VLM2VEC: TRAINING VISION-LANGUAGE MODELS FOR MASSIVE MULTIMODAL EMBEDDING TASKS

Ziyan Jiang[1*], **Rui Meng**[2], **Xinyi Yang**[2], **Semih Yavuz**[2], **Yingbo Zhou**[2], **Wenhu Chen**[1]
[1]University of Waterloo, [2]Salesforce Research
ziyanjiang528@gmail.com, ruimeng@salesforce.com, wenhuchen@uwaterloo.ca

https://tiger-ai-lab.github.io/VLM2Vec/

ABSTRACT

Embedding models play a crucial role in a variety of downstream tasks, including semantic similarity, information retrieval, and clustering. While there has been a surge of interest in developing universal text embedding models that generalize across tasks (e.g., MTEB), progress in learning universal multimodal embedding models has been comparatively slow, despite their importance and practical applications. In this work, we explore the potential of building universal multimodal embeddings capable of handling a broad range of downstream tasks. Our contributions are two fold: (1) we propose MMEB (Massive Multimodal Embedding Benchmark), which covers 4 meta-tasks (i.e. classification, visual question answering, multimodal retrieval, and visual grounding) and 36 datasets, including 20 training datasets and 16 evaluation datasets covering both in-distribution and out-of-distribution tasks, and (2) VLM2VEC (Vision-Language Model → Vector), a contrastive training framework that converts any vision-language model into an embedding model via contrastive training on MMEB. Unlike previous models such as CLIP and BLIP, which encode text and images independently without task-specific guidance, VLM2VEC can process any combination of images and text while incorporating task instructions to generate a fixed-dimensional vector. We develop a series of VLM2VEC models based on state-of-the-art VLMs, including Phi-3.5-V, LLaVA-1.6, and Qwen2-VL, and evaluate them on MMEB's benchmark. With LoRA tuning, VLM2VEC achieves a 10% to 20% improvement over existing multimodal embedding models on MMEB's evaluation sets. Our findings reveal that VLMs are surprisingly strong embedding models.

1 INTRODUCTION

Embeddings, or distributed representations, encode inputs (whether text or images) as fixed-dimensional vectors, enabling a range of downstream tasks. Since the advent of Word2Vec (Mikolov, 2013) and GloVe (Pennington et al., 2014), substantial research efforts have focused on learning textual embeddings (Kiros et al., 2015; Conneau et al., 2017) and image embeddings (Radford et al., 2021; Li et al., 2022; Jia et al., 2021; Yu et al., 2022). These embeddings facilitate a variety of applications, including textual and visual semantic similarity (Agirre et al., 2012; Marelli et al., 2014; Chechik et al., 2010; Cer et al., 2017), information retrieval (Mitra et al., 2017; Karpukhin et al., 2020; Lin et al., 2014), automatic evaluation (Zhang et al., 2020; Sellam et al., 2020), prompt retrieval for in-context learning (Liu et al., 2022; Rubin et al., 2022; Hongjin et al., 2022), and retrieval-augmented generation (Lewis et al., 2020; Guu et al., 2020; Izacard & Grave, 2020). A recent shift in research has focused on developing universal embeddings that can generalize across a wide range of tasks. For instance, Muennighoff et al. (2023) introduced MTEB (Massive Text Embedding Benchmark) to comprehensively assess text embeddings across tasks such as classification and clustering. MTEB has become the standard for evaluating universal text embeddings. Recent works (Wang et al., 2022a; Su et al., 2023; Wang et al., 2024a; Springer et al., 2024; BehnamGhader et al., 2024) have demonstrated promising results on the MTEB benchmark. However, progress in multimodal embeddings has been relatively slower. Despite advancements

[*]Work done during an internship at University of Waterloo in collaboration with Salesforce Research. Corresponding authors are Ziyan Jiang, Rui Meng and Wenhu Chen

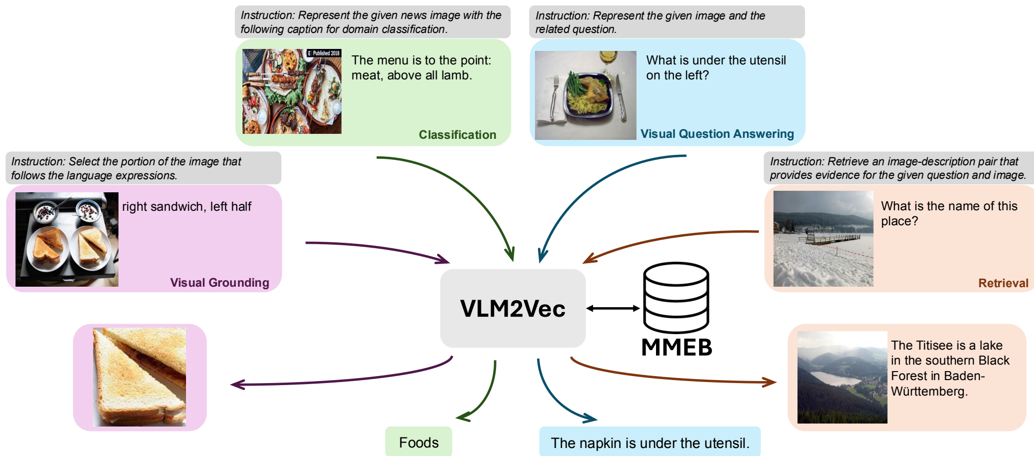

Figure 1: We develop a universal multimodal embedding benchmark, MMEB, along with VLM2VEC, an embedding model adapted from vision-language models (VLMs). VLM2VEC is capable of following instructions and performing various multimodal embedding tasks, accommodating any combination of image and text modalities.

in text embeddings, the lack of both benchmarks and methodologies in the multimodal embedding domain remains a challenge.

