U-MARVEL: Unveiling Key Factors for Universal Multimodal Retrieval via Embedding Learning with MLLMs

Xiaojie Li, Chu Li, Shi-Zhe Chen, Xi Chen

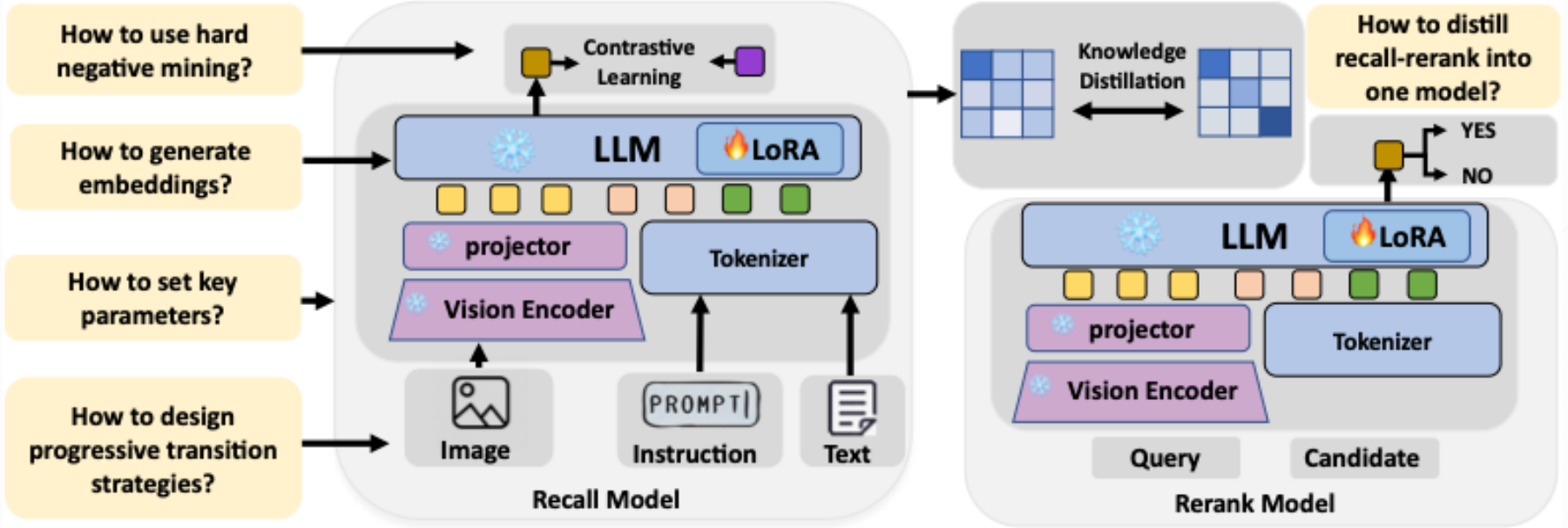
Tencent PCG | Nanjing University | ByteDance

CONTENTS

01. Introduction

Background

Universal Multimodal Retrieval (UMR) handles complex queries and candidates across diverse modalities. Despite the success of MLLM-based retrievers, the design principles for their embedding capabilities are not well understood.



To fill this gap, we conduct a comprehensive study on:

- **Model Adaptation:** How to effectively transform decoder-only MLLMs into embedders.
- **Training Strategies:** The impact of batch size, learning rate, and hard negative mining.
- **Efficiency:** Distilling the "recall-then-rerank" paradigm into a single model

Contribution

Based on these insights, we have introduced **U-MARVEL**, which outperforms current competitors at various scales on the M-BEIR benchmark (e.g., 4B, 7B) by a large margin and shows superior zero-shot performance across various tasks.

02. Recipe for Building U-MARVEL

1. How to Adapt MLLMs into Embedding Models?

Embedding Extraction

Finding1: Generating embeddings with bidirectional attention and mean pooling outperforms the common approach of using compression prompts with the last token mechanism.

Instruction Integration

Finding2: Masking instruction tokens during mean pooling enhances embedding performance.

Progressive Transition

Finding3: Progressive transition effectively adapts decoder-only MLLMs to embedding models through stepwise training..

