Compressing then Matching: An Efficient Pre-training Paradigm for Multimodal Embedding

Da Li [1,2*] **Yuxiao Luo** [3*] **Keping Bi**[1,2 †] **Jiafeng Guo**[1,2†] **Wei Yuan**[3‡]
Biao Yang[3] **Yan Wang**[3] **Fan Yang**[3] **Tingting Gao**[3] **Guorui Zhou**[3]
[1]State Key Laboratory of AI Safety, Institute of Computing Technology,
Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences [3] Kuaishou Technology
{lida21s, bikeping,guojiafeng}@ict.ac.cn
{luoyuxiao, yuanwei05, yangbiao, wangyan33, yangfan}@kuaishou.com

Abstract

Multimodal Large Language Models advance multimodal representation learning by acquiring transferable semantic embeddings, thereby substantially enhancing performance across a range of vision-language tasks, including cross-modal retrieval, clustering, and classification. An effective embedding is expected to comprehensively preserve the semantic content of the input while simultaneously emphasizing features that are discriminative for downstream tasks. Recent approaches demonstrate that MLLMs can be adapted into competitive embedding models via large-scale contrastive learning, enabling the simultaneous optimization of two complementary objectives. We argue that the two aforementioned objectives can be decoupled: a comprehensive understanding of the input enables the embedding model to achieve superior performance on downstream tasks via contrastive learning. In this paper, we propose **CoMa**, a compressed pre-training phase, which serves as a warm-up stage for contrastive learning. Experiments demonstrate that with only a small amount of pre-training data, we can transform an MLLM into a competitive embedding model. CoMa achieves new state-of-the-art results among MLLMs of comparable size on the MMEB, realizing optimization in both efficiency and effectiveness. Our project is available at https://github.com/Trustworthy-Information-Access/CoMa.

1 Introduction

Multimodal embedding is a core research area in artificial intelligence. It integrates heterogeneous data from modalities such as text, images, audio, and video to build cross-modal representations with rich semantics. These representations play an essential role in enabling key downstream tasks like image-text retrieval (Wang et al., 2024b; Tang et al., 2025), Retrieval Augmented Generation (RAG) (Yao et al., 2025; Gao et al., 2024), and Visual Question Answering (VQA) (Gardères et al., 2020; Chun et al., 2021).

As a cornerstone paradigm, contrastive learning-based dual-encoder models (e.g., CLIP (Radford et al., 2021), BLIP (Li et al., 2022), ALBEF (Li et al., 2021)) have achieved remarkable results by aligning heterogeneous modalities through large-scale paired data. These models primarily focus on aligning global semantics while neglecting the fine-grained semantic correspondences between local components. This deficiency leads to suboptimal performance in downstream tasks such as visual grounding (Xiao et al., 2024) and attribute-focused image retrieval (Li et al., 2025).

Multimodal Large Language Models (MLLMs) have demonstrated significant advances in generalized vision-language understanding. Unlike CLIP-based models that encode text and images separately before alignment, MLLMs like Qwen-VL (Bai et al., 2025) and LLaVA-OneVision (Li et al., 2024) utilize interleaved text-image sequences as input. This approach enables direct capture of fine-grained semantic correspondences between text and images. VLM2Vec (Jiang et al., 2025) proposes a contrastive training framework that transforms MLLMs into general multimodal embedding models. E5-V (Jiang et al., 2024) maps different modality inputs to a unified embedding space through specially formatted prompts, eliminating modality gaps without relying on multimodal training data. GME (Zhang et al., 2025) advances multimodal embedding by introducing large-scale, high-quality synthetic multimodal datasets, significantly improving cross-modal retrieval performance in MLLMs.

While multimodal embeddings have achieved substantial performance improvements through

[*]Contributed equally
[†]Corresponding authors
[‡]Project leader

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3707–3718
July 2-7, 2026

MLLMs, these advances remain predominantly data-dependent rather than methodologically grounded. MLLMs are inherently constrained by their autoregressive next token prediction objective, which fundamentally differs from the task-related application format associated with embeddings. Contrastive learning based on massive data is not an efficient method for achieving transformation between the two task paradigms. Some studies have attempted to address this issue by proposing a pre-training stage for multimodal embeddings to achieve efficient transformation. UniME (Gu et al., 2025) optimizes the language component embeddings in MLLMs by distilling knowledge from a text embedding model. MoCa (Chen et al., 2025a) replaces the causal attention mechanism of MLLM with bidirectional attention and designs a mask-based pre-training task to optimize the embedding learning process. The success of the above methods also heavily relies on high-quality, relevant data. In this paper, we propose a simple and effective pre-training strategy to reduce reliance on high-quality data.

