MoCa: Modality-aware Continual Pre-training Makes Better Bidirectional Multimodal Embeddings

Haonan Chen[1*], **Hong Liu**[2], **Yuping Luo, Liang Wang**[3],
Nan Yang[3], **Furu Wei**[3], **Zhicheng Dou**[1†]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Stanford University [3]Microsoft Corporation
{hnchen,dou}@ruc.edu.cn
hliu99@cs.stanford.edu, yupingl@cs.princeton.edu
{wangliang,nanya,fuwei}@microsoft.com

Abstract

Multimodal embedding models, built upon causal Vision Language Models (VLMs), have shown promise in various tasks. However, current approaches face three limitations: causal attention in VLM backbones is suboptimal for embedding tasks; scalability issues due to reliance on high-quality labeled paired data for contrastive learning; and limited diversity in training objectives and data. To address these issues, we propose MoCa, a two-stage framework for transforming pre-trained VLMs into bidirectional multimodal embedding models. The first stage, Modality-aware Continual Pre-training, introduces a joint reconstruction objective that simultaneously denoises interleaved texts and images, enhancing bidirectional context-aware reasoning. The second stage, Heterogeneous Contrastive Fine-tuning, leverages diverse, semantically rich multimodal data beyond simple image-caption pairs to enhance generalization and alignment. Our method addresses the stated limitations by introducing bidirectional attention through continual pre-training, scaling effectively with massive unlabeled datasets via joint reconstruction objectives, and utilizing diverse multimodal data for enhanced representation robustness. Experiments demonstrate that MoCa consistently improves performance across MMEB and ViDoRe-v2 benchmarks, achieving new state-of-the-arts, and exhibits strong scalability with both model size and training data on MMEB. We have released the model weights and data on our project page https://haon-chen.github.io/MoCa/.

1 Introduction

Multimodal embedding models have achieved significant improvements in various tasks including multimodal classification, visual question answering, and document retrieval [Jiang et al., 2024, Zhang et al., 2024, Chen et al., 2025a, Ma et al., 2024a]. These models are built on Vision Language Models (VLMs), such as Phi-V [Abdin et al., 2024], LLaVA [Liu et al., 2023], and Qwen-VL [Bai et al., 2025], which demonstrate strong generation and cross-modal comprehension capabilities. Recent methods apply contrastive learning on image-text pairs to off-the-shelf VLMs to align modalities and improve cross-modal representation quality [Jiang et al., 2024, Chen et al., 2025a].

Despite the success of current multimodal embedding models, three main limitations remain: (1) **Causal attention of pre-trained VLMs might be suboptimal for embedding models.** Mainstream multimodal embedding models [Jiang et al., 2024, Chen et al., 2025a, Zhang et al., 2024] inherit causal attention from their VLM backbones. However, studies on text embedding models [BehnamGhader et al., 2024, Li et al., 2023b, Lee et al., 2025] have shown that bidirectional attention typically produces superior embeddings compared to causal attention. Furthermore, bidirectional embedding models with mean pooling offer practical benefits, such as late chunking [Günther et al., 2024]. (2) **Contrastive learning is hard to scale without labeled pair data.** Contrastive learning fundamentally depends on diverse high-quality image-text pairs, which limits its scalability. Although large datasets of image-caption pairs exist [Schuhmann et al., 2021, Tschannen et al., 2025], curating diverse and high-quality multimodal pairs remains resource-intensive. Moreover, contrastive learning cannot leverage the vast amount of unpaired multimodal data available on the internet. (3) **Lack of diversity in training objectives and data distribution leads to suboptimal cross-modal alignment.** Prior works such as Jiang et al. [2024], Chen et al. [2025a] typically fine-tune Vision-Language Models (VLMs) with a single contrastive objective applied to a narrow range of data, *i.e.*, mostly short image-caption pairs. This setup fails to fully exploit the rich cross-modal

[*] Work done during Haonan's internship at Microsoft Research Asia, [†] Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 810–823

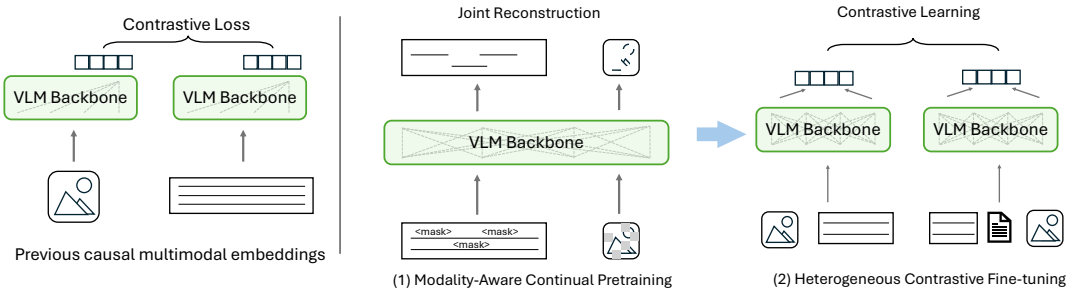

Figure 1: **Comparison of VLM-based multimodal embedding models.** Left: Previous single-stage contrastive learning with mainly image-caption pairs and causal attention. Right: MoCa. In *modality-aware continual pre-training*, we optimize a joint bidirectional reconstruction objective to denoise interleaved texts and images. In *heterogeneous contrastive fine-tuning*, we improve cross-modal fusion with diverse image and text contexts.

reasoning capabilities that pre-trained VLMs offer. As a result, the learned embedding models often overfit to the training distribution and struggle to generalize to more complex or diverse scenarios.

Recent advances in text embedding models, such as LLM2Vec [BehnamGhader et al., 2024], adapt pre-trained language models using bidirectional Masked Language Modeling (MLM) [Devlin et al., 2019]. Despite the success of Continual Pre-Training (CPT) in text-only domains, its potential remains underexplored for multimodal embeddings. Moreover, as shown in Section 3.3, MLM alone is insufficient for mixed-modality inputs, motivating modality-aware, bidirectional objectives to jointly process interleaved image and text signals.

