

A DETAILS OF MMEB

In this section, we provide additional details about our proposed benchmark, MMEB (Massive Multimodal Embedding Benchmark). Section A.1 outlines the specifics of the 36 datasets used in the MMEB benchmark. Section A.2 explains the process for determining the number of candidates in MMEB.

A.1 DATASET DETAILS

A.1.1 CLASSIFICATION

There are a total of 10 datasets for classification tasks.

ImageNet-1K (Deng et al., 2009) The dataset is a large-scale dataset commonly used in image classification, consisting of over 1 million images across 1K different classes.

ImageNet-A (Hendrycks et al., 2021b) The dataset contains images from a distribution unlike the ImageNet training distribution. ImageNet-A examples belong to ImageNet classes, but the examples are harder and can cause mistakes across various models. They cause consistent classification mistakes due to scene complications encountered in the long tail of scene configurations and by exploiting classifier blind spots.

ImageNet-R (Hendrycks et al., 2021a) The dataset contains set of images labeled with ImageNet labels obtained by collecting art, cartoons, deviantart, graffiti, embroidery, graphics, origami, paintings, patterns, plastic objects, plush objects, sculptures, sketches, tattoos, toys, and video game renditions of ImageNet classes.

VOC2007 (Everingham et al., 2014) The dataset focuses on recognizing objects in realistic scenarios and contains 20 object classes.

N24News (Wang et al., 2021) The dataset is sourced from the New York Times and consists of 24 categories, with each news article containing both text and image information. The task is to classify the given news image and its accompanying text into one of the 24 categories.

HatefulMemes (Kiela et al., 2020) The dataset proposes a new challenge set for multimodal classification, focusing on detecting hate speech in multimodal memes.

Place365 (Zhou et al., 2017) The dataset is a repository of 10 million scene photographs, labeled with scene semantic categories, comprising a large and diverse list of the types of environments encountered in the world.

SUN397 (Xiao et al., 2010) The dataset is a dataset for scene recognition consisting of 397 categories.

ObjectNet (Barbu et al., 2019) The dataset is a crowd-sourced test set of 50K images featuring objects in unusual poses and cluttered scenes, designed to challenge recognition performance. It includes controls for rotation, background, and viewpoint, and covers 313 object classes.

Country-211 (Radford et al., 2021) The dataset is designed to assess the geolocation capability of visual representations. It filters the YFCC100M dataset to find 211 countries that have at least 300 photos with GPS coordinates.

A.1.2 VISUAL QUESTION ANSWERING (VQA)

There are a total of 10 datasets for VQA tasks.

OK-VQA (Marino et al., 2019) The dataset includes questions that require external resources for answers.

A-OKVQA (Schwenk et al., 2022) The dataset is an augmented successor of OK-VQA, requiring a broad base of commonsense and world knowledge to answer. The questions generally cannot be answered by simply querying a knowledge base, and instead require some form of commonsense reasoning about the scene depicted in the image.

DocVQA (Mathew et al., 2021) The dataset contains questions for document analysis and recognition over document images of various types and content.

InfographicsVQA (Mathew et al., 2022) The dataset comprises a diverse collection of infographics accompanied by natural language question and answer annotations. The questions require methods capable of jointly reasoning over the document layout, textual content, graphical elements, and data visualizations.

ChartQA (Masry et al., 2022) The dataset is designed for question answering about charts, with a focus on visual and logical reasoning applied to real-world charts.

ScienceQA (Lu et al., 2022) The dataset contains questions with diverse science topics and annotations of their answers with corresponding lectures and explanations.

Visual7W-telling (Zhu et al., 2016) The dataset establishes a semantic link between textual descriptions and image regions through object-level grounding. It has two types of questions: “telling” and “pointing”. It leverages the six W questions (what, where, when, who, why, and how) to systematically examine a model’s capability for visual understanding through telling questions. Additionally, a seventh “which” question is appended for visual answers as pointing questions. We use “Visual7W-telling” in our VQA category and “Visual7W-pointing” in our visual grounding category.