We consider that a good embedding should possess two key characteristics: (1) **Comprehensive Information Coverage**: A good embedding should encompass as much of the input information as possible. (2) **Distinguishing Features**: Information relevant to matching should be highlighted within the embedding. The previous methods assumed that contrastive learning could achieve the simultaneous optimization of two objectives. Therefore, the optimization process requires a large amount of data. We attempt to decompose the optimization of them through a compressed pre-training task. During the compression pre-training phase, we divided the input into three parts: the input image, a set of learnable compression tokens, and image-based dialogue. By modifying the attention mechanism, we constrained the compressed tokens to extract information only from the image. Then we trained MLLMs to recover information from these compressed tokens, thereby completing the dialogue generation. At this stage, MLLMs are encouraged to generate comprehensive and rich compressed representations to address a wide variety of questions. During the contrastive learning phase, MLLMs focus on compressing token embeddings that are relevant to matching, thereby enhancing retrieval performance. Unlike other pre-training methods, our approach requires a smaller amount of data. The effectiveness of our proposed pre-training method hinges on whether the dialogue data is complex and diverse. To this end, we designed a high-quality data generation strategy that enables MLLM to automatically generate multi-turn dialogue data from a single image. This further reduces our reliance on data sources. Experimental results demonstrate that our approach achieves comparable performance to other pre-training methods while utilizing only approximately 10% of the training data volume required by other pre-training methods. Our main contributions are summarized as follows:

- We propose a compressed pre-training strategy combined with downstream contrastive learning, successfully transforming MLLMs into competitive multimodal embedding models.
- To reduce reliance on high-quality data during the pretraining phase, we propose an automated data synthesis method to supply data for our pre-training.
- Extensive experiments show that our pre-training strategy is simple and efficient. Training MLLMs with LoRA on small datasets can achieve competitive performance. We also conducted extensive analyses to show how it takes effect.

2 Related Work

2.1 Vision-Language Models for Multimodal Embedding

Embedding models are fundamental components of numerous downstream applications, including retrieval, clustering, and classification. Early works like CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and their variants primarily focused on learning universal representations from large-scale image-text pairs. These models encode images and text separately, then align them in a latent space. As a result, they primarily focus on aligning global semantics while neglecting the fine-grained semantic correspondences between local regions. Multimodal large language models process data from different modalities through an interleaved text-image format. They provide powerful backbones for multimodal embedding models. VLM2Vec (Jiang et al., 2025; Meng et al., 2025) transforms multimodal large language models (MLLMs), such as Phi-3.5-vision (Abdin et al., 2024), LLaVA-1.6 (Liu et al., 2024), and Qwen-VL (Wang et al., 2024a; Bai

et al., 2025), into competitive multimodal embedding models through large-scale contrastive learning. E5-V (Jiang et al., 2024) employs specially designed prompts to project inputs from diverse modalities into a unified representation space, effectively bridging modality gaps without multimodal training data. GME (Zhang et al., 2025) enhances multimodal embeddings by introducing large-scale, high-quality synthetic multimodal datasets, which mitigate modality imbalance in training and substantially boost the cross-modal retrieval performance of MLLMs.

2.2 Pretraining for MultiModal Embedding

While contrastive learning can effectively learn global alignment across modalities, it struggles to support deep cross-modal integration and fine-grained semantic understanding. To overcome this limitation, the objectives different from contrastive learning are integrated into the model's pre-training process to enhance performance.

In multimodal learning, LXMERT (Tan and Bansal, 2019) and UNITER (Chen et al., 2020) apply Masked Language Modeling (MLM) during pretraining to jointly learn image-text representations. Research on multimodal pre-training has also explored reconstruction-based objectives (Kim et al., 2021; Bao et al., 2022; Chen et al., 2025b; Wu et al., 2024; Chen et al., 2025c; Ge et al., 2025). Based on CLIP, ALBEF (Li et al., 2021) and ViLT (Kim et al., 2021) introduce Image-Text Matching and Masked Language Model tasks to learn multimodal fusion and fine-grained representations. BLIP (Li et al., 2022) and CoCa (Yu et al., 2022) integrate the language modeling task into the pretraining process of embeddings, leveraging the model's generative capabilities to optimize embedding performance.

Multimodal embedding models based on MLLMs have achieved significant progress across various evaluation tasks through large-scale contrastive learning. MLLMs are primarily designed to perform text generation tasks, differing from embedding applications. This gap highlights the need for designing pre-training methods tailored to embedding tasks. UniME (Gu et al., 2025) introduces a pretraining stage to enhance the multimodal embedding capabilities of the model. It involves pretraining with textual discriminative knowledge distillation, where knowledge is transferred from a powerful LLM-based teacher embedding model to strengthen the language component of MLLMs. Considering the efficiency of bidirectional attention for data encoding, MoCa (Chen et al., 2025a) introduces a modality-aware continual pre-training stage. This phase employs a joint reconstruction objective that denoises interleaved text-image inputs using both Masked Language Modeling (MLM) and Masked Autoencoding (MAE), effectively enhancing the model's capacity for bidirectional contextual representation and cross-modal alignment.

3 Method

3.1 Preliminary

3.1.1 Multimodal Embedding Models

VLMs split images I and text T into patches and tokens, respectively, map them into the same embedding space through a text encoder and a visual encoder, and finally generate unified hidden state features $H \in \mathbb{R}^{L\times N\times D}$. Most multimodal embedding models derive a holistic representation of the input from these hidden states. For causal models, the hidden states in each layer depend only on previous hidden states, and many works select the last token of the final layer—typically corresponding to the [EOS] token. In contrast, some VLM models convert causal attention into bidirectional attention, thereby gaining access to all input information. Consequently, they often use mean pooling over all sequence representations or train a special token to aggregate the final embedding.

