MATE: Meet At The Embedding - Connecting Images with Long Texts

Anonymous ACL submission

Abstract

While advancements in Vision Language Models (VLMs) have significantly improved the alignment of visual and textual data, these models primarily focus on aligning images with short descriptive captions. This focus limits their ability to handle complex text interactions, particularly with longer texts such as lengthy captions or documents, which have not been extensively explored yet. In this paper, we introduce Meet At The Embedding (MATE), a novel approach that combines the capabilities of VLMs with Large Language Models (LLMs) to overcome this challenge without the need for additional image-long text pairs. Specifically, we replace the text encoder of the VLM with a pretrained LLM-based encoder that excels in understanding long texts. To bridge the gap between VLM and LLM, MATE incorporates a projection module that is trained in a multi-stage manner. It starts by aligning the embeddings from the VLM text encoder with those from the LLM using extensive text pairs. This module is then employed to seamlessly align image embeddings closely with LLM embeddings. We propose two new cross-modal retrieval benchmarks to assess the task of connecting images with long texts (lengthy captions / documents). Extensive experimental results demonstrate that MATE effectively connects images with long texts, uncovering diverse semantic relationships.

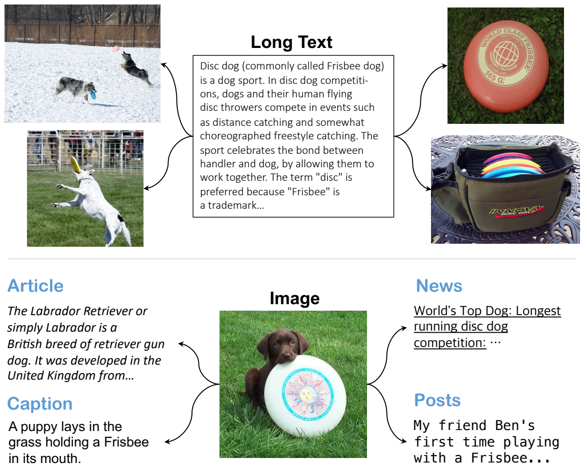


Figure 1: A long text can be linked with different images (above) and an image can be associated with various domains of texts (below). To facilitate these cross-modal interactions, it is essential to establish a robust connection between the embeddings of individual modality samples, while ensuring that both are contextually aligned and semantically rich.

1 Introduction

Recent advancements in Vision Language Models (VLMs) such as CLIP (Radford et al., 2021) and others (Schuhmann et al., 2022; Jia et al., 2021; Li and et al., 2022) have successfully connected visual and textual data by embedding them into a shared space. These models exhibit robust generalization across various visual domains, including medical imaging, art, and remote sensing (Lin et al., 2023; Liu et al., 2023; Conde and Turgutlu, 2021; Hentschel et al., 2022; Singha et al., 2023; Li et al., 2023). The core strength of VLMs stems from leveraging extensive image-caption pairs to obtain generalized and robust representations across diverse visual domains.

Despite their success, most text encoders in current VLMs are primarily designed for direct alignment between short captions and corresponding images. For instance, the text encoder in CLIP has a maximum context length of 77, and this limitation also applies to its longer caption-based variants (Yang et al., 2023; Fan et al., 2024; Zheng et al., 2024). As a result, these encoders struggle to fully comprehend the rich textual context of longer texts, such as captions exceeding 77 tokens or entire documents, that are related to images. Moreover, the reliance on caption-only training samples limits the ability to connect images with texts from various domains. As shown in Figure 1, there are many practical applications in associating images with various long texts which remain largely unexplored, prompting us to investigate this area further.

In this work, we introduce a novel method named *Meet At The Embedding* (MATE), which aligns embeddings to connect images and long texts. MATE leverages a Large Language Model (LLM) and VLMs without requiring additional image-long text pairs. Specifically, MATE aligns image embeddings from a VLM with text embeddings from a pretrained LLM-based encoder (Wang et al., 2023), thereby enhancing image-long text interactions. The LLM-based encoder, trained on diverse text domains, develops a robust understanding of language and advanced reasoning capabilities for handling long texts. We leverage this capability to understand long texts and produce discriminative embeddings for retrieval.

Our MATE model consists of the LLM encoder and the VLM's image encoder, with an additional projection module that converts image embeddings into LLM-aligned embeddings. MATE progressively aligns the VLM embeddings with the LLM embeddings through a multi-stage process: *text-to-LLM alignment* and *image-to-LLM alignment*. In the text-to-LLM alignment stage, we first pre-train the projection module with large-scale captions to align the VLM text encoder with the LLM encoder. Then, we fine-tune the module using query-document pairs (Nguyen et al., 2016) that contain rich textual information, inputting queries to the VLM text encoder and documents to the LLM. In the image-to-LLM alignment stage, we adapt this text-trained module to the VLM image encoder, aligning image embeddings with LLM embeddings using a minimal set of image-caption pairs. This approach effectively connects images with long texts without requiring direct image-long text pairs.

Furthermore, we introduce two new image-long text retrieval evaluation benchmarks: one for images paired with detailed, human-annotated lengthy captions (Onoe et al., 2024) or generative model produced lengthy captions (Zheng et al., 2024), and another for images associated with documents, using pairs sourced from Wikipedia (Chen et al., 2023b; Hu et al., 2023). The results demonstrate that our MATE method effectively links images with long texts and uncovers diverse semantic relationships. This capability enhances intuitive retrieval outcomes and advances our understanding of integrating complex textual and visual information, paving the way for diverse applications, including multi-lingual cases.

We summarize our contributions as:

- To the best of our knowledge, this is the first approach that addresses cross-modal interaction at the image-long text level including documents, establishing a new research topic in the field.
- We introduce the *Meet At The Embedding* (MATE) method, which efficiently aligns VLM and LLM embeddings to facilitate connections between images and long texts.
- With our newly introduced benchmarks, we demonstrate the superior performance of the MATE method in cross-modal retrieval.

2 Related Work

Embedding-based Representation Learning. By mapping given input samples into an embedding space, embedding-based representation learning methods have been actively explored in the fields of language (Su et al., 2023; Wang et al., 2022), vision (Qian et al., 2021; Chen et al., 2020b; Zhang et al., 2022), audio (Jansen et al., 2018) and many others. Various models have achieved significant success by incorporating diverse intra-modality samples at scale across different domains. These models facilitate single-modality and multi-domain representation learning, resulting in enhanced interactions.