Current research in multimodal embeddings faces two primary limitations: (1) existing studies typically evaluate visual embeddings on isolated tasks, such as ImageNet classification (Deng et al., 2009; Hendrycks et al., 2021a;b) or MSCOCO/Flickr retrieval (Lin et al., 2014; Plummer et al., 2015); (2) most existing models, such as CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and SigLIP (Zhai et al., 2023), either process text and images separately or perform shallow fusion of visual and textual information (Wei et al., 2023), limiting their ability to fully capture the relationships between text and image modalities. Furthermore, these models exhibit limited reasoning and generalization capabilities, particularly in zero-shot scenarios for complex reasoning tasks.

In this paper, we attempt to build an universal multimodal embedding framework to pave road for the future research, which consists of two efforts:

- MMEB: We introduce a novel benchmark, MMEB (Massive Multimodal Embedding Benchmark), which includes 36 datasets spanning four meta-task categories: classification, visual question answering, retrieval, and visual grounding. MMEB provides a comprehensive framework for training and evaluating embedding models across various combinations of text and image modalities. All tasks are reformulated as ranking tasks, where the model follows instructions, processes a query, and selects the correct target from a set of candidates. The query and target can be an image, text, or a combination of both. MMEB is divided into 20 in-distribution datasets, which can be used for training, and 16 out-of-distribution datasets, reserved for evaluation.

- VLM2VEC: We adopt the pre-trained vision-language models like Phi-3.5-V (Abdin et al., 2024) and LLaVA-1.6 (Li et al., 2024) as the backbone for VLM2VEC. In contrast to other multimodal embedding models like UniIR (Wei et al., 2023) and MagicLens (Zhang et al., 2024), which rely on late fusion of CLIP (Radford et al., 2021) features, our approach leverages the deep integration of vision and language features within a transformer architecture. There are several advantages to this approach: (1) VLMs are trained on massive multimodal datasets and can handle any combination of images and text, as well as high-resolution images and long text inputs; (2) vision and language features are deeply fused in the transformer model, improving the model's ability to capture cross-modal relationships; and (3) these models are well-suited for generalizing across diverse tasks, particularly those requiring instruction-following capabilities. These factors make VLM2VEC an ideal choice for task generalization. We trained VLM2VEC on the 20 MMEB training datasets using contrastive learning and compared its performance with various baselines.

Following extensive contrastive training, **VLM2VEC can handle any combination of images and text, producing fixed-dimensional vectors**. We evaluate VLM2VEC against a wide array of multimodal embedding models, including CLIP (Radford et al., 2021), BLIP2 (Li et al., 2023a), SigLIP (Zhai et al., 2023), MagicLens (Zhang et al., 2024), UniIR (Wei et al., 2023) and E5-V (Jiang et al., 2024), demonstrating consistent improvements across all task categories. Notably, compared to the best baseline model without fine-tuning, our model achieves a 21.1 point improvement (from 44.7 to 65.8) across all 36 MMEB datasets and a 16.1-point increase (from 41.7 to 57.8) on 16 out-of-distribution datasets for zero-shot evaluation. Compared to the best baseline model with fine-tuning, our model achieves a 18.6 point improvement (from 47.2 to 65.8) across all 36 MMEB datasets and a 14.7-point increase (from 43.1 to 57.8) on 16 out-of-distribution datasets for zero-shot evaluation. Moreover, as a general multimodal representation model, VLM2VEC can still achieve competitive zero-shot T2I (Text-to-Image) and I2T (Image-to-Text) performance on Flickr30K compared to existing CLIP-like models, as presented in Table 11.

2 MMEB: A BENCHMARK FOR MULTIMODAL EMBEDDINGS

2.1 DATASET OVERVIEW

We present MMEB (Massive Multimodal Embedding Benchmark), a comprehensive benchmark designed to evaluate multimodal embeddings across a diverse set of tasks. MMEB consists of 36 datasets organized into four meta-tasks: classification, visual question answering, retrieval, and visual grounding. Each task is reformulated as a ranking problem, where the model is provided with an instruction and a query (which may consist of text, images, or both) and is tasked with selecting the correct answer from a set of candidates. These candidates could be text, images, or additional instructions. The datasets are divided into two categories: 20 in-distribution datasets for training and 16 out-of-distribution datasets for evaluation. We report performance metrics across all 36 tasks. An overview of MMEB is provided in Figure 2 and the dataset statistics are provided in Table 1.

The embedding models are supposed to compress the query side into a vector and the target candidates into a set of vectors. The candidate with the highest dot-product will be selected as the prediction for evaluation. We measure the Precision@1 to reflect the percentage of top candidate matching the groundtruth. For the number of target candidates, a higher count could increase evaluation costs and hinder rapid model iteration, while a lower count might make the benchmark too simple and prone to saturation. To strike a balance between these extremes, we have chosen 1,000 candidates. Further details about this decision can be found in Section A.2.