Table1 : Incremental improvements of MLLM adaptation

ID	Methods	Local Avg.	Global Avg.
0	Baseline (Causal-attn + Last token)	56.6	54.8
1	Bid-attn and Mean Pooling	57.2	55.2
2	ID 1 + Instruction Integration	57.3	55.5
3	ID 2 + Progressive Transition	57.7	55.8

02. Recipe for Building U-MARVEL

2. How to Train MLLM-based Embedders by InfoNCE?

Hyperparameter Synergy

Finding3: Increasing batch size yields performance gains, but these improvements plateau without appropriate learning rate scaling. Additionally, learnable temperature parameters play a pivotal role in enhancing the effectiveness of contrastive learning.

Hard Negative Mining

Fingding4: Filtering Hard negatives may hinder convergence during training. Filtering false negatives and mixing random in-batch negatives help balance difficulty and improve performance.

Table: Incremental improvements from training recipes

ID	Methods	Local Avg.	Global Avg.
4	ID 3 + Hyperparameter Synergy	**60.1**	-
5	ID 4 + Continual Training with Hard Negative Mining	**61.7**	**59.9**

02. Recipe for Building U-MARVEL

3. Will Reranker Distillation Improve Performance?

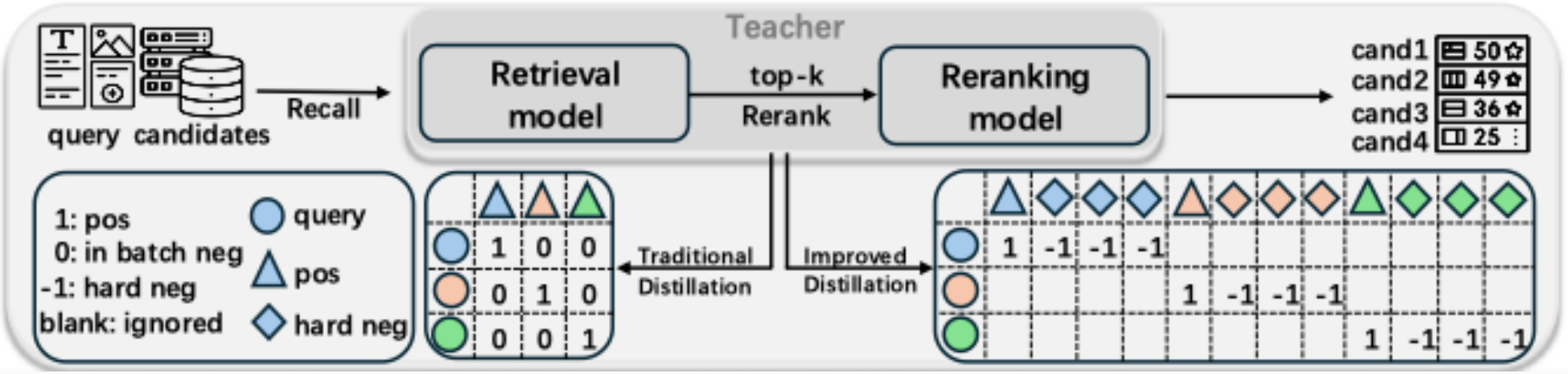


Figure 2: Distillation Strategy Illustration

- **The Bottleneck of Traditional Distillation:** Standard approaches compute similarity matrices using all *in-batch negatives*. For an MLLM-based reranker, evaluating all $\mathcal{O}(n^2)$ pairs incurs prohibitive computational costs (e.g., theoretically > 340 hours), making distillation practically infeasible.
- **Our Improved Distillation Strategy:** We reconstruct the distillation samples to consist strictly of the query, its positive match, and the **top-k hard negatives** retrieved by the recall model. We then align the student's probability distribution with the teacher's ensemble scores using KL divergence.

Cost Reduction
~96%

Diversity Increase
26x

ID	Methods	Local Avg.	Global Avg.
5	Hard Negative Mining	**61.7**	**59.9**
6	U-MARVEL (ID 5 + Improved Distillation)	**63.2**	**60.7**

03. Proposed Framework: U-MARVEL

Motivated by the findings presented above, we introduce a unified framework, termed **U-MARVEL** (**U**niversal **M**ultimod**A**l **R**etrie**V**al via **E**mbedding **L**earning), which consists of three key stages:

1. **Progressive Transition:** The model is progressively fine-tuned on the increasing complexity levels of the retrieval data, enabling it to gradually adapt to retrieval tasks.
2. **Hard Negative Mining & Fusion Reranker Model:** Building on progressive transition, we first train a hard negative model for recall and a reranker model for ranking, then linearly combine them to obtain a powerful recall-reranker model.
3. **Improved Distillation:** We perform improved knowledge distillation on the recall-reranker model to achieve single-stage efficiency.

