

# WeMM-Embedding: WeChat Multi-Modal Embedding Technical Report

Junjie Zhou\*, Ke Mei\*<sup>†</sup>, Lei Li\*, Tianyi Wang, Fengyun Rao<sup>‡</sup>, Jing LYU  
WeChat Vision, Tencent Inc.

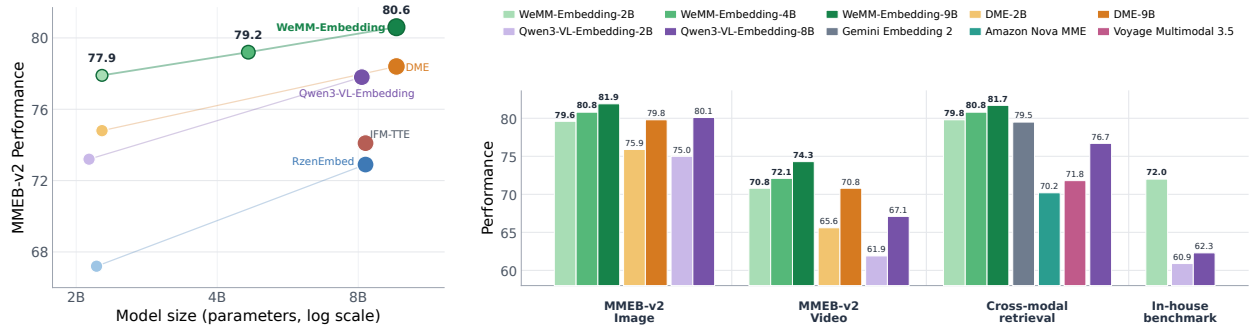
## Abstract

Universal multimodal embeddings are becoming a core component of modern AI systems, enabling heterogeneous content to be represented in a shared space for applications such as retrieval, recommendation, classification, and agentic systems. In this report, we present **WeMM-Embedding**, a family of universal multimodal embedding models supporting text, images, videos, visual documents, and arbitrarily interleaved multimodal inputs with flexible output dimensions. The family comprises 2B, 4B, and 9B variants and is trained in two stages: a large-scale multimodal alignment stage, followed by a refinement stage using curated data, fine-grained relevance supervision, and cross-scale knowledge transfer. Across extensive evaluations, WeMM-Embedding achieves leading performance on multiple public benchmarks. Notably, the 2B variant already surpasses the previously leading 8B open-source baseline on MMEB-v2, while the 9B variant further achieves a new state-of-the-art overall score of 80.6. WeMM-Embedding also demonstrates strong practical performance across WeChat applications, with substantial gains on a 26-task in-house benchmark and consistent improvements across 14 online A/B tests. It has been deployed at scale across recommendation and search applications, including WeChat Channels, Official Accounts, Moments, and e-commerce services. We have released the model weights and code to facilitate future research.

🤖 <https://huggingface.co/collections/tencent/wemm-embedding>

🔗 <https://github.com/Tencent/WeMM-Embedding>

**Date:** August 25, 2026



**Figure 1 Performance and efficiency overview of WeMM-Embedding.** **Left:** MMEB-v2 overall performance across different model sizes, compared with representative baselines. **Right:** Aggregate performance on the MMEB-v2 image and video subsets, a 12-dataset cross-modal retrieval suite, and the 26-task in-house benchmark.

\*Core contributors. <sup>†</sup>Project leader. <sup>‡</sup>Corresponding author: [fengyunrao@tencent.com](mailto:fengyunrao@tencent.com)

## 1 Introduction

Multimodal embedding models have become fundamental components of modern AI systems, owing to their ability to map heterogeneous inputs, such as text, images, and videos, into a shared dense representation space. Such representations support a broad range of downstream applications, including classification [12, 20], any-to-any retrieval [2, 25, 48], clustering [43], recommendation [3, 13], and agentic systems [9, 14, 46]. Early CLIP-style models [33, 37, 49] established the effectiveness of large-scale cross-modal alignment through modality-specific dual encoders. However, their modality-specific encoding pathways do not naturally support the joint representation of inputs that combine multiple modalities, such as interleaved text–image documents, composed multimodal queries, or videos paired with transcripts. Subsequent work incorporated pretrained visual encoders as visual tokenizers for text encoders, extending this paradigm to broader any-to-any matching [51, 54]. Nevertheless, these encoder-based approaches remain limited in generality as multimodal embedding expands toward increasingly diverse tasks and input compositions.

The emergence of Multimodal Large Language Models (MLLMs) [1, 8, 24] has driven rapid progress in universal multimodal embedding models. MLLMs naturally support arbitrary interleaved combinations of text, images, and videos, while their broad pretrained capabilities provide a strong foundation for representation learning across diverse tasks. Motivated by these advantages, early studies explored adapting the hidden states of MLLMs into general-purpose embeddings and demonstrated the feasibility of continued contrastive training on paired multimodal data [19, 23]. More recent advances in large-scale paired multimodal data synthesis and knowledge distillation have further improved model performance, making MLLM-based embedding an increasingly prominent paradigm for universal multimodal representation learning [4, 22, 50, 52].

In this work, we introduce **WeMM-Embedding**, a family of universal multimodal embedding models spanning 2B, 4B, and 9B scales. To build a strong and general-purpose multimodal embedding family, we devise a two-stage training strategy that progressively moves from broad multimodal alignment to finer-grained relevance learning. In the first stage, the models are trained on several hundred million heterogeneous pairs spanning diverse modalities, tasks, and domains, establishing broad multimodal coverage and a strong initial representation space. In the second stage, we further refine the models on a carefully curated corpus with improved semantic balance, higher data quality, and more challenging negatives, while introducing richer relevance supervision and cross-scale knowledge transfer to strengthen fine-grained matching. This progressive training strategy allows WeMM-Embedding to retain broad capability coverage while continuously improving relevance modeling and representation quality.

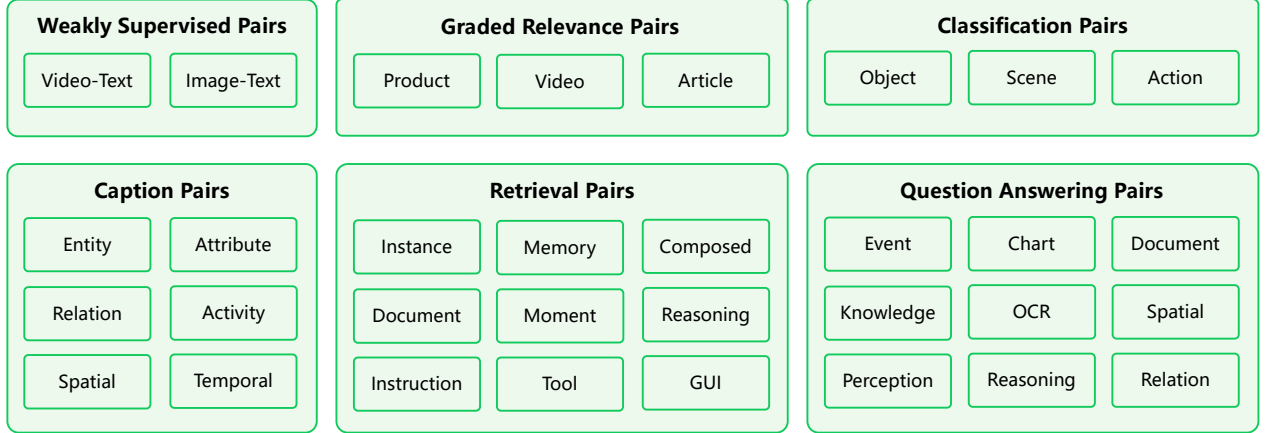
WeMM-Embedding achieves state-of-the-art performance across a broad range of public benchmarks and shows strong practical performance in real-world applications. As summarized in Figure 1, the compact 2B variant already surpasses previously leading 8B open-source baselines on MMEB-v2 [28], highlighting the strong parameter efficiency of WeMM-Embedding. Scaling to 9B further raises the overall score to 80.6, ranking first on the official MMEB-v2 leaderboard and outperforming all listed open-source and proprietary models.<sup>1</sup> On the cross-modal retrieval suite reported in Gemini Embedding 2 [35], the 2B model also compares favorably with leading proprietary models. Beyond public benchmark evaluations, WeMM-Embedding also demonstrates strong practical performance, with substantial gains on a 26-task in-house benchmark and consistent improvements across 14 online A/B tests across WeChat applications. It has also been deployed at scale across recommendation and search applications, including WeChat Channels, Official Accounts, Moments, and e-commerce services. Together, these results establish a new performance–efficiency frontier for universal multimodal embedding models and demonstrate strong practical value in large-scale deployment.

