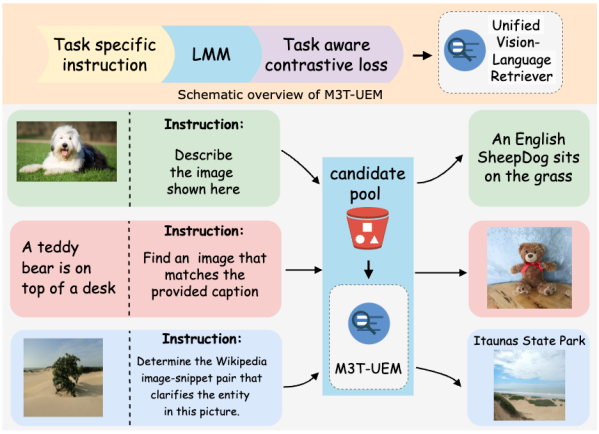
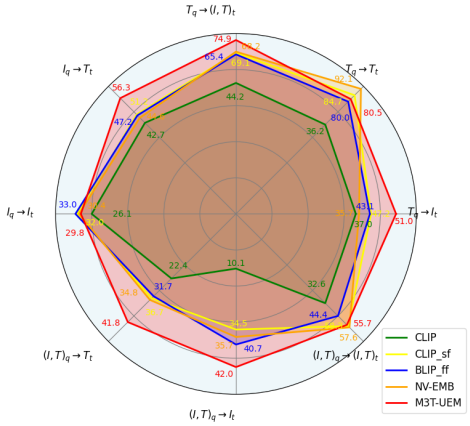
Multi-Modal Multi-Task Unified Embedding Model (M3T-UEM): A Task-Adaptive Representation Learning Framework

Rohan Sharma[1, 2] Changyou Chen[*1, 2] Feng-Ju Chang[*1] Seongjun Yun[*1]
Xiaohu Xie[*1] Rui Meng[1] Dehong Xu[1, 3] Alejandro Mottini[1] Qingjun Cui[1]
[1]Amazon [2]University at Buffalo [3] University of California, Los Angeles



(a) The M3T-UEM Framework



(b) Performance Comparison

Figure 1. M3T-UEM advances the LMM-based retriever with the designated task-specific instructions (see Supplementary 8, Table 10), a multiple-token summarization mechanism (see Fig. 2), and the task-aware contrastive loss (section 3.2). (a) The visualizations of the M3T-UEM framework's capabilities; (b) Multi-modal retrieval performance comparisons on M-BEIR to the CLIP and LMM based approaches.

Abstract

We present Multi-Modal Multi-Task Unified Embedding Model (M3T-UEM), a framework that advances vision-language matching and retrieval by leveraging a large language model (LLM) backbone. While concurrent LLM-based approaches have demonstrated impressive capabilities in multimodal and multitask scenarios; our work introduces novel mechanisms for task-adaptive learning and embedding extraction that further enhance the potential of LLM-based retrieval systems. Our key technical contribution lies in the development of a task-aware contrastive learning framework with an automated Bayesian weighing mechanism. This approach provides a principled way to balance multiple tasks during training, departing from conventional contrastive learning strategies. We further enhance the framework through a multiple token summarization strategy and an auxiliary language modeling objective, which together significantly improve retrieval performance. Comprehensive experiments on M-BEIR and ICinW benchmarks demonstrate the effectiveness of M3T-UEM, showing competitive or superior performance compared to both traditional encoder-based methods and recent LLM-based approaches. Furthermore, we demonstrate particular strengths in handling compositional conceptual changes and multilingual scenarios owing to the incorporation of an LLM backbone where the method drastically outperforms CLIP in zero-shot settings, often by orders of magnitude. [*]

1. Introduction

In the digital era, Large Multi-modal Models (LMMs), typically built upon LLMs, have become widespread due to their advanced reasoning capabilities. Their applications range from generating contextual image-based dialogues [2, 38, 44] to video understanding [17] and object segmentation [30]. Simultaneously, multimodal retrieval has

[*]Core Authors

emerged as a critical and pervasive technology, enabling users to access diverse information with rich context across numerous large-scale applications [13]. This has led to increased research investment across various sectors, including e-commerce, social media, and entertainment, all seeking to enhance their retrieval capabilities in recommender systems through multimodal and LMM integration [13, 52]. These developments have enabled sophisticated applications that handle complex query formats, such as retrieving paragraphs from image queries and processing mixed text-image inputs. This evolution has catalyzed recent adaptations of LLMs and LMMs for retrieval applications, as demonstrated in works like VLM2VEC [28], MM-Embed [35], NV-Embed [31], and MM-GEM [42].

In concordance with these recent efforts, we propose M3T-UEM, a framework that advances multi-modal retrieval by leveraging a pre-trained LLM through task-specific instructions as well as task-aware contrastive loss. Our architecture integrates visual information via a vision encoder coupled with a pre-trained QFormer and specialized projector layers. We introduce two key innovations: (1) a multiple-token summarization mechanism combined with an auxiliary language modeling objective that significantly enhances retrieval performance (Sec. 3.1), and (2) a novel task-specific contrastive learning approach (Sec. 3.2).

While existing methods rely on standard InfoNCE loss for contrastive learning, this approach faces fundamental limitations in multi-task scenarios where different objectives compete for optimization. The challenge becomes particularly acute when dealing with diverse vision-language tasks, each requiring different levels of attention to various semantic aspects. To address this, our framework introduces an automated Bayesian mechanism (Algorithm 1) that dynamically weighs different tasks during contrastive training.

The motivation for this innovation stems from several critical limitations of standard InfoNCE loss: First, it suffers from biased gradients [6, 55], particularly problematic when tasks have varying difficulties or data distributions. Second, it lacks robust mechanisms for handling noisy data pairs [48], which are inevitable in real-world multi-modal datasets. Most importantly, its single-task nature makes it ill-suited for balancing multiple competing objectives, often leading to suboptimal performance across tasks. Our task-aware InfoNCE loss directly addresses these challenges through a principled Bayesian framework, employing a stochastic Expectation Maximization algorithm that automatically adapts to task-specific requirements. This theoretical advancement, combined with our LMM architecture, enables superior performance in challenging scenarios such as zero-shot learning, multilingual retrieval, and compositional understanding - capabilities that effectively leverage the rich world knowledge and linguistic understanding inherent in pre-trained LLMs.

Our contributions are summarized as: ❶ We propose the M3T-UEM a LMM capable of multi-task and muti-modal retrieval, by leveraging an existing pre-trained LLM and incorporating vision modality and a multiple summarization mechanism ❷ We propose the task-aware InfoNCE loss which integrates task and sample specific weights into a bayesian formulation optimized through a stochastic EM mechanism enabling automated emphasis over tasks and samples. ❸ Through extensive evaluation over a multiplicity of scenarios, we demonstrate the merits of our architecture and the task-aware InfoNCE. Notably, M3T-UEM outperforms LLM based approaches [31, 35] in M-BEIR by over 1.1% on average and significantly outperforms CLIP over zero-shot, multilingual and compositional retrieval tasks, due to the inherent knowledge in LLMs.

2. Related Work

Relevant research pertains Vision-Language Pretrained Models (VLMs), which facilitate large-scale multimodal retrieval and generally fall into three categories: generative [38, 69], embedding-based [25, 50, 63], and hybrid models [33, 34, 42]. CLIP [50] jointly trains text and image encoders, whereas hybrid approaches such as BLIP [33, 34] and BLIP-2 [34] integrate vision encoders with language models unifying generative and embedding functionalities, with the embeddings serving as the visio-lingual channel. Modern vision-grounded text generation methods such as LLaVA [38] and FROMAGe [10] ground language models visually without an explicit sharing of the weights between the modalities. A recent emergence in retrieval methods [28, 31, 35, 42] is characterized by a unified embedding approach for vision and text based modalities, facilitated often by a backbone LLM or LMM, wherein we innovate through strategic architectural choices, demonstrably enhancing performance. Additionally, these methods and others aimed at representation learning [3, 5, 7, 10, 11, 15, 26, 51, 55, 57, 66] typically leverage the InfoNCE loss [7, 47] which may encounter fundamental limitations with false positives and negatives its task-agnostic treatment of data. Methods such as weighted InfoNCE [11] aim to refine similarity scoring yet remain limited. To address these limitations, our proposed M3T-UEM framework introduces a task-aware contrastive loss that adaptively optimizes similarity weighting and model parameters, enhancing multi-modal contrastive learning. A comprehensive overview of related works can be found in Supplementary 6.

