UniMoCo: Unified Modality Completion for Robust Multi-Modal Embeddings

Anonymous ACL submission

Abstract

Current vision-language models have been explored for multi-modal embedding tasks like information retrieval. However, they face significant challenges in real-world queries and targets involving diverse modality combinations, as existing approaches often fail to align all modality combinations within a unified embedding space during training, leading to degraded performance on rare modality patterns during inference. To address this fundamental limitation, we propose UniMoCo, a novel architecture featuring a modality-completion module that generates visual features from text, thereby ensuring modality completeness for both queries and targets. Additionally, UniMoCo incorporates a specialized training strategy that aligns embeddings from both original and modality-completed inputs, thus ensuring consistent and robust embeddings for diverse modality combinations. Comprehensive experiments demonstrate that UniMoCo outperforms previous methods while exhibiting consistent robustness across diverse settings. Furthermore, we identify and quantify the inherent bias in conventional approaches caused by imbalanced modality combinations in training data, showing that our modality-completion paradigm effectively mitigates this limitation.

1 Introduction

Multi-modal embedding methods encode inputs with different modalities (such as text and image) into representations in an unified high-dimensional vector space, facilitating downstream tasks such as image classification (Deng et al., 2009), information retrieval (Datta et al., 2008; Gordo et al., 2016), retrieval augmented generation (Zhao et al., 2023), visual-language alignment (Plummer et al., 2015; Lin et al., 2014), etc. Previous models such as CLIP (Radford et al., 2021), BLIP (Li et al., 2022), SigLIP (Zhai et al., 2023) and ALIGN (Jia et al., 2021) aim to learn unified multi-modal representations by aligning visual and textual modalities through large-scale pretraining on paired image-text data, enabling cross-modal understanding and multi-modal embedding task applications. However, these models usually adopt the dual-encoder architecture with shallow or even no fusion of the visual and textual features, making fine-grained cross-modal reasoning (e.g., spatial relationships or detailed text-image interactions) less effective, limiting their application in complicated multi-modal embedding scenarios.

Recently with the rapid advancement of large vision language models (LVLMs) (Zhu et al., 2023; Liu et al., 2023; Chen et al., 2024; Wang et al., 2024), the extraordinary visual-textual understanding and reasoning capabilities of LVLMs have been unleashed for multi-modal representation learning and embedding tasks adaption. VLM2VEC (Jiang et al., 2024) introduced massive multimodal embedding benchmark (MMEB), a comprehensive benchmark for multi-modal embedding tasks covering classification, retrieval, vision question answering (VQA) and visual grounding. Other works propose specific training strategies (Liu et al., 2024; Gu et al., 2025; Lan et al., 2025; Zhou et al., 2024b) or data augmentation techniques (Lin et al., 2024; Zhou et al., 2024a) to train LVLMs for embedding adaption. Despite these advancements, adapting LVLMs for embedding tasks reveals limited performance in real-world applications, where queries and targets often feature diverse and incomplete modality combinations. The challenge of missing text is effectively addressed through the use of task-specific instructions, a method supported by prior work (Jiang et al., 2024), and by the near-ubiquity of textual system prompts in practice. Therefore, the absence of the visual modality emerges as the more frequent and critical challenge.

Consequently, current methods struggle to align all modality combinations into a semantically coherent and unifiedembedding space during training. This misalignment, largely due to imbalanced com-

binations in training data, degrades performance when the model encounters underrepresented combinations during inference.

In this work, we propose **UniMoCo**, a model architecture with Unified Modality Completion designed to learn robust multi-modal embeddings. It consists of two main components: a backbone LVLM and a modality-completion module. During training and inference, the modality-completion module synthesizes visual embeddings directly from the input text whenever the visual modality is absent, which ensures that multi-modal representations remain complete even under modality absence. Moreover, we design a complementary training strategy that integrates contrastive loss to enhance representation quality by bringing matching query-target pairs closer while pushing non-matching ones apart, and employs auxiliary loss to maintain consistency between pseudo and real visual embeddings, thereby improving the reliability of synthesized embeddings when visual inputs are missing. Together, the UniMoCo architecture and its tailored training approach enable effective handling of diverse modality combinations in the embedding space, boosting the model's adaptability and robustness in real multi-modal scenarios. Our contributions can be summarized as follows:

- We propose UniMoCo, a LVLM-based framework augmented with a modality-completion module to produce robust multi-modal embeddings for diverse embedding tasks.
- We develop an effective training strategy combining contrastive learning with auxiliary losses to maximize UniMoCo's potential.
- We evaluate our method on MMEB, which shows existing methods are prone to modality combination bias while UniMoCo not only mitigates this issue but also delivers superior performance across multiple tasks.

2 Related Work

2.1 Text Representation Learning

Text embeddings are extensively utilized across diverse natural language processing tasks, including text classification, retrieval, and QA. Current approaches to learning text representation can be broadly categorized into task-specific and general-purpose paradigms. Early research primarily focused on developing specialized architectures for distinct applications: works (Huang et al., 2019; Karpukhin et al., 2020; Hao et al., 2019) targeted QA systems, while (Wang et al., 2018; Sinoara et al., 2019) addressed classification tasks, and (Huang et al., 2020; Cakaloglu et al., 2020; Zhan et al., 2020) specialized in retrieval scenarios. Recent advancements have shifted toward developing general-purpose embedding models with broader applicability. Multiple studies (Giorgi et al., 2020; Wang et al., 2022; Li et al., 2023b; BehnamGhader et al., 2024) have successfully employed contrastive learning frameworks for this objective. Concurrently, innovative approaches (Su et al., 2022; Asai et al., 2022) incorporate task-specific instructions alongside input text during encoding, enabling unified handling of multiple downstream tasks. Recent research has extended the application scope of decoder-only large language models (LLMs) traditionally used for generation, with several studies successfully utilized them as embedding models (Wang et al., 2023; Li et al., 2023b; BehnamGhader et al., 2024; Lee et al., 2024) with promising results.

2.2 Multi-modal Representation Learning

Unlike text embedding, multi-modal representations enable broader applicability across diverse tasks (Kim et al., 2020; Khare et al., 2021; Ge et al., 2021; Lin et al., 2022; Xia et al., 2023; Li et al., 2024; Zhou et al., 2024b), yet their learning poses greater challenges due to the complexity of aligning different modalities. Prior approaches predominantly leverage encoder-based architectures such as CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) to project different modalities into a unified space to align multi-modal inputs. Recent advances like VLMVec (Jiang et al., 2024) introduce LVLMs as backbones for a more generalized framework. Subsequent advancements focus on refining contrastive learning objectives (Lan et al., 2025; Yu et al., 2025), enhancing data quality via synthetic datasets (Zhou et al., 2024a; Chen et al., 2025), or optimizing training strategies (Liu et al., 2024; Gu et al., 2025; Zhou et al., 2024b). However, these efforts primarily target training methodologies rather than architectural innovations and overlook the critical limitation of incomplete modality combinations, which our work systematically addresses through novel structural improvements.