VizWiz (Gurari et al., 2018) The dataset originates from a natural visual question answering scenario, where blind individuals captured images and recorded spoken questions about them, along with 10 crowdsourced answers for each visual question. For our task, we select only the answerable questions.

TextVQA (Singh et al., 2019) The dataset is designed to benchmark visual reasoning based on text within images. Models need to read and reason about the text in images to answer related questions.

GQA (Hudson & Manning, 2019) The dataset is designed for real-world visual reasoning and compositional question answering. It uses real images from the Visual Genome dataset. Each image is accompanied by scene graph annotations that describe the classes and attributes of objects in the scene, as well as their pairwise relationships.

A.1.3 RETRIEVAL

There are a total of 12 datasets for retrieval tasks.

VisDial (Das et al., 2017) The dataset features dialogues created by two Amazon Mechanical Turk workers. One worker takes the role of the “questioner”, who only sees the text description of an image, while the other plays the “answerer”, who has access to the image. They engage in a 10-round Q&A session about the image. We repurpose this dataset as a retrieval task, where the goal is to retrieve the image based on the given dialogue.

CIRR (Liu et al., 2021) The dataset is designed for the task of composed image retrieval. It consists of pairs of real-life reference and target images, along with a modification sentence that describes the changes made between the two images.

FashionIQ (Wu et al., 2021) The dataset contains images of fashion products with crowd-sourced descriptions highlighting the differences between these products. Similar to CIRR, FashionIQ can also be used for the task of composed image retrieval, where each test case consists of a pair of reference and target images, along with a modification sentence that describes the changes between the two images.

VisualNews (Liu et al., 2020) The dataset contains publicly available news image paired with captions. We split this task into two setups: “**VisualNews_i2t**”, which retrieves the caption given the news image and “**VisualNews_t2i**”, which retrieves the news image given the caption.

MSCOCO (Lin et al., 2014) The dataset is a well-known image caption dataset. Similar to VisualNews, WE split this task into two setups: “**MSCOCO_i2t**”, which retrieves the caption given the image and “**MSCOCO_t2i**”, which retrieves the image given the caption.

WebQA (Chang et al., 2022) The dataset is a multihop, multimodal QA dataset that requires retrieving a Wikipedia page to answer a given question. We use the Wikipedia page’s image and text descriptions as the candidates for retrieval.

NIGHTS (Fu et al., 2023) The dataset contains human similarity judgments on image pairs that are alike in various ways. The original dataset consists of triplets: a reference image and two perturbed versions, along with human judgments indicating which version is most similar to the reference. Following M-BEIR (Wei et al., 2023), we refactor this dataset into a retrieval task to match pairwise images, where the reference image serves as the query, and the perturbed version that aligns with human judgment is the target.

OVEN (Hu et al., 2023) The dataset contains instances that include an image and a visual recognition text question. Additionally, each instance provides a related Wikipedia image along with its corresponding text description (the Wikipedia title and the first 100 tokens of its summary) as a reference for answering the question, which we treat as the target candidate.

EDIS (Liu et al., 2023) The dataset is a cross-modal image search in the news domain. This dataset contains entity-rich queries, requiring the model to understand both entities and events from the text queries. The candidate consists of the news image and its accompanying headline.

Wiki-SS-NQ (Ma et al., 2024a) The dataset is another retrieval-based VQA dataset. Unlike the original Natural Questions dataset (Kwiatkowski et al., 2019a), which uses a Wikipedia paragraph to answer the question, this dataset leverages Wiki-SS, utilizing Wikipedia page screenshots as the corpus. The screenshot provides more comprehensive information than a plain Wikipedia paragraph.

For **CIRR**, **FashionIQ**, **VisualNews**, **MSCOCO**, **WebQA**, **NIGHTS**, **OVEN** and **EDIS**, we use the processed versions from M-BEIR (Wei et al., 2023).

A.1.4 VISUAL GROUNDING

There are a total of 4 datasets for visual grounding tasks.

MSCOCO (Lin et al., 2014) The dataset includes an object detection task, which involves recognizing an object from a given class in an image. We have repurposed this task into a ranking problem within the MMEB format. The query consists of the image and the object name, while the target is the cropped image of the specified object. We gather distractors from other objects in the same image as well as from different images. We discard test cases where the object is too small.

