

# mmE5: Improving Multimodal Multilingual Embeddings via High-quality Synthetic Data

Haonan Chen<sup>1\*</sup>, Liang Wang<sup>2</sup>, Nan Yang<sup>2</sup>, Yutao Zhu<sup>1</sup>

Ziliang Zhao<sup>1</sup>, Furu Wei<sup>2</sup>, Zhicheng Dou<sup>1</sup>

<sup>1</sup>Gaoling School of Artificial Intelligence, Renmin University of China

<sup>2</sup>Microsoft Corporation

{hnchen, dou}@ruc.edu.cn

{wangliang, nanya, fuwei}@microsoft.com

<https://github.com/haon-chen/mmE5>

## Abstract

Multimodal embedding models have gained significant attention for their ability to map data from different modalities, such as text and images, into a unified representation space. However, the limited labeled multimodal data often hinders embedding performance. Recent approaches have leveraged data synthesis to address this problem, yet the quality of synthetic data remains a critical bottleneck. In this work, we identify three criteria for high-quality synthetic multimodal data. First, **broad scope** ensures that the generated data covers diverse tasks and modalities, making it applicable to various downstream scenarios. Second, **robust cross-modal alignment** makes different modalities semantically consistent. Third, **high fidelity** ensures that the synthetic data maintains realistic details to enhance its reliability. Guided by these principles, we synthesize datasets that: (1) cover a wide range of tasks, modality combinations, and languages, (2) are generated via a deep thinking process within a single pass of a multimodal large language model, and (3) incorporate real-world images with accurate and relevant texts, ensuring fidelity through self-evaluation and refinement. Leveraging these high-quality synthetic and labeled datasets, we train a **multimodal multilingual E5** model mmE5. Extensive experiments demonstrate that mmE5 achieves state-of-the-art performance on the MMEB Benchmark and superior multilingual performance on the XTD benchmark. Our codes, datasets, and models are released in <https://github.com/haon-chen/mmE5>.

## 1 Introduction

Multimodal embedding models encode multimedia inputs, such as images and text, into latent vector representations. They have demonstrated effectiveness across diverse downstream tasks, including

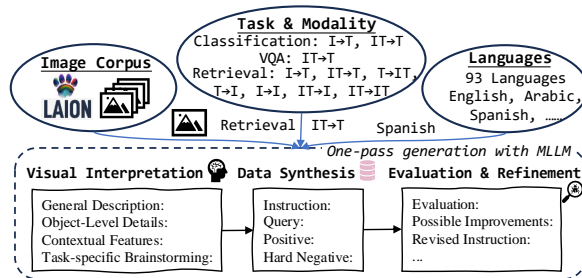


Figure 1: An illustration of our data synthesis framework. “X→Y” denotes a modality combination, where “X” represents the query side and “Y” denotes the target side. “T” denotes text and “I” denotes image.

classification (Deng et al., 2009), visual question answering (VQA) (Singh et al., 2019), and cross-modal retrieval (Hu et al., 2023). Prior studies have focused on training multimodal embedding models using simple text-image pre-trained models such as CLIP (Radford et al., 2021). More recently, researchers have turned to multimodal large language models (MLLMs), including LLaVA (Liu et al., 2023a) and Phi (Abdin et al., 2024), to develop universal embedding models.

These vision-language models (VLMs) mostly rely on high-quality human-labeled datasets to achieve robust embedding capabilities. Such datasets suffer from data scarcity because they require high costs of multimodal annotations. To address this, researchers have leveraged the advanced language modeling capabilities of large language models (LLMs) and MLLMs to synthesize datasets for fine-tuning multimodal embedding models (Zhang et al., 2024a; Zhou et al., 2024a; Zhang et al., 2024b). However, existing works lack a comprehensive exploration into the quality of synthetic embedding data. Typically, most data generated by them are limited to specific modality types of English retrieval tasks, harming the generalization capabilities of the embedding models.

After analyzing common application scenarios of multimodal embedding models, we identify

\*Work done during Haonan’s internship at MSR Asia. Prof. Zhicheng Dou is the corresponding author.

three key criteria and introduce a data synthesis framework guided by these principles: **(1) Broad scope.** Multimodal embedding models are commonly employed in tasks such as classification, visual question answering (VQA), and retrieval, which require understanding various input combinations of text and images. Additionally, multilingual contexts are increasingly popular in daily scenarios. As shown in Figure 1, our framework synthesizes datasets covering three tasks, seven modality combinations, and 93 languages, ensuring that models trained on it generalize effectively across diverse scenarios. **(2) Robust cross-modal alignment.** In multimodal tasks, models must understand and align information across different modalities to generate meaningful representations. Without accurate cross-modal alignment, embeddings may fail to capture the underlying relationships, leading to poor performance in downstream tasks. To synthesize data of robust cross-modal alignment, our framework incorporates a deep thinking process. Specifically, for each sampled image, we first employ an MLLM to interpret it from four perspectives before generating data: general information, object-level description, contextual background information, and task-specific brainstorming, *i.e.*, how the image relates to the given task. Additionally, the entire data synthesis process is executed within a single pass of an MLLM. By this, the MLLM can “see” the images at the whole time, avoiding potential information loss that might occur due to multiple I/O steps in previous works (Zhou et al., 2024a; Zhang et al., 2024b). **(3) High fidelity.** The individual quality of each modality (*e.g.*, real images, high-quality instructions, queries, and hard negatives) determines the overall usefulness of the dataset. To enhance fidelity, our framework uses real images sampled from an open-source corpus (LAION-400m (Schuhmann et al., 2021)) as the input images. We also apply a series of quality control measures, such as self-evaluation and refinement, ensuring that the synthetic components accurately reflect real-world distributions and maintain strong cross-modal alignment.

With the synthesized data ready, we train a **multimodal multilingual E5** model (mmE5). It achieves state-of-the-art performance on the 36 datasets of MMEB (Jiang et al., 2024b), using 45 times less training data than the previous SOTA model MMRet (Zhou et al., 2024a) (560K compared to 26M) in a zero-shot setting. After incor-

porating labeled data, mmE5 still demonstrates the best performance. Besides, mmE5 achieves the best results on the multilingual benchmark XTD (Aggarwal and Kale, 2020), demonstrating its superior multilingual capabilities.

In summary, our contributions are as follows:

- Based on our analysis of common scenarios for multimodal embedding models, we identify three key criteria of high-quality synthetic data: broad scope, robust cross-modal alignment, and high fidelity.
- We introduce a data synthesis framework guided by the proposed principles. This framework leverages an MLLM to produce high-quality synthetic datasets that cover a wide range of tasks, modality combinations, and languages. It ensures robust cross-modal alignment through a comprehensive multi-aspect interpretation process and maintains high fidelity by employing self-evaluation and refinement mechanisms.
- Compared to the previous leading model, mmE5 achieves SOTA performance on the MMEB benchmark while using 45× less synthetic data in both zero-shot and supervised settings. mmE5 also demonstrates superior multilingual capabilities on the XTD benchmark.

## 2 Related Work

**Multimodal Embedding** Previous studies, such as CLIP (Radford et al., 2021), Align (Jia et al., 2021), BLIP (Li et al., 2022), and CoCa (Yu et al., 2022), have employed large-scale weakly supervised data to learn separate multimodal representations through pre-training. Some works attempt to obtain universal embeddings for texts and images utilizing existing CLIP-like models (Wei et al., 2024; Liu et al., 2023b; Zhou et al., 2024b,c). For instance, UniIR (Wei et al., 2024) integrates separate embeddings from different modalities into unified features. Recent approaches finetune MLLMs to leverage their multimodal reasoning capabilities for obtaining universal representations (Jiang et al., 2024a,b; Zhang et al., 2024b; Zhou et al., 2024a; Lin et al., 2024). For example, VLM2Vec (Jiang et al., 2024b) utilizes instruction tuning to transform MLLMs into embedding models.

**Synthetic Data** The generation of synthetic data has been extensively explored for text embedding

Method	# Languages	Task	Modality Combinations	w/ MLLM	One Pass	Self-evaluation
MagicLens	1 (English)	Retrieval	IT→I	×	✓	×
MegaPairs	1 (English)	Retrieval	IT→I	✓	×	×
GME	1 (English)	Retrieval	T→IT, IT→IT	×	×	×
mmE5 (Ours)	93 (English, Spanish, etc.)	Classification, VQA, Retrieval	IT→I, T→IT, IT→IT, I→I, I→T, IT→T, T→I	✓	✓	✓

Table 1: Comparison of the synthetic datasets in our work with those from previous methods. Our synthetic datasets encompass 93 languages, two additional tasks, and a wider range of modality combinations. “IT→T” denotes a modality combination, where “IT” denotes images and texts on the query side and “T” denotes texts on the target side. The entire data synthesis process is executed within a single pass of an MLLM, thereby avoiding potential information loss and ensuring robust cross-modal alignment. We also employ real images and self-evaluation to maintain fidelity.

tasks (Wang et al., 2024a; Chen et al., 2024; Li et al., 2024b). With the recent emergence of MLLMs like Phi-3.5-V (Abdin et al., 2024) and LLaVA (Liu et al., 2023a), along with diffusion models such as Stable Diffusion (Rombach et al., 2022), researchers have been focusing on synthesizing data to address the scarcity of multimodal instruction-tuning datasets. For example, MagicLens (Zhang et al., 2024a) utilizes co-existing images from the same webpage and an LLM to create multimodal data triplets (query image, instruction, relevant image), *i.e.*, IT→I paradigm. MegaPairs (Zhou et al., 2024a) aims to synthesize more diverse data triplets by retrieving relevant images from different perspectives. GME (Zhang et al., 2024b) employs an LLM and a diffusion model to generate a fused modality dataset that includes both T→IT and IT→IT types. Table 1 presents a comparison of the synthesized data in this study with that of previous works.

