Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval

Supplementary Material

In the following, we present further experiments and details about the proposed multimodal retriever ReT.

A. Additional Details on Experimental Setup

A.1. Adding Images to Documents

In the original M2KR benchmark [32], while queries are always text-image pairs, documents entail only textual data (*i.e.*, $\mathrm{d} = d^T$). To enable the training of retrieval models capable of understanding multimodal documents, we enrich the splits of OVEN [16], InfoSeek [6], E-VQA [35], and OKVQA [34] by adding images in their respective documents (*i.e.*, $\mathrm{d} = (d^T, d^V)$). We only include images from the official collections found in literature, using the same knowledge base proposed along the original dataset when possible. These enriched documents, along with our models, will be publicly released. The number of document images collected for each dataset is detailed in Table 6.

OVEN. The dataset features a knowledge base of 6 million Wikipedia pages (Wiki6M), each uniquely identified by a `WikidataID` and potentially accompanied by an image. Text passages in M2KR are assigned a unique identifier in the format `Wiki6M_WikidataID`, allowing us to use the `WikidataID` to associate the corresponding page in Wiki6M. If the page includes an image, it is selected as d^V in the multimodal document.

InfoSeek. The dataset is derived from OVEN and, as such, shares the same knowledge base. Each M2KR document begins with a `WikiTitle`, followed by the textual content. We use the `WikiTitle` as a query to match the corresponding page in Wiki6M and select its image as d^V.

E-VQA. The official knowledge base of E-VQA comprehends 2M Wikipedia articles. Similarly to InfoSeek, for each document, whose ID is in the form `WikiWeb_WikiTitle_WikiSectionID`, we search for the page with the same `WikiTitle` in the 2M collection. However, the knowledge base of E-VQA divides a page into multiple sections, and each section may have a dedicated image. Consequently, given a matching page, we look for the image related to the section identified by `WikiSectionID`. If that image is missing, we consider the first available image on the same page as d^V.

OKVQA. The documents of OKVQA in M2KR are extracted from Wikipedia articles as well, according to [31]. The content of each document is preceded by the title of the article, as in InfoSeek. Given that no other source of images is available, we search into Wiki6M for Wikipedia pages whose title matches the `WikiTitle` found on each document. For a matching page, we attach its image to the corresponding OKVQA document. We highlight a possible drawback concerning document images in OKVQA. In particular, we find that 15% and 4% of them are also found on the documents of OVEN and InfoSeek, respectively. Given that the train split of OKVQA is much smaller compared to those two datasets (see Table 6), it may happen that, during training, those images are more easily associated with textual documents from OVEN and InfoSeek, ignoring OKVQA. A potential confirmation of this hypothesis is that OKVQA is the only dataset where we report superior performance when training without images on the document side (see Table 3, middle).

	Train		Val		Test	
Dataset	*w/ Images*	**Total**	*w/ Images*	**Total**	*w/ Images*	**Total**
OVEN	325,746	339,137	113,723	119,136	3,067	3,192
InfoSeek	647,308	676,441	-	-	79,914	98,276
E-VQA	164,518	167,369	9,722	9,852	3,660	3,750
OKVQA	4,253	9,009	2,415	5,046	2,415	5,046

Table 6. Number of document with images and total number of documents in each split of OVEN, InfoSeek, E-VQA, and OKVQA.

	Text Encoder			Visual Encoder		
Backbone	$\overline{L}$	L	**Layer Indices**	$\overline{L}$	L	**Layer Indices**
CLIP ViT-B	12	12	all	12	12	all
CLIP ViT-L	12	12	all	24	12	0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22
OpenCLIP ViT-H	24	12	1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23	32	12	0, 2, 5, 8, 11, 14, 17, 20, 23, 26, 29, 31
OpenCLIP ViT-G	32	16	0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 31	48	16	2, 5, 8, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47

Table 7. Indices of the layers selected for each backbone encoder to be used as input to the recurrent cell. We denote the original depth of a backbone, measured in number of layers, as $\overline{L}$, and the selected depth for a given configuration of ReT as L.