3.1.2 Contrastive Learning

In the optimization of embedding models, contrastive learning serves as the core training paradigm. Its fundamental principle involves learning discriminative representations by pulling semantically similar samples closer together while pushing irrelevant samples farther apart. This paradigm has been widely applied in downstream tasks such as retrieval and recommendation systems. During training, each instance is structured as a tuple $(q, d^+, \{d_1^-, ..., d_K^-\})$ where q is the query, d^+ is a positive item, and $\{d_1^-, ..., d_{|K|}^-\}$ corresponds to a set of K negative samples within the same batch. The widely used objective function in this setting is the InfoNCE (He et al., 2020), defined as:

$$\mathcal{L} = -\log \frac{e^{sim(h_q, h_{d^+})/\tau}}{e^{sim(h_q, h_{d^+})/\tau} + \sum_{i=1}^{K} e^{sim(h_q, h_{d_i^-})/\tau}},$$

where $sim(\cdot)$ denotes the similarity function and τ is the temperature.

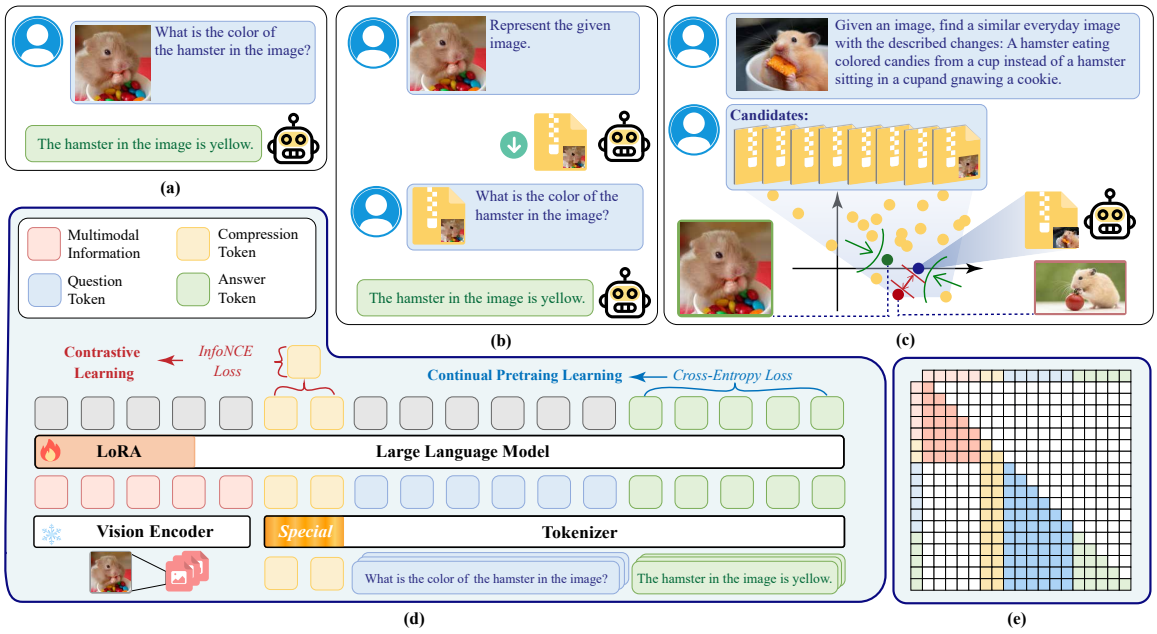


Figure 1: Architecture of CoMa, the top part demonstrates three training stages: **(a) Instruction-Tuning**, **(b) Continual PreTraining** and **(c) Contrastive Learning**. In **(a)** and **(b)**, our continual pretraining stage inherits the same format of Question-Answering (QA) task as the upstream stage, while the answer in the upstream depends on the whole image, and our stage depends on the Compression Tokens which condense the image. In **(c)**, the contrastive learning stage inherits the compression tokens as representations of multimodal input to apply to upstream retrieval-based tasks. The bottom part shows the implementation of our continual pretraining stage: **(d)** shows the model architecture where all the inputs are concatenated to calculate simultaneously, and the dependencies are driven by our modified causal attention masks in **(e)**.

3.2 Compression Pretraining

We argue that a good representation should first incorporate as much of the input information as possible, then highlight the part of it related to matching. However, previous studies (Chen et al., 2024; Ye et al., 2025) reveal substantial spatial and semantic redundancies in existing representations. To reduce these redundancies and improve representation efficiency, we designed an additional compressed pre-training stage to train multimodal embedding models to extract comprehensive information from the input.

3.2.1 Compression based on Question-Answer

Question-answering data formats are widely used in training large language models due to their flexibility, particularly in the critical stages of SFT and RLHF. Benefiting from their formalization capabilities, common tasks such as classification and summarization can be uniformly converted into question-answering formats. Furthermore, some complex evaluations for MLLMs are also conducted through question-answering.