### 3 Methodology: mmE5

In this section, we present our method, which synthesizes high-quality multimodal data for the further finetuning of our embedding model mmE5. As shown in Figure 2, our method consists of five stages: (1) Initially, for each data sample to be synthesized, we configure the specifics of the task, modality combination, language, and input images. (2) We employ an MLLM to generate multi-grained descriptions for the input images, ensuring that the synthesized texts are well-aligned with the images. (3) Utilizing this MLLM, we synthesize text data based on both the images and their descriptions. (4) The MLLM then evaluates its synthesized data from multiple perspectives, offering revised

data to enhance cross-modal alignment and fidelity. (5) Finally, the synthesized texts and images are used to finetune an MLLM specifically for embedding tasks. To minimize potential information loss, stages (2), (3), and (4) are executed within a single pass of the MLLM.

#### 3.1 Preliminaries

An MLLM can accept text, image, or text-image pairs as input, allowing both the query side  $q$  and the document side  $d$  to be multimodal. Inspired by existing works on synthetic text embedding data (Wang et al., 2024a; Chen et al., 2024), each data sample we generate is a quadruple of (task instruction, query, positive document, hard negative document), denoted as  $(t, q, d^+, d^-)$ . For each data piece, we first sample images from the large-scale open-source image corpus LAION-400M (Schuhmann et al., 2021) as the query image, positive image, and hard negative image  $(q_i, d_i^+, d_i^-)$ . Then, with these three images as input, an MLLM  $\pi_\theta$  can synthesize a multimodal embedding data sample  $y \sim \pi_\theta(y \mid q_i, d_i^+, d_i^-)$ , where  $y = (t, q_t, d_t^+, d_t^-)$ . As a result, the synthetic data can have a maximum of seven elements:  $\{t, (q_t, q_i), (d_t^+, d_i^+), (d_t^-, d_i^-)\}$ . More data examples can be found in Appendix D.

#### 3.2 Data Synthesis Framework

Guided by the principles of high-quality synthetic multimodal data, *i.e.*, broad scope, robust cross-modal alignment, and high fidelity, we introduce a data synthesis framework. This framework is designed to synthesize high-quality data that transforms an MLLM for downstream embedding tasks.

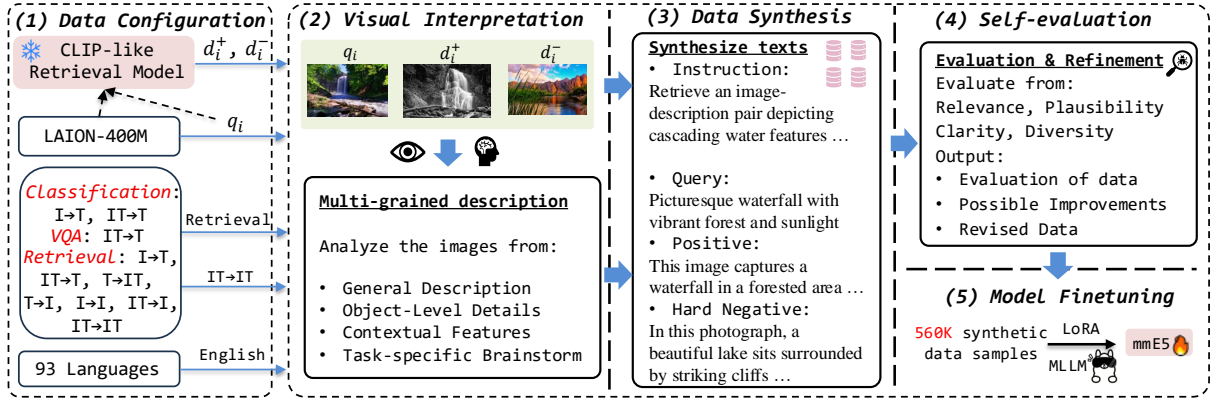


Figure 2: An illustration of our method. We take the generation of an IT→IT retrieval data sample as an example.

### 3.2.1 Data Configuration

To prepare for the data synthesis process, we configure the input data from three aspects:

**Task and Modality Combination** We aim to synthesize data with a broad scope by generating beyond simple retrieval data of IT→IT and T→IT types. Our data cover three key multimodal embedding tasks identified by previous work (Jiang et al., 2024b): classification, VQA, and retrieval. After selecting a task for synthesis, we will sample a modality combination with respect to the specific task, such as choosing from seven possible combinations for the retrieval task type. Note that we only synthesize data of modality types that are included in the MMEB benchmark (Jiang et al., 2024b), which can cover most scenarios.

**Image** Despite the powerful multimodal capabilities of modern MLLMs (e.g., GPT-4o, Llama-3.2 (Meta, 2024), and Llava-1.6), most cannot generate images, and those that can often produce low-fidelity images (Zhou et al., 2024b). Following previous works (Zhang et al., 2024a; Zhou et al., 2024a), we sample real images from the LAION-400M corpus (Schuhmann et al., 2021). First, we will sample a query image from the corpus ( $q_i \in \mathcal{I}$ ). Then, for the modality types involving images on the document side (e.g., IT→IT), we use a small embedding model, jina-clip-v2 (Koukounas et al., 2024), to retrieve a similar positive image  $d_i^+$  and a hard negative image  $d_i^-$  efficiently.

**Language** Most existing models only focus on high-source languages like English, harming the multilingual ability of embedding models. To synthesize multilingual data, we sample languages from the language list of XLM-R (Conneau et al., 2020) during configuration. In order to facilitate the common usage scenarios, we give high-source languages higher weights. Note that the generated

task instruction will always be in English for effective instruction tuning.

### 3.2.2 One-pass Generation with MLLM

With the data configuration ready, we introduce a deep thinking process that involves interpreting input images, generating data, and performing self-evaluation. To ensure that the MLLM always takes the image context into account, we execute this entire process in a single pass.

**Multi-aspect Visual Interpretation** To obtain a comprehensive understanding of the images, the MLLM  $\pi_\theta$  first analyzes them from multiple perspectives: (1) the general information, (2) detailed description of the objects present, (3) contextual background information, and (4) potential connections between the image and the text that may be synthesized. The deep understanding of the images enables  $\pi_\theta$  to produce texts that are closely aligned with the visual content, thereby enhancing the cross-modal alignment.

**Synthesizing Data** Using the images and their descriptions as input, we prompt  $\pi_\theta$  to synthesize texts ( $t, q_t, d_t^+, d_t^-$ ). Specifically, the text instruction  $t$  is expected to connect  $q_i$  with  $d_i^+$ .<sup>1</sup> The query and document texts should be relevant to their respective images. Note that the input and output formats for the synthetic data may vary depending on the combination of modalities. For example, for I→IT and T→IT types, there can be no query text and image, respectively.

**Self-evaluation** To further enhance the quality of the synthetic data,  $\pi_\theta$  evaluates the data it synthesizes from: (1) the relevance of the texts to their corresponding images, (2) the plausibility of hard negatives, (3) the clarity of  $t$ , and (4) the diversity

<sup>1</sup>Because of limited space, full prompts are omitted in this section. The complete prompts can be found in Appendix C.