MMEB offers a wide range of tasks from various domains, such as common, news, Wikipedia, web, and fashion. The benchmark incorporates diverse combinations of modalities for both queries and targets, including text, images, and text-image pairs. Additionally, tasks are designed to follow different types of instructions. For instance, tasks may involve object recognition (e.g., *"Identify the object shown in the image."*), retrieval (e.g., *"Find an image that matches the given caption."*), or visual grounding (e.g., *"Select the portion of the image that answers the question."*). Examples for each dataset in MMEB are provided in Tables 7, 8, 9 and 10. The diversity in MMEB makes it an ideal testbed for universal embeddings.

2.2 META-TASK AND DATASET DESIGN

MMEB is organized into four primary meta-task categories:

Classification The query consists of an instruction, an image, optionally accompanied by related text, while the target is the class label. The number of candidates equals the number of classes.

Visual Question Answering The query consists of an instruction, an image, and a piece of text as the question, while the target is the answer. Each query has 1 ground truth and 999 distractors as candidates.

Information Retrieval Both the query and target sides can involve a combination of text, images, and instructions. Each query has 1 ground truth and 999 distractors as candidates.

Visual Grounding The category is adapted from object detection tasks. The query combines an instruction (e.g., "Select the portion of the image that isolates the object of the given label: red apple") with the full image. This instruction guides the model to focus on a specific object within the image. Each candidate corresponds to cropped regions (bounding boxes) of the image, including

Table 1: The statistics of MMEB: 36 datasets across 4 meta-task categories, with 20 in-distribution datasets used for training and 16 out-of-distribution datasets used exclusively for evaluation.

Meta-Task	Dataset	Query	Target	OOD?	#Training	#Eval	#Candidates
Classification (10 Tasks)	ImageNet-1K	I	T		100K	1000	1000
	N24News	I + T	I		49K	1000	24
	HatefulMemes	I	T		8K	1000	2
	VOC2007	I	T		8K	1000	20
	SUN397	I	T		20K	1000	397
	Place365	I	T	✓	-	1000	365
	ImageNet-A	I	T	✓	-	1000	1000
	ImageNet-R	I	T	✓	-	1000	200
	ObjectNet	I	T	✓	-	1000	313
	Country-211	I	T	✓	-	1000	211
VQA (10 Tasks)	OK-VQA	I + T	T		9K	1000	1000
	A-OKVQA	I + T	T		17K	1000	1000
	DocVQA	I + T	T		40K	1000	1000
	InfographicVQA	I + T	T		24K	1000	1000
	ChartQA	I + T	T		28K	1000	1000
	Visual7W	I + T	T		70K	1000	1000
	ScienceQA	I + T	T	✓	-	1000	1000
	VizWiz	I + T	T	✓	-	1000	1000
	GQA	I + T	T	✓	-	1000	1000
	TextVQA	I + T	T	✓	-	1000	1000
Retrieval (12 Tasks)	VisDial	T	I		123K	1000	1000
	CIRR	I + T	I		26K	1000	1000
	VisualNews_t2i	T	I		100K	1000	1000
	VisualNews_i2t	I	T		100K	1000	1000
	MSCOCO_t2i	T	I		100K	1000	1000
	MSCOCO_i2t	I	T		113K	1000	1000
	NIGHTS	I	I		16K	1000	1000
	WebQA	T	I + T		17K	1000	1000
	OVEN	I + T	I + T	✓	-	1000	1000
	FashionIQ	I + T	I	✓	-	1000	1000
	EDIS	T	I + T	✓	-	1000	1000
	Wiki-SS-NQ	T	I	✓	-	1000	1000
Visual Grounding (4 Tasks)	MSCOCO	I + T	I		100K	1000	1000
	Visual7W-Pointing	I + T	I	✓	-	1000	1000
	RefCOCO	I + T	I	✓	-	1000	1000
	RefCOCO-Matching	I + T	I + T	✓	-	1000	1000

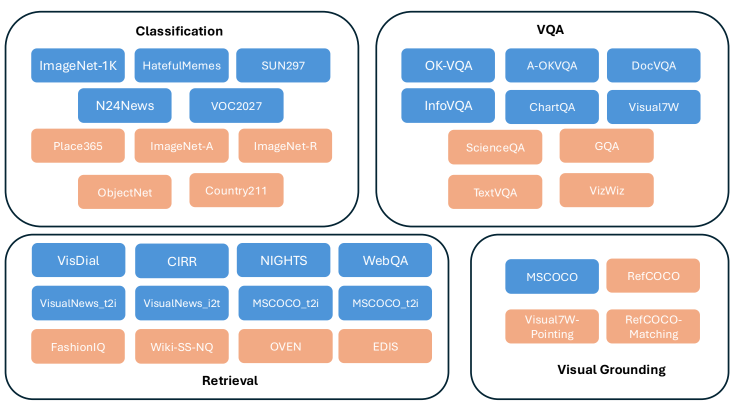

Figure 2: An overview of the tasks and datasets in MMEB. MMEB includes four meta-tasks and 36 datasets: 20 in-distribution datasets (blue) used for training and 16 out-of-distribution (orange) datasets used exclusively for evaluation.

both the object of interest and distractor regions. Each query includes 1,000 candidates: 1 ground truth and 999 distractors. These distractors may include hard negatives from the same object class, other objects in the image, or random objects from different images.

Further details on dataset processing can be found in Section A.1.