RefCOCO (Kazemzadeh et al., 2014) The dataset includes an object detection task that requires more reasoning than MSCOCO. Unlike simply identifying the object class, the RefCOCO dataset uses language expressions to refer to specific objects within an image. In our MMEB, we have two tasks related to RefCOCO: “**RefCOCO**” and “**RefCOCO-Matching**”. In “**RefCOCO**”, the query consists of the image and the language expressions referring to a specific object, while the target is the cropped image of that object. In “**RefCOCO-Matching**”, both the query and the target contain the image and the language expressions referring to a specific object, where the two objects are identical.

Visual7W-pointing (Zhu et al., 2016) The dataset establishes a semantic link between textual descriptions and image regions through object-level grounding. It has two types of questions: “telling” and “pointing”. It leverages the six W questions (what, where, when, who, why, and how) to systematically examine a model’s capability for visual understanding through telling questions. Additionally, a seventh “which” question is appended for visual answers as pointing questions. We use “Visual7W-telling” in our VQA category and “Visual7W-pointing” in our visual grounding category.

A.2 SELECTION OF NUMBER OF CANDIDATES

A large number of candidates can make the benchmark more challenging and realistic. However, we also considered the computational cost when designing the benchmark. Choosing an excessively large number of candidates could result in very high inference costs, which may hinder rapid model iteration. As shown in Table 5, we compare the performance of V_{LM2VEC} with different numbers of candidates in the MMEB benchmark. The results show that if the number of candidates is too small, the benchmark becomes saturated quickly. To balance evaluation cost with benchmark difficulty, we selected 1,000 as the optimal number of candidates.

Table 5: We compare the performance of V_{LM2VEC} using different numbers of candidates in MMEB. To balance evaluation cost with benchmark difficulty, we selected 1,000 as the optimal number of candidates.

#Candidates	Meta-Task Average Score				Average Score		
	Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# of datasets →	10	10	12	4	20	16	36
100	54.8	81.8	86.1	89.6	85.2	65.9	76.6
500	54.8	65.9	72.6	82.8	74.6	57.3	66.9
1000	54.8	54.9	62.3	79.5	66.5	52.0	60.1
2000	54.8	50.1	56.7	71.0	62.2	48.0	55.9
5000	54.8	41.3	46.5	65.3	54.5	43.2	49.5

Table 6: The detailed results of the baselines and our V_{LM2VEC} on MMEB, which includes 20 in-distribution datasets and 16 out-of-distribution datasets. The out-of-distribution datasets are highlighted with a yellow background in the table. We include only the best version of V_{LM2VEC} in the table, which uses LLaVA-1.6 as backbone.