On the other hand, VLMs (Radford et al., 2021; Schuhmann et al., 2022; Jia et al., 2021; Li and et al., 2022) have emerged as powerful tools for bridging the modality gap between visual and textual data. These models utilize dual-encoder architectures to encode images and text separately, effectively aligning them within a common embedding space that provides robust representations. However, unlike the diverse images in the VLM training sets, the text component is often limited to short descriptive captions. This limitation may restrict the depth of textual understanding and contextual richness that the models can achieve. Efforts such as (Yang et al., 2023; Fan et al., 2024; Zheng et al., 2024) have been made to mitigate this issue by rewriting captions to be lengthy and informative. Nevertheless, these methods still face limitations because they require a costly captioning process, and the resulting captions are still short, at most 77 tokens. The longer caption-version CLIP (Zhang et al., 2024) was also developed, but it is still limited to 248 tokens, which is insufficient.

Additionally, these models rely solely on image-caption pairs, which lack the capability to incorporate complex reasoning that can be obtained from dense text. In this work, we propose a new efficient approach that connects a powerful LLM-based encoder (Wang et al., 2023) with the VLM image encoder, not only enhancing the textual understanding capability but also enabling robust connections between long texts and images.

Vision Language Cross-Modal Retrieval. The primary application of embedding-based representation learning models is information retrieval, which leverages embeddings to assess the similarity between query and gallery samples. Effective embedding models generate discriminative embeddings by grasping the underlying semantics of data samples, thereby enhancing the accuracy of retrieval results. Many existing methods in image and text retrieval focus on short captions related to images or vice versa, or on composing image queries with brief textual modifications to retrieve related images (Chen et al., 2020a; Li et al., 2019a; Long et al., 2024; Jang and Lim, 2024). We identify a gap in cross-modal retrieval between images and long texts (lengthy captions / documents), where significant potential remains unexplored. To this end, we propose new image and document retrieval experiments involving lengthy captions (Zheng et al., 2024; Onoe et al., 2024) and Wikipedia-style documents (Chen et al., 2023b; Hu et al., 2023). These necessitate a comprehensive understanding of the long texts to accurately match related images from a large-scale database, and our MATE approach achieves the best retrieval results, demonstrating superior performance in understanding complex cross-modal interactions.

3 Method

In this section, we present our MATE method, which aims to establish image-long text alignment by employing a VLM image encoder and a pre-trained LLM-based encoder. It should be noted that MATE does not require additional image-long text pairs for training. The pre-trained CLIP (Schuhmann et al., 2022) and LLM-based E5 (Wang et al., 2023) are utilized as our baseline models. First, we investigate how these models are trained to distribute embeddings (in Section 3.1) to assess the feasibility of connecting these models. Next, we outline the multi-stage training strategy (in Section 3.2) that efficiently achieves our goal.

3.1 Preliminary

Renowned by CLIP, VLM models are trained using a large dataset $\mathcal{D}_v = \{(x_n, t_n)\}_{n=1}^N$ consisting of pairs of images (x_n) and their corresponding captions (t_n). These models utilize an image encoder E_I and a text encoder E_T, which generate the image embedding $\mathbf{v} \in \mathbb{R}^{k_a} : \mathbf{v} = E_I(x)$ and the text embedding $\mathbf{w} \in \mathbb{R}^{k_a} : \mathbf{w} = E_T(t)$, both in the same dimension k_a. All embeddings are typically $l2$-normalized to compute cosine similarity easily.

Then, the InfoNCE loss (also known as a contrastive loss) (Oord et al., 2018) is utilized to update trainable parameters of both modality encoders as:

$$\mathcal{L}_{VLM} = \mathcal{L}_{nce}(\mathbf{v}, \mathbf{w}) + \mathcal{L}_{nce}(\mathbf{w}, \mathbf{v}) \tag{1}$$

where $\mathcal{L}_{nce}$ is computed with the given embedding vectors $\mathbf{x}$ and $\mathbf{y}$ as:

$$\mathcal{L}_{nce} = -\sum_{i=1}^{N_B} \log \frac{\exp\left(\mathbf{x}_i^T \cdot \mathbf{y}_i / \tau\right)}{\sum_{j=1}^{N_B} \exp\left(\mathbf{x}_i^T \cdot \mathbf{y}_j / \tau\right)} \tag{2}$$

for N_B number of image-text pairs with temperature τ. This training objective results in an image and its corresponding caption being aligned, while those that are not paired are distanced.

Similarly, the LLM-based encoder E_5 is also updated using a contrastive approach. Unlike VLM, it utilizes a query (q_n)-document (d_n) paired text-only dataset $\mathcal{D}_l = \{(q_n, d_n)\}_{n=1}^N$, where the query represents relatively shorter text compared to the document. The query embedding $\mathbf{q} \in \mathbb{R}^{k_b} : \mathbf{q} = E_5(q)$ and the document embedding $\mathbf{d} \in \mathbb{R}^{k_b} : \mathbf{d} = E_5(d)$ are obtained with E_5 as k_b-dimensional, $l2$-normalized vectors.

The training loss for the LLM encoder is applied as:

$$\mathcal{L}_{LLM} = \mathcal{L}_{nce}(\mathbf{q}, \mathbf{d}) \tag{3}$$

which leads to embeddings of the query and its corresponding document to be closely aligned, while non-paired instances become distant. Note that both VLM and LLM embedding spaces are developed in a contrastive manner, and are presumed to share some common representations.

