This is the peer reviewd version of the followng article:

Recurrence Meets Transformers for Universal Multimodal Retrieval / Caffagni, D., Sarto, S., Cornia, M., Baraldi, L., Cucchiara, R.. - In: IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. - ISSN 0162-8828. - (2026), pp. 1-23. [10.1109/tpami.2026.3729293]

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

29/10/2026 12:00

(Article begins on next page)

Recurrence Meets Transformers for Universal Multimodal Retrieval

Davide Caffagni*, Sara Sarto*, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara

***Abstract*—With the rapid advancement of multimodal retrieval and its application in LLMs and multimodal LLMs, increasingly complex retrieval tasks have emerged. Existing methods predominantly rely on task-specific fine-tuning of vision-language models and are limited to single-modality queries or documents. In this paper, we propose ReT-2, a unified retrieval model that supports multimodal queries, composed of both images and text, and searches across multimodal document collections where text and images coexist. ReT-2 leverages multi-layer representations and a recurrent Transformer architecture with LSTM-inspired gating mechanisms to dynamically integrate information across layers and modalities, capturing fine-grained visual and textual details. We evaluate ReT-2 on the challenging M2KR, M-BEIR, and MMEB benchmarks across different retrieval configurations. Results demonstrate that ReT-2 consistently achieves state-of-the-art performance across diverse settings, while offering faster inference and reduced memory usage compared to prior approaches. When integrated into retrieval-augmented generation pipelines, ReT-2 also improves downstream performance on Encyclopedic-VQA and InfoSeek datasets. Our source code and trained models are publicly available at: https://github.com/aimagelab/ReT-2.**

***Index Terms*—Multimodal Retrieval, Recurrence-Augmented Transformers, Retrieval-Augmented Generation.**

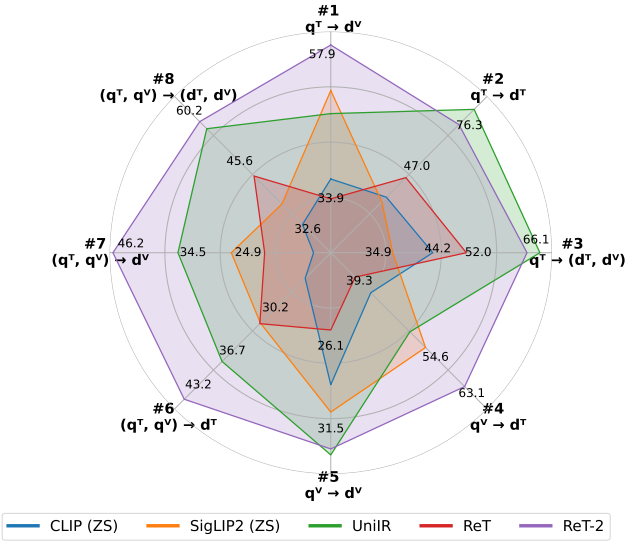


Fig. 1. In this work, we present Recurrence-enhanced Transformer (**ReT-2**), a novel retrieval approach supporting different tasks and data configurations, from cross-modal image-to-text retrieval – *i.e.*, $q^V \rightarrow d^T$, to multimodal text–image-to-text–image retrieval – *i.e.*, $(q^T, q^V) \rightarrow (d^T, d^V)$. The plot shows average results on the M-BEIR benchmark tasks [1], highlighting the performance gains of the proposed method over its previous version (*i.e.*, ReT [2]) and other state-of-the-art methods.

I. Introduction

INFORMATION retrieval is a fundamental and challenging task that entails identifying relevant content from large and heterogeneous data collections to satisfy user information needs. Early approaches predominantly focused on unimodal retrieval, where queries and retrievable items belonged to the same modality, such as text or images [3]–[5]. In recent years, however, the field has progressively shifted towards multimodal data [6]–[8], reflecting the growing presence of images, text, and other media in real-world applications. The advent of vision-language models, including CLIP [9], ALIGN [10] and other variants [11]–[13], further enabled effective cross-modal retrieval, allowing, for example, natural language queries to retrieve relevant images or vice versa.

At the same time, driven by the advent of Multimodal Large Language Models (MLLMs) [14]–[17] and the growing prominence of visual question answering tasks, there is an increasing demand for retrieval models capable of handling *multimodal queries* and retrieving *multimodal documents*, where multiple modalities coexist both within the query and the items to be retrieved. A typical example involves a query combining an image and a related text question, or specifying which part of the image should be retrieved [1], [18]–[20]. Despite significant progress in multimodal retrieval, existing state-of-the-art methods are largely limited to single-modality queries and documents, and thus fail to fully satisfy the flexibility required by modern applications.

To address these challenges, in this paper, we propose a retrieval approach that natively supports multimodal queries and documents (consisting of both text and images), and that can also handle scenarios with missing modalities from either query or document side. Our approach enables a more general retrieval paradigm (*i.e.*, universal multimodal retrieval), where multiple modalities and diverse retrieval tasks can be accommodated within a single unified framework. Unlike previous approaches that rely on feature fusion from only the last layer of vision-language backbones [1], our method exploits multi-layer representations for both modalities. We argue that explicitly incorporating features from shallower layers allows the model to better capture the wide variety of multimodal queries and documents, including fine-grained visual or textual details that are often lost in deeper layers. Moreover, we complement this design with an analysis of layer

*The first two authors equally contributed to this research.

D. Caffagni, S. Sarto, L. Baraldi, and R. Cucchiara are with the Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Italy (e-mail: {davide.caffagni, sara.sarto, lorenzo.baraldi, rita.cucchiara}@unimore.it).

M. Cornia is with the Department of Education and Humanities, University of Modena and Reggio Emilia, Italy (e-mail: marcella.cornia@unimore.it).

activations, which allows us to identify and prune redundant layers, thereby reducing computational overhead while simultaneously enhancing robustness.

To achieve these goals, we design a Transformer-based recurrent cell that, at each layer, merges features from the visual and textual backbones with its internal states. Inspired by the gating mechanism of an LSTM [21], our model employs a forget gate to control how much information to retain from shallower layers, while textual and visual input gates modulate the unimodal information flow. This design enables our model to dynamically determine which layers and modalities are most informative for encoding each query or document. In stark contrast to previous work that combines Transformers and recurrence architectures [22], [23], we do not utilize recurrence to model temporal dependencies among tokens. Rather, we exploit a Transformer-based recurrent cell to model the interaction across representations from multiple layers of pre-trained backbones. The recurrent design allows us to save parameters, as multiple layers are processed by the same, shared Transformer-based cell. On the other hand, while in a standard Transformer all the layer representations participate in the residual stream, the input and forget gates of our cell dynamically select which modality and which depth of representation can enter the hidden state.

Our proposed model, which we call **ReT-2** (Recurrence-enhanced Transformer), is experimentally evaluated on the challenging M2KR benchmark [18], which integrates a diverse collection of datasets adapted for multimodal retrieval. To further broaden the evaluation, we extend the M2KR benchmark by augmenting the OVEN [24], InfoSeek [19], Encyclopedic-VQA [20], and OKVQA [25] splits to incorporate images within the reference documents. In addition, we conduct experiments on the M-BEIR benchmark [1], focusing particularly on settings where certain modalities are absent. Across more than eight distinct retrieval configurations, including text-to-image, text–image-to-image, and text–image-to-text–image retrieval, ReT-2 consistently demonstrates strong and stable performance, as summarized in Fig. 1. Across all these experiments, including zero-shot evaluations on the MMEB benchmark [26], our results highlight not only the effectiveness of our approach in conventional retrieval scenarios, but also its ability to generalize to highly compositional and underexplored multimodal configurations.

Finally, we demonstrate the utility of ReT-2 as a retrieval backbone for retrieval-augmented generation in knowledge-intensive visual question answering. In this setting, many questions can not be answered without retrieving external multimodal knowledge, making retrieval quality a decisive factor. Our experiments, conducted on Encyclopedic-VQA [20] and InfoSeek [19], show that ReT-2 provides more effective retrieval support compared to alternative retrieval methods, enabling off-the-shelf MLLMs [16], [17] to achieve higher answer accuracy without task-specific fine-tuning.

Beyond retrieval accuracy and its effectiveness when employed as retrieval backbone for downstream tasks, we further assess the computational efficiency of ReT-2 in comparison to existing state-of-the-art methods. Our analysis reveals that the benefits of ReT-2 are not confined to retrieval effectiveness, but also extend to practical efficiency, achieving faster inference and reduced memory usage than competing methods.

In summary, our main contributions are as follows:

- We present ReT-2, a unified retrieval model that supports multimodal queries and documents, equipped with a recurrence-enhanced Transformer cell that integrates visual and textual features via LSTM-inspired gating.
- Unlike prior approaches that rely only on final-layer features, we exploit multi-layer representations and introduce a pruning strategy to remove redundant layers, improving both robustness and efficiency.
- Extensive experiments on the M2KR, M-BEIR, and MMEB benchmarks demonstrate state-of-the-art performance across a wide range of multimodal retrieval tasks.
- We further show that ReT-2 boosts retrieval-augmented generation for knowledge-intensive VQA, enabling off-the-shelf MLLMs to achieve higher answer accuracy.

This work is an extended and improved version of our earlier conference paper [2]. Compared to our previous approach (*i.e.*, ReT), the current work provides a deeper architectural analysis and introduces several architectural modifications that collectively improve both efficiency and robustness. These advances lead to a conceptually simpler yet more effective model, allowing ReT-2 to set a stronger foundation for universal multimodal retrieval and its downstream applications.

II. Related Work

From Unimodal Retrieval to Cross-Modal Retrieval. Classical retrieval methods were largely unimodal, focusing on either text-based document search or content-based image retrieval [3]–[5]. While effective in their domains, they lacked the ability to bridge modalities. The advent of large-scale vision-language datasets [7], [8] and dual-encoder models such as CLIP [9] and its variants [12], [13], [27] marked a turning point, enabling contrastive learning to align images and text in a shared embedding space. Despite this progress, these models are typically evaluated on relatively small benchmarks like Flickr30k [28] and COCO [6], which emphasize simple queries and limit generalization. Building on these advances, more complex retrieval scenarios have emerged, including composed image retrieval [29], long-text-to-image retrieval [30], and multimodal query-to-multimodal document retrieval [19], [24]. Current methods typically address such tasks through specialized fine-tuning [31]–[33], but a universal framework capable of seamlessly accommodating diverse query and document modalities remains an open challenge.

Universal Multimodal Retrieval. With a rising demand for multimodal retrieval systems, the ability to handle complex multimodal queries has become essential. This trend has led to the development of specialized benchmarks for multimodal retrieval, supporting diverse tasks and data configurations. For instance, M2KR [18] combines several datasets for this task. Similarly, the large-scale M-BEIR benchmark [1] covers a wide range of domains and image sources, enabling comprehensive evaluation of multimodal retrieval approaches.

The challenge of developing robust multimodal representations remains a foundational question in multimodal learning, driving research toward effective strategies for encoding

queries and documents across modalities. UniIR [1] integrates modalities using features from the last layer of pre-trained models [9], [34], aiming to build a unified retriever for diverse tasks. Similarly, GENIUS [35] is a flexible generative retrieval framework that converts multimodal inputs into discrete representations and enhances generalization through query-target interpolation. Meanwhile, models like FLMR [36] and PreFLMR [18] explore a late-interaction paradigm [37], where multimodal queries and text-only documents are encoded independently into sets of latent tokens, and relevance scores are computed by aggregating token-level similarities.

With the recent advancements in LLMs [38], [39], research has increasingly turned to multimodal models to align visual and textual modalities via visual instruction tuning [14]. Despite these advances, the potential of MLLMs for universal retrieval tasks remains relatively underexplored, with approaches such as LamRA [40] and MM-Embed [41] attempting to repurpose MLLMs for the task. However, employing MLLMs in this setting typically requires multi-stage finetuning of a large number of parameters, leading to substantial training costs and limited inference efficiency. To address these challenges, recent works explore more efficient strategies, such as PUMA [42], which prunes parameters to reduce computational overhead, and JFE [43], which leverages early visual-textual fusion to enhance cross-modal understanding.

In contrast, ReT-2 is designed to efficiently integrate multi-layer visual and textual features thanks to a recurrent-enhanced architecture, achieving performance comparable to MLLM-based models while avoiding their high computational costs.

Recurrence-Augmented Transformers. Transformer architectures [44] have achieved impressive results across diverse domains, from natural language understanding [38], [45], [46] to computer vision [47]–[50]. However, their quadratic complexity with respect to input length has driven research into alternative designs. One line of work integrates recurrent mechanisms within Transformer models, interleaving Transformer layers with recurrent neural networks to balance attention with sequential processing [51]–[53]. Other approaches, such as the R-Transformer [22], incorporate local recurrent cells to enable parallel computation. The Block-Recurrent Transformer [23], for instance, embeds recurrent dynamics inspired by LSTM cells [21] directly within the Transformer framework. Unlike prior works that primarily use recurrence to reduce the computational cost of modeling temporal dependencies, in this paper, we exploit recurrence to enable multi-layer and multimodal feature integration, aiming to enhance performance on multimodal retrieval tasks.

