Appendix

A. Data Instructions

For train data, in-domain test data and zero-shot Instance-Location Search task, the query instruction is: "Find me an image containing the object in the given image with the following caption". To create the final query text, the provided caption is concatenated with questions generated by the MLLM. For zero-shot Instance Search task, the query text is: "Given the [instance name] in the image, find a new image containing the [instance name]". For candidates that only image modality are available, the instruction is: "Represent the given image".

B. More Details of IDMR dataset

In this section, we provide statistics of our IDMR zero-shot test data in Tab. 6 We also visualize the examples of the synthetic training data (Tab. 7) and zero-shot test data (Tab. 8) for Instance-Driven Multimodal Retrieval.

C. Qualitative Results of IDMR model

Top-5 retrieved images of our IDMR model and VLM2Vec [15] on IDMR in-domain test set and zero-shot test data are shown in Fig. 6 and Fig. 7 respectively.

D. Full Results on MMEB

In Tab. 9, we present detailed results on the MMEB benchmark [15]. This benchmark includes 20 in-distribution datasets and 16 out-of-distribution (OOD) datasets, with the OOD datasets highlighted in blue in the table. Our 26B IDMR model achieves state-of-the-art performance across these tasks. Additionally, our 8B model outperforms MMRet, even though MMRet has a comparable model size and is trained on 26M data.

Base Dataset	Qry. Text	Qry. Image	# Triplets	# Candidate Images
LaSOT [8]	Instance Search	Crop	1400	20
		Full	1400	20
	Location conditioned Instance Search	Crop	1400	1000
		Full	1400	1000
EPIC-KITCHENS-100 [7]	Instance Search	Crop	120	100
		Full	120	100
	Location conditioned Instance Search	Crop	120	100
		Full	120	100

Table 6. Statistics of the four zero-shot subsets in IDMR-bench.

Dataset	Query Image	Query Text	Traget Image
COCO [21]		Find me an image containing the object in the given image with the following caption: The surfboard is being held by a person wearing an orange shirt and a beanie, with a sandy path and trees in the background.	
COCO [21]		Find me an image containing the object in the given image with the following caption: The clock is in the middle of the building.	
COCO [21]		Find me an image containing the object in the given image with the following caption: The baseball glove is located on the pitcher's left hand.	
Objects365 [30]		Find me an image containing the object in the given image with the following caption: The fan is located on the ceiling near a person playing guitar.	
Objects365 [30]		Find me an image containing the object in the given image with the following caption: The drum is located in front of the group of people dressed in kilts.	
Objects365 [30]		Find me an image containing the object in the given image with the following caption: The cow is near the shoreline of a beach with waves rolling in.	
OpenImages [16]		Find me an image containing the object in the given image with the following caption: The Jug is placed on a table, next to a plate of food and a glass.	
OpenImages [16]		Find me an image containing the object in the given image with the following caption: The Jet ski is on a body of water near a dock and a house with a stone wall.	
OpenImages [16]		Find me an image containing the object in the given image with the following caption: The Lizard is situated on a dirt ground near a wooden structure, surrounded by trees and other reptiles.	

Table 7. Examples of IDMR training dataset.

Dataset	Subtask	Query Image	Query Text	Traget Image
LaSOT [8]	Instance		Given the sheep in the image, find an everyday image that contains the sheep.	
LaSOT [8]	Instance		Given the bicycle in the image, find an everyday image that contains the bicycle.	
LaSOT [8]	Location		Find me an image containing the object in the given image with the following caption: Find a picture that the umbrella is being held by a woman on a paved path with greenery in the background.	
LaSOT [8]	Location		Find me an image containing the object in the given image with the following caption: The cat is sitting on a patch of dirt or gravel.	
Kitchens [7]	Instance		Given the knife in the image, find an everyday image that contains the knife.	
Kitchens [7]	Instance		Given the bottle in the image, find an everyday image that contains the bottle.	
Kitchens [7]	Location		Find me an image containing the object in the given image with the following caption: The Bottle is located in the bottom right corner of the refrigerator door, next to a jar.	
Kitchens [7]	Location		Find me an image containing the object in the given image with the following caption: The Knife is located on a black cutting board, with a blue kitchen appliance to its right and a white juicer to its left.	