3 VLM2VEC: TRANSFORMING LVMS TO EMBEDDERS

3.1 CONTRASTIVE TRAINING

We develop VLM2VEC, a contrastive training framework designed to convert any state-of-the-art vision-language model into an embedding model, as illustrated in Figure 3. A relevant query-target pair is denoted as (q, t^+). Both q and t^+ could be either single image, text or single image + text. We define $q : (q_t, q_i)$ and $t^+ : (t_t^+, t_i^+)$.

We then apply the instruction to the original query q to generate a new one q_{inst}:

$$q_{\text{inst}} = \texttt{[IMAGE\_TOKEN]}\,\text{Instruct: }\{\textit{task\_definition}\}\ \backslash n\ \text{Query: }\{q\} \tag{1}$$

where "{*task_definition*}" is a placeholder for a one-sentence description of the embedding task. To enhance the embedding model's generalizability by better understanding instructions, we have crafted task-specific instructions, as shown in Tables 7, 8, 9 and 10.

Given a pretrained VLM, we feed query and target into it to obtain the query and target embeddings $(\mathbf{h}_{q_{\text{inst}}}, \mathbf{h}_{t^+})$ by taking the last layer vector representation of the last token. To train the embedding model, we adopt the standard InfoNCE loss $\mathcal{L}$ over the in-batch negatives and hard negatives:

$$\min\ \ \mathcal{L} = -\log \frac{\phi(\mathbf{h}_{q_{\text{inst}}}, \mathbf{h}_{t^+})}{\phi(\mathbf{h}_{q_{\text{inst}}}, \mathbf{h}_{t^+}) + \sum_{t^- \in \mathbb{N}} \phi(\mathbf{h}_{q_{\text{inst}}}, \mathbf{h}_{t^-})} \tag{2}$$

where $\mathbb{N}$ denotes the set of all negatives, and $\phi(\mathbf{h}_q, \mathbf{h}_t)$ is a function that computes the matching score between query q and target t. In this paper, we adopt the temperature-scaled cosine similarity function as $\phi(\mathbf{h}_q, \mathbf{h}_t) = \exp(\frac{1}{\tau}\cos(\mathbf{h}_q, \mathbf{h}_t))$,where τ is a temperature hyper-parameter.

3.2 INCREASING BATCH SIZE THROUGH GRADCACHE

Since hard negatives are often difficult or ambiguous to collect for most multimodal datasets, using larger batch sizes becomes crucial. This increases the number of in-batch random negatives, which in turn helps improve the performance of the embedding model.

A bottleneck lies in the GPU memory that limits us from increasing the batch size and the number of in-batch random negatives during training, as each training instance may include one image (either from the query or target side) or multiple images (from both query and target sides), resulting in substantial memory consumption. We apply GradCache (Gao et al., 2021a), a gradient caching technique that decouples backpropagation between contrastive loss and the encoder, removing encoder backward pass data dependency along the batch dimension.

Mathematically, supposed we have a large batch of queries $\mathcal{Q}$, and we divide it into a set of sub-batches, each of which can fit into memory for gradient computation: $\mathcal{Q} = \{\hat{Q}_1, \hat{Q}_2, \ldots\}$. There are two major steps: "Representation Gradient Computation and Caching" and "Sub-batch Gradient Accumulation". First, gradient tensors within each subbatch is calculated and stored: $\mathbf{u}_i = \frac{\partial \mathcal{L}}{\partial f(q_i)}, \quad \mathbf{v}_i = \frac{\partial \mathcal{L}}{\partial f(d_i)}$.

Then gradients are accumulated for encoder parameters across all sub-batches:

$$\frac{\partial \mathcal{L}}{\partial \Theta} = \sum_{\hat{Q}_j \in \mathbb{Q}} \sum_{q_i \in \hat{Q}_j} \mathbf{u}_i \frac{\partial f(q_i)}{\partial \Theta} + \sum_{\hat{D}_j \in \mathbb{D}} \sum_{d_i \in \hat{D}_j} \mathbf{v}_i \frac{\partial f(d_i)}{\partial \Theta} \tag{3}$$

4 EXPERIMENTS

In this section, we adopt Phi-3.5-V (Abdin et al., 2024), LLaVA-1.6 (Li et al., 2024) and Qwen2-VL-7B-Instruct (Wang et al., 2024b) as the backbone VLMs, with training conducted via either full model fine-tuning or LoRA. The temperature for the loss function is set to 0.02, with a batch size of 1,024, a maximum text length of 256 tokens, and 2K training steps. The LoRA variant uses a rank of 8. For VLM2VEC leveraging Phi-3.5-V as the backbone, we configure the number of sub-image

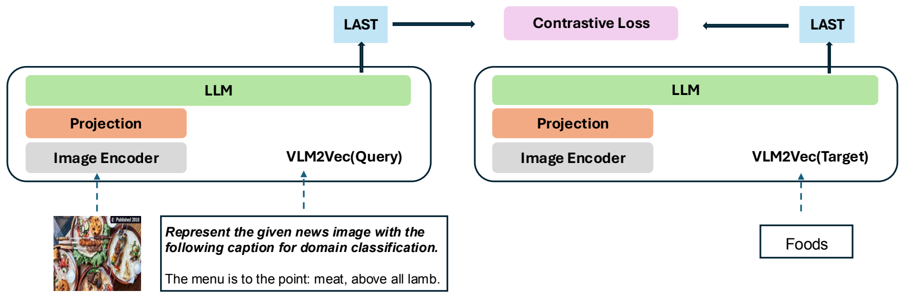

Figure 3: VLM2VEC uses a VLM as the backbone to deeply integrate image and text features. It is trained with a contrastive loss between the query and target, following task-specific instructions. The training data consists of diverse combinations of modalities on both the query and target sides, which may include images, text, or image-text pairs.