	CLIP	OpenCLIP	SigLIP	BLIP2	MagicLens	E5-V	UniIR	V_{LM2VEC}
Classification (10 tasks)								
ImageNet-1K	55.8	63.5	45.4	10.3	48.0	9.6	58.3	74.5
N24News	34.7	38.6	13.9	36.0	33.7	23.4	42.5	80.3
HatefulMemes	51.1	51.7	47.2	49.6	49.0	49.7	56.4	67.9
VOC2007	50.7	52.4	64.3	52.1	51.6	49.9	66.2	91.5
SUN397	43.4	68.8	39.6	34.5	57.0	33.1	63.2	75.8
Place365	28.5	37.8	20.0	21.5	31.5	8.6	36.5	44.0
ImageNet-A	25.5	14.2	42.6	3.2	8.0	2.0	9.8	43.6
ImageNet-R	75.6	83.0	75.0	39.7	70.9	30.8	66.2	79.8
ObjectNet	43.4	51.4	40.3	20.6	31.6	7.5	32.2	39.6
Country-211	19.2	16.8	14.2	2.5	6.2	3.1	11.3	14.7
<i>All Classification</i>	42.8	47.8	40.3	27.0	38.8	21.8	44.3	61.2
VQA (10 tasks)								
OK-VQA	7.5	11.5	2.4	8.7	12.7	8.9	25.4	69.0
A-OKVQA	3.8	3.3	1.5	3.2	2.9	5.9	8.8	54.4
DocVQA	4.0	5.3	4.2	2.6	3.0	1.7	6.2	52.0
InfographicsVQA	4.6	4.6	2.7	2.0	5.9	2.3	4.6	30.7
ChartQA	1.4	1.5	3.0	0.5	0.9	2.4	1.6	34.8
Visual7W	4.0	2.6	1.2	1.3	2.5	5.8	14.5	49.8
ScienceQA	9.4	10.2	7.9	6.8	5.2	3.6	12.8	42.1
VizWiz	8.2	6.6	2.3	4.0	1.7	2.6	24.3	43.0
GQA	41.3	52.5	57.5	9.7	43.5	7.8	48.8	61.2
TextVQA	7.0	10.9	1.0	3.3	4.6	8.2	15.1	62.0
<i>All VQA</i>	9.1	10.9	8.4	4.2	8.3	4.9	16.2	49.9
Retrieval (12 tasks)								
VisDial	30.7	25.4	21.5	18.0	24.8	9.2	42.2	80.9
CIRR	12.6	15.4	15.1	9.8	39.1	6.1	51.3	49.9
VisualNews.t2i	78.9	74.0	51.0	48.1	50.7	13.5	74.3	75.4
VisualNews.i2t	79.6	78.0	52.4	13.5	21.1	8.1	76.8	80.0
MSCOCO.t2i	59.5	63.6	58.3	53.7	54.1	20.7	68.5	75.7
MSCOCO.i2t	57.7	62.1	55.0	20.3	40.0	14.0	72.1	73.1
NIGHTS	60.4	66.1	62.9	56.5	58.1	4.2	66.2	65.5
WebQA	67.5	62.1	58.1	55.4	43.0	17.7	89.6	87.6
FashionIQ	11.4	13.8	20.1	9.3	11.2	2.8	40.2	16.2
Wiki-SS-NQ	55.0	44.6	55.1	28.7	18.7	8.6	12.2	60.2
OVEN	41.1	45.0	56.0	39.5	1.6	5.9	69.4	56.5
EDIS	81.0	77.5	23.6	54.4	62.6	26.8	79.2	87.8
<i>All Retrieval</i>	53.0	52.3	31.6	33.9	35.4	11.5	61.8	67.4
Visual Grounding (4 tasks)								
MSCOCO	33.8	34.5	46.4	28.9	22.1	10.8	46.6	80.6
RefCOCO	56.9	54.2	70.8	47.4	22.8	11.9	67.8	88.7
RefCOCO-matching	61.3	68.3	50.8	59.5	35.6	38.9	62.9	84.0
Visual7W-pointing	55.1	56.3	70.1	52.0	23.4	14.3	71.3	90.9
<i>All Visual Grounding</i>	51.8	53.3	59.5	47.0	26.0	19.0	65.3	86.1
Final Score (36 tasks)								
All	37.8	39.7	34.8	25.2	27.8	13.3	44.7	62.9
All IND	37.1	39.3	32.3	25.3	31.0	14.9	47.1	67.5
All OOD	38.7	40.2	38.0	25.1	23.7	11.5	41.7	57.1

Table 7: Examples of datasets in MMEB (Part 1 of 4). *Instructions* are written in italic font style.











Category	Dataset	Query Text	Query Image	Target Text	Target Image
Classification	ImageNet-1K (Deng et al., 2009)	<i>Represent the given image for classification</i>		Italian greyhound	-
	ImageNet-A (Hendrycks et al., 2021b)	<i>Represent the given image for classification.</i>		sea anemone, anemone	-
	ImageNet-R (Hendrycks et al., 2021a)	<i>Represent the given image for classification.</i>		baseball player	-
	N24News (Wang et al., 2021)	<i>Represent the given news image with the following caption for domain classification.</i> Ms. Goodman styled Amber Valletta with wings for a 1993 shoot by Peter Lindbergh for Harper's Bazaar.		Style	-
	VOC2007 (Everingham et al., 2014)	<i>Identify the object shown in the image.</i>		bus	-
	SUN397 (Xiao et al., 2010)	<i>Identify the scene shown in the image.</i>		firing range indoor	-
	ObjectNet (Barbu et al., 2019)	<i>Identify the object shown in the image.</i>		mug	-
	Country-211 (Radford et al., 2021)	<i>Identify the country depicted in the image.</i>		China	-
	HatefulMemes (Kiela et al., 2020)	<i>Represent the given image for binary classification to determine whether it constitutes hateful speech or not.</i>		No	-
	Place365 (Zhou et al., 2017)	<i>Identify the scene shown in the image.</i>		Airport Terminal	-

Table 8: Examples of datasets in MMEB (Part 2 of 4). *Instructions* are written in italic font style.