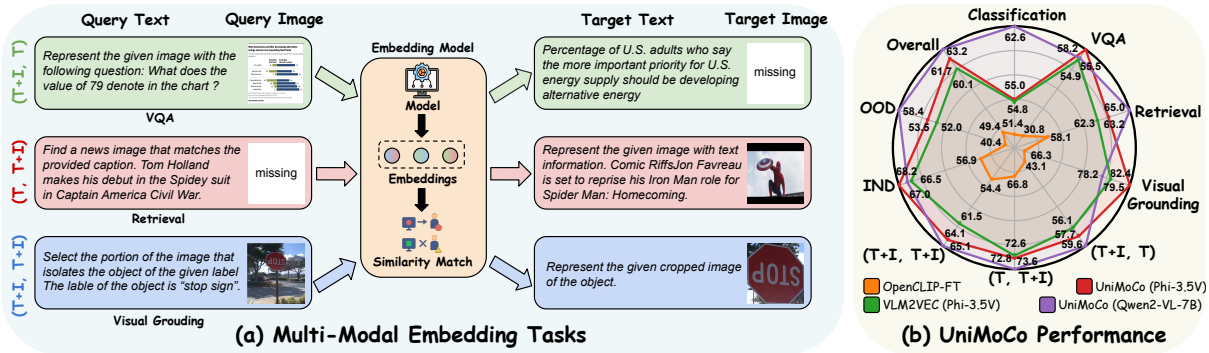

Figure 1: (a) Multi-modal embedding tasks involve three common modality combinations, illustrated with examples from the MMEB (Jiang et al., 2024): $(\mathrm{T}+\mathrm{I}, \mathrm{T})$, $(\mathrm{T}, \mathrm{T}+\mathrm{I})$, and $(\mathrm{T}+\mathrm{I}, \mathrm{T}+\mathrm{I})$. Specifically, $(\mathrm{T}+\mathrm{I}, \mathrm{T})$ denotes a query containing both text and image, while the target is text-only; the other combinations follow similarly. A multi-modal embedding model encodes both query and target into a unified space and then conduct tasks through similarity matching. (b) UniMoCo's performance vs. other embedding models on MMEB.

2.3 Modality Missing

Real-world multi-modal applications often suffer from missing modalities during training or inference, which can significantly degrade performance (Wu et al., 2024). Early dual-encoder models like CLIP, BLIP, and FLAVA (Radford et al., 2021; Li et al., 2022; Singh et al., 2022) tackle this by learning a shared modality-invariant space. However, their reliance on complete pre-training data introduces biases toward dominant modality combinations (Ma et al., 2022, 2021).

Generative approaches synthesize proxy modalities to reconstruct missing inputs (Zhou et al., 2024a; Chen et al., 2025; Feng et al., 2024), yet the quality of the approaches is set by the off-the-shelf frozen generators. Specialized expert-based methods like Flex-MoE (Yun et al., 2024) use dynamic routing, while transformer-based approaches (Jiang et al., 2024; Lee et al., 2023; Yang et al., 2024) leverage prompt or adapter tuning for flexible inputs. However, fixed dropout strategies of these methods often fail to generalize to rare modality combinations. In contrast, we propose a lightweight modality-completion module integrated with a unified LVLM backbone, synthesizing missing visual embeddings to ensure consistent alignment across all modality combinations.

3 Methodology

3.1 Problem Definition

In this work, we propose UniMoCo, a unified multi-modal embedding model that projects both the query and candidate targets into a shared high-dimensional embedding space. Our model accepts both textual and optional visual inputs, encoding them into compact and expressive embeddings $E \in \mathbb{R}^d$, where d represents the embedding dimension. These unified representations are designed to capture discriminative features. Within this space, similarity matching is performed to identify the candidate target embedding closest to the query embedding, making it the most suitable match.

The input data (for both query and target) comprises two elements: (1) text, which includes a task-defining **instruction** (e.g., "*Find an image matching the fashion image and style note.*") and specific **content** (e.g., "*Shiny silver material with short sleeves and a fit-and-flare silhouette.*"); (2) optional images. They are processed simultaneously to generate embeddings.

Given a query q and n candidate targets $\{c_1, c_2, \ldots, c_n\}$, we compute their respective embeddings $E_q = f(q)$ and $E_{c_i} = f(c_i)$ for $i = 1, \ldots, n$, where $f(\cdot)$ represents the embedding model. The optimal match c^* is determined by selecting the candidate with the highest similarity:

$$c^* = \arg\max_{c_i} \mathrm{sim}(E_q, E_{c_i}),$$

where $\mathrm{sim}(E_q, E_{c_i})$ is typically implemented as temperature-scaled cosine similarity function.

3.2 UniMoCo Architecture

Figure 2 presents our UniMoCo architecture, which utilizes an LVLM as its backbone with three components: LLM, vision encoder, and projector. To ensure robust handling of diverse input modality combinations, we incorporate task-specific instructions into every input. When the textual content is

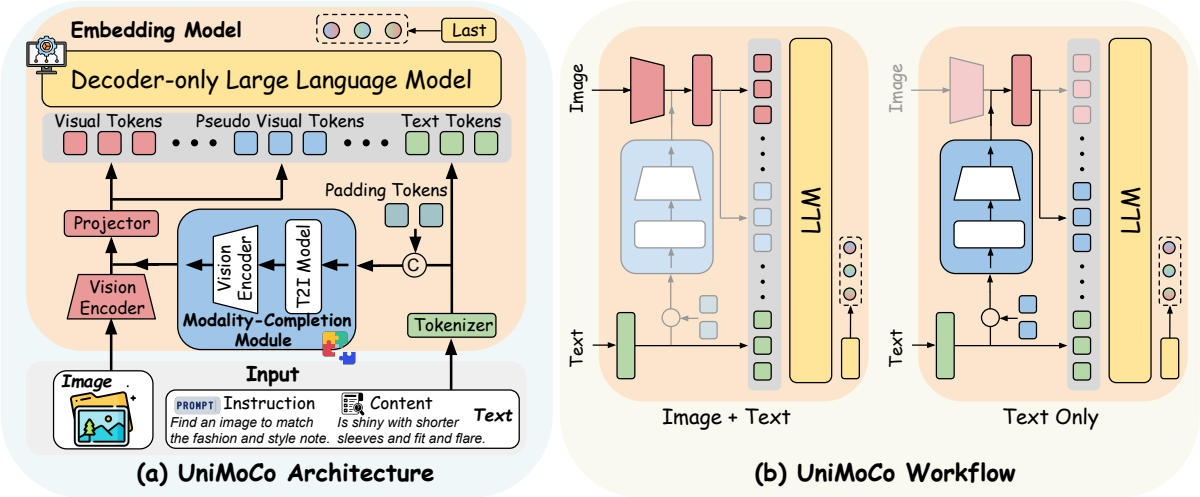

Figure 2: (a) UniMoco architecture. Processes image/text inputs through an LLM, with the final output token as the unified embedding. (b) UniMoCo workflow illustration. The left panel shows image-text processing while the right panel shows text-only input processing. Grayed-out icons indicate inactive modules in each scenario. This unified workflow supports both training and inference phases.

missing, the instruction itself serves as the available linguistic input, thereby compensating for the absence. For cases lacking visual input, we introduce a dedicated text-to-image (T2I) module that utilizes a compact language model to directly convert text into pseudo visual embeddings. This approach avoids the computational overhead and embedding-task misalignment associated with conventional diffusion-based T2I methods (Zhou et al., 2024a; Chen et al., 2025; Feng et al., 2024), while effectively maintaining cross-modal alignment and functional coherence across all input conditions without redundant processing.

