MuCo: Multi-turn Contrastive Learning for Multimodal Embedding Model

Supplementary Material

Appendix

We include additional materials in this document.

A. Additional clarification for Tab. 5

Regarding the batch size scaling experiments in Tab. 5 of the main paper, we clarify that the learning rates were scaled according to the batch size: $5e^{-5}$ for batch size 2,048, $1e^{-4}$ for 4,096, and $2e^{-4}$ for batch sizes 7,168 and 8,192.

B. Detailed comparison of pretraining datasets

We include this section to further substantiate the effectiveness and efficiency of our M3T dataset, providing a more detailed analysis than the main paper. As shown in Tab. A, we present a comprehensive statistical and performance comparison with mmE5 and MegaPairs. This comparison highlights M3T's superior efficiency (evident in its simplified IT-2-T structure and zero hard negatives) and its effectiveness (achieving state-of-the-art zeroshot and fine-tuning performance on the MMEB and M-BEIR benchmarks).

Differences in meta-Task composition. The datasets differ in their meta-task composition. Both mmE5 and our M3T include samples for CLS, VQA, and RET, which are core categories from the MMEB benchmark. Notably, M3T features a highly diverse VQA set (25.5M pairs) categorized into global, local, and creative sub-tasks. In contrast, MegaPairs consists solely of retrieval samples, specifically for the IT-2-I modality. This compositional difference is clearly reflected in the zeroshot (ZS) MMEB benchmark performance. As shown in Tab. A, M3T and mmE5, which align with the MMEB benchmark's core categories, significantly outperform MegaPairs in the ZS setting (Overall scores: 58.2 and 55.6 vs. 41.5). Interestingly, this trend shifts after fine-tuning (FT), where MegaPairs (68.7) surpasses mmE5 (68.6), suggesting that Overall FT performance correlates strongly with data scale. Therefore, we scaled M3T to 5M images, achieving the highest performance in both zeroshot and fine-tuning. One notable observation is that despite M3T's large volume of VQA data, it does not lead to disproportionately high VQA performance at the expense of other tasks. Instead, it contributes to a robust improvement across all categories. We attribute this balanced enhancement to our **MuCo** framework, which effectively leverages this task diversity to refine the quality of the initial turn's embedding.

Differences in modality composition and robust generalization. A unique characteristic of M3T is its exclusive reliance on the IT-2-T (Image+Text query to Text target) modality structure, unlike mmE5 or MegaPairs which utilize diverse combinations. Despite this constraint, M3T demonstrates remarkable generalization in the M-BEIR benchmark. In the zeroshot setting, M3T achieves the highest Overall score (37.8), surpassing both mmE5 (37.3) and MegaPairs (35.7). Specifically, M3T secures the best performance in $S \rightarrow S$ (38.2) and notably in $M \rightarrow M$ tasks (48.3). While M3T shows a slightly lower score in the zeroshot $M \rightarrow S$ setting (27.0) compared to MegaPairs (29.6), this gap is effectively bridged after fine-tuning, where M3T achieves comparable performance (45.5 vs. 45.8) and dominates across all other metrics. Crucially, the robustness of our learned representations is highlighted in the global retrieval setting, where the candidate pool includes all datasets combined. As shown in Table A, while the inclusion of massive distractors causes a performance drop across all models, training with M3T consistently outperforms baselines in the global setting (Overall 17.4 vs. 15.8 for mmE5). This suggests that training with M3T learns a globally discriminative embedding space. Even without domain-specific tuning, our model effectively mitigates inter-task interference and maintains semantic separability against distractors from heterogeneous tasks, a capability that is further amplified after fine-tuning (Global Overall 51.6).

Computational Efficiency and Training Statistics. To analyze efficiency from a computational perspective, we list the statistical characteristics observed during actual training in Tab. A. All metrics were measured using the Qwen2-VL-2B model with a batch size of 1,024, specifically breaking down the computational load into query, positive target, and explicit hard-negative batches. A key distinction is that M3T does not utilize explicitly mined hard negatives. Consequently, M3T exhibits a lower total token count compared to mmE5, leading to a reduced time per iteration (8.41s for M3T vs. 15.33s for mmE5). It is crucial to note that the token counts reported for M3T encompass the simultaneous processing of all seven multi-turn pairs, whereas the metrics for other datasets only reflect the processing of single pairs. MegaPairs shows the lowest token counts, GFLOPs, and time per iteration; this is primarily because it employs a fixed image resolution of 512×512, whereas

Table A. **Detailed comparison of pretraining datasets.** We compare our M3T with mmE5 [5] and MegaPairs [44] across dataset statistics, training efficiency, and downstream performance. Metrics marked with † are measured using the **MuCo**-2B model with 1024 batch size on 32 A100 GPUs. For M-BEIR, values in parentheses indicate results evaluated using local candidate pools specific to each of the 16 datasets.

Metric	mmE5	MegaPairs	M3T (ours)
# of Samples			
# of samples	560,000	26,235,105	5,103,183
# of Pairs per Meta-task			
# of **All** pairs	560,000	26,235,105	5,103,183
# of **CLS** pairs	140,000	0	5,103,183
# of **VQA** pairs	140,000	0	25,515,915
# of **RET** pairs	280,000	26,235,105	5,103,183
# of Pairs per modality			
# of T-2-T	0	0	0
# of T-2-I	14,090	0	0
# of T-2-IT	14,081	0	0
# of I-2-T	224,217	0	0
# of I-2-I	27,988	0	0
# of IT-2-T	195,783	0	35,722,281
# of IT-2-I	56,185	26,235,105	0
# of IT-2-IT	27,656	0	0
Batch metrics and training time for 1024 batch			
Avg batch Qry tokens†	1,318,912	447,488	1,252,352
Avg batch Pos tokens†	918,528	432,128	596,992
Avg batch HN tokens †	828,416	361,472	0
GFLOPs per Iteration†	37.5	13.7	18.6
Second per Iteration†	15.33	7.51	8.41
Total Iterations†	547	25,620	4,984
Total Hours†	2.3	53.4	11.6
Zeroshot setting on *MMEB* benchmark			
ZS MMEB (CLS)	51.1	50.4	53.6
ZS MMEB (VQA)	58.8	21.6	59.9
ZS MMEB (RET)	53.4	45.2	55.2
ZS MMEB (GRD)	65.2	57.5	74.6
ZS MMEB (Overall)	55.6	41.5	58.2
Fine-tuning on *MMEB* benchmark			
FT MMEB (CLS)	65.4	65.9	66.2
FT MMEB (VQA)	65.6	64.2	65.6
FT MMEB (RET)	69.5	69.7	70.1
FT MMEB (GRD)	81.1	83.7	85.8
FT MMEB (Overall)	68.6	68.7	69.5
Zeroshot setting on *M-BEIR* benchmark			
ZS M-BEIR ($S \to S$)	12.7(36.8)	11.7(34.6)	15.1(38.2)
ZS M-BEIR ($S \to M$)	21.7(48.6)	14.6(45.9)	21.7(47.5)
ZS M-BEIR ($M \to S$)	9.0(27.8)	16.7(29.6)	11.0(27.0)
ZS M-BEIR ($M \to M$)	35.9(47.0)	32.6(42.1)	35.4(48.3)
ZS M-BEIR (Overall)	15.8(37.3)	15.9(35.7)	17.4(37.8)
Fine-tuning on *M-BEIR* benchmark			
FT M-BEIR ($S \to S$)	47.1(50.6)	48.1(50.6)	49.1(51.4)
FT M-BEIR ($S \to M$)	65.8(66.5)	66.0(66.9)	68.2(68.3)
FT M-BEIR ($M \to S$)	40.9(45.5)	42.2(45.8)	41.6(45.5)
FT M-BEIR ($M \to M$)	61.5(65.8)	62.8(66.8)	64.8(68.7)
FT M-BEIR (Overall)	49.7(53.2)	49.7(53.5)	51.6(54.2)

