CREM: Compression-Driven Representation Enhancement for Multimodal Retrieval and Comprehension

Supplementary Material

A. More Implementation Details

Training Details As shown in Tab. 1, we follow most hyperparameter configurations from [3, 6]. A cosine annealing learning rate schedule is adopted for both retrieval and generation tasks. We apply LoRA with rank 16 and alpha 64, targeting the projection layers of query, key, value, and output. This preserves the model's comprehension capacity without degrading retrieval performance. Images are processed using dynamic resolution and MRoPE, with the number of image tokens constrained between 256 and 1280. Training is conducted for 2000 steps with a 100-step warmup. All experiments are run on 8 NVIDIA A800 GPUs. Details of the retrieval training datasets are listed in Tab. 5.

Table 1. Training Hyperparameters and Computational Requirements for Retrieval and Generation Tasks.

Hyperparameter	Retrieval	Generation
Training Samples	1.1M	665K
Batch Size	1024	128
Resize Tokens	(256, 1280)	
Learning rate	5×10^{-5}	
Optimizer	AdamW	
Learning Rate Decay	cosine	
Loss Weight	1	0.5
Warmup Steps	100	
LoRA Rank	16	
LoRA Alpha	64	
LoRA Target Modules	q, k, v, o	
Temperature τ	0.02	
Training Steps	2000	
GPU Configuration	8×A800	

Evaluation Details For retrieval evaluation, we use the MMEB test set shown in Tab. 5. Generation tasks are evaluated based on VLMEvalKit [1], which supports over 80 benchmarks. We report MMBench [5] results based on its English version, as our model is trained on corpora only in English.

B. More Results and Analysis

Detailed Results on MMEB Per-task results on MMEB across 36 tasks are presented in Tab. 6. Some potentially stronger baselines are excluded due to incomplete score reporting. CREM 2B and CREM 7B variants shown are trained with high-resolution images, using up to 1280 vision tokens as in UNITE [4].

Cross-domain Evaluations on More Benchmarks Following the evaluation protocol in UniME [2], we further validate the cross-domain generalization of our model across three distinct dimensions: short-caption retrieval, long-caption retrieval, and compositional reasoning. As summarized in Tab. 2, our model, particularly the CREM-7B variant, demonstrates superior cross-domain generalization, outperforming LLaVA-OV-based UniME in most metrics. This performance edge is particularly evident in complex long-form text matching and fine-grained compositional understanding, highlighting the effectiveness of our compression-driven representation in capturing and preserving intricate multimodal semantics.

Ablations on Training Hyperparameters As shown in Tab. 3, we conduct an ablation study on the loss weights for retrieval (α_r) and generation (α_g), as well as the probability p used in stochastic compression-driven language modeling. We observe that reducing the generation loss weight yields only a marginal impact on generative quality while slightly improving embedding performance. Based on these observations, we set $\alpha_r = 1$ and $\alpha_g = 0.5$ to achieve a balanced trade-off between retrieval accuracy and comprehension ability. Furthermore, applying compression-driven training for all samples weakens native generative capability and provides only limited gains in retrieval performance, whereas pure generative training does not enhance representation quality, lacks compression ability, and leads to inferior retrieval results. Therefore, we set $p = 0.5$ to balance generative fidelity and retrieval effectiveness.

Ablations on Pooling Methods As shown in Tab. 4, we evaluate several pooling strategies for aggregating information from chorus tokens into a single retrieval embedding. Using a mlp network followed by mean pooling degrades performance, and attention pooling with a learnable query token also underperforms compared to simple mean pooling. We attribute this to weak supervision signals during contrastive training, which may hinder the pooling module from effectively learning to extract informative features across tokens.

Table 2. **Zero-shot cross-domain retrieval performance.** Results are reported as average Recall@1 across short-caption, long-caption, and compositional retrieval benchmarks. The best results are in bold.

Models	Short Caption Retrieval				Long Caption Retrieval				Compositional Retrieval		
	Flickr30K		COCO		ShareGPT4V		Urban1K		SugarCrepe		
	$q^t \to c^i$	$q^i \to c^t$	$q^t \to c^i$	$q^i \to c^t$	$q^t \to c^i$	$q^i \to c^t$	$q^t \to c^i$	$q^i \to c^t$	Replace	Swap	Add
VLM2Vec [3]	76.0	90.6	46.8	66.6	85.8	90.7	84.7	90.8	85.8	66.3	86.5
UniME [2]	**83.3**	94.4	54.8	74.0	**93.9**	89.3	94.3	95.5	80.5	65.5	82.2
CREM 2B	81.4	91.9	55.8	71.6	90.3	89.4	94.0	94.3	87.9	66.1	91.5
CREM 7B	82.9	**95.5**	**56.4**	**76.1**	90.7	**90.8**	**97.6**	**98.0**	**88.7**	**74.0**	**91.7**

Table 3. **Ablation on Loss Weights and Compression Probability.** α_r and α_g denote the loss weights for retrieval and generation objectives, respectively, while p represents the probability of applying compression-driven generation.