III. Background

Problem Formulation. In our setting, both queries $\mathrm{q} = (q^T, q^V)$ and documents $\mathrm{d} = (d^T, d^V)$ are structured as paired image-text instances. The textual component of each query, q^T, typically comprises an instruction, *e.g.* "Utilizing the given image, obtain documents that respond to the following question", followed by a question specific to the associated query image q^V. The goal is to retrieve documents that are most relevant to the given query. Each document consists of a textual response d^T that addresses the query and may optionally include a corresponding image d^V.

This task presents several significant challenges. It demands the ability to interpret fine-grained visual and linguistic cues, establish coherent alignment between multimodal semantics in the query and candidate documents, and perform reasoning across visual and textual modalities. Furthermore, the presence or absence of images in documents introduces variability in the retrieval signal, making the matching process more complex.

ReT: Recurrent Transformer with Fine-grained Late Interaction. ReT [2] is a multimodal retrieval model that introduces a novel Transformer-based recurrent cell designed to fuse multimodal features from both queries and documents, enabling the computation of fine-grained similarity scores between them. The model leverages pre-trained visual and textual backbones, aggregating features across multiple layers to construct rich representations of each modality. While leveraging a Transformer architecture, ReT incorporates a learnable gating mechanism inspired by LSTMs [21] to regulate information flow across layers. At each recurrent step, it combines its internal state with the visual and textual features extracted from the current backbone layer, treating lower-level features as the past. This mechanism enables the model to selectively preserve or discard earlier-layer information, allowing it to emphasize more meaningful high-level features during fusion.

The model comprises two dedicated encoders for queries and documents, $\text{ReT}_{\mathbf{Q}}$ and $\text{ReT}_{\mathbf{D}}$, which share the same architecture but maintain separate learnable parameters that are optimized jointly. Each encoder integrates a recurrent cell with pre-trained visual and textual backbones. Specifically, for each modality $m \in T, V$, the unimodal backbone produces a set of activations $E^m(q^m) = E_l^m(q^m)_{l=1}^L$, where $E_l^m(q^m) \in \mathbb{R}^{N \times d}$ denotes the features extracted from the l-th layer and L is the total number of layers. At each layer l, the recurrent cell performs *feature fusion* [1] over three inputs: the hidden state from the previous step, $\mathbf{h}_l \in \mathbb{R}^{k \times d}$, and the visual and textual representations $\mathbf{E}_l^V$ and $\mathbf{E}_l^T$. For the initial step, the hidden state $\mathbf{h}_0$ is initialized with $k = 32$ learnable vectors.

Formally, given a query-document pair (q, d), each side uses distinct learnable input tokens. The final outputs are:

$$\mathbf{Q} = \text{ReT}_{\mathbf{Q}}(\mathrm{q}) \in \mathbb{R}^{k \times \overline{d}} \tag{1}$$

$$\mathbf{D} = \text{ReT}_{\mathbf{D}}(\mathrm{d})^{\intercal} \in \mathbb{R}^{\overline{d} \times k}, \tag{2}$$

where $\overline{d}$ is the dimension after projection.

During training, these representations are used to compute a fine-grained late-interaction [54] relevance score:

$$s(\mathbf{Q}, \mathbf{D}) = \sum_{i=1}^{k} \max_{j=1\ldots k} \mathbf{Q}_i \cdot \mathbf{D}_j. \tag{3}$$

Here, similarity is computed as the dot product between the i-th query and j-th document token. The $\max$ operator ensures that only the most relevant document tokens contribute to the score of each query token, effectively filtering out locally irrelevant matches. Training is performed by jointly optimizing both the query and the document encoder with an InfoNCE loss [9], where global query-document cosine similarities are replaced with the score defined in Eq. 3.

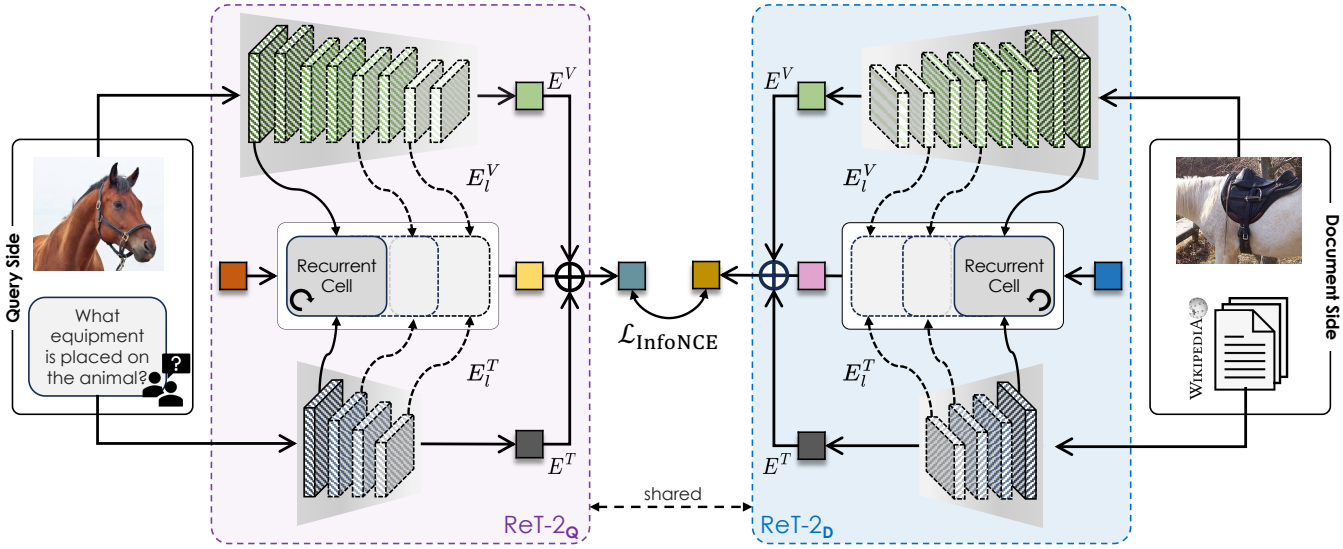


Fig. 2. Overview of the proposed Recurrence-enhanced Transformer (**ReT-2**) for universal multimodal retrieval.

Limitations of ReT. Our previous work, ReT, demonstrated strong retrieval performance, validating the effectiveness of recurrent multimodal fusion. However, there remains room for improvement in both efficiency and efficacy. Given the recurrent nature of the architecture, reducing the number of fused layers could lead to faster inference, thereby improving computational efficiency and reducing redundant representations across depth [55], [56]. Beyond efficiency, processing every backbone layer also lengthens the recurrent chain that low-level information must survive and repeatedly exposes the cell's gates to near-identical, residual-stream-redundant evidence across consecutive layers, both of which increase the risk of overwriting or replaying the same fusion decision rather than genuinely refining it (cf. Appendix B). Additionally, ReT encodes queries and documents into 32×128 matrices (*i.e.*, $k \times \overline{d}$). We empirically observe that these matrices suffer from rank collapse [57], where their rows converge to a uniform representation, undermining the purpose of leveraging multiple embeddings to capture diverse nuances of the input. This is consistent with the theoretical account of rank collapse in self-attention networks [58], which we relate directly to the recurrent cell in Appendix B. This raises the question of whether a single, larger embedding is better than a small embedding matrix for multimodal retrieval.

IV. Proposed Method

In this section, we introduce an enhanced variant of ReT, referred to as **ReT-2**, which is specifically designed to address the limitations identified in the original model. ReT-2 aims to improve retrieval effectiveness and efficiency when dealing with heterogeneous data sources in large-scale, multimodal collections of entities. A graphical overview of the architecture of our ReT-2 model is shown in Fig. 2.

A. Overall Architecture

In our ReT-2 model, the architecture retains two dedicated encoders for queries and documents. However, in contrast to the previous version (which employed separate parameter sets optimized jointly), ReT-2 introduces a unified encoder architecture with shared weights for both modalities. Specifically, each encoder comprises a recurrent fusion cell coupled with pre-trained, learnable visual and textual backbones. This parameter sharing not only reduces model complexity and reduces overfitting, but also encourages consistent representation learning across queries and documents. This design rests on the assumption that queries and documents are represented through compatible combinations of visual and textual inputs and must be mapped into a common semantic space for their symmetric dot-product similarity to be meaningful; the same assumption should hold, more broadly, for any symmetric or approximately symmetric dual-encoder retrieval setting.

In the following, we retain the notation introduced in Sec. III and denote the cross-attention [44] between two matrices $\mathbf{x}$ and $\mathbf{y}$, as $\texttt{Attention}(\mathbf{x}, \mathbf{y})$.

Recurrent Cell. The architecture of the recurrent cell is illustrated in Fig. 3. Within the cell, the input hidden state $\boldsymbol{h}_l$ is processed through three parallel branches. The first branch retains the candidate hidden state $\mathbf{c}_l$ of the recurrent cell. Notably, for layers $l \geq 1$, $\boldsymbol{h}_l$ encodes accumulated, layer-specific representations of both the image and text. Rather than processing all layers of the visual and textual backbones, we consistently sample three representative layers: one from the lower (early), one from the middle, and one from the upper (final) sections of each backbone. This approach ensures a balanced capture of low-, mid-, and high-level features while maintaining computational efficiency and architectural compatibility across backbones of varying depth.

To effectively incorporate contextual information from both modalities, the remaining two branches perform feature fusion between $\boldsymbol{h}_l$ and the unimodal visual and textual representations extracted from the l-th layer of their respective backbones.

Specifically, we employ two independent cross-attention modules to fuse the normalized input $\hat{\boldsymbol{h}}_l$ with the visual and

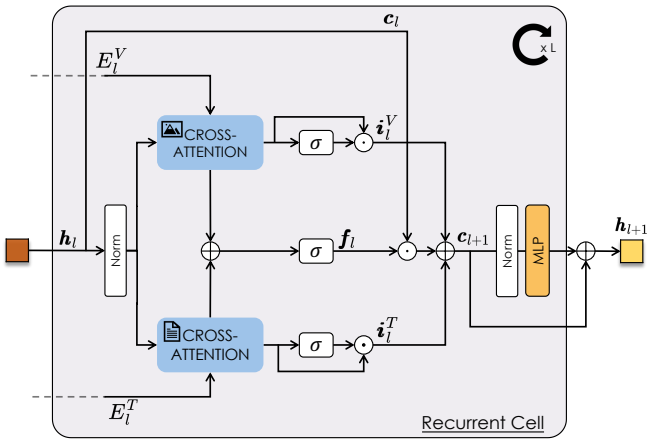


Fig. 3. Graphical illustration of the proposed recurrent cell for multimodal retrieval, which integrates layer-specific textual and visual features into a matrix-form hidden state.

textual representations, respectively, as

$$\boldsymbol{z}_l^m = \texttt{Attention}\left(\hat{\boldsymbol{h}}_l,\, E_l^m\right), \tag{4}$$

where $m \in T, V$ and $\hat{\boldsymbol{h}}_l = \texttt{LayerNorm}(\boldsymbol{h}_l)$.

The outputs of the three branches are combined to compute the updated internal state of the recurrent cell. This state is formed as a gated linear combination of the candidate state $\mathbf{c}_l$ and the outputs from the two feature fusion branches, denoted as $\boldsymbol{z}_l^T$ and $\boldsymbol{z}_l^V$. The combination is modulated by a set of learnable forget and input gates.

In detail, the forget gate $\boldsymbol{f}_l$ controls the extent to which information from earlier applications of the recurrent cell (corresponding to shallower layers, or the "past") is retained in the current step, based on the ongoing multimodal interaction $\boldsymbol{z}_l^m$. In parallel, the input gates $\boldsymbol{i}_l^m$ regulate the influence of the unimodal features from the current (l-th) layer. This mechanism allows the model to attenuate noisy or less relevant high-level representations when fine-grained visual or textual details (*e.g.*, colors or shapes) are more pertinent to the query. Formally, the next candidate state is obtained as

$$\boldsymbol{c}_{l+1} = \mathbf{c}_l \odot \boldsymbol{f}_l + \boldsymbol{z}_l^T \odot \boldsymbol{i}_l^T + \boldsymbol{z}_l^V \odot \boldsymbol{i}_l^V, \tag{5}$$

where $\boldsymbol{f}_l$, $\boldsymbol{i}_l^T$ and $\boldsymbol{i}_l^V$ indicate the learnable sigmoidal gates. In particular, these are computed as follows:

$$\begin{aligned} \boldsymbol{f}_l &= \sigma\left(\boldsymbol{W}_f^T \cdot \boldsymbol{z}_l^T + \boldsymbol{W}_f^V \cdot \boldsymbol{z}_l^V + b_f\right), \\ \boldsymbol{i}_l^m &= \sigma\left(\boldsymbol{W}_i^m \cdot \boldsymbol{z}_l^m + b_i\right), \end{aligned} \tag{6}$$

where $\boldsymbol{W}_f^T, \boldsymbol{W}_f^V, \boldsymbol{W}_i^m$ are trainable weight matrices, and b_f, b_i are fixed scalar biases.