Category	Dataset	Query Text	Query Image	Target Text	Target Image
VQA	OK-VQA (Marino et al., 2019)	<i>Represent the given image with the following question.</i> What breed of dog is this?		chihuahua	-
	A-OKVQA (Schwenk et al., 2022)	<i>Represent the given image with the following question.</i> What is the metal basket near the net used to hold?		tennis balls	-
	DocVQA (Mathew et al., 2021)	<i>Represent the given image with the following question.</i> What is name of university?		university of california	-
	InfographicsVQA (Mathew et al., 2022)	<i>Represent the given image with the following question.</i> Which social platform has heavy female audience?		pinterest	-
	ChartQA (Masry et al., 2022)	<i>Represent the given image with the following question.</i> How many food item is shown in the bar graph?		14	-
	ScienceQA (Lu et al., 2022)	<i>Represent the given image with the following question.</i> Which of these states is farthest north?		South Carolina	-
	Visual7W-telling (Zhu et al., 2016)	<i>Represent the given image with the following question.</i> Where is the man sitting?		At the computer	-
	VizWiz (Gurari et al., 2018)	<i>Represent the given image with the following question.</i> Can you tell me what this medicine is please?		night time	-
	GQA (Hudson & Manning, 2019)	<i>Represent the given image with the following question.</i> What is under the utensil on the left?		The napkin is under the utensil.	-
	TextVQA (Singh et al., 2019)	<i>Represent the given image with the following question.</i> What is the brand of this camera?		dakota	-

Table 9: Examples of datasets in MMEB (Part 3 of 4). *Instructions* are written in italic font style.









Category	Dataset	Query Text	Query Image	Target Text	Target Image
Retrieval	VisDial (Das et al., 2017)	<i>Represent the given dialogue about an image, which is used for image retrieval.</i> Q:do you see a lot of people A:just 3 Q:what is the tennis player wearing A:white tennis dress Q:what color is her tennis racket A:black Q:is she wearing a hat A:a visor Q:is she close to the net A:no Q:do you see another player A:no Q:do you see a tennis bag A:no	-	<i>Represent the given image.</i>	
	VisualNews.i2t (Liu et al., 2020)	<i>Retrieve an image of this news caption.</i> US goalkeeper Hope Solo makes a save.	-	<i>Represent the given image.</i>	
	MSCOCO.i2t (Lin et al., 2014)	<i>Find me an everyday image that matches the given caption.</i> Man riding a motor bike on a dirt road on the countryside.	-	<i>Represent the given image.</i>	
	WebQA (Chang et al., 2022)	<i>Find a Wikipedia image-passage pair that answers this question.</i> Do both the Hays County Courthouse in San Marcos, Texas and the Ike Wood House at 227 Mitchell Street in San Marcos, Texas have six columns on their front entrance?	-	<i>Represent the given Wikipedia image with related text information.</i> Hays County Courthouse (2018), San Marcos, TX The Hays County Courthouse in San Marcos, Texas. Listed on the National Register of Historic Places. 227 Mitchell, San Marcos, Texas Ike Wood House at 227 Mitchell Street in San Marcos, Texas.	
	EDIS (Liu et al., 2023)	<i>Find a news image that matches the provided caption.</i> Tom Holland makes his debut in the Spidey suit in Captain America Civil War.	-	<i>Represent the given image with related text information.</i> Comic RiffsJon Favreau is set to reprise his Iron Man role for Spider Man: Homecoming.	
	Wiki-SS-NQ (Ma et al., 2024a)	<i>Find the document screenshot that can answer the given query.</i>	-	<i>Represent the given document screenshot.</i>	
	VisualNews.i2t (Liu et al., 2020)	<i>Find a caption for the news in the given photo.</i>		-	Canadian Prime Minister Stephen Harper shakes hands with President Obama during the North American Leaders Summit in Toluca Mexico in February 2014.
	MSCOCO.i2t (Lin et al., 2014)	<i>Find an image caption describing the given everyday image.</i>		-	A man on a bicycle riding next to a train.