crops to 4. For VLM2VEC using LLaVA-1.6 and Qwen2-VL as the backbone, we resize the input images to a uniform resolution, employing two setups: a high-resolution configuration of 1344×1344 and a low-resolution configuration of 336×336. For the 20 training datasets, we randomly select up to 100K data points. When using GradCache, we set a sub-batch size accordingly, with the total batch size accumulated to 1,024. All experiments were run on 8 H100 GPUs. We report Precision@1 for all models in Table 2. It measures the ratio of positive candidates being ranked in the top place for all queries.

4.1 BASELINES

Four groups of baselines are reported in this study.

CLIP-family: We utilize vision/language encoders such as CLIP (Radford et al., 2021), OpenCLIP (Cherti et al., 2023), SigLIP (Zhai et al., 2023), and BLIP2 (Li et al., 2023a) as our baseline. Due to the length limitations of the text encoder, some queries or target text in certain tasks may be truncated. We apply score-level fusion by combining multimodal features using element-wise addition with equal weights ($w_1 = w_2 = 1$). We do not use instructions, as they could potentially degrade performance. For more details, please refer to Section 4.3.4.

UniIR: UniIR (Wei et al., 2023) is a unified, instruction-guided multimodal retriever designed to handle eight different retrieval tasks across multiple modalities. The model builds on CLIP and BLIP, employing shallow fusion techniques such as score-level and feature-level fusion to integrate modalities. In this study, we use the CLIP_SF and BLIP_FF variations as baselines.

MagicLens: MagicLens (Zhang et al., 2024) is a self-supervised image retrieval model capable of handling open-ended instructions. It utilizes a dual-encoder architecture with shared parameters, initializing the vision and language encoders with either CoCa or CLIP. The model uses a multi-head attention pooler to unify multimodal inputs into a single embedding. For this study, we report results using the CLIP-Large backbone.

E5-V: E5-V (Jiang et al., 2024) is a contemporary model that also leverages vision-language models for multimodal embedding tasks. It proposes a single-modality training approach, where the model is trained exclusively on text pairs. In contrast, our model is trained on multimodal pairs, which include various combinations of image and text modalities on both the query and target sides.

For all our baselines, we first use their original versions. Additionally, we have fine-tuned both CLIP and OpenCLIP on MMEB training datasets. We adopt the same experimental configurations as VLM2VEC to ensure a fair comparison. For the remaining baseline models, UniIR and MagicLens also utilize a shallow fusion approach based on CLIP models, with their primary contribution being the datasets they were trained on. E5-V proposes training exclusively on text pairs, making it unsuitable for fine-tuning on our datasets. Therefore, we have not included the fine-tuned versions of these three models in this comparison.

Table 2: Results on the MMEB benchmark. The scores are averaged per meta-task. For detailed scores per dataset, see Table 6. We include baselines with and without fine-tuning on MMEB training datasets and our models with three different backbones using a batch size of 1,024.

Model	**Per Meta-Task Score**				**Average Score**		
	Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# of datasets →	10	10	12	4	20	16	36
Baseline Models (No Fine-tuning on MMEB Training)							
CLIP (Radford et al., 2021)	42.8	9.1	53.0	51.8	37.1	38.7	37.8
BLIP2 (Li et al., 2023a)	27.0	4.2	33.9	47.0	25.3	25.1	25.2
SigLIP (Zhai et al., 2023)	40.3	8.4	31.6	59.5	32.3	38.0	34.8
OpenCLIP (Cherti et al., 2023)	47.8	10.9	52.3	53.3	39.3	40.2	39.7
UniIR (BLIP_FF) (Wei et al., 2023)	42.1	15.0	60.1	62.2	44.7	40.4	42.8
UniIR (CLIP_SF) (Wei et al., 2023)	<u>44.3</u>	<u>16.2</u>	<u>61.8</u>	<u>65.3</u>	<u>47.1</u>	<u>41.7</u>	<u>44.7</u>
E5-V (Jiang et al., 2024)	21.8	4.9	11.5	19.0	14.9	11.5	13.3
Magiclens (Zhang et al., 2024)	38.8	8.3	35.4	26.0	31.0	23.7	27.8
Baseline Models (Fine-tuning on MMEB Training)							
CLIP	55.2	19.7	53.2	62.2	47.6	42.8	45.4
OpenCLIP	<u>56.0</u>	<u>21.9</u>	<u>55.4</u>	<u>64.1</u>	<u>50.5</u>	<u>43.1</u>	<u>47.2</u>
Ours (VLM2VEC)							
Phi-3.5-V, Full-model fine-tuned (#crop=4)	52.8	50.3	57.8	72.3	62.8	47.4	55.9
Phi-3.5-V, LoRA (#crop=4)	54.8	54.9	62.3	79.5	66.5	52.0	60.1
LLaVA-1.6, LoRA (low-res)	54.7	50.3	56.2	64.0	61.0	47.5	55.0
LLaVA-1.6, LoRA (high-res)	61.2	49.9	67.4	**86.1**	67.5	57.1	62.9
Qwen2-VL, LoRA (high-res)	**62.6**	**57.8**	**69.9**	81.7	**72.2**	**57.8**	**65.8**
Δ - Best baseline (No Fine-tuning)	+17.9	+41.6	+8.1	+25.1	+25.1	+16.1	+21.1
Δ - Best baseline (Fine-tuning)	+6.6	+35.9	+14.5	+22.0	+21.7	+14.7	+18.6