both M3T and mmE5 support variable resolutions. Regarding total training time for one epoch, MegaPairs takes the longest (53.4 hours) due to its massive scale, followed by M3T (11.6 hours), with mmE5 being the fastest (2.3 hours) due to its smaller dataset size. However, when normalizing for scale, the efficiency of M3T becomes evident. If mmE5 were scaled to match the volume of MegaPairs or M3T, its estimated training time would surge from 2.3 hours to approximately 109 hours and 21.2 hours (nearly double the time required for M3T), respectively. This confirms that the M3T framework is explicitly designed to maximize both performance and computational efficiency.

Table B. **MMEB datasets. Boldface** indicates out-of-distribution evaluation data, while the others represent in-distribution evaluation.

Tasks	Datasets
Classification	ImageNet-1K [8], N24News [38], HatefulMemes [19], VOC2007 [9], SUN397 [41], **ImageNet-A** [15], **ImageNet-R** [14], **Place365** [43], **ObjectNet** [3], **Country-211** [34]
VQA	OK-VQA [30], A-OKVQA [35], DocVQA [32], InfographicsVQA [33], ChartQA [31], Visual7W-telling [45], **ScienceQA** [28], **VizWiz** [12], **TextVQA** [37], **GQA** [17]
Retrieval	VisDial [7], CIRR [27], VisualNews [24], MSCOCO [23], NIGHTS [10], WebQA [4], **FashionIQ** [40], **Wiki-SS-NQ** [29], **OVEN** [16], **EDIS** [25]
Visual Grounding	MSCOCO [23], **RefCOCO** [18], **RefCOCO-matching** [18], **Visual7W-pointing** [45]

Table C. **M-BEIR datasets.** S: Single modality (text or image), M: Multi-modality (text and image).

Tasks	Datasets
$S \to S$	VisualNews(*T2I*) [24], MSCOCO(*T2I*) [23], Fashion-200K(*T2I*) [13], WebQA(*T2T*) [4], VisualNews(*I2T*) [24], MSCOCO(*I2T*) [23], Fashion-200K(*I2T*) [13], NIGHTS(*I2I*) [10]
$S \to M$	EDIS(*T2IT*) [25], WebQA(*T2IT*) [4]
$M \to S$	OVEN(*IT2T*) [16], InfoSeek(*IT2T*) [6], FashionIQ(*IT2I*) [40], CIRR(*IT2I*) [27]
$M \to M$	OVEN(*IT2IT*) [16], InfoSeek(*IT2IT*) [6]

C. Details of benchmark datasets for fine-tuning.

In this work, we evaluate our proposed method and compare it with previous methods using two representative benchmarks for universal multimodal embeddings: MMEB and M-BEIR. As detailed in Tab. B, MMEB comprises four meta-tasks: Classification, Visual Question Answering, Retrieval, and Visual Grounding, consisting of 10, 10, 12, and 4 datasets, respectively. As shown in Tab. C, M-BEIR con-

MLLM (Dense caption)

Dense caption

This interior shot captures a spacious, warm, open-plan living room featuring polished, dark-stained hardwood floors and neutral beige walls that rise to a high, white vaulted ceiling with a wooden ceiling fan and a decorative bird painting on a beam. In the foreground, the back of a large, L-shaped tan leather sofa anchors the space. On the left, a prominent red and brown brick fireplace with a lit fire serves as a focal point, its rustic wooden mantel holding a large round clock with Roman numerals. To the right of the fireplace is a white-framed doorway and a floor lamp. On the far-right wall, a low, light-wood bookshelf is filled with media and topped by a whimsical painting of a dog in an aviator cap, a small "HUG A DOG" sign, and various picture frames. In the background, the open kitchen is brightly lit, showing light wood cabinetry, dark countertops, and a breakfast bar with two stools.

LLM (Multiple pair generator)

Classification

Query: Identify the primary room type and its main features.
Target: A living room with a brick fireplace and an attached open-plan kitchen.

Retrieval

Query: A depiction of the given image.
Target: This interior shot captures a spacious, warm, open-plan living room featuring polished, …

Creative VQA

Query: Based on the lit fire and the overall decor, what is the intended atmosphere of this space?
Target: The space is designed to feel cozy, warm, and inviting, likely for comfortable family gatherings.

Global VQA

Query: Describe the ceiling and floor of the main living area.
Target: The room has a high, white vaulted ceiling with a wooden ceiling fan, and dark-stained hardwood floors

Query: What is the layout between the living room and the kitchen?
Target: It is an open-plan layout, with the living room flowing directly into the kitchen area, which features a breakfast bar.

Local VQA

Query: What is the prominent art piece on the right wall and what is below it?
Target: There is a large, whimsical painting of a dog wearing an aviator cap, and a small sign below it reads 'HUG A DOG'.

Query: Describe the fireplace and the main object on its mantel.
Target: It is a fireplace made of red and brown bricks with a thick wooden mantel, upon which sits a large, round clock.

Figure A. **Overview of the M3T Dataset Synthesis Pipeline.** The data synthesis proceeds in two stages. First, an MLLM (Qwen2.5-VL-75B) is used to generate a comprehensive, objective dense caption of an image's rich visual information. Second, an LLM (OSS-20B) takes the dense caption as input to synthesize a diverse set of text-only query-positive pairs aligned with seven meta-tasks. Specifically, these include one for classification, one retrieval, and five distinct VQA pairs (two global VQA, two local VQA, and one creative VQA). For the retrieval task, the dense caption itself serves as the positive target, which is abbreviated with "..." for clarity in the figure.

sists of 16 datasets. Unlike MMEB, M-BEIR focuses exclusively on retrieval tasks, serving as a specialized benchmark to analyze retrieval performance across diverse modalities.