3. The Proposed M3T-UEM

3.1. Model Design

The M3T-UEM architecture Our M3T-UEM model leverages a pretrained LLM backbone as a unified encoder to embed multimodal data within a shared representation

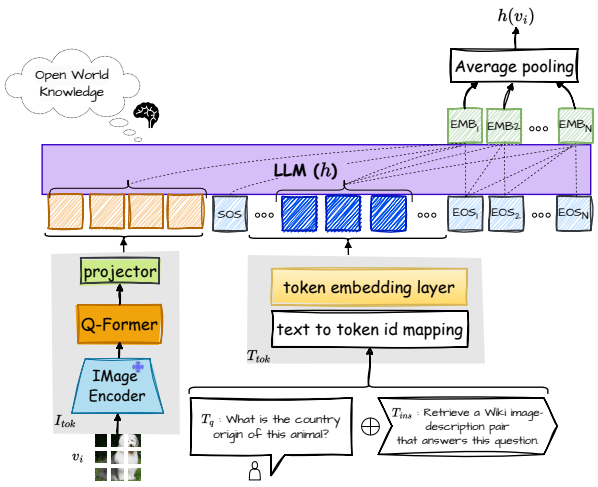


Figure 2. The M3T-UEM architecture: A decoder-only pretrained LLM processes image tokens (I_{tok}) from a vision embedder, and the text tokens (T_{tok}) from a text embedder. We prepend the start-of-sentence ($[SOS]$) token, and append N $[EOS]$ tokens to obtain the embedding of a mixed modality query input.

space. The model architecture is illustrated in Figure 2. In M3T-UEM, all modality data are transformed into a single token sequence using the LLM vocabulary, then processed as input tokens by the backbone. To map vision tokens into the LLM token space, we adopt a Q-Former adapter [34]. Although alternative MLP based adapters [38] are available, we chose Q-Former for its flexibility in controlling the length of the vision input token sequence. To enable the LLM as an effective embedding model, we introduce additional embedding tokens ($[EOS]$) at the end of each input sequence, which collaboratively capture different aspects of the representation. This design reduces the risk of losing useful information due to the heavy reliance on a single token which our ablations confirm (Section 4.4).

Multi-task design A distinctive feature of M3T-UEM compared to traditional embedding learning frameworks [33, 50] is its design to perform multi-task embedding learning, with input data composed of interleaved modalities from various datasets. To this end, we utilize the multi-task capabilities of the LLM by incorporating task-specific instructions (Supplementary 8) to distinguish between tasks. The text input format follows the pattern "Task Instruction + Text Input", guiding the model to differentiate embeddings based on the task requirements.

Remark 1 *Our distinction from the concurrent works using LLM-based architectures for multi-modal embedding learning, VLM2VEC [28], MM-GEM [42], MM-Embed [35] and NV-Embed [31] are, 1) our model is trained on larger-scale datasets incorporating diverse multi-task instructions and we explore the benefits of LLM backbones with regards to compositional and multi-lingual understanding (Section 4); 2) our framework introduces a novel task-aware contrastive loss, which addresses some inherent limitations of traditional contrastive losses (Section 3.2);*

3.2. Task-Aware Contrastive Learning

Multi-modal contrastive learning Contrastive learning is a potent and widely adopted [33, 50] framework for learning generalizable representations. In traditional multi-modal settings, the goal is to match modality embeddings after encoding by the modality-specific encoders. Consider multimodal contrastive learning on image-text pairs $\{(\mathbf{v}_i, \mathbf{t}_i)\}_{i=1}^N$, where $\mathbf{v}_i$ represents the i-th data point in the image modality, and $\mathbf{t}_i$ is the corresponding data point in text modality. In traditional encoder-based contrastive learning, each modality is encoded into a shared embedding space as $\mathbf{z}_{v_i} = f(\mathbf{v}_i)$ and $\mathbf{z}_{t_i} = g(\mathbf{t}_i)$, where $f(\cdot)$ and $g(\cdot)$ are the encoders for the image and text modalities, respectively. The standard multimodal contrastive learning aims to train these encoders by optimizing the following contrastive loss[†]: $\bar{\mathcal{L}}_{\text{con}} \triangleq -\frac{1}{N}\sum_{i=1}^N \log \bar{\mathcal{L}}_i$, where

$$\bar{\mathcal{L}}_i \triangleq \frac{\mathsf{sim}(f(\mathbf{v}_i), g(\mathbf{t}_i))}{\mathsf{sim}(f(\mathbf{v}_i), g(\mathbf{t}_i)) + \sum_{k=1}^K \mathsf{sim}(f(\mathbf{v}_i), g(\mathbf{t}_k))},$$

where $\mathsf{sim}(\cdot,\cdot)$ denotes a similarity metric between two embedding vectors, with K as the number of negative samples for $\mathbf{v}_i$ in a minibatch. We use $\mathsf{sim}(\mathbf{a}, \mathbf{b}) \triangleq e^{\mathbf{a}^T \mathbf{b}/\mu}$ in this paper with μ being the temperature parameter. For notation simplicity, we use the following: $s_i^+ \triangleq \mathsf{sim}(f(\mathbf{v}_i), g(\mathbf{t}_i))$ and $s_{ik}^- \triangleq \mathsf{sim}(f(\mathbf{v}_i), g(\mathbf{t}_k))$, so the above contrastive loss can be rewritten as $\bar{L}_i \triangleq s_i^+/(s_i^+ + \sum_{k=1}^K s_{ik}^-)$.

Task-aware contrastive learning In our M3T-UEM framework, all modalities are encoded with a shared LLM backbone whereby the image tokens generated with a pretrained vision encoder and Q-Former projector are followed by special SOS token and 16 EOS tokens forming a unified sequence. Denoting the encoding function as $h(\cdot)$, given an input sequence $\mathbf{x}$ (single or joint modality), $\mathbf{z} = h(\mathbf{x})$ represents the embedding of the input sequence. For paired data from different learning tasks (defined in Section 3.1), we treat them as query and target inputs, aiming to optimize $h(\cdot)$ to match queries with their corresponding targets via contrastive learning. A distinctive aspect of our model is its ability to balance multimodal, multi-task data, an essential capability given that real-world data often exhibit unbalanced distributions and noise. Traditional contrastive loss is ineffective under these conditions and can lead to suboptimal results. To address this, we propose a task-aware contrastive loss as illustrated in Figure 3.

Specifically, we first define a task indicator, $\tau : \mathbb{R}^d \to \mathbb{I}$, which maps a data point $\mathbf{v}_i$ or $\mathbf{t}_i$ to the corresponding

[†]For simplicity, we describe only the asymmetric contrastive loss here, although we use the symmetric one in our experiments.

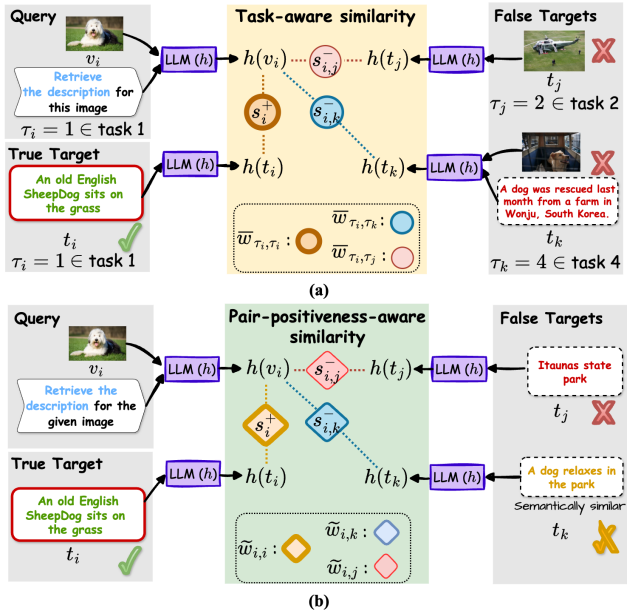


Figure 3. The proposed task-aware contrastive loss: Two essential weights are introduced to assess the positive and negative pair similarity scores. While the task-aware weights **(a)** combat the inter-task variances, e.g. to reduce the pair similarity score if the target modality is incorrect , the pair-positiveness-aware weights **(b)** account for intra-task semantic similarities between queries and targets and adjust the scale of a similarity scores accordingly.

task index, *e.g.*, $\tau(\mathbf{v}_i) = 1$ means $\mathbf{v}_i$ belongs to the first task. For notation simplicity, we denote $\tau(\mathbf{v}_i)$ or $\tau(\mathbf{t}_i)$ as τ_i. Then, we associate each task pair (i, j) with a random weight $w_{i,j}$ denoting the correlation/importance of the two task, *i.e.*, if the correlation is high, data from one task should be paid more attention in constructing the contrastive loss. To this end, a task-aware contrastive loss is defined as $\mathcal{L}_{\text{con}} \triangleq -\frac{1}{N}\sum_{i=1}^{N} \log \mathcal{L}_i$, where[‡]

$$\mathcal{L}_i \triangleq \frac{w_{\tau_i,\tau_i} s_i^+}{w_{\tau_i,\tau_i} s_i^+ + \sum_{k=1}^{K} w_{\tau_i,\tau_k} s_{ik}^-} = \frac{s_i^+}{s_i^+ + \sum_{k=1}^{K} \bar{w}_{\tau_i,\tau_k} s_{ik}^-},$$

and $\bar{w}_{\tau_i,\tau_k} \triangleq \frac{w_{\tau_i,\tau_k}}{w_{\tau_i,\tau_i}}$ reflects a task-wise importance score that will be automatically inferred during training. Additionally, inspired by [6, 48], we incorporate sample-specific weights $\tilde{w}_{ik}$ for each positive-negative pair $(\mathbf{v}_i, \mathbf{t}_k)$ for more flexible modeling, and formulate our final multi-task contrastive loss as: $\mathcal{L}_{\text{mcon}} = -\frac{1}{N}\sum_{i=1}^{N} \log \mathcal{L}_i$, with

$$\mathcal{L}_i \triangleq \frac{s_i^+}{s_i^+ + \sum_{k=1}^{K} (\bar{w}_{\tau_i,\tau_k} + \tilde{w}_{ik}) s_{ik}^-} \tag{1}$$