Key Results: Overall Performance

SOTA Performance

U-MARVEL significantly outperforms the baseline (LamRA-Ret) in both single-model and reranker-augmented settings on the M-BEIR benchmark.

Strong Generalization

Maintains SOTA performance across different backbones (e.g., Qwen3-VL-4B-Instruct), demonstrating robust generalization across model sizes.

Methods	$q^t \to c^i$			$q^t \to c^t$		$q^t \to (c^i, c^t)$		$q^i \to c^t$			$q^i \to c^i$		$(q^i, q^t) \to c^i$		$(q^i, q^t) \to c^t$		Avg.
	VisualNews R@5	MSCOCO R@5	Fashion200K R@10	WebQA R@5	EDIS R@5	WebQA R@5	VisualNews R@5	MSCOCO R@5	Fashion200K R@10	NIGHTS R@5	OVEN R@5	InfoSeek R@5	FashionIQ R@10	CIRR R@5	OVEN R@5	InfoSeek R@5	
Single model																	
CLIP-L (Radford et al. 2021)	43.3	61.1	6.6	36.2	43.3	45.1	41.3	79	7.7	26.1	24.2	20.5	7	13.2	38.8	26.4	32.5
SigLIP (Zhai et al. 2023)	30.1	75.7	**36.5**	39.8	27	43.5	30.8	88.2	34.2	28.9	29.7	25.1	14.4	22.7	41.7	27.4	37.2
UniIR-BLIP$_{FF}$ (Wei et al. 2024)	23.4	79.7	26.1	80	50.9	79.8	22.8	89.9	28.9	33	41	22.4	29.2	52.2	55.8	33	46.8
UniIR-CLIP$_{SF}$ (Wei et al. 2024)	42.6	81.1	18	84.7	59.4	78.7	43.1	92.3	18.3	32	45.5	27.9	24.4	44.6	67.6	48.9	50.6
LamRA-Ret (Liu et al. 2024b)	41.6	81.5	28.7	86	62.6	81.2	39.6	90.6	30.4	32.1	54.1	52.1	33.2	53.1	76.2	63.3	56.6
U-MARVEL(Qwen3VL-4B-Instruct)	36.2	82.6	27.9	96.8	66.2	86.7	35.7	**93.9**	26.5	33.3	57.9	56.8	36.0	60.5	73.3	70.1	58.8
U-MARVEL(Qwen2VL-7B-Instruct)	**47.3**	**84.4**	33.6	**97.1**	**78.8**	**88.5**	**47.3**	93.5	**35.1**	**34.2**	**62.5**	**58.3**	**36.4**	**60.7**	**79.4**	**74.7**	**63.2**
+Reranker																	
LamRA (Liu et al. 2024b)	48	85.2	32.9	96.7	75.8	87.7	48.6	92.3	36.1	33.5	59.2	**64.1**	37.8	**63.3**	79.2	78.3	63.7
U-MARVEL$^+$ (Qwen3VL-4B-Instruct)	42.0	85.5	29.5	98.1	69.1	88.8	41.3	93.0	30.7	**35.0**	59.0	58.0	38.0	64.1	78.1	75.1	61.6
U-MARVEL$^+$ (Qwen2VL-7B-Instruct)	**49.4**	**85.6**	34.2	**98.5**	**81.4**	**89.4**	**50.5**	88.4	**37.7**	34.7	**63.7**	62.9	**38.2**	63.2	**80.8**	**78.9**	**64.8**

Table 7: Comparisons with SoTA approaches on M-BEIR benchmark in local pool setting.

Key Results: Overall Performance

SOTA Performance

U-MARVEL significantly outperforms the baseline (LamRA-Ret) in both single-model and reranker-augmented settings on the M-BEIR benchmark.

Strong Generalization

Maintains SOTA performance across different backbones (e.g., Qwen3-VL-4B-Instruct), demonstrating robust generalization across model sizes.