Table 10: Examples of datasets in MMEB (Part 4 of 4). *Instructions* are written in italic font style.

















Category	Dataset	Query Text	Query Image	Target Text	Target Image
Retrieval	CIRR (Liu et al., 2021)	<i>Given an image, find a similar everyday image with the described changes. Show three bottles of soft drink.</i>		<i>Represent the given image.</i>	
	FashionIQ (Wu et al., 2021)	<i>Find an image to match the fashion image and style note. Is shiny and silver with shorter sleeves and fit and flare.</i>		<i>Represent the given image.</i>	
	NIGHTS (Fu et al., 2023)	<i>Find a day-to-day image that looks similar to the provided image.</i>		<i>Represent the given image.</i>	
	OVEN (Hu et al., 2023)	<i>Retrieve a Wikipedia image-description pair that provides evidence for the question of this image. What is the name of this place?</i>		<i>Represent the given Wikipedia image with related text information. Titisee. The Titisee is a lake in the southern Black Forest in Baden-Württemberg. It covers an area of 1.3 (km2) and is an average of 20 (m) deep. It owes its formation to the Feldberg glacier, the moraines of which were formed in the Pleistocene epoch and nowadays form the shores of the lake. The lake's outflow, at 840 (m) above sea level, is the River Gutach, which merges with the Haslach stream below Kappel to form the Wutach. The waters of the Titisee thus drain eventually into the Upper Rhine between Tiengen and Waldshut. On the north shore lies the.</i>	
Grounding	MSCOCO (Lin et al., 2014)	<i>Select the portion of the image that isolates the object of the given label The label of the object is "stop sign".</i>		<i>Represent the given cropped image of the object.</i>	
	Visual7W-Pointing (Zhu et al., 2016)	<i>Select the portion of the image that answers the given question. Which door is behind a person sitting on a bench?</i>		<i>Represent the given cropped image of the object.</i>	
	RefCOCO (Kazemzadeh et al., 2014)	<i>Select the portion of the image that follows the language expressions. man in black coat</i>		<i>Represent the given cropped image of the object.</i>	
	RefCOCO-Matching (Kazemzadeh et al., 2014)	<i>Select the portion of the image that follows the language expressions. kid on right in back, blondish hair</i>		<i>Select the portion of the image that follows the language expressions. top right kid</i>	

Table 11: Zero-shot text-image retrieval performance on Flickr30K. As a general multimodal representation model, V_{LM2VEC} can still achieve competitive T2I (Text-to-Image) and I2T (Image-to-Text) scores when compared to existing CLIP-like models. The baseline numbers are sourced from Sun et al. (2023) and Zhang et al. (2024). We use the best version of V_{LM2VEC} here, which is built upon the LLaVA-1.6 backbone.

Model	image retrieval			text retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
OpenAI CLIP-B/16	62.1	85.6	91.8	81.9	96.2	98.8
Open CLIP-B/16	69.8	90.4	94.6	86.3	97.9	99.4
EVA-02-CLIP-B/16	71.2	91.0	94.7	85.7	96.7	98.9
OpenAI CLIP-L/14	65.2	87.3	92.0	85.2	97.3	99.0
Open CLIP-L/14	75.0	92.5	95.6	88.7	98.4	99.2
EVA-02-CLIP-L/14	77.3	93.6	96.8	89.7	98.6	99.2
MagicLens-B	76.2	93.7	96.5	87.9	97.7	99.5
MagicLens-L	79.7	95.0	97.4	89.6	98.7	99.4
V_{LM2VEC} (LLaVA-1.6)	80.3	95.0	97.4	94.6	99.5	99.8