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.