Methods	$q^t \to c^i$			$q^t \to c^t$		$q^t \to (c^i, c^t)$		$q^i \to c^t$			$q^i \to c^i$		$(q^i, q^t) \to c^i$		$(q^i, q^t) \to c^t$		Avg.
	VisualNews R@5	MSCOCO R@5	Fashion200K R@10	WebQA R@5	EDIS R@5	WebQA R@5	VisualNews R@5	MSCOCO R@5	Fashion200K R@10	NIGHTS R@5	OVEN R@5	InfoSeek R@5	FashionIQ R@10	CIRR R@5	OVEN R@5	InfoSeek R@5	
Single model																	
UniIR-BLIP$_{FF}$ (Wei et al. 2024)	23	75.6	25.4	79.5	50.3	79.7	21.1	88.8	27.6	33	38.7	19.7	28.5	51.4	57.8	27.7	45.5
UniIR-CLIP$_{SF}$ (Wei et al. 2024)	42.6	**77.9**	17.8	84.7	59.4	78.8	42.8	92.3	17.9	32	39.2	24	24.3	43.9	60.2	44.6	48.9
MM-Embed (Lin et al. 2024)	41	71.3	17.1	95.9	68.8	85	41.3	90.1	18.4	32.4	42.1	42.3	25.7	50	64.1	57.7	52.7
LamRA-Ret (Liu et al. 2024b)	41.3	75.4	28.7	85.8	62.5	81	39.3	90.4	30.4	32.1	48.4	48.7	33.1	50.5	70	60	54.9
U-MARVEL(Qwen3VL-4B-Instruct)	36.0	73.2	27.9	96.6	65.9	86.3	35.6	**93.9**	26.3	33.2	53.7	47.7	35.8	**57.2**	67.1	62.7	56.2
U-MARVEL(Qwen2VL-7B-Instruct)	**47.2**	72.8	**33.3**	**96.7**	**78.7**	**87.7**	**47.2**	93.5	**34.9**	**34.0**	**58.3**	**52.2**	**36.0**	56.0	**73.1**	**69.2**	**60.7**
+Reranker																	
LamRA (Liu et al. 2024b)	46.9	**78**	32.5	96.5	74.4	87.1	47.6	92.4	36.6	34.2	54	**58.7**	37.4	59.7	72.6	**74**	61.4
U-MARVEL^{+} (Qwen3VL-4B-Instruct)	41.2	71.1	29.7	98.0	69.7	**88.8**	40.8	90.8	30.6	34.1	53.7	51.0	37.4	**60.7**	71.4	70.2	58.7
U-MARVEL^{+} (Qwen2VL-7B-Instruct)	**48.8**	70.1	**33.8**	**98.3**	**80.8**	88.3	**49.8**	86.0	**36.8**	**34.8**	**58.7**	56.9	**37.4**	58.4	**74.9**	73.8	**61.8**

Table 8: Comparisons with SoTA approaches on M-BEIR benchmark in global pool setting.

Key Results: Zero-shot Evaluation

Zero-shot Capabilities

We evaluated U-MARVEL on unseen datasets for image-text and text-to-video retrieval.

Strong Transferability

U-MARVEL achieves state-of-the-art zero-shot performance, demonstrating robust transferability to new tasks.