Models	Per Meta-Task Score				Average Score		
	Class.	VQA	Retr.	Ground.	IND	OOD	Overall
<i>Zero-shot Setting Models</i>							
CLIP (Radford et al., 2021)	42.8	9.1	53.0	51.8	-	-	37.8
BLIP2 (Li et al., 2023)	27.0	4.2	33.9	47.0	-	-	25.2
SigLIP (Zhai et al., 2023)	40.3	8.4	31.6	59.5	-	-	34.8
OpenCLIP (Cherti et al., 2023)	47.8	10.9	52.3	53.3	-	-	39.7
E5-V (Jiang et al., 2024a)	21.8	4.9	11.5	19.0	-	-	13.3
MagicLens (Zhang et al., 2024a)	38.8	8.3	35.4	26.0	-	-	27.8
MMRet (w/ 26M synthetic data)	47.2	18.4	<b>56.5</b>	62.2	-	-	44.0
mmE5 (w/ 560K synthetic data)	<b>60.6</b>	<b>55.7</b>	<u>54.7</u>	<b>72.4</b>	-	-	<b>58.6</b>
<i>Partially Supervised Finetuning Models<sup>†</sup></i>							
UniIR (Wei et al., 2024)	42.1	15.0	60.1	62.2	-	-	42.8
MM-EMBED (Lin et al., 2024)	48.1	32.2	63.8	57.8	-	-	50.0
GME (Zhang et al., 2024b)	56.9	41.2	67.8	53.4	-	-	55.8
<i>Supervised Finetuning Models</i>							
CLIP (Radford et al., 2021)	55.2	19.7	53.2	62.2	47.6	42.8	45.4
OpenCLIP (Cherti et al., 2023)	56.0	21.9	55.4	64.1	50.5	43.1	47.2
VLM2Vec (Jiang et al., 2024b)	<u>61.2</u>	49.9	67.4	<u>86.1</u>	67.5	57.1	62.9
MMRet (Zhou et al., 2024a)	56.0	<u>57.4</u>	<u>69.9</u>	83.6	<u>68.0</u>	<u>59.1</u>	<u>64.1</u>
mmE5 (w/ synthetic data + labeled data)	<b>67.6</b>	<b>62.7</b>	<b>71.0</b>	<b>89.7</b>	<b>72.4</b>	<b>66.6</b>	<b>69.8</b>

Table 2: Results on MMEB benchmark, consisting of 36 tasks across four types: classification (Class.), VQA, retrieval (Retr.), and visual grounding (Ground.). <sup>†</sup> UniIR, MM-EMBED, and GME are not strictly zero-shot models. UniIR and MM-EMBED are trained on the MBEIR dataset (Wei et al., 2024), which includes 10 retrieval datasets included in the MMEB. Similarly, GME is trained on the UMRB dataset (Zhang et al., 2024b), which shares 14 datasets with the MMEB. For VLM2Vec, we use the LLaVA-based version with high-resolution images reported in its original paper. The second-best performances are underlined, and the best performances are in bold.

(creativity) of the synthesized data. Following this evaluation,  $\pi_\theta$  provides suggestions for potential improvements. Finally, a revised version of each data sample is produced and utilized for the subsequent contrastive training phase.

### 3.3 Finetuning Embedding Model mmE5

Following previous works of instruction-tuned text embedding models (Xiao et al., 2024; Li et al., 2024a) and multimodal embedding models (Jiang et al., 2024b), we apply an instruction template on each query: [IMAGE]  $\{t\} \setminus n \{q_t\} \{q_i\}$ , where “[IMAGE]” is the image token that varies from different MLLMs. We then append an “[EOS]” token to each query and document. The representation of each input in an MLLM is derived from the output of the “[EOS]” token from the final layer.

We utilize the InfoNCE loss (van den Oord et al., 2018) to perform the standard contrastive learning objective on our synthetic data  $\mathcal{D}$ :

$$\mathcal{L} = -\log \frac{\phi(\mathbf{q}, \mathbf{d}^+)}{\phi(\mathbf{q}, \mathbf{d}^+) + \sum_{\mathbf{d}^- \in \mathcal{N}} \phi(\mathbf{q}, \mathbf{d}^-)}, \quad (1)$$

where  $\mathbf{q}$  is the encoded multimodal query,  $\mathbf{d}$  represents the encoded document, and  $\mathcal{N}$  denotes the set of negative documents. The function  $\phi(\cdot) =$

$\exp(\cos(\cdot)/\tau)$ , where  $\cos(\cdot)$  denotes cosine similarity, and  $\tau$  is a temperature hyperparameter.

## 4 Experiments

### 4.1 Experimental Setup

We synthesize a total of 560K multimodal embedding data samples. The MLLM utilized for data synthesis is *GPT-4o-2024-08-06*. The backbone model for mmE5 is Llama-3.2-11B-Vision<sup>2</sup>. For finetuning mmE5, we employed LoRA (Hu et al., 2022) with a rank of 8. We evaluate the general embedding performance in terms of Precision@1 on the MMEB benchmark (Jiang et al., 2024b). This benchmark comprises 36 multimodal embedding tasks across four categories: classification (10), VQA (10), retrieval (12), and visual grounding (4). Our synthetic dataset is distributed among classification, VQA, and retrieval tasks in a 1:1:2 ratio. We synthesize more retrieval data since this type contains more kinds of modality combinations. We do not synthesize visual grounding data since they are relatively simpler for MLLM based on the MMEB results. To evaluate multilingual multimodal capabilities, we conducted tests using the

<sup>2</sup><https://huggingface.co/meta-llama/Llama-3.2-11B-Vision>

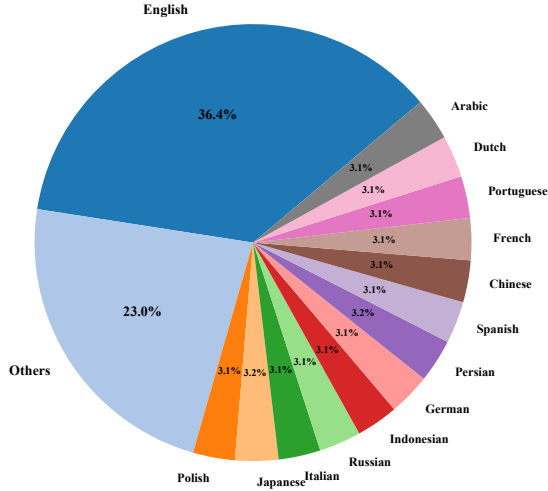


Figure 3: Distribution of languages in the synthetic data.

XTD benchmark (Aggarwal and Kale, 2020). Following MURAL (Jain et al., 2021), we conduct experiments on seven languages of XTD and report Recall@10 results. Additional details regarding the synthetic data, prompts, and implementation can be found in Appendix A, B, and C, respectively.

## 4.2 Results on MMEB

The overall results on the MMEB benchmark are presented in Table 2. mmE5 achieves the best performance on both zero-shot setting (with synthetic data only) and supervised setting (with IND training datasets of MMEB). This demonstrates the quality of our synthetic data and the effectiveness of our multimodal embedding model. Furthermore, we can make the following observations: (1) mmE5 generalizes well on all four kinds of tasks. This demonstrates the broad scope of our synthetic multimodal embedding data in terms of task types. (2) With only 560K synthetic data, mmE5 manages to perform better than MMRet which uses 26M data. This proves the quality of our synthetic data again. (3) Intriguingly, mmE5 underperforms MMRet on retrieval tasks in a zero-shot setting. This is because MMRet is trained on 26M pure retrieval data, which makes it perform well on retrieval tasks, but generalizes poorly on other task types.

## 4.3 Multilingual Performance on XTD

We synthesize a multilingual multimodal dataset that consists of 93 languages, in order to train our embedding model mmE5 to generalize across more languages. The language distribution of our dataset is presented in Figure 3. Notably, the dataset primarily consists of English data samples, facilitating

Model	it	es	ru	zh	pl	tr	ko	Avg.
ALIGN (Jia et al., 2021)	87.9	88.8	82.3	86.5	79.8	73.5	76.6	82.2
MURAL (Jain et al., 2021)	91.8	92.9	87.2	89.7	91.0	89.5	88.1	90.0
VLM2Vec (Jiang et al., 2024b)	83.7	87.1	86.7	92.8	76.1	37.2	63.9	75.4
jina (Koukounas et al., 2024)	93.6	94.1	89.8	91.8	94.3	92.7	90.1	92.3
M-CLIP (Carlsson et al., 2022)	93.1	93.6	90.0	94.0	94.3	93.1	89.0	92.4
GME (Zhang et al., 2024b)	95.1	96.4	92.3	96.4	94.9	89.8	93.6	94.1
mmE5 (full)	96.1	96.2	93.3	96.3	95.4	93.6	96.0	<b>95.3</b>
w/ synthetic data only	90.9	89.6	86.3	90.2	90.3	87.2	86.7	88.7
w/ english synthetic data	86.3	86.3	84.2	88.8	84.9	81.0	84.4	85.1

Table 3: Results on XTD benchmark, a text-to-image retrieval task covering seven languages.

common usage scenarios. For the 75 low-resource languages, we evenly synthesize data samples to obtain a balanced multilingual dataset that supports comprehensive cross-linguistic generalization.

To evaluate the multilingual capability of mmE5, we conduct experiments across seven languages on a text-to-image retrieval benchmark XTD. As presented in Table 3, mmE5 outperforms other models in terms of overall performance on all languages, demonstrating its superior multilingual multimodal embedding capability. The following observations can be made: (1) The multilingual performance of multimodal embedding models is largely dependent on their foundational models. For example, jina-clip-v2 and M-CLIP outperform VLM2Vec-LLaVA, despite VLM2Vec’s strong performance on MMEB. GME exhibits robust performance on XTD, which can be attributed to the powerful multilingual MLLM, Qwen2-VL (Wang et al., 2024b). (2) The performance of mmE5 declines when labeled data is omitted, indicating that general multimodal capabilities remain essential for multilingual retrieval tasks. (3) In a zero-shot setting, mmE5 trained on multilingual synthetic data (mmE5 w/ synthetic data only) outperforms mmE5 with the same amount of English synthetic data (mmE5 w/ english synthetic data). This suggests that the extensive language coverage provided by our synthetic data enhances the multilingual capabilities of embedding models.