We argue that leveraging question-answering data as a supervisory signal offers a novel approach for optimizing the compression capabilities of multimodal embedding models. Given an image, when questions are diverse and complex, MLLMs must comprehensively and accurately understand the image information to provide corresponding answers. MLLMs trained through this method can serve as a superior backbone for multimodal embedding models.

3.2.2 Compression Mechanism

For an input image I and its corresponding question Q and answer A, we first insert a set of compression tokens $C = [C_1, ..., C_K]$ behind the image, where K is significantly smaller than the length of the image input. The input is serialized as:

$$\begin{aligned}\text{Input} = &< |\text{image\_pad}| >, \ldots, < |\text{image\_pad}| >, \\ &< C_1 >, \ldots, < C_K >, [\text{Question}], [\text{Answer}].\end{aligned}$$

Unlike the training process of SFT, our pretraining process aims to maximize the following objective:

$$P(A \mid \mathcal{C} \oplus Question;\ \theta).$$

Benefiting from casual attention, compressed tokens extract information from images to support

the subsequent training of question-answering. The pretraining method we propose effectively bridges the gap between instruction models and embedding models. It not only fully leverages the knowledge stored in model parameters but also is similar in task form to SFT. Unlike SFT, which relies on high-quality question-answering data (typically requiring questions to be as complex as possible and answers to be strictly accurate) for training, our proposed compression pre-training method emphasizes comprehensive coverage and diversity of questions without strictly demanding answer accuracy.

After pretraining, the contrastive learning phase proceeds without any conversational components. We extract representations from the final hidden states corresponding to the compression tokens, apply mean pooling to aggregate these features, and utilize contrastive learning to align the multimodal embeddings.

3.2.3 Attention Mask-Guided Information Compression

In our implementation, the compression capability is integrated into the calculation of the QA loss. As illustrated in Figure 1(d), the multimodal input, compression tokens C, and conversational inputs $Q \oplus A$ are concatenated into a single sequence. A key design consideration is the dependency structure among these three components: the compression tokens naturally depend on the multimodal input, following standard practice. Crucially, the conversational segment depends solely on the compression tokens, implying that the hidden states of the conversational part cannot be computed using information from the input segment. To enforce this dependency structure, we modify the causal attention mask. As shown in Figure 1**(e)**, we mask the lower triangular region between the conversational and information segments, setting the corresponding attention scores to zero.

3.3 Automatic Data Generation

Existing pre-training methods often exhibit strong dependence on both the quantity and diversity of training data. Our proposed pre-training paradigm leverages complex, multi-source QA data, which places higher demands on data quality. Drawing inspiration from related work such as Self-Instruct (Wang et al., 2023), we explore the potential of MLLMs to generate diverse, high-quality training samples autonomously.

Given the retrieval-oriented characteristics of our task and the objectives of our pre-training, we avoid random instruction generation and instead prioritize the coverage and comprehensiveness of instructions with respect to image content. We employed Qwen2.5-VL-7B to generate three to five potential questions for an image at random, and instructed Qwen2.5-VL-7B to provide answers to the generated questions through a multi-turn dialogue format. Detailed instructions are provided in the Figure 6, and statistical information of the pre-training data is presented in the Table 1.

Table 1: Statistics of Pretraining Data.

Dataset \ # Turns	3	4	5	Total
CIRR	79	138	16,423	16,640
HatefulMemes	74	26	8,400	8,500
MSCOCO	706	399	2,2507	2,3612
MSCOCO_i2t	105	64	29,830	29,999
MSCOCO_t2i	115	54	29,831	30,000
N24News	59	42	29,899	30,000
SUN397	9	14	19,827	19,850
VOC2007	227	156	6,293	6,676
Visual7W	23	25	14,318	14,366
WebQA	17	7	12,849	12,873
Total	1,739	1,295	219,482	222,516

4 Experiments

4.1 Datasets

Both the pre-training and contrastive learning data for CoMa originate from the MMEB-V1 (Jiang et al., 2025), which comprises 36 datasets categorized into 4 meta-tasks: classification, visual question answering, retrieval, and visual grounding. We randomly sampled approximately 220K examples from the MMEB-V1 training set and constructed a pre-training dataset based on the images it contained, following the method described in Section 3.3. Statistical information about the pre-training dataset is shown in Appendix Table 1. After pretraining is completed, we use only the training set of MMEB-V1 for contrastive learning.

4.2 Training Procedure

We employ the Qwen2.5-VL as the backbone for our multimodal embedding model. Our training procedure consists of two stages: compression pre-training followed by contrastive learning. To handle inputs from multiple modalities and accommodate images of varying sizes, we employ dynamic

Table 2: **Results on MMEB-V1.** "IND" denotes in-distribution, and "OOD" refers to out-of-distribution. **Bold** and underline indicate the optimal and suboptimal performance, respectively.