## 2 Data Construction

To support general-purpose multimodal representation learning, we collect and synthesize several hundred million training examples spanning diverse modalities, domains, and tasks. We express these heterogeneous examples in a unified pair-based format, allowing them to be incorporated into a common training framework while preserving their original supervision signals. We further construct a curated collection that complements the large-scale data with greater semantic diversity, higher data quality, and more informative supervision.

---

<sup>1</sup>Leaderboard status as of August 24, 2026, according to the [official MMEB leaderboard](#).



**Figure 2 Overview of our multimodal training data.** Major data families and representative coverage across diverse task settings and content domains.

## 2.1 Large-Scale Training Data Collection

Our training data are collected from public datasets and web-scale weakly supervised sources, and are further supplemented with task-oriented synthetic data and in-house collections. The resulting data span text, images, videos, and their interleaved combinations across diverse domains and task settings.

**Unified Pair-Based Format.** Despite substantial variation in input structures and supervision signals, these heterogeneous tasks can all be formulated as matching a source instance against one or more target candidates. Accordingly, we represent each training example in a unified pair-based format:

$$z_i = (I_i, q_i, c_i, \mathcal{N}_i, y_i), \quad (1)$$

where  $I_i$  denotes an optional task-specific instruction,  $q_i$  denotes the source instance,  $c_i$  denotes the paired target,  $\mathcal{N}_i$  denotes an optional set of explicit hard negatives, and  $y_i$  denotes an optional graded relevance score. Both  $q_i$  and  $c_i$  may contain text, images, videos, or interleaved combinations of these modalities. Each element of  $\mathcal{N}_i$  is drawn from the same target candidate space as  $c_i$  and has the corresponding target-side modality structure. When provided,  $I_i$  specifies the intended matching relation and helps distinguish tasks with similar input modalities. For standard source–target pairs,  $c_i$  serves as the positive target, while both  $\mathcal{N}_i$  and  $y_i$  may be omitted. When explicit hard negatives are available,  $\mathcal{N}_i$  contains candidates that are related to  $q_i$  but less relevant than  $c_i$ . For datasets with graded relevance annotations,  $y_i$  retains the original ordinal or continuous relevance score assigned to the pair  $(q_i, c_i)$  and is subsequently used to construct relative ordering constraints during training.

Under this formulation, examples from different tasks share a common source–target structure and can be organized into task-specific batches within a unified multi-task training pipeline. For standard pairwise data, targets from other examples in the same batch serve as in-batch negatives. For tasks with shared target spaces, such as classification, repeated targets may introduce false negatives; these collisions are excluded through semantic target masking, as detailed in Section 3.2.

**Data Coverage and Composition.** As shown in Figure 2, our training data cover diverse forms of supervision, task settings, and content domains. The training corpus includes several major types of data and supervision, with representative examples illustrated in the figure. We describe the main components below:

- **Weakly Supervised Pairs.** We collect large-scale image–text and video–text pairs from public datasets and web-scale weakly supervised sources, where visual and textual content are associated through

naturally occurring correspondence rather than explicit visual descriptions. These pairs provide broad but relatively coarse supervision across diverse visual content.

- **Caption Pairs.** We pair images and videos with captions that explicitly describe their visual content, ranging from concise summaries to detailed descriptions of entities, attributes, relations, spatial context, actions, and events. Compared with weakly supervised pairs, these examples provide more direct and fine-grained visual–language correspondence.
- **Retrieval Pairs.** Retrieval examples associate queries with relevant candidates across text, images, videos, and interleaved multimodal inputs. They range from conventional unimodal and cross-modal matching to more challenging settings involving composed queries, reasoning, instructions, long contexts, spatial grounding, temporal localization, and agent-related retrieval.
- **Classification Pairs.** We reformulate classification datasets as source–label pairs, where the target is represented by either a class name or a natural-language description of the corresponding category. These examples cover a variety of image- and video-based recognition tasks, including object, scene, and action classification.
- **Multimodal Question–Answer Pairs.** We pair textual or multimodal questions with answer targets across a range of capabilities, including visual perception, relational and spatial understanding, optical character recognition, knowledge-intensive understanding, reasoning, document and chart comprehension, and event understanding.
- **Graded Relevance Pairs.** In the large-scale collection, graded supervision takes the form of manually assigned discrete relevance levels for source–target pairs. These labels distinguish multiple degrees of relevance beyond binary matching and support ranking-oriented training [17]. Reranker-derived relevance scores are introduced separately during the subsequent data curation stage.

## 2.2 Curated Data Construction

Alongside the large-scale collection, we construct a curated dataset approximately one tenth its size. The curated dataset focuses on improving semantic balance and data quality while introducing more informative supervision. Its construction includes Semantic-ID-guided resampling, quality control, and selective hard-negative enrichment.

**Semantic-ID-Guided Resampling.** Although the large-scale collection covers a broad range of domains and tasks, its semantic distribution remains skewed toward frequent patterns. Inspired by Semantic IDs [34], we derive a discrete identifier for each source–target pair to characterize this distribution and guide resampling. For each pair, the side with the longer serialized token sequence is encoded using an intermediate checkpoint of WeMM-Embedding. We then fit a three-level residual k-means quantizer (RQ-KMeans) [26] to the resulting representations. Each representation is subsequently mapped through the learned codebooks to obtain a three-element Semantic ID. We use the assignment density at each codebook level to guide resampling. Examples associated with densely populated codes are sampled at lower rates, whereas those mapped to less populated codes are retained at higher rates. This reduces repeated exposure to frequent semantic patterns without enforcing a uniform distribution over Semantic IDs.

**Data Quality Refinement.** The resampled training examples are further refined using a multimodal large language model. Based on the corresponding task context, the model assesses whether each source–target pair reflects the intended matching relation and filters out mismatched examples. We also refine the textual fields where necessary. For example, weakly supervised image–text and video–text pairs offer broad coverage but may contain noisy or factually inaccurate descriptions; in such cases, the model corrects factual inconsistencies while preserving the original alt-text style and level of detail.

**Hard-Negative Construction.** A subset of the refined training examples is further enriched with explicit hard negatives to improve discrimination among semantically similar candidates. The construction procedure varies with the target modality. For text targets, multimodal large language models generate plausible but

incorrect candidates based on the source and positive target. For image and video targets, intermediate checkpoints of our WeMM-Embedding models are used to retrieve semantically similar candidates from task-specific candidate pools. For a smaller subset of the mined candidate sets, reranking models are further used to assign relevance scores, providing finer-grained supervision among difficult candidates. The corresponding training objective is introduced in Section 3.2.

### 3 Modeling and Training Strategy

WeMM-Embedding comprises three universal multimodal embedding models with 2B, 4B, and 9B parameters, built on the corresponding natively multimodal Qwen3.5 backbones [39]. This section first introduces how heterogeneous multimodal inputs are encoded into dense representations, and then presents the two-stage strategy used to train the model family.

#### 3.1 Modeling

WeMM-Embedding is built upon the Qwen3.5 architecture. Benefiting from its native support for heterogeneous and interleaved multimodal inputs, WeMM-Embedding can encode arbitrary combinations of text, images, and videos into unified dense representations. We adopt last-token pooling by appending a dedicated `<embedding>` token to the input sequence and using its final-layer hidden state as the output representation.

Formally, an input instance may contain an optional task-specific instruction together with an ordered sequence of multimodal segments:

$$\mathcal{D} = \langle I_{\text{inst}}, x_1, x_2, \dots, x_m \rangle, \quad (2)$$

where  $I_{\text{inst}}$  denotes the optional instruction and each  $x_i$  may correspond to text, an image, or a video. Textual segments are converted into token embeddings through the native tokenizer and embedding layer, while visual segments are converted into visual token representations through the native visual processing pipeline. The textual and visual tokens are arranged according to the original order of the input segments, followed by a dedicated `<embedding>` token:

$$\mathbf{S} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N, \mathbf{z}_{\text{emb}}], \quad (3)$$

where each  $\mathbf{z}_i$  denotes a textual or visual token embedding, and  $\mathbf{z}_{\text{emb}}$  denotes the token embedding of `<embedding>`. The resulting multimodal sequence is then processed by the LLM backbone  $G_\theta$ :

$$\mathbf{H} = G_\theta(\mathbf{S}) = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N, \mathbf{h}_{\text{emb}}]. \quad (4)$$

By default, the `<embedding>` token is placed at the end of the sequence and attends to all preceding textual and visual content under the native causal attention mask. The same causal formulation also allows multiple `<embedding>` tokens to be inserted at different sequence positions. For example, when a video is followed by its automatic speech recognition (ASR) transcript, placing one token after the video tokens and another at the end of the sequence enables video-only and joint video-text representations to be extracted within a single forward pass, supporting downstream applications with different modality requirements.