In this work, we introduce a two-stage framework, MoCa, to transform pre-trained VLMs into effective bidirectional multimodal embedding models. As illustrated in Figure 1, our approach consists of two stages: (1) **Modality-aware Continual Pre-training** and (2) **Heterogeneous Contrastive Fine-tuning**. In the first stage, we introduce a joint reconstruction objective that requires the model to simultaneously denoise interleaved text and image inputs, which encourages the model to jointly reason across modalities. For text, we apply masked language modeling (MLM), where masked tokens are predicted using the full multimodal context. For images, we adopt masked autoencoding (MAE): a subset of image patches are randomly masked and reconstructed by a lightweight decoder conditioned on image and text contexts. In the second stage, as opposed to previous works which used mainly image-caption pairs, we add diverse heterogeneous data including (i) long-form query-document pairs, supporting document-level understanding and complex reasoning over extended context, (ii) curated multimodal pairs, offering various visual and textual contexts beyond image captions of specific distributions, and (iii) real-world text pairs, enhancing linguistic representations across diverse domains.

Together, the two stages directly address the limitations outlined above. Stage one tackles Limitations (1) and (2) by using massive unlabeled interleaved data and enhancing bidirectional, context-aware reasoning across modalities. It also partially mitigates Limitation (3) by applying joint reconstruction on diverse multimodal inputs. Stage two directly addresses Limitation (3) by introducing heterogeneous and semantically rich multimodal pairs to enhance generalization and alignment across various domains.

We conduct experiments with MoCa and verify that our trained model consistently improves performance across MMEB [Jiang et al., 2024] and ViDoRe-v2 [Faysse et al., 2025] benchmarks. Besides, it demonstrates strong scalability with respect to both model size and training data on MMEB. Specifically, after continual pre-training on only 30B tokens, our 3B model matches or surpasses the performance of competitive 7B baselines. When scaled to 7B parameters, our model sets new state-of-the-art results on MMEB.

In summary, our contributions are as follows.

- We are the first to propose a continual pre-training (CPT) approach with unlabeled data to adapt VLMs to bidirectional embedding models and demonstrate its strong scalability with respect to model and corpus sizes.
- We also show that contrastive fine-tuning with heterogeneous data and cross-modal interactions enhances model generalization.
- With the two techniques combined, our frame-

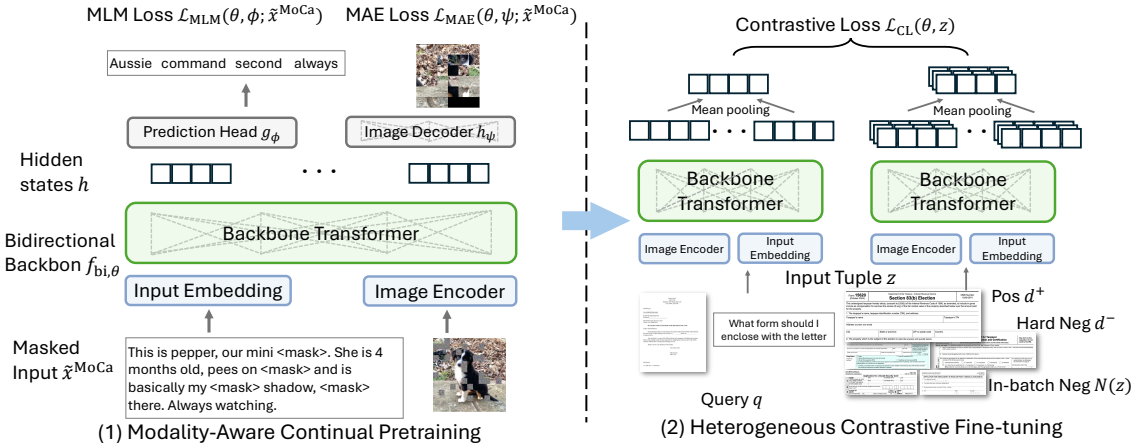

Figure 2: **MoCa.** (1) In *modality-aware continual pre-training*, the VLM backbone is trained to jointly reconstruct masked texts and images based on interleaved multimodal context with masked language modeling and masked autoencoding. (2) In *heterogeneous contrastive fine-tuning*, the VLM backbone from the previous stage is further fine-tuned with contrastive loss on a broad range of heterogeneous data.

work, MoCa, consistently improves performance on various benchmarks and achieves state-of-the-art performance on MMEB.

2 Method: MoCa

In this section, we present MoCa, which transforms pre-trained VLMs into powerful bidirectional multimodal embedding models. As illustrated in Figure 2, our method comprises two stages: (1) **Modality-aware Continual Pre-training**, where the model learns to reconstruct masked texts and images with a joint denoising objective. The objective leverages MLM and MAE, which enables the model to shift from causal to bidirectional attention with better representation quality. (2) **Heterogeneous Contrastive Fine-tuning**, where we perform contrastive fine-tuning with a diverse set of multimodal pairs spanning long-form query-document pairs, curated multimodal pairs and real-world text pairs. This stage further aligns vision and language embeddings while improving generalization across varied real-world tasks.

2.1 Preliminaries

Vision Language Model Backbones. VLMs are widely adopted as the backbones of multimodal embedding models [Jiang et al., 2024, Zhang et al., 2024, Chen et al., 2025a]. Consider a multimodal input of length T, denoted by $x = [x_1, \ldots, x_T]$, where each $x_i, i \in [T]$, can be either a discrete text token or a continuous image patch. A VLM backbone first maps the input to the same space with input embedding layers for text tokens and visual encoders for image patches. The input embeddings are then passed through the backbone transformer to obtain hidden states $f_\theta^{\text{causal}}(x) \in \mathbb{R}^{T \times d}$ where θ denotes the model parameters, and $f_\theta^{\text{causal}}(\cdot)$ refers to the VLM backbone with causal attention.

Multimodal embedding models. Most existing models inherit the causal attention from VLM backbones. Therefore, $(f_\theta^{\text{causal}}(x))_j$ only depends on $[x_1, \ldots, x_j]$. To extract embeddings from the backbone, a natural choice would be the hidden states corresponding to the last (EOS) token, i.e. $\text{Emb}_\theta^{\text{causal}}(x) := (f_\theta^{\text{causal}}(x))_T$.