The modality-completion module is further enhanced through the addition of a supplementary vision encoder. This architectural decision stems from our observation that embeddings produced by the T2I model exhibit incomplete consistency with those generated by the original vision encoder processing real images. The additional encoder serves to better capture and represent the characteristics of our pseudo visual embeddings. Furthermore, when processing text tokens within this module, we concatenate them with padding tokens to maintain a fixed input length that matches the number of visual tokens produced by the primary vision encoder from authentic images.[1] A comprehensive analysis of this module's components and their contributions is provided in Section 4.4.

Figure 2 details the operational workflow for various input scenarios. When presented with complete multi-modal inputs (Case 1), the system bypasses the completion module entirely, functioning as a conventional LVLM. In situations where image data is unavailable (Case 2), the textual input is simultaneously processed by both the LLM and our completion module, with the latter generating pseudo visual tokens to substitute for the missing image representations. This dual-path approach ensures robust performance across different inputs while maintaining model consistency.

3.3 UniMoCo Training Strategy

As illustrated in Figure 3, our framework employs two complementary loss functions. In conventional multi-modal embedding frameworks (Jiang et al., 2024), contrastive learning aims to learn discriminative representations by minimizing distance between query-target positive pairs while maximizing separation from negatives. For a batch of B query-target pairs, the InfoNCE loss is defined as:

$$\mathcal{L}_1 = -\frac{1}{B}\sum_{i=1}^{B}\log\left(\frac{\exp(\mathrm{sim}(E_q^{(i)}, E_{c^+}^{(i)})/\tau}{\sum_{j=1}^{B}\exp(\mathrm{sim}(E_q^{(i)}, E_{c_j}^{(i)})/\tau)}\right), \tag{1}$$

where $E_q^{(i)}, E_{c^+}^{(i)} \in \mathbb{R}^d$ denote the i-th query and positive target embeddings respectively, B represents batch size, and $\tau > 0$ is the temperature.

However, modality-complete and modality-missing inputs produce fundamentally misaligned embeddings in the latent space, so standard contrastive loss fails to reconcile these divergent rep-

[1]For details of the padding token design, see Appendix C.1.

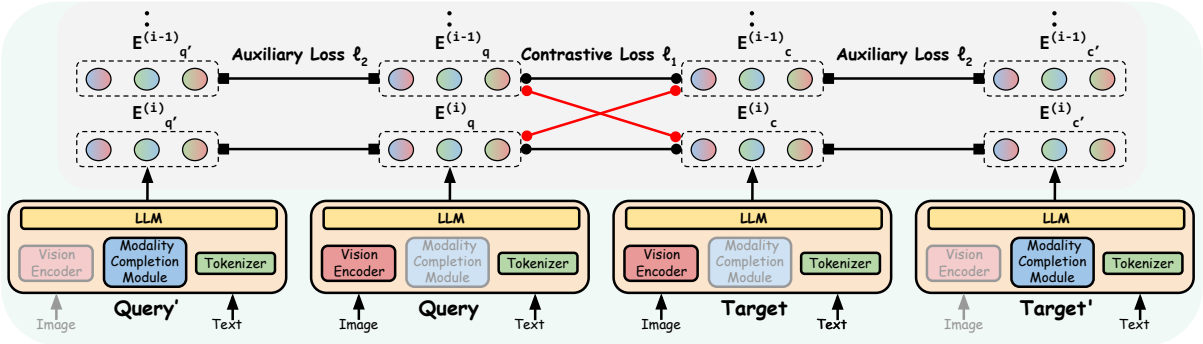

Figure 3: UniMoCo training combines a primary contrastive loss ($\mathcal{L}_1$) with auxiliary loss ($\mathcal{L}_2$). Black lines indicate positive pairs to be pulled closer in the embedding space, while red lines denote negative pairs requiring separation.

resentations, preventing LVLMs from learning a unified embedding space. To address the gap between two types of inputs and enable consistent cross-modal projection, we introduce an auxiliary loss tailored for multi-modal inputs. Given input q (or c_i) containing both image and text modalities, we construct q' (or $c_i^{'}$) by removing the image component and generating pseudo visual tokens through our completion module.

$$\mathcal{L}_2 = \frac{1}{B}\sum_{i=1}^{B}\left[\mathcal{H}(E_c^{(i)}, E_{c'}^{(i)}) + \mathcal{H}(E_q^{(i)}, E_{q'}^{(i)})\right], \tag{2}$$

where $\mathcal{H}(\cdot,\cdot)$ denotes the cross-entropy function. This proposed loss objective naturally accommodates uni-modal cases: when the input q lacks visual content, q' becomes identical to q, nullifying their cross-entropy contribution.

The composite loss function combines these objectives through linear combination:

$$\mathcal{L} = \mathcal{L}_1 + \alpha\mathcal{L}_2, \tag{3}$$

where $\alpha \in \mathbb{R}^+$ controls the relative importance of cross-modal alignment. Our framework jointly optimizes two objectives: discriminative embedding learning $\mathcal{L}_1$ and modality-invariant representation learning via $\mathcal{L}_2$. For complete modalities, $\mathcal{L}_1$ reduces distances between positive pairs while separating negative pairs. When modalities are missing, $\mathcal{L}_2$ bridges pseudo image embeddings with real counterparts. The orthogonal combination of them constructs a unified embedding space that preserves discriminative power across modalities and robustness to missing-modality data.

4 Evaluation

4.1 Setup

In our study, we employ Phi-3.5V (Abdin et al., 2024) and Qwen2-VL-7B (Wang et al., 2024) as foundational LVLMs, while utilizing Phi-1.5 and Qwen2-1.5B as their T2I counterparts. The loss function incorporates a temperature parameter of 0.02 and a hyperparameter α of 0.1. For Phi-3.5V based models, we process each image into 4 sub-image crops, whereas for Qwen2-VL-7B based models, all images are resized to a standardized resolution of 672×672. For details of experimental settings, please refer to Appendix A.

For training data, we utilize the MMEB (Jiang et al., 2024), comprising 20 diverse datasets across four domains: classification, retrieval, VQA, and visual grounding. By subsampling datasets exceeding 50K instances into 50K samples, we construct a final training set of 662K samples.

The evaluation framework also builds upon MMEB, incorporating both in-distribution (20 datasets) and out-of-distribution (16 datasets) test sets, with each dataset containing 1K samples. We employ Precision@1 as the primary metric for assessing model performance across all benchmarks.

