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

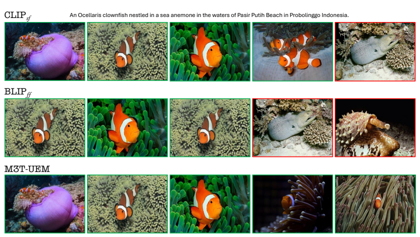
Rohan Sharma, Changyou Chen, Feng-Ju Chang, Seongjun Yun, Xiaohu Xie, Rui Meng, Dehong Xu, Alejandro Mottini, Qingjun Cui

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Introduction

M3T-UEM is a unified large language model–based framework for **multi-modal and multi-task retrieval**, introducing a **task-aware Bayesian contrastive loss** and **multi-token summarization** mechanism that deliver **state-of-the-art performance across multi-task, multi-modal, multilingual, compositional, and zero-shot retrieval benchmarks.**

Illustration and Algorithm

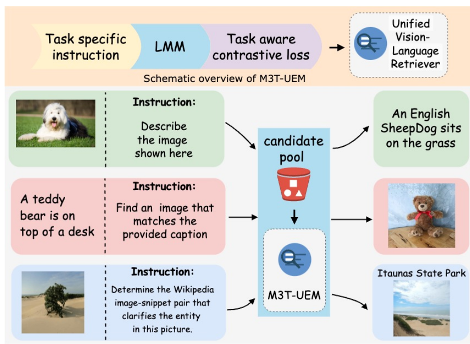


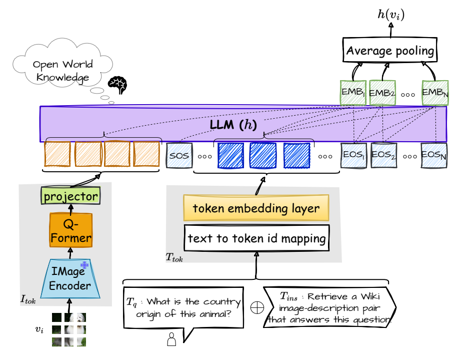
M3T-UEM shows superior retrieval in confounding scenarios.

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)

Multi-Task Learning Framework





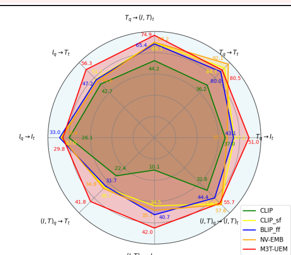
Evaluation

Image Classification in the Wild over 20 benchmark datasets. **: 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**

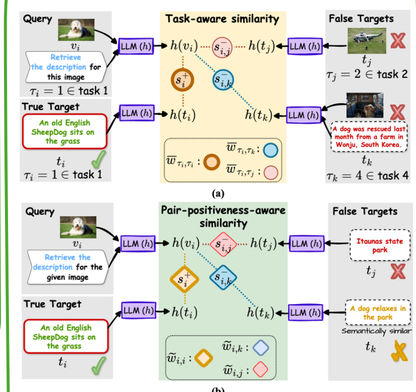
Table 4. **Compositionality:** The image-caption-matching accuracy (%) for the SugarCrepe (SC) and WinoGround datasets.

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



Multi-modal retrieval performance comparisons on M-BEIR to the CLIP and LMM based approaches.

Weighted Contrastive Learning



$$\mathcal{L}_{\text{mcon}} = -\frac{1}{N}\sum_{i=1}^{N} \log \mathcal{L}_i, \text{ with}$$

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

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

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

Introducing u_i, we have the joint

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

And thereafter, the posteriors for w_τ, u_i are

$$p(\bar{w}_{\tau_i,\tau_k} | \mathcal{D}, \{u_i\}) \quad (3)$$

$$= \text{Gamma}(1 + a_\tau, b_\tau + \sum_{i'}\sum_{k'} 1_{\tau_{i'}=\tau_i} 1_{\tau_{k'}=\tau_k} u_{i'} s_{i'k'}^-), \text{ and}$$

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

$$p(u_i | \mathcal{D}, \{\bar{w}_{\tau_i,\tau_k}\}, \{\tilde{w}_{ik}\})$$

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