For better representation quality, we introduce bidirectional VLM backbone $f_\theta^{\text{bi}}(\cdot)$, which essentially removes the attention causal masks from $f_\theta^{\text{causal}}(\cdot)$. To extract embeddings from $f_\theta^{\text{bi}}(\cdot)$, we adopt mean pooling. Therefore we have $\text{Emb}_\theta^{\text{bi}}(x) := \frac{1}{T}\sum_{i=1}^{T}(f_\theta^{\text{bi}}(x))_i$.

Contrastive learning. Based on the causal multimodal embedding models initialized with pre-trained VLMs above, previous works follow the recipe of text embedding models [Wang et al., 2022, Li et al., 2023b, Lee et al., 2025, Xiao et al., 2024] to align image and text embeddings. Each training example is a tuple $(q, d^+, \{d_1^-, \ldots, d_K^-\})$, where q is the *query*, d^+ is a *positive document*, and $\{d_1^-, \ldots, d_K^-\}$ is a set of *hard negative documents*.

2.2 Modality-aware Continual Pre-training

This stage enhances the bidirectional representation capability of a pre-trained VLM with joint denoising objectives over interleaved texts and images. As shown in Figure 2 (Left), we adopt two complementary objectives: masked language modeling

(MLM) [Devlin et al., 2019] for texts and masked autoencoding (MAE) [He et al., 2022] for images.

Masked Language Modeling. We apply MLM to the text tokens in the input x. Specifically, we randomly sample a subset of text tokens $\tilde{T}^{\text{MLM}} \subset \{1, \ldots, T\}$ and replace each $x_i,\ i \in \tilde{T}^{\text{MLM}}$, with a special mask token <|mask|>. The resulting input with masked text tokens is denoted as $\tilde{x}^{\text{MLM}}$.

The VLM encoder then processes the sequence with masked text tokens with bidirectional attention, enabling each masked position to attend to all visible text tokens and image patches. This facilitates contextual learning that captures both intra- and cross-modal dependencies. The model is trained to accurately predict each masked token $x_i,\ i \in \tilde{T}^{\text{MLM}}$, based on the surrounding context. Following BehnamGhader et al. [2024], we predict the masked token at position i using the previous token $i-1$ (*i.e.*, shift the labels) to align with the training recipe of most auto-regressive VLMs. Suppose the MLM prediction head is $g_\phi(\cdot)$ parameterized by ϕ. The MLM loss is then computed with cross-entropy over the masked positions:

$$\mathcal{L}_{\text{MLM}}(\theta, \phi) = \sum_{i \in \tilde{T}^{\text{MLM}}} \ell_{\text{CE}}((p_{\theta,\phi})_{i-1}, x_i), \quad (1)$$

where $(p_{\theta,\phi})_{i-1} := g_\phi((f_{\text{bi},\theta}(\tilde{x}^{\text{MLM}}))_{i-1})$ is the model's predicted distribution over the vocabulary at position i. This objective encourages the model to leverage both local and global contexts to recover masked information, enhancing its ability as a bidirectional encoder for embedding tasks.

Masked Autoencoding. Similar to MLM on text tokens, we want a denoising objective to reconstruct image patches based on text and image context. Inspired by He et al. [2022], we mask image patches and feed them to the bidirectional VLM backbone. On top of the VLM hidden states, we use a light-weight decoder to predict the original patches. Concretely, given input with masked text tokens $\tilde{x}^{\text{MLM}}$, we further randomly sample a subset of image patches $\tilde{T}^{\text{MAE}} \subset \{1, \ldots, T\}$ and replace each $x_i,\ i \in \tilde{T}^{\text{MAE}}$ with vectors sampled from a unit Gaussian distribution. The resulting input with masked text and images is denoted as $\tilde{x}^{\texttt{MoCa}}$.

The model is trained to predict each masked patch $x_i,\ i \in \tilde{T}^{\text{MAE}}$ based on the surrounding multimodal context. Following He et al. [2022], we add a shallow transformer h_ψ as the image patch decoder on top of the VLM encoder. The MAE loss is computed with MSE over the masked patches:

$$\mathcal{L}_{\text{MAE}}(\theta, \psi) = \sum_{i \in \tilde{T}^{\text{MAE}}} \ell_{\text{MSE}}(\hat{x}_i, x_i). \quad (2)$$

$\hat{x}_i = h_\psi(f_{\text{bi},\theta}(\tilde{x}^{\texttt{MoCa}}))_i$ is the reconstructed patch.

The final modality-aware continual pre-training objective is a weighted sum of MLM and MAE on the sequence with masked texts and images,

$$\mathcal{L}_{\texttt{MoCa}}(\theta, \phi, \psi) = \mathcal{L}_{\text{MLM}}(\theta, \phi) + w\mathcal{L}_{\text{MAE}}(\theta, \psi),$$

where w is the weight to balance these two losses[1].

Efficient implementation. In practice, we use data parallel to distribute workloads across GPUs. To achieve load balancing, we calculate the compute cost of each sequence based on sequence length and image sizes and implement a sequence packing algorithm to make sure all GPUs process a batch with almost the same compute cost.

2.3 Heterogeneous Contrastive Fine-tuning

Following the CPT stage, we fine-tune the model with a contrastive objective over a broad range of heterogeneous data. Unlike prior methods [Jiang et al., 2024, Chen et al., 2025a] that primarily rely on image-caption pairs, we leverage more diverse sources to improve robustness. As shown in Figure 2, this includes: **(1) Long-form multimodal pairs**, which consist of document-level inputs containing both images and extended text. These samples support complex cross-modal reasoning and coherence over long contexts. **(2) Curated multimodal pairs** that offer varied and high-quality alignments beyond typical captioning. **(3) Text-only pairs**, sampled from large-scale retrieval datasets, which enhance the model's ability to encode fine-grained semantic differences in language.