For evaluation baselines, we employ several multi-modal embedding models, including CLIP (Radford et al., 2021), OpenCLIP (Cherti et al., 2023), BLIP2 (Li et al., 2023a), and SigLIP (Zhai et al., 2023), all of which utilize vision or language encoders to generate feature representations. Additionally, we incorporate UniIR (Wei et al., 2024) and VLM2VEC (Jiang et al., 2024) [2], two models specifically designed for multi-modal embedding tasks. It should be noted

[2]Here we only include Phi-3.5V variant of VLM2VEC. The rationale for this selection can be found in Appendix B.

that our comparison focuses solely on architectural differences, excluding methods that optimize embeddings from other perspectives (Lan et al., 2025; Yu et al., 2025; Zhou et al., 2024a; Chen et al., 2025; Liu et al., 2024; Gu et al., 2025). These approaches are orthogonal to our work and could complement our architecture in the future.

4.2 Main Results

The results presented in Table 1 demonstrate that our proposed UniMoCo framework outperforms all baselines on the MMEB. Detailed results appear in Appendix B. Notably, the Qwen2-VL variant achieves the best overall performance with a score of 63.2, comprising 67.0 on in-distribution datasets and 58.4 on out-of-distribution datasets. When examining performance across different modality combinations, this variant attains scores of 59.6, 73.6, and 65.1 for $(\mathrm{T}+\mathrm{I}, \mathrm{T})$, $(\mathrm{T}, \mathrm{T}+\mathrm{I})$, and $(\mathrm{T}+\mathrm{I}, \mathrm{T}+\mathrm{I})$ tasks respectively.

Our analysis reveals significant improvements over existing approaches. Compared to the strongest baseline without fine-tuning, UniMoCo shows substantial gains of 12.8, 43.2, 4.9, and 20.2 points on classification, VQA, retrieval, and grounding meta-tasks, respectively. Compared to fine-tuned baselines, our method maintains consistent improvements of 7.8, 3.3, 2.7, and 2.9 points.

We observe that $(\mathrm{T}+\mathrm{I}, \mathrm{T})$ tasks dominates the majority of the datasets in the MMEB training set (13 out of 20). A similar trend emerges when comparing Phi-3.5V based VLMVEC and our UniMoCo: while VLMVEC performs comparably to UniMoCo on $(\mathrm{T}+\mathrm{I}, \mathrm{T})$ tasks, our method significantly outperforms it on other modality combinations. This reveals that the failure of traditional architectures to properly align all modality completions results in a bias toward frequent training combinations. In comparison, UniMoCo overcomes this challenge by unifying all modalities within one aligned architecture, resulting in uniformly better performance across all combinations. Additional analysis can be found in Section 4.3.

4.3 Modality Combination Bias Analysis

This study investigates whether traditional model architectures exhibit inherent biases toward specific modality combinations and evaluates the potential of our proposed UniMoCo framework in addressing this limitation. To ensure a fair comparison, we conduct extensive experiments using both VLM2VEC and our UniMoCo approach, with Phi-3.5V serving as the backbone LVLM. For our experimental setup, we construct a specialized training set from MMEB's retrieval-related classes, comprising three distinct subsets corresponding to different modality combinations. To control data distribution, we create three training set variants, each with half the samples from one modality combination and the rest equally split between the other two. Models trained on these imbalanced sets are evaluated on MMEB, thus enabling rigorous examination of modality preferences.

As demonstrated in Figure 4, traditional approaches are indeed susceptible to imbalanced modality combinations in the training data. VLM2VEC performs strongly when evaluated on tasks matching its dominant training modality combination. For instance, when trained on $(\mathrm{T}, \mathrm{T}+\mathrm{I})$ dominated datasets, VLM2VEC achieves a score of 62.9, approaching UniMoCo's performance of 63.4. However, its performance significantly deteriorates on other combinations (42.8 vs. 45.0 and 20.1 vs. 30.9 for UniMoCo). Furthermore, VLM2VEC trained on other modality distributions show markedly reduced performance on $(\mathrm{T}, \mathrm{T}+\mathrm{I})$ evaluation tasks, scoring only 60.5 and 61.0, respectively. This limitation remains consistent across all evaluations examined.

In contrast, our UniMoCo framework demonstrates remarkable robustness, maintaining consistently high performance across all evaluation benchmarks regardless of the dominant modality combination in the training data. These results clearly indicate that UniMoCo effectively addresses the critical limitation of modality bias.

4.4 Ablation Study

4.4.1 Modality-Completion Module

To thoroughly examine the design choices for our modality-completion module, we compare various architectural variants including configurations with and without padding tokens as well as additional vision encoders as proposed in Section 3.2.

Table 2 reveals that simply using a single T2I model without proper modifications leads to significant performance degradation. This occurs because the T2I model produces output embeddings with lengths corresponding to the input query text. In typical scenarios where query texts rarely exceed 40 tokens, this produces short pseudo visual embeddings that must be matched against real image embeddings containing at least 576 patch tokens. Such

Table 1: Evaluation results on the MMEB, displaying average meta-task scores. Baseline comparisons include both fine-tuned (FT) and non-FT variants on training data, alongside our Phi-3.5V and Qwen2-VL-7B variants. The notation $(\mathrm{T}+\mathrm{I}, \mathrm{T})$ stands for datasets containing multi-modal queries with text targets,. IND/OOD distinguishes between in-distribution and out-of-distribution datasets. The optimal results highlighted in **bold** and the strongest baseline performances (both FT and non-FT variants) are indicated with underlines.

Models	Per Meta-Task Score				Average Score					
	Classification	VQA	Retrieval	Grounding	$(\mathrm{T}+\mathrm{I}, \mathrm{T})$	$(\mathrm{T}, \mathrm{T}+\mathrm{I})$	$(\mathrm{T}+\mathrm{I}, \mathrm{T}+\mathrm{I})$	IND	OOD	Overall
# of Datasets →	10	10	12	4	22	6	8	20	16	36
w/o Fine-tuning on MMEB										
CLIP (Radford et al., 2021)	42.8	9.1	53.0	51.8	29.8	62.1	41.6	37.1	38.7	37.8
OpenCLIP (Cherti et al., 2023)	47.8	10.9	52.3	53.3	33.1	57.9	44.2	39.3	40.2	39.7
BLIP2 (Li et al., 2023a)	27.0	4.2	33.9	47.0	15.7	43.1	37.9	25.3	25.1	25.2
SigLIP (Zhai et al., 2023)	40.3	8.4	31.6	59.5	27.0	44.6	49.0	32.3	38.0	34.8
UniIR (Wei et al., 2024)	42.1	15.0	60.1	62.2	32.5	58.2	59.7	44.7	40.4	42.8
Δ w/o fine-tune	↑12.8	↑43.2	↑4.9	↑20.2	↑26.5	↑15.4	↑5.4	↑23.5	↑18.0	↑20.4
w/ Fine-tuning on MMEB										
CLIP-FT	50.0	27.0	55.3	64.8	40.3	64.7	49.3	52.2	38.9	47.0
OpenCLIP-FT	51.4	30.8	58.1	66.3	43.1	66.8	54.4	56.9	40.4	49.6
VLM2VEC (Phi-3.5V) (Jiang et al., 2024)	54.8	54.9	62.3	79.5	56.1	72.6	61.5	66.5	52.0	60.1
UniMoCo (Phi-3.5V)	55.0	**58.2**	63.2	**82.4**	57.7	72.8	64.1	**68.2**	53.5	61.7
UniMoCo (Qwen2-VL-7B)	**62.6**	55.5	**65.0**	78.2	**59.6**	**73.6**	**65.1**	67.0	**58.4**	**63.2**
Δ w/ fine-tune	↑7.8	↑3.3	↑2.7	↑2.9	↑3.5	↑1.0	↑3.6	↑1.7	↑6.4	↑3.1