3.2 Multi-stage Alignment

When building a connection between the VLM image encoder and the LLM encoder, we could consider utilizing image-long text pairs for training. However, these pairs are scarce due to the complexity of labeling, as defining what constitutes relevant

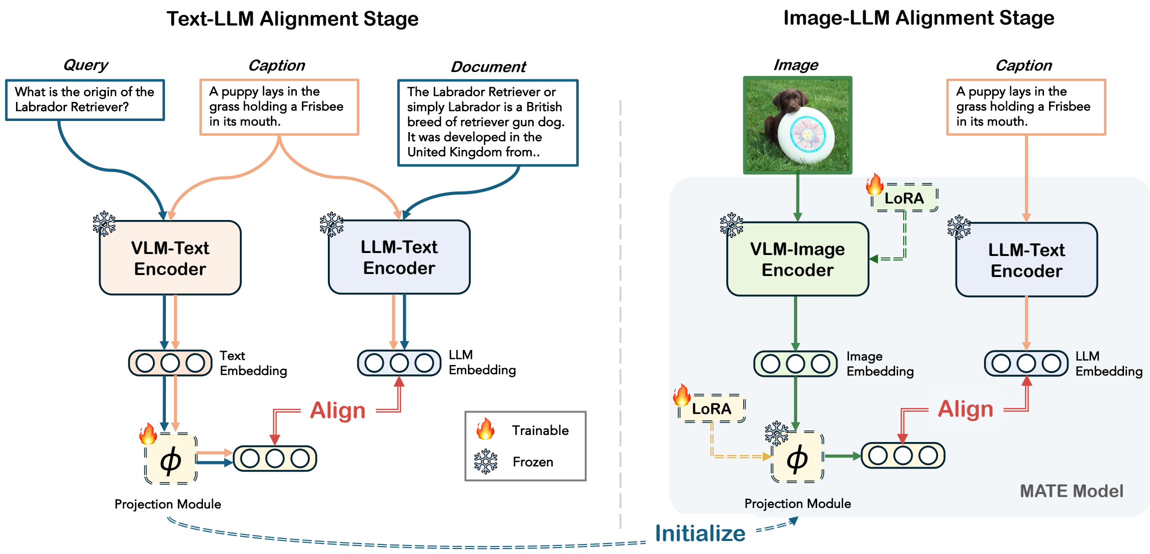


Figure 2: Training pipeline of MATE: Two separate stages are applied with text-only or image-text pairs.

pairs is challenging. Thus, our idea is to train indirectly using existing datasets of image-caption pairs and query-document pairs in a multi-stage manner. This multi-stage approach is beneficial as it allows for incremental learning, where each stage builds upon the knowledge acquired in the previous one, transitioning from query-document (short text-long text) to image-caption. As a result, MATE can perform image-long text retrieval without directly relying on image-long text pairs. We achieve this by first aligning the text encoder of the VLM with the LLM (Section 3.2.1), and then connecting the image encoder of the VLM with the LLM (Section 3.2.2), as shown in Figure 2.

Here, we employ an additional projection module ϕ, due to the differences in dimensionality and representation between VLM and LLM embeddings. This module consists of a few linear layers that project VLM embeddings into the LLM embedding space. Specifically, ϕ takes VLM embeddings as inputs and produces either $\mathbf{u}$ or $\bar{\mathbf{u}}$, where $\mathbf{u} = \phi(\mathbf{v})$ and $\bar{\mathbf{u}} = \phi(\mathbf{w})$. Both $\mathbf{u}$ and $\bar{\mathbf{u}}$ are embedding vectors with the same k_b-dimensionality as the LLM embeddings $\mathbf{d}$.

3.2.1 Text-to-LLM Alignment

First, we pre-train the module ϕ by utilizing the VLM text encoder E_T and the LLM encoder E_5 with a large-scale text-only dataset of captions (t), to reduce the gap between embeddings of VLM and LLM. We train ϕ to align $\bar{\mathbf{u}}$, where $\bar{\mathbf{u}} = \phi(\mathbf{w})$ and $\mathbf{w} = E_T(t)$, with $\bar{\mathbf{d}}$, where $\bar{\mathbf{d}} = E_5(t)$, in a contrastive manner using Equation 3.

Then, we fine-tune ϕ with a text dataset configured with query-document pairs to provide further context of long texts. This process helps ϕ to better understand and align the nuances between related texts, enhancing its ability to accurately match VLM embeddings with the most relevant documents. Similar to the pre-training stage, we utilize E_T and E_5 with the query-document pairs (q, d) to train ϕ to align $\bar{\mathbf{u}}$ and $\bar{\mathbf{d}}$ with Equation 3. We utilize the same number of caption pairs as query-document pairs in a training batch to ensure that ϕ remains robust across diverse captions.

Throughout these processes, we freeze the parameters of E_5 and E_T to preserve the original generalized representation of LLM embeddings and ensure smooth integration with the corresponding VLM image encoder E_I in the subsequent stage.

3.2.2 Image-to-LLM Alignment

With ϕ trained on text-only data in the previous stage, we initialize the parameters of the same architecture ϕ in this stage to transfer dense textual knowledge. Additionally, we apply LoRA (Hu et al., 2021) parameters to both ϕ and E_I to keep the original parameters and train the entire model efficiently. LoRA facilitates fine-tuning by introducing trainable low-rank matrices that adapt the original weights of the model without directly modifying them. This approach helps preserve the original model's capabilities, allowing ϕ to retain its understanding of query-document relationships.

Given a minimal set of image-caption pairs (x, t), we aim to robustly connect image embeddings to LLM embeddings. Specifically, we seek to align $\mathbf{u}$, where $\mathbf{u} = \phi(\mathbf{v})$ and $\mathbf{v} = E_I(x)$, with $\mathbf{d}$,

Dataset	Maximum	Minimum	Average
MSMARCO	807 / 465	9 / 11	81.48 / 90.27
DOCCI-Train	565 / 456	35 / 35	139.27 / 138.86
Oven	1837 / 2136	12 / 15	271.18 / 304.70
Infoseek	1514 / 1788	30 / 33	335.11 / 378.46

Table 1: Token count statistics per image with two different tokenizers: VLM (CLIP) / LLM (Mistral).

where $\mathbf{d} = E_T(t)$. The learning is conducted using the VLM training objective as defined in Equation 1. Ultimately, by utilizing a trained image encoder and projection module with the LLM, MATE can project both image and text into the LLM embedding space. This integration allows for seamless interactions between the visual data represented by VLM image embeddings and the textual data encapsulated in LLM-based representations.

4 Experiments

4.1 Setup

Datasets. For MATE model training, we utilize the datasets as: text-only datasets for Section 3.2.1 include a standard subset of image-caption pairs from the BLIP (Li and et al., 2022) pre-training stage, specifically 16M out of a total of 115M, where only the captions are used for pre-training. We use the 532K query-document pairs from MSMARCO (Nguyen et al., 2016) passage retrieval dataset for fine-tuning. For Section 3.2.2, we use the 585K image-caption pairs from LLaVA-alignment (Liu et al., 2024), which is collected from the CC3M (Sharma et al., 2018) dataset.