The final-layer hidden state corresponding to each `<embedding>` token is L2-normalized to obtain the output representation:

$$\mathbf{e}_{\mathcal{D}} = \frac{\mathbf{h}_{\text{emb}}}{\|\mathbf{h}_{\text{emb}}\|_2}. \quad (5)$$

In addition, WeMM-Embedding supports flexible embedding dimensions through Matryoshka Representation Learning (MRL) [21]. Given the final hidden state  $\mathbf{h}_{\text{emb}} \in \mathbb{R}^D$ , an embedding of dimension  $d \leq D$  is obtained by retaining its first  $d$  dimensions and applying L2 normalization:

$$\mathbf{e}_{\mathcal{D}}^{(d)} = \frac{\mathbf{h}_{\text{emb},1:d}}{\|\mathbf{h}_{\text{emb},1:d}\|_2}, \quad d \in \mathcal{D}_{\text{MRL}}, \quad (6)$$

where  $\mathcal{D}_{\text{MRL}}$  denotes the predefined set of supported embedding dimensions. At inference time, embeddings at all supported dimensions can be obtained from a single forward pass through prefix truncation and re-normalization.

## 3.2 Training Strategy

We train WeMM-Embedding using a two-stage strategy. The first stage establishes a general multimodal embedding space through large-scale multi-task alignment. The second stage continues training on curated data, combining contrastive learning with explicit hard negatives, selective reranker supervision, and embedding distillation from a larger model. Both stages operate on the unified pair-based representation introduced in Section 2.1, while each batch uses the objective supported by its supervision signals.

### 3.2.1 Stage 1: Large-Scale Multimodal Alignment

In the first stage, we train WeMM-Embedding on the large-scale collection described in Section 2.1, comprising several hundred million source–target pairs across diverse modalities, domains, and tasks. Each batch is constructed from a consistent data source, while batches from different tasks are interleaved throughout training. Standard paired examples are optimized with contrastive learning [30], whereas examples carrying native graded relevance annotations use a score-gap-weighted CoSENT-style ranking objective [16].

**Contrastive Learning.** For standard paired data, we optimize source–target alignment using an InfoNCE objective [30]. Each batch is drawn from a single dataset, keeping the task definition and candidate space consistent within the batch. For each source, its paired target is treated as the positive, while targets associated with other examples serve as in-batch negatives. When explicit hard negatives are available, they are incorporated into the same negative pool. This construction provides informative task-consistent negatives, but may also introduce collisions when different examples contain nearly identical sources or targets. We therefore apply duplicate-aware masking on both sides of the pair before computing the loss.

Let  $\mathcal{C}_j = \{c_j^+\} \cup \mathcal{N}_j$  denote the candidates associated with source  $q_j$ , where  $\mathcal{N}_j$  is empty for ordinary paired examples. Given a batch of  $B$  source–target pairs, the contrastive objective is defined as

$$\mathcal{L}_{\text{CL}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(s(q_i, c_i^+)/\tau)}{\exp(s(q_i, c_i^+)/\tau) + \sum_{j=1}^B \sum_{c \in \mathcal{C}_j \setminus \{c_i^+\}} M_{i,j,c} \exp(s(q_i, c)/\tau)}, \quad (7)$$

where  $s(\cdot, \cdot)$  denotes cosine similarity between normalized embeddings and  $\tau$  is a learnable temperature parameter. The duplicate-aware mask is defined as

$$M_{i,j,c} = \begin{cases} 0, & \text{if } j \neq i \text{ and } s(q_i, q_j) > \tau_{\text{dup}}, \\ & \text{or if } s(c_i^+, c) > \tau_{\text{dup}}, \\ 1, & \text{otherwise,} \end{cases} \quad (8)$$

where  $\tau_{\text{dup}}$  is the similarity threshold used to identify near-duplicate representations. Source-side masking excludes all candidates associated with a near-duplicate source, while target-side masking excludes candidates that closely match the current positive target. The latter applies to both in-batch targets and explicit hard negatives.

**Graded Relevance Learning.** Within the large-scale collection, a subset of source–target pairs is annotated with manually assigned discrete relevance levels. Such annotations are particularly useful for item-to-item retrieval and recommendation-oriented scenarios, where different pairs may exhibit varying degrees of relatedness. Treating these examples as ordinary positive pairs under the contrastive objective would ignore this graded structure. We therefore adopt a score-gap-weighted CoSENT-style objective [16], which optimizes the relative ordering of pairwise similarities and assigns greater importance to comparisons with larger relevance gaps.

Formally, given a graded-relevance batch

$$\mathcal{B}_{\text{rel}} = \{(q_i, c_i, y_i)\}_{i=1}^G, \quad (9)$$

let  $s_i = s(q_i, c_i)$  denote the predicted similarity of the  $i$ -th source–target pair. For two examples  $i$  and  $j$ , we define the relevance-gap weight as

$$w_{ij} = \max(|y_i - y_j|, \epsilon), \quad (10)$$

where  $\epsilon$  is a small positive constant. The ranking loss associated with example  $i$  is

$$\begin{aligned} \mathcal{L}_i^{\text{Rel}} = \log & \left[ 1 + \sum_{j: y_i > y_j} w_{ij} \exp(\gamma[s_j - s_i]) \right. \\ & \left. + \sum_{j: y_j > y_i} w_{ij} \exp(\gamma[s_i - s_j]) \right], \end{aligned} \quad (11)$$

and the batch objective is

$$\mathcal{L}_{\text{Rel}} = \frac{1}{G} \sum_{i=1}^G \mathcal{L}_i^{\text{Rel}}. \quad (12)$$

This objective encourages pairs with higher relevance levels to receive higher similarities and gives greater weight to comparisons with larger label gaps. Pairs with equal labels impose no ordering constraint.

**Matryoshka Representation Learning.** We apply Matryoshka Representation Learning [21] to both contrastive and graded-relevance training. For each batch, the objective associated with its supervision is evaluated independently at every supported embedding dimension:

$$\mathcal{L}_X^{\text{MRL}} = \sum_{d \in \mathcal{D}_{\text{MRL}}} \alpha_d \mathcal{L}_X^{(d)}, \quad X \in \{\text{CL}, \text{Rel}\}, \quad (13)$$

where  $\mathcal{L}_X^{(d)}$  is computed using embeddings independently truncated and normalized at dimension  $d$ , and  $\alpha_d$  denotes the corresponding loss weight. Thus, each interleaved batch is optimized using the multi-dimensional form of its corresponding task objective.

### 3.2.2 Stage 2: Curated Fine-Tuning and Distillation

Following large-scale multimodal alignment, WeMM-Embedding is further trained on the curated dataset described in Section 2.2. The contrastive and graded-relevance objectives remain in use, while additional supervision is provided by reranking models and a larger embedding teacher. Reranking models provide query-specific ordering signals over mined candidates for selected tasks, whereas the embedding teacher transfers the similarity structure induced within each training batch.



**Reranker Supervision.** We train dedicated multimodal rerankers to provide fine-grained ordering supervision over query-specific candidate sets, each containing an annotated positive target and mined hard negatives. Building on the score-gap-weighted CoSENT formulation used for graded relevance learning in Section 3.2.1, we replace the manually assigned relevance levels with reranker scores and construct comparisons only among candidates associated with the same query. Specifically, for source  $q_b$  with candidate set  $\mathcal{C}_b = \{c_{b,1}, \dots, c_{b,K}\}$ , the reranker scores  $\hat{y}_{b,k}$  induce the ordered pair set

$$\hat{\mathcal{P}}_b = \{(i, j) \mid \hat{y}_{b,i} > \hat{y}_{b,j}\}. \quad (14)$$

The corresponding objective is

$$\mathcal{L}_{\text{Rank}} = \frac{1}{B} \sum_{b=1}^B \log \left[ 1 + \sum_{(i,j) \in \hat{\mathcal{P}}_b} \omega_{b,ij} \exp(\gamma[s(q_b, c_{b,j}) - s(q_b, c_{b,i})]) \right], \quad (15)$$

where  $\omega_{b,ij} = \max(|\hat{y}_{b,i} - \hat{y}_{b,j}|, \epsilon)$  weights each comparison according to the reranker score gap. Reranker-scored batches use  $\mathcal{L}_{\text{Rank}}$  in place of the standard contrastive objective.