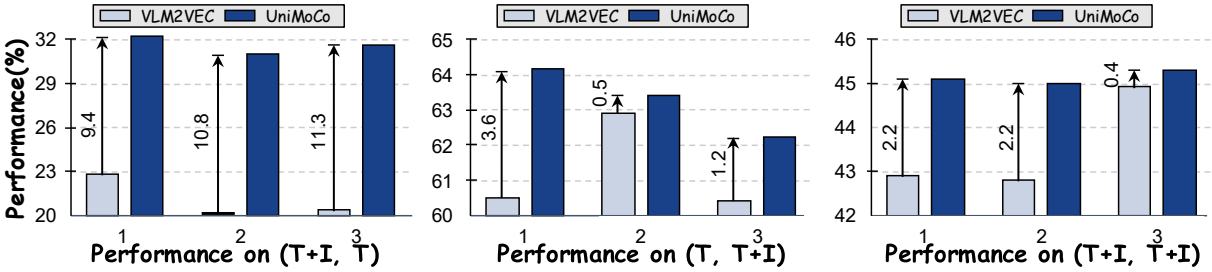

Figure 4: Performance of models trained on skewed training data distributions. The x-axis labels (1, 2, 3) correspond to training sets dominated by $(\mathrm{T}+\mathrm{I}, \mathrm{T})$, $(\mathrm{T}, \mathrm{T}+\mathrm{I})$, and $(\mathrm{T}+\mathrm{I}, \mathrm{T}+\mathrm{I})$ combinations respectively.

Table 2: Evaluation results on MMEB across different architectural designs. Baseline represents VLM2VEC; other configurations show UniMoCo variants.

Models	Average Score					
	$(\mathrm{T}+\mathrm{I}, \mathrm{T}+\mathrm{I})$	$(\mathrm{T}+\mathrm{I}, \mathrm{T})$	$(\mathrm{T}, \mathrm{T}+\mathrm{I})$	IND	OOD	Overall
# of Datasets →	22	6	8	20	16	36
Baseline	52.9	68.4	60.0	62.4	50.1	56.9
T2I Only	49.6	64.7	57.2	58.9	48.4	53.6
w/ Encoder	51.9	66.5	57.6	61.2	48.1	55.4
w/ Padding	53.3	69.5	**60.5**	63.5	49.8	57.4
w/ Encoder + Padding	**54.0**	**70.1**	60.5	**64.3**	50.1	**58.0**

Table 3: Performance evaluation across varying T2I model scales on MMEB, demonstrating how model scale impacts multi-modal representation learning.

T2I Model Size	Average Score					
	$(\mathrm{T}+\mathrm{I}, \mathrm{T})$	$(\mathrm{T}, \mathrm{T}+\mathrm{I})$	$(\mathrm{T}+\mathrm{I}, \mathrm{T}+\mathrm{I})$	IND	OOD	Overall
# of Datasets →	22	6	8	20	16	36
0.5B	47.5	68.4	53.4	58.9	44.0	52.3
1.5B	54.0	70.1	60.5	64.3	50.1	58.0
7B	**56.3**	**73.3**	**62.2**	**67.2**	**52.0**	**60.4**

significant length discrepancy creates challenges for computing meaningful similarity between these representations. Our padding strategy using optimized prompts and dummy tokens to align the T2I model's input length with real image token counts can generate more representative pseudo visual embeddings. As shown in Table 2, this approach consistently enhances performance across all modality combinations by better leveraging the T2I model's latent capabilities.

Since pseudo visual embeddings still differ from their real counterparts, employing a dedicated vision encoder to process these embeddings yields additional gains (53.6 to 55.4). However, this gain is outweighed by padding (55.4 vs. 57.4) due to the persistent issue of dimensional mismatch within the T2I processing pipeline.

Furthermore, the synergistic combination of both techniques yields superior improvements over the baseline, particularly for the $(\mathrm{T}+\mathrm{I}, \mathrm{T}+\mathrm{I})$ and $(\mathrm{T}, \mathrm{T}+\mathrm{I})$ tasks that heavily depend on robust modality completion. These results validate our approach's capability to effectively align diverse modality combinations with robust performance.

Table 4: Performance evaluation showing the impact of α, which balances contrastive loss $\mathcal{L}_1$ and auxiliary loss $\mathcal{L}_2$ in our objective function.

Configurations	Average Score					
	$(\mathrm{T+I, T})$	$(\mathrm{T, T+I})$	$(\mathrm{T+I, T+I})$	IND	OOD	Overall
# of Datasets →	22	6	8	20	16	36
$\alpha = 0.0$	54.0	70.1	60.5	64.4	50.0	58.0
$\alpha = 0.1$	54.6	71.1	60.5	64.5	51.1	58.5
$\alpha = 0.2$	**55.3**	**71.8**	**61.1**	**64.8**	**52.1**	**59.2**
$\alpha = 0.3$	55.2	71.0	60.6	65.0	51.3	58.9
$\alpha = 0.4$	54.5	71.6	60.5	64.7	50.8	58.5

4.4.2 T2I Model

Table 3 investigates how model scale affects multi-modal embedding performance using Qwen2-VL-7B as the LVLM backbone with three language model variants (Qwen2-0.5B, Qwen2-1.5B, Qwen2-7B). Our experiments demonstrate a clear relationship between model size and task performance, showing that larger T2I models consistently improve UniMoCo's embedding quality in all benchmarks and modality combinations. This phenomenon aligns with established scaling laws (Jia et al., 2021), as increased model capacity improves domain adaptation and semantic representation capabilities critical for generating high-fidelity pseudo visual embeddings from textual inputs. The 7B variant demonstrates significantly improved performance by achieving greater accuracy in text-to-embedding translation, especially in tasks that demand precise latent space mapping. These findings underscore the importance of model scale in multi-modal systems, demonstrating that parameter expansion in specialized components can significantly boost effectiveness.

4.4.3 Training Strategy

As demonstrated in Table 4, the integration of auxiliary loss for multi-modal inputs significantly enhances model performance from all aspects. It simultaneously enhances downstream task performance, improves generalization on OOD, and strengthens robustness against diverse modality combinations. Our experiments show consistent performance gains when increasing the hyperparameter α from 0.0 to 0.2. We attribute it to the auxiliary loss's role in accelerating the convergence of the modality-completion module, thereby generating higher-quality pseudo visual embeddings that facilitate better query-target matching. However, further increasing α beyond 0.2 to 0.4 results in performance degradation. This suggests that excessive weighting of the auxiliary loss may shift the primary training objective away from the target of learning effective multi-modal embeddings. Through comprehensive evaluation, we identify $\alpha = 0.2$ as the optimal balance point that maintains the primary training objective while preserving enhanced transfer capabilities.