Models	#Params	Per Meta-Task Score				Average Score		
		Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# of Datasets →		10	10	12	4	20	16	36
CLIP (ViT-L)	0.4B	55.2	19.7	53.2	62.2	47.6	42.8	45.4
OpenCLIP (ViT-L)	0.4B	41.5	6.9	44.6	53.5	32.8	36.0	36.6
GME (Qwen2-VL)	2B	56.9	41.2	67.8	53.4	-	-	-
UNITE (Qwen2-VL)	2B	63.2	55.9	65.4	75.6	65.8	60.1	63.3
VLM2Vec (Qwen2.5-VL)	3B	55.3	57.3	62.7	73.2	-	-	60.3
E5-V (LLaVA-1.6)	7B	39.7	10.8	39.4	60.2	34.2	33.9	33.9
MMRet (LLaVA-1.6)	7B	56.0	57.4	69.9	83.6	68.0	59.1	65.8
VLM2Vec (Qwen2-VL)	7B	62.6	57.8	69.9	81.7	72.2	57.8	65.8
CAFe (LLaVA-OV)	7B	65.2	65.6	70.0	91.2	**75.8**	62.4	69.8
UNITE (Qwen2-VL)	7B	**68.3**	65.1	71.6	84.8	73.6	66.3	70.3
mmE5 (Llama-3.2-Vision)	11B	67.6	62.8	70.9	89.7	72.3	66.7	69.8
UniME (Phi3.5-V)	4.2B	54.8	55.9	64.5	81.8	68.2	52.7	64.2
UniME (LLaVA-1.6)	7B	60.6	52.9	67.9	85.1	68.4	57.9	66.6
MoCa (Qwen2.5-VL)	3B	59.8	62.9	70.6	88.6	72.3	61.5	67.5
MoCa (Qwen2.5-VL)	7B	65.8	64.7	**75.0**	**92.4**	74.7	**67.6**	71.5
CoMa (Qwen2.5-VL)	3B	61.3	65.1	70.0	82.7	71.3	61.6	67.5
CoMa (Qwen2.5-VL)	7B	67.4	**70.6**	72.4	87.6	75.2	**67.6**	**72.2**

resolution via MRoPE (Wang et al., 2024a), limiting the maximum number of vision tokens to 1024. The number of compression tokens is 32. Due to computational resource constraints, we set the batch size to 256 during the pretraining phase. In the contrastive learning phase, we scaled it up to 1024 using the GradCache (Gao et al., 2021). For all experiments, we used LoRA (Hu et al., 2021) (rank=16) and a learning rate of 5e-5 in both stages, and our GPU requirements are only one-quarter of those for MoCa (Chen et al., 2025a).

4.3 Evaluation and Metrics

We evaluate the performance of CoMa on the MMEB-V1 benchmark, which provides evaluation benchmarks across four meta-tasks consistent with those in the training set. The benchmark comprises 36 evaluation datasets, including 20 in-distribution and 16 out-of-distribution subsets. We employ Precision@1 as our evaluation metric, emphasizing top-ranked results to reflect their applicability in real-world scenarios.

5 Overall Performance

We compared the performance of CoMa against other competitive multimodal embedding models on the MMEB. The results are shown in Table 2. We categorize these baselines into two groups: one directly employs contrastive learning, while the other incorporates an additional pre-training stage. For the same backbone, introducing additional pre-training can effectively improve its retrieval performance. This highlights the importance of the pre-training stage when converting MLLMs into multimodal embedding models. Compared to other pre-training approaches, the compressed pre-training stage proposed in CoMa demonstrates superior effectiveness. Experimental results show that CoMa achieves optimal or near-optimal levels across multiple key metrics, further validating the effectiveness of the compressed pre-training strategy. In terms of training efficiency, CoMa employs the LoRA for training. Compared to the best baseline MoCa, CoMa utilizes only 300 million tokens during the pre-training phase, significantly fewer than the 30 billion tokens required by MoCa. Furthermore, during the contrastive learning phase, CoMa achieves state-of-the-art performance using only half the training data of MoCa and with a smaller batch size. This demonstrates that CoMa is both simple and efficient, significantly reducing computational resource requirements without sacrificing performance.

6 Further Analysis

6.1 Scaling of Compression Tokens

We utilised 32 compressed tokens to extract information from the multimodal input. To investigate the impact of the number of tokens on performance,

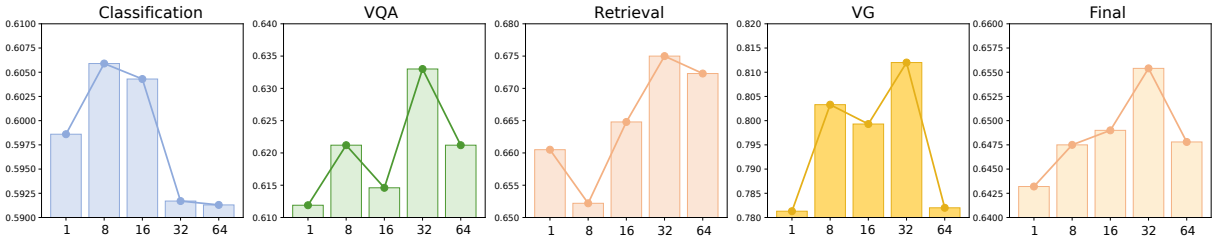
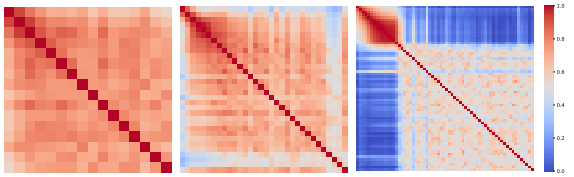


Figure 2: The impact of different numbers of compressed tokens on performance.

we adjusted the number of compressed tokens in CoMa, retrained the multimodal embedding model, and compared performance across different tasks. We employed Qwen2.5-VL-3B as the backbone of CoMa for analysis. Considering training efficiency, we only used 500K training samples randomly sampled from the MMEB-V1 during the contrastive learning phase. The results are shown in Figure 2.