4.2 Main Result

From Table 2, the best variant of VLM2VEC leverages Qwen2-VL, is trained with LoRA, and processes input images at a relatively high resolution of 1344 × 1344. It achieves an average precision@1 of 65.8% across all 36 datasets from MMEB. Additionally, it maintains an average precision@1 of 57.8% on 16 out-of-distribution tasks in zero-shot evaluation, suggesting strong generalization ability. These results indicate that when trained on datasets spanning diverse task categories, domains, and modality combinations, VLM2VEC effectively follows instructions to align visual and textual representations while generalizing well to unseen tasks.

Furthermore, VLM2VEC using Phi-3.5-V and LLaVA-1.6 also outperforms baseline models. Notably, LLaVA-1.6 has a transparent pre-training data recipe with minimal overlap with MMEB's OOD datasets, reinforcing that the strong zero-shot performance of VLM2VEC is not due to prior exposure of its backbone to the OOD datasets. When using the same backbone, the full fine-tuning variant achieves slightly lower scores than the LoRA version. For a detailed discussion comparing full fine-tuning and LoRA, please refer to Section 4.3.1.

Compared to other baseline models, with or without fine-tuning on MMEB training data, our model demonstrates consistent improvements. Compared to the best baseline model without fine-tuning, our model achieves a 21.1 point improvement (from 44.7 to 65.8) across all 36 MMEB datasets and a 16.1-point increase (from 41.7 to 57.8) on 16 out-of-distribution datasets for zero-shot evaluation. Compared to the best baseline model with fine-tuning, our model achieves a 18.6 point improvement (from 47.2 to 65.8) across all 36 MMEB datasets and a 14.7-point increase (from 43.1 to 57.8) on 16 out-of-distribution datasets for zero-shot evaluation. Additionally, unlike the baseline models, which fail to demonstrate reasonable performance across all different task categories, VLM2VEC achieves strong performance across all four meta-task categories. This highlights its capability to handle a wide range of multimodal embedding tasks effectively.

4.3 Result Analysis

To train an effective and generalizable multimodal embedding, various factors need to be considered, ranging from the data to the training setup. In this section, we present detailed ablation studies on these factors. We will discuss two training setups: Full Fine-Tuning vs. LoRA, along with Training parameters, and two topics related to data: Meta-task generalization and Impact of instructions.

4.3.1 FULL FINE-TUNING VS. LORA

When fine-tuning the VLMs, a key decision is whether to conduct full fine-tuning, which updates all parameters in the model, or to use a parameter-efficient method such as LoRA. We compare the performance of fully fine-tuned VLM2VEC with its LoRA variants at different ranks. The training and data setups are kept consistent across all models. We observe that LoRA achieves better performance when the rank is appropriately configured.

Table 3: We compare the performance of fully fine-tuned VLM2VEC with its LoRA variants at different ranks. LoRA can achieve better performance when the rank is appropriately configured. All the models utilize Phi-3.5-V as their backbone.

Model	**Meta-Task Average Score**				**Average Score**		
	Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# of datasets →	10	10	12	4	20	16	36
Full Fine-Tuning (bs=256)	50.4	46.4	52.6	68.6	57.9	44.7	52.0
LoRA r = 4 (bs=256)	52.7	**53.6**	**60.1**	**80.2**	**64.9**	**50.4**	**58.4**
LoRA r = 8 (bs=256)	**52.9**	52.5	60.3	80.0	64.2	50.8	58.2
LoRA r = 16 (bs=256)	51.1	40.5	52.0	72.5	54.9	45.8	50.8
LoRA r = 32 (bs=256)	50.6	47.8	53.9	72.5	58.9	46.5	53.4

4.3.2 TRAINING PARAMETERS

During our experiments, we identified three key parameters that significantly impact the performance of VLM2VEC: training batch size, the number of sub-image crops, and the number of training steps. In Figure 4, we observe that the final performance gradually improves as we increase the batch size, training step size, and number of sub-image crops. We particularly want to highlight the impact of batch size. Due to the lack of hard negatives, using a large batch size with plenty of random negatives, supported by the GradCache technique, plays a crucial role in enhancing the performance of VLM2VEC, as discussed in Section 3.2.

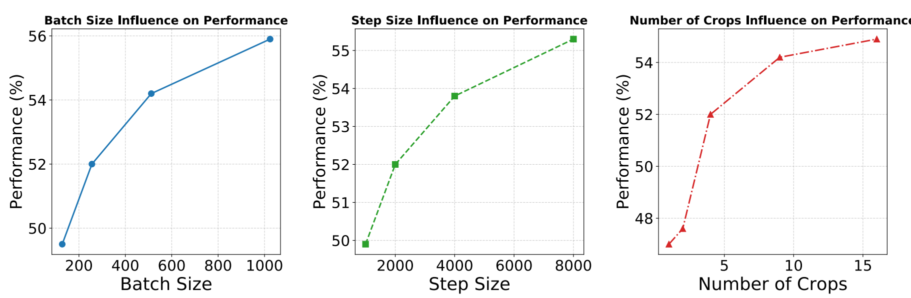

Figure 4: The figures demonstrate the influence of the training setup on VLM2VEC's final performance. Here, we examine the effects of training batch size, the number of sub-image crops, and the number of training steps. All the models utilize Phi-3.5-V as their backbone.