The updated state $\boldsymbol{c}_{l+1}$ undergoes layer normalization and is passed through a residual two-layer feed-forward network to produce the output of the recurrent cell, as

$$\boldsymbol{h}_{l+1} = \boldsymbol{c}_{l+1} + \texttt{MLP}\left(\texttt{LayerNorm}\left(\boldsymbol{c}_{l+1}\right)\right). \tag{7}$$

After going through different layers of the backbones, the output from the last iteration of the recurrent cell, $\boldsymbol{h}_L \in \mathbb{R}^{k \times d}$ (where $k = 1$ in our novel formulation), consists of a latent token that serves to compute query-document relevance scores. Specifically, the output $\boldsymbol{h}_L$ is transformed into a different vector space through a linear projection $\mathbf{W}_{final} \in \mathbb{R}^{d \times \overline{d}}$, *i.e.*

$$\overline{\boldsymbol{h}}_L = \boldsymbol{h}_L \cdot \mathbf{W}_{final}. \tag{8}$$

Global Feature Injection. At the output of the recurrent cell, $\overline{\boldsymbol{h}}_L$ encodes multimodal information that integrates details from multiple levels of abstraction. However, retaining access to the raw global features provides a broader contextual representation of the query or document. To leverage this complementary information, we augment the multimodal representation $\overline{\boldsymbol{h}}_L$ with the unimodal outputs of the visual and textual backbones, denoted as E^V and E^T, respectively. These typically correspond to the `CLS` visual pooler token and the `EOS` textual pooler token. The integration is performed by summing the global features with the output of the recurrent cell, obtaining the final representation of the query as

$$\overline{\boldsymbol{h}}_L = \overline{\boldsymbol{h}}_L + E^V(q^V) + E^T(q^T). \tag{9}$$

The recurrent and global paths are intended to be complementary: the recurrent cell selectively integrates layer-specific cues, which can come at the expense of broader semantic context, while the pooler tokens directly preserve such context. This complementarity is expected to hold whenever the global representation is itself well-trained and not already redundant with the fine-grained evidence accumulated by the recurrent branch, a property we expect to transfer to other architectures that aggregate hierarchical or multi-layer features.

Missing Modalities. In practical use cases, queries and documents are not always multimodal. For instance, a user can formulate a purely textual query, asking to retrieve an image-text pair, as well as an image alone. ReT-2 handles these scenarios by using an empty string to overcome a missing text, and a black picture in place of a missing image. Next, the input and forget gate components pertaining to the missing modality (cf. Eq. 6) and the unimodal global features extracted from the placeholder input (cf. Eq. 9) are manually set to zero.

B. Training Procedure

Given a query-document pair (q, d), along with a learnable token in input for both the query and document sides, we denote the corresponding final output $\overline{\boldsymbol{h}}_L$ of the query and the document encoders as

$$\mathbf{Q} = \text{ReT-2}_{\mathbf{Q}}(\mathrm{q}) \in \mathbb{R}^{k \times \overline{d}} \tag{10}$$

$$\mathbf{D} = \text{ReT-2}_{\mathbf{D}}(\mathrm{d})^{\intercal} \in \mathbb{R}^{\overline{d} \times k}, \tag{11}$$

where $\text{ReT-2}_{\mathbf{Q}} = \text{ReT-2}_{\mathbf{D}}$ in our shared implementation.

Training is performed by optimizing both the query and the document encoder with the InfoNCE loss [9], where query-document similarity is computed as the dot-product between the query and the document token.

C. Summary: From ReT to ReT-2

Overall, ReT-2 incorporates several targeted changes to enhance the efficiency, robustness, and simplicity of the original ReT architecture:

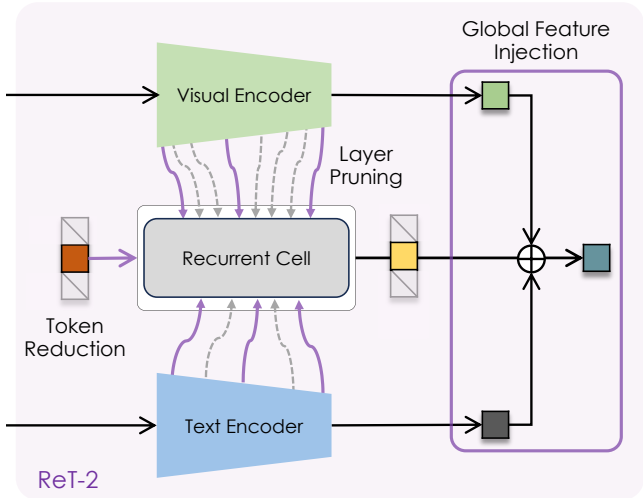


Fig. 4. Visualization of the differences between the previous method (*i.e.*, ReT) and the newly proposed ReT-2. The new method introduces three key modifications: (i) **token reduction**, instead of multiple input tokens, only a single token is used; (ii) **layer pruning**, rather than using all textual and visual layers, we now select only three representative layers (early, middle, and late), independent of the architecture; and (iii) **global feature injection**, a newly added module that integrates global information to enhance the representation. Highlighted regions indicate the most significant differences.

- **Shared Query-Document Encoder**: Unlike ReT, which used separate encoders with distinct learnable parameters for queries and documents, ReT-2 adopts a shared architecture with tied weights, promoting consistency and reducing model complexity.
- **Token Reduction**: The number of input tokens is reduced from 32 to a single token for each query or document. This design choice addresses the issue of rank collapse observed in the output embeddings and encourages the model to produce more compact and meaningful representations. The resulting token can be interpreted as a task-specific `[CLS]`-like representation, recurrently accumulated rather than computed from the raw input.
- **Simplified Contrastive Objective**: The use of a single token per side eliminates the need for the fine-grained contrastive loss used in ReT. Instead, we apply a standard InfoNCE loss directly over the single fused token from both the query and document, significantly simplifying the retrieval pipeline and improving inference efficiency.
- **Layer Pruning**: Rather than relying on all layers of the textual and visual backbones or attempting to explicitly align architectures with different depths, we always sample three layers: one from the lower (early), one from the middle, and one from the upper (final) part of each backbone. This strategy ensures compatibility and stability, especially when backbones differ in depth.
- **Global Feature Injection**: To enhance contextual understanding, ReT-2 integrates global feature representations alongside layer-specific features. This injection of global context helps the model capture general information, further improving retrieval accuracy and robustness.

A visual summary of the modifications and improvements implemented in ReT-2 is provided in Fig. 4.

TABLE I
SELECTED LAYERS FOR EACH BACKBONE IN RET-2. L DENOTES THE DEPTH OF EACH BACKBONE, MEASURED IN NUMBER OF LAYERS.

	Text Encoder		Visual Encoder	
Backbone	L	**Layer Indices**	L	**Layer Indices**
CLIP ViT-B	12	3, 7, 11	12	3, 7, 11
ColBERTv2	12	3, 7, 11	-	-
CLIP ViT-L	12	3, 7, 11	24	3, 18, 23
SigLIP2 ViT-L	24	3, 18, 23	24	3, 18, 23
OpenCLIP ViT-H	24	3, 18, 23	32	4, 25, 31

V. EXPERIMENTS ON MULTIMODAL RETRIEVAL

A. Datasets and Evaluation Metrics

We evaluate our models on the M2KR [18], M-BEIR [1], and MMEB [26] benchmarks, which provide a diverse, large-scale collection of datasets and task configurations for comprehensive evaluation of multimodal retrieval performance.

M2KR Benchmark. M2KR integrates heterogeneous sources, including WIT [59], IGLUE [60], KVQA [61], CC3M [7], MSMARCO [62], OVEN [24], LLaVA [63], InfoSeek [19], Encyclopedic-VQA [20], and OKVQA [25]. These datasets span a wide range of domains, enabling robust evaluation of retrieval models under varying degrees of complexity and multimodal reasoning. To better align with our setting, where both queries and documents are multimodal, we augment the M2KR splits of OVEN, InfoSeek, Encyclopedic-VQA, and OKVQA by enriching the reference documents with associated images [2], thereby enabling a more effective evaluation of models that rely on both textual and visual signals during retrieval. In our experiments, we employ training, validation, and test splits used in previous works [2], [18].

M-BEIR Benchmark. M-BEIR comprises eight retrieval tasks and ten different datasets, with around 1.5M human-authored queries and a pool of 5.6M candidate documents. The benchmark spans diverse sources, including everyday images, fashion products, Wikipedia entries, and news articles. In addition to standard multimodal settings, it includes tasks with missing modalities on either the query or document side, enabling evaluation under incomplete conditions. To ensure consistency between training and testing, M-BEIR adapts datasets originally designed for different tasks, including OVEN [24], EDIS [64], CIRR [29], FashionIQ [65], COCO [6], Fashion200k [66], Visual News [67], and NIGHTS [68]. Moreover, M-BEIR defines a *global* retrieval scenario, where candidates are retrieved from the full 5.6M pool encompassing all tasks and datasets, and a *local* one, which restricts candidates to the task-specific pool provided by each dataset. In this paper, we report results on the M-BEIR$_{\text{local}}$ setting, for fair comparison with existing state-of-the-art retrieval models.

MMEB Benchmark. The Massive Multimodal Embedding Benchmark (MMEB) comprises 36 datasets spanning four meta-task categories: classification (10 tasks), visual question answering (10 tasks), multimodal retrieval (12 tasks), and visual grounding (4 tasks). In contrast to M2KR and M-BEIR, which focus on retrieval, MMEB enables a broader evaluation of multimodal embedding quality across diverse downstream

TABLE II
ABLATION STUDY RESULTS ON THE M2KR BENCHMARK. ALL EXPERIMENTS ARE WITH CLIP VIT-L FOR BOTH VISUAL AND TEXTUAL ENCODERS.

	WIT	**IGLUE**	**KVQA**	**OVEN**	**LLaVA**	**InfoSeek**		**E-VQA**		**OKVQA**		
Model	R@10	R@1	R@5	R@5	R@1	R@5	PR@5	R@5	PR@5	R@5	PR@5	**Avg**
ReT [2]	73.4	81.8	63.5	82.0	79.9	47.0	60.5	44.5	57.9	20.2	66.2	61.5
+ global features	79.3	81.6	65.8	82.8	82.1	46.6	60.8	43.3	58.0	17.4	64.8	62.0
+ score fusion (32 tokens)	79.5	81.9	66.7	83.4	81.0	42.7	57.5	43.4	57.1	17.5	65.1	61.4
+ shared architecture (32 tokens)	78.0	83.4	66.7	83.5	84.2	48.0	59.9	48.0	61.1	13.3	62.8	62.6
+ shared architecture (16 tokens)	78.4	82.2	66.1	83.6	83.2	47.5	60.1	48.4	61.7	13.1	63.6	62.5
+ shared architecture (8 tokens)	78.1	82.6	62.4	83.5	82.7	47.2	60.9	48.4	61.1	14.2	65.6	62.4
+ shared architecture (4 tokens)	78.7	82.2	63.5	82.9	82.3	48.5	61.3	47.5	60.8	13.5	63.1	62.2
+ shared architecture (single token)	78.3	81.9	66.5	84.1	84.1	48.2	60.5	48.7	60.9	12.9	65.0	62.8
+ non-shared architecture (single token)	79.8	82.5	67.8	84.4	81.4	46.4	59.0	42.6	56.7	17.1	65.6	62.1
+ layer pruning	77.9	82.2	63.3	84.3	82.9	50.1	62.2	47.7	60.7	14.6	66.0	62.9
+ global features (**ReT-2**)	81.1	82.9	72.3	83.1	83.8	48.0	61.0	49.7	62.6	15.2	65.9	64.1
- recurrent cell	69.5	74.2	59.9	75.8	81.9	37.4	52.5	46.6	59.0	14.1	60.0	57.3
- shared architecture	75.5	81.1	70.0	78.7	79.1	50.3	62.2	35.8	51.5	13.9	63.5	60.1
- score fusion (32 tokens)	77.9	81.6	60.9	82.8	81.9	46.2	60.2	47.8	59.3	12.9	62.4	61.3
- token reduction (32 tokens)	81.3	82.8	70.0	83.1	83.2	48.0	62.4	50.0	61.5	13.6	63.9	63.6
- layer pruning	80.7	81.8	71.1	82.8	83.5	48.1	61.7	49.6	61.9	14.8	65.6	63.8

tasks. We use MMEB solely for evaluation, reporting performance in a zero-shot setting.

Evaluation Metrics. Following the evaluation protocol of M2KR, we assess model performance using recall@K (*i.e.*, the percentage of queries for which the target document falls within the top-K most similar documents). The value of K is determined based on the experimental setup of each sub-dataset. For VQA splits, we also report the pseudo recall metric, as proposed in [18], which considers a retrieved document relevant whenever it contains the answer. For M-BEIR, we adhere to the original evaluation protocol and report standard recall@K values accordingly (using $K = 5$ for most datasets, and $K = 10$ for Fashion200k and FashionIQ). For MMEB, we report recall@K with $K = 1$ for all categories, which is equivalent to precision@1 used in the original benchmark.