(a) (b) (c)

Figure 3: Similarity between compression tokens with different numbers: (a) 16, (b) 32, and (c) 64.

We found that, regardless of the number of compressed tokens used, CoMa outperforms the baseline (VLM2Vec based on the same backbone) without additional pretraining. This demonstrates the effectiveness of the compressed pre-training strategy. As the number of compressed tokens increases, CoMa's average performance across different tasks follows an initial upward trend followed by a decline: performance gradually improves in the initial phase, but beyond 32 tokens, performance actually decreases as the token count increases. It is evident that as the number of compression tokens increases, the capacity of the compression space expands, thereby enabling the compression tokens to incorporate a greater amount of information. To investigate why performance declines as the number of compressed tokens decreases, we evaluated the pairwise similarity between compressed tokens and presented the results in Figure 3. They are calculated on 100 randomly sampled instances. Compared to 16 tokens, 32 tokens provide sufficient space for compressing input information. The compression space offered by 64 tokens contains redundant information (Dark Blue Area) that may interfere with matching, leading to performance degradation. The number of compressed tokens is closely related to the amount of data.

6.2 The Impact of Different Pretraining Methods

To explore what data formats and training methods enable CoMa to learn compression capabilities more effectively, we designed several ablation experiments concerning the quantity mapping between images and questions, including: a multi-turn conversation format (one image to multiple questions) and its variant, the single-turn conversation format (one image to one question). In addition, we altered the content of the question-answer pairs, replacing them with descriptions and captions to analyse the impact of different content types on performance. In terms of training methods, we employed KL as an alternative to the cross-entropy loss in Next Token Prediction for comparison. Regarding the experimental setting, for contrastive learning, we only used 500K training samples consistent with the ablation experiments in Section 6.1, and other parameter settings were kept the same as the main experiment in Section 4.2. The results under different settings are shown in Table 3.

Table 3: Performance under different formats and training methods. Optimal results are displayed in **bold**, and the suboptimal results are shown with underlining.

Format	Loss	Per Meta-Task				Avg
		Classification	VQA	Retrieval	VG	
Multi Turn		59.2	**63.3**	67.5	**81.2**	**65.5**
Single Turn		59.6	62.2	66.3	79.6	64.8
Description	CE	59.9	61.6	66.7	80.8	64.9
Caption		**60.7**	62.1	**67.5**	77.2	65.2
Multi Turn	KL	58.1	61.7	67.4	77.6	64.4

We found that using a multi-turn dialogue format yields better results than a single-turn dialogue format. A reasonable hypothesis is that compression differs from instruction tuning in that compression is inherently a lossy process, which means some

detailed information must be discarded during compression. Under this premise, setting single questions may lead to excessive focus on the detailed aspects of the image. While multiple questions focus on different aspects of the same image, the model may autonomously balance what needs to be compressed and what can be discarded to minimise the information loss. We also found that replacing dialogue with detailed image descriptions or captions did not yield satisfactory results. This comparison demonstrates the critical importance of information coverage for pre-training. For massive datasets, image caption serves as a suboptimal yet efficient alternative solution.

6.3 Effect of Compression Pretraining

The compression pretraining strategy we propose is close to supervised fine-tuning in the training paradigm, while it is similar to embedding models in terms of functionality. To explore the underlying mechanisms of compression pre-training, we analyzed how representations evolve for the same input across different training stages and visualized these changes in Figure 4. This result provides an intuitive illustration of the model's representational evolution.

(a) (b) (c)

Figure 4: Representations of queries and targets across Three Stages: (a) Base, (b) Compression Pretraining, and (c) Contrastive Learning. Representations in (a) are extracted from the base model via the [EOS] token, while the others are via compression tokens. The representations are decomposed using Principal Component Analysis (PCA).

It is evident that after compressed pre-training, the representation of the same input becomes closer to the representation obtained after final contrastive learning. The compressed pre-training method bridges the gap between instruction models and embedding models, effectively reducing the fitting difficulty in contrastive learning.

6.4 Is Distillation Better For Compression?

Unlike standard instruction-tuning, our approach focuses on optimising the compression tokens in the latent space. It seems to work better when using distillation training methods, such as applying the KL divergence loss, as it allows for fine-grained optimisation. However, as shown in the Table 3, the training based on cross-entropy loss actually brings better performance. Intuitively, KL divergence requires complete consistency between output distributions, which seems overly strict supervision for inherently lossy tasks like compression. What we need is to strike a balance between compression capability and generalisation ability, but the KL criterion may limit the generalisation capability of the model. We cannot require the distribution of the model's output to be consistent with that before compression; instead, we can only supervise the expected answers, which is achieved by the standard cross-entropy loss.