4.3.3 META-TASK GENERALIZATION

We have demonstrated that VLM2VEC has the potential to transfer to out-of-distribution datasets after being trained on a diverse range of in-distribution datasets, with the instruction-following settings. An interesting question arises as to whether focusing on a specific meta-task can enhance the model's overall generalizability. We have trained three models, each focused solely on one meta-task (classification, visual question answering, and retrieval). Visual grounding was not included due to the limited number of training datasets. We then evaluated the models' transferability to other meta-tasks. We refer to these three models as VLM2VEC $_{\text{RET}}$, trained on 8 retrieval tasks, VLM2VEC $_{\text{VQA}}$, trained on 6 visual question answering tasks, and VLM2VEC $_{\text{CLS}}$, trained on 5 classification tasks.

Figure 5 illustrates the generalizability of these three models on unseen meta-tasks. We could observe that VLM2VEC $_{\text{RET}}$ has better generalizablilty on other meta-task, compared with other two models, especially on visual grounding categories. The reason is that retrieval tasks involve a more diverse combination of text and visual modalities from both the query and target sides, which helps the model generalize better to unseen meta-tasks. This observation highlights the benefits of using more diverse tasks in the VLM2VEC training process.

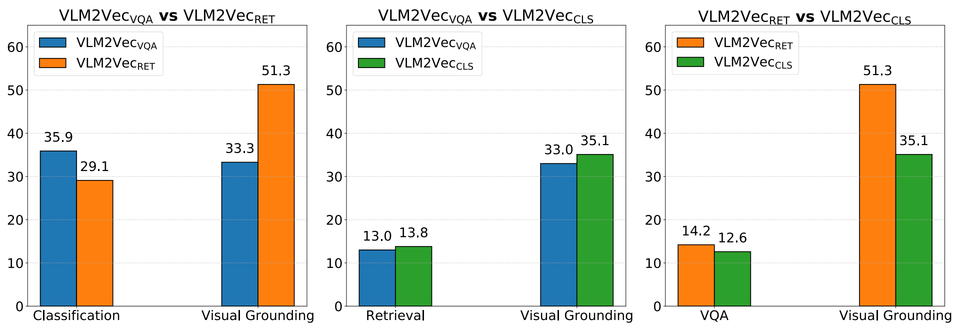

Figure 5: The figures show the generalization ability of models trained on one meta-task to other unseen meta-tasks. For example, the first subplot compares the performance of VLM2VEC trained exclusively on VQA datasets with VLM2VEC trained exclusively on retrieval datasets across the other two meta-task categories: classification and visual grounding. Overall, VLM2VEC trained on retrieval tasks demonstrate better generalization ability because retrieval tasks involve a more diverse combination of text and visual modalities from both the query and target sides. VLM2VEC utilizes Phi-3.5-V as its backbone.

4.3.4 IMPACT OF INSTRUCTIONS

Previous studies have shown the influence of instructions on addressing various tasks. VLM2VEC, which leverages a VLM as its backbone and is trained on large-scale datasets with instructions, is expected to better generalize across tasks and improve performance in multimodal embedding tasks. In this section, we evaluate the performance of both CLIP and VLM2VEC with and without task-specific instructions to quantify the impact of incorporating instructions into the embedding process. As shown in Table 4, incorporating instructions reduces the CLIP model's performance by 29.4%, while our VLM2VEC achieves a 49.4% improvement. This highlights how a VLM backbone enhances the embedding model's instruction-following capability and emphasizes the advantages of instruction-guided embeddings.

Table 4: Comparison of CLIP and our VLM2VEC with and without task-specific instructions. Incorporating instructions could decrease CLIP's performance by 29.4%, whereas our VLM2VEC achieves a 49.4% improvement. VLM2VEC utilizes Phi-3.5-V as its backbone.

Model	**Meta-Task Average Score**				**Average Score**		
	Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# of datasets →	10	10	12	4	20	16	36
CILP							
w/o instruction	42.8	9.1	53.0	51.8	37.1	38.7	37.8
w/ instruction	17.4	8.0	41.3	52.9	23.8	30.3	26.7
Δ	-59.3%	-12.1%	-22.1%	2.1%	-35.8%	-21.7%	-29.4%
Ours (VLM2VEC)							
w/o instruction	36.7	33.5	31.1	44.3	37.3	31.6	34.8
w/ instruction	50.4	46.4	52.6	68.6	57.9	44.7	52.0
Δ	37.3%	38.5%	69.1%	54.9%	55.2%	41.5%	49.4%

5 RELATED WORK

5.1 TEXT EMBEDDING

Text embeddings have demonstrated significant potential in powering downstream applications such as information retrieval (Karpukhin et al., 2020; Xiong et al., 2020), text similarity (Gao et al., 2021b), prompt retrieval for in-context learning (Hongjin et al., 2022), and classification (Logeswaran & Lee, 2018; Reimers & Gurevych, 2019). Early work focused on creating effective embeddings for specific tasks. With the rise of pretrained language models, efforts have shifted toward developing universal embedding models capable of handling a wide range of embedding tasks. Studies such as GTR (Ni et al., 2022) and E5 (Wang et al., 2022a) leveraged large amounts of noisy paired data to pretrain and fine-tune dense retrievers. More recent works like TART (Asai et al., 2022) and InstructOR (Su et al., 2023) introduced natural language prompts to guide embedding models in producing task-relevant embeddings. Building on this, models like E5Mistral(Wang et al., 2024a), SFR-Embedding(Meng et al., 2024), RepLLaMA(Ma et al., 2024b), GTE-Qwen2(Li et al., 2023b), and NV-Embed (Lee et al., 2024) have utilized pretrained large language models (LLMs) as their backbone, fine-tuning them with multi-task data and instructions. These models have delivered significant improvements over earlier approaches that did not use LLMs for initialization or instruction tuning.