Methods	$q^i \rightarrow c^t$			$q^t \rightarrow c^i$			$(q^i, q^t) \rightarrow c^i$		$q^{dialog} \rightarrow c^i$	$(q^i \oplus q^t) \rightarrow c^i$	ITM	
	ShareGPT4V R@1	Urban-1K* R@1	Flickr R@1	ShareGPT4V R@1	Urban-1K* R@1	Flickr R@1	CIRCO* MAP@5	GeneCIS* R@1	Visual Dialog* R@1	Multi-round FashionIQ* R@5	CC-Neg Acc.	Sugar-Crepe* Acc.
Single model												
CLIP-L (Radford et al. 2021)	84	52.8	67.3	81.8	68.7	87.2	4	13.3	23.7	17.7	66.7	73
UniIR-CLIP$_{SF}$ (Wei et al. 2024)	85.8	75	78.7	84.1	78.4	**94.2**	12.5	16.8	26.8	39.4	79.9	80.3
E5-V (Jiang et al. 2024a)	86.7	84	79.5	84	82.4	88.2	24.8	18.5	54.6	19.2	83.2	84.7
MagicLens-L (Zhang et al. 2024b)	85.5	59.3	72.5	60.9	24.2	84.6	29.6	16.3	28	22.6	62.7	75.9
VLM2Vec (Jiang et al. 2024b)	90.7	90.8	76	85.8	84.7	90.6	-	-	-	-	-	79.5
UniME (Gu et al. 2025)	**97.2**	95.9	81.9	93.9	95.2	93.4	-	-	-	-	-	85
LamRA-Ret (Liu et al. 2024b)	93.3	95.1	82.8	88.1	94.3	92.7	33.2	18.9	62.8	60.9	79.6	85.8
U-MARVEL(Qwen3VL-4B-Instruct)	96.4	96.5	79.3	96.7	**98.2**	87.4	33.3	18.4	64.5	63.8	76.4	**89.0**
U-MARVEL(Qwen2VL-7B-Instruct)	96.4	**96.7**	**85.4**	**97.2**	93.5	93.3	**36.2**	**19.1**	**70.3**	**65.7**	**84.5**	87.9
+Reranker												
LamRA (Liu et al. 2024b)	**97.9**	**98.8**	88.1	96.5	98	**97.6**	42.8	**24.8**	70.9	63.9	85.9	**93.5**
U-MARVEL$^+$(Qwen3VL-4B-Instruct)	97.8	98.4	86.4	**99.0**	99.0	93.3	42.3	21.6	74.2	65.1	76.7	95.2
U-MARVEL$^+$(Qwen2VL-7B-Instruct)	97.8	97.7	**88.5**	98.9	**98.2**	95.1	**46.0**	22.6	**77.6**	**66.3**	**86.1**	93.4

* indicates that the images in these datasets are sourced from COCO or FashionIQ.

Table 9: Comparisons with SoTA approaches on zero-shot image and text benchmarks

Key Results: Zero-shot Evaluation

Zero-shot Capabilities

We evaluated U-MARVEL on unseen datasets for image-text and text-to-video retrieval.

Strong Transferability

U-MARVEL achieves state-of-the-art zero-shot performance, demonstrating robust transferability to new tasks.

Methods	MSR-VTT			MSVD		
	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot (finetuned with text-video data)						
InternVideo (Wang et al. 2022)	40.0	65.3	74.1	43.4	69.9	79.1
ViCLIP (Wang et al. 2023)	42.4	-	-	49.1	-	-
UMT-L (Li et al. 2023b)	42.6	64.4	73.1	49.9	77.7	85.3
InternVideo2$_{s2}$-6B (Wang et al. 2024b)	55.9	78.3	85.1	59.3	84.4	89.6
Zero-shot (finetuned only with text-image data)						
VLM2Vec (Jiang et al. 2024b)	43.5	69.3	78.9	49.5	77.5	85.7
LamRA (Liu et al. 2024b)	44.7	68.6	78.6	52.4	79.8	87.0
LLaVE-7B (Lan et al. 2025)	46.8	71.1	80.0	52.9	80.1	87.0
U-MARVEL(Qwen3VL-4B-Instruct)	34.1	54.4	64.1	44.0	70.3	79.1
U-MARVEL(Qwen2VL-7B-Instruct)	**47.2**	**72.0**	**80.2**	**54.6**	**80.9**	**87.7**
U-MARVEL^{+}(Qwen3VL-4B-Instruct)	47.0	69.7	78.5	53.1	79.9	87.1
U-MARVEL^{+}(Qwen2VL-7B-Instruct)	34.5	54.2	64.1	44.1	70.4	79.1

Table 10: Comparisons with SoTA approaches on zero-shot text-to-video retrieval benchmarks.

Conclusion & Acknowledgements

Conclusion

- We conducted a comprehensive study to uncover key factors for effective UMR with MLLMs.
- We introduced U-MARVEL, a unified framework that achieves SOTA performance on M-BEIR and strong zero-shot generalization.

Acknowledgements

Code Repository:

https://github.com/chaxjli/U-MARVEL

Contact Us:

chaxjli@tencent.com

shizhechen@tencent.com

jasonxchen@tencent.com

lichu@bytedance.com

Thank you for your attention!