A.2. Architectural Details

Handling Different Depths of the Backbones. According to the notation defined in Sec. 3.1, in ReT we assume that the textual and visual backbones have the same number of layers L. However, in practice, this condition is rarely satisfied. For instance, our typical configuration is based on CLIP ViT-L, which features 12 layers in the text encoder and 24 in the visual encoder. To address this discrepancy, we opt for selecting an equal number of layers from each unimodal encoder. Whenever possible, we sample the indices of the selected layers with a uniform interval. For

CLIP ViT-L, this approach involves selecting one layer every two from the visual encoder. If a uniform interval is unfeasible, we manually adjust the number of layers. For OpenCLIP ViT-H and OpenCLIP ViT-G, we include 12 and 16 layers respectively from each backbone. Please refer to Table 7 for the exact layer indices selected for each configuration of ReT. With these premises, we highlight that the second and third ablation experiments concerning the *Effect of Recurrence* in Table 3 (bottom) assume that the number of layers for each backbone is already adjusted to the same depth L. That is, when we apply recurrence on the first four layers, it means that the recurrent cell operates on the first, third, fifth, and seventh layers of the visual backbone (*i.e.*, the layer indices are 0, 2, 4, and 6).

Training Data. Our models are trained on the full M2KR dataset, with the exception of MSMARCO, as it does not include images on the query side. Following PreFLMR [32], we adjust the sampling proportions for each dataset. First, because InfoSeek contains many questions related to the same Wikipedia entity, we downsample the training split to obtain a roughly equal number of questions for each entity (around 76k samples). Second, we upsample E-VQA and OKVQA: duplicating E-VQA samples and repeating OKVQA samples nine times per epoch. Notably, duplicating E-VQA does not negatively affect performance on other datasets, nor does it lead to overfitting on its test set. For OKVQA, the upsampling is justified by the limited size of its training split, and we apply the same upsampling factor used in the third training stage of PreFLMR.

Additional Hyperparameters. The learnable gates of ReT can be adjusted by setting the forget and input gate bias, *i.e.*, b_f and b_i respectively (cf. Eq. 4). A common choice is to choose b_f equal to 1 and b_i equal to -1, which has the effect of biasing the model toward remembering the past. However, we empirically found no benefit in such a choice, as ReT natively learns to keep information from its past, that is, to retain information from shallower layers. So, we set both the forget and input gate bias to zero. Following prior work [17], the weights of the learnable matrices $\boldsymbol{W}_F^T$, $\boldsymbol{W}_f^V$, $\boldsymbol{W}_i^V$, and $\boldsymbol{W}_i^V$ in Eq. 4, that modulate the gates, are initially drawn from a truncated normal distribution.

Efficient Indexing for Inference. For evaluating our models, we index the multimodal document collections of the target benchmark using PLAID [41], which supports the fine-grained late-interaction paradigm. We refer readers to the original paper for detailed information about the indexing process. The only exception is when using CLIP for indexing the knowledge base for retrieval-augmented VQA in Table 5. Since CLIP generates a single embedding for both queries and documents, we build a Faiss index [20] to enable efficient dot-product similarity searches.

		InfoSeek (top-1)			InfoSeek (top-3)		
MLLM	**Retriever**	Un-Q	Un-E	All	Un-Q	Un-E	All
LLaVA-v1.5	-	6.9	7.3	7.1	6.9	7.3	7.1
LLaVA-v1.5	CLIP	18.6	17.6	18.1	21.0	20.1	20.6
LLaVA-v1.5	PreFLMR	17.4	15.8	16.6	19.3	17.4	18.3
LLaVA-v1.5	**ReT (Ours)**	**24.1**	**18.1**	**20.7**	**28.1**	**21.1**	**24.1**
LLaVA-MORE	-	7.3	7.4	7.4	7.3	7.4	7.4
LLaVA-MORE	CLIP	16.9	16.1	16.5	19.9	18.7	19.3
LLaVA-MORE	PreFLMR	17.1	15.4	16.2	19.2	17.2	18.1
LLaVA-MORE	**ReT (Ours)**	**23.8**	**16.8**	**19.7**	**28.5**	**20.3**	**23.8**

Table 8. Retrieval-augmented generation results on InfoSeek when using different information retrievers.