## 4.4 Application to Other Base MLLM

We train mmE5 based on the powerful MLLM LLaMA-3.2-Vision, which is instruction-tuned and effective in interpreting multimodal inputs. Notably, our synthetic data and training paradigm can effectively transform other foundation MLLMs into embedding models. We use both our synthetic data and labeled data to train LLaVA-1.6<sup>3</sup> and Phi-

<sup>3</sup><https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf>

Base MLLM	Avg. on MMEB
Phi-3.5-V (Abdin et al., 2024)	61.0
LLaVA-1.6 (Liu et al., 2023a)	65.8
LLaMA-3-Vision (Meta, 2024) (Ours)	<b>69.8</b>
<i>Baselines (For Reference)</i>	
VLM2Vec (Phi-3.5-V)	60.1
VLM2Vec (LLaVA-1.6)	62.9
MMRet (LLaVA-1.6)	64.1
VLM2Vec (LLaMA-3.2)	64.8

Table 4: Performances of mmE5 with different MLLMs.

3.5-V<sup>4</sup>. The performances of mmE5 with different foundation MLLMs are presented in Table 4. The results show that models trained using our method consistently outperform baseline models built on the same foundational MLLMs. This indicates that our synthetic data can effectively enhance the capability of MLLMs to embed multimodal inputs.

#### 4.5 Discussions of Data Synthesis Process

In this section, we will further investigate the data synthesis process via zero-shot experiments.

##### 4.5.1 Ablation Studies

To evaluate each component of our data synthesis framework, we conduct ablation studies of mmE5:

**Deep Thinking Process** To synthesize high-quality data, we introduce a deep thinking process to boost data synthesis. As presented in Table 5, the performance of mmE5 declines when the Visual Interpretation and Self-evaluation components are excluded. For example, mmE5 performs worse when utilizing the original data compared to the revised data. This indicates that the self-evaluation mechanism can enhance data fidelity, facilitating the training of a more robust embedding model.

**Embedding Task Types** In order to expand the scope of data, we synthesize data across three task types: classification, VQA, and retrieval. The performance of mmE5 decreases after each type of multimodal embedding data is omitted, demonstrating that our diverse synthetic data can facilitate model generalization. Intriguingly, the performance drops the least after removing the retrieval data, which is inconsistent with previous research (Jiang et al., 2024b). One possible explanation is that our backbone, Llama-3.2 Vision, inherently exhibits more robust retrieval capabilities than Phi-3.5-V.

<sup>4</sup><https://huggingface.co/microsoft/Phi-3.5-vision-instruct>

Model	Avg. on MMEB
mmE5 (280K synthetic data only)	<b>57.4</b>
w/o. Visual Interpretation	57.2
w/o. Self-evaluation	56.0
w/o. Classification Data	52.5
w/o. VQA Data	55.1
w/o. Retrieval Data	56.5
w/ IT2I only (MagicLens & MegaPairs)	30.1
w/ IT2IT & T2IT only (GME)	28.6
w/o. Hard Negative	56.2
w/ English Data only (280K)	57.6
w/o. English Data (280K)	56.9

Table 5: Performances of ablated models on MMEB. For efficient test, we conduct zero-shot experiments on 280K synthetic data, which has the same tasks, modality types and languages as the full synthetic data.

**Modality Combinations** Most prior works focus on one or two modality types, such as “IT2I” (e.g., MagicLens (Zhang et al., 2024a) and MegaPairs (Zhou et al., 2024a)) or “IT2IT & T2IT” (e.g., GME (Zhang et al., 2024b)). We propose to synthesize data across various modality combinations to enhance the diversity of our synthetic dataset, i.e., the scope of our synthetic multimodal data. To evaluate the impact of these additional modality combinations, we train mmE5 with the same amount of datasets that contain types “IT2I” or “IT2IT & T2IT” only. The performance of mmE5 significantly decreased when limited to these combinations from previous works, which indicates that the additional modalities enable our embedding model to generalize more effectively across different combinations and task types.

**Hard Negative** Each sample in our synthetic dataset incorporates a hard negative document to help mmE5 learn subtle differences. After excluding the hard negatives, the model’s performance drops significantly, which demonstrates the importance of this technique for contrastive learning.

**Language** To investigate the impact of linguistic diversity on model performance on English benchmarks, we conducted experiments using synthetic data in two configurations: English-only and non-English languages only. Our model, mmE5, demonstrated a slight performance advantage with English-only synthetic data, although the difference was minimal. Nonetheless, mmE5 achieved satisfactory results with 280K data samples from languages other than English. This suggests that

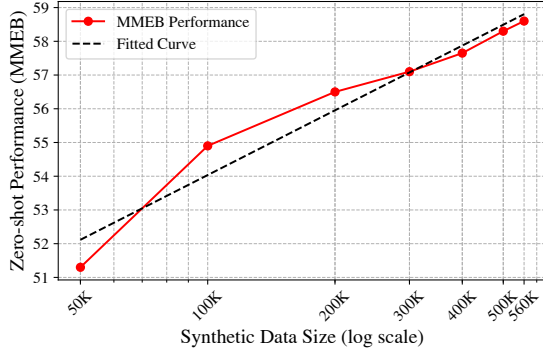


Figure 4: The impact of synthetic data size on multi-modal embedding performance on MMEB.

our multilingual dataset enhances the embedding model’s ability to generalize effectively in both multilingual and English-only contexts.

#### 4.5.2 Scaling Effect

The scaling effect is an important aspect of synthetic data generation for multimodal embedding models (Zhang et al., 2024b; Zhou et al., 2024a). It explores how the performance of the model varies with the size of synthetic datasets. Besides, the data synthesis and training processes demand significant computational resources and time. Therefore, studying the scaling effect allows us to identify the point of diminishing returns, ensuring that resources are utilized efficiently without overproducing redundant data.

In this section, we further investigate the performance of mmE5 using synthetic datasets of varying sizes. Specifically, we conduct zero-shot experiments on MMEB to analyze the scaling effect. As illustrated in Figure 4, mmE5 consistently achieves better performance with increased training data, demonstrating the high quality of our synthetic data again. This paradigm also indicates a linear-log relationship between the model performance and data size, consistent with previous works of text embedding (Chen et al., 2024) and dense retrieval (Fang et al., 2024). This finding facilitates the balancing of the cost and the multimodal embedding model performance for future works.

#### 4.6 Hyperparameter Analysis

In order to analyze the training process of our multimodal embedding model, we perform experiments with mmE5 using various training settings. For efficiency, we report zero-shot results for mmE5 trained with 280K synthetic data. Note that we tune these hyperparameters on evaluation datasets comprising

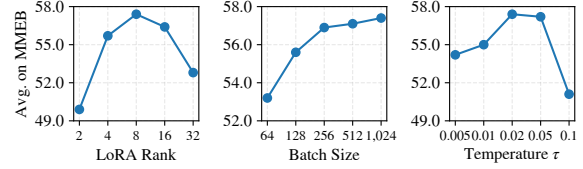


Figure 5: The zero-shot performances of mmE5 with different training settings on MMEB (280K synthetic data for efficient test).

1K samples from each training set. However, for consistency with previous experiments, we present results on the MMEB test sets.

**LoRA Rank** denotes the rank of the additional low-rank matrices in LoRA. This parameter influences the number of parameters added to the original model, balancing the model’s capacity and computational efficiency. As shown in the left part of Figure 5, the performance of mmE5 initially improves then drops. This demonstrates a trade-off: a lower rank reduces memory and computation but may lead to underfitting if  $r$  is too small, whereas a higher rank risks harming the pre-trained multimodal reasoning capabilities of MLLM.

**Training Batch Size** In contrastive learning, batch size plays a critical role because it directly affects the number of negative samples available for training. As presented in the middle part of Figure 5, the performance of mmE5 consistently increases with larger batch size. However, large batches demand significantly more GPU memory, *i.e.*, more computational resources.

**Temperature** The temperature parameter  $\tau$  in the InfoNCE loss (Equation 1) influences the separation between positive and negative samples in the embedding space. We can observe that mmE5’s performance first improves then declines with larger temperature. This pattern suggests a trade-off: a low  $\tau$  forces the model to strongly penalize near-positive negatives, which can lead to overfitting, while a high  $\tau$  leads to a more uniform distribution of embeddings, which may hinder the effective separation of positive and negative samples.

## 5 Conclusion

In this work, we synthesize high-quality multimodal multilingual data to train the model mmE5. We first define high-quality multimodal synthetic data based on three criteria: broad scope, robust cross-modal alignment, and high fidelity. Then, we develop a data synthesis framework guided by these principles. Finally, we train a multimodal multi-



lingual embedding model using the high-quality synthetic data. mmE5 achieves SOTA performances on both the general benchmark MMEB and the multilingual benchmark XTD.