To evaluate MATE and other models for the new image-long text cross-modal retrieval tasks, we reconfigure existing image-lengthy caption paired datasets: *DOCCI* (Onoe et al., 2024) and *CC3M-long* (Zheng et al., 2024), and Wikipedia-based image-document paired datasets: *Infoseek* (Chen et al., 2023b) and *Oven* (Hu et al., 2023).

Specifically, DOCCI contains about 1.5K high-resolution images accompanied by human-annotated, detailed descriptive captions. DOCCI is divided into a training set of 9.6K pairs and a test set of 5.1K pairs. We use the test set for image-lengthy caption retrieval experiments. CC3M-long features images and model-generated lengthy captions from three different large multi-modal models (Liu et al., 2024; Chen et al., 2023a; Dai et al., 2024). We use 5K pairs of the Share-GPT4V-generated version for evaluation, ensuring no images overlap with the LLaVA-alignment dataset.

For image-document retrieval tests, we adopt Infoseek (Chen et al., 2023b) and Oven (Hu et al., 2023) datasets provided by (Wei et al., 2023). Both datasets include triplets of images, query text, and document passages. We merge the passages to reconstruct the original lengthy documents. As a result, the Infoseek dataset comprises 1.8K documents with 9.6K related images, averaging 5.3 paired images per document. The Oven dataset includes 3.5K documents with 37.6K related images, averaging 10.7 paired images per document. Examples can be found in Appendix A.

To further investigate whether the length of text in each dataset is sufficient to be defined as long texts, we report token count statistics using the tokenizers from CLIP (Radford et al., 2021) and Mistral (Jiang et al., 2023) in Table 1. The average token counts across all datasets exceed the CLIP text encoder's maximum capacity of 77 tokens.

Evaluation Metrics. Following standards in retrieval evaluation (Radford et al., 2021; Li et al., 2019a; Jang and Lim, 2024), we report image-lengthy caption retrieval results using recall scores at top K (R@K) and employ mean Average Precision (mAP@K) for image-document retrieval to better assess multi-positive connections.

Implementation Details. In this paper, we employ the baseline VLM with CLIP-ViT-G/14 (Cherti et al., 2023), which utilizes Transformer-based image and text encoders. For the LLM-based encoder, we use the instruction-tuned Mistral 7B (Jiang et al., 2023) and the fine-tuned E5 (Wang et al., 2023) model as a baseline with the final embedding dimension of $k_b = 4,096$. Pretrained weights provided by HuggingFace[1] (Wolf et al., 2020) are applied to models as: `laion/CLIP-ViT-bigG-14-laion2B-39B-b160k`, `intfloat/e5-mistral-7b-instruct`. The projection module ϕ comprises three linear layers, each followed by layer normalization and GELU (Hendrycks and Gimpel, 2016) activation. The intermediate hidden dimension of the linear layers is set to four times the dimensionality of the output embedding. We employ additional LoRA (Hu et al., 2021) parameters for the image encoder and ϕ in Section 3.2.2, configured as follows: $\text{LoRA}_\alpha = 16$, rank $= 16$, and dropout $= 0.1$.

For training, we use 8 A100-80GB GPUs for training and evaluation. The AdamW optimizer (Loshchilov and Hutter, 2017) is employed with

[1]https://huggingface.co/models

Type	Method	*Caption Query, Image Gallery*				*Image Query, Caption Gallery*			
		R@1	R@5	R@25	R@50	R@1	R@5	R@25	R@50
Results on DOCCI test									
Zero-shot	CLIP (Cherti et al., 2023)	12.16	27.04	46.96	56.92	16.86	35.49	56.04	65.47
	Long-CLIP (Zhang et al., 2024)	45.24	71.76	89.35	93.75	38.59	69.04	89.88	95.35
	ALIGN (Jia et al., 2021)	62.37	85.31	96.27	98.10	59.88	82.65	94.25	96.61
	BLIP (Li and et al., 2022)	54.10	79.55	93.27	96.22	54.69	80.29	94.33	96.96
	MATE	**73.45**	**93.78**	**98.94**	**99.67**	**62.86**	**87.98**	**97.67**	**99.22**
Fine-tuned on DOCCI Train	ALIGN (Jia et al., 2021)	70.20	90.75	98.06	99.16	67.22	88.47	97.29	98.78
	BLIP-336 (Li and et al., 2022)	79.98	95.80	99.57	99.86	67.06	90.04	98.53	99.49
	MATE-336	81.84	97.16	99.80	99.98	74.35	94.53	99.57	99.86
	MATE-448	**84.55**	**97.80**	**99.88**	**99.98**	**76.55**	**95.82**	**99.67**	**99.90**
Results on CC3M-long test									
Zero-shot	CLIP (Cherti et al., 2023)	3.46	7.54	15.32	19.68	9.96	21.64	38.62	46.16
	Long-CLIP (Zhang et al., 2024)	54.06	75.42	87.66	90.84	51.34	73.46	87.32	90.80
	ALIGN (Jia et al., 2021)	56.80	75.58	86.62	90.24	58.54	76.92	88.18	91.38
	BLIP (Li and et al., 2022)	47.00	67.16	82.26	86.76	58.20	78.64	89.26	91.98
	MATE	**59.54**	**78.50**	**89.72**	**92.92**	**62.24**	**81.00**	**91.10**	**94.08**

Table 2: Image and lengthy caption cross-modal retrieval results on DOCCI test set and CC3M-long test set. The numbers '336' and '448' beside methods denote the image resolutions used for fine-tuning.

a learning rate of 1e-4 and a batch size of 4,096 for the text-to-LLM training stage, and a learning rate of 3e-5 with a batch size of 512 for the image-to-LLM training stage. The temperature τ for the InfoNCE loss is fixed at 0.02, and we iterate the model for 1 epoch for the pre-training stage, and 3 epochs for the fine-tuning stages.

For evaluation, we compare MATE model with four VLMs: CLIP (CLIP-ViT-G/14 (Cherti et al., 2023)) and Long-CLIP (Zhang et al., 2024), both interpolated in their positional encoding to process lengthy texts up to 2,048 tokens, and ALIGN (Jia et al., 2021) and BLIP (Li and et al., 2022), which are based on BERT (Devlin et al., 2018) with a maximum token length of 512. For Long-CLIP, we use the LongCLIP-L model provided by the authors. For ALIGN, we utilize the Huggingface weights from `kakaobrain/align-base`, and for BLIP, we use the official model with ViT-L, pretrained on 129M samples. For MATE, CLIP, and Long-CLIP, we process entire documents, while for ALIGN and BLIP, we truncate documents that exceed 512 tokens due to their token length limitations. We ensure all artifacts used in our paper adhere to their specific licensing terms, permitting research use.