A.3. Retrieval-Augmented VQA

In the paper, we evaluate ReT for retrieval-augmented VQA on InfoSeek, which requires answering visual questions related to Wikipedia entities. The dataset has two subsets: unseen question and unseen entity, which entail either questions or entities not found in the training set. Answers are found on the entity's Wikipedia page within the 6M knowledge base [16] used by InfoSeek and OVEN in M2KR. Given the difficulty of the task, retrieval is essential.

Controlled Knowledge Base. To build our experimental knowledge base, we select from the original 6M collection all the Wikipedia pages whose entity is referenced in a question from the validation split. To expand it, we randomly sample additional entities, totalling 50K Wikipedia entities. Each Wikipedia page is chunked into shorter text passages of approximately 100 words [21]. Following M2KR, we prepend the Wikipedia title of the corresponding page to each passage, ending up with the format: `title: WikiTitle content: [...]`. If the page includes an image, we attach it to every passage obtained from the same page, so that it can be used by ReT as the document image d^V. Our knowledge base includes 525,177 passages and will be open-sourced to support future research.

Retrieval Settings. We index our controlled knowledge base with either CLIP, PreFLMR, or ReT. Note that ReT is the only model that can leverage document images. On the query side, we prepend the textual query of a visual question from InfoSeek with a randomly chosen instruction, according to the M2KR format [1]. When embedding queries with CLIP, we ignore the question and only process the image with the visual encoder. To generate answers, we include the top-k retrieved textual passages in the prompt of the chosen answer generator (*e.g.*, a pre-trained LLM such as LLaMA-3). We experiment with k equal to either 1 or 3. The prompt used is the following:

[1] For instance:
`Using the provided image, obtain documents that address the subsequent question: {question}.`

Model	WIT R@10	IGLUE R@1	KVQA R@5	OVEN R@5	LLaVA R@1	InfoSeek R@5	InfoSeek PR@5	E-VQA R@5	E-VQA PR@5	OKVQA R@5	OKVQA PR@5
Comparison with Naive Multi-Layer Feature Fusion											
CLIP (Multi-Layer Feature Fusion)	59.9	72.1	38.4	65.8	59.9	30.2	48.6	21.4	42.2	10.4	59.5
ReT (Ours)	**73.4**	**81.8**	**63.5**	**82.0**	**79.9**	**47.0**	**60.5**	**44.5**	**57.9**	**20.2**	**66.2**
Effect of Recurrence on Query and Document Sides											
w/o recurrence (query side)	**75.4**	81.6	58.8	79.6	75.9	33.3	52.1	35.1	50.9	15.4	60.3
w/o recurrence (document side)	74.0	79.4	62.7	80.2	76.8	**50.0**	**61.9**	35.5	51.6	14.9	63.9
ReT (Ours)	73.4	**81.8**	**63.5**	**82.0**	**79.9**	47.0	60.5	**44.5**	**57.9**	**20.2**	**66.2**
Effect of Text Encoder											
w/ ColBERTv2	**73.9**	79.3	48.6	79.6	79.6	40.0	58.9	43.4	**59.0**	19.0	64.1
ReT (Ours)	73.4	**81.8**	**63.5**	**82.0**	**79.9**	**47.0**	**60.5**	**44.5**	57.9	**20.2**	**66.2**

Table 9. Additional ablation study results. All experiments are with CLIP ViT-L for both visual and textual encoders.

```
Answer the question given the following
context.
Context:
{C1}
##
{...}
##
{Ck}
##
Question: {question}
If the context does not help with the
question, try to shortly answer it
anyway.
Do not answer anything else.
Short answer:
```

We omit the *system* preamble because it is specific to the given LLM. Additionally, we augment the prompt with 3-shot examples, one for each question type that can be found in InfoSeek [6]: string, numerical, and date. The examples are randomly drawn from the training set.

Experiments with Multimodal LLMs. Table 8 shows additional retrieval-augmented VQA results on InfoSeek. Differently from Table 5, here we use a Multimodal LLM (MLLM) [4] that can *see* the query image as the answer generator. In particular, we employ the popular LLaVA-v1.5 [33], based on Vicuna-7B, and the more recent LLaVA-MORE[2], which is based on LLaMA-3.1-8B [11]. In general, MLLMs achieve better performance than LLMs, even though also in this context removing retrieval is severely detrimental. ReT confirms itself as the best information retriever, outscoring PreFLMR by as many as 5.8 points in the top-3 settings when paired with LLaVA-v1.5.