B. Implementation Details

In our experiments, we evaluate multiple configurations of both visual and textual backbones. For the visual encoder, we consider CLIP ViT-B/32, CLIP ViT-L/14 [9], SigLIP2 ViT-L/14 [13], and OpenCLIP ViT-H/14 [11]. For the textual encoder, we use the corresponding CLIP/SigLIP variants as well as ColBERTv2 [54]. Following the methodology described in Sec. IV, we retain only three representative layers from each backbone, corresponding to early, intermediate, and late stages. The specific layers selected for each configuration are detailed in Table I.

Our models are trained in mixed precision with the Adam optimizer on 4 NVIDIA A100 64GB GPUs for up to 24 hours. When adding global features, we always unfreeze the pooling layer of the backbones, if present. This corresponds to the visual and textual linear projections for CLIP-based and ColBERTv2 models, and to the attention pooling layers for SigLIP2. Following ReT, the recurrent cell of ReT-2 operates with a hidden size d equal to 1,024 and with the biases b_i and b_f equal to zero. The dimension of $\mathbf{W}_{final}$ (cf. Eq. 8) is set to match d with the dimension of the global features. When unfreezing the unimodal backbones, we activate gradient checkpointing, and we downscale their learning rate by 0.05 compared to the recurrent cell for stability. At test time, we index passages using the Faiss library [69] for fast retrieval.

For M2KR, we use the same training recipe as ReT [2], setting the learning rate to 5×10^{-5} with a cosine scheduler and a batch size of 512, training for 75k steps. We observe that training further leads to overfitting on some benchmarks, particularly severe on InfoSeek. For M-BEIR, we train for 20 epochs with a batch size of 768, using the data sampling strategy proposed in [43]. The learning rate is linearly ramped up to 1×10^{-4} within the first 300 steps, and then decays accordingly to a cosine schedule.

C. Ablation Studies and Analyses

The original ReT model employs 32 input tokens and a dedicated recurrent cell on both the query and document sides. During training, the output tokens are used to compute a fine-grained late-interaction relevance score, following [37], [54] (cf. Sec. III). Table II presents ablation studies supporting the architectural modifications introduced in ReT-2. Results are reported on the M2KR benchmark, using CLIP ViT-L as visual and textual backbones.

Score Fusion. We first assess the impact of replacing the fine-grained late-interaction relevance score computation with a score fusion strategy. In practice, rather than computing 32 $\times$ 32 dot products for each query-document pair, we sum the rows of the output matrix of ReT before the late-interaction projection to dimension 128, obtaining a single embedding token, typically of a size varying from 768 to 1,024, to compute the query-document similarity via dot product. Note that this is equivalent to substituting the $\max$ operator in Eq. 3 with a new summation over j (see [1] for more details). However, thanks to the distributive property of the dot product, we do not need to compute the 32×32 similarity matrix

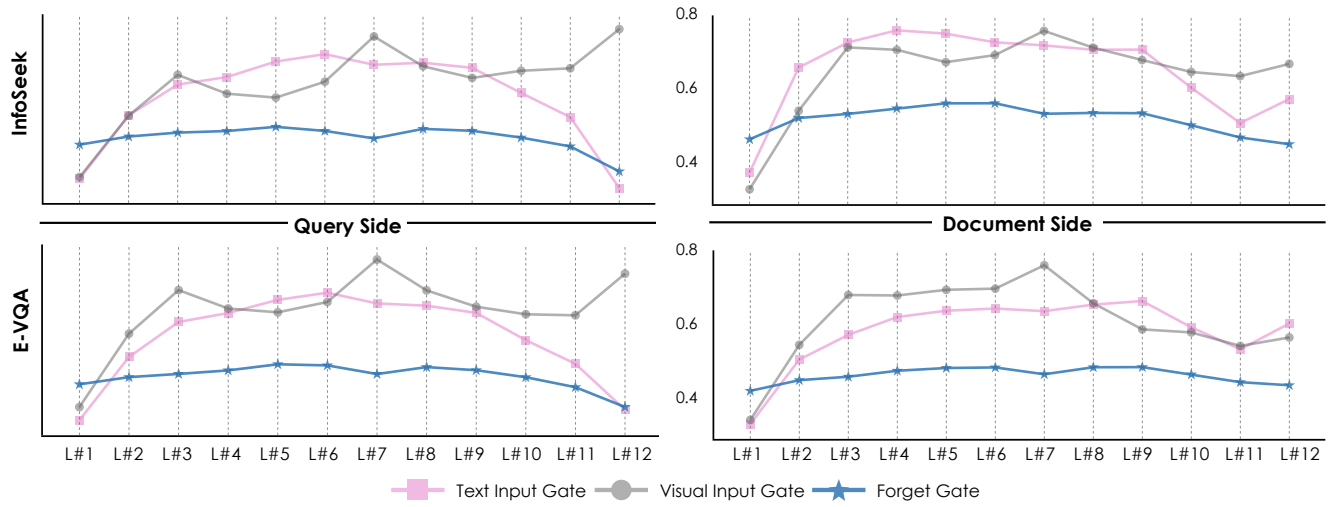


Fig. 5. Analysis of average gate activation over 2k examples from the InfoSeek and Encyclopedic-VQA test split of the M2KR benchmark.

TABLE III
PERFORMANCE VARIATION ACROSS DIFFERENT LAYER CONFIGURATIONS ON THE M2KR BENCHMARK. ALL EXPERIMENTS ARE WITH CLIP VIT-L FOR BOTH VISUAL AND TEXTUAL ENCODERS (*i.e.*, $L = 24$ FOR VISUAL, $L = 12$ FOR TEXTUAL).

	Layers		WIT	IGLUE	KVQA	OVEN	LLaVA	InfoSeek		E-VQA		OKVQA		
Model	**Text**	**Vision**	R@10	R@1	R@5	R@5	R@1	R@5	PR@5	R@5	PR@5	R@5	PR@5	**Avg**
ReT-2 (early only)	3	3	79.3	82.5	71.4	80.9	81.2	49.0	62.0	45.1	57.9	13.3	63.3	62.3
ReT-2 (middle only)	7	18	80.2	83.2	70.4	82.1	80.6	50.2	61.9	48.9	61.3	13.0	63.5	63.2
ReT-2 (late only)	11	23	80.7	83.2	71.0	82.7	80.7	48.4	60.4	46.7	59.6	12.1	63.1	62.6
ReT-2 (early+middle)	3, 7	3, 18	80.7	82.8	65.5	83.2	82.9	49.6	62.5	49.7	61.7	13.9	65.0	63.4
ReT-2 (early+late)	3, 11	3, 23	80.3	83.2	71.1	83.1	83.0	48.0	62.4	50.0	61.5	14.3	64.9	63.8
ReT-2 (middle+late)	7, 11	18, 23	81.2	82.6	72.0	83.5	83.2	48.6	61.1	48.5	61.3	15.0	64.3	63.8
ReT-2	3, 7, 11	3, 18, 23	81.1	82.9	72.3	83.1	83.8	48.0	61.0	49.7	62.6	15.2	65.9	64.1
- layer pruning	all	odds	80.7	81.8	71.1	82.8	83.5	48.1	61.7	49.6	61.9	14.8	65.6	63.8

explicitly. This shift enables faster and more memory-efficient training, as well as quicker inference retrieval, with minimal change in performance, as the average retrieval score moves from 61.5 to 61.4 – *i.e.*, *score fusion (32 tokens)*.

Sharing Weights. Building on the score fusion model, we experiment with sharing the weights between the query and document encoders, essentially setting $\text{ReT}_{\mathbf{Q}} = \text{ReT}_{\mathbf{D}}$. Apart from saving memory during training, switching to a shared architecture – *i.e.*, *shared architecture (32 tokens)* – raises the average score to 62.6, with an improvement of +1.2 points compared to having separate encoders. As most substantial gains come from InfoSeek and Encyclopedic-VQA, which present tens to hundreds of questions for the same Wikipedia entity, we credit the shared architecture approach for reducing overfitting on entities seen during training.

Token Reduction. The next change arises from an analysis of the output matrix of ReT, revealing that it suffers from rank collapse. Empirically, we register the rank collapse score [57] of the 32-row matrix generated by ReT when embedding samples from the InfoSeek test split of M2KR. The last recurrent step of ReT outputs $32 \times 1{,}024$ matrices. For them, we register an average rank collapse score of 0.18 when embedding queries and 0.22 when embedding documents. After applying the late-interaction linear projection to 32×128 dimensions, the average rank collapse scores further plummet to 0.09 and 0.11. Ideally, those scores would tend to 1.0, and our analysis indicates that the 32 token embeddings of the output matrix converge to a unified representation. Consequently, the purpose of using multiple token embeddings to represent inputs is questionable. This motivates the exploration of a token reduction strategy, by applying score fusion to a number of tokens equal to 16, 8, 4, and 1 (*i.e.*, no score fusion at all). While reducing the number of tokens initially seems to degrade performance, we register an average improvement of +0.2 points when switching from 32 tokens to a single one – *i.e.*, *shared architecture (single token)*. Notably, this happens along with a reduction in trainable parameters and less computation, as with a single token, there is no need to apply self-attention in the recurrent cell of ReT. For completeness, we also include the single token version of ReT without sharing weights between the query and document encoders – *i.e.*, *non-shared architecture (single token)*. We provide a theoretical account of this rank-collapse behavior, connecting it to self-attention networks [58] and to the role of the forget gate in Eq. (5), in Appendix B.

Layer Pruning. Driven by the computational constraints of ReT, primarily due to the recurrent cell being applied to a predefined number of layers ranging from 12 to 16, we explore a layer pruning strategy to improve efficiency. In detail, we

TABLE IV
EXPERIMENTAL RESULTS ON THE M2KR BENCHMARK [18], COMPARING RET-2 TO BASELINES AND COMPETITORS WHEN VARYING THE VISUAL BACKBONE. BOLD FONT DENOTES THE BEST RESULTS UNDER THE SAME BACKBONE. THE † MARKER DENOTES OUR REPRODUCTIONS.

		WIT	**IGLUE**	**KVQA**	**OVEN**	**LLaVA**	**InfoSeek**		**E-VQA**		**OKVQA**		
Model	**Backbone**	R@10	R@1	R@5	R@5	R@1	R@5	PR@5	R@5	PR@5	R@5	PR@5	**Avg**
CLIP (ZS)	CLIP ViT-B	48.9	63.1	**57.8**	58.1	33.0	33.6	47.4	0.13	12.1	0.52	49.9	36.8
FLMR [36]	CLIP ViT-B	23.8	-	31.9	40.5	56.4	-	47.1	-	-	-	68.1	-
PreFLMR [18]	CLIP ViT-B	41.7	57.3	28.6	46.3	67.2	26.0	48.8	**55.0**	**67.9**	**27.2**	**66.1**	48.4
ReT [2]	CLIP ViT-B	60.1	73.9	26.9	72.9	76.6	30.2	48.1	33.0	48.9	13.9	58.3	49.3
ReT-2 (Ours)	CLIP ViT-B	**68.3**	**76.1**	56.6	**73.8**	**81.2**	**36.9**	**52.7**	36.1	52.9	12.0	60.7	**55.2**
ReT-2 (Ours)	CLIP ViT-B 🔥	**73.7**	**77.7**	**66.6**	**77.3**	**86.0**	**38.3**	**53.8**	42.0	57.6	14.9	62.6	**59.1**
CLIP (ZS)	CLIP ViT-L	65.9	74.9	73.3	68.5	36.6	48.0	58.4	0.17	12.0	0.59	49.2	45.0
PreFLMR [18]	CLIP ViT-L	60.5	69.2	43.6	59.8	71.8	37.4	57.9	**60.9**	**70.8**	**31.4**	**68.5**	57.4
ReT [2]	CLIP ViT-L	73.4	81.8	63.5	82.0	79.9	47.0	60.5	44.5	57.9	20.2	66.2	61.5
ReT-2 (Ours)	CLIP ViT-L	**81.1**	**82.9**	**72.3**	**83.1**	**83.8**	**48.0**	**61.0**	49.7	62.6	15.2	65.9	**64.1**
ReT-2 (Ours)	CLIP ViT-L 🔥	**86.1**	**84.4**	**78.1**	**86.8**	**88.6**	**49.1**	**62.3**	56.4	67.4	20.0	67.8	**67.9**
SigLIP2 (ZS)	SigLIP2 ViT-L	51.9	60.0	**48.4**	74.3	41.1	51.4	60.4	19.5	33.2	6.1	50.1	45.1
PreFLMR [18]†	SigLIP2 ViT-L	68.3	**76.1**	39.1	71.5	73.5	42.9	59.5	**51.6**	**64.1**	**17.8**	**70.6**	57.7
ReT [2]	SigLIP2 ViT-L	65.7	71.8	34.8	81.1	75.1	42.2	56.4	35.2	51.2	15.4	63.3	53.8
ReT-2 (Ours)	SigLIP2 ViT-L	**70.3**	71.2	48.2	**85.3**	**81.8**	**57.1**	**65.5**	44.5	58.1	10.8	61.5	**59.5**
ReT-2 (Ours)	SigLIP2 ViT-L 🔥	**80.6**	**79.4**	**61.8**	**88.8**	**89.4**	**59.7**	**67.7**	**51.6**	63.5	**21.2**	70.5	**66.7**
OpenCLIP (ZS)	OpenCLIP ViT-H	74.2	78.2	**68.0**	78.4	45.3	**53.2**	61.3	20.8	33.3	7.3	63.9	53.1
PreFLMR [18]	OpenCLIP ViT-H	60.5	71.2	39.4	61.5	72.3	39.2	59.5	**62.5**	**71.7**	**30.2**	**68.1**	57.8
ReT [2]	OpenCLIP ViT-H	71.4	80.0	59.3	83.0	79.8	47.3	60.7	44.8	57.8	18.2	63.4	60.5
ReT-2 (Ours)	OpenCLIP ViT-H	**80.2**	**82.3**	66.2	**83.3**	**86.1**	52.8	**63.1**	45.9	59.3	14.4	64.0	**63.4**
ReT-2 (Ours)	OpenCLIP ViT-H 🔥	**85.5**	**84.2**	**75.8**	**88.4**	**91.1**	**58.0**	**66.7**	58.9	69.3	18.3	65.1	**69.2**