4.2 Results on Image-Lengthy Caption

DOCCI-test. The image-lengthy caption retrieval results on the DOCCI test set are reported in Table 2. We categorize the methods into two groups: zero-shot, which includes the original VLM models and our MATE model, and the fine-tuned version, which is trained on the DOCCI training set images and captions. In the zero-shot scenario, CLIP shows the lowest performance due to its training on shorter captions of less than 77 tokens, while the average token count in the DOCCI dataset is significantly higher. ALIGN achieves better scores than Long-CLIP and BLIP primarily due to its ability to process larger images of width and height of 289 compared to 224 of others, and the fact that the images in the DOCCI dataset are mostly of much higher resolution. Despite using the same CLIP image encoder, our MATE model achieves significantly better retrieval results by successfully leveraging the LLM encoder.

In terms of the fine-tuned case, we train the models using the fine-tuning setup for retrieval proposed in BLIP (Li and et al., 2022). We fine-tune ALIGN with images of width and height of 289 due to its architectural constraints, and utilize larger scale images, 336 or 448, to fine-tune BLIP and MATE to determine whether the models can be improved with more visual information. We observe that all models show improved retrieval scores, with BLIP outperforming ALIGN by processing larger images. Notably, MATE demonstrates a significant performance gain and achieves the best results when the largest images are used. This demonstrates that MATE is effective at leveraging increased visual details for enhanced performance.

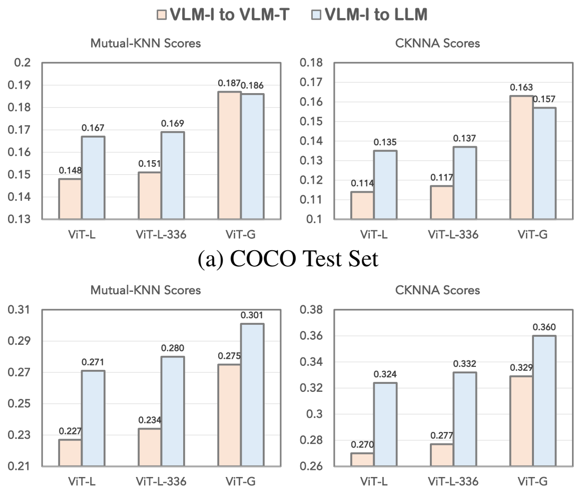
CC3M-long. The experimental results on CC3M-long test set with model-generated captions are presented in Table 2. Similar to the observations in human-annotated captions, our MATE achieves the best retrieval performance. Compared to CLIP, MATE shows an impressive average improvement

Method	Document Query, Image Gallery				Image Query, Document Gallery			
	mAP@5	mAP@10	mAP@25	mAP@50	mAP@5	mAP@10	mAP@25	mAP@50
Results on Infoseek								
CLIP (Cherti et al., 2023)	2.78	3.89	5.25	6.08	15.13	16.13	16.80	17.06
Long-CLIP (Zhang et al., 2024)	10.03	13.46	17.67	19.60	30.60	32.34	33.22	33.49
ALIGN (Jia et al., 2021)	9.06	12.06	15.96	18.01	29.78	31.33	32.22	32.49
BLIP (Li and et al., 2022)	6.23	8.25	11.04	12.42	25.37	26.98	28.03	28.36
MATE	**14.51**	**19.29**	**24.95**	**27.44**	**37.71**	**39.80**	**40.87**	**41.14**
Results on Oven								
CLIP (Cherti et al., 2023)	1.88	2.75	4.19	5.02	13.54	14.39	14.95	15.17
Long-CLIP (Zhang et al., 2024)	4.54	7.12	11.06	13.00	24.85	26.27	27.23	27.53
ALIGN (Jia et al., 2021)	5.72	8.50	12.61	14.69	26.92	28.25	29.08	29.35
BLIP (Li and et al., 2022)	3.44	5.23	8.07	9.58	21.61	22.95	23.88	24.22
MATE	**8.54**	**12.98**	**19.74**	**22.52**	**34.60**	**36.30**	**37.34**	**37.67**

Table 3: Image and document cross-modal retrieval results on Infoseek and Oven datasets.

Model	Image Resolution	Pre-train Data Size	Encoder Model Size	Embedding Dimension (k_a)
ViT-L	224	400M	300M	768
ViT-L-336	336	400M	303M	768
ViT-G	224	2B	1.8B	1280

Table 4: Details of CLIP variants' image encoder.



(a) COCO Test Set

(b) DOCCI Test Set

Figure 3: Measuring alignment between embeddings of VLM image with VLM text (VLM-I to VLM-T), and VLM image with LLM text (VLM-I to LLM). The higher score indicates a closer alignment.

of approximately 60.8 pp across all recall metrics. When compared to the second-best performing model, ALIGN, MATE still exhibits a notable average improvement of around 3.11 pp although MATE uses smaller scale images. These results highlight MATE's robustness and accuracy in capturing exact matches from cross-modal samples, which is crucial as the reliance on generative models grows and the need for effective evaluation mechanisms becomes more pronounced.

4.3 Results on Image-Document

Infoseek. The image-document retrieval results on the Infoseek dataset, as detailed in Table 3, highlight the outstanding performance of the MATE model in both retrieval scenarios. MATE significantly outperforms other models, achieving an average improvement of approximately 17 pp and 23.6 pp over CLIP, and 6.36 pp and 7.47 pp over Long-CLIP, across all evaluated metrics, respectively. This is particularly notable in the challenging environment of matching documents to images and vice versa, where MATE leads with the highest mAP scores across all evaluated metrics. This underscores MATE's advanced effectiveness in navigating and extracting relevant information across different media types, setting a new benchmark for accuracy in cross-modal retrieval tasks.