Remark 2 *Compared to recent work using a sample-wise weighting scheme [48], our approach with loss in (1) 1) simplifies the original loss by consolidating redundant positive weights w_{τ_i,τ_i} into the negative weights, enhancing training stability; and 2) integrates both task-wise and sample-wise adaptations, making it the first to introduce such modeling in contrastive learning, thus offering greater generalizability and improved performance.*

By incorporating these task-aware and sample-specific weights, our multi-task loss automatically balances data from different tasks (through task-aware weights $\bar{w}_{\tau_i,\tau_k}$) and handles potential noisy positive-negative data pairs (via pairwise weights $\tilde{w}_{ik}$). This capability is crucial for learning robust multimodal representations in multi-task settings, a challenge not yet addressed in existing work. Finally, our LLM backbone naturally incorporates an autoregressive language-model loss, $\mathcal{L}_{\text{lm}}$, which serves as a regularizer to balance both embedding and generation qualities and our final loss for M3T-UEM is given by

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mcon}} + \lambda \mathcal{L}_{\text{lm}}\,, \tag{2}$$

where λ is a hyperparameter set to 0.1 in our experiments.

Optimization Optimizing our task-aware contrastive loss (Eq. (1)) presents two main challenges: 1) The optimal solution tends toward a degenerate case where all weights $\bar{w}_{\tau_i,\tau_k}$ and $\tilde{w}_{ik}$ are zero, undesirable because it disregards negative data pairs, making them contribute nothing to the learning process; 2) Direct optimization of all weights via stochastic gradient descent is infeasible, as the number of weights grows quadratically with respect to the data size. Fortunately, we can leverage the stochastic expectation-maximization (EM) approach [46], to alternatively sample the weights given the LLM, and optimize the LLM backbone based on Eq. (1) using the sampled weights. We therefore reformulate the loss in Eq. (1) as a likelihood function in a probabilistic framework and introduce appropriate priors for the weights to enable efficient posterior inference. We then augment the likelihood in Eq. (1) with auxiliary random variables $\{u_i\}_{i=1}^N$ and using the Gamma identity [6], we can express Eq. (1) as a joint distribution over the data $\mathcal{D}$ and auxiliary variables u_i, conditioned on the random weights $\{\bar{w}_{\tau_i,\tau_k}, \tilde{w}_{ik}\}$, as follows:

$$p(\mathcal{D}, \{u_i\} | \{\bar{w}_{\tau_i,\tau_k}\}, \{\tilde{w}_{ik}\}) \propto s_i^+ e^{-u_i s_i^+} \prod_{k=1}^{K} e^{-u_i(\bar{w}_{\tau_i,\tau_k} + \tilde{w}_{ik}) s_{ik}^-}$$

We introduce Gamma priors[§] for the weights $\{\bar{w}_{\tau_i,\tau_k}, \tilde{w}_{ik}\}$, defined as $p(\bar{w}_{\tau_i,\tau_k}) = \mathsf{Gamma}(a_\tau, b_\tau)$ and $p(\tilde{w}_{ik}) = \mathsf{Gamma}(a, b)$. This allows the posterior distributions $p(\bar{w}_{\tau_i,\tau_k}|\mathcal{D}, u_i)$ and $p(\tilde{w}_{ik}|\mathcal{D}, u_i)$ to also follow Gamma distributions that can be directly sampled from:

$$p(\bar{w}_{\tau_i,\tau_k}|\mathcal{D}, \{u_i\}) \tag{3}$$
$$= \mathsf{Gamma}(1 + a_\tau, b_\tau + \sum_{i'}\sum_{k'} 1_{\tau_{i'}=\tau_i} 1_{\tau_{k'}=\tau_k} u_{i'} s_{i'k'}^-), \text{ and}$$

$$p(w_{ik}|\mathcal{D}, u_i) = \mathsf{Gamma}(1 + a, b + u_i s_{ik}^-)\,, \tag{4}$$

[‡]We consider a single positive pair in each $\mathcal{L}_i$ here. Derivations to handle multiple positive pairs are given in Supplementary 7.

[§]We use the shape-rate parameterization for the Gamma distribution.

Algorithm 1 Stochastic EM for Learning M3T-UEM

for iter **do**
  for ns = $1, \cdots, M$ **do**
    Sample $\{u_i\}$ from the posteriors Eq. (5).
    Sample $\{\bar{w}_{\tau_i,\tau_k}\}$ from the posteriors Eq. (3).
    Sample $\{\tilde{w}_{ik}\}$ from the posteriors Eq. (4).
  end for
  Based on the sampled weights $\{\bar{w}_{\tau_i,\tau_k}\}$ and $\{\tilde{w}_{ik}\}$, optimize the model parameter with the proposed task-aware contrastive loss in Eq. (2) with SGD.
end for

where $1_{a=b} = 1$ if a equals b and 0 otherwise. In addition, conditioned on the weights, the posterior distribution of u_i also follows simple Gamma distributions:

$$p(u_i|\mathcal{D}, \{\bar{w}_{\tau_i,\tau_k}\}, \{\tilde{w}_{ik}\}) = \mathsf{Gamma}(1, s_i^+ + \sum_{k=1}^{K} (\bar{w}_{\tau_i,\tau_k} + \tilde{w}_{ik}) s_{ik}^-) \,. \quad (5)$$

Consequently, a stochastic EM algorithm can be applied to alternately infer the random weights and optimize the model parameters (LLM weights). The specific algorithm is outlined in Algorithm 1, with more detailed explanations and derivations provided in the supplementary 7.

4. Experiments

We use the e5-Mistral-7b-instruct [59] as the backbone and adapt the best out-of-the-box version of the BLIP-2 framework containing the Q-Former and the ViT-g-14 vision encoder which we retain, ensuring the preservation of alignment between the two components. A comprehensive list of parameters are enumerated in Table 1.

Table 1. **Model and Training Hyperparameters.** Key parameters for backbone, training, and optimization.

Component	Details
Backbone	e5-Mistral-7B-Instruct [59]
Vision Encoder	ViT-g-14 + Q-Former (BLIP-2 [34])
Learning Rate	Stage 1: 2×10^{-3}, Stage 2: 1×10^{-4}
Decay & Temp.	$\gamma = 0.9999$, $\mu = 0.01$
LoRA	Rank: 32, Scaling: $\alpha = 32$
Training Steps	Stage 1: 7K, Stage 2: 14K
EOS Tokens	16, Mean-Pooling
Hardware	64× NVIDIA A100, Batch: 5120

4.1. Training Procedure

In line with prior work [33, 34, 38], we conduct the training process in two stages, using Eq. (2), as described below.

Stage 1: Initializing cross-modality alignment Herein, we aim to warm up the framework to the two modalities and therefore train the components responsible for the alignment between image and LLM token spaces using the loss 2. Specifically, we optimize (1) the Q-former, (2) the projection layer; and (3) the language modeling head, which also ensures seamless vision-conditioned text generation. Low-Rank Adaptation [21] is employed, resulting in a total of 109M trainable parameters amounting merely to $\simeq 1.4\%$ of the total number of model parameters.

Stage 2: Refining multimodal representations We minimally LoRA-finetune specific target modules, which include key projection layers in the LLM, in addition to the parameters tuned in stage 1, thereby further conditioning for the unification of multimodal representation. This stage indulges 200M parameters ($\simeq 2.5\%$ of the total). The training procedure encompasses eight multimodal retrieval tasks for which we craft customized instructions, specific to the task of retrieval, as presented in the Supplementary, Table 10. Each task describes a specific retrieval scenario across the image ($\mathcal{I}$) and text ($\mathcal{T}$) modalities, with 8 varieties in the types of queries and targets.

4.2. Datasets

Training: We use a combination of the LAION 400M [53] and CC3M [54] datasets, in addition to the recently curated M-BEIR benchmark datasets [60]. M-BEIR integrates ten diverse datasets, spanning domains such as everyday imagery, fashion, Wikipedia entries, and news articles, suited for retrieval using the human-authored instructions consisting of 1.5 mill. queries and a pool of 5.6 mil. retrieval candidates. We use weighted sampling during training using dataset sizes, therefore mitigating biases and overfitting.

Evaluation: We assess the model under a variety of scenarios. Evaluations over the M-BEIR for **multimodal multi-task retrieval**, cover eight tasks as defined in Supplementary 8. We follow the standard retrieval evaluation metric, *Recall*@5 and *Recall*@10 in keeping with the dataset specific practices. Additionally, we assess **zero-shot classification** performance using the "Image Classification in the Wild" (ICinW) benchmark [32], consisting of 20 datasets designed to assess models' ability towards categorization of images captured in diverse and real-world conditions.