sample a total of three layers, corresponding to the early, middle, and late stages of both the visual and textual backbone[1]. This strategy guarantees to preserve information from different abstraction levels, and it has been recently proven effective for the visual-language alignment of MLLMs [70]. Our choice is further supported by an empirical analysis of the average gate activations of ReT, conducted on the InfoSeek and Encyclopedic-VQA test splits of M2KR. As shown in Fig. 5, the visual input gate exhibits three prominent activation peaks, aligning with the selected layer groups. On the other hand, the textual input gate has a smoother behavior, peaking mainly across early-to-middle stages, thus highlighting the importance of including low-level textual features. Quantitative experimental results validate the effectiveness of this pruning strategy: not only does it preserve retrieval performance, but it also yields a +0.1 points improvement in accuracy. Beyond this empirical evidence, Appendix B provides a theoretical motivation for this design, arguing that a shorter recurrent chain is easier to retain over [21], [71], [72], and that consecutive backbone layers, being connected by residual paths, tend to carry redundant evidence [73]–[75], exposing the cell's gates to correlated – rather than complementary – decisions when no pruning is applied.

Global Feature Injection. Finally, we incorporate global features, obtaining our final ReT-2 model. In detail, we apply score fusion by summing the multimodal, multilayer token coming from the recurrent cell with the pooler token of the visual backbone and the one from the textual backbone. This raises the average score to 64.1, with a +2.6 points improvement over ReT. For fairness of comparison, we apply global feature injection to ReT as well (*i.e.*, gray row). In this setting, the pooler tokens are first projected to dimension 128 and then concatenated to the 32 tokens of the recurrent cell. We highlight that, even in this scenario, global features raise the performance of ReT, while still falling behind ReT-2. We also point out that the favorable impact of adding global features plays a synergistic role with the multi-layer, multimodal features generated by the recurrent cell. Indeed, if we drop the recurrent cell and only optimize the pre-trained backbones to generate better global features, the performance lowers considerably – *i.e.*, - *recurrent cell*.

Per-Component Contributions. To clarify the role of each design choice, we repeat the ablation studies presented so far, this time starting from our final configuration, removing one component at a time independently (Table II, last four rows). First, using dedicated encoders for queries and documents (*i.e.*, - *shared architecture*) drops the overall score by -4.0 points. Second, replacing the score fusion paradigm with the original fine-grained late-interaction scheme of ReT (*i.e.*, - *score fusion (32 tokens)*) drops the average score by -2.8 points. Third, setting the number of learnable tokens back to 32 (*i.e.*, - *token reduction (32 tokens)*) yields a modest drop of 0.5 points, suggesting limited benefit from multiple tokens despite the additional computational cost. Finally, applying the recurrent cell to all layers (*i.e.*, - *layer pruning*) rather than choosing a single layer from the early, middle, and late sections of the backbones shows comparable results (-0.3 points), justifying our layer pruning design to save time and computation.

Layer Selection. To validate the robustness of our layer selection strategy, we additionally evaluate alternative layer

[1]Because in Table II ReT-2 is paired with CLIP ViT-L/14, it follows that we employ the third, eighteenth, and second last layer from the visual backbone, and the third, seventh, and second last layer from the textual backbone. We refer to Table I for the layer selection in backbones with different depths.

TABLE V
EXPERIMENTAL RESULTS ON THE M2KR BENCHMARK [18], COMPARING RET-2 TO BASELINES AND COMPETITORS WHEN EMPLOYING COLBERTV2 [54] AS TEXTUAL BACKBONE. † INDICATES OUR REPRODUCTIONS.

		WIT	IGLUE	KVQA	OVEN	LLaVA	InfoSeek		E-VQA		OKVQA		
Model	**Backbone**	R@10	R@1	R@5	R@5	R@1	R@5	PR@5	R@5	PR@5	R@5	PR@5	**Avg**
PreFLMR [18]	CLIP ViT-L	60.5	69.2	43.6	59.8	71.8	37.4	57.9	**60.9**	**70.8**	**31.4**	**68.5**	57.4
ReT [2]	CLIP ViT-L	73.9	79.3	48.6	79.6	79.6	40.0	58.9	43.4	59.0	19.0	64.1	58.7
ReT-2 (Ours)	CLIP ViT-L	**78.6**	**80.3**	**48.8**	**81.2**	**80.3**	**50.9**	**64.9**	47.1	62.1	14.8	62.1	**61.0**
ReT-2 (Ours)	CLIP ViT-L 🔥	**81.9**	**81.0**	**62.9**	**83.7**	**84.8**	**52.1**	**66.2**	55.1	67.8	16.7	64.2	**65.1**
PreFLMR [18]†	SigLIP2 ViT-L	68.3	76.1	39.1	71.5	73.5	42.9	59.5	**51.6**	**64.1**	**17.8**	**70.6**	57.7
ReT [2]	SigLIP2 ViT-L	65.7	71.8	34.8	81.8	75.1	42.2	56.4	35.2	51.2	15.4	63.3	53.9
ReT-2 (Ours)	SigLIP2 ViT-L	**78.9**	**79.1**	**48.6**	**84.4**	**83.0**	**53.7**	**66.3**	49.1	63.2	15.2	61.9	**62.1**
ReT-2 (Ours)	SigLIP2 ViT-L 🔥	**82.7**	**82.5**	**56.4**	**86.6**	**86.1**	**59.4**	**68.6**	**56.5**	**68.6**	17.3	67.4	**66.5**

TABLE VI
RESULTS ON THE M2KR BENCHMARK WHEN TRAINING RET-2 WITH THE 3-STAGE PIPELINE PROPOSED IN [18]. OUR ORIGINAL SINGLE-STAGE VERSION IS REPORTED IN GRAY. ALL MODELS EMPLOY CLIP VIT-L/14 AS VISUAL BACKBONE WHICH IS KEPT FROZEN IN ALL TRAINING STAGES.

			WIT	IGLUE	KVQA	OVEN	LLaVA	InfoSeek		E-VQA		OKVQA		
Model	**Textual Backbone**	**GPU Hrs**	R@10	R@1	R@5	R@5	R@1	R@5	PR@5	R@5	PR@5	R@5	PR@5	**Avg**
PreFLMR [18]	ColBERTv2 🔥	864	60.5	69.2	43.6	59.8	71.8	37.4	57.9	60.9	70.8	**31.4**	68.5	57.4
ReT-2 (Ours)	ColBERTv2 🔥	260	77.9	79.3	57.0	81.6	80.9	**51.7**	**65.7**	**62.0**	**71.8**	25.3	69.1	65.7
ReT-2 (Ours)	CLIP ViT-L 🔥	212	**81.9**	**83.7**	**75.0**	**84.4**	**84.9**	50.6	64.6	61.9	70.3	26.4	**72.8**	**68.8**
ReT-2 (Ours)	CLIP ViT-L	80	81.1	82.9	72.3	83.1	83.8	48.0	61.0	49.7	62.6	15.2	65.9	64.1

configurations in Table III. In particular, we evaluate models that use only shallow (early), middle, or deep (late) layers, as well as combinations of these (*e.g.*, early+middle, early+late, and middle+late). The results show that relying on a single depth consistently underperforms, while combining layers improves performance. Note that, using a single layer is equivalent to *bypassing the recurrence at all*, as the network showed in Fig. 3 runs only once. These results confirm that the chosen layers are meaningful, providing complementary information from multiple depths of the network.

In summary, our final model, ReT-2, fuses multimodal and multi-layer features into a single learnable token, shares parameters between the query and document encoders, and incorporates global features. This design achieves superior performance without relying on the computationally expensive fine-grained contrastive loss, and is adopted as the final model for all subsequent experiments.

D. Comparison with the State of the Art

Results on the M2KR Benchmark. Table IV presents a comparison of our proposed method, ReT-2, against a zero-shot CLIP baseline and other retrieval approaches. These include FLMR [36] and PreFLMR [18], two multimodal retrieval models trained on M2KR. Both models adopt a multimodal query and a text-only document setting. FLMR relies on the `CLS` token for image representation, whereas PreFLMR enriches visual information using patch embeddings from the penultimate layer, capturing more fine-grained features. For reference, we also report results from our earlier model, ReT [2]. We also include a variant of ReT-2 in which the visual and textual backbones are unfrozen during training (🔥).

Across all datasets and backbones, ReT-2 consistently outperforms the original ReT. For example, on WIT with a frozen CLIP ViT-L backbone, ReT-2 achieves a substantial gain of +7.7 points over ReT (81.1 vs. 73.4). When compared to other state-of-the-art methods, ReT-2 achieves the best average performance in most settings, with the only exceptions being Encyclopedic-VQA and OKVQA, where PreFLMR slightly outperforms it. In this regard, we notice that PreFLMR employs a three-stage training pipeline, with the second stage being dedicated to Encyclopedic-VQA and the third stage entailing a careful balancing and resampling of each sub-dataset. In contrast, our ReT-2 models are trained in a single-stage run on the entire M2KR dataset. The trainable variant of ReT-2 (🔥) further boosts performance – for instance, with the SigLIP2 backbone, the trainable version delivers an average improvement of +7.2 points. Similar trends are observed across all backbone architectures: CLIP ViT-B shows improvement from 55.2 to 59.1, CLIP ViT-L from 64.1 to 67.9, and OpenCLIP ViT-H from 63.4 to 69.2. Finally, scaling the unfrozen visual backbone also correlates with stronger retrieval results: average performance increases from 59.1 with CLIP ViT-B, to 67.9 with CLIP ViT-L, and 69.2 with OpenCLIP ViT-H. In contrast, when the backbones are frozen, we observe a similar trend to that reported in both ReT and PreFLMR: the larger OpenCLIP ViT-H underperforms relative to the smaller CLIP ViT-L, suggesting that the benefits of scaling depend on the dataset and experimental setting.

To provide a fairer comparison with the original PreFLMR model, which uses ColBERTv2 [54] as its textual backbone, in Table V we also report the results obtained when replacing the textual backbone in both ReT and ReT-2 with ColBERTv2. This evaluation is conducted using both CLIP and SigLIP2 ViT-L visual backbones, ensuring consistency and comparability across architectures. As it can be seen, the performance trends remain consistent: ReT-2 continues to outperform both the original ReT and PreFLMR, even when matched on backbone architecture. The largest improvements

TABLE VII
EXPERIMENTAL RESULTS ON THE M-BEIR$_{\text{LOCAL}}$ BENCHMARK [1]. † INDICATES OUR REPRODUCTIONS, AND GRAY DENOTES MLLM-BASED METHODS.