Table 8. Examples of IDMR zero-shot testing dataset.

Figure 6. Visualization of the Top-5 results from our model v.s. VLM2Vec on zero-shot test data, with the correct target image highlighted in green.

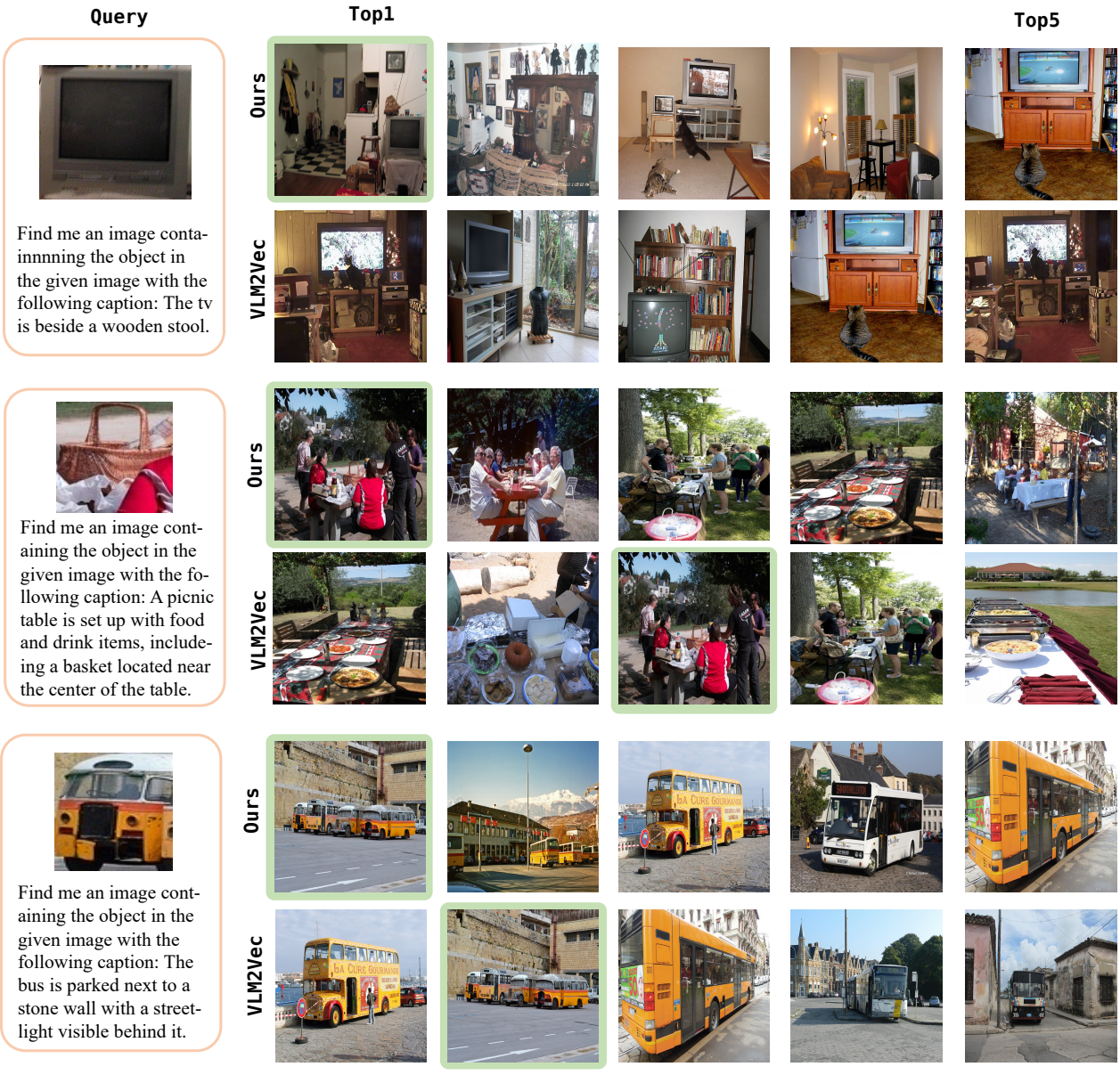

Figure 7. Visualization of the Top-5 results from our model v.s. VLM2Vec on in-domain test data, with the correct target image highlighted in green.

Task	Fine-Tune		Train from MLLM		
	VLM2Vec	MMRet	Ours(8B)	Ours(26B)	Ours(8B w/o. 557K synthetic data)
Classification (10 tasks)					
ImageNet-1K	65.6	58.8	70.5	80.6	70.7
N24News	79.5	71.3	80.2	81.6	79.6
HatefulMemes	67.1	53.7	72.9	72.3	70.5
VOC2007	88.6	85.0	86.1	92.7	87.3
SUN397	72.7	70.0	77.3	78.8	0.78
Place365	42.6	43.0	44.2	38.9	44.0
ImageNet-A	19.3	36.1	39.3	63.6	36.3
ImageNet-R	70.2	71.6	71.6	84.0	72.2
ObjectNet	29.5	55.8	26.2	50.5	32.3
Country-211	13.0	14.7	14.7	20.3	14.7
All Classification	54.8	56.0	58.3	66.33	58.6
VQA (10 tasks)					
OK-VQA	63.2	73.3	68.9	71.0	69.1
A-OKVQA	50.2	56.7	56.6	59.2	56.8
DocVQA	78.4	78.5	73.0	75.1	72.1
InfographicsVQA	40.8	39.3	40.9	44.6	43.3