Furthermore, we highlight the benefits of leveraging pretrained LLMs by evaluating over zero-shot **compositionality prediction** using the SUGARCREPE [20] and WINOGROUND [58] datasets. Specifically, SUGARCREPE assesses models for image-caption matching, where the task is to predict the correct caption or image among distractors with subtle compositional changes in concepts like **Replace**, **Swap**, and **Add object, attribute, and relation**. Similarly, the WINOGROUND dataset evaluates visio-linguistic compositional reasoning, requiring matching of images to the right captions among the others with identical words but in different orders. Prediction accuracies (top-1

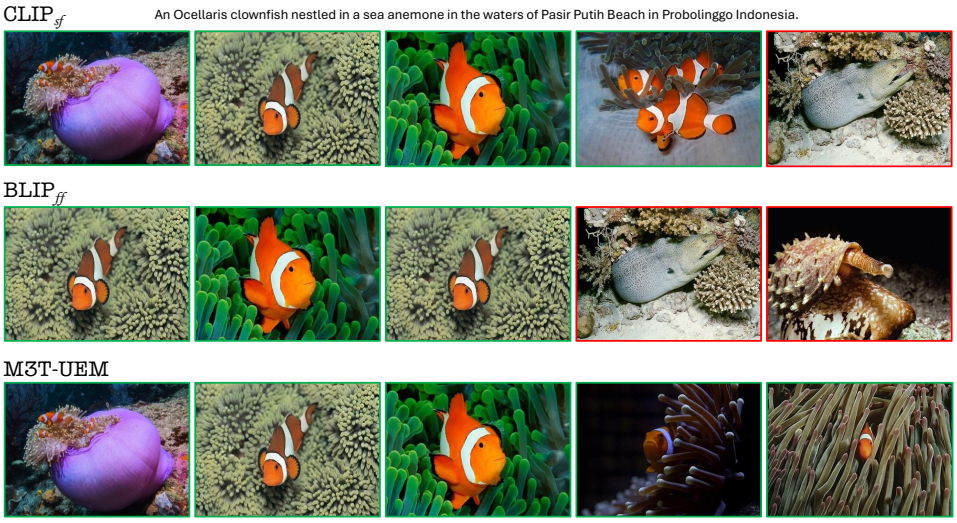


Figure 4. **Qualitative illustration:** Baseline models in the M-BEIR benchmark, such as CLIP$_{SF}$ and BLIP$_{SF}$, often retrieve images with broadly relevant visual features yet miss fine-grained semantic accuracy. For instance, they may return marine animals like moray eels instead of the correct species, Ocellaris clownfish, due to shared underwater context and similar ecological cues. M3T-UEM resolves such ambiguities by leveraging task-adaptive understanding to ground retrievals in both taxonomic and relational semantics, consistently selecting images that align with the full intent of the textual prompt demonstrating that it is better at finding nemo.

$k-$NN) are reported for both datasets treating all classification/prediction tasks as image-caption matching. Finally, **multilingual zero-shot retrieval** using the Flickr datasets [18, 64] and the XTD200 dataset [1] are conducted with comparisons against recent methods using the former. We emphasize comparisons against ViT-g-14 as it is employed as the vision encoder in our framework.

4.3. Results and Discussions

4.3.1. Multi-Modal Multi-Task Retrieval

A comprehensive evaluation over the M-BEIR retrieval benchmark is presented in Table 2, incorporating zero-shot and multi-task tuned CLIP/BLIP models [60]. Additionally, we incorporate a breadth of contemporary LMM based arts, NV-Embed [31] (upon multimodal adaptation), MM-Embed [35] and LLaVA based fine-tuned methods (Appendix 6.1). M3T-UEM (TA) demonstrates the best overall retrieval performance. We additionally compare the task-aware (TA) variant against the standard contrastive variant (STD) noting that the task-aware loss enables better recalls, further testifying to the merits of incorporating an automated regime of selective task importance based weighing within a ubiquitous contrastive loss function. We attribute the improved performance to the integration of richer image-caption datasets (LAION and CC3M) but also to the model's task awareness in distinguishing fine-grained multimodal relationships across datasets of varying scales. Qualitative evaluation consists of the top-5 retrievals ($\mathcal{T}_q \rightarrow \mathcal{I}_t$) as compared against CLIP$_{sf}$ and BLIP$_{sf}$ in Figure 4. While CLIP$_{SF}$ and BLIP$_{SF}$ retrieve visually similar underwater scenes, they often miss taxonomic cues, returning moray eels instead of Ocellaris clownfish, whereas M3T-UEM aligns with both context and species-level semantics.

4.3.2. Zero-Shot Image Classification

Table 3 compares M3T-UEM to MM-GEM and multiple CLIP and OpenCLIP based modesl over the ICinW benchmark consisting of 20 datasets with intricate visual cues and challenging classification boundaries. M3T-UEM consistently demonstrates superior performance, achieving the highest average accuracy of 67.51% and testifying to the benefits of a task-aware approach. M3T-UEM manages to focus effectively on relevant features during training and generalizes well in out-of-distribution zero-shot contexts with new, unseen classes. Additional zero-shot evaluations are under supplementary Section 9.1.

4.3.3. Compositionality Prediction

A curious trait of incorporating LLMs into retrieval tasks that we aim to highlight is demonstrated in Table 4. By leveraging the pretrained LLM's structured knowledge of object relationships, attributes, and contextual hierarchies, M3T-UEM achieves strong performance across the SUGARCREPE datasets, (detailed further in Appendix section 9.2) particularly excelling in text-driven compositional reasoning. On tasks like *"Replace"* and *"Add,"* M3T-UEM outperforms OpenCLIP ViT-g-14, demonstrating superior relational understanding with **88.9% text retrieval accuracy** compared to OpenCLIP's **81.7%** on *"Replace"*. While both models achieve competitive results in image retrieval, M3T-UEM demonstrates superior comprehension of image queries. This enhancement stems from the refined image embedding process, where ViT representations are

Table 2. **M-BEIR Retrieval:** Performance Comparisons to the SoTA zero-shot CLIP/BLIP and UniIR models, CLIP_{sf}, BLIP_{ff}, their multi-task variants (MT), and recent LMM based methods. STD and TA stand for M3T-UEM trained with standard InfoNCE loss and task aware (TA) loss, respectively. $Recall@5$ is measured except for FashionIQ and Fashion200K datasets, where we report $Recall@10$. LL-E: LLaVA-E, LL-P: LLaVA-P, MM-E: MM-Embed, NV-E: NV-Embed.

Task	Dataset	SoTA Zero-Shot		MT		UniIR		SoTA LMM				M3T-UEM	
		CLIP	BLIP2	CLIP_{sf}	BLIP_{ff}	CLIP_{sf}	BLIP_{ff}	LL-E	LL-P	MM-E	NV-E	TA	STD
$(\mathcal{T}_q \to \mathcal{I}_t)$	VisualNews [37]	43.3	16.7	40.6	22.8	42.6	23.4	33.2	34.2	41.0	32.1	40.1	**43.4**
	MSCOCO [36]	61.1	63.8	79.9	78.3	81.1	79.7	69.3	70.8	71.3	64.6	**82.1**	81.9
	Fashion200K [16]	6.6	14.0	16.8	25.8	18.0	26.1	13.5	13.3	17.1	10.4	30.7	27.7
$(\mathcal{T}_q \to \mathcal{T}_t)$	WebQA [4]	36.2	38.6	83.7	77.9	84.7	80.0	88.6	88.8	**95.9**	92.1	80.5	80.9
$(\mathcal{T}_q \to (\mathcal{I},\mathcal{T})_t)$	EDIS [39]	43.3	26.9	57.4	51.2	59.4	50.9	55.9	56.6	**68.8**	55.1	67.8	63.5
	WebQA [4]	45.1	24.5	76.7	79.2	78.7	79.8	80.3	81.6	**85.0**	81.3	82.0	80.1
$\mathcal{I}_q \to \mathcal{T}_t$	VisualNews [37]	41.3	15.0	40.0	20.9	43.1	22.8	32.4	33.3	41.3	30.4	**44.4**	43.9
	MSCOCO [36]	79.0	80.0	90.3	85.8	92.3	89.9	91.8	92.2	90.1	90.3	**93.4**	91.7
	Fashion200K [16]	7.7	14.2	18.4	27.4	18.3	28.9	13.9	14.7	18.4	13.2	31.0	28.3
$\mathcal{I}_q \to \mathcal{I}_t$	NIGHTS [14]	26.1	25.4	31.1	31.5	32.0	**33.0**	31.8	30.7	32.4	30.4	29.8	28.1
$(\mathcal{I},\mathcal{T})_q \to \mathcal{T}_t$	OVEN [22]	24.2	12.2	46.6	42.8	45.5	41.0	37.9	39.1	42.1	36.3	51.7	**51.8**
	InfoSeek [8]	20.5	5.5	28.3	23.9	27.9	22.4	31.0	32.9	**42.3**	33.3	31.9	27.7
$(\mathcal{I},\mathcal{T})_q \to \mathcal{I}_t$	FashionIQ [61]	7.0	4.4	23.2	28.4	24.4	29.2	27.4	27.0	25.7	26.0	**31.4**	29.5
	CIRR [40]	13.2	11.8	38.7	48.6	44.6	52.2	48.1	45.4	50.0	45.3	**52.5**	52.3
$(\mathcal{I},\mathcal{T})_q \to (\mathcal{I},\mathcal{T})_t$	OVEN [22]	38.8	27.3	69.0	56.3	67.6	55.8	61.6	62.6	64.1	61.7	71.4	**72.8**
	InfoSeek [8]	26.4	15.8	49.2	32.9	48.9	33.0	50.3	50.0	**57.7**	53.4	40.0	39.1
Average		32.5	24.8	49.4	45.8	50.6	46.8	47.9	48.3	52.7	47.2	**53.9**	52.7