Notably, the most significant improvements occur in $(\mathrm{T+I, T})$ and $(\mathrm{T, T+I})$ datasets, which aligns with our design since these scenarios involve missing image modality where the auxiliary loss practically bridges the gap between generated pseudo visual embeddings and real visual embeddings. The observed improvement in $(\mathrm{T+I, T+I})$ configurations may stem from the auxiliary loss's secondary benefit of aligning embedding spaces between different vision encoders. This alignment enables language models to process inputs from more consistent embedding space, thereby improving their processing ability. The consistent performance gains across different modality combinations validate that appropriate auxiliary supervision can simultaneously enhance both modality completion and cross-modal alignment.

4.4.4 Efficiency Analysis

To comprehensively evaluate our framework, we quantify the overhead of Phi-3.5V based UniMoCo against the baseline (w/o the completion module and $\mathcal{L}_2$ loss) under the same backbone.

During training, UniMoCo requires approximately 35% more latency and consumes 20% higher peak memory. In inference, however, the overhead drops significantly to just 8% added latency and 10% memory increase. The cost disparity between training and inference arises because $\mathcal{L}_2$ loss generates more activations during training, consuming more memory and latency, whereas inference involves no loss computation and thus incurs less overhead. We argue that this moderate and acceptable increase in computational resources, particularly the minor overhead at inference, is a justifiable trade-off for UniMoCo's significant and consistent performance gains.

5 Conclusion

In this paper, we propose UniMoCo which incoraprates LVLM with modality completion effectively handles diverse modality combinations in embedding tasks. It outperforms existing methods and reduces bias caused by modality imbalance in training data. This approach ensures robust and consistent performance across various scenarios.

Limitation

Our study addresses modality combination bias but still leaves several areas for future work. We focus on structural innovations without fully exploring other approaches like training data enhancement or contrastive learning optimization. Combining these with our architecture could produce more robust multi-modal representations. Furthermore, our experiments primarily use the MMEB, but testing on diverse benchmarks would better validate generalizability. These potential extensions suggest a clear pathway for advancing UniMoCo's performance.

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.

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

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

Tolgahan Cakaloglu, Christian Szegedy, and Xiaowei Xu. 2020. Text embeddings for retrieval from a large knowledge base. In *International Conference on Research Challenges in Information Science*, pages 338–351. Springer.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829.

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):1–60.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Tiantian Feng, Daniel Yang, Digbalay Bose, and Shrikanth Narayanan. 2024. Can text-to-image model assist multimodal learning for visual recognition with visual modality missing? *arxiv preprint arXiv: 2402.09036*.

Xuri Ge, Fuhai Chen, Joemon M Jose, Zhilong Ji, Zhongqin Wu, and Xiao Liu. 2021. Structured multi-modal feature embedding and alignment for image-sentence retrieval. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5185–5193.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*.

Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. 2016. Deep image retrieval: Learning global representations for image search. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 241–257. Springer.

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. *arXiv preprint arXiv:2504.17432*.

Yu Hao, Xien Liu, Ji Wu, and Ping Lv. 2019. Exploiting sentence embedding for medical question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 938–945.

Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2553–2561.

Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge graph embedding based question answering. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 105–113.

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

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. 2021. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1033–1036. IEEE.

Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan Plummer. 2020. Mule: Multimodal universal language embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11254–11261.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. 2025. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. *arXiv preprint arXiv:2503.04812*.

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

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. 2023. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*.

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

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

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

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, and 1 others. 2024. Groundinggpt: Language enhanced multimodal grounding model. *arXiv preprint arXiv:2401.06071*.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2024. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multi-modal contrastive representation learning for entity alignment. *arXiv preprint arXiv:2209.00891*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yikun Liu, Pingan Chen, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. 2024. Lamra: Large multimodal model as your advanced retrieval assistant. *arXiv preprint arXiv:2412.01720*.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 18156–18165.

Mengmeng Ma, Jian Ren, Long Zhao, S. Tulyakov, Cathy Wu, and Xi Peng. 2021. SMIL: Multimodal learning with severely missing modality. *arxiv preprint arXiv: 2103.05677*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 15617–15629.

Roberta A Sinoara, Jose Camacho-Collados, Rafael G Rossi, Roberto Navigli, and Solange O Rezende. 2019. Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems*, 163:955–971.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. *arXiv preprint arXiv:1805.04174*.

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. 2024. Deep multimodal learning with missing modality: A survey. *arxiv preprint arXiv:2409.07825*.

Wei Xia, Tianxiu Wang, Quanxue Gao, Ming Yang, and Xinbo Gao. 2023. Graph embedding contrastive multi-modal representation learning for clustering. *IEEE Transactions on Image Processing*, 32:1170–1183.

Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. 2024. MMA: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 23826–23837.

Hao Yu, Zhuokai Zhao, Shen Yan, Lukasz Korycki, Jianyu Wang, Baosheng He, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, and Hanchao Yu. 2025. Cafe: Unifying representation and generation with contrastive-autoregressive finetuning. *arXiv preprint arXiv:2503.19900*.

Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. 2024. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. Repbert: Contextualized text embeddings for first-stage retrieval. *arXiv preprint arXiv:2006.15498*.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and 1 others. 2023. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*.

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

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024b. Vista: visualized text embedding for universal multi-modal retrieval. *arXiv preprint arXiv:2406.04292*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Details of Evaluation Settings

A.1 Evaluation Datasets

For the main experiments, UniMoCo is trained on 20 diverse datasets encompassing multiple tasks: classification (ImageNet-1K, N24News, HatefulMemes, VOC2007, SUN397), VQA (OK-VQA, A-OKVQA, DocVQA, InfographicsVQA, ChartQA, Visual7, ScienceQA), (VisDial, CIRR, VisualNews_i2t, VisualNews_t2i, MSCOCO_t2i, MSCOCO_i2t, NIGHTS, WebQA), and visual grounding (MSCOCO). Evaluation is conducted on 36 test datasets as specified in Table 6, with all datasets sourced from MMEB (Jiang et al., 2024).

For additional experiments, we maintain the same training configuration except in the modality combination bias analysis. In this specific evaluation, we select three representative base datasets: VisualNews_i2t $(\mathrm{T}+\mathrm{I}, \mathrm{T})$, VisDial $(\mathrm{T}, \mathrm{T}+\mathrm{I})$, and NIGHTS $(\mathrm{T}+\mathrm{I}, \mathrm{T}+\mathrm{I})$. We construct three distinct training sets by varying the sample distributions: (1) 10,000 samples from VisualNews_i2t with 5,000 each from VisDial and NIGHTS; (2) 10,000 from VisDial with 5,000 each from VisualNews_i2t and NIGHTS; (3) 10,000 from NIGHTS with 5,000 each from VisDial and VisualNews_i2t. Both our approach and the VLM2VEC baseline are trained on these configurations, with evaluation performed across all 36 MMEB datasets to produce the results discussed in Section 4.3.