The Figure 5 shows an example of training loss distribution under the same settings and instance. It can be seen that both cross-entropy and KL divergence focus most of their loss on the similar tokens (those related to the answer). In comparison, cross-entropy is more concentrated, while KL divergence distributes considerable loss to more tokens, where many of them are unimportant for compression, such as [EOS].

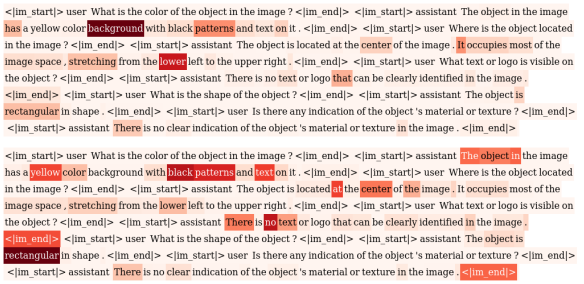


Figure 5: Loss distribution across tokens for Cross-Entropy (Top) and KL Divergence (Bottom).

7 Conclusion

In this paper, we propose CoMa, which decouples the compression and matching functionalities within multimodal embedding models by introducing an additional compression pre-training process. Experiments demonstrate that CoMa is both straightforward and effective, achieving state-of-the-art performance on MMEB-V1. The pre-training process of CoMa utilises only images as compressed input. However, CoMa is not limited to processing images. It can also handle multimodal data such as plain text and video. We will explore

the impact of compressing different multimodal data on CoMa's performance in future work.

Limitations

Constrained by training resources, CoMa was pre-trained and contrastively learned only on a limited amount of data. Its performance upper bound remains to be further explored following subsequent increases in data scale.

Acknowledgments

This work was funded by New Generation Artificial Intelligence-National Science and Technology Major Project of No. 2025ZD0123301, the National Natural Science Foundation of China (NSFC) under Grants No. 62302486 and No. 62441229, the Innovation Project of ICT CAS under Grants No. E361140, and the Strategic Priority Research Program of the CAS under Grants No. XDB0680102.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, 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, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*, volume 35, pages 32897–32912. Curran Associates, Inc.

Haonan Chen, Hong Liu, Yuping Luo, Liang Wang, Nan Yang, Furu Wei, and Zhicheng Dou. 2025a. Moca: Modality-aware continual pre-training makes better bidirectional multimodal embeddings. *Preprint*, arXiv:2506.23115.

Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. 2025b. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. *Preprint*, arXiv:2502.08468.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. *Preprint*, arXiv:2403.06764.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025c. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *Preprint*, arXiv:2501.17811.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Computer Vision – ECCV 2020*, pages 104–120, Cham. Springer International Publishing.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic embeddings for cross-modal retrieval. *Preprint*, arXiv:2101.05068.

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

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. ConceptBert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498, Online. Association for Computational Linguistics.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2025. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *Preprint*, arXiv:2404.14396.

Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. 2025. Breaking the modality barrier: Universal embedding learning with multimodal llms. *Preprint*, arXiv:2504.17432.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. *Preprint*, arXiv:1911.05722.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

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

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *Preprint*, arXiv:2410.05160.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Preprint*, arXiv:2201.12086.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Preprint*, arXiv:2107.07651.

Siting Li, Xiang Gao, and Simon Shaolei Du. 2025. Highlighting what matters: Promptable embeddings for attribute-focused image retrieval. *Preprint*, arXiv:2505.15877.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. 2025. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *Preprint*, arXiv:2507.04590.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. 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)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Gaopeng Gou, and Qi Wu. 2025. Missing target-relevant information prediction with world model for accurate zero-shot composed image retrieval. *Preprint*, arXiv:2503.17109.

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. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. *Preprint*, arXiv:2212.10560.

Zhen Wang, Da Li, Yulin Su, Min Yang, Minghui Qiu, and Walton Wang. 2024b. Fashionlogo: Prompting multimodal large language models for fashion logo embeddings. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4113–4117. ACM.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. 2024. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *Preprint*, arXiv:2410.13848.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. 2024. Towards visual grounding: A survey. *Preprint*, arXiv:2412.20206.

Sijia Yao, Pengcheng Huang, Zhenghao Liu, Yu Gu, Yukun Yan, Shi Yu, and Ge Yu. 2025. Expandr: Teaching dense retrievers beyond queries with llm guidance. *Preprint*, arXiv:2502.17057.

Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, and Yansong Tang. 2025. Voco-llama: Towards vision compression with large language models. *Preprint*, arXiv:2406.12275.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *Preprint*, arXiv:2205.01917.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025. Gme: Improving universal multimodal retrieval by multimodal llms. *Preprint*, arXiv:2412.16855.