Each training instance z is structured as a tuple $(q, d^+, \{d_1^-, \ldots, d_K^-\})$, and both query and documents may be either unimodal or multimodal. For each sample z, let $N(z) = \{d_1^-, \ldots, d_K^-\} \cup N_{\text{in}}(z)$ be the set of negative documents to contrast, where $N_{\text{in}}(z)$ contains other documents in the same batch. Denote by $\mathbf{q}$, $\mathbf{d}$ the embeddings of q, d, respectively, *i.e.*, $\mathbf{q} = \text{Emb}_\theta^{\text{bi}}(q)$ and $\mathbf{d} = \text{Emb}_\theta^{\text{bi}}(d)$. The contrastive loss is then defined as:

$$\mathcal{L}_{\text{CL}}(\theta) = -\log \frac{\Phi(\mathbf{q}, \mathbf{d}^+)}{\Phi(\mathbf{q}, \mathbf{d}^+) + \sum_{\mathbf{d} \in N(z)} \Phi(\mathbf{q}, \mathbf{d})}, \quad (3)$$

[1]MLM and MAE losses are calculated on the same input $\tilde{x}^{\texttt{MoCa}}$ containing masked images and masked texts.

where $\Phi(\cdot,\cdot)$ is a similarity function defined as: $\Phi(\mathbf{a},\mathbf{b}) = \exp(\cos(\mathbf{a},\mathbf{b})/\tau)$, with $\cos(\cdot,\cdot)$ denoting cosine similarity and τ is the temperature.

Task-aware batching. In Heterogeneous Contrastive Fine-tuning, documents from different tasks can vary significantly. For example, a puppy is completely different from a screenshot of an arXiv paper. This makes distinguishing them trivial, which diminishes the benefit of in-batch negatives. To address this, we adopt task-aware batching [Li et al., 2023b]. By ensuring that all instances within the same batch originate from the same task, we enable harder in-batch negative samples, leading to improved representation quality.

3 Experiments

3.1 Experimental Setup

3.1.1 Modality-aware Continual Pre-training

Training Data. We incorporate three categories of training corpora as the CPT dataset: (1) **Text-only data** from DCLM [Li et al., 2024], (2) **Common image-text pairs** from PixelProse [Singla et al., 2024] (CommonPool, CC12M, and RedCaps), MAmmoTH-VL-Instruct [Guo et al., 2024], and MMEB training set [Jiang et al., 2024], and (3) **Document-level multimodal data** from DocMatix [Laurençon et al., 2024], VisRAG [Yu et al., 2025], and ColPali training set [Faysse et al., 2025]. For each dataset, we randomly sample 500K instances, resulting in a total corpus of ~30B tokens.
Implementation Details. We adopt Qwen-2.5-VL [Bai et al., 2025] as the VLM backbone. We use a maximum input sequence length of 2048 tokens and a micro-batch size of 12,800 across 32 H100 GPUs. The learning rate is set to 2×10^{-6}. For MLM, we apply a mask ratio of 0.4 for `MoCa-3B` and 0.6 for `MoCa-7B`. For MAE, we use a masking ratio of 0.5 for `MoCa-3B` and 0.6 for `MoCa-7B`. The lightweight image decoder for MAE (h_ψ) is initialized from the middle layer of the Qwen-2.5-VL backbone to improve loss stability. The MAE loss $\mathcal{L}_{\text{MAE}}$ is weighted by $w = 0.5$ to balance with the MLM loss $\mathcal{L}_{\text{MLM}}$.

3.1.2 Heterogeneous Contrastive Learning

Training Data. We use three categories of datasets for contrastive learning: (1) **Long-form multimodal pairs** from VisRAG [Yu et al., 2025] and the ViDoRe training set [Faysse et al., 2025], (2) **Common multimodal pairs** from the training sets of MMEB [Jiang et al., 2024] and mmE5 [Chen et al., 2025a], and (3) **Text-only pairs** from the large-scale dense retrieval dataset E5 [Wang et al., 2022]. From each dataset, we randomly sample 50K instances, resulting in a total of approximately 2M contrastive training pairs.
Implementation Details. We use a batch size of 2048. The learning rate is set to 1×10^{-5}, and the temperature τ in the contrastive loss is set to 0.03. Each query is accompanied by two hard negative document pairs, which is curated by mmE5 [Chen et al., 2025a].

The CPT and CL stages utilize the training data from benchmarks. This practice is standard in multimodal embedding works, such as mmE5 [Chen et al., 2025a] and ColPali [Faysse et al., 2025].

3.1.3 Evaluation

Massive Multimodal Embedding Benchmark (MMEB). We assess the general embedding quality with MMEB [Jiang et al., 2024], reporting results with Precision@1. MMEB includes 36 multimodal tasks across four types: 10 classification tasks, 10 visual question answering (VQA) tasks, 12 retrieval tasks, and 4 visual grounding tasks.
Visual Document Retrieval (ViDoRe-v2). We evaluate the performance on visual document retrieval with NDCG@5 on ViDoRe-v2. This is designed to test retrieval systems across diverse document types, tasks, languages, and settings.

3.2 Overall Results

We present the overall multimodal embedding performance on MMEB in Table 1 and the results on long-form document-level retrieval from ViDoRe-v2 in Table 2. Our model, `MoCa`, consistently outperforms all strong baselines on both benchmarks, demonstrating the effectiveness of our proposed framework. Several key observations can be drawn from the results: (1) The combination of a VLM backbone with bidirectional attention, Continual Pre-training (CPT), and heterogeneous Contrastive Learning (CL) consistently yields superior performance. Specifically, the configuration "bidirectional + CPT + CL" outperforms both "causal + CL" and "bidirectional + CL". This confirms the importance of modality-aware pre-training in unlocking the full potential of bidirectional architectures. (2) After continual pre-training on 30B tokens, our 3B model with bidirectional adaptation surpasses 7B baselines trained only with contrastive learning. (3) Scaling our approach to a 7B model leads to

Table 1: MMEB results. In addition to existing baselines, we evaluate three variants of Qwen-2.5-VL with different model sizes and attention mechanisms. We highlight the best scores in **bold** and the second-best with an underline.