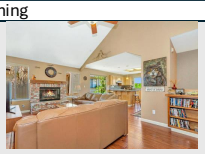
D. Details of synthesizing our M3T dataset

We introduce the **M3T** (Multi-modal multi-turn) dataset, a large-scale dataset designed for pretraining robust multimodal embedding models. Here, we describe the synthesis process of our M3T dataset.

The M3T data synthesis pipeline is intentionally straightforward as depicted in Fig. A. The construction process proceeds in two main steps. The first step involves dense image captioning using a state-of-the-art MLLM, Qwen2.5-VL-75B [2]. For the first step, we randomly sample 5 million images from DataComp [11], selecting only images where at least one spatial dimension is 512 pixels or larger, a threshold established empirically. We observe this is a crucial factor for the MLLM to accurately recognize small-sized objects and perform Optical Character Recognition (OCR) on text within the images during the first step. Guided by a prompt (Fig. B), the MLLM is instructed to produce objective, direct descriptions that identify key objects, their essential attributes, and their spatial relationships. This initial step effectively distills the rich visual information of an image into a dense caption, focusing on observable content.

In the second step, we use the generated dense captions as input to a LLMs, OSS-20B [1], to synthesize a diverse set of query-positive pairs for each image. We design these synthesized pairs to align with the core categories of the MMEB benchmark. Specifically, we generate the following pairs for each image: 1) a query for classifying the image's dominant semantic content, 2) one retrieval query, for which the dense caption serves as the positive target, and 3) five distinct Visual Question Answering (VQA) pairs. The VQA pairs were designed to encourage different reasoning abilities: two pairs require a holistic understanding of the entire scene, two pairs require focusing on localized details, and

Image to text dense captioning



Role:
You are an AI that concisely and accurately describes visual scenes.

Task:
Generate a dense caption that clearly identifies the key objects, figures, and actions in the scene. Describe their essential attributes and their spatial relationships in detail.

Rules:
- Conciseness and Clarity: Use simple, direct sentences. Prioritize key information and avoid overly elaborate or poetic descriptions.
- Direct Description: Describe the scene directly.
- No Self-Reference: Do not refer to the input as an "image," "picture," or "photo." Avoid all phrases like "The image shows," "In this picture," etc.
- Content-Only: Focus exclusively on the visual elements.

Final Command:
Your response must consist only of the descriptive text. Do not explain yourself or output anything else.

Figure B. **Prompt used for dense image captioning with an MLLM.** The prompt is structured into four sections: Role, Task, Rules, and Final command. The core objective is to instruct the MLLM to generate a dense, objective description capturing all key visual information.

Table D. **Additional ablation study of the subsequent turn design for fine-tuning on single-pair dataset.** 'Self-reconstruction template' implies using the initial turn for the subsequent turn. Notably, simple text patterns for the special mask token yield comparable performance. ' ' is 3 space characters.

Setup	MMEB
Self-reconstruction template	68.5
Simple text pattern for masking (' ')	69.4
Simple text pattern for masking ('___')	69.4
Special token for masking (<|mask|>)	69.5

one creative pair demands reasoning beyond literal description. Notably, this entire set of query-positive pairs is synthesized in a single pass using one comprehensive prompt (Fig. C) for the LLM.

This two-stage approach is intentional, as it is designed to create an efficient and extensible pipeline. By isolating the computationally intensive image-to-text conversion as a distinct, one-time preliminary step, the subsequent synthesis of query-positive pairs becomes a flexible and inexpensive text-only operation. Consequently, our 5-million-image dataset yields a total of 35 million query-positive pairs with 5 million images.

Table E. **Various augmented data combinations (Eq. 6).** We report MMEB performance for various data combinations using **MuCo**-2B model. Note that Case 1 represents the baseline result obtained by fine-tuning without applying the **MuCo** strategy for fine-tuning.

Case	(q, p)	(q, p')	(q', p)	(q', p')	MMEB
1	✓	-	-	-	68.1
2	-	✓	-	-	67.7
3	-	-	✓	-	67.9
4	-	-	-	✓	68.0
5	✓	✓	-	-	69.1
6	✓	-	✓	-	68.9
7	✓	-	-	✓	68.9
8	-	✓	✓	-	68.6
9	-	✓	-	✓	68.7
10	-	-	✓	✓	68.7
11	✓	✓	✓	✓	69.5

E. Additional ablation studies

Impact of pretraining and fine-tuning stages. To investigate the individual contributions of the pretraining and fine-tuning stages, we conduct an ablation study reported in Tab. F for MMEB and Tab. G for M-BEIR. Interestingly, our results demonstrate that the **MuCo** fine-tuning strategy alone is highly effective; even without pretraining, it surpasses previous state-of-the-art methods. For instance, on MMEB, **MuCo**-2B (fine-tuning only) achieves 68.4%, outperforming the fully trained B3-2B (68.1%). Similarly, **MuCo**-7B (fine-tuning only) reaches 72.6%, exceeding B3-7B (72.0%). A similar trend is observed on M-BEIR, where our fine-tuning-only models consistently outperform strong baselines like LamRA-Ret. Furthermore, incorporating the pretraining stage with our M3T dataset provides a significant performance boost. This confirms that the multi-turn learning signals are beneficial in both stages. Consequently, the combined effect of leveraging the rich, dialogue-driven context from M3T during pretraining and the adaptive multi-turn reconstruction during fine-tuning maximizes the model's representational power, leading to the most robust performance.

Additional ablation study of the subsequent turn design for fine-tuning on single-pair dataset. We present an additional ablation study on subsequent turn and token design in Tab. D. The 'self-reconstruction template', which utilizes the initial turn for its subsequent turn, achieves a score of 68.5, matching the 'rephrasing template' result in Tab. 8 from the main paper. This suggests that self-reconstruction functions as a semantic refinement process similar to rephrasing. Furthermore, substituting the special mask token with simple text patterns yields comparable performance. This demonstrates that the model can effectively interpret the reconstruction instruction from the prompt context alone during fine-tuning.

Various augmented data combination for fine-tuning.

Table F. **Impact of pretraining and fine-tuning on MMEB.** We report Precision@1 (%) results across four categories: Classification (**CLS**), Visual Question Answering (**VQA**), Retrieval (**RET**), and Visual Grounding (**GRD**). **ID** and **OOD** denote in-distribution and out-of-distribution averages, respectively. Notably, our model using **MuCo** fine-tuning alone already surpasses previous state-of-the-art performance.