Empirically, we find that reranking candidates retrieved by our embedding models does not consistently improve performance across multimodal tasks. Similar to observations reported in prior work [22], stable gains are observed only on a limited subset of tasks. We therefore restrict reranker supervision to settings where it yields reliable improvements.

**Embedding Distillation.** To obtain teacher supervision with broader coverage, we additionally distill from a larger model in the same WeMM-Embedding family. Unlike reranker supervision, which relies on task-specific models and preconstructed candidate sets, embedding distillation derives online soft targets from the batch-wise source-target similarity distributions produced by the teacher. It therefore requires no additional offline relevance annotations and can be applied across heterogeneous batch types.

Formally, for a batch containing  $B$  sources and a target pool of size  $K$ , let  $\mathbf{A}^T, \mathbf{A}^S \in \mathbb{R}^{B \times K}$  denote the source-to-target similarity matrices produced by the teacher and student:

$$A_{ij}^T = \frac{s_T(q_i, c_j)}{\tau_T}, \quad A_{ij}^S = \frac{s_S(q_i, c_j)}{\tau_S}, \quad (16)$$

where  $\tau_T$  and  $\tau_S$  are the teacher and student temperatures. We further construct the reverse similarity matrices between the positive targets and the sources:

$$\bar{A}_{ij}^T = \frac{s_T(c_i^+, q_j)}{\tau_T}, \quad \bar{A}_{ij}^S = \frac{s_S(c_i^+, q_j)}{\tau_S}. \quad (17)$$

The corresponding row-wise relation distributions are

$$\mathbf{P}_{T,i}^{q \rightarrow c} = \text{Softmax}(\mathbf{A}_{i,:}^T), \quad \mathbf{P}_{S,i}^{q \rightarrow c} = \text{Softmax}(\mathbf{A}_{i,:}^S), \quad (18)$$

$$\mathbf{P}_{T,i}^{c \rightarrow q} = \text{Softmax}(\bar{\mathbf{A}}_{i,:}^T), \quad \mathbf{P}_{S,i}^{c \rightarrow q} = \text{Softmax}(\bar{\mathbf{A}}_{i,:}^S). \quad (19)$$

We define the bidirectional embedding-distillation objective as

$$\mathcal{L}_{\text{Emb}} = \frac{1}{2B} \sum_{i=1}^B \left[ D_{\text{KL}}(\mathbf{P}_{T,i}^{q \rightarrow c} \parallel \mathbf{P}_{S,i}^{q \rightarrow c}) + D_{\text{KL}}(\mathbf{P}_{T,i}^{c \rightarrow q} \parallel \mathbf{P}_{S,i}^{c \rightarrow q}) \right]. \quad (20)$$



The two terms align the teacher and student similarity distributions in the source-to-target and target-to-source directions, respectively. Unlike the one-hot supervision used in standard contrastive learning, the teacher distributions retain relative similarity differences among candidates, providing softer and more structured targets for transferring semantic relations. Empirically, this supervision is particularly valuable for our compact WeMM-Embedding variants and contributes substantially to their performance gains.

**Training Configuration.** For each Stage 2 batch, the task objective is selected according to its available supervision:

$$\mathcal{L}_{\text{Task}} = \begin{cases} \mathcal{L}_{\text{CL}}^{\text{MRL}}, & \text{for standard paired or hard-negative batches,} \\ \mathcal{L}_{\text{Rel}}^{\text{MRL}}, & \text{for graded-relevance batches,} \\ \mathcal{L}_{\text{Rank}}^{\text{MRL}}, & \text{for reranker-scored batches.} \end{cases} \quad (21)$$

Reranker-scored batches use the ranking objective in place of the standard contrastive objective. Embedding distillation, by contrast, is added to the selected task objective whenever a larger embedding teacher is available. The overall Stage 2 objective is therefore

$$\mathcal{L}_{\text{Stage2}} = \mathcal{L}_{\text{Task}} + \lambda_{\text{Emb}} \mathcal{L}_{\text{Emb}}, \quad (22)$$

where  $\lambda_{\text{Emb}}$  denotes the weight assigned to the embedding-distillation loss.

For the 2B and 4B variants, the frozen 9B WeMM-Embedding model serves as the embedding teacher during Stage 2. Since no larger embedding teacher is available for the 9B variant, we instead train multiple specialized Stage 2 variants using complementary data mixtures and training configurations, and combine them through model merging [47] to obtain the final 9B model.

## 4 Experiments and Analysis

We evaluate WeMM-Embedding across a broad range of public and in-house benchmarks covering image, video, visual-document, text, and agent-oriented retrieval. Our experiments cover the MMEB benchmark series [15, 28], a collection of widely used cross-modal retrieval benchmarks reported in the Gemini Embedding 2 study [35], and an in-house benchmark comprising 26 tasks derived from real-world applications within WeChat. We compare WeMM-Embedding with representative state-of-the-art multimodal embedding models, including both open-source and proprietary commercial models.

### 4.1 Performance on the MMEB Series

The MMEB series provides a comprehensive evaluation across diverse modalities, task formulations, and retrieval scenarios. Specifically, MMEB-v2 [28] evaluates general multimodal embedding capabilities across images, videos, and visual documents, while MMEB-v3 [15] further extends the evaluation to audio, complex text retrieval, and agent-centric scenarios. In this section, we follow the benchmark-defined evaluation metrics, using Hit@1 for image, video, audio, and agent tasks and NDCG@5 for text and visual-document tasks.

#### 4.1.1 General Multimodal Representation

We first report the performance of WeMM-Embedding on MMEB-v2, which comprises 78 datasets spanning images, videos, and visual documents and covers diverse tasks including classification, question answering, retrieval, and visual grounding [28]. As shown in Table 1, WeMM-Embedding-2B achieves an overall score of 77.9, outperforming Qwen3-VL-Embedding-2B [22] and DME-2B [6] by 4.7 and 3.1 points, respectively, and slightly surpassing Qwen3-VL-Embedding-8B. The 4B variant further improves the score to 79.2, outperforming all compared 8B–9B baselines. Scaling to 9B further raises the overall score to 80.6, ranking first on the official MMEB-v2 leaderboard and outperforming all listed open-source and proprietary models.