Models	Size	Per Meta-Task Score				Average Score		
		Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
Existing Baselines								
CLIP [Radford et al., 2021]	428M	55.2	19.7	53.2	62.2	47.6	42.8	45.4
BLIP2 [Li et al., 2023a]	428M	27.0	4.2	33.9	47.0	–	–	25.2
SigLIP [Zhai et al., 2023]	652M	40.3	8.4	31.6	59.5	–	–	34.8
GME [Zhang et al., 2024]	7B	56.9	41.2	67.8	53.4	–	–	55.8
MM-EMBED [Lin et al., 2024]	7B	48.1	32.2	63.8	57.8	-	-	50.0
VLM2Vec [Jiang et al., 2024]	7B	61.2	49.9	67.4	86.1	67.5	57.1	62.9
MMRet [Zhou et al., 2024]	7B	56.0	57.4	69.9	83.6	68.0	59.1	64.1
mmE5 [Chen et al., 2025a]	11B	**67.6**	62.7	71.0	89.7	72.4	66.6	69.8
Variants of Qwen-2.5-VL with only Contrastive Learning on MMEB Training Set								
causal attn.	3B	59.8	63.8	68.2	83.8	72.7	58.5	66.4
bidirectional attn.	3B	59.1	60.6	68.7	83.4	71.7	57.6	65.4
bidirectional attn.	7B	60.5	62.2	70.5	85.6	72.6	60.1	67.1
Ours								
MoCa-3B	3B	59.8	62.9	70.6	88.6	72.3	61.5	67.5
MoCa-7B	7B	65.8	**64.7**	**75.0**	**92.4**	**74.7**	**67.6**	**71.5**

Table 2: Results on ViDoRe-v2. "Syn" denotes synthetic data, "Mul" indicates multilingual tasks, and "Bio" refers to biomedical domains. The best results are shown in **bold**, while the second-best are underlined.

Models	Size	ESG_Human	Eco_Mul	Bio	ESG_Syn	ESG_Syn_Mul	Bio_Mul	Eco	Avg.
Existing Baselines									
SigLIP [Zhai et al., 2023]	652M	28.8	14.0	33.8	19.8	21.9	18.2	29.8	23.8
VLM2Vec [Jiang et al., 2024]	7B	33.9	42.0	38.8	36.7	38.4	29.7	51.4	38.7
VisRAG-Ret [Yu et al., 2025]	3B	53.7	48.7	54.8	45.9	46.4	47.7	59.6	51.0
GME [Zhang et al., 2024]	7B	**65.8**	56.2	**64.0**	54.3	**56.7**	55.1	62.9	59.3
mmE5 [Chen et al., 2025a]	11B	52.8	44.3	51.3	55.1	54.7	46.8	48.6	50.5
Ours									
MoCa-3B	3B	63.3	57.3	62.5	**58.3**	54.8	59.8	62.8	**59.8**
MoCa-7B	7B	58.8	**57.6**	63.2	55.3	51.4	**61.3**	**63.8**	58.8

substantial improvements across all MMEB task categories, establishing new state-of-the-art results on MMEB. (4) Although the 7B model performs better on MMEB overall, the 3B model achieves slightly higher results on ViDoRe-v2. This is likely because ViDoRe-v2 includes fewer training and evaluation samples, and the smaller model is less prone to overfitting in such low-resource settings. Across the full MMEB benchmark, however, larger models show consistent improvements.

3.3 Ablation Study on CPT & CL stages

To understand the contribution of each major design choice in our framework, we conduct ablation studies on both the *Modality-aware Continual Pre-training* and *Heterogeneous Contrastive Fine-tuning* stages. As shown in Table 3, removing any key component leads to a consistent performance drop on both benchmarks (MMEB and Vidore-v2), demonstrating the importance of each part.

Table 3: Performances of the ablated models. We evaluate the contribution of each component in our framework by removing key elements from both stages.

Model	MMEB	ViDoRe-v2
MoCa-3B	**67.5**	**59.8**
Modality-aware Continual Pre-training Stage		
w/o. MLM	66.2	57.2
w/o. MAE	66.8	56.9
w/o. CPT (MLM & MAE)	65.8	56.2
Heterogeneous Contrastive Fine-tuning Stage		
w/o. Text-only Pairs	67.1	59.2
w/o. Document Retrieval Pairs	66.9	45.7
w/o. Task-aware Batching	67.2	58.8

Modality-aware Continual Pre-training. The ablation of either the *Masked Language Modeling* (MLM) or the *Masked Autoencoding* (MAE) objective results in a performance decline, indicating both text and image reconstruction are essential for learning modality-specific representations.

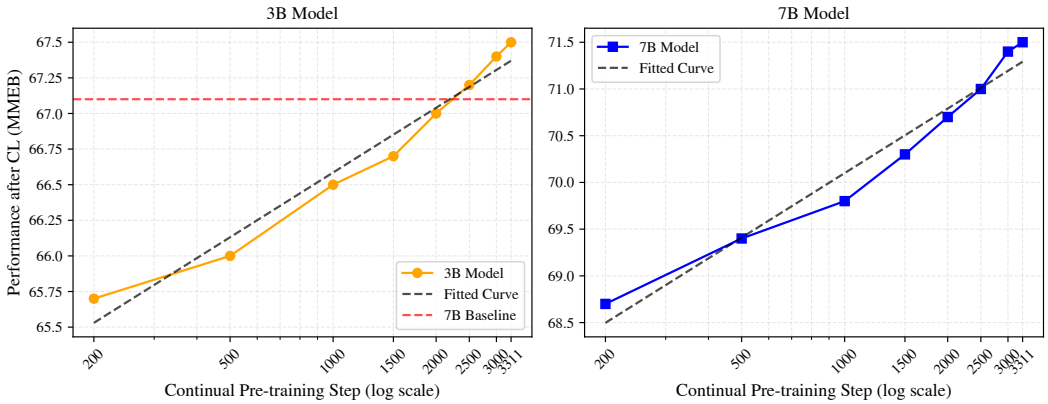

Figure 3: Scaling effect of our CPT stage on downstream performance. We evaluate MMEB performance after CL using checkpoints (left: 3B, right: 7B) from different steps of CPT.