## Limitations

Our work has several limitations that we intend to resolve in future research:

1. Our model currently relies on the proprietary MLLM GPT-4o for synthesizing multimodal data. Future work should explore aligning smaller MLLMs with the knowledge from GPT-like models to achieve more efficient data synthesis.
2. mmE5 focus on text and image modalities. Future models should aim to extend coverage to additional modalities, such as audio and video.
3. Due to the cost limitation and the observed scaling effect, we limited the amount of data produced for model training. Future research may consider increasing data size while preserving diversity to optimize model performance.

## Acknowledgments

This work was supported by Beijing Natural Science Foundation No. L233008, Beijing Municipal Science and Technology Project No. Z231100010323009, National Natural Science Foundation of China No. 62272467. The work was partially done at the Beijing Key Laboratory of Research on Large Models and Intelligent Governance and Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE.

## 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, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, 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 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. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *CoRR*, abs/2404.14219.
- Pranav Aggarwal and Ajinkya Kale. 2020. [Towards zero-shot cross-lingual image retrieval](#). *CoRR*, abs/2012.05107.
- Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. [Cross-lingual and multilingual CLIP](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 6848–6854. European Language Resources Association.
- Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. 2024. [Little giants: Synthesizing high-quality embedding data at scale](#). *CoRR*, abs/2410.18634.
- Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. [Reproducible scaling laws for contrastive language-image learning](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2818–2829. IEEE.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. [Imagenet: A large-scale hierarchical image database](#). In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.
- Yan Fang, Jingtao Zhan, Qingyao Ai, Jiaxin Mao, Weihang Su, Jia Chen, and Yiqun Liu. 2024. [Scaling laws for dense retrieval](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research*

- and Development in Information Retrieval, *SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 1339–1349. ACM.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. [Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 12031–12041. IEEE.
- Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. [MURAL: multimodal, multitask retrieval across languages](#). *CoRR*, abs/2109.05125.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. [Scaling up visual and vision-language representation learning with noisy text supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.
- Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024a. [E5-V: universal embeddings with multimodal large language models](#). *CoRR*, abs/2407.12580.
- Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. 2024b. [Vlm2vec: Training vision-language models for massive multimodal embedding tasks](#). *arXiv preprint arXiv:2410.05160*.
- Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Günther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, and Han Xiao. 2024. [jina-clip-v2: Multilingual multimodal embeddings for text and images](#).
- Chaofan Li, Zheng Liu, Shitao Xiao, Yingxia Shao, and Defu Lian. 2024a. [Llama2vec: Unsupervised adaptation of large language models for dense retrieval](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3490–3500. Association for Computational Linguistics.
- Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, and Furu Wei. 2024b. [Synthetic data \(almost\) from scratch: Generalized instruction tuning for language models](#). *CoRR*, abs/2402.13064.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. [BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. [BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.
- Sheng-chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2024. [Mm-embed: Universal multimodal retrieval with multimodal llms](#).
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. [Improved baselines with visual instruction tuning](#).
- Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2023b. [Universal vision-language dense retrieval: Learning A unified representation space for multi-modal retrieval](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. <https://ai.meta.com/blog/meta-llama-3/>.
- 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. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. [High-resolution image synthesis with latent diffusion models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.

- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. [LAION-400M: open dataset of clip-filtered 400 million image-text pairs](#). *CoRR*, abs/2111.02114.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. [Towards VQA models that can read](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.
- Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. [Representation learning with contrastive predictive coding](#). *CoRR*, abs/1807.03748.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. [Improving text embeddings with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11897–11916. Association for Computational Linguistics.
- 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. 2024b. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *CoRR*, abs/2409.12191.
- Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhui Chen. 2024. [Uniir: Training and benchmarking universal multimodal information retrievers](#). In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXVII*, volume 15145 of *Lecture Notes in Computer Science*, pages 387–404. Springer.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-pack: Packed resources for general chinese embeddings](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 641–649. ACM.
- Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. [Coca: Contrastive captioners are image-text foundation models](#). *Trans. Mach. Learn. Res.*, 2022.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. [Sigmoid loss for language image pre-training](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE.
- Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhui Chen, Yu Su, and Ming-Wei Chang. 2024a. [Magiclens: Self-supervised image retrieval with open-ended instructions](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024b. [Gme: Improving universal multimodal retrieval by multimodal llms](#).
- Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. 2024a. [Megapairs: Massive data synthesis for universal multimodal retrieval](#). *arXiv preprint arXiv:2412.14475*.
- Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024b. [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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3185–3200. Association for Computational Linguistics.
- Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. 2024c. [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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14608–14624. Association for Computational Linguistics.

## Appendix

### A Details about Synthetic Data

Task	Modality combination	# Samples
Classification	image-to-text	126,177
	(image,text)-to-text	13,823
Retrieval	image-to-text	98,040
	(image,text)-to-text	41,960
	(image,text)-to-image	56,185
	image-to-image	27,988
	(image,text)-to-(image,text)	27,656
	text-to-image	14,090
VQA	text-to-(image,text)	14,081
	(image,text)-to-text	140,000

Table 6: Statistics of the multimodal synthetic data used for training mmE5.

In this study, we introduce a synthetic multimodal multilingual embedding dataset designed to facilitate model learning. This section delves into the details of our synthetic dataset. The dataset is comprised of three distinct tasks and seven modality combinations, totaling 560K data samples. Table 6 provides a detailed statistical overview of our synthetic data, categorized by tasks and modalities.

### B Implementation Details

#### B.1 Data Synthesis

For the data synthesis process, we employ the MLLM *GPT-4o-2024-08-06* model to generate data samples. Both the temperature and top-p parameters are set to 1.0 to ensure diverse and coherent outputs. Our image corpus is sourced from LAION-400m (Schuhmann et al., 2021), from which we exclude images that are either corrupted or have inaccessible URLs. Each synthetic data sample incorporates one image sampled from this corpus as the query image. For modality combinations that include images on the document side, we utilize the jina-clip-v2<sup>5</sup> model to retrieve a similar image, along with a hard negative image, to serve as additional inputs.

#### B.2 Finetuning Embedding Model

We train mmE5 using the open-source MLLM, Llama-3.2-11B-Vision<sup>6</sup>. The training is conducted on 64 NVIDIA A100 GPUs, each equipped with

<sup>5</sup><https://huggingface.co/jinaai/jina-clip-v2>

<sup>6</sup><https://huggingface.co/meta-llama/Llama-3.2-11B-Vision>

40GB of memory. To optimize GPU memory usage, we employ gradient checkpointing and set the gradient accumulation steps to 4. The model is trained with a learning rate of  $2e-5$  for one epoch, utilizing both synthetic and labeled data. LoRA (Hu et al., 2022) is applied to the MLLM with a rank of 8. Each training sample incorporates one hard negative document. Hard negatives are mined for each subset of MMEB using VLM2Vec-LoRA<sup>7</sup>, with the 70th position in the ranking list selected as the hard negative sample.

### C Prompts

We use different prompts of data synthesis for different tasks. For retrieval task, we design two prompts for modality combinations that involve images on the document side or not. Let us take the prompt of generating classification data for an example to illustrate the prompt design.

First, we sample a modality combination from {image-to-text, (image,text)-to-text}. If the query side does not include texts, the “input\_text” of the classification data sample will be an empty string. Similarly, for modalities of retrieval task that do not include document texts, the “positive\_document” and “hard\_negative\_document” will be empty. Following previous works of synthesizing text embedding data (Wang et al., 2024a; Chen et al., 2024), we will randomly select a clarity and difficulty setting to enhance diversity.

Then, for the multi-aspect visual description process, we ask the MLLM to explicitly **include four perspectives of description**. Besides, for the data synthesis process, we also ask the MLLM to **follow some specific guidelines**. Furthermore, the MLLM will **evaluate the initially generated data from several aspects** and provide “possible\_improvements”. Finally, the revised version of data will be used as the output data sample. Note that there are no task instructions generated for the VQA task, since they are all fixed as “Represent the given image with the following question:”.

### D Data Examples

In this section, we present the examples of the synthetic multimodal embedding data for Retrieval (Figure 6 and Figure 7), Classification (Figure 8), and VQA (Figure 9) tasks.

<sup>7</sup><https://huggingface.co/TIGER-Lab/VLM2Vec-LoRA>



### Prompt: Synthesizing Classification Data

Your mission is to first produce a detailed visual description of the image (within 300 words), identifying all potential aspects for generating high-quality data for a {image-to-text, (image,text)-to-text} classification task.

Based on the description, brainstorm a potentially useful task.

Here are a few examples for your reference: {example tasks}

Then, you should write one multi-modal classification example for this task in JSON format. The JSON object must contain the following keys:

- "description": a string, your detailed visual description, listing all required elements.
- "task\_instruction": a string, describing the classification task.
- "input\_text": {"an empty string", "a string the input text specified by the classification task"}.
- "label": a string, the correct label of the image and input\_text (if not empty) based on the task instruction.
- "misleading\_label": a string, an incorrect label that is related to the task.
- "evaluation": a string, a brief summary of the evaluation of data quality.
- "possible\_improvements": a string, suggestions for improving the data based on the guidelines.
- "revised\_task\_instruction": the revised task instruction.
- "revised\_input\_text": the revised input text, {"an empty string", "a string the input text specified by the classification task"}.
- "revised\_label": the revised label.
- "revised\_misleading\_label": the revised misleading label.