Table 3. **ICinW Benchmark.** Evaluation using zero-shot classification accuracy (%). The datasets correspond to **C101**: Caltech101, **C10**: CIFAR10, **C100**: CIFAR100, **C211**: Country211, **DTex**: DescriTextures, **EST**: EuroSAT, **FER**: FER2013, **FGVC**: FGVC Aircraft, **OxP**: Oxford Pets, **VOC**: VOC2007, **F101**: Food101, **GT**: GTSRB, **OxF**: Oxford Flowers, **R45**: RESISC45, **HM**: HatefulMemes, **RST**: Rendered SST2, **KIT**: KITTI, **MNT**: MNIST, **PC**: PatchCamelyon, **StC**: Stanford Cars and datasets respectively. **: CLIP; *: Open CLIP

Method	C101	C10	C100	C211	DTex	EST	FER	FGVC	OxP	VOC	F101	GT	OxF	R45	HM	RST	KIT	MNT	PC	StC	**Mean Acc.**
ViT-L **	93.0	94.0	67.4	28.1	52.6	49.5	45.5	25.7	92.2	79.5	90.2	52.9	71.4	68.9	62.3	59.9	20.5	64.4	58.4	67.4	61.8
ViT-L *	94.1	96.0	82.5	25.4	61.5	65.1	47.7	32.4	92.9	80.7	89.9	56.5	74.2	68.9	72.1	60.6	22.5	65.2	57.2	91.4	66.1
ViT-g-14 *	94.4	97.1	83.9	28.8	68.3	64.5	48.1	37.8	94.3	85.8	91.6	46.6	78.1	72.6	53.3	64.6	18.2	68.4	55.1	92.9	67.2
ViT-H-14 *	84.7	97.4	84.7	29.9	67.9	71.7	50.6	42.6	94.3	77.6	92.7	54.4	79.9	70.6	53.1	64.1	11.1	72.8	53.6	93.5	67.3
MM-GEM	92.7	97.0	82.8	26.0	67.2	69.5	47.4	31.9	90.6	80.3	89.8	54.3	69.8	68.9	61.5	61.5	26.2	69.5	50.5	89.3	66.3
M3T-UEM	92.8	98.6	88.2	24.5	65.5	71.1	57.6	25.9	86.9	84.8	90.3	50.1	74.7	70.0	58.3	61.9	28.8	68.9	69.1	82.1	**67.5**

effectively mapped into the LLM embedding space. Similarly, in WINOGROUND, M3T-UEM surpasses OpenCLIP in text retrieval while maintaining comparable image retrieval accuracy. These results highlight M3T-UEM's **advantage in relational reasoning**, reinforcing the benefits of LLM-based multimodal alignment in capturing fine-grained compositional structures.

Table 4. **Compositionality:** The image-caption-matching accuracy (%) for the SUGARCREPE (SC) and WINOGROUND datasets.

Dataset	M3T-UEM		ViT-g-14	
	$\mathcal{T}_q \to \mathcal{I}_t$	$\mathcal{I}_q \to \mathcal{T}_t$	$\mathcal{T}_q \to \mathcal{I}_t$	$\mathcal{I}_q \to \mathcal{T}_t$
SC - Replace	100.0	88.9	100.0	81.7
SC - Swap	100.0	68.8	100.0	62.9
SC - Add	100.0	87.5	100.0	83.3
WinoGround	13.0	34.5	11.2	28.0
Average	**78.2**	**69.9**	77.8	64.0

4.3.4. Multilingual Zero-Shot Retrieval

Table 6 compares the zero-shot retrieval over Flickr30k [64] with contemporary arts where M3T-UEM performs at par with the SoTA methods. However, the comparative performance of M3T-UEM and OpenCLIP ViT-g-14 on multilingual zero-shot retrieval tasks in Table 5 highlights the strength of M3T-UEM's LLM architecture, which leverages multilingual capabilities despite being fine-tuned solely in English. While both models perform similarly in English, M3T-UEM achieves higher recalls. In non-English contexts, M3T-UEM consistently outperforms OpenCLIP, especially in languages like Chinese, Japanese, and Swedish, achieving **recalls as high as 81.5% and 85.2%** for Swedish compared to OpenCLIP's **26.6% and 37.1%**. The ViT-g-14 model struggles particularly with non-Latin scripts (Japanese and Chinese), underscoring M3T-UEM's superior **cross-lingual generalization** and adaptability.

Table 5. **Multilingual Zero-Shot Retrieval:** *Recall*@5 for image-text retrievals on the Flickr8k [18], Flickr30k [64], and XTD200 [1] datasets in different languages.

Dataset	M3T-UEM		ViT-g-14		Language
	$\mathcal{T}_q \rightarrow \mathcal{I}_t$	$\mathcal{I}_q \rightarrow \mathcal{T}_t$	$\mathcal{T}_q \rightarrow \mathcal{I}_t$	$\mathcal{I}_q \rightarrow \mathcal{T}_t$	
Flickr8k	91.2	97.8	90.4	96.2	English
	14.6	39.0	1.0	3.7	Chinese
Flickr30k	93.2	98.5	91.6	98.3	English
	14.9	47.5	0.9	4.8	Chinese
XTD200	93.7	94.8	87.9	89.6	English
	43.4	51.7	9.8	16.1	Japanese
	81.5	85.2	26.6	37.1	Swedish
Average	**72.9**	**77.2**	41.4	47.6	All

Table 6. **Zero-Shot Retrieval on Flickr 30K:** *Recall*@5 evaluation using the Flickr30k [64]. Closely matching model sizes for each method are chosen.

Model	$\mathcal{T}_q \rightarrow \mathcal{I}_t$	$\mathcal{I}_q \rightarrow \mathcal{T}_t$	Average ↑
TIGeR [49]	91.8	–	–
E5V [27]	82.8	90.4	86.6
VLM2VEC [28]	92.8	98.7	95.7
MM-GEM [42]	92.6	99.0	95.8
LLM2CLIP [23]	83.8	93.9	88.9
MagicLens [68]	93.7	97.7	95.7
M3T-UEM	93.2	98.5	**95.9**

Table 7. **Ablations:** Retrieval performance average over M-BEIR benchmark ablating various design components. Differences against the best variant are reported in red.

TA Loss	Two Stage	16xEOS	LM-Loss	Retrieval Avg.
✔	✔	✔	✔	38.0
✘	✔	✔	✔	37.4 (−0.6)
✔	✘	✔	✔	35.7 (−2.3)
✔	✔	✘	✔	37.6 (−0.3)
✔	✔	✔	✘	37.9 (−0.1)

4.4. Ablation Study

Table 7 illustrates the ablations for the key design choices on retrieval performance, averaged over the M-BEIR benchmark. For this study we conduct smaller scale ablations and train the stage 2 model for 5k steps. For the one vs two-stage test, we equate the computational expense (FLOPs) accounting for the difference in trainable parameters (109M vs 200M), and train the single-stage model for 8.8k steps leading to $C \times 1766$ FLOPs in both cases. We witness a marked drop in performance with this setup, underscoring the critical need for alignment of the vision-based modules with the LLM. Additionally, we find that incorporation of multiple EOS tokens endows the architecture with enhanced representations whilst using only one leads to a drop of 0.3 in retrieval performance, highlighting the merits of this design aspect where multiple tasks with interleaving modalities are involved. A more detailed ablation is provided in supplementary 10, where we further explore this aspect of our method. Task awareness additionally lends a performance boost of 0.6% under these settings. Furthermore, we find that incorporation of $\mathcal{L}_{lm}$ leads to marginal differences in performance, benefiting the method by 0.1 points, corroborating similar recent findings [42].

4.5. Inference Latency vs Performance

An evolving trend of the utility of LLMs towards numerous applications has revived the questions of efficiency vs performance. In Table 8, we explore this trade off by comparing against the comparable model size of VLM2VEC (Phi-3.5V) [¶] [28] in addition to the CLIP-based models – ViT-g-14 (our vision encoder choice in Table 1). We use a single A-100 GPU with a batch size of 64. We also analyze the average zero-shot classification accuracy. We indeed realize the performance vs throughput tradeoffs with the ViT models being more efficient. However, coupled with the superior average performance of M3T-UEM, the multi-task capabilities and the inherent benefits of LLMs as explored in sections 4.3.3 and 4.3.4, in addition to a growing literature in throughput optimization of LLMs [19, 56], we project an optimistic trend towards adaptation of LLMs for retrieval.