Heterogeneous Contrastive Fine-tuning.[2] We then evaluate the effect of different data type used during contrastive fine-tuning. Removing *text-only pairs* results in a noticeable drop, showing their importance for maintaining strong language representations. Excluding *long-form document retrieval pairs* (VisRAG and the training set of ColPali) also hurts performance, especially on ViDoRe, demonstrating their value in supporting extended contexts. Finally, removing the *task-aware batching* technique leads to performance degradation. This technique helps prevent the model from overfitting to task-specific patterns and encourages it to learn more sample-discriminative representations.

These findings suggest that both stages are necessary. Omitting CPT yields a weaker initialization and degraded performance (see w/o CPT). Besides, CL is what turns a base model into an embedding model by aligning modalities for retrieval. MoCa combines these two indispensable stages: CPT first builds a strong, context-aware bidirectional encoder, and CL then aligns the multimodal embedding space, enabling high-quality retrieval.

3.4 Data Scaling of Continual Pre-training

To evaluate the impact of data scaling on CPT, we analyze how downstream performance changes with increased CPT steps. Specifically, we perform contrastive learning on multiple checkpoints along a single CPT trajectory for 3B and 7B models.

As shown in Figure 3, downstream performance on MMEB improves consistently as the number of CPT steps increases. Notably, after approximately 2,200 steps (~20B tokens), the 3B model achieves performance on par with the 7B baseline trained without CPT. This demonstrates that our CPT stage substantially enhances the quality of bidirectional representations, which in turn improves alignment during the CL stage.

While constrained by computational resources, our findings suggest that further scaling of CPT with more data and training steps can continue to improve model performance. This insight provides practical guidance for balancing training cost with expected gains in future work.

3.5 Hyperparameter Analysis

To further understand the training process of the CPT stage, we conduct experiments of hyperparameter analysis and present the results in Figure 4. We evaluate the performance of MoCa (3B) on MMEB using models trained with a fixed amount of data. We select hyperparameter values based on performance on validation sets, each containing 1K samples drawn from the corresponding training data. To maintain consistency with earlier experiments, we report the results on the MMEB test set.

Mask Ratio. We examine the effect of different masking probabilities for both masked language modeling (MLM) and masked autoencoding (MAE). Increasing the mask ratio generally encourages the model to rely more on contextual signals across modalities, but overly high ratios can lead to degraded learning due to excessive information removal and low signal-to-noise ratio. Our experiments show that moderate masking rates (MLM: 40%, MAE: 50%) strike a good balance, enabling strong cross-modal reasoning without making the model to forget relevant input tokens or patches.

Loss Weight. We also study the weight assigned to the MAE loss in the overall CPT objective. A balanced combination is crucial: if the MAE loss

[2]Without CL the model lacks retrieval capability; MoCa w/o. CL ablation is therefore not reported.

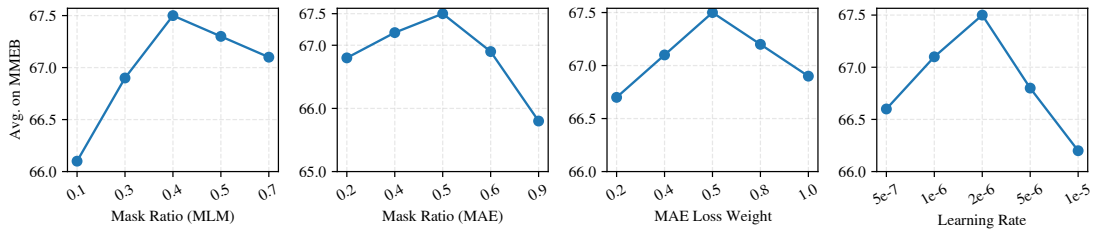

Figure 4: The performances of MoCa (3B) with different CPT settings on MMEB.

Table 4: Generalization of MoCa across different VLM backbones. "Baseline" refers to models trained only with CL. "MoCa" refers to models trained with CPT.

VLM Backbone	Size	Method	MMEB	ViDoRe-v2
Qwen-2.5-VL	3B	Baseline	65.8	56.2
		MoCa	67.5 ↑	59.8 ↑
LLaVA-1.6	7B	Baseline	64.9	54.7
		MoCa	66.2 ↑	57.7 ↑
Phi-3.5-V	3B	Baseline	60.5	51.7
		MoCa	63.8 ↑	55.9 ↑

is underweighted, the model may fail to integrate visual semantics effectively; if overweighted, it may affect the language modeling objective. We find that a MAE loss weight of 0.5 provides the best aligning with both visual and textual objectives.

Learning Rate. The learning rate is a critical factor influencing the stability and convergence speed of continual pretraining. A lower learning rate can lead to underfitting, especially in early training stages, while an overly large learning rate may disrupt the pre-trained knowledge and destabilize training. Through empirical tuning, we find that a learning rate of 2×10^{-6} provides a stable optimization process, allowing the model to adapt effectively to the new denoising objectives without catastrophic forgetting.

3.6 Generalization to Other VLM Backbones

To validate that our MoCa framework can be easily transformed to other backbones, we apply MoCa to multiple causal VLMs. For each VLM, we first train a baseline using only our CL stage. We then run the full two-stage MoCa (CPT → CL) versions.

As shown in Table 4, the full MoCa delivers consistent gains over the CL-only baselines across all backbones on both benchmarks. This confirms that MoCa is a generalizable framework for transforming various off-the-shelf causal VLMs into powerful bidirectional multimodal embedding models.

4 Related Work

Multimodal Embedding. Multimodal embedding models represent inputs from different modalities in a shared space to support cross-modal understanding. ALIGN [Jia et al., 2021], BLIP [Li et al., 2022], and CLIP [Radford et al., 2021] adopt dual-encoder architectures. They encode each modality separately and align their outputs using contrastive learning. Recent works build on pre-trained VLMs [Jiang et al., 2024, Zhang et al., 2024, Chen et al., 2025a]. Texts and images share the same encoder. Despite various techniques to improve contrastive learning [Lan et al., 2025], most approaches still rely on causal models without exploring the advantages of bidirectional architectures.