**For the description, please include the following elements:**

- General Description: Provide an overall summary of the image, including the primary objects, scene, and notable features.
- Object-Level Details: Identify the individual objects in the image, their attributes (e.g., color, size, position), and their relationships to one another.
- Contextual Features: Describe the scene or environment, including background details, lighting, and any actions taking place.
- Task-specific Brainstorming: Analyze explore how this image could relate to text (e.g., captions, contextual descriptions).

**Please adhere to the following guidelines:**

- Task should be suitable for the given image.
- Avoid generate task similar to classification of sentiment / subject / study field / genre / main topic / spam / urgency / language.
- The "input\_text" should be {"less than 10", "at least 10", "at least 50", "at least 100", "at least 200"} words and diverse in expression (if not empty).
- The "misleading\_label" must be a valid label for the given task, but not as appropriate as the "label" for the image.
- The text of "task\_instruction" should be in English and others fields should be in {language}.
- Avoid including the values of the "label" and "misleading\_label" fields in the "input\_text" (if not empty), that would make the task too easy.
- The "input\_text" (if not empty) is {"clear", "understandable with some effort", "ambiguous"} and requires {"high school", "college", "PhD"} level education to comprehend.
- **When generating the data, please evaluate the following aspects:**
  1. Relevance: Are the generated input texts and labels (if not empty) tightly connected to their corresponding image and task objectives? Does the task instruction effectively link the query image with the positive label?
  2. Plausibility: Are misleading labels sufficiently relevant to the image or labels while remaining definitively incorrect? Could they mislead the model?
  3. Clarity: Is the generated task clear and unambiguous, providing sufficient instruction to connect the query image with the label, without being overly specific or abstract?
  4. Diversity: Does the generated data introduce variation in task instructions, texts (if not empty), and labels to avoid repetitive patterns in the dataset?
- Provide a detailed evaluation of the data based on the above criteria. For each criterion, explain specific flaws or strengths.
- Suggest specific revisions to address any identified weaknesses, ensuring the revised data better aligns with the guidelines and task objectives.
- Avoid revisions that overly simplify the task instruction, text (if not empty), or labels, as this may reduce their utility for training.
- Ensure that revised data maintains consistency with the corresponding image content and classification task requirements.

Your output must always be a JSON object only. Do not explain yourself or output anything else. Be creative!

### Prompt: Synthesizing VQA Data

Your mission is to first produce detailed visual descriptions of the image (within 300 words), identifying all potential aspects for generating high-quality data for a visual QA task.

Based on the description, write one visual QA example based on the given image in JSON format. The JSON object must contain the following keys:

- "description": a string, your detailed visual description, listing all required elements.
- "question": a string, specifying the question based on the image content.
- "positive\_answer": a string, the correct answer for the question based on the image content.
- "hard\_negative\_answer": a string, an incorrect answer that appears plausible but is ultimately wrong.
- "evaluation": a string, a brief summary of the evaluation of data quality.
- "possible\_improvements": a string, suggestions for improving the data based on the guidelines.
- "revised\_question": the revised question.
- "revised\_positive\_answer": the revised positive answer.
- "revised\_hard\_negative\_answer": the revised hard negative answer.

**For the description, please include the following elements:**

- General Description: Provide an overall summary of the image, including the primary objects, scene, and notable features.
- Object-Level Details: Identify the individual objects in the image, their attributes (e.g., color, size, position), and their relationships to one another.
- Contextual Features: Describe the scene or environment, including background details, lighting, and any actions taking place.
- Task-specific Brainstorming: Analyze explore how this image could relate to text (e.g., captions, contextual descriptions).

**Please adhere to the following guidelines:**

- The "question" should be { "less than 10", "at least 10", "at least 50", "at least 100", "at least 200" } words and diverse in expression.
- The "hard\_negative\_answer" must be plausible but less appropriate than the "positive\_answer".
- The values for all fields should be in {language}.
- Avoid including explicit hints in the question that make the answer too obvious.
- The "question" (if not empty) is { "clear", "understandable with some effort", "ambiguous" } and requires { "high school", "college", "PhD" } level education to comprehend.

**When generating the data, please evaluate the following aspects:**

1. Relevance: Are the generated question and answers tightly linked to the image content and consistent with the task requirements?
  2. Plausibility: Does the "hard\_negative\_answer" closely resemble the "positive\_answer" while remaining definitively incorrect? Could it mislead the model?
  3. Diversity: Does the generated data introduce variation in questions, and answers to avoid repetitive patterns in the dataset?
- Provide a detailed evaluation of the data based on the above criteria. For each criterion, explain specific flaws or strengths.
  - Suggest specific revisions to address any identified weaknesses, ensuring the revised data better aligns with the guidelines and task objectives.
  - Avoid revisions that overly simplify or trivialize the "question".
  - Ensure revised data maintain consistency with the image content and task-specific requirements.

Your output must always be a JSON object only. Do not explain yourself or output anything else. Be creative!

### Prompt: Synthesizing Retrieval Data (Only Query Image)

Your mission is to first produce a detailed visual description of the image (within 300 words), identifying all potential aspects for generating high-quality data for a {image-to-text, (image,text)-to-text} retrieval task.

Based on the description, brainstorm a potentially useful task.

Here are a few examples for your reference: {example tasks}

Then, you should write one retrieval example for this task in JSON format. The JSON object must contain the following keys:

- "description": a string, your detailed visual description, listing all required elements.
- "task\_instruction": a string, describing the retrieval task.
- "query": {"an empty string", "a random user search query specified by the retrieval task and the query image."}
- "positive\_document": a string, the relevant document for the query image content.
- "hard\_negative\_document": a string, a hard negative document that only appears relevant to the query image content.
- "evaluation": a string, a brief summary of the evaluation of data quality.
- "possible\_improvements": a string, suggestions for improving the data based on the guidelines.
- "revised\_task\_instruction": the revised task instruction.
- "revised\_query": the revised query, {"an empty string", "a random user search query specified by the retrieval task and the query image."}.
- "revised\_positive\_document": the revised positive document, a string, the relevant document for the query image content.
- "revised\_hard\_negative\_document": the revised hard negative document, a string, a hard negative document that only appears relevant to the query image content.

**For the description, please include the following elements:**

- General Description: Provide an overall summary of the image, including the primary objects, scene, and notable features.
- Object-Level Details: Identify the individual objects in the image, their attributes (e.g., color, size, position), and their relationships to one another.
- Contextual Features: Describe the scene or environment, including background details, lighting, and any actions taking place.
- Task-specific Brainstorming: Analyze explore how this image could relate to text (e.g., captions, contextual descriptions).

**Please adhere to the following guidelines:**

- The task should involve both query and documents (positive and hard negative, if not empty). It must directly indicate the relation without being overly detailed or abstract.
- The query (if not empty) should be {"extremely long-tail", "long-tail", "common"}, {"less than 5 words", "5 to 15 words", "at least 10 words"}, {"clear", "understandable with some effort", "ambiguous"}, and diverse in topic.
- All documents (if not empty) must be created independent of the query. Avoid copying the query verbatim. It's acceptable if some parts of the "positive\_document" are not topically related to the query.
- All documents (if not empty) should be at least {"10", "30", "200", "300"} words long.
- The "hard\_negative\_document" (if not empty) contains some useful information, but it should be less useful or comprehensive compared to the "positive\_document".
- The text of "task\_instruction" should be in English and others fields should be in {language}.
- Do not provide any explanation in any document (if not empty) on why it is relevant or not relevant to the query.
- Do not use the word "query" or "document" in the generated content.
- Both the query and documents (if not empty) require {"high school", "college", "PhD"} level education to understand.
- **When generating the data, please evaluate the following aspects:**
  1. Relevance: Are the generated query and documents (if not empty) tightly connected to their corresponding image and task objectives? Does the task instruction effectively link the query image with the positive text?
  2. Plausibility: Are hard negatives sufficiently similar to the query or positive examples while remaining definitively incorrect? Could they mislead the model?
  3. Clarity: Is the generated task clear and unambiguous, providing sufficient instruction to connect the query image with the positive document, without being overly specific or abstract?
  4. Diversity: Does the generated data introduce variation in task instructions, queries, and documents to avoid repetitive patterns in the dataset?
- Provide a detailed evaluation of the data based on the above criteria. For each criterion, explain specific flaws or strengths.
- Suggest specific revisions to address any identified weaknesses, ensuring the revised data better aligns with the guidelines and task objectives.
- Avoid revisions that overly simplify the task instruction, query, or documents, as this may reduce their utility for training.
- Ensure that revised data maintains consistency with the corresponding image content and retrieval task requirements.

Your output must always be a JSON object only. Do not explain yourself or output anything else. Be creative!