Pretraining	Fine-tuning	CLS	VQA	RET	GRD	ID	OOD	Overall
MuCo-2B								
✓	–	53.6	59.9	55.2	74.6	–	–	58.2
–	✓	62.4	64.5	70.2	85.1	72.9	62.2	68.4
✓	✓	66.2	65.6	70.1	85.8	72.9	65.0	69.5
MuCo-7B								
✓	–	56.0	64.7	58.9	75.7	–	–	61.6
–	✓	68.6	70.7	72.0	89.7	76.7	67.6	72.6
✓	✓	68.3	71.9	73.7	90.9	77.3	69.1	73.6

Table G. **Impact of pretraining and fine-tuning on M-BEIR.** We report average Recall (%) results where S and M denote Single modality (text or image) and Multi-modality (text and image), respectively. The arrow ($\rightarrow$) indicates the 'query $\rightarrow$ target' direction.

Pretraining	Fine-tuning	$S \rightarrow S$	$S \rightarrow M$	$M \rightarrow S$	$M \rightarrow M$	Overall
MuCo-2B						
✓	–	15.1	21.7	11.0	35.4	17.4
–	✓	48.5	67.3	41.2	63.5	50.9
✓	✓	49.1	68.2	41.6	64.8	51.6
MuCo-7B						
✓	–	16.7	29.1	11.9	33.7	19.2
–	✓	52.8	70.4	46.3	69.3	55.5
✓	✓	54.0	71.6	47.4	70.4	56.6

We analyze the impact of different pair combinations in Eq. 6. Tab. E shows the result. The baseline (Case 1), which uses only the initial (q, p) pairs, achieves 68.1. Interestingly, using only one of the subsequent pairs (Cases 2-4) results in slightly worse performance (67.7-68.0) than the baseline. This suggests that while the subsequent turn provides accumulative supervision, this signal alone is insufficient to effectively train the initial embedding (which is used for inference). However, performance significantly improves (e.g., 69.1 in Case 5) when the initial (q, p) pair is combined with any augmented pair (Cases 5-7). This validates our hypothesis: while the initial embedding is explicitly trained, the accumulative supervision from the subsequent augmented turns acts as a powerful refiner. As expected, Cases 8-10 perform worse than Cases 5-7 because they omit the crucial training signal for the initial turn. As expected, our full approach (Case 11), which utilizes all four combinations, achieves the highest performance (69.5).

F. Extended main tables

In this section, we present the comprehensive experimental results for the MMEB and M-BEIR benchmarks, expanding upon the summarized findings in the main text. Due to space constraints, the main paper reported only aggregated performance metrics. Here, we provide the full breakdown: Tab. J details the Precision@1 scores for all 36 individual datasets within the MMEB benchmark, covering both In-Distribution and Out-of-Distribution splits. Tab. H presents the detailed retrieval performance for the M-BEIR benchmark, reporting Recall metrics (Recall@5 or Recall@10) for each of the 16 datasets across diverse retrieval tasks. We also report quantitative results (M-BEIR$_{\text{local}}$) evaluated using local candidate pools specific to each of the 16 datasets in Tab. I.

G. M3T examples

We present examples of our dataset in Fig. D and Fig. E.

Table H. **Extended main table on M-BEIR.** The first row specifies the retrieval task, where q and c denote query and candidate, with superscripts t and i representing text and image modalities, respectively. Abbreviations include VN (VisualNews), F200K (Fashion200K), InfoS (InfoSeek), and FIQ (FashionIQ). We report Recall@10 for FashionIQ and Fashion200K, and Recall@5 for all other datasets.

Methods	Size	$q^t \to c^i$			$q^t \to c^t$	$q^t \to (c^i, c^t)$		$q^i \to c^t$			$q^i \to c^i$	$(q^i, q^t) \to c^t$		$(q^i, q^t) \to c^i$		$(q^i, q^t) \to (c^i, c^t)$		Overall
		VN	COCO	F200K	WebQA	EDIS	WebQA	VN	COCO	F200K	NIGHTS	OVEN	InfoS	FIQ	CIRR	OVEN	InfoS	
		R@5	R@5	R@10	R@5	R@5	R@5	R@5	R@5	R@10	R@5	R@5	R@5	R@10	R@5	R@5	R@5	
Zeroshot setting (pretrained) on M-BEIR benchmark																		
CLIP-L [34]	428M	0.0	0.0	0.0	32.1	6.7	5.5	0.0	0.0	0.0	25.3	0.0	0.0	4.4	5.4	24.5	22.1	9.9
SigLIP [42]	652M	0.0	0.0	0.0	34.0	1.1	2.1	0.0	0.0	0.0	28.7	0.0	0.0	4.8	7.1	27.2	24.3	8.1
BLIP [20]	470M	0.0	0.0	0.0	38.1	0.0	0.0	0.0	0.0	0.0	25.1	0.0	0.0	2.2	7.4	10.1	7.9	5.7
BLIP2 [21]	210M	0.0	0.0	0.0	35.2	0.0	0.0	0.0	0.0	0.0	24.0	0.0	0.0	3.9	6.2	13.8	11.4	5.9
MuCo-2B	2B	3.4	6.7	0.2	64.4	28.4	15.0	1.9	17.6	0.2	26.0	7.3	8.6	10.2	18.1	38.0	32.9	17.4
MuCo-7B	7B	4.1	7.6	0.1	72.6	31.3	26.9	2.0	20.1	0.2	26.6	8.2	8.7	10.3	20.5	36.0	31.5	19.2
Fine-tuning on M-BEIR benchmark ($< 7B$ Models)																		
UniIR [39]	428M	42.6	77.9	17.8	84.7	59.4	78.8	42.8	92.3	17.9	32.0	39.2	24.0	24.3	43.9	60.2	44.6	48.9
LamRA-Ret [26]	2B	31.5	70.8	23.1	82.9	54.2	78.3	31.3	88.3	24.5	28.8	45.8	42.4	30.2	45.5	67.0	55.1	50.0
MuCo-2B	2B	32.3	63.3	27.9	89.0	54.5	82.0	33.7	90.8	26.0	29.9	45.0	44.5	29.9	47.1	66.3	63.3	51.6
Fine-tuning on M-BEIR benchmark ($\geq 7B$ Models)																		
MM-Embed [22]	7B	41.0	71.3	17.1	95.9	68.8	85.0	41.3	90.1	18.4	32.4	42.1	42.3	25.7	50.0	64.1	57.7	52.7
LamRA-Ret [26]	7B	41.3	75.4	28.7	85.8	62.5	81.0	39.3	90.4	30.4	32.1	48.4	48.7	33.1	50.5	70.0	60.0	54.9
M3Task-UEM [36]	7B	40.1	82.1	30.7	80.5	67.8	82.0	44.4	93.4	31.0	29.8	51.7	31.9	31.4	52.5	71.4	40.0	53.9
MuCo-7B	7B	41.7	74.0	32.3	91.8	59.7	83.6	39.2	92.3	31.6	29.5	51.1	51.7	33.3	53.4	72.0	68.8	56.6

Table I. **Extended main table on M-BEIR$_{\text{local}}$.** This table presents quantitative results evaluated using **local candidate pools** specific to each of the 16 datasets. The first row specifies the retrieval task, where q and c denote query and candidate, with superscripts t and i representing text and image modalities, respectively. Abbreviations include VN (VisualNews), F200K (Fashion200K), InfoS (InfoSeek), and FIQ (FashionIQ). We report Recall@10 for FashionIQ and Fashion200K, and Recall@5 for all other datasets.