Instruction
You are an advanced AI assistant trained to analyze images and generate meaningful questions that capture their most important information. Analyze the given image and generate 3-5 specific questions that capture its most important visual information for retrieval purposes.
Each question should:
1. Focus on distinct key elements (objects, actions, settings).
2. Be clear and answerable from visual content alone.
3. Avoid subjective interpretations.
Consider, but not limited to the following questions:
1. Main objects and their attributes (type, color, position)
2. The Scene context (time/weather if apparent, location)
3. Visible text/logos
4. Notable relationships between elements
[Output]:

Figure 6: Prompt for Data Generation

Table 4: **Detailed MMEB-V1 Results.** Performance of baselines and CoMa variants across 20 in-distribution (IND) and 16 out-of-distribution (OOD) datasets. OOD datasets are highlighted with a yellow background.

	CLIP	VLM2Vec	MMRet	UniME	mmE5	MoCa-7B	CoMa -3B	CoMa -7B
Classification (10 tasks)								
ImageNet-1K	55.8	74.5	58.8	71.3	77.8	78.0	79.2	82.0
N24News	34.7	80.3	71.3	79.5	81.7	81.5	79.0	79.6
HatefulMemes	51.1	67.9	53.7	64.6	64.2	77.6	68.5	70.9
VOC2007	50.7	91.5	85.0	90.4	91.0	90.0	80.4	83.1
SUN397	43.4	75.8	70.0	75.9	77.7	76.8	70.8	76.7
Place365	28.5	44.0	43.0	45.6	43	43.0	37.8	46.3
ImageNet-A	25.5	43.6	36.1	45.5	56.3	52.7	43.8	54.8
ImageNet-R	75.6	79.8	71.6	78.4	86.3	83.0	81.0	85.2
ObjectNet	43.4	39.6	55.8	36.4	62.5	45.2	56.2	67.3
Country-211	19.2	14.7	14.7	18.7	35.4	30.4	21.2	28.4
All Classification	42.8	61.2	56.0	60.6	67.6	65.8	61.8	67.4
VQA (10 tasks)								
OK-VQA	7.5	69.0	73.3	68.3	67.6	36.9	64.7	71.4
A-OKVQA	3.8	54.4	56.7	58.7	56.1	57.1	54.6	62.5
DocVQA	4.0	52.0	78.5	67.6	90.3	94.3	94.0	95.9
InfographicsVQA	4.6	30.7	39.3	37.0	56.5	77.2	74.2	80.2
ChartQA	1.4	34.8	41.7	33.4	50.5	69.8	67.3	75.0
Visual7W	4.0	49.8	49.5	51.7	51.9	58.5	54.1	58.1
ScienceQA	9.4	42.1	45.2	40.5	55.8	59.2	46.6	57.5
VizWiz	8.2	43.0	51.7	42.7	52.8	46.2	51.4	54.7
GQA	41.3	61.2	59.0	63.6	61.7	71.6	56.0	65.2
TextVQA	7.0	62.0	79.0	65.2	83.3	75.8	78.7	85.4
All VQA	9.1	49.9	57.4	52.9	62.6	64.7	64.2	70.6
Retrieval (12 tasks)								
VisDial	30.7	80.9	83.0	79.7	74.1	84.5	81.0	82.0
CIRR	12.6	49.9	61.4	52.2	54.7	53.4	58.7	60.8
VisualNews_t2i	78.9	75.4	74.2	74.8	77.6	78.2	74.5	77.8
VisualNews_i2t	79.6	80.0	78.1	78.8	83.3	83.1	77.5	79.3
MSCOCO_t2i	59.5	75.7	78.6	74.9	76.4	79.8	73.6	77.0
MSCOCO_i2t	57.7	73.1	72.4	73.8	73.2	73.9	72.7	75.1
NIGHTS	60.4	65.5	68.3	66.2	68.3	66.7	65.6	67.6
WebQA	67.5	87.6	90.2	89.8	88.0	91.4	88.8	90.3
FashionIQ	11.4	16.2	54.9	16.5	28.8	28.9	21.5	26.4
Wiki-SS-NQ	55.0	60.2	24.9	66.6	65.8	82.7	66.4	64.1
OVEN	41.1	56.5	87.5	55.7	77.5	80.4	70.0	77.3
EDIS	81.0	87.8	65.6	86.2	83.7	96.9	86.0	91.0
All Retrieval	53.0	67.4	69.9	67.9	71.0	75.0	69.7	72.4
Visual Grounding (4 tasks)								
MSCOCO	33.8	80.6	76.8	76.5	53.7	84.6	69.4	73.2
RefCOCO	56.9	88.7	89.8	89.3	92.7	94.0	90.0	94.8
RefCOCO-matching	61.3	84.0	90.6	90.6	88.8	95.5	92.6	93.5
Visual7W-pointing	55.1	90.9	77.0	84.1	92.3	95.3	85.0	88.9
All Visual Grounding	51.8	86.1	83.6	85.1	89.6	92.4	84.3	87.6
Final Score (36 tasks)								
All	37.8	62.9	64.1	66.6	69.8	71.5	67.5	72.2
All IND	37.1	67.5	59.1	68.4	72.3	74.7	71.6	75.2
All OOD	38.7	57.1	68.0	57.9	66.7	67.6	61.5	67.6