| Model                           | Size      | AVG         | Image       |             |             |             |             | Video       |             |             |             |             | VisDoc      |
|---------------------------------|-----------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                                 |           |             | Overall     | CLS         | QA          | Ret         | GD          | Overall     | CLS         | QA          | V-Ret       | M-Ret       | Overall     |
| # datasets →                    |           | 78          | 36          | 10          | 10          | 12          | 4           | 18          | 5           | 5           | 5           | 3           | 24          |
| VLM2Vec [19]                    | 2B        | 47.8        | 59.7        | 58.7        | 49.3        | 65.0        | 72.9        | 29.0        | 33.4        | 30.5        | 20.6        | 33.0        | 44.0        |
| GME [50]                        | 2B        | 55.4        | 51.9        | 54.4        | 29.9        | 66.9        | 55.5        | 33.9        | 34.9        | 42.0        | 25.6        | 32.4        | 76.8        |
| VLM2Vec-V2 [28]                 | 2B        | 59.3        | 64.9        | 62.9        | 56.3        | 69.5        | 77.3        | 34.9        | 39.3        | 34.3        | 28.8        | 38.5        | 69.2        |
| Ops-MM-embedding-v1             | 2B        | 64.6        | 69.0        | 68.1        | 65.1        | 69.2        | 80.9        | 47.6        | 53.6        | 55.7        | 41.8        | 33.7        | 70.8        |
| RzenEmbed [18]                  | 2B        | 67.2        | 72.3        | 68.5        | 66.3        | 74.5        | 90.3        | 47.3        | 50.4        | 49.7        | 46.6        | 38.9        | 74.5        |
| Qwen3-VL-Embedding [22]         | 2B        | 73.2        | 75.0        | 70.3        | 74.3        | 74.9        | 88.6        | 61.9        | 71.9        | 64.9        | 53.9        | 53.3        | 79.2        |
| DME-Small [6] <sup>†</sup>      | 2B        | 74.8        | 75.9        | 68.6        | 76.2        | 75.6        | 94.3        | 65.6        | 84.5        | 61.9        | 55.5        | 57.4        | 79.9        |
| <b>WeMM-Embedding</b>           | <b>2B</b> | <b>77.9</b> | <b>79.6</b> | <b>74.7</b> | <b>79.2</b> | <b>79.3</b> | <b>94.0</b> | <b>70.8</b> | <b>84.9</b> | <b>71.7</b> | <b>63.4</b> | <b>58.4</b> | <b>80.7</b> |
| <b>WeMM-Embedding</b>           | <b>4B</b> | <b>79.2</b> | <b>80.8</b> | <b>75.1</b> | <b>81.0</b> | <b>80.8</b> | <b>95.0</b> | <b>72.1</b> | <b>86.8</b> | <b>72.5</b> | <b>65.2</b> | <b>58.6</b> | <b>82.0</b> |
| VLM2Vec [19]                    | 8B        | 53.2        | 65.5        | 62.7        | 56.9        | 69.4        | 82.2        | 34.0        | 39.1        | 30.0        | 29.0        | 40.6        | 49.1        |
| GME [50]                        | 8B        | 59.2        | 56.0        | 57.7        | 34.7        | 71.2        | 59.3        | 38.6        | 37.4        | 50.4        | 28.4        | 38.2        | 79.3        |
| Ops-MM-embedding-v1             | 8B        | 68.9        | 72.7        | 69.7        | 69.6        | 73.1        | 87.2        | 53.8        | 59.7        | 62.2        | 45.7        | 43.2        | 74.4        |
| RzenEmbed [18]                  | 8B        | 72.9        | 75.9        | 70.6        | 71.7        | 78.5        | 92.1        | 55.7        | 58.8        | 63.5        | 51.0        | 45.5        | 81.3        |
| IFM-TTE [11]                    | 8B        | 74.1        | 77.9        | 76.7        | 78.5        | 74.6        | 89.3        | 59.2        | 60.5        | 67.9        | 51.7        | 54.9        | 79.5        |
| Qwen3-VL-Embedding [22]         | 8B        | 77.8        | <u>80.1</u> | 74.2        | 81.1        | 80.2        | 92.3        | 67.1        | 78.4        | 71.0        | 58.7        | 56.1        | <u>82.4</u> |
| DME-Medium [6] <sup>†</sup>     | 9B        | <u>78.4</u> | 79.8        | 74.5        | 80.9        | 78.2        | 94.6        | <u>70.8</u> | 87.7        | 71.0        | 61.0        | 58.5        | <u>82.0</u> |
| <b>WeMM-Embedding</b>           | <b>9B</b> | <b>80.6</b> | <b>81.9</b> | <b>76.2</b> | <b>82.2</b> | <b>81.7</b> | <b>95.6</b> | <b>74.3</b> | <b>87.4</b> | <b>77.7</b> | <b>67.4</b> | <b>58.5</b> | <b>83.3</b> |
| Seed-1.6-Embedding <sup>*</sup> | —         | 76.9        | 78.0        | 75.1        | 74.9        | 79.3        | 89.1        | 67.7        | 85.2        | 66.7        | 59.1        | 54.8        | 82.2        |
| QQMM-embed-v4 <sup>†</sup>      | —         | 78.3        | 82.0        | 76.8        | 83.3        | 80.7        | 96.1        | 67.4        | 77.8        | 71.9        | 60.0        | 54.9        | 80.9        |
| Octen-VL-Large <sup>*</sup>     | —         | 80.1        | 81.9        | 75.7        | 84.5        | 80.6        | 94.4        | 76.0        | 87.1        | 82.1        | 68.0        | 60.4        | 80.5        |
| DME-Large [6] <sup>†</sup>      | —         | 80.2        | 81.1        | 75.4        | 82.9        | 79.6        | 95.2        | 74.4        | 90.1        | 75.7        | 64.5        | 62.2        | 83.4        |

**Table 1** Benchmarking results on MMEB-v2. Baseline results are taken from the [official MMEB leaderboard](#). <sup>†</sup>Closed-source leaderboard submission without publicly released model weights or a public inference endpoint. <sup>\*</sup>Proprietary commercial model with an undisclosed parameter count. CLS: classification; QA: question answering; Ret: retrieval; GD: visual grounding; V-Ret: video retrieval; M-Ret: moment retrieval.

#### 4.1.2 Text and Agent-Centric Tasks

We further evaluate WeMM-Embedding on MMEB-v3 [15], which extends the benchmark with complex text retrieval, agent-centric tasks, audio evaluation, and the MCMR image-retrieval task. Among the newly added tasks, 53 text tasks cover reasoning retrieval, instruction following, long-context retrieval, multi-condition retrieval, and general retrieval, while 47 agent tasks cover tool, GUI, and memory retrieval. When computing V3-All, unsupported tasks are assigned a score of zero; accordingly, the 11 audio tasks are scored as zero for the current WeMM-Embedding models, which do not support audio input. Baseline per-task results on the newly added tasks are obtained from the MMEB-v3 paper or score files submitted to the official MMEB leaderboard.<sup>2</sup> For VLM2Vec [19], VLM2Vec-V2 [28], GME [50], and Qwen3-VL-Embedding [22], we recompute V3-All by combining their originally reported MMEB-v2 results with the corresponding results on the newly added MMEB-v3 tasks.

As shown in Table 2, WeMM-Embedding-2B already outperforms all compared baselines on MMEB-v3, achieving 56.0 on V3-All. It also attains the highest scores among all baselines on the Text and Agent task groups, with 45.3 and 45.1, respectively. The 4B and 9B variants achieve V3-All scores of 58.2 and 59.5, respectively, further widening the gap over existing models.

## 4.2 Cross-Modal Retrieval Evaluation

The cross-modal evaluation comprises 12 widely used public benchmarks that measure semantic alignment between text and visual content across images, videos, and visual documents. For the three proprietary models, Gemini Embedding 2 [35], Amazon Nova MME [32], and Voyage Multimodal 3.5 [40], we use the results reported in the Gemini Embedding 2 technical report [35]. Qwen3-VL-Embedding and WeMM-Embedding are evaluated by us on the corresponding public evaluation sets using the metric specified for each benchmark.

<sup>2</sup>Leaderboard results are based on submissions available as of August 24, 2026.