Multimodal Continual Pre-training Continual pre-training (CPT) involves further training of pre-trained models with additional data or new objectives to adapt to specific domains and tasks [Ke et al., 2023]. In multimodal learning, LXMERT [Tan and Bansal, 2019] and UNITER [Chen et al., 2020] apply MLM during pre-training to learn image-text representations with relatively small transformers. Another line of works explore multimodal CPT with reconstruction objectives [Kim et al., 2021, Ge et al., 2024, Zhao et al., 2025, Wang et al., 2024]. ViLT [Kim et al., 2021] uses MLM and Image-Text Matching (ITM). More recently, Janus [Wu et al., 2024, Ma et al., 2024b, Chen et al., 2025b] apply reconstruction losses to both text tokens and image pixels.

5 Conclusion

We proposed a two-stage framework for multimodal embeddings, combining modality-aware continual pre-training and heterogeneous contrastive fine-tuning. Our method MoCa leverages bidirectional attention mechanisms and joint reconstruction objectives to enhance cross-modal interactions. Additionally, by incorporating diverse and extensive multimodal data, our framework significantly improves the robustness and generaliza-

tion of embedding models. Experiments show that our approach achieves state-of-the-art performance, demonstrating strong scalability with respect to both model and data size on MMEB.

Limitations

While MoCa demonstrates promising performance, several limitations remain:

1. The current framework focuses primarily on image-text pre-training. Extending continual pre-training to include additional modalities—such as video, speech, or structured data—remains an open challenge and may improve cross-modal generalization.

2. The model employs standard denoising objectives and encoder architectures. Exploring more advanced denoising strategies or unified encoder-decoder designs could lead to better representation quality and training efficiency.

3. Our evaluation is limited to common benchmarks. Assessing MoCa on more complex real-world tasks, such as multi-hop retrieval or interleaved-input reasoning, is necessary to fully understand its robustness and practical value.

Acknowledgments

This work was supported by National Natural Science Foundation of China No. 62272467. The work was partially done at the 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 Santacroce, 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. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219, 2024. doi: 10.48550/ARXIV.2404.14219. URL https://doi.org/10.48550/arXiv.2404.14219.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025. doi: 10.48550/ARXIV.2502.13923. URL https://doi.org/10.48550/arXiv.2502.13923.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders. *CoRR*, abs/2404.05961, 2024. doi: 10.48550/ARXIV.2404.05961. URL https://doi.org/10.48550/arXiv.2404.05961.

Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. *CoRR*, abs/2502.08468, 2025a. doi: 10.48550/ARXIV.2502.08468. URL https://doi.org/10.48550/arXiv.2502.08468.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *CoRR*, abs/2501.17811, 2025b. doi: 10.48550/ARXIV.2501.17811. URL https://doi.org/10.48550/arXiv.2501.17811.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020. doi: 10.1007/978-3-030-58577-8_7. URL https://doi.org/10.1007/978-3-030-58577-8_7.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In

Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=ogjBpZ8uSi.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. SEED-X: multimodal models with unified multi-granularity comprehension and generation. *CoRR*, abs/2404.14396, 2024. doi: 10.48550/ARXIV.2404.14396. URL https://doi.org/10.48550/arXiv.2404.14396.

Michael Günther, Isabelle Mohr, Bo Wang, and Han Xiao. Late chunking: Contextual chunk embeddings using long-context embedding models. *CoRR*, abs/2409.04701, 2024. doi: 10.48550/ARXIV.2409.04701. URL https://doi.org/10.48550/arXiv.2409.04701.

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *CoRR*, abs/2412.05237, 2024. doi: 10.48550/ARXIV.2412.05237. URL https://doi.org/10.48550/arXiv.2412.05237.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01553. URL https://doi.org/10.1109/CVPR52688.2022.01553.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *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, 2021. URL http://proceedings.mlr.press/v139/jia21b.html.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=m_GDIItaI3o.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang, editors, *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 5583–5594. PMLR, 2021. URL http://proceedings.mlr.press/v139/kim21k.html.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. *CoRR*, abs/2503.04812, 2025. doi: 10.48550/ARXIV.2503.04812. URL https://doi.org/10.48550/arXiv.2503.04812.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *CoRR*, abs/2408.12637, 2024. doi: 10.48550/ARXIV.2408.12637. URL https://doi.org/10.48550/arXiv.2408.12637.

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. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=lgsyLSsDRe.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee F. Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah M. Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Raghavi Chandu, Thao Nguyen, Igor Vasiljevic, Sham M. Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alex Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models. In Amir Globersons, Lester

Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *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, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *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, 2023a. URL https://proceedings.mlr.press/v202/li23q.html.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *CoRR*, abs/2308.03281, 2023b. doi: 10.48550/ARXIV.2308.03281. URL https://doi.org/10.48550/arXiv.2308.03281.

Sheng-chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms, 2024. URL https://arxiv.org/abs/2411.02571.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 6492–6505. Association for Computational Linguistics, 2024a. URL https://aclanthology.org/2024.emnlp-main.373.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *CoRR*, abs/2411.07975, 2024b. doi: 10.48550/ARXIV.2411.07975. URL https://doi.org/10.48550/arXiv.2411.07975.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *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, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114, 2021. URL https://arxiv.org/abs/2111.02114.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions. *CoRR*, abs/2406.10328, 2024. doi: 10.48550/ARXIV.2406.10328. URL https://doi.org/10.48550/arXiv.2406.10328.

Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1514. URL https://doi.org/10.18653/v1/D19-1514.

Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786, 2025. doi: 10.48550/ARXIV.2502.14786. URL https://doi.org/10.48550/arXiv.2502.14786.

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. *CoRR*, abs/2212.03533, 2022. doi: 10.48550/ARXIV.2212.03533. URL https://doi.org/10.48550/arXiv.2212.03533.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang,

Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *CoRR*, abs/2409.18869, 2024. doi: 10.48550/ARXIV.2409.18869. URL https://doi.org/10.48550/arXiv.2409.18869.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *CoRR*, abs/2410.13848, 2024. doi: 10.48550/ARXIV.2410.13848. URL https://doi.org/10.48550/arXiv.2410.13848.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *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, 2024. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=zG459X3Xge.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 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, 2023. doi: 10.1109/ICCV51070.2023.01100. URL https://doi.org/10.1109/ICCV51070.2023.01100.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms, 2024. URL http://arxiv.org/abs/2412.16855.