Methods	Size	$q^t \to c^i$			$q^t \to c^t$	$q^t \to (c^i, c^t)$		$q^i \to c^t$			$q^i \to c^i$	$(q^i, q^t) \to c^t$		$(q^i, q^t) \to c^i$		$(q^i, q^t) \to (c^i, c^t)$		Overall
		VN	COCO	F200K	WebQA	EDIS	WebQA	VN	COCO	F200K	NIGHTS	OVEN	InfoS	FIQ	CIRR	OVEN	InfoS	
		R@5	R@5	R@10	R@5	R@5	R@5	R@5	R@5	R@10	R@5	R@5	R@5	R@10	R@5	R@5	R@5	
Zeroshot setting (pretrained) on M-BEIR benchmark																		
CLIP-L [34]	428M	43.3	61.1	6.6	36.2	43.3	45.1	41.3	79.0	7.7	26.1	24.2	20.5	7.0	13.2	38.8	26.4	32.5
SigLIP [42]	652M	30.1	75.7	36.5	39.8	27.0	43.5	30.8	88.2	34.2	28.9	29.7	25.1	14.4	22.7	41.7	27.4	37.2
BLIP [20]	470M	16.4	74.4	15.9	44.9	26.8	20.3	17.2	83.2	19.9	27.4	16.1	10.2	2.3	10.6	27.4	16.6	26.8
BLIP2 [21]	210M	16.7	63.8	14.0	38.6	26.9	24.5	15.0	80.0	14.2	25.4	12.2	5.5	4.4	11.8	27.38	15.8	24.8
MuCo-2B	2B	18.5	66.0	13.8	71.0	46.0	49.0	17.8	79.4	10.9	28.0	31.6	39.3	9.8	27.4	48.2	48.3	37.8
MuCo-7B	7B	18.8	68.6	15.6	79.1	49.7	58.2	18.5	80.9	13.0	29.7	37.6	41.1	12.8	29.1	49.5	49.1	40.7
Fine-tuning on M-BEIR benchmark ($< 7B$ Models)																		
UniIR [39]	428M	42.6	81.1	18.0	84.7	59.4	78.7	43.1	92.3	18.3	32.0	45.5	27.9	24.4	44.6	67.6	48.9	50.6
LamRA-Ret [26]	2B	30.8	79.7	25.1	80.8	54.3	77.8	31.2	88.9	27.1	28.7	51.1	44.2	28.9	47.7	72.3	60.8	51.8
MuCo-2B	2B	33.2	80.3	28.2	89.5	55.3	81.3	32.8	90.4	26.8	30.0	51.9	48.0	30.7	51.4	72.2	65.2	54.2
Fine-tuning on M-BEIR benchmark ($\geq 7B$ Models)																		
LamRA-Ret [26]	7B	41.6	81.5	28.7	86.0	62.6	81.2	39.6	90.6	30.4	30.4	32.1	54.1	52.1	33.1	76.2	63.3	56.6
MuCo-7B	7B	42.2	83.3	33.2	90.9	60.2	83.1	40.6	92.5	32.2	28.6	57.3	55.8	33.6	56.7	76.2	72.3	58.7

Table J. **Extended main table on MMEB.** This table details the results for both baselines and **MuCo** on the MMEB benchmark. The evaluation covers 20 in-distribution and 16 out-of-distribution datasets, where out-of-distribution entries are distinguished by a yellow background. We specifically feature our strongest model **MuCo**-7B