		#1			#2	#3		#4			#5	#6		#7		#8		
	Backbone	**VN**	**COCO**	**F200k**	**WQA**	**EDIS**	**WQA**	**VN**	**COCO**	**F200k**	**NIGHTS**	**OVEN**	**InfoSeek**	**FIQ**	**CIRR**	**OVEN**	**InfoSeek**	**Avg**
CLIP (ZS)	CLIP ViT-L	43.4	61.1	6.6	36.2	43.3	45.1	41.3	79.0	7.7	26.1	24.2	20.5	7.0	13.2	38.8	26.4	32.5
SigLIP2 (ZS)	SigLIP2 ViT-L	40.0	77.5	34.8	33.7	27.3	42.5	40.4	88.1	35.3	28.4	30.0	30.2	20.4	29.3	41.9	34.3	39.6
PreFLMR [18]	CLIP ViT-L	-	-	-	68.1	21.8	37.6	0.1	8.1	0.0	-	19.9	21.7	-	-	27.4	23.5	-
ReT [2]	CLIP ViT-L	23.2	66.3	12.3	47.0	47.1	56.9	23.0	85.5	9.5	21.5	39.0	21.4	10.6	27.1	57.3	33.9	36.3
ReT [2]	CLIP ViT-L 🔥	24.2	72.8	14.5	54.3	48.5	65.6	24.1	87.6	15.7	25.6	37.5	20.2	13.0	37.2	56.3	35.2	39.5
GENIUS [35]	CLIP ViT-L 🔥	27.4	78.0	16.2	44.6	44.3	60.6	28.4	91.1	16.3	30.2	41.9	20.7	19.3	39.5	52.5	30.1	40.1
UniIR [1]	BLIP ViT-L 🔥	23.4	79.7	26.1	80.0	50.9	79.8	22.8	89.9	28.9	**33.0**	41.0	22.4	29.2	52.2	55.8	33.0	46.8
UniIR [1]	CLIP ViT-L 🔥	42.6	81.1	18.0	84.7	59.4	78.7	43.1	92.3	18.3	32.0	45.5	27.9	24.4	44.6	67.6	48.9	50.6
UniIR [1]†	SigLIP2 ViT-L 🔥	29.4	78.1	21.6	75.3	49.9	77.6	33.0	91.1	44.5	29.5	52.9	27.9	33.1	54.0	71.2	**50.7**	51.2
ReT-2 (Ours)	CLIP ViT-L 🔥	**47.3**	80.2	21.1	**86.0**	**56.7**	**80.2**	**46.8**	91.6	22.7	31.5	48.7	27.5	23.8	44.3	69.1	47.0	51.5
ReT-2 (Ours)	SigLIP2 ViT-L 🔥	38.9	**84.8**	**50.0**	76.3	53.7	78.4	42.0	**95.0**	**52.2**	31.5	**54.1**	**32.3**	**35.3**	**57.1**	**72.1**	48.3	**56.4**
MM-Embed [41]	LLaVA-NeXT-7B	41.0	71.3	17.1	95.9	68.8	85.0	41.3	90.1	18.4	32.4	42.1	42.3	25.7	50.0	64.1	57.7	52.7
JFE [43]	PaliGemma-3B	34.6	78.5	37.2	88.7	54.3	82.4	33.1	90.0	36.9	27.8	46.0	35.6	31.8	54.0	72.7	61.1	54.0
PUMA [42]	Qwen2-VL-7B	35.7	79.5	25.8	86.2	35.2	90.1	29.0	31.4	58.2	78.4	52.7	48.3	30.6	49.9	74.0	65.2	54.4
LamRA [40]	Qwen2-VL-7B	41.6	81.5	28.7	86.0	62.6	81.2	39.6	90.6	30.4	32.1	54.1	52.1	33.2	53.1	76.2	63.3	56.6

TABLE VIII
EXPERIMENTAL RESULTS ON THE MMEB BENCHMARK [26] IN A ZERO-SHOT SETTING (NO TRAINING ON THE MMEB SPLIT).

Model	**Classification**	**VQA**	**Retrieval**	**Grounding**	**All**
CLIP [9]	42.8	9.1	53.0	51.8	37.8
BLIP-2 [76]	27.0	4.2	33.9	47.0	25.2
MagicLens [77]	38.8	8.3	35.4	26.0	27.8
E5-V [78]	21.8	4.9	11.5	19.0	13.3
UniIR [1]	44.3	16.2	61.8	62.2	44.3
ReT-2 (CLIP)	44.2	16.3	64.3	62.8	45.0
ReT-2 (SigLIP2)	**55.3**	**16.7**	**64.5**	**59.2**	**48.1**

are again observed with the trainable variant of ReT-2, yielding average gains of +6.4 and +12.6 points over ReT when using CLIP and SigLIP2 ViT-L, respectively.

Additionally, since ReT-2 models are trained in a single run on all the data from M2KR, in Table VI we explore the more complex training scheme proposed by PreFLMR [18]. This consists of a 3-stage training: first, it begins with an extensive pre-training on the full M2KR data, followed by a fine-tuning on in-domain data from Encyclopedic-VQA [20]; finally, a new training is run over the full M2KR data, this time with a careful balancing of the data sampled from each dataset. In particular, data from OKVQA are up-sampled 10 times, while data from Encyclopedic-VQA is upsampled twice. Moreover, the text encoder is kept frozen during the first pre-training run, while it is free during the subsequent stages. We train variants of ReT-2 following these settings, matching the total amount of training samples seen by PreFLMR at each stage. As shown, while the gain on all datasets are preserved, the fine-tuning on Encyclopedic-VQA data greatly increases the performance on this dataset, surpassing the results of PreFLMR. Similarly, the scores on OKVQA are now much closer to PreFLMR in terms of R@5, and even better concerning PR@5, probably caused by the upsample of OKVQA data in the final training stage. These results indicate that the training mixture plays a key role in optimizing multimodal retrievers. Nevertheless, our single-stage training strategy (*i.e.*, last gray row) remains simpler, avoids dataset rebalancing, and reduces training cost, while still achieving state-of-the-art performance. Fine-tuning on in-domain data can further benefit domain-specific scenarios.

Results on M-BEIR Benchmark. In Table VII, we further evaluate the generalization capability of our proposed approach on M-BEIR$_{\text{local}}$. The benchmark comprises eight distinct tasks, each presenting different modality configurations and challenges[2]. In this setting, we compare ReT-2 with zero-shot baselines and competitors like UniIR [1], GENIUS [35], and the previous version of our model (*i.e.*, ReT). Specifically, UniIR proposes strategies for encoding multimodal queries and documents, by leveraging pre-trained models like CLIP and BLIP [34] to integrate different modalities. In this table, we also include our reproduction of UniIR using the SigLIP2 backbone to ensure a fair and consistent comparison. GENIUS, on the other hand, is a versatile generative retrieval framework that discretizes multimodal inputs. As additional competitors, we include retrieval models based on MLLMs, such as MM-Embed [41], JFE [43], PUMA [42], and LamRA [40]. Due to their significantly larger model sizes and parameter counts, these methods are not directly comparable to ours.

The results show that ReT-2, using both the CLIP and SigLIP2 ViT-L backbones, significantly outperforms not only the original ReT version but also all other competitors. For instance, the SigLIP2 variant of ReT-2 achieves a notable improvement of +5.2 points over UniIR using the same backbone. Remarkably, despite being smaller in size and not relying on an LLM, the variant of ReT-2 based on SigLIP2 delivers the best overall performance compared to nearly all MLLM-based competitors, falling just short of the LamRA model, which achieves only a +0.2-points average improvement.

Results on MMEB Benchmark. We further evaluate ReT-2 on MMEB [26] which spans diverse multimodal tasks across four different categories: classification, visual question answering, retrieval, and visual grounding. We focus on ReT-2 models trained on M-BEIR, as it better reflects the heterogeneous settings of MMEB. For comparison, we consider models reported in [26] that are not specifically fine-tuned on MMEB, including CLIP [9], BLIP-2 [76], MagicLens [77], E5-V [78], and UniIR [1]. Table VIII report the results averaged on each category. As shown, ReT-2 with the CLIP backbone already outperforms all competing methods, achieving a score on

[2]A detailed description of each task is provided in Appendix A.

TABLE IX
COMPARISON OF TRAINING RESOURCES AND INFERENCE TIMES BETWEEN RET-2 AND COMPETING METHODS.

	Training Info			Inference Time (ms)			
Model	Backbones	#GPUs	Hrs	Forward	Retrieval	All ↓	#Tokens
CLIP (ZS)	T ❄ V ❄	-	-	18.6	0.7	19.3	1
SigLIP2 (ZS)	T ❄ V ❄	-	-	19.2	0.8	20.0	1
PreFLMR [18]	T 🔥 V ❄	4	864	32.7	406.1	438.8	320
UniIR [1]	T 🔥 V 🔥	8	72	23.8	0.8	24.6	1
LamRA [40]	MLLM 🔥	16	N/A	52.7	1.5	54.2	1
ReT [2]	T ❄ V ❄	4	80	31.4	3.5	34.9	32
ReT-2 (Ours)	T ❄ V ❄	4	80	26.8	0.8	27.6	1
ReT-2 (Ours)	T 🔥 V 🔥	4	160	26.8	0.8	27.6	1

the retrieval category of 64.3 against 61.8 of UniIR, which represents the strongest baseline. Switching to the SigLIP2 backbone further boosts the retrieval score to 64.5, and raises the overall average to 48.1, confirming that the gains observed on M2KR and M-BEIR generalize to this broader benchmark.

E. Computational Analysis

In Table IX, we provide a computational analysis of ReT-2 and competitors in terms of resource demand for training and inference speed. The analysis employs a subset of the InfoSeek dataset comprising 100k image-text passages and 4.7k image-text queries. For CLIP ViT-L and SigLIP2, which we include as baselines for image-text retrieval, we mask out text on the query side and images on the document side. For ReT and PreFLMR, we follow the implementation in [37] to index passages, enabling efficient fine-grained late-interaction retrieval through GPU acceleration. This implementation runs the forward pass of the models in full precision, so we stick with full precision to measure the forward time of all the models. An exception is LamRA, which we run in half precision to account for the additional memory requirements due to its 7B MLLM backbone. For the other methods, we build a `GpuIndexFlat` using the Faiss library. All experiments are run on a single NVIDIA A100 GPU (64GB of VRAM).

Notably, ReT-2 benefits from the introduced layer pruning strategy and the use of a single input token to embed queries and documents, resulting in significantly faster forward and retrieval times compared to ReT and PreFLMR, which rely on the more computationally intensive fine-grained late-interaction paradigm. Compared with UniIR, ReT-2 demonstrates competitive retrieval speed while generally requiring equal or lower training resources, depending on whether the unimodal backbones are trained together with the recurrent retrieval cell or kept frozen. It is worth noting that LamRA takes nearly twice the forward and retrieval time of ReT-2, not to mention the additional storage required for saving embeddings of size 3,584 rather than 768 as in our model. Ultimately, the decision to rely on MLLMs rather than smaller encoders based on vision-language foundation models is a trade-off between performance and efficiency.

VI. EXPERIMENTS ON RETRIEVAL-AUGMENTED VQA

As a more realistic use case, we evaluate our approach for retrieval-augmented generation in knowledge-intensive VQA, where an off-the-shelf MLLM must answer visual questions requiring detailed knowledge of a specific entity (*e.g.*, the subject of a Wikipedia page). Since such questions are often unanswerable without external knowledge, we assess the effectiveness of ReT-2 in retrieving relevant context to help the MLLM answer the questions correctly.

A. Datasets and Evaluation Metrics

Encyclopedic-VQA Dataset [20]. It contains visual questions related to a Wikipedia entity. The test set counts 5,750 questions, of which 1,000 are two-hop questions, meaning that two Wikipedia pages should be retrieved sequentially, with the correct answer lying in the second one. Results are evaluated in terms of accuracy, with an answer being counted as correct if its BEM score [79] with respect to the ground-truth is higher than 0.5. The official knowledge base consists of 2M Wikipedia pages, each divided into several sections, possibly attached with an image. As we are interested in the multimodal retrieval task, we split each Wikipedia page into multiple image-text documents, with the text being the content of the section. Concerning the visual component, we have three scenarios: we select the image directly linked to the specific section; if unavailable, we fall back to the first image associated with the entire Wikipedia page, often corresponding to the first picture appearing in the web page; if neither option exists, we omit the image, obtaining a text-only document. Following this protocol, we collect 15.9M documents in total.

InfoSeek Dataset [19]. Similarly, InfoSeek entails visual questions about Wikipedia entities. The test set annotations have not been publicly released, so we report the performance on the 73,620 questions of the validation set. A question can be of type *string*, *numeric*, or *time*, and is marked as either *unseen question* or *unseen entity*, based on the given question or the referring Wikipedia entity being not present in the training set. The evaluation metric is the harmonic mean between the accuracy on the unseen question and unseen entity splits, both computed with an exact matching criterion. The official knowledge base of InfoSeek contains as many as 6M Wikipedia pages. However, only a subset of them is typically used for evaluation. Thus, to be consistent with prior research, we stick with the 100k pages in the knowledge base proposed in previous works [80], [81]. Different from Encyclopedic-VQA, these pages are not divided into sections, so we follow ReT [2] and split each page into chunks of 100 words with the format `Title: [WikiTitle]; Content: [...]`. If an image is available for a given Wikipedia page, we attach it to all of its text chunks, creating image-text documents. This process builds up a knowledge base of 1.02M documents.

B. Implementation Details

We run experiments with two different MLLMs, namely LLaVA-MORE-8B [16], built upon LLaMA-3.1-8B [39], and the more recent Qwen2.5-VL-7B [17]. In all experiments, we prompt[3] the MLLM with the text content of the top-k retrieved documents, using k equal to 3. Because InfoSeek relies on

[3]The exact prompt used in our experiments is reported in Appendix A.