| Model                    | Size      | V3-All      | Text        |      |      |      |      |      | Agent       |      |      |        | MCMR        | Audio       |
|--------------------------|-----------|-------------|-------------|------|------|------|------|------|-------------|------|------|--------|-------------|-------------|
|                          |           |             | Overall     | RR   | IF   | LC   | MC   | GR   | Overall     | Tool | GUI  | Memory |             |             |
| # tasks →                |           | 190         | 53          | 20   | 9    | 6    | 5    | 13   | 47          | 35   | 8    | 4      | 1           | 11          |
| VLM2Vec-V2 [28]          | 2B        | 38.3        | 24.5        | 7.8  | 29.2 | 11.5 | 50.3 | 41.2 | 28.7        | 27.6 | 36.2 | 23.3   | 4.1         | 0.0         |
| Omni-Embed-Nemotron [45] | 3B        | 43.5        | 39.2        | 17.2 | 42.8 | 40.0 | 69.7 | 56.9 | 36.5        | 38.1 | 32.0 | 32.2   | 26.1        | <b>36.5</b> |
| E5-Omni [5]              | 3B        | 44.6        | 26.7        | 11.5 | 44.4 | 8.3  | 60.7 | 34.4 | 36.9        | 37.7 | 35.6 | 32.3   | 31.9        | <u>30.8</u> |
| Qwen3-VL-Embedding [22]  | 2B        | 50.9        | 39.2        | 16.6 | 40.6 | 53.1 | 61.2 | 58.0 | 39.3        | 42.6 | 30.4 | 28.4   | 42.0        | 0.0         |
| <b>WeMM-Embedding</b>    | <b>2B</b> | <u>56.0</u> | <u>45.3</u> | 26.3 | 49.3 | 61.9 | 59.7 | 58.7 | <u>45.1</u> | 46.2 | 38.9 | 47.2   | <b>42.5</b> | 0.0         |
| <b>WeMM-Embedding</b>    | <b>4B</b> | <b>58.2</b> | <b>47.9</b> | 30.0 | 52.0 | 62.4 | 60.1 | 61.1 | <b>49.0</b> | 50.6 | 41.5 | 50.4   | <u>41.9</u> | 0.0         |
| WAVE [38]                | 7B        | 26.3        | 13.7        | 5.9  | 31.3 | 2.6  | 22.9 | 18.6 | 11.3        | 11.9 | 11.8 | 5.7    | 8.9         | 31.8        |
| VLM2Vec [19]             | 8B        | 32.9        | 22.2        | 7.2  | 28.1 | 5.9  | 40.9 | 41.7 | 19.7        | 19.8 | 21.4 | 15.9   | 0.9         | 0.0         |
| LCO-Embedding-Omni [42]  | 7B        | 40.6        | 32.4        | 9.5  | 52.0 | 12.6 | 61.4 | 52.1 | 27.8        | 29.0 | 25.0 | 23.0   | 20.0        | <b>43.2</b> |
| GME [50]                 | 8B        | 43.6        | 37.1        | 12.5 | 52.4 | 17.8 | 59.0 | 62.5 | 35.6        | 39.0 | 30.0 | 17.1   | 27.3        | 0.0         |
| E5-Omni [5]              | 7B        | 47.1        | 26.9        | 11.1 | 42.1 | 9.7  | 56.9 | 35.9 | 36.7        | 37.4 | 38.0 | 27.3   | 41.1        | <u>43.0</u> |
| Tianmu-Emb-Uni           | 8B        | 53.3        | <u>43.6</u> | 20.7 | 47.1 | 58.5 | 63.4 | 62.0 | <u>39.4</u> | 42.2 | 35.6 | 22.6   | 38.8        | 38.9        |
| Qwen3-VL-Embedding [22]  | 8B        | <u>53.5</u> | 42.5        | 18.2 | 44.8 | 58.0 | 61.2 | 61.2 | 38.4        | 41.3 | 33.5 | 22.8   | 38.0        | 0.0         |
| <b>WeMM-Embedding</b>    | <b>9B</b> | <b>59.5</b> | <b>48.8</b> | 31.8 | 52.8 | 64.5 | 58.3 | 61.1 | <b>51.0</b> | 53.0 | 43.3 | 48.9   | <b>49.3</b> | 0.0         |

**Table 2** Evaluation results on MMEB-v3. V3-All averages all 190 tasks, comprising the 78 MMEB-v2 tasks, 53 Text tasks, 47 Agent tasks, 11 Audio tasks, and MCMR [10]. Following the evaluation protocol defined in the MMEB-v3 paper [15], Text results are reported using NDCG@5. Unsupported tasks are assigned a score of zero. RR: reasoning retrieval; IF: instruction following; LC: long-context retrieval; MC: multi-condition retrieval; GR: general retrieval.

As shown in Table 3, WeMM-Embedding-2B achieves an overall score of 79.8 across the 12 tasks, outperforming all compared open-source baselines and comparing favorably with leading proprietary models. Scaling the model to 4B and 9B further improves the overall score to 80.8 and 81.7, respectively.

|                               |                | Gemini<br>Embedding 2 <sup>†</sup> | Amazon<br>Nova MME <sup>†</sup> | Voyage<br>Multimodal 3.5 <sup>†</sup> | Qwen3-VL-<br>Embedding | WeMM-<br>Embedding | WeMM-<br>Embedding | WeMM-<br>Embedding |
|-------------------------------|----------------|------------------------------------|---------------------------------|---------------------------------------|------------------------|--------------------|--------------------|--------------------|
| Size                          |                | –                                  | –                               | –                                     | 8B                     | 2B                 | 4B                 | 9B                 |
| Text → Image<br>Recall@1      | MSCOCO [7]     | 62.9                               | 57.2                            | 58.1                                  | 58.7                   | 62.6               | 63.5               | 64.6               |
|                               | Flickr30k [31] | 89.1                               | 81.6                            | 89.9                                  | 84.6                   | 86.9               | 87.0               | 87.8               |
|                               | DOCCI [29]     | 93.4                               | 84.0                            | 83.8                                  | 95.0                   | 94.0               | 95.8               | 96.7               |
|                               | TextCaps [36]  | 89.6                               | 76.0                            | 79.4                                  | 79.3                   | 82.5               | 85.8               | 85.2               |
|                               | <b>Mean</b>    | <b>83.8</b>                        | 74.7                            | 77.8                                  | 79.4                   | 81.5               | 83.0               | <u>83.6</u>        |
| Image → Text<br>Recall@1      | MSCOCO [7]     | 78.8                               | 68.3                            | 74.5                                  | 71.1                   | 78.8               | 79.4               | 79.3               |
|                               | Flickr30k [31] | 97.4                               | 87.5                            | 94.5                                  | 90.7                   | 95.6               | 96.6               | 94.4               |
|                               | DOCCI [29]     | 91.3                               | 76.5                            | 77.4                                  | 92.8                   | 90.6               | 93.7               | 92.8               |
|                               | TextCaps [36]  | 97.4                               | 88.9                            | 88.6                                  | 91.9                   | 93.3               | 96.0               | 94.8               |
|                               | <b>Mean</b>    | 91.2                               | 80.3                            | 83.8                                  | 86.6                   | 89.6               | <u>91.4</u>        | <b>90.3</b>        |
| Text → Video<br>NDCG@10       | VATEX [41]     | 68.8                               | 60.3                            | 55.2                                  | 68.5                   | 74.6               | 75.9               | 78.4               |
|                               | MSR-VTT [44]   | 68.0                               | 67.0                            | 63.0                                  | 70.6                   | 72.7               | 71.7               | 74.1               |
|                               | YouCook2 [53]  | 52.5                               | 34.7                            | 31.4                                  | 54.7                   | 63.9               | 62.2               | 66.3               |
|                               | <b>Mean</b>    | 63.1                               | 54.0                            | 49.9                                  | 64.6                   | <u>70.4</u>        | 69.9               | <b>72.9</b>        |
| Document Retrieval<br>NDCG@10 |                | ViDoRe V2 [27]                     | 64.9                            | 60.6                                  | <u>65.5</u>            | 62.8               | 61.4               | 62.1               |
| <b>AVG</b>                    |                | 79.5                               | 70.2                            | 71.8                                  | 76.7                   | 79.8               | <u>80.8</u>        | <b>81.7</b>        |

**Table 3** Cross-modal retrieval results on 12 public benchmarks. AVG denotes the average across the 12 datasets. <sup>†</sup>Proprietary commercial model with an undisclosed parameter count.

### 4.3 In-House Evaluation

We further evaluate WeMM-Embedding on an in-house benchmark comprising 26 tasks derived from real-world applications within WeChat. The benchmark covers five categories: classification, search, cross-domain content matching, article relevance, and video relevance. As shown in Table 4, WeMM-Embedding-2B substantially outperforms the representative open-source baseline and achieves higher scores across all five categories.

Beyond offline evaluation, WeMM-Embedding has been deployed at scale in recommendation systems spanning

| Model                   | Size      | AVG         | Classification | Search      | Cross-DM    | Article Rel. | Video Rel.  |
|-------------------------|-----------|-------------|----------------|-------------|-------------|--------------|-------------|
| # tasks →               |           | 26          | 4              | 7           | 4           | 4            | 7           |
| Qwen3-VL-Embedding [22] | 2B        | 60.9        | 60.1           | 51.9        | 55.5        | 75.5         | 64.9        |
| <b>WeMM-Embedding</b>   | <b>2B</b> | <b>72.0</b> | <b>72.6</b>    | <b>65.5</b> | <b>73.7</b> | <b>86.4</b>  | <b>68.8</b> |

**Table 4** Evaluation results on the in-house benchmark. AVG denotes the average over all 26 tasks. Cross-DM denotes cross-domain content matching; Article Rel. denotes article relevance; Video Rel. denotes video relevance.