α_r	α_g	p	Generation		Retrieval
			Nat.	Comp.	MMEB
1	1	1	52.4	45.3	66.4
1	1	0.5	53.1	44.5	66.3
1	1	0	52.9	-	65.3
1	**0.5**	**0.5**	**53.1**	**44.2**	**66.7**
0.5	1	0.5	53.2	45.0	65.5

Table 4. **Ablation on Pooling Methods.** Evaluation of various pooling strategies used to aggregate chorus tokens for retrieval. "CLS": classification, "QA": question answering, "RET": retrieval, "GD": grounding.

Pooling	CLS	QA	RET	GD	Overall
MLP	64.6	60.7	67.6	76.0	65.8
Attention	63.6	59.4	67.7	76.7	65.3
Mean	**65.8**	**60.7**	**68.3**	**78.9**	**66.7**

C. Limitations

Although our framework achieves impressive performance on both retrieval and generation, several limitations remain. The fixed number of chorus tokens after training limits its adaptability, as optimal compression may vary between tasks. Dynamic or task-aware token allocation is a promising direction for future work. Compression-based inference leads to poor performance in OCR tasks, likely due to impaired fine-grained visual understanding, which could be alleviated by incorporating more OCR-specific data. Additionally, our training is primarily based on MMEB and ShareGPT-4V, and incorporating more diverse, high-quality retrieval and generation data may further improve generalization.

References

[1] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024. 1

[2] Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. Breaking the modality barrier: Universal embedding learning with multimodal llms. *arXiv preprint arXiv:2504.17432*, 2025. 1, 2

[3] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024. 1, 2

[4] Fanheng Kong, Jingyuan Zhang, Yahui Liu, Hongzhi Zhang, Shi Feng, Xiaocui Yang, Daling Wang, Yu Tian, Fuzheng Zhang, Guorui Zhou, et al. Modality curation: Building universal embeddings for advanced multimodal information retrieval. *arXiv preprint arXiv:2505.19650*, 2025. 1

[5] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 1

[6] Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025. 1

Table 5. **Details of Retrieval Data.** MMEB consists of 36 datasets across four meta-task categories. Of these, 20 in-distribution datasets are used for training and 16 out-of-distribution datasets are reserved for evaluation.

Meta-Task	Dataset	Query→Target	Distribution Type	#Training	#Eval	#Candidates
Classification (10 Tasks)	ImageNet-1K	I→T	IND	100K	1000	1000
	N24News	I + T→I	IND	49K	1000	24
	HatefulMemes	I→T	IND	8K	1000	2
	VOC2007	I→T	IND	8K	1000	20
	SUN397	I→T	IND	20K	1000	397
	Place365	I→T	OOD	-	1000	365
	ImageNet-A	I→T	OOD	-	1000	1000
	ImageNet-R	I→T	OOD	-	1000	200
	ObjectNet	I→T	OOD	-	1000	313
	Country-211	I→T	OOD	-	1000	211
VQA (10 Tasks)	OK-VQA	I + T→T	IND	9K	1000	1000
	A-OKVQA	I + T→T	IND	17K	1000	1000
	DocVQA	I + T→T	IND	40K	1000	1000
	InfographicVQA	I + T→T	IND	24K	1000	1000
	ChartQA	I + T→T	IND	28K	1000	1000
	Visual7W	I + T→T	IND	70K	1000	1000
	ScienceQA	I + T→T	OOD	-	1000	1000
	VizWiz	I + T→T	OOD	-	1000	1000
	GQA	I + T→T	OOD	-	1000	1000
	TextVQA	I + T→T	OOD	-	1000	1000
Retrieval (12 Tasks)	VisDial	T→I	IND	123K	1000	1000
	CIRR	I + T→I	IND	26K	1000	1000
	VisualNews_t2i	T→I	IND	100K	1000	1000
	VisualNews_i2t	I→T	IND	100K	1000	1000
	MSCOCO_t2i	T→I	IND	100K	1000	1000
	MSCOCO_i2t	I→T	IND	113K	1000	1000
	NIGHTS	I→I	IND	16K	1000	1000
	WebQA	T→I + T	IND	17K	1000	1000
	OVEN	I + T→I + T	OOD	-	1000	1000
	FashionIQ	I + T→I	OOD	-	1000	1000
	EDIS	T→I + T	OOD	-	1000	1000
	Wiki-SS-NQ	T →I	OOD	-	1000	1000
Visual Grounding (4 Tasks)	MSCOCO	I + T→I	IND	100K	1000	1000
	Visual7W-Pointing	I + T→I	OOD	-	1000	1000
	RefCOCO	I + T→I	OOD	-	1000	1000
	RefCOCO-Matching	I + T→I + T	OOD	-	1000	1000

Table 6. **Detailed MMEB Results.** Performance of baseline models and our CREM across 20 in-distribution (IND) and 16 out-of-distribution (OOD) datasets. OOD datasets are highlighted with a yellow background. For each baseline, we report the strongest variant with complete evaluation metrics: VLM2Vec 7B (LLaVA-1.6), MMRet 7B (LLaVA-1.6), UniME 7B (LLaVA-1.6), mmE5 11B (Llama-3.2-Vision), and UNITE 7B (Qwen2-VL).