Yue Zhao, Fuzhao Xue, Scott Reed, Linxi Fan, Yuke Zhu, Jan Kautz, Zhiding Yu, Philipp Krähenbühl, and De-An Huang. QLIP: text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation. *CoRR*, abs/2502.05178, 2025. doi: 10.48550/ARXIV.2502.05178. URL https://doi.org/10.48550/arXiv.2502.05178.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. Megapairs: Massive data synthesis for universal multimodal retrieval. *arXiv preprint arXiv:2412.14475*, 2024.

Appendix

A Detailed Results on MMEB-V2

We present the detailed results of MoCa and baseline models on the MMEB benchmark [Jiang et al., 2024] in Table 5, covering 36 tasks across four categories: classification, VQA, retrieval, and visual grounding.

Table 5: Detailed performance of multimodal models on 36 MMEB tasks [Jiang et al., 2024]. We show results of baseline models and our method (MoCa) at 3B and 7B scales.

Task	CLIP	OpenCLIP	SigLIP	BLIP2	VLM2Vec	MMRet	mmE5	MoCa (3B)	MoCa (7B)
Classification (10 tasks)									
ImageNet-1K	55.8	63.5	45.4	10.3	74.5	58.8	77.6	75.4	78.0
N24News	34.7	38.6	13.9	36.0	80.3	71.3	82.1	80.9	81.5
HatefulMemes	51.1	51.7	47.2	49.6	67.9	53.7	64.3	70.6	77.6
VOC2007	50.7	52.4	64.3	52.1	91.5	85.0	91.0	87.0	90.0
SUN397	43.4	68.8	39.6	34.5	75.8	70.0	77.9	74.8	76.8
Place365	28.5	37.8	20.0	21.5	44.0	43.0	42.6	38.8	43.0
ImageNet-A	25.5	14.2	42.6	3.2	43.6	36.1	56.7	39.7	52.7
ImageNet-R	75.6	83.0	75.0	39.7	79.8	71.6	86.3	75.4	83.0
ObjectNet	43.4	51.4	40.3	20.6	39.6	55.8	62.2	31.3	45.2
Country-211	19.2	16.8	14.2	2.5	14.7	14.7	34.8	24.0	30.4
All Classification	42.8	47.8	40.3	27.0	61.2	56.0	67.6	59.8	65.8
VQA (10 tasks)									
OK-VQA	7.5	11.5	2.4	8.7	69.0	73.3	67.9	40.0	36.9
A-OKVQA	3.8	3.3	1.5	3.2	54.4	56.7	56.4	54.6	57.1
DocVQA	4.0	5.3	4.2	2.6	52.0	78.5	90.3	93.0	94.3
InfographicsVQA	4.6	4.6	2.7	2.0	30.7	39.3	56.2	67.7	77.2
ChartQA	1.4	1.5	3.0	0.5	34.8	41.7	50.3	64.1	69.8
Visual7W	4.0	2.6	1.2	1.3	49.8	49.5	51.9	61.6	58.5
ScienceQA	9.4	10.2	7.9	6.8	42.1	45.2	55.7	45.4	59.2
VizWiz	8.2	6.6	2.3	4.0	43.0	51.7	52.8	52.3	46.2
GQA	41.3	52.5	57.5	9.7	61.2	59.0	62.1	66.9	71.6
TextVQA	7.0	10.9	1.0	3.3	62.0	79.0	83.5	83.1	75.8
Avg. VQA	9.1	10.9	8.4	4.2	49.9	57.4	62.7	62.9	64.7
Retrieval (12 tasks)									
VisDial	30.7	25.4	21.5	18.0	80.9	83.0	73.7	80.5	84.5
CIRR	12.6	15.4	15.1	9.8	49.9	61.4	54.9	55.7	53.4
VisualNews_t2i	78.9	74.0	51.0	48.1	75.4	74.2	77.7	74.4	78.2
VisualNews_i2t	79.6	78.0	52.4	13.5	80.0	78.1	83.4	77.8	83.1
MSCOCO_t2i	59.5	63.6	58.3	53.7	75.7	78.6	76.2	76.4	79.8
MSCOCO_i2t	57.7	62.1	55.0	20.3	73.1	72.4	73.6	72.6	73.9
NIGHTS	60.4	66.1	62.9	56.5	65.5	68.3	68.8	67.4	66.7
WebQA	67.5	62.1	58.1	55.4	87.6	90.2	88.1	90.6	91.4
FashionIQ	11.4	13.8	20.1	9.3	16.2	54.9	28.6	22.2	28.9
Wiki-SS-NQ	55.0	44.6	55.1	28.7	60.2	24.9	65.2	73.3	82.7
OVEN	41.1	45.0	56.0	39.5	56.5	87.5	77.3	75.9	80.4
EDIS	81.0	77.5	23.6	54.4	87.8	65.6	83.6	80.8	96.9
Avg. Retrieval	53.0	52.3	31.6	33.9	67.4	69.9	71.0	70.6	75.0
Visual Grounding (4 tasks)									
MSCOCO	33.8	34.5	46.4	28.9	80.6	76.8	85.0	80.2	84.6
RefCOCO	56.9	54.2	70.8	47.4	88.7	89.8	92.7	92.1	94.0
RefCOCO-matching	61.3	68.3	50.8	59.5	84.0	90.6	88.9	92.8	95.5
Visual7W-pointing	55.1	56.3	70.1	52.0	90.9	77.0	92.3	89.5	95.3
Avg. Grounding	51.8	53.3	59.5	47.0	86.1	83.6	89.7	88.7	92.4
Final Score (36 tasks)									
All IND Avg.	37.1	39.3	32.3	25.3	67.5	59.1	72.4	72.3	74.7
All OOD Avg.	38.7	40.2	38.0	25.1	57.1	68.0	66.6	61.5	67.6
All Tasks Avg.	37.8	39.7	34.8	25.2	62.9	64.1	69.8	67.5	71.5