WeChat Official Accounts, WeChat Channels, and e-commerce content. Its multimodal representations combine complementary signals from text, cover images, and video frames, and are used at multiple stages of the recommendation pipeline, including candidate retrieval, ranking feature construction, user sequence modeling, and cross-domain content understanding. Semantic IDs derived from these representations further provide compact discrete features for indexing and sequence modeling. To date, WeMM-Embedding has delivered consistent gains in 14 online A/B tests across these systems, with the corresponding improvements subsequently rolled out in production. These deployments have improved content matching, recommendation quality, and user engagement, with notable gains for long-tail and newly published content.

WeMM-Embedding has also been deployed in WeChat search, supporting semantic retrieval across diverse content sources including WeChat Channels videos, WeChat Official Accounts articles, and WeChat Moments. The model supports both unimodal and cross-modal retrieval and improves semantic relevance and retrieval quality across text, image, and video content. Taken together, the in-house evaluation and production deployments show that the strong performance of WeMM-Embedding extends beyond public benchmarks to diverse real-world tasks and large-scale recommendation and search systems.

## 4.4 Further Analysis

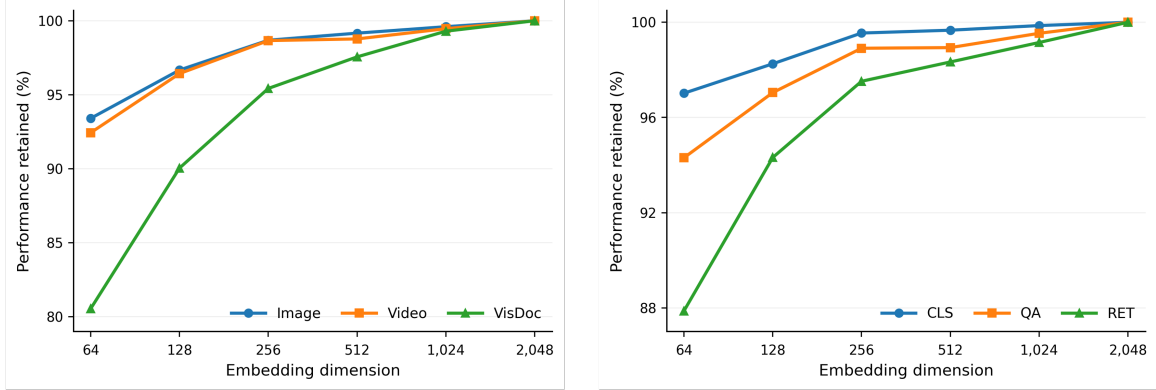
In this section, we first analyze the performance of the learned Matryoshka representations across different embedding dimensions, then revisit key Stage-1 design choices, and finally present a cumulative analysis of the strategies adopted for Stage-2 training.

### 4.4.1 MRL Analysis

WeMM-Embedding incorporates Matryoshka Representation Learning (MRL) [21], allowing a single model to produce nested representations at multiple embedding dimensions. We evaluate the 2B model at dimensions ranging from 64 to 2,048 on the corresponding task subsets of MMEB-v2 to examine how dimensionality reduction affects different modalities and task types. For each evaluation group, we report the proportion of performance retained relative to its result at 2,048 dimensions.

As shown in the left panel of Figure 3, performance on image and video tasks follows closely matched trends as the embedding dimension decreases. At 256 dimensions, the model retains 98.7% of its 2,048-dimensional performance on both image and video tasks; the corresponding retention rates rise to 99.2% and 98.8% at 512 dimensions, respectively. Visual-document tasks exhibit greater sensitivity to dimensionality reduction, which may reflect the higher information density of visual documents, particularly their dense textual content.

The right panel of Figure 3 further shows different sensitivity patterns across task types. Classification is the least sensitive to dimensionality reduction, question answering exhibits moderate degradation, and retrieval is affected most substantially, especially at 64 and 128 dimensions. Once the embedding dimension reaches 256, however, all three task groups retain more than 97% of their respective full-dimensional performance, and the gains from further increasing the dimension become progressively smaller. Taken together, these results suggest that 256- or 512-dimensional representations provide practical choices for efficiency-sensitive applications while preserving most of the performance achieved at 2,048 dimensions.



**Figure 3 MRL analysis of WeMM-Embedding-2B on MMEB-v2. Left:** Performance retained on the image, video, and visual-document subsets across embedding dimensions. **Right:** Performance retained for classification (CLS), question answering (QA), and retrieval (RET), each averaged over the corresponding image and video tasks.

#### 4.4.2 Stage-1 Design Analysis

We conduct a small-scale ablation study with a 2B model to examine some of the key Stage-1 design choices, including task-specific instructions, task-consistent batching, and duplicate-aware masking. The resulting variants are evaluated on MMEB-v2 under a common experimental protocol.

| Stage-1 Configuration                 | AVG         | Image       | Video       | VisDoc      |
|---------------------------------------|-------------|-------------|-------------|-------------|
| <b>Full configuration</b>             | <b>71.9</b> | <b>76.1</b> | <b>59.1</b> | <b>75.3</b> |
| <i>w/o</i> task-specific instructions | 71.1        | 75.7        | 58.2        | 73.8        |
| <i>w/o</i> task-consistent batching   | 68.5        | 71.5        | 56.3        | 73.1        |
| <i>w/o</i> duplicate-aware masking    | 71.4        | 75.3        | 58.6        | 75.1        |

**Table 5** Results on MMEB-v2 from a small-scale Stage-1 ablation study using a 2B model. AVG denotes the average across all 78 MMEB-v2 tasks.

As shown in Table 5, removing task-specific instructions leads to a 0.8-point decrease in the overall score, with the largest degradation observed on visual-document tasks. This result indicates that explicit instructions help the model capture task-specific matching objectives across heterogeneous tasks. Task-consistent batching has the largest impact among the evaluated designs. Replacing it with mixed sampling results in a 3.4-point decrease in overall performance, suggesting that batches constructed from a consistent task and candidate space provide more informative in-batch negatives. Duplicate-aware masking also contributes to performance, as its removal leads to a 0.5-point decrease in the overall score. By filtering duplicate or semantically equivalent targets from the negative set, it makes the unified contrastive objective better suited to tasks with repeated labels, such as classification.

#### 4.4.3 Cumulative Stage-2 Analysis

We further conduct a cumulative study with the 2B model to examine several key strategies used in Stage-2 training, including curated data, reranker supervision, embedding-teacher distillation, and an expanded visual input budget. Starting from the Stage-1 checkpoint, these strategies are introduced sequentially, and each resulting configuration is evaluated on MMEB-v2.

As shown in Table 6, cumulatively incorporating our Stage-2 strategies improves the overall score by 2.2 points. First, curated data yields clear gains on both image and video tasks, highlighting the value of balanced resampling, quality refinement, and hard-negative enrichment when constructing a compact Stage-2 training set from the large-scale multimodal corpus. Building on this, reranker supervision provides an additional improvement, with its largest benefit observed on visual-document tasks. Embedding-teacher distillation

| Cumulative Configuration                      | AVG         | Image       | Video       | VisDoc      |
|-----------------------------------------------|-------------|-------------|-------------|-------------|
| Stage-1 checkpoint                            | 75.7        | 77.7        | 67.5        | 78.9        |
| + curated data                                | 76.6        | 78.9        | 69.2        | 78.5        |
| + reranker supervision                        | 76.7        | 78.8        | 69.4        | 79.1        |
| + embedding-teacher distillation              | 77.6        | 79.4        | 70.0        | 80.4        |
| <b>+ expanded visual input budget (final)</b> | <b>77.9</b> | <b>79.6</b> | <b>70.8</b> | <b>80.7</b> |

**Table 6** Cumulative Stage-2 results of the 2B model on MMEB-v2. Each row reports the configuration after introducing the listed strategy. AVG denotes the average across all 78 MMEB-v2 tasks.

further improves performance across all three domains, demonstrating that dense similarity supervision transfers effectively across heterogeneous tasks and contributes to the strong performance of the compact model. Finally, expanding the visual input budget through higher-resolution inputs and denser video frame sampling brings a further improvement, particularly on video tasks. Overall, these substantial gains suggest that carefully designed supervision signals remain essential for further refining universal multimodal embedding models once large-scale multimodal alignment has been established.