Oven. More challenging experiments conducted on the Oven dataset, which contains a far more extensive collection of images and documents, are shown in Table 3. The results demonstrate the superior performance of MATE across all metrics compared to other methods. Specifically, MATE significantly outperforms other models, achieving an average improvement of approximately 12.49 pp and 21.97 pp over CLIP, and 5.57 pp and 8.08 pp over ALIGN, across all evaluated metrics, respectively. This highlights MATE's robustness and effectiveness in handling complex cross-modal image-to-document retrieval tasks involving diverse and large-scale gallery samples.

4.4 Further Analysis

Investigation on Choice of Image Encoder. We measure the alignment between three CLIP variants, as detailed in Table 4, and the LLM using the metrics proposed in (Huh et al., 2024), to determine which one is the most feasible for connection. The scores are reported in Figure 3 using the image-short caption pairs from the COCO test set (Lin et al., 2014) and the image-lengthy caption pairs

Configurations	Document Query, Image Gallery				Image Query, Document Gallery			
	mAP@5	mAP@10	mAP@25	mAP@50	mAP@5	mAP@10	mAP@25	mAP@50
(a) Single linear layer w.o. ϕ	9.76	12.92	17.19	19.35	29.03	31.04	32.19	32.51
(b) ϕ w.o. pre-training in 3.2.1	12.54	16.76	21.84	24.21	34.92	37.10	38.18	38.48
(c) ϕ w.o. fine-tuning in 3.2.1	13.36	17.68	22.81	25.23	35.90	37.94	39.07	39.37
(d) Image encoder: ViT-L	13.02	17.11	22.44	24.85	36.23	38.31	39.34	39.64
(e) Image encoder: ViT-L-336	13.06	17.21	22.52	24.95	36.31	38.40	39.46	39.76
(f) More Image-caption pairs	14.41	18.82	24.06	26.34	36.86	39.01	40.05	40.34
(g) With all proposals	**14.51**	**19.29**	**24.95**	**27.44**	**37.71**	**39.80**	**40.87**	**41.14**

Table 5: Ablation study results on Infoseek dataset. 'w.o.' denotes without.

Method	R@1	R@5	R@25	R@50
Chinese Caption Query, Image Gallery				
w/o Fine-tuning on Chinese				
CLIP (Cherti et al., 2023)	0.25	0.93	3.16	5.54
Long-CLIP (Cherti et al., 2023)	0.02	0.11	0.55	1.02
ALIGN (Jia et al., 2021)	0.40	1.36	5.22	8.70
BLIP (Li and et al., 2022)	0.11	0.45	1.91	3.57
MATE	33.64	61.12	84.61	92.91
w/ Fine-tuning on Chinese				
CN-CLIP (Yang et al., 2022)	37.63	64.49	87.65	94.72
Image Query, Chinese Caption Gallery				
w/o Fine-tuning on Chinese				
CLIP (Cherti et al., 2023)	0.76	2.31	7.41	11.82
Long-CLIP (Cherti et al., 2023)	0.02	0.17	0.59	1.12
ALIGN (Jia et al., 2021)	0.93	3.08	9.13	14.37
BLIP (Li and et al., 2022)	0.34	1.25	4.27	7.05
MATE	31.05	57.72	84.59	92.76
w/ Fine-tuning on Chinese				
CN-CLIP (Yang et al., 2022)	36.44	63.07	86.93	94.04

Table 6: Image and Chinese caption cross-modal retrieval results on COCO-CN (Li et al., 2019b) dataset.

from the DOCCI test set. Three key observations emerge from the results. First, larger encoder sizes yield higher alignment scores. Second, lengthy captions result in higher scores. Lastly, and most interestingly, the alignment score of the VLM image to LLM generally exceeds that of the VLM image to VLM text and it is dominant for lengthy captions (DOCCI). Based on these findings, we hypothesize that the LLM encoder shares more common representations with the larger VLM image encoder. Consequently, we select the ViT-G image encoder as our baseline for image-long text connection.

Ablation Study. To validate the proposed schemes of MATE, we perform an ablation study as shown in Table 5. We experiment with configurations (a, b, c) to evaluate the impact of the multi-stage training strategy. For (a), we directly connect the VLM image encoder with the LLM encoder without utilizing ϕ. For (b) and (c), we either remove the pretraining with large-scale captions or omit the fine-tuning with query-document pairs, respectively. The results confirm that combining all training procedures significantly contributes to performance gains. In experiments (d, e), we test different image encoders and find that the choice of ViT-G achieves the best performance. In (f), we increase the number of image-caption pairs utilized in Section 3.2.2 from 0.58M to 3M and observe that the performance is either saturated or slightly degraded, indicating that MATE does not require an excessive number of image-caption pairs to achieve optimal performance. Overall, the optimal performance is achieved when all proposed components are integrated.

Multilingual Capability. We test MATE's cross-modal retrieval with Chinese captions and images from the CN-COCO dataset (Li et al., 2019b), which includes 4.5K pairs. Despite not being trained on image-Chinese caption pairs, MATE shows decent performance and closely matches to Chinese caption-based CN-CLIP (Yang et al., 2022), while other image-English caption-based methods do not perform as well, as shown in Table 6. This success can be attributed to the multilingual capabilities of the LLM encoder, enabling MATE to effectively retrieve relevant content across different languages without specific training, thus highlighting its broad applicability.

5 Conclusion

In this paper, we introduce MATE, a novel method that effectively bridges the gap between images and extensive texts without paired data. MATE integrates a pretrained LLM-based text encoder with a VLM-based image encoder to efficiently align image embeddings with text embeddings. The process begins by aligning VLM text embeddings with LLM embeddings using extensive text pairs, followed by aligning image embeddings with these LLM embeddings. We also introduce new benchmarks to test image-long text retrieval tasks, demonstrating that MATE effectively connects images with extensive texts. This work pioneers a new direction for research in cross-modal interactions.

Limitations

The proposed MATE approach, while innovative in bridging VLMs with LLMs to handle complex text-image interactions, presents certain limitations that warrant further exploration. Primarily, the reliance on a projection module to align embeddings from different models introduces potential challenges in maintaining semantic consistency across modalities, especially when scaling to diverse and extensive datasets. Additionally, the effectiveness of MATE in real-world scenarios where data may not be as cleanly labeled or structured as the datasets used in training remains to be thoroughly evaluated. On the broader impact front, MATE has the potential to significantly enhance the accessibility and interpretability of visual content across various domains, by enabling more nuanced and context-aware image-text associations.

References

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020a. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *CVPR*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *ICML*. PMLR.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023b. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *CVPR*.