Regarding modality combinations, the datasets can be categorized into three types based on their input-output configurations: The $(\mathrm{T}+\mathrm{I}, \mathrm{T})$ group includes ImageNet-1K, N24News, HatefulMemes, VOC2007, SUN397, Place365, ImageNet-A, ImageNet-R, ObjectNet, Country211, VisualNews_i2t, MSCOCO_i2t, OK-VQA, A-OKVQA, DocVQA, InfographicsVQA, ChartQA, Visual7, ScienceQA, VizWiz, GQA, TextVQA. The $(\mathrm{T}, \mathrm{T}+\mathrm{I})$ category comprises VisDial, VisualNews_t2i, MSCOCO_t2i, WebQA, Wiki-SS-NQ, EDIS. Lastly, the $(\mathrm{T}+\mathrm{I}, \mathrm{T}+\mathrm{I})$ combination contains CIRR, NIGHTS, FashionIQ, OVEN, MSCOCO, RefCOCO, RefCOCO-Matching, Visual7W-Pointing. Among the training datasets, 13 belong to $(\mathrm{T}+\mathrm{I}, \mathrm{T})$, 3 to $(\mathrm{T}, \mathrm{T}+\mathrm{I})$, and 4 to $(\mathrm{T}+\mathrm{I}, \mathrm{T}+\mathrm{I})$. The predominance of $(\mathrm{T}+\mathrm{I}, \mathrm{T})$ datasets in the training set introduces a bias, as discussed in Section 4.3.

A.2 Hyperparameters and Computational Requirements

Table 5 presents our detailed setting during training and test.

B Specific Results on MMEB

Table 6 provides comprehensive experimental results corresponding to the data presented in Table 1. It is worth noting that VLMVEC (Jiang et al., 2024) has also introduced model variants based on LLaVA-1.6 and Qwen2-VL, which were not included in our comparative analysis for the following reasons. Firstly, as documented in their materials, these models were trained using a different dataset configuration, employing 100K samples per dataset compared to our 50K sample size. Secondly, they adopted a higher input resolution of 1344×1344 pixels, which differs from our standardized resolution of 672×672 pixels. Due to these substantial differences in training settings and model configurations, we considered a direct performance comparison would not yield fair or meaningful results, thus justifying their exclusion from our evaluations.

C Further Analysis

C.1 Design of Padding Tokens

To ensure consistent input lengths for the modality-completion module, padding tokens are applied to align textual inputs with the fixed number of image tokens (576 tokens for Phi-3.5V-based UniMoCo). The formatted prompt structure follows this template:

$$q_t^{'} = [\mathrm{PaddingPrompt}] \| q_t \| [\mathrm{END}] \| [\mathrm{dummytokens}] \tag{4}$$

Here, q_t' represents the processed input to the modality-completion module, while q_t denotes the original textual input. The padding instruction states: *"Image modality is missing in this case. We employ a text-to-image model to generate a highly detailed visual description based on the given instruction and query. The characters following [END] serve as placeholders. Query: "*. The number of dummy tokens after [END] is calculated as $N = 576 - \ell(\mathrm{P_{pad}}) - \ell(q_t) - 1$ padding, ensuring the total length of q_t' matches the required 576 tokens. This padding mechanism maintains structural consistency between textual and visual inputs during processing.

Table 5: Hyperparameters and computational requirements for UniMoCo (Phi-3.5V) and UniMoCo (Qwen2-VL-7B) during training and test.

Hyperparameter	UniMoCo (Phi-3.5V)	UniMoCo (Qwen2-VL-7B)
Training Setting		
Resolution	336×336	672×672
Training samples	662K	
Number of Samples per Dataset	50K	
Batch size	1024	
Learning rate	6×10^{-5}	1×10^{-4}
LoRA rank	8	
Steps	2K	
GPU configuration	8×A100	
Precision	BF16	
Training time	~135 hours	~185 hours
Test Setting		
Test samples	36K	
Number of Samples per Dataset	1K	
Batch size	16	
GPU configuration	1×A100	
Precision	BF16	
Test time	~3 hours	~10 hours

We investigated various approaches for padding token formatting in our experiments. First, we designed task-specific padding prompts where each category (Classification, Retrieval, VQA, or Grounding) received distinct prompts structured as $[\mathrm{Label}][\mathrm{PaddingPrompt}]$ where label corresponds to the category name, denoted as Pad-1 in Table 7. Second, we examined varying padding lengths across task categories, with classification tasks padded to half the standard 576 tokens while other tasks retained full padding (Pad-2). This adjustment was based on the observation that classification targets typically involve concise labels (one or two words) that might benefit from shorter pseudo visual embeddings. However, both methods showed minimal performance improvements and sometimes caused degradation, prompting their abandonment.

We further explored padding length variations, testing configurations including half-length padding (288 tokens, denoted as Pad-3). This adjustment demonstrated contrasting effects across different models: while it adversely affected the performance of Phi-3.5V based UniMoCo, it proved beneficial for the Qwen2-VL-7B based implementation. The latter model, operating at 672×672 resolution, typically requires $576 \times 4 = 2304$ image tokens, but the half-length strategy effectively reduced this requirement to 1152 tokens, resulting in improved efficiency and performance.

Given the substantial variability in input text lengths—ranging from brief single-word labels to comprehensive descriptive responses—we introduced a threshold-based discrimination system (Pad-4) to dynamically adjust padding. Inputs shorter than the predetermined threshold received half-padding, whereas longer inputs retained full-length padding. However, after extensive evaluation across multiple threshold values, we observed negligible performance differences, ultimately leading to the abandonment of this adaptive padding approach.

In light of these findings, we evaluated computational overhead. Although long padding tokens increase latency—as language model processing time scales with sequence length—the overall impact of padding is modest. This stems from: (1) the modality-completion module being substantially smaller than the main LVLM backbone, and (2) the LVLM already handling lengthy sequences of visual, pseudo, and text tokens. Consequently, we can pay less attention to the computational overhead of padding and our padding length selection prioritized model performance over minimal com-

Table 6: The detailed results of the baselines and our UniMoCo on MMEB. OOD are highlighted with a yellow background in the table. Here UniMoCo-1 uses Phi-3.5V as backbone LVLM while UniMoCo-2 uses Qwen2-VL-7B as backbone LVLM.