ChartQA	59.0	41.7	62.9	64.6	61.0
Visual7W	47.7	49.5	52.1	54.9	51.6
ScienceQA	43.4	45.2	52.1	54.7	53.8
VizWiz	39.2	51.7	44.6	47.1	45.5
GQA	60.7	59.0	61.2	71.0	58.8
TextVQA	66.1	79.0	73.5	77.0	74.6
All VQA	54.9	57.4	58.6	61.9	58.7
Retrieval (12 tasks)					
VisDial	73.3	83.0	80.7	81.5	80.5
CIRR	47.8	61.4	54.0	57.6	55.6
VisualNews_{t2i}	67.2	74.2	73.3	78.5	74.0
VisualNews_{i2t}	70.7	78.1	76.9	80.6	76.6
MSCOCO_{t2i}	70.6	78.6	76.9	79.1	75.8
MSCOCO_{i2t}	66.5	72.4	73.7	75.4	73.5
NIGHTS	66.1	68.3	67.9	68.6	67.0
WebQA	88.1	90.2	89.6	89.0	89.2
FashionIQ	12.9	54.9	20.6	21.0	23.7
Wiki-SS-NQ	56.6	24.9	64.0	66.9	65.2
OVEN	47.3	87.5	58.2	67.4	58.3
EDIS	79.9	65.6	88.7	87.6	88.6
All Retrieval	62.3	69.9	68.7	71.1	69.0
Visual Grounding (4 tasks)					
MSCOCO	67.3	76.8	75.2	81.5	71.6
RefCOCO	84.7	89.8	89.3	91.7	86.7
RefCOCO-matching	79.2	90.6	88.1	88.1	86.9
Visual7W-pointing	86.8	77.0	89.9	93.1	84.1
All Visual Grounding	79.5	83.6	85.6	88.6	82.3
Final Score (36 tasks)					
All IND	66.5	59.1	70.5	73.4	70.2
All OOD	52.0	68.0	57.9	63.4	57.9
All	60.1	64.1	64.9	69.2	64.7

Table 9. Performance of each task on MMEB.