## Prompt: Synthesizing Retrieval Data (With Document Images)

Your mission is to first produce detailed visual descriptions of the images (within 600 words), identifying all potential aspects for generating high-quality data for a `{(image,text)-to-image, image-to-image, (image,text)-to-(image,text), text-to-image, text-to-(image,text)}` retrieval task that involves both query and document images.

Based on the description, brainstorm a potentially useful task.

Here are a few examples for your reference: `{example tasks}`

Then, you should write one retrieval example for this task in JSON format. The JSON object must contain the following keys:

- "description": a string, your detailed visual description, listing all required elements.
- "task\_instruction": a string, describing the retrieval task.
- "query": `{"an empty string", "a random user search query specified by the retrieval task and the query image."}`
- "positive\_document": `{"an empty string", "a string, the relevant document for the query based on the query text and image content"}`
- "hard\_negative\_document": `{"an empty string", "a string, a hard negative document that only appears relevant to the query and the query image content."}`
- "evaluation": a string, a brief summary of the evaluation of data quality.
- "possible\_improvements": a string, suggestions for improving the data based on the guidelines.
- "revised\_task\_instruction": the revised task instruction.
- "revised\_query": the revised query, `{"an empty string", "a random user search query specified by the retrieval task and the query image."}`.
- "revised\_positive\_document": the revised positive document, a string, `{"an empty string", "a string, the relevant document for the query based on the query text and image content"}`
- "revised\_hard\_negative\_document": the revised hard negative document, `{"an empty string", "a string, a hard negative document that only appears relevant to the query and the query image content."}`

**For the description, please include the following elements:**

- General Description: Provide an overall summary of the image, including the primary objects, scene, and notable features.
- Object-Level Details: Identify the individual objects in the image, their attributes (e.g., color, size, position), and their relationships to one another.
- Contextual Features: Describe the scene or environment, including background details, lighting, and any actions taking place.
- Task-specific Brainstorming: Analyze explore how this image could relate to text (e.g., captions, contextual descriptions).

**Please adhere to the following guidelines:**

- The task must connect the query image and positive image through their content. It must directly indicate the relation without being overly detailed or abstract.
- The query (if not empty) should be `{"extremely long-tail", "long-tail", "common"}`, `{"less than 5 words", "5 to 15 words", "at least 10 words"}`, `{"clear", "understandable with some effort", "ambiguous"}`, and diverse in topic.
- The query (if not empty) should effectively associate the query image with the positive image.
- All documents (if not empty) must be created independent of the query. Avoid copying the query verbatim. It's acceptable if some parts of the "positive\_document" are not topically related to the query.
- All documents (if not empty) should be at least `{"10", "30", "200", "300"}` words long.
- The "hard\_negative\_document" (if not empty) contains some useful information, but it should be less useful or comprehensive compared to the "positive\_document".
- The text of "task\_instruction" should be in English and others fields should be in `{language}`.
- Do not provide any explanation in any document (if not empty) on why it is relevant or not relevant to the query.
- Do not use the word "query" or "document" in the generated content.
- Both the query and documents (if not empty) require `{"high school", "college", "PhD"}` level education to understand.
- **When generating the data, please evaluate the following aspects:**
  1. Relevance: Are the generated query and documents (if present) tightly linked to their corresponding images? Does the task instruction effectively connect the query image to the positive image?
  2. Plausibility: Are the negative examples, including hard negatives, realistic and similar enough to the positive image to challenge the model, while still being definitively incorrect?
  3. Clarity: Is the generated task clear and unambiguous, providing sufficient instruction to connect the query image with the positive image, without being overly specific or abstract?
  4. Diversity: Does the generated data introduce variation in task instructions, queries, and documents to avoid repetitive patterns in the dataset?
- Provide a detailed evaluation of the data based on the above criteria. For each criterion, explain specific flaws or strengths.
- Suggest specific revisions to address any identified weaknesses, ensuring the revised data better aligns with the guidelines and task objectives.
- Avoid revisions that overly simplify the task or create unrealistic connections between the query and positive image.
- Ensure that revised data maintains consistency with the corresponding image content and retrieval task requirements.

Your output must always be a JSON object only. Do not explain yourself or output anything else. Be creative!



### An Example of Synthetic Retrieval IT2IT Data (Description & Initial Data)




Images (Query; Positive; Hard Negative)	  
Description	<p><b>The first image</b> portrays a vibrant outdoor scene of a powerful waterfall cascading down a ledge surrounded by lush foliage. The waterfall is centrally located, with a stream flowing in the foreground over several rocks and branches. The image captures a vivid display of sunlight filtering through dense greenery, with bright shades of green and the mist from the waterfall adding to the natural beauty. The scene suggests a secluded, serene environment.</p> <p><b>The second image</b> presents a cascading waterfall, framed in monochrome, offering a sense of timelessness and tranquility. The waterfall descends through rocky formations before settling into a small pool at the bottom. The foreground features smooth rocks, polished by flowing water, which lends texture and depth to the black-and-white photograph. The dark shadows and bright highlights create stark contrasts, enhancing the visual impact. The surrounding area is densely wooded, indicating a secluded spot within a forest.</p> <p><b>The third image</b> displays a vividly colorful landscape dominated by an expansive lake flanked by tall, rugged cliffs. The cliffs exhibit a range of warm hues, from deep reds to oranges, basked in the glow of a low-hanging sun. The sky is a brilliant blue with scattered clouds, which together with the reflections on the water, create a picturesque scene. Reeds in the foreground frame the serene water, whose stillness mirrors the sky and cliffs, suggesting peace and prismatic beauty.</p>
Task Instruction	Retrieve an image-description pair depicting cascading water features in different lighting and weather conditions.
Query	Picturesque waterfall with vibrant forest and sunlight
Positive document	This image captures a waterfall in a forested area, presented in a monochrome format. It highlights the natural setting with cascading water over rocky terrain, creating an air of serenity. The use of black and white emphasizes light contrasts and textures of the water and surroundings, offering a classic, peaceful portrayal of nature unaffected by human interference, emphasizing timeless tranquility.
Hard negative document	In this photograph, a beautiful lake sits surrounded by striking cliffs under a bright blue sky. The cliffs are illuminated by the sun, showcasing a range of warm colors that contrast with clear blue reflections on the water surface. Framing the lake are tall reeds, contributing to a sense of isolation and calm are very good.

Figure 6: An example of the synthetic Retrieval IT2IT data (part 1). This part includes the input images, the multi-aspect descriptions, and the initially generated data.

An Example of Synthetic Retrieval IT2IT Data (Evaluation & Refinement)	
Evaluation	<p><b>Relevance:</b> The task effectively links the images through their feature of cascading water though the third doesn't fit.</p> <p><b>Plausibility:</b> The negative stands apart from the cascading water concept but holds some scenic elements.</p> <p><b>Clarity:</b> The task is clearly defined.</p> <p><b>Diversity:</b> There is a good variation in image scenes using water as a theme.</p>
Possible Improvements	Ensure the negative image relates closer to cascading water but under different weather conditions or lighting. Provide more differentiation between positive and negative descriptions. Revise watermark elements for image clarity.
Revised Task Instruction	Retrieve an image-description pair depicting waterfalls in different lighting or weather conditions.
Revised Query	Vibrant sunlight waterfall with lush surrounding forest
Revised Positive document	This image captures a waterfall in a forested area, presented in a monochrome format. It highlights the natural setting with cascading water over rocky terrain, creating an air of serenity. The use of black and white emphasizes light contrasts and textures of the water and surroundings, offering a classic, peaceful portrayal of nature unaffected by human interference, emphasizing timeless tranquility.
Revised Hard negative document	The photo showcases a breathtaking desert landscape under a vibrant sky. The stillness of the clear, reflective lake beneath towering orange-red cliffs suggests peace. Foregrounding tall reeds framing this water body adds depth without cascading water, contrasting yet resembling a waterfall.

Figure 7: An example of the synthetic Retrieval IT2IT data (part 2). This part includes the evaluation, possible improvements, and the revised data.

## An Example of Synthetic Classification IT2T Data


Image (Query)	
Description	The image is a movie poster featuring a dramatic scene with a dusky, smoky background. The foreground shows a man in a sharp, dark suit and a woman in a leather outfit, both shown from the waist up. They are positioned closely, facing the viewer. Above them, the text 'NICOLAS CAGE' is prominently displayed. The background depicts a dystopian, smoky landscape with industrial structures, flames, and dark billowing smoke, giving an impression of chaos and destruction. The middle lower section features the title 'THE HUMANITY BUREAU' in bold, white capital letters. Below the title, there are smaller texts with names 'Sarah Lind', 'Hugh Dillon', and more, alongside the phrase 'Escape from New America'. The color scheme is predominantly dark, with flashes of fiery orange.
Task Instruction	Identify the genre this movie poster belongs to
Input Text	Destruction backdrop
Label	Science Fiction
Misleading Label	Romantic Comedy
Evaluation	<p><b>Relevance:</b> The poster's imagery and text signaling a dystopian theme are closely related to the 'Science Fiction' label.</p> <p><b>Plausibility:</b> The misleading label 'Romantic Comedy' might be incorrect due to the intense imagery of destruction, however, it still bears a relationship to movie genres.</p> <p><b>Clarity:</b> The task is clear in instructing the identification of film genre based on visual cues.</p> <p><b>Diversity:</b> The data introduces variety in illustration, steering clear of repetitive genre classification.</p>
Possible Improvements	To enhance data quality, add more challenging input text variations to test genre recognition effectiveness and to diversify task styles.
Revised Task Instruction	Determine the movie genre based on the visual elements in the poster.
Revised Input Text	Chaotic landscape scene
Revised Label	Science Fiction
Revised Misleading Label	Adventure

Figure 8: An example of the synthetic Classification IT2T data.