B. Additional Experimental Results

Comparison with Naive Multi-Layer Feature Fusion. The intuition of combining features from multiple layers is, in principle, independent from the proposed recurrent cell architecture behind ReT. We thus propose a simpler baseline, that first collects the features from each layer of the CLIP textual and visual encoder. Then, the features from different layers are averaged and transformed with a learnable linear projection, obtaining a sequence of textual and visual tokens. These final representations are fused together via cross-attention, resulting in a multimodal sequence of tokens for fine-grained late-interaction retrieval. We report the results of this experiment in Table 9 (top). According to that, it can be concluded that fusing features from multiple layers by simply taking the mean is not competitive compared to the recurrent approach proposed by ReT.

Effect of Recurrence on Query and Document Sides. From a qualitative analysis of the forget and input gate activations of ReT (see Fig. 4), we observe that the learned gates behave more uniformly across different layers when encoding multimodal documents. On the contrary, when processing multimodal queries, the gates experience larger fluctuations. We thus raise the question of whether the flexibility of the proposed multi-layer feature fusion strategy is necessary also to encode multimodal documents. To address this point, we train a variant of ReT, where we replace the recurrent cell in the document encoder with a cross-attention layer that is applied only to the features from the last layer of the pre-trained backbones. This is equivalent to adopting the proposed architecture of ReT to encode queries, and the CLIP (Feature Fusion) architecture proposed in Sec. 4 as a baseline to process documents. For completeness, we also try the opposite, *i.e.*, training CLIP (Feature Fusion) on the query side and ReT on the documents. According to Table 9 (middle), with the exception of WIT, replacing the recurrent cell with cross-attention on the query side seriously degrades performance. Doing that on the document side results in a noticeable improvement on InfoSeek but, in general, we still observe a gap in favor of the original ReT model, where two different recurrent cells independently work on both sides.

Effect of Text Encoder. We also train a version of ReT where we replace the CLIP textual encoder with ColBERTv2 [42], which has a larger context length of up to

[2] https://github.com/aimagelab/LLaVA-MORE

512 tokens, compared to the 77 tokens of CLIP. As shown in Table 9 (bottom), this choice translates into a minor improvement of $+0.5$ R@10 points on WIT, and $+1.1$ PR@5 points on E-VQA. However, by sticking with the CLIP textual encoder, ReT enjoys more robust performance across the entire M2KR benchmark, picking with a gain of $+14.9$ and $+7.0$ R@5 points on KVQA and InfoSeek.

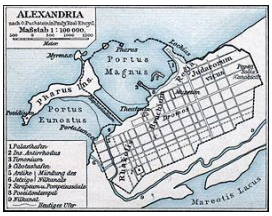
C. Qualitative Results and Visualizations

Results on M2KR. In Fig. 5 we present a qualitative comparison between PreFLMR and ReT. To ensure a direct evaluation with PreFLMR, we focus exclusively on datasets from the M2KR benchmark that exclude document images. As it can be seen, ReT consistently retrieves more contextually accurate and detailed information to address the given queries. In contrast, Fig. 6 reports a qualitative analysis of the difference between PreFLMR and ReT, where the latter can exploit document images, when available. The examples demonstrate that accessing document images significantly improves the ability of the model to answer questions accurately, leveraging the visual context to complement textual information.

Qualitative Analysis of Gate Activations. In Fig. 7 we show additional activation patterns of our learnable sigmoid gates within the recurrent cell of ReT. On the left, we show gate behavior for a single query-document encoding, while on the right, we show average activations over 2k samples from the OVEN (top), E-VQA (middle), and OKVQA (bottom) split of M2KR. When encoding queries, the behavior of the gates shows a balanced contribution from both the visual and textual components, with the forget gate never dropping to zero, indicating that both modalities consistently provide useful information. Interestingly, the impact of each layer varies significantly, with noticeable fluctuations in gate values across the layers, highlighting the dynamic interaction between modalities throughout the encoding process. Notably, on the OVEN dataset, the contribution from visual information is higher than that from textual input compared to the other two datasets, suggesting a greater reliance on visual input in this setting.