TABLE X
VQA ACCURACY SCORES ON THE ENCYCLOPEDIC-VQA TEST SET AND THE INFOSEEK VALIDATION SET COMPARING OFF-THE-SHELF MLLMS WITH DIFFERENT RETRIEVAL MODELS.

		E-VQA		InfoSeek		
Model	**Retrieval Model**	Single-Hop	All	Un-Q	Un-E	All
BLIP-2 [76]	-	12.6	12.4	12.7	12.3	12.5
InstructBLIP [82]	-	11.9	12.0	8.9	7.4	8.1
LLaVA-1.5-7B [15]	-	16.3	16.9	9.6	9.4	9.5
LLaVA-MORE-8B [16]	-	13.8	14.9	8.9	8.0	8.4
LLaVA-MORE-8B [16]	CLIP ViT-L	17.9	19.0	14.5	13.6	14.1
LLaVA-MORE-8B [16]	SigLIP2 ViT-L	17.5	18.6	16.0	15.1	15.5
LLaVA-MORE-8B [16]	PreFLMR [18]	27.8	26.9	13.0	11.7	12.3
LLaVA-MORE-8B [16]	ReT [2]	21.9	21.8	21.1	15.0	17.5
LLaVA-MORE-8B [16]	UniIR [1]	16.9	18.2	**25.1**	18.8	21.5
LLaVA-MORE-8B [16]	**ReT-2 (Ours)**	**28.5**	**27.1**	24.3	**21.5**	**22.8**
Qwen2.5-VL-7B [17]	-	19.8	19.7	18.6	18.1	18.3
Qwen2.5-VL-7B [17]	CLIP ViT-L	19.5	20.4	18.7	17.9	18.3
Qwen2.5-VL-7B [17]	SigLIP2 ViT-L	20.1	20.9	19.8	19.5	19.7
Qwen2.5-VL-7B [17]	PreFLMR [18]	**34.4**	**33.0**	18.0	15.8	16.8
Qwen2.5-VL-7B [17]	ReT [2]	26.6	26.2	24.5	17.9	20.7
Qwen2.5-VL-7B [17]	UniIR [1]	18.6	19.2	**29.0**	22.4	25.3
Qwen2.5-VL-7B [17]	**ReT-2 (Ours)**	33.5	31.6	27.9	**25.1**	**26.4**

TABLE XI
VQA ACCURACY SCORES ON THE ENCYCLOPEDIC-VQA TEST SET AND INFOSEEK VALIDATION SET COMPARING TASK-SPECIFIC MODELS EQUIPPED WITH DIFFERENT RETRIEVAL MODULES.

	Retrieval		E-VQA		InfoSeek		
Model	**Model**	**Size**	Single-Hop	All	Un-Q	Un-E	All
Wiki-LLaVA-7B [80]	CLIP+Contr.	0.6B	17.7	20.3	30.1	27.8	28.9
EchoSight-8B [83]	EVA-CLIP	8B	26.4	24.9	30.0	30.7	30.4
VLM-PRF [84]	EVA-CLIP	8B	40.1	39.2	43.5	42.1	42.5
ReflectiVA-3B [81]	EVA-CLIP	8B	33.7	35.2	39.6	38.1	38.9
	PreFLMR [18]	0.5B	28.5	31.6	32.7	29.0	30.8
	UniIR [1]	0.4B	36.9	38.3	**36.9**	32.1	34.3
	ReT-2 (Ours)	0.5B	**38.7**	**40.1**	36.5	**33.5**	**34.9**
ReAG-3B [85]	EVA-CLIP	8B	41.3	42.9	43.7	42.9	43.3
	PreFLMR [18]	0.5B	28.5	31.4	34.0	30.7	32.3
	UniIR [1]	0.4B	38.0	39.2	41.3	36.8	38.9
	ReT-2 (Ours)	0.5B	**38.7**	**40.0**	**43.2**	**39.9**	**41.5**
ReflectiVA-7B [81]	EVA-CLIP	8B	36.8	36.8	43.5	44.3	43.9
	PreFLMR [18]	0.5B	28.7	31.5	35.4	32.9	34.1
	UniIR [1]	0.4B	36.4	**38.9**	37.8	36.7	37.2
	ReT-2 (Ours)	0.5B	**36.5**	38.2	**38.5**	**37.6**	**38.0**
ReAG-7B [85]	EVA-CLIP	8B	44.9	47.0	48.3	46.2	47.2
	PreFLMR [18]	0.5B	28.4	31.6	39.8	35.4	37.5
	UniIR [1]	0.4B	42.6	44.1	46.1	40.6	43.2
	ReT-2 (Ours)	0.5B	**42.9**	**44.7**	**47.6**	**43.5**	**45.5**

exact matching to evaluate answers, we prepend its prompt with 3-shot examples, one for each question type, to teach the MLLM how to format the answer. Generation is done through beam search decoding, using a beam size of 5 for LLaVA-MORE and a beam size of 3 for Qwen2.5-VL, limiting the number of generated tokens to 20 due to memory constraints.

For retrieval, we build image-text queries with the image being taken directly from the visual question, while we prepend the question with an instruction taken from the M2KR templates [18]. Finally, in Encyclopedic-VQA, we do not differentiate between single and two-hop questions, running a single-step retrieval even in a two-hop context.

C. Experimental Results

For these experiments, we compare ReT-2 against other multimodal retrievers, such as ReT, UniIR, and PreFLMR, and also include the results of the original CLIP and SigLIP2 models as baselines. Note that for all considered models, we employ their best-performing configurations in this setting. The experimental results are reported in Table X, which also includes off-the-shelf MLLMs without retrieval as references, namely BLIP-2 [76], InstructBLIP [82], and LLaVA-1.5-7B [15], as well as the two MLLMs used in our experiments (*i.e.*, LLaVA-MORE-8B [16] and Qwen2.5-VL-7B [17]).

First of all, we observe that in general, both LLaVA-MORE and Qwen2.5-VL benefit from retrieval-augmented generation, demonstrating the challenge posed by Encyclopedic-VQA and InfoSeek. When LLaVA-MORE is used as the generator, ReT-2 stands out as the best multimodal retriever across both benchmarks, even outscoring PreFLMR on Encyclopedic-VQA, despite PreFLMR having undergone a dedicated training stage on that dataset. This suggests that at a large scale, the fine-grained late-interaction mechanism may be exposed to the size of the knowledge base more severely than single-token retrieval. Switching to Qwen2.5-VL, the results are better than LLaVA-MORE, testifying the superior capabilities of this more recent MLLM. In this context, ReT falls slightly behind PreFLMR on Encyclopedic-VQA, but compensates for that by confirming itself as the best retriever on InfoSeek, scoring 9.6 points higher than PreFLMR, which even underperforms compared to Qwen2.5-VL without retrieval. Overall, these results confirm the effectiveness of our approach, showing that off-the-shelf MLLMs can achieve competitive performance in knowledge-intensive VQA without task-specific fine-tuning.

D. ReT-2 as Retriever for Task-Specific Methods

In the experiments of Table X on general-purpose MLLMs, the considered retrieval models are specifically designed to directly provide the MLLM with a chunk of text from the knowledge base. In contrast, state-of-the-art methods for knowledge-intensive VQA [80], [81], [83], [85] typically involve fine-tuning the MLLM and, in many cases, a two-stage retrieval pipeline. In this setting, the first stage retrieves multimodal candidate documents, while the second stage refines the selection by extracting the most relevant textual passages. As a result, the role of the retriever is primarily to identify the entity referred to by the multimodal query. This entity is usually associated with a document (*e.g.*, a full Wikipedia page), which is then processed by an additional module for filtering and refinement. For instance, ReflectiVA [81] employs the MLLM itself to select relevant passages from the retrieved document, whereas ReAG [85] introduces a secondary LLM (*i.e.*, the critic) to filter irrelevant content. Moreover, these models are typically fine-tuned on the target datasets to better exploit the retrieved knowledge.

Accordingly, while previous experiments evaluate retrievers based on their ability to return *fine-grained* passages likely to contain the answer, here we instead assess their effectiveness as *coarse* search engines, retrieving broadly relevant documents for subsequent refinement by downstream components. Table XI reports the results of this analysis, where we compare ReT-2, PreFLMR, and UniIR as coarse retrieval

engines within two state-of-the-art pipelines, ReflectiVA and ReAG, both based on Qwen2.5-VL [17] at the 3B and 7B scales. We follow the original implementation settings and knowledge bases for both Encyclopedic-VQA (2M Wikipedia pages) and InfoSeek (100k Wikipedia pages). For each query, the top-k Wikipedia pages are retrieved based on the page summary and the associated primary image, when available (with $k = 20$ following [85]). As a reference, we report the performance of the original models equipped with EVA-CLIP 8B as retriever, which is nearly $16\times$ larger than ReT-2 in terms of parameters and requires approximately $3.5\times$ more storage for embeddings. Under this setting, ReT-2 consistently outperforms PreFLMR and UniIR across both pipelines and model scales, achieving performance comparable to significantly larger retrieval backbones and further demonstrating its effectiveness in downstream retrieval scenarios.

VII. Conclusion

In this work, we introduced ReT-2, a recurrent Transformer-based retrieval model that unifies multimodal queries and documents within a single framework. By combining multi-layer visual and textual representations through a gated recurrent cell, ReT-2 achieves robust retrieval performance across diverse multimodal settings, as shown on the M2KR and M-BEIR benchmarks. Furthermore, it proves to be a powerful retrieval component for retrieval-augmented generation, enabling off-the-shelf MLLMs to achieve superior accuracy on knowledge-intensive VQA tasks. Our analysis also shows that ReT-2 offers not only accuracy improvements but also significant efficiency gains, with faster inference and reduced memory usage compared to existing methods. Overall, we believe that leveraging multi-layer features with recurrent integration offers a promising direction toward more scalable, robust, and practical multimodal retrieval systems.

Acknowledgments

This work has been partially supported by the EU Horizon project ELLIOT (GA No. 101214398) and by the EuroHPC JU projects MINERVA (GA No. 101182737) and IT4LIA (GA No. 101234224). We also acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources.

References

[1] C. Wei, Y. Chen, H. Chen, H. Hu, G. Zhang, J. Fu, A. Ritter, and W. Chen, "UniIR: Training and Benchmarking Universal Multimodal Information Retrievers," in *ECCV*, 2024.

[2] D. Caffagni, S. Sarto, M. Cornia, L. Baraldi, and R. Cucchiara, "Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval," in *CVPR*, 2025.

[3] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave, "Unsupervised Dense Information Retrieval with Contrastive Learning," *arXiv preprint arXiv:2112.09118*, 2021.

[4] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. PAMI*, vol. 40, no. 5, pp. 1224–1244, 2017.

[5] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *CVPR*, 2017.

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *ECCV*, 2014.

[7] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning," in *ACL*, 2018.

[8] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "LAION-5B: An open large-scale dataset for training next generation image-text models," in *NeurIPS*, 2022.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *ICML*, 2021.

[10] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021.

[11] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image Learning," in *CVPR*, 2023.

[12] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid Loss for Language Image Pre-Training," in *ICCV*, 2023.

[13] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans *et al.*, "SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features," *arXiv preprint arXiv:2502.14786*, 2025.

[14] D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, M. Cornia, and R. Cucchiara, "The Revolution of Multimodal Large Language Models: A Survey," in *ACL Findings*, 2024.

[15] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved Baselines with Visual Instruction Tuning," in *CVPR*, 2024.

[16] F. Cocchi, N. Moratelli, D. Caffagni, S. Sarto, L. Baraldi, M. Cornia, and R. Cucchiara, "LLaVA-MORE: A Comparative Study of LLMs and Visual Backbones for Enhanced Visual Instruction Tuning," in *ICCV Workshops*, 2025.

[17] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2.5-VL Technical Report," *arXiv preprint arXiv:2502.13923*, 2025.

[18] W. Lin, J. Mei, J. Chen, and B. Byrne, "PreFLMR: Scaling Up Fine-Grained Late-Interaction Multi-modal Retrievers," in *ACL*, 2024.

[19] Y. Chen, H. Hu, Y. Luan, H. Sun, S. Changpinyo, A. Ritter, and M.-W. Chang, "Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions?" in *EMNLP*, 2023.

[20] T. Mensink, J. Uijlings, L. Castrejon, A. Goel, F. Cadar, H. Zhou, F. Sha, A. Araujo, and V. Ferrari, "Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories," in *ICCV*, 2023.

[21] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, 1997.

[22] Z. Wang, Y. Ma, Z. Liu, and J. Tang, "R-Transformer: Recurrent Neural Network Enhanced Transformer," *arXiv preprint arXiv:1907.05572*, 2019.

[23] D. Hutchins, I. Schlag, Y. Wu, E. Dyer, and B. Neyshabur, "Block-Recurrent Transformers," in *NeurIPS*, 2022.

[24] H. Hu, Y. Luan, Y. Chen, U. Khandelwal, M. Joshi, K. Lee *et al.*, "Open-domain Visual Entity Recognition: Towards Recognizing Millions of Wikipedia Entities," in *CVPR*, 2023.

[25] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge," in *CVPR*, 2019.