	Zeroshot				Fine-tuning							
	CLIP	**MMRet**	**mmE5**	**MuCo**	**VLM2VEC**	**MMRet**	**mmE5**	**LLaVE**	**UniME**	**B3**	**MoCa**	**MuCo**
Classification (10 tasks)												
ImageNet-1K	55.8	49.1	68.8	62.7	74.5	58.8	77.8	77.1	71.3	84.3	78.0	82.9
N24News	34.7	45.8	54.5	48.1	80.3	71.3	81.7	82.1	79.5	81.6	81.5	82.0
HatefulMemes	51.1	51.0	55.0	58.9	67.9	53.7	64.2	74.3	64.6	64.2	77.6	73.4
VOC2007	50.7	74.6	73.9	67.1	91.5	85.0	91.0	90.3	90.4	89.7	90.0	86.5
SUN397	43.4	60.1	72.7	73.1	75.8	70.0	77.7	79.1	75.9	82.8	76.8	80.7
Place365	28.5	35.3	39.7	40.8	44.0	43.0	43.0	45.0	45.6	47.9	43.0	46.0
ImageNet-A	25.5	31.6	46.1	24.0	43.6	36.1	56.3	51.6	45.5	56.5	52.7	61.8
ImageNet-R	75.6	66.2	86.2	89.5	79.8	71.6	86.3	90.9	78.4	91.9	83.0	91.9
ObjectNet	43.4	49.2	74.8	70.9	39.6	55.8	62.5	46.2	36.4	73.2	45.2	49.9
Country-211	19.2	9.3	35.1	25.2	14.7	14.7	35.4	20.1	18.7	27.8	30.4	28.3
All Classification	42.8	47.2	60.7	62.6	61.2	56.0	67.6	65.7	60.6	70.0	65.8	68.3
VQA (10 tasks)												
OK-VQA	7.5	28.0	56.6	62.6	69.0	73.3	67.6	71.1	68.3	71.6	36.9	72.7
A-OKVQA	3.8	11.6	50.0	53.4	54.4	56.7	56.1	70.8	58.7	59.5	57.1	60.4
DocVQA	4.0	12.6	81.3	92.0	52.0	78.5	90.3	90.3	67.6	94.7	94.3	95.4
InfographicsVQA	4.6	10.6	44.0	68.6	30.7	39.3	56.5	53.5	37.0	68.9	77.2	78.0
ChartQA	1.4	2.4	35.2	50.6	34.8	41.7	50.5	62.2	33.4	59.8	69.8	68.2
Visual7W	4.0	9.0	40.4	48.4	49.8	49.5	51.9	55.8	51.7	55.9	58.5	64.1
ScienceQA	9.4	23.3	47.3	52.9	42.1	45.2	55.8	54.4	40.5	51.7	59.2	57.3
VizWiz	8.2	25.9	54.0	50.1	43.0	51.7	52.8	48.5	42.7	50.6	46.2	55.0
GQA	41.3	41.3	65.4	84.4	61.2	59.0	61.7	68.4	63.6	67.5	71.6	81.4
TextVQA	7.0	18.9	83.1	84.4	62.0	79.0	83.3	79.4	65.2	85.1	75.8	86.4
All VQA	9.1	18.4	55.7	65.0	49.9	57.4	62.6	65.4	52.9	66.5	64.7	71.9
Retrieval (12 tasks)												
VisDial	30.7	62.6	39.1	65.0	80.9	83.0	74.1	83.0	79.7	86.1	84.5	85.3
CIRR	12.6	65.7	41.6	27.6	49.9	61.4	54.7	54.5	52.2	65.8	53.4	54.1
VisualNews_t2i	78.9	45.7	51.2	48.7	75.4	74.2	77.6	76.6	74.8	80.7	78.2	82.9
VisualNews_i2t	79.6	33.4	64.9	60.4	80.0	78.1	83.3	81.2	78.8	84.5	83.1	85.6
MSCOCO_t2i	59.5	68.7	55.0	67.5	75.7	78.6	76.4	78.9	74.9	79.8	79.8	79.3
MSCOCO_i2t	57.7	56.7	59.1	62.3	73.1	72.4	73.2	74.7	73.8	76.7	73.9	77.2
NIGHTS	60.4	59.4	58.9	64.6	65.5	68.3	68.3	67.0	66.2	67.4	66.7	66.6
WebQA	67.5	76.3	82.9	83.7	87.6	90.2	88.0	90.4	89.8	90.4	91.4	89.2
FashionIQ	11.4	31.5	21.6	17.9	16.2	54.9	28.8	23.3	16.5	28.2	28.9	24.7
Wiki-SS-NQ	55.0	25.4	58.8	70.6	60.2	24.9	65.8	63.9	66.6	69.5	82.7	73.5
OVEN	41.1	73.0	67.6	62.3	56.5	87.5	77.5	68.0	55.7	70.6	80.4	74.5
EDIS	81.0	59.9	55.2	76.7	87.8	65.6	83.7	89.1	86.2	88.7	96.9	92.0
All Retrieval	53.0	56.5	54.7	58.9	67.4	69.9	71.0	70.9	67.9	74.1	75.0	73.7
Visual Grounding (4 tasks)												
MSCOCO	33.8	42.7	59.0	53.6	80.6	76.8	53.7	87.0	76.5	74.4	84.6	80.6
RefCOCO	56.9	69.3	78.9	79.3	88.7	89.8	92.7	95.4	89.3	92.9	94.0	94.2
RefCOCO-matching	61.3	63.2	80.8	91.5	84.0	90.6	88.8	92.8	90.6	91.2	95.5	93.6
Visual7W-pointing	55.1	73.5	71.2	78.5	90.9	77.0	92.3	92.5	84.1	80.5	95.3	95.1
All Visual Grounding	51.8	62.2	72.5	75.7	86.1	83.6	89.6	91.9	85.1	84.6	92.4	90.9
Final Score (36 tasks)												
All IND	37.1	43.5	57.2	60.9	67.5	59.1	72.3	64.4	68.4	75.9	74.7	77.3
All OOD	38.7	44.3	60.4	62.4	57.1	68.0	66.7	75.0	57.9	67.1	67.6	69.1
All	37.8	44.0	58.6	61.6	62.9	64.1	69.8	70.3	66.6	72.0	71.5	73.6

Multiple pairs synthesis

Role:
You are an expert data generator for training advanced vision-language models. You meticulously create structured data based on descriptive input.

Task:
Your task is to generate a structured dataset in JSON format based on a provided scene description. The dataset must contain question-answer pairs for three distinct tasks: classification, Visual Question Answering (VQA), and retrieval.
- **Classification**: Generate one question that categorizes the main subject, setting, or event, along with a concise answer (which can be a single word or a short sentence).
- **VQA**: Generate five distinct question-answer pairs. Prioritize questions that explore the relationships, interactions, or comparative attributes between different subjects and objects, rather than simple identification queries (e.g., avoid "What color is the car?"):
    - **Two global** questions and answers that cover the overall context.
    - **Two local** questions and answers that focus on specific, localized details. If the context allows for spatial distinction, these questions must refer to distinct areas (e.g., "on the left", "at the top", "in the middle of").
    - **One creative** question and answer that requires inference or imagination based on the context.
- **Retrieval**: Formulate a concise and simple query whose answer would be a detailed description of the entire scene. Write queries naturally without explicitly using spatial positioning terms like "background" or "foreground."
    - Any **statement-style query** or **question-style query**.

Rules:
- The output must be a single, raw JSON object and nothing else.
- All generated text must be in English.
- Use the exact JSON keys as specified in the format below.
- All answers and negative answers can be a single word or a short sentence.
- **Query Style Variation**: For the "query_cls" and all five "query_vqa_*" fields, you must generate a mix of question-style and statement-style queries. Do not exclusively use one style.
- **Secure Answer**: All queries **must not** contain the answer of it. It should be formulated in a way that someone cannot guess the answer from the query alone, without seeing the context.
- **Information Gap Principle**: Every query, regardless of its style (question or statement), must create an "information gap" that can only be filled by the answer from the context. The query itself should never contain the core information that the answer is supposed to provide.
    - The query ASKS FOR information; the answer PROVIDES it.
    - Avoid the pattern where the query PROVIDES information and the answer simply REPHRASES it.
- **Relevance**: All questions and answers must be factually grounded in and directly derivable from the provided description. Do not invent details not present in the context.

(Cond.) Multiple pairs synthesis

- **Strict Grounding and No Invention Principle**: Every generated element (query, answer) must be strictly and explicitly verifiable from the provided description. Do not infer, assume, or add information that is not explicitly stated.
    - This is especially critical for **spatial** details. If the description does not specify an object's location (e.g., 'on the left', 'at the top', 'in the background'), you must not generate a query that asks for such information or an answer that states it.
- **Clarity**: Frame questions that are clear, specific, and unambiguous, ensuring they are grammatically correct and easily understandable.
- **Diversity**: Vary the sentence structures and vocabulary used across all generated questions and answers. Avoid using repetitive patterns or templates.
- **No Self-Reference**: Do not refer to the context as an "image," "picture," "photo," "scene," or any similar noun. Avoid phrases like "This shows," "What is depicted," or using words like "query" or "documents". Frame all questions and answers from the perspective of a direct observer.
- **Expert-Level Inquiry**: All queries and answers should be formulated to challenge an individual with a graduate-level education or higher.