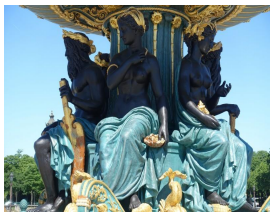
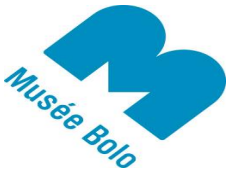
On the document side, overall, the model initially relies more heavily on the textual component in the first layers, with gate activations becoming more stable and balanced in later layers. Notably, in the OKVQA dataset, the gap between activation values is larger compared to the other two datasets. This is especially evident in the single example on the left side, where the image document provides little contribution to answering the question: consequently, the visual input gate maintains low scores across all layers, while the textual information receives greater emphasis.

Results on Retrieval-Augmented VQA. Some qualitative results on sample image-question pairs from InfoSeek are reported in Fig. 8, comparing the answers generated by augmenting LLaVA-MORE with different information retrieved by ReT against those retrieved by the original CLIP and PreFLMR models. The results highlight that ReT consistently retrieves more accurate and precise documents that help address specific multimodal questions, outperforming the other approaches. This demonstrates the effectiveness of our architecture in enhancing downstream performance.

WIT		IGLUE	
			
Could you elucidate the document associated with this image?	*Provide information about the document linked to this image.*	*Identify the document that is connected to this image.*	*What document is represented by this image?*
PreFLMR [32]: title: Mel Hunter section title: Science Fiction Illustration hierarchical section title: Mel Hunter Biography [...]	**PreFLMR [32]**: title: University of Nebraska Press hierarchical section title: University of Nebraska Press caption attribution [...]	**PreFLMR [32]**: title: Laodicea in Syria section title: History hierarchical section title: Laodicea in Syria History caption: Map showing [...]	**PreFLMR [32]**: title: China National Highway 318 hierarchical section title: China National Highway 318 caption: China National [...]
ReT (Ours) : title: Topopolis hierarchical section title: Topopolis caption: Interior view of an O'Neill cylinder space habitat, similar [...]	**ReT (Ours)** : title: WSRU hierarchical section title: WSRU caption: WRSK's "grunge" logo, 1996, Logo for the former WRSK-FM [...]	**ReT (Ours)** : title: Library of Alexandria section title: Early expansion and organization hierarchical section title: Library of [...]	**ReT (Ours)** : title: China National Highway 318 hierarchical section title: China National Highway 318 caption: China National [...]

KVQA		LLaVA	
			
Provide a brief description of the image and the relevant details of the person in the image.	*Provide a brief description of the image and the relevant details of the person in the image.*	*What does the background of the image include?*	*What activity might be good for this dog and why?*
PreFLMR [32]: This is an image of Lynn Swann and HHS Secretary Tommy Thompson. Tommy Thompson went to University [...]	**PreFLMR [32]**: This is an image of Plumb attending a "Heroes for Autism" event in Hollywood, California, April 2009. Eve Plumb [...]	**PreFLMR [32]**: The background of the image features a wall or a fence that is filled with graffiti art, giving the scene an urban vibe.	**PreFLMR [32]**: The dog's owner is likely engaging in this activity of playing catch with the orange Frisbee to provide physical exercise [...]
ReT (Ours) : This is an image of (l. to r.) The two Milk producers Bruce Cohen and Dan Jinks, Stuart, director Gus Van Sant [...]	**ReT (Ours)** : This is an image of Comic-Con 2012. Alex Borstein went to San Francisco State University, date of birth is 1973-02-15 [...]	**ReT (Ours)** : The background of the image includes a blue wall and a reclaimed building in a city.	**ReT (Ours)** : A suitable activity for this dog would be playing Frisbee, as shown in the image. Playing Frisbee not only provides the dog [...]

Figure 5. Qualitative results on M2KR, for datasets that do not include document images.