Table 8. **Efficiency vs Performance.** Throughput, accuracy and memory usage using a single A-100 GPU and a batch size of 64.

Model	Throughput (Samples/sec) ↑			Avg. Acc. ↑	Memory ↓
	CIFAR-100	SUN397	Country211	%	MB
ViT-g-14	20.1	19.5	19.9	59.2	8,573
VLM2VEC	16.4	15.5	16.2	38.5	35,248
M3TUEM	18.9	16.9	17.1	62.4	21,778

5. Conclusion

We introduced M3T-UEM, a multi-modal multi-task embedding framework that enhances multi-modal retrieval and classification by employing a pretrained LLM as a unified backbone across vision-language modalities. Our approach streamlines the adaptation of pretrained LLMs for multi-modal embedding applications, establishing a foundation for future research in modality integration. The proposed task-aware contrastive loss mechanism significantly improves M3T-UEM's capability to handle complex multi-modal matching scenarios. Through comprehensive empirical evaluation, M3T-UEM demonstrates consistent performance gains over CLIP-based and LMM embedding approaches across diverse tasks. These results establish a new benchmark in unified, large-scale multi-modal representation learning and open promising research in multi-modal task-aware learning, including potential extensions to additional modalities such as audio.

[¶] Evaluated using the official repo. The consumption patterns could be attributed to a larger image size (336×336) used in the Phi-3.5V model.

References

[1] Pranav Aggarwal and Ajinkya Kale. Towards zero-shot cross-lingual image retrieval. *arXiv preprint arXiv:2012.05107*, 2020. 6, 8

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 1

[4] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504, 2022. 7

[5] Changyou Chen, Jianyi Zhang, Yi Xu, Liqun Chen, Jiali Duan, Yiran Chen, Son Tran, Belinda Zeng, and Trishul Chilimbi. Why do we need large batchsizes in contrastive learning? a gradient-bias perspective. *Advances in Neural Information Processing Systems*, 35:33860–33875, 2022. 2, 1

[6] Changyou Chen, Jianyi Zhang, Yi Xu, Liqun Chen, Jiali Duan, Yiran Chen, Son Dinh Tran, Belinda Zeng, and Trishul Chilimbi. Why do we need large batchsizes in contrastive learning? a gradient-bias perspective. In *Advances in Neural Information Processing Systems*, 2022. 2, 4, 1

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 1

[8] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*, 2023. 7

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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, 2024. 1

[10] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 539–546. IEEE, 2005. 2, 1

[11] Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16670–16681, 2022. 2, 1

[12] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 215–223, Fort Lauderdale, FL, USA, 2011. PMLR. 4

[13] Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. A review of modern recommender systems using generative models (gen-recsys). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6448–6458, 2024. 2

[14] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 7

[15] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022. 2, 1

[16] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471, 2017. 7

[17] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514, 2024. 1

[18] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 6, 8

[19] Connor Holmes, Masahiro Tanaka, Michael Wyatt, Ammar Ahmad Awan, Jeff Rasley, Samyam Rajbhandari, Reza Yazdani Aminabadi, Heyang Qin, Arash Bakhtiari, Lev Kurilenko, et al. Deepspeed-fastgen: High-throughput text generation for llms via mii and deepspeed-inference. *arXiv preprint arXiv:2401.08671*, 2024. 8

[20] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality, 2023. 5

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[22] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075, 2023. 7

[23] Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Xiyang Dai, Dongdong Chen,

Chong Luo, and Lili Qiu. Llm2clip: Powerful language model unlock richer visual representation, 2024. 8, 1

[24] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 4

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2, 1

[26] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7661–7671, 2023. 2, 1

[27] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024. 8, 1

[28] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. 2024. 2, 3, 8, 1

[29] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 4

[30] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 1

[31] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024. 2, 3, 6, 1

[32] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35:9287–9301, 2022. 5

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2, 3, 5, 1

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2, 3, 5, 1

[35] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-EMBED: UNIVERSAL MULTIMODAL RETRIEVAL WITH MULTIMODAL LLMS. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 6, 1

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7

[37] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*, 2020. 7

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2, 3, 5

[39] Siqi Liu, Weixi Feng, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. Edis: Entity-driven image search over multimodal web content. *arXiv preprint arXiv:2305.13631*, 2023. 7

[40] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021. 7

[41] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2022. 1

[42] Feipeng Ma, Hongwei Xue, Guangting Wang, Yizhou Zhou, Fengyun Rao, Shilin Yan, Yueyi Zhang, Siying Wu, Mike Zheng Shou, and Xiaoyan Sun. Multi-modal generative embedding model. *arXiv preprint arXiv:2405.19333*, 2024. 2, 3, 8, 1

[43] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017. 4

[44] Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*, 2024. 1

[45] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. 4

[46] Søren Nielsen. The stochastic em algorithm: estimation and asymptotic results. *Bernoulli*, 6:457–489, 2000. 4

[47] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 1

[48] Jiayu Qin, Jian Chen, Rohan Sharma, Jingchen Sun, and Changyou Chen. A probability contrastive learning framework for 3d molecular representation learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2, 4, 1

[49] Leigang Qu, Haochuan Li, Tan Wang, Wenjie Wang, Yongqi Li, Liqiang Nie, and Tat-Seng Chua. Unified text-to-image generation and retrieval. *arXiv preprint arXiv:2406.05814*, 2024. 8

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 1

[51] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020. 2, 1

[52] KJ Sankalp, Sai Naveena BV, Charith Chandra Sai Balne, Vinodh Kumar Sunkara, Sreyoshi Bhaduri, Vinija Jain, and Aman Chadha. Advancements in modern recommender systems: Industrial applications in social media, e-commerce, entertainment, and beyond. 2024. 2

[53] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5, 2, 3

[54] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 5, 2, 3

[55] Rohan Sharma, Kaiyi Ji, Changyou Chen, et al. Auc-cl: A batchsize-robust framework for self-supervised contrastive representation learning. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 1

[56] Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Beidi Chen, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. Flexgen: High-throughput generative inference of large language models with a single gpu. In *International Conference on Machine Learning*, pages 31094–31116. PMLR, 2023. 8

[57] Rakshith Sharma Srinivasa, Jaejin Cho, Chouchang Yang, Yashas Malur Saidutta, Ching-Hua Lee, Yilin Shen, and Hongxia Jin. Cwcl: Cross-modal transfer with continuously weighted contrastive loss. *Advances in Neural Information Processing Systems*, 36, 2023. 2, 1

[58] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022. 5

[59] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. 5

[60] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*, 2023. 5, 6, 1, 2, 3

[61] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. 7

[62] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 4

[63] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. 2, 1

[64] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 6, 7, 8

[65] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1

[66] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021. 2, 1

[67] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. The visual task adaptation benchmark. 2019. 4

[68] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*, 2024. 8, 1

[69] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2, 1

Multi-Modal Multi-Task Unified Embedding Model (M3T-UEM): A Task-Adaptive Representation Learning Framework

Supplementary Material

6. Related Work

Vision-Language Pretrained Models (VLMs) form the backbone of modern large-scale multimodal retrieval systems. These models are generally categorized into generative, embedding, or hybrid models. Generative models frame retrieval tasks as autoregressive generation [38, 69], while embedding models capture the global representation of each modality, showing higher effectiveness for cross-modal retrieval and open-set classification [25, 50, 63]. Hybrid models like BLIP [33, 34] and MM-GEM [42] define a text encoder with a decoder LLM, balancing both generative and embedding functionalities.

Modality Unification has been a longstanding objective in the quest for a universal multimodal retrieval system. Approaches range from jointly training image and text encoders, as in CLIP [50], to sharing parameters between text encoders and decoders, as seen in BLIP [33, 34] and InternVL [9]. Some methods, like CoCa [65], further split the decoder into uni-modal and multi-modal components. However, none of these approaches focus on unifying or sharing weights between the image encoder and the text encoder. FROMAGe [10] grounds a language model in the visual domain by fine-tuning input and output linear layers while keeping the core language model frozen, using image features from a pretrained encoder. In contrast, our approach generates image features within the LLM itself by inputting a concatenation of the image encoder outputs and designed instructions, utilizing a shared backbone across mixed modalities.

Contrastive Learning has become the *de facto* approach for learning joint representations across multiple modalities [3, 5, 7, 10, 11, 15, 26, 51, 55, 57, 66]. The widely adopted InfoNCE loss [7, 47] treats each positive or negative pair equally, making it task-unaware and sensitive to false positive and negative pair data. This has led to various adaptations, including [11], to address these issues. A related work is [57], which introduces a weighted version of InfoNCE, though these weights are predefined and deterministic at the sample level, limiting their ability to fine-tune attention to different similarity scores in a multi-task setting. In contrast, our M3T-UEM framework, inspired and generalizing the recent flexible contrastive learning techniques [6, 48], employs a task-aware contrastive loss that jointly optimizes similarity-score-level weights and LLM model parameters, enabling more granular and adaptive control over contrastive learning in multimodal contexts.