JSON Format:
You should answer in this JSON format:
```json
{
"query_cls": "",
"answer_cls": "",
"negative_answer_cls": "",
"query_vqa_global1": "",
"answer_vqa_global1": "",
"query_vqa_global2": "",
"answer_vqa_global2": "",
"query_vqa_local1": "",
"answer_vqa_local1": "",
"query_vqa_local2": "",
"answer_vqa_local2": "",
"query_vqa_creative": "",
"answer_vqa_creative": "",
"query_retrieval": "",
}
```

Image description:
{dense_caption_here}

Final command:
Generate the JSON object based on the description provided above. Adhere strictly to all rules. Do not output any text, explanation, or code block formatting before or after the JSON object.

Figure C. **Prompt used for synthesizing multiple query-target pairs with the LLM.** For clarity, the prompt is plotted in two columns. Unlike the prompt for dense image captioning (Fig. B), this prompt is empirically longer and applies more detailed, structured rules to generate the seven distinct pairs from an input dense caption.

Classification

Query: What type of dessert is served on the white plate?
Target: Smoothie.

Global VQA

Query: Which topping appears on top of the pink smoothie?
Target: Crumbled nuts.

Query: Which color, other than white, dominates the surface where the items are placed?
Target: Red.

Local VQA

Query: What color are the cookies positioned to the right of the glass?
Target: Golden-brown

Query: Which fruit in the background is ripe?
Target: Banana

Creative VQA

Query: To add crunch to the smoothie, which item in the setting would serve that purpose?
Target: Golden-brown cookies

Retrieval

Query: Describe the arrangement of food items presented.
Target: A glass filled with a pink smoothie or yogurt-based dessert is placed on a white plate atop a red checkered tablecloth. The smoothie is topped with crumbled nuts and garnished with a dark berry. To the right of the glass, there are two round, golden-brown cookies resting on the tablecloth. In the background, a ripe banana and a small pink bowl containing dark berries are visible, adding to the overall fresh and healthy presentation. The setting appears bright and clean, emphasizing the vibrant colors of the food items.

Classification

Query: What type of jewelry is being described?
Target: Ring.

Global VQA

Query: Explain the shape contrast between the center stone and the surrounding diamonds.
Target: The center square-cut gemstone contrasts with the round diamonds.

Query: State the style of the band on this piece.
Target: It has a sleek, double-banded design.

Local VQA

Query: Identify the feature that secures the central gemstone.
Target: It is held by four prongs.

Query: Specify how the surrounding diamonds are set.
Target: They are set in a pave arrangement.

Creative VQA

Query: What narrative could one infer about the wearer based on the diamond setting?
Target: Perhaps the wearer values sharp elegance, akin to disciplined thoughts, reflected in the square-cut stone.

Retrieval

Query: Provide a detailed description of the jewelry item presented.
Target: A ring featuring a prominent square-cut gemstone at its center, surrounded by a halo of smaller round diamonds. The band is crafted from polished silver or white metal with a sleek, double-banded design. The central gemstone is held securely by four prongs, and the surrounding diamonds are set in a pave arrangement, enhancing the overall brilliance and elegance of the piece. The ring's surface reflects light, emphasizing its polished finish and intricate details.

Classification

Query: Identify the primary type of vegetation present.
Target: Bamboo forest.

Global VQA

Query: Describe the overall vertical structure of the area.
Target: Tall, slender, straight bamboo stalks.

Query: Explain how light interacts with the canopy above.
Target: Filtering sunlight into dappled patterns.

Local VQA

Query: What color variations are observed along the bamboo stalks?
Target: Varying shades of green and gray.

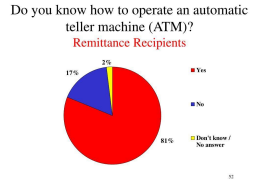
Query: Summarize the appearance of the canopy overhead.
Target: Lush and vibrant.

Creative VQA

Query: If the forest were a musical composition, which instrument would most aptly represent the bamboo stalks?
Target: A high-pitched harp

Retrieval

Query: Detail the forest scene.
Target: A dense bamboo forest is captured from a low-angle perspective, looking upward through the towering, slender trunks of bamboo trees. The bamboo stalks are tall, straight, and segmented, with varying shades of green and gray. The canopy above is lush and vibrant, filtering sunlight into dappled patterns. The scene conveys a sense of height and natural density, emphasizing the verticality of the bamboo forest.



Classification

Query: What is the chart's main topic?
Target: ATM knowledge.

Global VQA

Query: How many respondents indicated they understand ATM operations?
Target: 81%.

Query: What percentage of respondents chose an option other than 'Yes'?
Target: 19%.

Local VQA

Query: Which color segment represents respondents who do not know how to use an ATM?
Target: Blue.

Query: Which segment percentage reflects respondents who answered 'Don't know / No answer'?
Target: 2%.

Creative VQA

Query: If this survey had been conducted 20 years ago when ATMs were less common, how would the red section likely change?
Target: It would be likely smaller than 81%.

Retrieval

Query: What does the chart about ATM usage show?
Target: A pie chart titled "Do you know how to operate an automatic teller machine (ATM)?" is displayed, focusing specifically on "Remittance Recipients." The chart is divided into three segments: a red segment comprising 81% of respondents labeled "Yes," indicating they know how to operate an ATM; a blue segment comprising 17% labeled "No"; and a yellow segment comprising 2% labeled "Don't know / No answer." A legend on the right side of the chart corresponds to these segments using colored squares—red for "Yes," blue for "No," and yellow for "Don't know / No answer." The overall layout is clean and organized, featuring clear labels and percentages for each category, with the slide number "52" located in the bottom right corner.

Figure D. **Examples of our M3T.**

Classification

Query: What product category does this showcase belong to?
Target: Wall stickers

Global VQA

Query: How do the wall stickers interact with the surrounding furniture?
Target: They provide a decorative backdrop that enhances the furniture.

Query: What color palette is chosen for the circular decals across the display?
Target: They feature lime-green, white, dark brown, black, and red hues.

Local VQA

Query: In the upper left setting, how are the decals arranged?
Target: They are installed in a grid pattern.

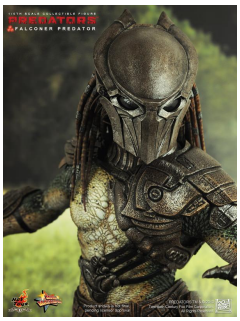
Query: Regarding the close-up, what details are visible in the rings?
Target: Concentric rings in black and red.

Creative VQA

Query: Which sticker design would best complement a blue sofa?
Target: White circular decals.