5.2 MULTIMODAL EMBEDDINGS

Multimodal embeddings have long been a significant research challenge. Early works like CLIP (Radford et al., 2021), BLIP (Li et al., 2022; 2023a), Align (Jia et al., 2021), SigLIP (Zhai et al., 2023), SimVLM Wang et al. (2022b) and CoCa (Yu et al., 2022) primarily focused on learning universal representations from large-scale, weakly supervised image-text pairs. These models generally encode images and text separately, projecting them into a shared space. This approach has laid the groundwork for more recent multimodal models like LLaVA (Liu et al., 2024).

Most research on universal multimodal embeddings involves fine-tuning models like CLIP or BLIP, typically using simple fusion mechanisms to combine visual and language information. For instance, UniIR (Wei et al., 2023) creates multimodal embeddings by simply adding text and visual features, while MagicLens (Zhang et al., 2024) employs shallow self-attention layers to integrate these features more effectively. The study most similar to ours is E5-V (Jiang et al., 2024), a contemporary work that fine-tunes a vision-language model using only text training data.

5.3 EMBEDDING BENCHMARKS

Significant efforts have been made to develop benchmarks for evaluating retrieval systems. For text retrieval models, MS MARCO (Nguyen et al., 2016) and Natural Questions (Kwiatkowski et al., 2019b) are two of the most widely used benchmarks in general domains. To broaden the evaluation across more diverse domains, BEIR (Thakur et al.) was introduced, incorporating 18 datasets from various fields. Building on this, MTEB (Muennighoff et al., 2023) further expands BEIR's scope by adding more tasks, such as classification, clustering, and semantic textual similarity (STS).

For multimodal retrieval, several benchmarks have been introduced to evaluate model performance across different modalities. MBEIR (Wei et al., 2023) includes 8 tasks and 16 datasets, designed to test models' ability to retrieve information based on various forms of queries and instructions.

6 CONCLUSION

In this paper, we aim to build the first large-scale multimodal embedding framework, comprising two main components: MMEB and VLM2VEC. MMEB includes 36 datasets across four meta-task categories, providing a comprehensive and diverse framework for training and evaluating embedding models. VLM2VEC leverages VLMs as a backbone to deeply fuse visual and textual spaces, enhancing generalization to unseen tasks through instruction following.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. SemEval-2012 task 6: A pilot on semantic textual similarity. In Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret (eds.), **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 385–393, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics. URL `https://aclanthology.org/S12-1051`.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. Task-aware retrieval with instructions. *arXiv preprint arXiv:2211.09260*, 2022.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens (eds.), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL `https://aclanthology.org/S17-2001`.

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*, pp. 16495–16504, 2022.

Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3), 2010.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 670–680, 2017.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 326–335, 2017.

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*, pp. 248–255. Ieee, 2009.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111:98 – 136, 2014. URL `https://api.semanticscholar.org/CorpusID:207252270`.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. Scaling deep contrastive learning batch size under memory limited setup. *arXiv preprint arXiv:2101.06983*, 2021a.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, 2021b.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

SU Hongjin, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*, 2022.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12065–12075, 2023.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering, 2020. URL https://arxiv.org/abs/2007.0128.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. *Advances in neural information processing systems*, 28, 2015.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019a.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019b.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*, 2020.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *DeeLIO 2022*, pp. 100, 2022.

Siqi Liu, Weixi Feng, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. Edis: Entity-driven image search over multimodal web content. *arXiv preprint arXiv:2305.13631*, 2023.

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*, pp. 2125–2134, 2021.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJvJXZb0W.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. *arXiv preprint arXiv:2406.11251*, 2024a.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2421–2425, 2024b.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 216–223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/363_Paper.pdf.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

Rui Meng, Ye Liu, Shafiq Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. Sfr-embedding-2: Advanced text embedding with multi-stage training, 2024. URL https://huggingface.co/Salesforce/SFR-Embedding-2_R.

Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th international conference on world wide web*, pp. 1291–1299, 2017.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, 2023.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. November 2016. URL https://www.microsoft.com/en-us/research/publication/ms-marco-human-generated-machine-reading-comprehension-dataset/.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, et al. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9844–9855, 2022.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

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*, pp. 2641–2649, 2015.

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

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2655–2671, 2022.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, 2020.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. Repetition improves language model embeddings. *arXiv preprint arXiv:2402.15449*, 2024.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1102–1121, 2023.

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.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022a.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. N24news: A new dataset for multimodal news classification. *arXiv preprint arXiv:2108.13327*, 2021.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2022b.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*, 2023.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11307–11317, 2021.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

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*, pp. 11975–11986, 2023.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4995–5004, 2016.