6.1. Baselines

Baselines: We survey a breadth of contemporary arts suitable for comparison studies. M-BEIR retrieval is compared against the UniIR baselines [60]. Additionally, we incorporate the evaluation of recent LMM based methods NV-Embed [31], MM-Embed [35] and LLaVA based fine-tuned methods wherein LLaVA-E uses a similar EOS embedding for summarization whereas LLaVA-P is instructed for summarization using the last token. For the ICinW benchmark, we incorporate MM-GEM [42] as a baseline with more zero-shot comparisons against VLM2VEC [28], LLM2CLIP [23] and more [27, 41, 68]. We further incorporate ViT-g in multiple evaluations in order to elucidate the improvements made using our architecture, while leveraging it as our vision encoder.

7. Supplement for Task-Aware Contrastive Learning

7.1. Derivations to Handle Multiple Positive Pairs

To handel multiple positive pairs, we assume there are P positive pairs for each data point. Furthermore, to more closely connect positive data pairs, we assume the data from the same set of positive pairs share the same set of negative data pairs. Consequently, we define the task-aware contrastive loss with multiple positive pairs as $\mathcal{L}_{\text{con}} \triangleq -\frac{1}{NP}\sum_{i=1}^{N}\sum_{j=1}^{P}\log \mathcal{L}_{ij}$, where

$$\mathcal{L}_{ij} \triangleq \frac{w^{+}_{\tau_i,\tau_j} s^{+}_{ij}}{w^{+}_{\tau_i,\tau_j} s^{+}_{ij} + \sum_{k=1}^{K} w^{-}_{\tau_i,\tau_k} s^{-}_{ik}} = \frac{s^{+}_{ij}}{s^{+}_{ij} + \sum_{k=1}^{K} \bar{w}^{-}_{\tau_i,\tau_k} s^{-}_{ik}},$$

and $\bar{w}_{\tau_i,\tau_k} \triangleq \frac{w^{-}_{\tau_i,\tau_k}}{w^{+}_{\tau_i,\tau_j}}$ reflect task-wise importance scores that will be automatically inferred during training. Note all positive data with data i share the same set of negative pairs with similarity scores s^{-}_{ik}'s. Similar to the single positive pair case, we introduce data-wise weights $\{\tilde{w}^{-}_{ik}\}$ for more flexible modeling, resulting in the final loss as

$$\mathcal{L}_{ij} \triangleq \frac{s^{+}_{ij}}{s^{+}_{ij} + \sum_{k=1}^{K} (\bar{w}^{-}_{\tau_i,\tau_k} + \tilde{w}_{ik}) s^{-}_{ik}},$$

Now introducing an auxiliary random variable u_{ij} for each (i,j)-pair leads to an augmented likelihood distribution, defined as

$$p(\mathcal{D}, \{u_{ij}\} | \{\bar{w}_{\tau_i,\tau_k}\}, \{\tilde{w}_{ik}\}) \propto \prod_i \prod_j s_{ij}^+ e^{-u_{ij}\left(s_{ij}^+ + \sum_{k=1}^K (\bar{w}_{\tau_i,\tau_k}^- + \tilde{w}_{ik}) s_{ik}^-\right)}$$

Introducing Gamma priors for the weights $\{\bar{w}_{\tau_i,\tau_k}^-, \tilde{w}_{ik}\}$, denoted as $p(\bar{w}_{\tau_i,\tau_k}^-) = \mathsf{Gamma}(a_\tau, b_\tau)$ and $p(\tilde{w}_{ik}) = \mathsf{Gamma}(a, b)$, we have the joint posterior distribution for $\{u_{ij}\}$, $\{\bar{w}_{\tau_i,\tau_k}^-\}$, and $\{\tilde{w}_{ik}\}$ as

$$p(\{u_{ij}\}, \{\bar{w}_{\tau_i,\tau_k}^-\}, \{\tilde{w}_{ik}\} | \mathcal{D}) \propto \prod_i \prod_j s_{ij}^+ e^{-u_{ij}\left(s_{ij}^+ + \sum_{k=1}^K (\bar{w}_{\tau_i,\tau_k}^- + \tilde{w}_{ik}) s_{ik}^-\right)} p(\bar{w}_{\tau_i,\tau_k}^-) p(\tilde{w}_{ik}) \tag{6}$$

Based on the joint distribution (6), the posterior distribution for each random variable can be directly read out, as

$$p(\bar{w}_{\tau_i,\tau_k}^- | \mathcal{D}, \{u_i\}) = \mathsf{Gamma}(1 + a_\tau, b_\tau + \sum_{i'} \sum_{k'} \sum_{j=1}^P 1_{\tau_{i'}=\tau_i} 1_{\tau_{k'}=\tau_k} u_{i'j} s_{i'k'}^-)$$

$$p(w_{ik} | \mathcal{D}, u_i) = \mathsf{Gamma}(1 + a, b + \sum_{j=1}^P u_{ij} s_{ik}^-) \,,$$

$$p(u_{ij} | \mathcal{D}, \{\bar{w}_{\tau_i,\tau_k}^-\}, \{\tilde{w}_{ik}\}) = \mathsf{Gamma}(1, s_{ij}^+ + \sum_{k=1}^K (\bar{w}_{\tau_i,\tau_k}^- + \tilde{w}_{ik}) s_{ik}^-) \,.$$

7.2. Stochastic Expectation Maximization

Stochastic expectation maximization (sEM) is a stochastic version of the standard EM framework, which is introduced to efficiently learning a probability model with latent variables when dealing with large data.

Specifically, let $p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ represent a probability model, where each observation $\mathbf{x}$ (corresponding to the multi-modality input data in our case) has a corresponding latent variable $\mathbf{z}$ (corresponding to $\{u_i\}$, $\{\bar{w}_{\tau_i,\tau_k}\}$ and $\{\tilde{w}_{ik}\}$ in our case), with the global model parameter $\boldsymbol{\theta}$ (corresponding to the LLM parameter in our case). To learn the corresponding model, one standard paradigm is via maximum likelihood estimation, as

$$\max_{\boldsymbol{\theta}} \sum_i \log \int p(\mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\theta}) \mathrm{d}\, \mathbf{z}_i \,.$$

Due to the infeasibility of the integration, direct optimization of the likelihood is infeasible. The EM algorithm resolves this problem by optimizing an alternative objective function, a lower bound of the likelihood, by introducing an auxiliary distribution $q(\mathbf{z} \,|\, \mathbf{x})$ for the latent variable $\mathbf{z}$:

$$\max_{\boldsymbol{\theta}} \sum_i \int q(\mathbf{z}_i \,|\, \mathbf{x}_i) \log \frac{p(\mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\theta})}{q(\mathbf{z}_i \,|\, \mathbf{x}_i)} \mathrm{d}\, \mathbf{z}_i \,.$$

Consequently, the EM algorithm alternative between the following two steps: at iteration t

- **Expectation**: Conditioned on $\boldsymbol{\theta}_{t-1}$, estimate $q(\mathbf{z}_i \,|\, \mathbf{x}_i)$ for all training data.
- **Maximization**: Conditioned on the new estimated $q(\mathbf{z}_i \,|\, \mathbf{x}_i)$ and $\boldsymbol{\theta}_{t-1}$, maximize the following objective function to update $\boldsymbol{\theta}_{t-1}$:

$$\boldsymbol{\theta}_t = \arg\max_{\boldsymbol{\theta}_{t-1}} \sum_i \mathbb{E}_{q(\mathbf{z}_i \,|\, \mathbf{x}_i)} \left[\log p(\mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\theta}_{t-1})\right] \,. \tag{7}$$

Stochastic EM is an extension of EM at a big-data setting, where it is computationally infeasible to estimate $q(\mathbf{z}_i \,|\, \mathbf{x}_i)$ for all the data. To this end, one version of stochastic EM use samples from the posterior distribution $p(\mathbf{z}_i \,|\, \mathbf{x}_i; \boldsymbol{\theta})$ to replace $q(\mathbf{z}_i \,|\, \mathbf{x}_i)$ for a minibatch of data at each iteration, and approximate the expectation in (7) with sample averages, thus alternating between the following two steps:

- **Expectation**: Conditioned on $\boldsymbol{\theta}_{t-1}$, sample $\mathbf{z}_i \sim p(\mathbf{z}_i \,|\, \mathbf{x}_i; \boldsymbol{\theta}_{t-1})$ for the current minibatch of data.
- **Maximization**: Conditioned on the sampled $\mathbf{z}_i$'s and $\boldsymbol{\theta}_{t-1}$, maximize the following objective function to update $\boldsymbol{\theta}_{t-1}$:

$$\boldsymbol{\theta}_t = \arg\max_{\boldsymbol{\theta}_{t-1}} \sum_i \log p(\mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\theta}_{t-1}).$$

Apply the framework to our setting, we arrive at Algorithm 1 to optimize our proposed M3T-UEM framework.

7.3. Hyper-parameters Settings for Algorithm 1

We list the hyper-parameters to optimize Eq. (1), which are selected based on the validation set performances, as illustrated in Table 9.

Table 9. Hyper-parameters of Stochastic EM for Learning M3T-UEM

Hyper-parameter	**Value**
iter	5
M	# of samples in a batch (= N)
a_τ, b_τ in Eq. (3)	5
a, b in Eq. (4)	5

8. Designated Instructions

The detailed instructions applied to LAION 400M [53], CC3M [54] for creating the 8 multi-modal tasks are presented in Table 10. Note for M-BEIR [60] we use the instructions provided by the dataset itself [60].

