigate this, we implement a dataset resampling strategy to increase the sampling ratio of image-related data. This rebalancing allows the model to learn more effectively from the slower-converging tasks. As shown in the final experiment (Exp5), when all strategies are combined, resampling further boosts the performance to our best result of 67.2. This highlights the importance of balancing the training data exposure based on task-specific convergence rates.

4.5 Results of Model Souping

We also explore the effectiveness of model souping for LoRA adapters, which involves merging multiple specialized adapters into a single, generalized one. This consolidated adapter captures complementary knowledge, leading to improved performance. As shown in Table 4, this strategy yields the best overall score of 71.61.

5 Conclusion

In this paper, we present RzenEmbed, a novel unified framework that significantly advances multimodal embedding learning. By introducing a sophisticated two-stage training strategy, including a hardness-weighted InfoNCE loss with false negative mitigation, RzenEmbed effectively learns discriminative and universal representations across text, images, videos, and visual documents. Extensive experimental evaluations confirm RzenEmbed's superior performance. It achieves state-of-the-art results on the MMEB leaderboard, setting new records in visual document retrieval, video retrieval, and overall score. As a compact yet highly effective model, RzenEmbed provides a powerful solution to the growing need for

advanced multimodal retrieval in applications such as AI agents, multimodal search and recommendation, and Retrieval-Augmented Generation.

References

- Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norrick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacrose, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp A. Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219, 2024.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibor Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.
- Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey A. Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bosnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier J. Hénaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b VLM for transfer. *CoRR*, abs/2407.07726, 2024.
- ByteDance Seed. Seed1.6-embedding. Online, 2025. URL <https://seed1-6-embedding.github.io/>.
- Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. In *ACL (Findings)*, pp. 8254–8275. Association for Computational Linguistics, 2025.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV (17)*, volume 15075 of *Lecture Notes in Computer Science*, pp. 370–387. Springer, 2024.
- Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pp. 2818–2829. IEEE, 2023.
- 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*. OpenReview.net, 2025.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP (1)*, pp. 6894–6910. Association for Computational Linguistics, 2021.

- Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, Zhenchong Hu, Bo-Wen Zhang, Jijie Li, Dong Liang, Yingli Zhao, Yulong Ao, Yaoqi Liu, Fangxiang Feng, and Guang Liu. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *CoRR*, abs/2410.18558, 2024.
- Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. Breaking the modality barrier: Universal embedding learning with multimodal llms. *CoRR*, abs/2504.17432, 2025a.
- Tiancheng Gu, Kaicheng Yang, Kaichen Zhang, Xiang An, Ziyong Feng, Yueyi Zhang, Weidong Cai, Jiankang Deng, and Lidong Bing. Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning. *CoRR*, abs/2510.13515, 2025b.
- Wenyi Hong, Weihang Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, Lei Zhao, Zhuoyi Yang, Xiaotao Gu, Xiaohan Zhang, Guanyu Feng, Da Yin, Zihan Wang, Ji Qi, Xixuan Song, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Yuxiao Dong, and Jie Tang. Cogvlm2: Visual language models for image and video understanding. *CoRR*, abs/2408.16500, 2024.
- 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 *ICLR*. OpenReview.net, 2022.
- Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhua Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *ICLR*. OpenReview.net, 2025.
- Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL (1)*, pp. 1601–1611. Association for Computational Linguistics, 2017.
- Yeong-Joon Ju and Seong-Whan Lee. From generator to embedder: Harnessing innate abilities of multimodal llms via building zero-shot discriminative embedding model. *CoRR*, abs/2508.00955, 2025.
- Fanheng Kong, Jingyuan Zhang, Yahui Liu, Hongzhi Zhang, Shi Feng, Xiaocui Yang, Daling Wang, Yu Tian, Victoria W., Fuzheng Zhang, and Guorui Zhou. Modality curation: Building universal embeddings for advanced multimodal information retrieval. *CoRR*, abs/2505.19650, 2025.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466, 2019.
- Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. *CoRR*, abs/2503.04812, 2025.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025, 2025.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a.

- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *CoRR*, abs/2308.03281, 2023b.
- Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1504–1512, 2023c.
- Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal LLMS. In *ICLR*. OpenReview.net, 2025.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- Rui Meng, Ziyang Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhui Chen, and Semih Yavuz. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *CoRR*, abs/2507.04590, 2025.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@NIPS*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.
- Zi-Hao Qiu, Quanqi Hu, Zhuoning Yuan, Denny Zhou, Lijun Zhang, and Tianbao Yang. Not all semantics are created equal: Contrastive self-supervised learning with automatic temperature individualization. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28389–28421. PMLR, 2023.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, pp. 2383–2392. The Association for Computational Linguistics, 2016.
- Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *ICLR*. OpenReview.net, 2021.
- Evgenia Rusak, Patrik Reizinger, Attila Juhas, Oliver Bringmann, Roland S. Zimmermann, and Wieland Brendel. Infonce: Identifying the gap between theory and practice. In *AISTATS*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4159–4167. PMLR, 2025.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.
- Raghuvinder Thirukovalluru, Rui Meng, Ye Liu, Karthikeyan K, Mingyi Su, Ping Nie, Semih Yavuz, Yingbo Zhou, Wenhui Chen, and Bhuwan Dhingra. Breaking the batch barrier (B3) of contrastive learning via smart batch mining. *CoRR*, abs/2505.11293, 2025.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*, pp. 809–819. Association for Computational Linguistics, 2018.

- Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghu-ram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiye Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Ju-yeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Heseng Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Sugan- than, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi- Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yun-Hsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. Embeddinggemma: Powerful and lightweight text representations. *CoRR*, abs/2509.20354, 2025.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In *ACL (1)*, pp. 11897–11916. Association for Computational Linguistics, 2024a.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *CoRR*, abs/2409.12191, 2024b.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yanan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingdong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Bqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhao Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265, 2025.
- Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhao Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *ECCV (87)*, volume 15145 of *Lecture Notes in Computer Science*, pp. 387–404. Springer, 2024.
- Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, pp. 4818–4829. IEEE, 2024.
- Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. *arXiv preprint arXiv:2505.05071*, 2025.
- Youze Xue, Dian Li, and Gang Liu. Improve multi-modal embedding learning via explicit hard negative gradient amplifying. *CoRR*, abs/2506.02020, 2025.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhut- dinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi- hop question answering. In *EMNLP*, pp. 2369–2380. Association for Computational Linguistics, 2018.

- Hao Yu, Zhuokai Zhao, Shen Yan, Lukasz Korycki, Jianyu Wang, Baosheng He, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, and Hanchao Yu. Cafe: Unifying representation and generation with contrastive-autoregressive finetuning. *CoRR*, abs/2503.19900, 2025.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pp. 11941–11952. IEEE, 2023.
- Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhui Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. In *ICML*. OpenReview.net, 2024.
- 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 *CVPR*, pp. 9274–9285. Computer Vision Foundation / IEEE, 2025.
- 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 *ACL (1)*, pp. 19076–19095. Association for Computational Linguistics, 2025.