OVEN		InfoSeek	
			
Where is this place?	*what is the name of this building?*	*Which company manufactures this vehicle?*	*Where is the lake outflow to?*
PreFLMR [32]: A supercomputer is a computer with a high level of performance as compared to a general-purpose computer. [...]	**PreFLMR [32]**: The Fontaine Louvois is a monumental public fountain in Square Louvois on the rue Richelieu in the Second [...]	**PreFLMR [32]**: [...] On 28 May 2013, Aston Martin announced the V12 Vantage S - a sportier version of the V12 Vantage that [...]	**PreFLMR [32]**: The ships also connect to the Giessbachbahn, a funicular which climbs up to the famous Giessbach Falls. [...]
ReT (Ours) : The **Musée Bolo** or Swiss Museum of Computer Science, Digital Culture and Video Games is a private museum dedicated [...]	**ReT (Ours)** : The **Fontaines de la Concorde** are two monumental fountains located in the Place de la Concorde in the center of Paris. [...]	**ReT (Ours)** : SMG III gearbox with the E60 M5 that produces at 7,750 rpm and of torque at 6,100 rpm. **BMW** claimed performance [...]	**ReT (Ours)** : It flows out into a further stretch of the **Aare** at its western end. The culminating point of the lake's drainage basin is [...]
			

E-VQA		OKVQA	
			
What does this animal eat?	*When was this museum built?*	*What military branch do the men in the picture belong to?*	*What ingredient is missing from the picture to make a pb and j sandwich?*
PreFLMR [32]: At the larval stage, Pisaster ochraceus are filter feeders and their diet consists of plankton. As an adult, P. ochraceus feeds [...]	**PreFLMR [32]**: Built in 1860 by leading local industrialist Wilhelm Vedder, the original building was erected in 1860 with extensive [...]	**PreFLMR [32]**: [...] In modern usage "**navy**" used alone always denotes a military fleet, although the term "merchant **navy**" for a [...]	**PreFLMR [32]**: [...] sweet cream, is the highest grade of butter, has a sweet flavor, and is readily spreadable. If the butter is salted, the salt [...]
ReT (Ours) : They mainly feed on **sponges and small bacteria**. The sea star moves these tiny particles, which are captured in mucus and [...]	**ReT (Ours)** : The Kunsthalle Bern is a Kunsthalle (art exposition hall) on the Helvetiaplatz in Bern, Switzerland. It was built in **1917-1918** [...]	**ReT (Ours)** : A **navy** or maritime force is the branch of a nation's armed forces principally designated for naval and amphibious warfare [...]	**ReT (Ours)** : be substituted for the **jelly** component. On the flip side, the popularity of almond butter has inspired some aficionados to [...]
			

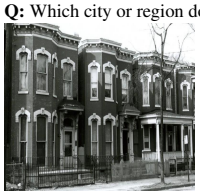
Figure 6. Qualitative results on M2KR, for datasets that include document images. We highlight the reference answer in bold font whenever it is found in the retrieved text. For ReT, we also add the document image attached to the text.

Figure 7. Qualitative analysis of gate activations. The left-side displays gate behavior during the encoding of a single query-document pair, while the right shows average activations over 2k examples from OVEN (top), E-VQA (middle), and OKVQA (bottom).

Q: What is this place named after?

CLIP [40]:
Moscow ✗
PreFLMR [32]:
Golden Gate ✗
ReT (Ours):
Izmaylovo ✓

Q: Which city or region does this place locate in?



CLIP [40]:
Bloomington, Illinois ✗
PreFLMR [32]:
Illinois ✗
ReT (Ours):
Richmond ✓

Q: Who is the founder of the organization in the image?

CLIP [40]:
Madeleine Paulson ✗
PreFLMR [32]:
William McPherson Allen ✗
ReT (Ours):
William Boeing ✓

Q: What is the closest parent taxonomy of this plant?

CLIP [40]:
Allium ✗
PreFLMR [32]:
Allium ✗
ReT (Ours):
Maianthemum ✓

Q: How many centimetre is the wingspan of this bird?

CLIP [40]:
20 ✗
PreFLMR [32]:
144-155 ✗
ReT (Ours):
100-105 ✓

Q: What is the length of this bridge in metre?

CLIP [40]:
No answer ✗
PreFLMR [32]:
178.4 ✗
ReT (Ours):
1056 ✓

Figure 8. VQA results on the validation split of InfoSeek, augmenting LLaVA-MORE with different information retrievers.