### An Example of Synthetic VQA IT2T Data


Image (Query)	
Description	The image features a young boy sitting in a rustic wooden crate filled with apples, with a specific focus on him biting into a green apple. The boy's light blonde, wavy hair frames his face as he appears engaged in eating. He is dressed in a striped shirt with blue, green, and yellow hues, paired with blue jeans, adding a playful and casual vibe. The wooden crate, slightly weathered and grayish, holds several red and green apples, emphasizing the theme of harvest. Surrounding the boy are lush apple trees, their branches laden with ripe red and green apples and vibrant green leaves. The background is filled with soft sunlight filtering through the orchard, creating a bright and cheerful setting. The natural tones of green and red apples contrast with the boy's colorful outfit, making him the focal point. The scene suggests a connection to outdoor activities like apple picking and highlights the freshness of rural environments.
Question	In the image, there is a child dressed in a brightly striped shirt standing near a large container filled with colorful fruits. Based on the scene depicted, what type of fruit is the child likely surrounded by, considering the visible tree and its fruit characteristics?
Positive Answer	Apples
Negative Answer	Bananas
Evaluation	<b>Relevance</b> of the question and answers is high, focusing specifically on the child and the surrounding fruits, while maintaining consistency with the image content. The hard negative answer, although certainly wrong, could mislead due to its <b>plausible</b> nature, but overall is less fitting given the absence of bananas in the scene. The data displays <b>diversity</b> through detailed description without repetitive elements.
Possible Improvements	Consider more deeply associating the description of the tree's characteristics to add context that makes the negative answer slightly more challenging, without sacrificing the overall accuracy of context. Ensuring a balance between misleading potential and implausibility would enhance this data further.
Revised Question	What type of fruit, seen in the trees and collected in the container, is prominently featured around the child who is standing next to the wooden box in this lush outdoor setting?
Revised Positive Answer	Apples
Revised Negative Answer	Oranges

Figure 9: An example of the synthetic VQA IT2T data.



## E Detailed Results

In this section, we present the detailed comparisons of mmE5 to baseline models on both zero-shot and supervised finetuning settings on the MMEB benchmark (Jiang et al., 2024b). Due to space limitation, we omit the detailed results of partially supervised finetuning models.

Task	Zero-shot Setting Models							Supervised Finetuning Models			
	CLIP	OpenCLIP	SigLIP	BLIP2	MagicLens	E5-V	MMRet	mmE5	VLM2Vec	MMRet	mmE5
<b>Classification (10 tasks)</b>											
ImageNet-1K	55.8	63.5	45.4	10.3	48.0	9.6	49.1	68.8	74.5	58.8	77.6
N24News	34.7	38.6	13.9	36.0	33.7	23.4	45.8	54.5	80.3	71.3	82.1
HatefulMemes	51.1	51.7	47.2	49.6	49.0	49.7	51.0	55.0	67.9	53.7	64.3
VOC2007	50.7	52.4	64.3	52.1	51.6	49.9	74.6	73.9	91.5	85.0	91.0
SUN397	43.4	68.8	39.6	34.5	57.0	33.1	60.1	72.7	75.8	70.0	77.9
Place365	28.5	37.8	20.0	21.5	31.5	8.6	35.3	39.7	44.0	43.0	42.6
ImageNet-A	25.5	14.2	42.6	3.2	8.0	2.0	31.6	46.1	43.6	36.1	56.7
ImageNet-R	75.6	83.0	75.0	39.7	70.9	30.8	66.2	86.2	79.8	71.6	86.3
ObjectNet	43.4	51.4	40.3	20.6	31.6	7.5	49.2	74.8	39.6	55.8	62.2
Country-211	19.2	16.8	14.2	2.5	6.2	3.1	9.3	35.1	14.7	14.7	34.8
All Classification	42.8	47.8	40.3	27.0	38.8	21.8	47.2	60.7	61.2	56.0	67.6
<b>VQA (10 tasks)</b>											
OK-VQA	7.5	11.5	2.4	8.7	12.7	8.9	28.0	56.6	69.0	73.3	67.9
A-OKVQA	3.8	3.3	1.5	3.2	2.9	5.9	11.6	50.0	54.4	56.7	56.4
DocVQA	4.0	5.3	4.2	2.6	3.0	1.7	12.6	81.3	52.0	78.5	90.3
InfographicsVQA	4.6	4.6	2.7	2.0	5.9	2.3	10.6	44.0	30.7	39.3	56.2
ChartQA	1.4	1.5	3.0	0.5	0.9	2.4	2.4	35.2	34.8	41.7	50.3
Visual7W	4.0	2.6	1.2	1.3	2.5	5.8	9.0	40.4	49.8	49.5	51.9
ScienceQA	9.4	10.2	7.9	6.8	5.2	3.6	23.3	47.3	42.1	45.2	55.7
VizWiz	8.2	6.6	2.3	4.0	1.7	2.6	25.9	54.0	43.0	51.7	52.8
GQA	41.3	52.5	57.5	9.7	43.5	7.8	41.3	65.4	61.2	59.0	62.1
TextVQA	7.0	10.9	1.0	3.3	4.6	3.2	18.9	83.1	62.0	79.0	83.5
Avg.	9.1	10.9	8.4	4.2	8.3	4.9	18.4	55.7	49.9	57.4	62.7
<b>Retrieval (12 tasks)</b>											
VisDial	30.7	25.4	21.5	18.0	24.8	9.2	62.6	39.1	80.9	83.0	73.7
CIRR	12.6	15.4	15.1	9.8	39.1	6.1	65.7	41.6	49.9	61.4	54.9
VisualNews_t2i	78.9	74.0	51.0	48.1	50.7	13.5	45.7	51.2	75.4	74.2	77.7
VisualNews_i2t	79.6	78.0	52.4	13.5	21.1	8.1	53.4	64.9	80.0	78.1	83.4
MSCOCO_t2i	59.5	63.6	58.3	53.7	54.1	20.7	68.7	55.0	75.7	78.6	76.2
MSCOCO_i2t	57.7	62.1	55.0	20.3	40.0	14.0	56.7	59.1	73.1	72.4	73.6
NIGHTS	60.4	66.1	62.9	56.5	58.1	4.2	59.4	58.9	65.5	68.3	68.8
WebQA	67.5	62.1	58.1	55.4	43.0	17.7	76.3	82.9	87.6	90.2	88.1
FashionIQ	11.4	13.8	20.1	9.3	11.2	2.8	31.5	21.6	16.2	54.9	28.6
Wiki-SS-NQ	55.0	44.6	55.1	28.7	18.7	8.6	25.4	58.8	60.2	24.9	65.2
OVEN	41.1	45.0	56.0	39.5	1.6	5.9	73.0	67.6	56.5	87.5	77.3
EDIS	81.0	77.5	23.6	54.4	62.6	26.8	59.9	55.2	87.8	65.6	83.6
Avg.	53.0	52.3	31.6	33.9	35.4	11.5	56.5	54.7	67.4	69.9	71.0
<b>Visual Grounding (4 tasks)</b>											
MSCOCO	33.8	34.5	46.4	28.9	22.1	10.8	42.7	59.0	80.6	76.8	85.0
RefCOCO	56.9	54.2	70.8	47.4	22.8	11.9	69.3	78.9	88.7	89.8	92.7
RefCOCO-matching	61.3	68.3	50.8	59.5	35.6	38.9	63.2	80.8	84.0	90.6	88.9
Visual7W-pointing	55.1	56.3	70.1	52.0	23.4	14.3	73.5	71.2	90.9	77.0	92.3
Avg.	51.8	53.3	59.5	47.0	26.0	19.0	62.2	72.5	86.1	83.6	89.7
<b>Final Score (36 tasks)</b>											
All IND Avg.	37.1	39.3	32.3	25.3	31.0	14.9	43.5	57.2	67.5	59.1	72.4
All OOD Avg.	38.7	40.2	38.0	25.1	23.7	11.5	44.3	60.4	57.1	68.0	66.6
All Avg.	37.8	39.7	34.8	25.2	27.8	13.3	44.0	58.6	62.9	64.1	69.8

Table 7: Detailed results of zero-shot setting and supervised setting models on each dataset of MMEB (Jiang et al., 2024b).