Marcos V Conde and Kerem Turgutlu. 2021. Clip-art: Contrastive pre-training for fine-grained art classification. In *CVPR*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2024. Improving clip training with language rewrites. *NeurIPS*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Simon Hentschel, Konstantin Kobs, and Andreas Hotho. 2022. Clip knows image aesthetics. *Frontiers in Artificial Intelligence*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *ICLR*.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *ICCV*.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Young Kyun Jang and Ser-nam Lim. 2024. Towards cross-modal backward-compatible representation learning for vision-language models. *arXiv preprint arXiv:2405.14715*.

Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel PW Ellis, Shawn Hershey, Jiayang Liu, R Channing Moore, and Rif A Saurous. 2018. Unsupervised learning of semantic audio representations. In *ICASSP*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Junnan Li and et al. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019a. Visual semantic reasoning for image-text matching. In *ICCV*.

Xiang Li, Congcong Wen, Yuan Hu, and Nan Zhou. 2023. Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision. *International Journal of Applied Earth Observation and Geoinformation*.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019b. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *NeurIPS*.

Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. 2023. Clip-driven universal model for organ segmentation and tumor detection. In *ICCV*.

Zijun Long, George Killick, Richard McCreadie, and Gerardo Aragon Camarasa. 2024. Multiway-adapter: Adapting multimodal large language models for scalable image-text retrieval. In *ICASSP*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *ICLR*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. 2024. Docci: Descriptions of connected and contrasting images. *arXiv preprint arXiv:2404.19753*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. 2021. Spatiotemporal contrastive video representation learning. In *CVPR*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PMLR.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Mainak Singha, Ankit Jha, Bhupendra Solanki, Shirsha Bose, and Biplab Banerjee. 2023. Applenet: Visual attention parameterized prompt learning for few-shot remote sensing image generalization using clip. In *CVPR*.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *ACL Findings*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2023. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP: system demonstrations*.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. 2023. Alip: Adaptive language-image pre-training with synthetic caption. In *ICCV*.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-clip: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. 2024. Dreamlip: Language-image pre-training with long captions. *arXiv preprint arXiv:2403.17007*.

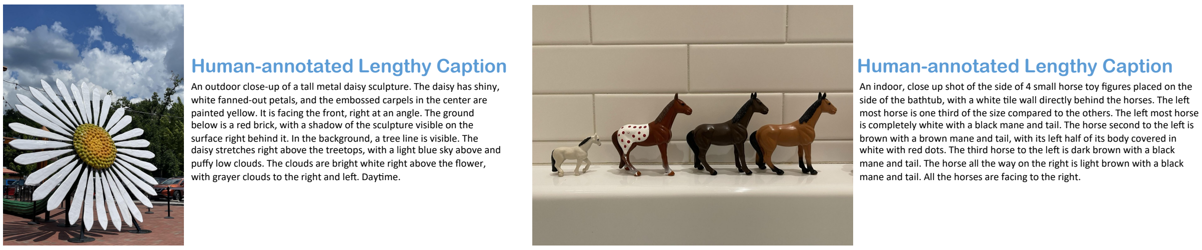


Figure 4: Examples of DOCCI test set of image-human annotated lengthy caption pairs.

Figure 5: Examples of CC3M-long test set of image-generated lengthy caption pairs.

A Appendix

Image-document Examples. We provide examples of configured benchmarks to evaluate MATE and others using image-lengthy caption pairs in Figures 4 and 5. Examples of image-document pairs are shown in Figures 6, 7, 8, and 9.

Figure 6: An example of Infoseek dataset of image-document pair.

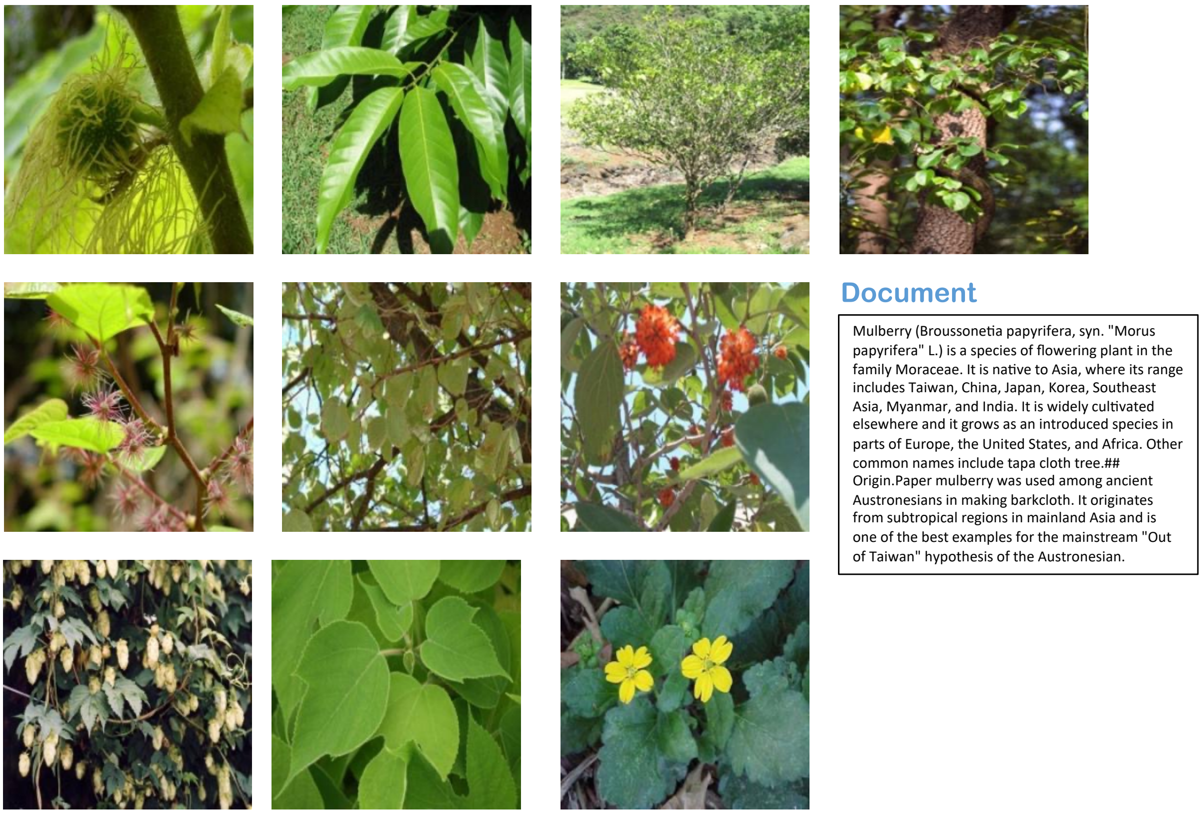


Figure 7: An example of Infoseek dataset of image-document pair.