Retrieval

Query: Describe the decor setup.
Target: The image presents concentric circular wall stickers showcased in four distinct settings: a bedroom with a green wall and lime-green grid decals, a living area with a blue wall and scattered white decals, a minimalist yellow room with dark brown decals, and a close-up detailing a black and red design. It offers extensive customization through a color chart that includes White, Black, Silver, Grey, Charcoal, Cream, Yellow, Gold, Light Brown, Brown, Copper, Burgundy, Lavender, Purple, Violet, Red, Soft Pink, Pink, Turquoise, Bright Green, Racing Green, Pastel Blue, Light Blue, and Dark Blue. The product is available in three size sets, Small (8 x 13 cm), Medium (16 x 13 cm), and Large (28 x 13 cm), with a note indicating that prices vary based on the chosen size, emphasizing versatility for interior decor.



Classification

Query: Identify the main category of the toy.
Target: Predator figure.

Global VQA

Query: Identify the accessory that conveys combat readiness.
Target: Bladed weapon

Query: Which visual trait demonstrates the figure's battle-worn appearance?
Target: Weathered armor plates.

Local VQA

Query: State the feature seen on the chest plate.
Target: Body armor with weathered plates.

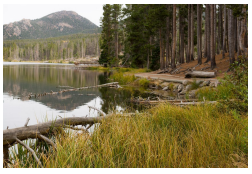
Query: What texture is visible beneath the armor?
Target: Scaly reptilian skin texture.

Creative VQA

Query: If this figure were to engage in a jungle hunt, which environmental element might it prioritize?
Target: Green foliage.

Retrieval

Query: What details can be found in the figure's overall depiction?
Target: This highly detailed 1/6th scale collectible figure depicts the "Falconer Predator" from the Predators franchise, featuring a sleek metallic helmet with angular eye slits, weathered battle-worn armor, and exposed scaly reptilian skin. Posed dynamically in a combat-ready stance with a bladed weapon in hand and utility equipment strapped to its waist and arms, the figure is set against a blurred background of green foliage and earthy tones suggestive of a jungle environment. The lighting accentuates the intricate craftsmanship and realistic textures of the figure, which includes official branding from Hot Toys and Twentieth Century Fox to capture the intense essence of the formidable alien hunter.



Classification

Query: What type of natural feature dominates the horizon?
Target: Mountain.

Global VQA

Query: Describe the overall natural setting portrayed.
Target: A tranquil lake within a forested hillside capped by a mountain under a hazy sky.

Query: Highlight the central pathway leading toward the mountain.
Target: The dirt path winds along the lakeshore, bordered by rocks and logs, flanked by evergreen trees.

Local VQA

Query: What materials frame the water's edge?
Target: Tall grasses and fallen tree branches.

Query: What type of vegetation lines the path toward the background?
Target: Dense evergreen trees.

Creative VQA

Query: If a traveler followed the winding path, what sense of motion might they experience?
Target: A quiet, gentle progression through a secluded, water-bordered landscape toward the mountain.

Retrieval

Query: Summarize the entire wilderness scene.
Target: A serene natural landscape features a calm lake reflecting the surrounding environment. In the foreground, tall grasses and fallen tree branches frame the water's edge. The middle ground includes a dirt path winding along the lakeshore, bordered by scattered rocks and logs. Dense evergreen trees line the path, leading toward the background where a forested hillside rises. Beyond the trees, a mountain with rocky slopes and sparse vegetation dominates the horizon under a hazy sky. The overall scene conveys tranquility and untouched wilderness.

Classification

Query: What is the slide about?
Target: Caffeine beliefs.

Global VQA

Query: What category of items are displayed around the central text?
Target: Caffeine-containing items

Query: What do the two bullet points beneath the title state?
Target: Enhances performance & mood; encourages wakefulness

Local VQA

Query: Which slogan is featured on the Lipton tea box?
Target: Feel good.

Query: What is printed on the NoDoz tablet box?
Target: NoDoz Maximum Strength.

Creative VQA

Query: Why is the pink Nescafé mug positioned prominently?
Target: Because it's named Nescafé, a coffee brand, distinguishing it from energy drinks

Retrieval

Query: What does the slide illustrate about caffeine items?
Target: The image features a slide titled "Common Beliefs & Claims about Caffeine," which presents a symmetrically arranged collection of caffeine-containing beverages and supplements against a clean white background to highlight their colorful packaging. The display includes beverages such as Monster Energy, Red Bull, V8 Plus Energy, Spike energy drink, Shasta Cola, and sparkling water, alongside a Lipton tea box, a Nescafé mug, and a coffee advertisement featuring a woman. These items are complemented by caffeine supplements like NoDoz Maximum Strength, Vivarin, and ProPlus. Centrally located text emphasizes the theme with claims that caffeine "enhances performance & mood" and "encourages wakefulness."

Figure E. **Examples of our M3T.**

References

[1] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

[3] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019. 2

[4] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504, 2022. 2

[5] Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8254–8275, 2025. 2

[6] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *EMNLP*, 2023. 2

[7] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 2

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[9] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 2

[10] Stephanie Fu, Netanel Y Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: learning new dimensions of human visual similarity using synthetic data. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 50742–50768, 2023. 2

[11] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023. 3

[12] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 2

[13] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471, 2017. 2

[14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 2

[15] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 2

[16] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075, 2023. 2

[17] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2

[18] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2

[19] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020. 2

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 6

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 6

[22] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. In *The Thirteenth International Conference on Learning Representations*, 2025. 6

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In

European conference on computer vision, pages 740–755. Springer, 2014. 2

[24] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6761–6771, 2021. 2

[25] Siqi Liu, Weixi Feng, Tsu-Jui Fu, Wenhu Chen, and William Wang. Edis: Entity-driven image search over multimodal web content. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4877–4894, 2023. 2

[26] Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4015–4025, 2025. 6

[27] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2125–2134, 2021. 2

[28] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 2

[29] Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505, 2024. 2

[30] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 2

[31] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022. 2

[32] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 2

[33] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 6

[35] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 2

[36] Rohan Sharma, Changyou Chen, Feng-Ju Chang, Seongjun Yun, Xiaohu Xie, Rui Meng, Dehong Xu, Alejandro Mottini, and Qingjun Cui. Multi-modal multi-task unified embedding model (m3t-uem): A task-adaptive representation learning framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22783–22793, 2025. 6

[37] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 2

[38] Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. N24news: A new dataset for multimodal news classification. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6768–6775, 2022. 2

[39] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer, 2024. 6

[40] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. 2

[41] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 2

[42] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 6

[43] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 2

[44] Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. Megapairs: Massive data synthesis for universal multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19076–19095, 2025. 2

[45] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016. 2