	CLIP	OpenCLIP	SigLIP	BLIP2	UniIR	VLM2VEC	UniMoCo-1	UniMoCo-2
Classification (10 tasks)								
ImageNet-1K	55.8	63.5	45.4	10.3	58.3	65.6	62.7	75.2
N24News	34.7	38.6	13.9	36.0	42.5	79.5	81.7	69.5
HatefulMemes	51.1	51.7	47.2	49.6	56.4	67.1	71.0	77.0
VOC2007	50.7	52.4	64.3	52.1	66.2	88.6	87.1	84.5
SUN397	43.4	68.8	39.6	34.5	63.2	72.7	69.7	74.1
Place365	28.5	37.8	20.0	21.5	36.5	42.6	42.7	44.0
ImageNet-A	25.5	14.2	42.6	3.2	9.8	19.3	23.0	47.3
ImageNet-R	75.6	83.0	75.0	39.7	66.2	70.2	72.2	84.1
ObjectNet	43.4	51.4	40.3	20.6	32.2	29.5	23.5	39.6
Country-211	19.2	16.8	14.2	2.5	11.3	13.0	16	29.6
All Classification	42.8	47.8	40.3	27.0	44.3	54.81	55.0	62.6
VQA (10 tasks)								
OK-VQA	7.5	11.5	2.4	8.7	25.4	63.2	65.5	65.0
A-OKVQA	3.8	3.3	1.5	3.2	8.8	50.2	54.0	55.6
DocVQA	4.0	5.3	4.2	2.6	6.2	78.4	78.5	83.6
InfographicsVQA	4.6	4.6	2.7	2.0	4.6	40.8	43.3	47.6
ChartQA	1.4	1.5	3.0	0.5	1.6	59.0	57.8	53.2
Visual7W	4.0	2.6	1.2	1.3	14.5	47.7	52.3	48
ScienceQA	9.4	10.2	7.9	6.8	43.4	42.1	51.2	44.9
VizWiz	8.2	6.6	2.3	4.0	24.3	39.2	40	41.4
GQA	41.3	52.5	57.5	9.7	48.8	60.7	69.1	42.2
TextVQA	7.0	10.9	1.0	3.3	15.1	66.1	70.5	73.4
All VQA	9.1	10.9	8.4	4.2	16.2	54.9	58.2	55.5
Retrieval (12 tasks)								
VisDial	30.7	25.4	21.5	18.0	42.2	73.3	75.5	72.6
CIRR	12.6	15.4	15.1	9.8	51.3	47.8	50.0	51.0
VisualNews_t2i	78.9	74.0	51.0	48.1	74.3	67.2	68.5	73.0
VisualNews_i2t	79.6	78.0	52.4	13.5	76.8	70.7	70.6	69.4
MSCOCO_t2i	59.5	63.6	58.3	53.7	68.5	70.6	71.7	70.7
MSCOCO_i2t	57.7	62.1	55.0	20.3	72.1	66.5	67.9	61.3
NIGHTS	60.4	66.1	62.9	56.5	66.2	66.1	67.5	67.9
WebQA	67.5	62.1	58.1	55.4	89.6	88.1	88.5	71.0
FashionIQ	11.4	13.8	20.1	9.3	40.2	12.9	16.1	21.2
Wiki-SS-NQ	55.0	44.6	55.1	28.7	12.2	56.6	59.5	66.9
OVEN	41.1	45.0	56.0	39.5	69.4	47.3	49.3	68.1
EDIS	81.0	77.5	23.6	54.4	79.2	79.9	73.2	87.3
All Retrieval	53.0	52.3	31.6	33.9	61.8	62.3	63.2	65.0
Visual Grounding (4 tasks)								
MSCOCO	33.8	34.5	46.4	28.9	46.6	67.3	79.7	68.5
RefCOCO	56.9	54.2	70.8	47.4	67.8	84.7	85.7	83.3
RefCOCO-matching	61.3	68.3	50.8	59.5	62.9	79.2	79.9	85.8
Visual7W-pointing	55.1	56.3	70.1	52.0	71.3	86.8	84.4	75.0
All Visual Grounding	51.8	53.3	59.5	47.0	65.3	79.5	82.4	78.2
Final Score (36 tasks)								
All	37.8	39.7	34.8	25.2	42.8	60.1	61.7	63.2
All IND	37.1	39.3	32.3	25.3	47.1	66.5	68.2	67.0
All OOD	38.7	40.2	38.0	25.1	41.7	52.0	53.5	58.4
All $(T+I, T)$	29.8	33.1	15.7	27.0	32.5	56.1	57.7	59.6
All $(T, T+I)$	62.1	57.9	43.1	44.6	58.2	72.6	72.8	73.6
All $(T+I, T+I)$	41.6	44.2	37.9	49.0	59.7	61.5	64.1	65.1

putational savings.

C.2 Choice of LoRA Rank

To reduce computational costs and training time, we employ LoRA for efficient fine-tuning of the models. We conduct experiments with different LoRA ranks using Phi-3.5V as the backbone LVLM. As demonstrated in Table 8, a rank of 8 achieves optimal performance across all tasks. Consequently, we adopt this configuration for subsequent evaluations, including the main results and ablation studies.

C.3 Text-missing Modality Analysis

In our design, we incorporate task-specific instructions (e.g., "Represent the given image for classification") to supply the textual content, thereby focusing primarily on visual-missing (V-missing) scenarios. This strategy follows VLM2VEC (Jiang et al., 2024), whose Table 4 reports performance gains exceeding 20% for CLIP and VLM2VEC. Furthermore, in real-world applications, system prompts are typically provided when employing LLMs or VLMs, making text-missing (T-missing) cases extremely uncommon. Therefore, we adopt this design and direct our main focus toward V-missing scenarios.

We acknowledge that the current architecture could be extended into a symmetric architecture, where T-missing and V-missing cases are handled uniformly via a modality-completion module that generates pseudo embeddings. Exploring such adaptations is reserved for future work.

C.4 Uni-modal Performance Analysis

Our benchmark does not include (T, T) tasks, as multi-modal embedding models—including UniMoCo and prior works (Jiang et al., 2024; Lan et al., 2025; Liu et al., 2024; Gu et al., 2025)—are generally designed for scenarios where images appear in either the query or target. For purely textual retrieval tasks, uni-modal text embedding models remain more appropriate.

Although UniMoCo is a multi-modal framework, it can also operate under uni-modal conditions. We evaluated its text-only performance on the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2022), comparing it with LLM2VEC (BehnamGhader et al., 2024), an LLM-based text embedding model. The results indicate that UniMoCo achieves competitive performance, with a performance gap of no more than 8% across tasks. Since UniMoCo is optimized for multi-modal retrieval, and its use of $\mathcal{L}_1$ and $\mathcal{L}_2$ loss may limit its effectiveness in purely textual tasks.

Table 7: Performance Comparison of Different Padding Strategies. For Pad-2 and Pad-3, we also tested alternative lengths such as quarter padding. In Pad-4, the token threshold is set to 40. While we experimented with other configurations for Pad-2/3/4, the current settings yielded the best results.

Methods	**Classification**	**VQA**	**Retrieval**	**Grounding**	**Overall**
Task-Specific Padding Formats					
Pad-1 (Category prompts)	51.5	54.3	58.9	79.9	57.9
Pad-2 (Variable lengths)	52.0	52.9	59.4	81.3	58.0
Padding Length Variations					
Pad-3 (Half-length)	51.6	54.9	58.8	**82.5**	58.3
Pad-4 (Length-adaptive)	52.2	54.6	**60.7**	80.6	58.9
Baseline Configuration					
Standard Padding	**52.7**	**55.4**	60.5	80.8	**59.2**

Table 8: Performance comparison of different LoRA ranks.

Methods	**Classification**	**VQA**	**Retrieval**	**Grounding**	**Overall**
$r = 4$	52.3	51.9	56.6	76.3	56.3
$r = 8$	52.7	**55.4**	**60.5**	**80.8**	**59.2**
$r = 16$	**53.3**	54.2	59.0	77.7	58.2
$r = 32$	53.3	55.4	58.5	75.4	58.1