## 5 Conclusion and Future Work

In this report, we introduced WeMM-Embedding, a family of universal multimodal embedding models spanning multiple scales. The family is trained with a two-stage strategy that progresses from broad multimodal alignment to finer-grained representation learning through curated data, richer relevance supervision, and cross-scale knowledge transfer. Extensive evaluations across public benchmarks demonstrate that WeMM-Embedding achieves state-of-the-art performance and establishes a new performance–efficiency frontier. Moreover, it delivers consistent gains on the 26-task in-house benchmark and across 14 online A/B tests. WeMM-Embedding has also been deployed at scale across multiple recommendation and search systems within WeChat, further demonstrating its practical value. In the future, we will continue to extend WeMM-Embedding toward omni-modal inputs, scale the model family to larger sizes, and further improve data curation and fine-grained relevance supervision.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. [arXiv preprint arXiv:2303.08774](#), 2023.
- [2] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16495–16504, 2022.
- [3] Ben Chen, Xian Guo, Siyuan Wang, Zihan Liang, Yue Lv, Yufei Ma, Xinlong Xiao, Bowen Xue, Xuxin Zhang, Ying Yang, et al. Onesearch: A preliminary exploration of the unified end-to-end generative framework for e-commerce search. [arXiv preprint arXiv:2509.03236](#), 2025.
- [4] 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. [arXiv preprint arXiv:2502.08468](#), 2025.
- [5] Haonan Chen, Sicheng Gao, Radu Timofte, Tetsuya Sakai, and Zhicheng Dou. e5-omni: Explicit cross-modal alignment for omni-modal embeddings. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 19430–19443, 2026.
- [6] Haonan Chen, Chu Li, Zhicheng Wang, Yuanwei Liu, Yuanjiang Wang, Shaohua Jiang, and Zhicheng Dou. Douyin multimodal embedding model technical report, 2026. URL <https://arxiv.org/abs/2608.02148>.
- [7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. [arXiv preprint arXiv:1504.00325](#), 2015.



- [8] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- [9] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.
- [10] Wei Chow, Yuan Gao, Linfeng Li, Xian Wang, Qi Xu, Hang Song, Lingdong Kong, Ran Zhou, Yi Zeng, Yidong Cai, et al. Merit: Multilingual semantic retrieval with interleaved multi-condition query. *Advances in Neural Information Processing Systems*, 38:74806–74867, 2025.
- [11] Xuanming Cui, Jianpeng Cheng, Hong-you Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Mishra, Taipeng Tian, Qi Guo, et al. Think then embed: Generative context improves multimodal embedding. In *International Conference on Learning Representations*, volume 2026, pages 2690–2709, 2026.
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [13] Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *CoRR*, abs/2502.18965, 2025. doi: 10.48550/ARXIV.2502.18965. URL <https://doi.org/10.48550/arXiv.2502.18965>.
- [14] Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. *arXiv preprint arXiv:2410.08182*, 2024.
- [15] Haohang Huang, Xuan Lu, Mingyi Su, Xuan Zhang, Ziyang Jiang, Ping Nie, Kai Zou, Tomas Pfister, Wenhui Chen, Wei Zhang, et al. Mmeh-v3: Measuring the performance gaps of omni-modality embedding models. *arXiv preprint arXiv:2604.23321*, 2026.
- [16] 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 Lang. Process.*, 32:2800–2813, 2024. doi: 10.1109/TASLP.2024.3402087. URL <https://doi.org/10.1109/TASLP.2024.3402087>.
- [17] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002. doi: 10.1145/582415.582418. URL <http://doi.acm.org/10.1145/582415.582418>.
- [18] 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.
- [19] Ziyang Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024.
- [20] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- [21] 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.
- [22] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibao Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *CoRR*, abs/2601.04720, 2026. doi: 10.48550/ARXIV.2601.04720. URL <https://doi.org/10.48550/arXiv.2601.04720>.
- [23] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*, 2024.
- [24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL [http://papers.nips.cc/paper\\_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html).



- [25] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2125–2134, 2021.
- [26] Xinchao Luo, Jiangxia Cao, Tianyu Sun, Jinkai Yu, Rui Huang, Wei Yuan, Hezheng Lin, Yichen Zheng, Shiyao Wang, Qigen Hu, Changqing Qiu, Jiaqi Zhang, Xu Zhang, Zhiheng Yan, Jingming Zhang, Simin Zhang, Mingxing Wen, Zhaojie Liu, and Guorui Zhou. QARM: quantitative alignment multi-modal recommendation at kuaishou. In Meeyoung Cha, Chanyoung Park, Noseong Park, Carl Yang, Senjuti Basu Roy, Jessie Li, Jaap Kamps, Kijung Shin, Bryan Hooi, and Lifang He, editors, Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM 2025, Seoul, Republic of Korea, November 10-14, 2025, pages 5915–5922. ACM, 2025. doi: 10.1145/3746252.3761502. URL <https://doi.org/10.1145/3746252.3761502>.
- [27] Quentin Macé, António Loison, and Manuel Faysse. Vidore benchmark v2: Raising the bar for visual retrieval. arXiv preprint arXiv:2505.17166, 2025.
- [28] Rui Meng, Ziyang 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.
- [29] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. In European Conference on Computer Vision, pages 291–309. Springer, 2024.
- [30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.
- [31] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision, pages 2641–2649, 2015.
- [32] Danilo Poccia. Amazon Nova Multimodal Embeddings: State-of-the-art embedding model for agentic RAG and semantic search, October 2025. URL <https://aws.amazon.com/blogs/aws/amazon-nova-multimodal-embeddings-now-available-in-amazon-bedrock/>.
- [33] 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, pages 8748–8763. PMLR, 2021.
- [34] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. Recommender systems with generative retrieval. Advances in Neural Information Processing Systems, 36:10299–10315, 2023.
- [35] Madhuri Shanbhogue, Zhe Li, Shanfeng Zhang, Gustavo Hernández Ábrego, Shih-Cheng Huang, Aashi Jain, Daniel Salz, Sonam Goenka, Chaitra Hegde, Ji Ma, et al. Gemini embedding 2: A native multimodal embedding model from gemini. arXiv preprint arXiv:2605.27295, 2026.
- [36] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In European conference on computer vision, pages 742–758. Springer, 2020.
- [37] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389, 2023.
- [38] Changli Tang, Qinfan Xiao, Ke Mei, Tianyi Wang, Fengyun Rao, and Chao Zhang. Wave: learning unified & versatile audio-visual embeddings with multimodal llm. In International Conference on Learning Representations, volume 2026, pages 61596–61612, 2026.
- [39] Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL <https://qwen.ai/blog?id=qwen3.5>.
- [40] Voyage AI. voyage-multimodal-3.5: A new multimodal retrieval frontier with video support, January 2026. URL <https://blog.voyageai.com/2026/01/15/voyage-multimodal-3-5/>.
- [41] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatec: A large-scale, high-quality multilingual dataset for video-and-language research. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 4580–4590. IEEE, 2019.

- [42] Chenghao Xiao, Hou Pong Ken Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. Scaling language-centric omnimodal representation learning. Advances in Neural Information Processing Systems, 38: 158370–158401, 2025.
- [43] Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth Enevoldsen, and Niklas Muennighoff. Mieb: Massive image embedding benchmark. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 22187–22198, 2025.
- [44] 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, pages 5288–5296, 2016.
- [45] Mengyao Xu, Wenfei Zhou, Yauhen Babakhin, Gabriel Moreira, Ronay Ak, Radek Osmulski, Bo Liu, Even Oldridge, and Benedikt Schifferer. Omni-embed-nemotron: A unified multimodal retrieval model for text, image, audio, and video. arXiv preprint arXiv:2510.03458, 2025.
- [46] Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. Advances in Neural Information Processing Systems, 38:17577–17604, 2025.
- [47] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. Advances in Neural Information Processing Systems, 36:7093–7115, 2023.
- [48] Huaying Yuan, Jian Ni, Zheng Liu, Yueze Wang, Junjie Zhou, Zhengyang Liang, Bo Zhao, Zhao Cao, Ji-Rong Wen, and Zhicheng Dou. Momentseeker: A task-oriented benchmark for long-video moment retrieval. Advances in Neural Information Processing Systems, 38, 2025.
- [49] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11975–11986, 2023.
- [50] 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 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9274–9285. IEEE, 2025.
- [51] Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. Vista: Visualized text embedding for universal multi-modal retrieval. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3185–3200, 2024.
- [52] 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), pages 19076–19095, 2025.
- [53] Luwei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.
- [54] Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. Marvel: unlocking the multi-modal capability of dense retrieval via visual module plugin. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14608–14624, 2024.