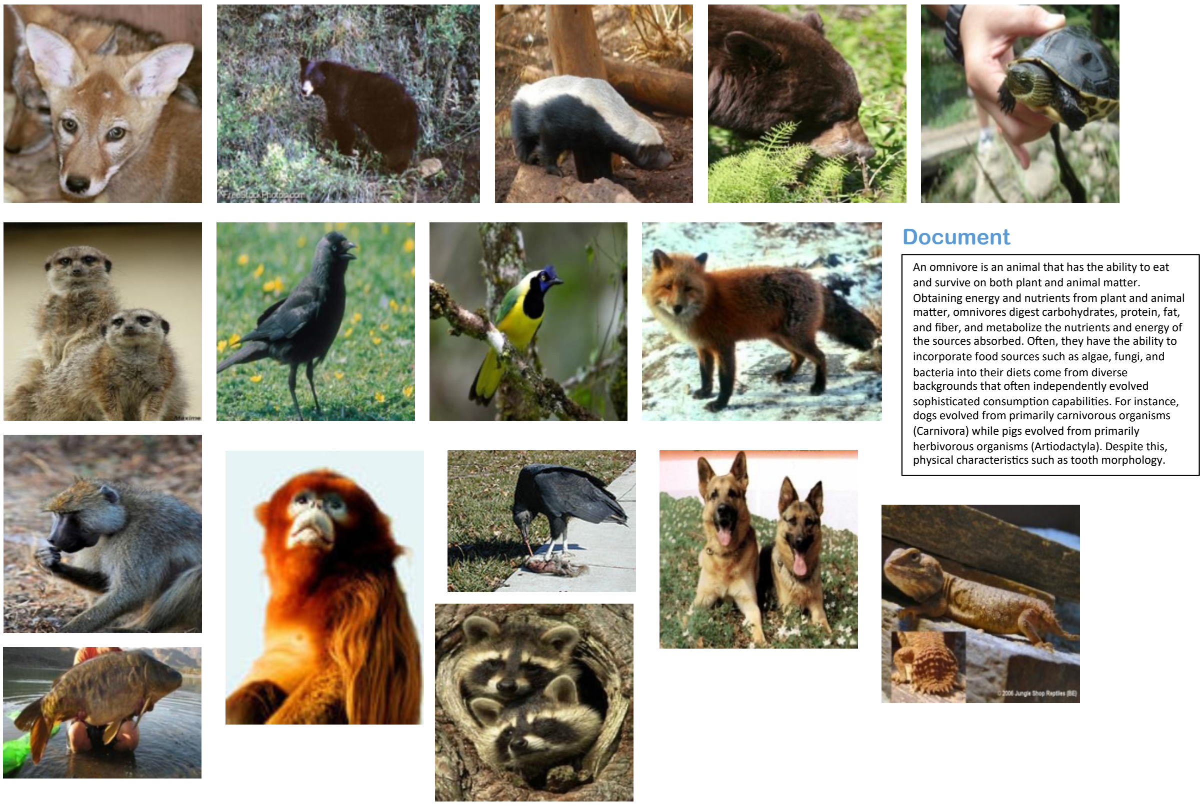


Figure 8: An example of Oven dataset of image-document pair.

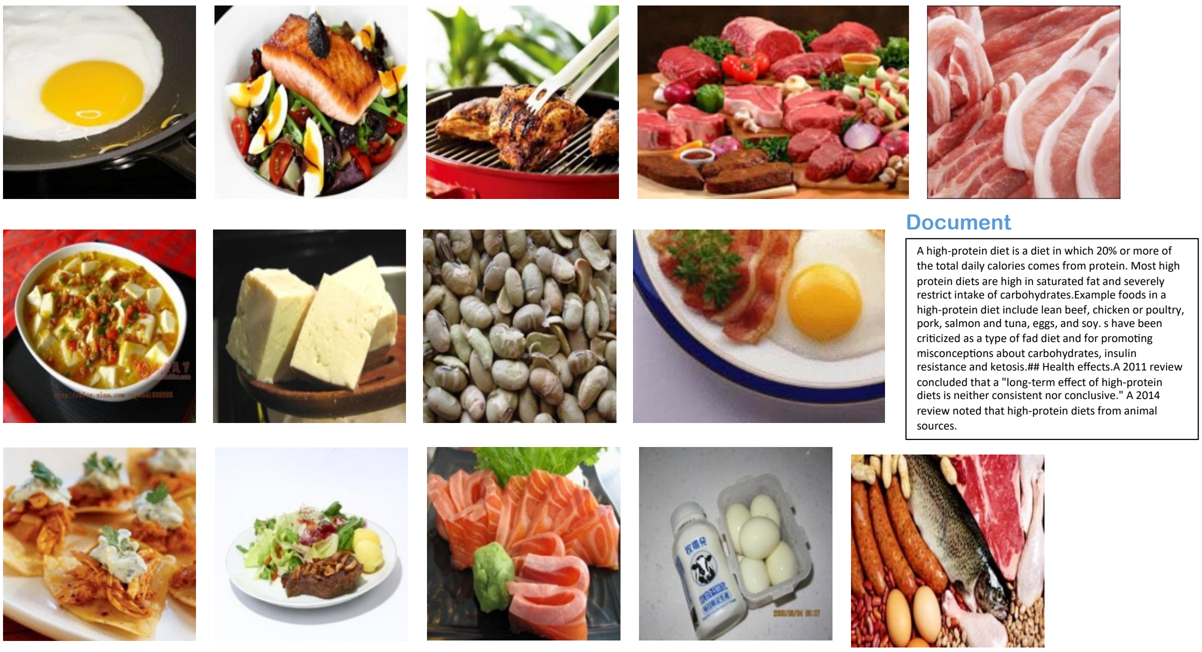


Figure 9: An example of Oven dataset of image-document pair.