	CLIP	VLM2Vec	MMRet	UniME	mmE5	UNITE	CREM 2B	CREM 7B
Classification (10 tasks)								
ImageNet-1K	55.8	74.5	58.8	71.3	77.8	80.2	82.6	83.6
N24News	34.7	80.3	71.3	79.5	81.7	80.3	77.8	81.2
HatefulMemes	51.1	67.9	53.7	64.6	64.2	67.1	62.1	65.0
VOC2007	50.7	91.5	85.0	90.4	91.0	84.9	80.8	85.6
SUN397	43.4	75.8	70.0	75.9	77.7	78.7	77.5	77.3
Place365	28.5	44.0	43.0	45.6	43	44.5	41.5	44.0
ImageNet-A	25.5	43.6	36.1	45.5	56.3	59.2	47.6	54.5
ImageNet-R	75.6	79.8	71.6	78.4	86.3	90.5	90.8	90.2
ObjectNet	43.4	39.6	55.8	36.4	62.5	68.1	72.2	71.4
Country-211	19.2	14.7	14.7	18.7	35.4	29.5	25.1	30.1
All Classification	42.8	61.2	56.0	60.6	67.6	68.3	65.8	68.3
VQA (10 tasks)								
OK-VQA	7.5	69.0	73.3	68.3	67.6	67.1	63.9	71.8
A-OKVQA	3.8	54.4	56.7	58.7	56.1	58.0	53.4	61.4
DocVQA	4.0	52.0	78.5	67.6	90.3	92.7	91.7	94.2
InfographicsVQA	4.6	30.7	39.3	37.0	56.5	71.3	65.2	76.2
ChartQA	1.4	34.8	41.7	33.4	50.5	63.2	53.3	67.2
Visual7W	4.0	49.8	49.5	51.7	51.9	54.9	52.7	57.4
ScienceQA	9.4	42.1	45.2	40.5	55.8	51.2	42.5	59.2
VizWiz	8.2	43.0	51.7	42.7	52.8	53.4	48.2	53.6
GQA	41.3	61.2	59.0	63.6	61.7	56.8	54.0	66.5
TextVQA	7.0	62.0	79.0	65.2	83.3	82.3	82.4	86.4
All VQA	9.1	49.9	57.4	52.9	62.6	65.1	60.7	69.4
Retrieval (12 tasks)								
VisDial	30.7	80.9	83.0	79.7	74.1	80.5	80.5	85.6
CIRR	12.6	49.9	61.4	52.2	54.7	51.6	58.2	62.5
VisualNews_t2i	78.9	75.4	74.2	74.8	77.6	79.3	71.3	79.5
VisualNews_i2t	79.6	80.0	78.1	78.8	83.3	82.4	76.9	83.2
MSCOCO_t2i	59.5	75.7	78.6	74.9	76.4	78.2	75.1	78.3
MSCOCO_i2t	57.7	73.1	72.4	73.8	73.2	74.3	72.2	73.8
NIGHTS	60.4	65.5	68.3	66.2	68.3	66.0	66.8	67.6
WebQA	67.5	87.6	90.2	89.8	88.0	87.0	89.3	90.8
FashionIQ	11.4	16.2	54.9	16.5	28.8	26.3	16.1	23.3
Wiki-SS-NQ	55.0	60.2	24.9	66.6	65.8	72.2	61.0	73.2
OVEN	41.1	56.5	87.5	55.7	77.5	73.1	67.4	73.1
EDIS	81.0	87.8	65.6	86.2	83.7	88.3	84.5	84.4
All Retrieval	53.0	67.4	69.9	67.9	71.0	71.6	68.3	72.9
Visual Grounding (4 tasks)								
MSCOCO	33.8	80.6	76.8	76.5	53.7	73.9	65.0	69.5
RefCOCO	56.9	88.7	89.8	89.3	92.7	89.2	86.5	92.0
RefCOCO-matching	61.3	84.0	90.6	90.6	88.8	90.1	89.3	93.6
Visual7W-pointing	55.1	90.9	77.0	84.1	92.3	86.1	74.7	89.3
All Visual Grounding	51.8	86.1	83.6	85.1	89.6	84.8	78.9	86.1
Final Score (36 tasks)								
All	37.8	62.9	64.1	66.6	69.8	70.3	66.7	72.1
All IND	37.1	67.5	59.1	68.4	72.3	73.6	70.8	75.6
All OOD	38.7	57.1	68.0	57.9	66.7	66.3	61.5	67.8