Table 10. Designated instructions for unifying different datasets – LAION 400M [53], CC3M [54], and M-BEIR [60] to create rich multi-modal retrieval tasks.

Task	Designed Instruction
1. $\mathcal{I}_q \rightarrow \mathcal{T}_t$	Retrieve the $description$ for a given $image$, picking randomly from one of the following each time: • Describe the image shown here. • What is the caption of the image. • Write a brief caption for the image.
2. $\mathcal{T}_q \rightarrow \mathcal{I}_t$	Identify the matching $image$ for a given description, picking randomly from one of the following each time: • Pick the image that matches this description. • What is the image that is described by the caption here. • Choose the correct image using this caption as the descriptive.
3. $\mathcal{I}_q \rightarrow \mathcal{I}_t$	Match a similar $image$ based on a provided $image$ reference, picking randomly from one of the following each time: • Pick the image that matches this image. • What is the image that looks like the image here. • Choose the correct image using this image as the reference.
4. $\mathcal{I}_q \rightarrow (\mathcal{I}, \mathcal{T})_t$	Retrieve the correct $image, caption$ pair for a given $image$, picking randomly from one of the following each time: • Pick the image-caption pair that matches this image. • What is the image-caption pair that looks like the image here. • Choose the correct image-caption pair using this image as the reference.
5. $(\mathcal{I}, \mathcal{T})_q \rightarrow \mathcal{I}_t$	Identify the matching image from a $image, caption$ pair, picking randomly from one of the following each time: • Pick the image that matches this image-caption pair. • What is the image that looks like the image-caption pair here. • Choose the correct image using this image-caption pair as the reference.
6. $(\mathcal{I}, \mathcal{T})_q \rightarrow \mathcal{T}_t$	Retrieve the matching $description$ from a $image, caption$ pair, picking randomly from one of the following each time: • Pick the caption that matches this image-caption pair. • What is the description that looks like the image-caption pair here. • Choose the correct text using this image-caption pair as the reference.
7. $\mathcal{T}_q \rightarrow (\mathcal{I}, \mathcal{T})_t$	Identify the correct $image, caption$ pair for the given $description$, picking randomly from one of the following each time: • Pick the image-caption pair that matches this caption. • What is the image-caption pair that looks like the caption here. • Choose the correct image-caption pair using this caption as the reference.
8. $(\mathcal{I}, \mathcal{T})_q \rightarrow (\mathcal{I}, \mathcal{T})_t$	Match a similar $image, caption$ pair using this $image, caption$ as reference, picking from one of the following each time: • Pick the image-caption pair that matches this image-caption pair". • What is the image-caption pair that looks like the image-caption pair here. • Choose the correct image-caption pair using this image-caption pair as the reference.

Table 11. **Zero-Shot Image Classification.** We include additional zero-shot evaluation metrics using the CLIP benchmark [24] and compare our model's performance against the CLIP ViT-g-14 across seven datasets. For the DSprites benchmark, we report the mean score across sub-tasks, including predictions of shape, scale, x- and y-positions, and orientation.

Method	STL10 [12]	CLEVR Counts [29]	CLEVR Distance [29]	Sun397 [62]	SVHN [45]	DMLab [67]	DSprites (Mean) [43]	Average
OpenCLIP ViT-g-14	98.9	19.5	17.1	69.8	51.9	18.1	11.9	41.02
M3T-UEM	95.5	17.1	19.9	74.6	58.5	20.8	10.8	42.46

Table 12. **Ablation:** Performance comparison across different variants of M3T-UEM with standard contrastive loss trained for 6.5k steps using varying numbers of EOS tokens. Note different from the main results, we use the NDCG@10 metric for this ablation study taken from our earlier evaluations. Thus, the numbers are not directly comparable to the main results.

Task	**Dataset**	**EOS=1**	**EOS=4**	**EOS=16**
$(\mathcal{T}_q \to \mathcal{T}_t)$	VisualNews	40.6	41.4	41.3
	MSCOCO	58.2	61.6	63.0
	Fashion200K	5.9	6.3	6.9
$(\mathcal{I}_q \to \mathcal{T}_t)$	WebQA	23.1	23.8	23.4
$(\mathcal{T}_q \to (\mathcal{I}, \mathcal{T})_t)$	EDIS	30.0	30.4	30.3
	WebQA	36.0	35.9	36.8
$(\mathcal{I}_q \to \mathcal{T}_t)$	VisualNews	38.0	41.7	44.3
	MSCOCO	22.8	22.5	24.9
	Fashion200K	75.5	75.0	75.2
$(\mathcal{I}_q \to \mathcal{I}_t)$	NIGHTS	50.0	48.4	50.8
$((\mathcal{I}, \mathcal{T})_q \to \mathcal{T}_t)$	OVEN	62.7	59.9	60.4
	InfoSeek	32.8	35.3	40.1
$((\mathcal{I}, \mathcal{T})_q \to \mathcal{I}_t)$	FashionIQ	34.7	37.8	42.0
	CIRR	89.4	91.0	92.1
$((\mathcal{I}, \mathcal{T})_q \to (\mathcal{I}_t, \mathcal{T}_t))$	OVEN	71.5	75.3	80.4
	InfoSeek	62.7	65.2	68.7
Average		**42.3**	**43.3**	**44.9**

9. Additional Results

9.1. Zero-Shot Image Classification

We report additional zero-shot image classification evaluations in Table 11, where we compare against the pre-trained ViT-g-14 from open-clip. The datasets include STL10 [12] for object recognition, CLEVR Counts and CLEVR Distance [29] for reasoning, SUN397 [62] for scene classification, SVHN [62] for digit recognition, DMLab [67] for reinforcement learning environments and DSprites [43], which involves shape, scale, position and orientation prediction. Again, M3T-UEM demonstrates strong performance across various benchmarks, where our model achieves a competitive mean score outperforming open-clip, showcasing its robustness in diverse zero-shot settings. [‖]

9.2. Compositionality

We evaluate M3T-UEM on compositionality benchmarks, as presented in Table 13. Leveraging the pretrained LLM's world knowledge of object relationships, attributes, and contextual hierarchies, M3T-UEM demonstrates robust performance across the SugarCrepe datasets, particularly excelling in text-based compositional variations. On tasks such as *"Replace Relation"* and *"Add Object,"* M3T-UEM outperforms OpenCLIP ViT-g-14, capturing nuanced relational shifts with 81.93% text retrieval accuracy compared to OpenCLIP's 68.35% on "Replace Relation". While both models achieve high image retrieval accuracy, M3T-UEM exhibits superior comprehension of complex text queries. Similarly, in WinoGround, M3T-UEM surpasses ViT-g-14 in text retrieval while maintaining comparable image retrieval performance. These results highlight M3T-UEM's enhanced capacity for relational reasoning, demonstrating the advantage of LLM-based alignment in handling intricate compositional challenges.

Table 13. **Compositionality:** The image-caption-matching accuracy (%) for the SugarCrepe (SC) and WinoGround datasets.

Dataset	**M3T-UEM**		**ViT-g-14**	
	$\mathcal{T}_q \to \mathcal{I}_t$	$\mathcal{I}_q \to \mathcal{T}_t$	$\mathcal{T}_q \to \mathcal{I}_t$	$\mathcal{I}_q \to \mathcal{T}_t$
SC - Replace Obj.	100.0	96.6	100.0	96.0
SC - Replace Rel.	100.0	81.9	100.0	68.3
SC - Replace Att.	100.0	88.2	100.0	80.7
SC - Swap Obj.	100.0	66.9	100.0	60.4
SC - Swap Att.	100.0	70.6	100.0	65.5
SC - Add Obj.	100.0	91.5	100.0	85.8
SC - Add Att.	100.0	83.5	100.0	80.9
WinoGround	13.0	34.5	11.2	28.0
Average	**89.12**	**75.91**	88.90	71.51

10. Additional "EOS" tokens

In the following, we conduct a thorough ablation over the number of "EOS" tokens and the resulting performance over the M-BEIR dataset. For this study, we conduct the second stage training for 6.5k steps and evaluate over the M-BEIR benchmark. The Table 12 illustrates the performance of M3T-UEM variants with varying numbers of EOS tokens across multiple modalities and tasks, including text-to-image, image-to-text, and multimodal transformations. The results highlight that increasing the number of EOS tokens consistently improves the average performance metrics. This improvement can be attributed to the rich and diverse nature of multimodal information, where each modality – text, image, or a combination – encodes distinct, com-

[‖]Independent evaluations were conducted separately for both models using the repository: https://github.com/LAION-AI/CLIP_benchmark

plex representations. Using multiple EOS tokens allows the model to better capture and align these representations during contrastive learning, effectively disentangling modality-specific and shared features. This flexibility is crucial for tasks requiring nuanced understanding and retrieval, such as identifying relationships across modalities or generating contextually aligned outputs. As the complexity of the encoded information increases, the additional tokens provide the capacity needed for robust multimodal integration, ensuring higher performance across datasets and tasks.