[26] Z. Jiang, R. Meng, X. Yang, S. Yavuz, Y. Zhou, and W. Chen, "VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks," in *ICLR*, 2025.

[27] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "EVA-CLIP: Improved Training Techniques for CLIP at Scale," *arXiv preprint arXiv:2303.15389*, 2023.

[28] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models," in *ICCV*, 2015.

[29] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould, "Image Retrieval on Real-Life Images With Pre-Trained Vision-and-Language Models," in *ICCV*, 2021.

[30] B. Zhang, P. Zhang, X. Dong, Y. Zang, and J. Wang, "Long-CLIP: Unlocking the Long-Text Capability of CLIP," in *ECCV*, 2024.

[31] A. Miech, J.-B. Alayrac, I. Laptev, J. Sivic, and A. Zisserman, "Thinking Fast and Slow: Efficient Text-to-Visual Retrieval With Transformers," in *CVPR*, 2021.

[32] A. Brown, W. Xie, V. Kalogeiton, and A. Zisserman, "Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval," in *ECCV*, 2020.

[33] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo, "Conditioned and Composed Image Retrieval Combining and Partially Fine-Tuning CLIP-Based Features," in *CVPR*, 2022.

[34] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," in *ICML*, 2022.

[35] S. Kim, X. Zhu, X. Lin, M. Bastan, D. Gray, and S. Kwak, "GENIUS: A Generative Framework for Universal Multimodal Search," in *CVPR*, 2025.

[36] W. Lin, J. Chen, J. Mei, A. Coca, and B. Byrne, "Fine-grained Late-interaction Multi-modal Retrieval for Retrieval Augmented Visual Question Answering," in *NeurIPS*, 2023.

[37] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in *ACM SIGIR*, 2020.

[38] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *NeurIPS*, 2020.

[39] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The Llama 3 Herd of Models," *arXiv preprint arXiv:2407.21783*, 2024.

[40] Y. Liu, Y. Zhang, J. Cai, X. Jiang, Y. Hu, J. Yao, Y. Wang, and W. Xie, "LamRA: Large Multimodal Model as Your Advanced Retrieval Assistant," in *CVPR*, 2025.

[41] S.-C. Lin, C. Lee, M. Shoeybi, J. Lin, B. Catanzaro, and W. Ping, "MM-Embed: Universal Multimodal Retrieval with Multimodal LLMs," in *ICLR*, 2025.

[42] Y. Lyu, R. Shao, G. Chen, Y. Zhu, W. Guan, and L. Nie, "PUMA: Layer-Pruned Language Model for Efficient Unified Multimodal Retrieval with Modality-Adaptive Learning," *ACM Multimedia*, 2025.

[43] L. Huang, Q. Wu, Z. Miao, and T. Yamasaki, "Joint Fusion and Encoding: Advancing Multimodal Retrieval from the Ground Up," *arXiv preprint arXiv:2502.20008*, 2025.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *NeurIPS*, 2017.

[45] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *EMNLP*, 2020.

[46] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971*, 2023.

[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ICLR*, 2021.

[48] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021.

[49] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling Vision Transformers," in *CVPR*, 2022.

[50] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM CSUR*, vol. 54, no. 10s, pp. 1–41, 2022.

[51] T. Lei, "When attention meets fast recurrence: Training language models with reduced compute," in *EMNLP*, 2021.

[52] T. Lei, Y. Zhang, S. I. Wang, H. Dai, and Y. Artzi, "Simple recurrent units for highly parallelizable recurrence," in *EMNLP*, 2017.

[53] A. Bapna, G. Foster, L. Jones, M. Hughes, M. Johnson, M. Chen, M. Schuster, N. J. Parmar *et al.*, "The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation," in *ACL*, 2018.

[54] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, and M. Zaharia, "ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction," in *NAACL*, 2022.

[55] F. Dalvi, H. Sajjad, N. Durrani, and Y. Belinkov, "Analyzing Redundancy in Pretrained Transformer Models," in *EMNLP*, 2020.

[56] S. He, G. Sun, Z. Shen, and A. Li, "What Matters in Transformers? Not All Attention is Needed," *arXiv preprint arXiv:2406.15786*, 2024.

[57] F. A. Joseph, J. Sieber, M. Zeilinger, and C. A. Alonso, "Lambda-Skip Connections: the Architectural Component that Prevents Rank Collapse," in *ICLR*, 2025.

[58] Y. Dong, J. B. Cordonnier, and A. Loukas, "Attention is not all you need: pure attention loses rank doubly exponentially with depth," in *ICML*, 2021.

[59] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning," in *ACM SIGIR*, 2021.

[60] E. Bugliarello, F. Liu, J. Pfeiffer, S. Reddy, D. Elliott, E. M. Ponti, and I. Vulić, "IGLUE: A Benchmark for Transfer Learning Across Modalities, Tasks, and Languages," in *ICML*, 2022.

[61] S. Shah, A. Mishra, N. Yadati, and P. P. Talukdar, "KVQA: Knowledge-Aware Visual Question Answering," in *AAAI*, 2019.

[62] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset," in *NeurIPS*, 2016.

[63] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," *NeurIPS*, 2024.

[64] S. Liu, W. Feng, T.-j. Fu, W. Chen, and W. Y. Wang, "EDIS: Entity-Driven Image Search over Multimodal Web Content," in *EMNLP*, 2023.

[65] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris, "Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback," in *CVPR*, 2021.

[66] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis, "Automatic Spatially-Aware Fashion Concept Discovery," in *ICCV*, 2017.

[67] F. Liu, Y. Wang, T. Wang, and V. Ordonez, "Visual News: Benchmark and Challenges in News Image Captioning," in *EMNLP*, 2021.

[68] S. Fu, N. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola, "DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data," in *NeurIPS*, 2023.

[69] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," *IEEE Trans. on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[70] J. Lin, H. Chen, Y. Fan, Y. Fan, X. Jin, H. Su, J. Fu, and X. Shen, "Multi-Layer Visual Feature Fusion in Multimodal LLMs: Methods, Analysis, and Best Practices," in *CVPR*, 2025.

[71] Y. Bengio, P. Simard, and P. Frasconi, "Learning Long-Term Dependencies with Gradient Descent is Difficult," *IEEE Transactions on Neural Networks*, 1994.

[72] R. Pascanu, T. Mikolov, and Y. Bengio, "On the Difficulty of Training Recurrent Neural Networks," in *ICML*, 2013.

[73] E. Voita, R. Sennrich, and I. Titov, "The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives," in *EMNLP*, 2019.

[74] A. Gromov, K. Tirumala, H. Shapourian, P. Glorioso, and D. A. Roberts, "The Unreasonable Ineffectiveness of the Deeper Layers," in *ICLR*, 2025.

[75] Q. Sun, M. Pickett, A. K. Nain, and L. Jones, "Transformer Layers as Painters," in *AAAI*, 2025.

[76] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," in *ICML*, 2023.

[77] K. Zhang, Y. Luan, H. Hu, K. Lee, S. Qiao, W. Chen, Y. Su, and M.-W. Chang, "MagicLens: Self-Supervised Image Retrieval with Open-Ended Instructions," in *ICML*, 2024.

[78] T. Jiang, M. Song, Z. Zhang, H. Huang, W. Deng, F. Sun, Q. Zhang *et al.*, "E5-V: Universal Embeddings with Multimodal Large Language Models," *arXiv preprint arXiv:2407.12580*, 2024.

[79] J. Bulian, C. Buck, W. Gajewski, B. Börschinger, and T. Schuster, "Tomayto, Tomahto. Beyond Token-level Answer Equivalence for Question Answering Evaluation," in *EMNLP*, 2022.

[80] D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, and R. Cucchiara, "Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs," in *CVPR Workshops*, 2024.

[81] F. Cocchi, N. Moratelli, M. Cornia, L. Baraldi, and R. Cucchiara, "Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering," in *CVPR*, 2025.

[82] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning," in *NeurIPS*, 2023.

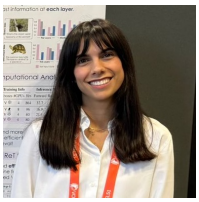
[83] Y. Yan and W. Xie, "EchoSight: Advancing Visual-Language Models with Wiki Knowledge," in *EMNLP Findings*, 2024.

[84] Y. Hong, J. Gu, Q. Yang, L. Fan, Y. Wu, Y. Wang, K. Ding, S. Xiang, and J. Ye, "Knowledge-based Visual Question Answer with Multimodal Processing, Retrieval and Filtering," in *NeurIPS*, 2025.

[85] A. Compagnoni, M. Morini, S. Sarto, F. Cocchi, D. Caffagni, M. Cornia, L. Baraldi, and R. Cucchiara, "ReAG: Reasoning-Augmented Generation for Knowledge-based Visual Question Answering," in *CVPR*, 2026.

[86] Y. Chang, M. Narang, H. Suzuki, G. Cao, J. Gao, and Y. Bisk, "WebQA: Multihop and Multimodal QA," in *CVPR*, 2022.

Davide Caffagni received the M.Sc. degree in Computer Engineering cum laude from the University of Modena and Reggio Emilia in 2023. He is currently pursuing a PhD in Information and Communication Technologies (ICT) at the University of Modena and Reggio Emilia. His research topics include Computer Vision and Natural Language Processing, with a focus on image captioning, multimodal large language models, and multimodal retrieval.



Sara Sarto received the Ph.D. degree cum laude in ICT from the University of Modena and Reggio Emilia in 2026. She is currently a Postdoctoral Researcher at the Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia. She was an Applied Scientist Intern at Amazon in 2025. Her research interests include vision-and-language models and multimodal learning, focusing on cross-modal and multimodal retrieval.

Marcella Cornia received the Ph.D. degree cum laude in ICT from the University of Modena and Reggio Emilia in 2020. She is currently an Associate Professor with the Department of Education and Humanities, University of Modena and Reggio Emilia. She has authored or coauthored more than 100 publications in scientific journals and international conference proceedings. Her research interests include vision-and-language tasks, generative AI, and multimodal learning. She is member of ELLIS and regularly serves as Area Chair/Reviewer for international conferences and journals.

Lorenzo Baraldi received the Ph.D. degree cum laude in ICT from the University of Modena and Reggio Emilia in 2018. He is currently an Associate Professor with the Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia. He was a Research Intern at Facebook AI Research (FAIR) in 2017. He has authored or coauthored more than 130 publications in scientific journals and international conference proceedings. His research interests include video understanding, deep learning, and multimedia. He is an ELLIS Scholar and regularly serves as Area Chair/Reviewer for international conferences and journals.

Rita Cucchiara received the Ph.D. degree in Computer Engineering from the University of Bologna in 1992. She is currently a Full Professor of Computer Engineering and the Rector of the University of Modena and Reggio Emilia, where she also heads the AImageLab Laboratory. She has authored or coauthored more than 600 papers in journals and international proceedings, and has been a coordinator of several projects in computer vision and pattern recognition. She also is ELLIS Fellow, director of the ELLIS Unit of Modena (Unimore), and Coordinator of the Unimore Unit of the National Ph.D. School in AI for Society.

APPENDIX A
ADDITIONAL IMPLEMENTATION DETAILS

Additional Details on M2KR. We report in Table XII the detailed composition of the M2KR benchmark. While the MSMARCO dataset (highlighted in gray in the table) is part of the original benchmark, we opted not to include it in our experiments due to its entirely textual nature (*i.e.*, $q^T \rightarrow d^T$). Other datasets, namely WIT, KVQA, IGLUE, and CC3M, are designed to evaluate the ability of the models to identify relevant documents given an input image (*i.e.*, $(q^V, q^T) \rightarrow d^T$). While CC3M is incorporated into the M2KR training set to enhance scene understanding, it is excluded from validation and test splits since its original focus is caption generation rather than retrieval. The IGLUE benchmark, a subset of WIT, is retained to ensure comparability with prior work. Finally, KVQA, originally introduced as a knowledge-based VQA task, is adapted to fit our setting.

In this setting, each item is paired with an instruction sampled from predefined templates. For instance, WIT and IGLUE use instructions like {`Image`} `Please describe the document that corresponds to this image`; KVQA uses {`Image`} `Provide a brief description of the image and the relevant details of the person in the image`; and CC3M uses templates similar to {`Image`} `Describe the image concisely`. For a detailed enumeration of all possible instruction templates, we refer the reader to the original M2KR paper [18].

The task $(q^V, q^T) \rightarrow (d^V, d^T)$ requires joint understanding of both images and text for accurate retrieval. In this setting, we use datasets such as OVEN, LLaVA, OKVQA, InfoSeek, and Encyclopedic-VQA. Training, validation, and test samples are downsampled from the original datasets. In the original M2KR benchmark, these datasets were not fully multimodal – *i.e.*, the document side did not include visual input. To better align with our setting, we augment these splits by enriching the reference documents with images, as described in Sec. V-A. This dataset version is available in our repository.

Additional Details on M-BEIR. The M-BEIR benchmark consists of eight multimodal retrieval tasks spanning ten datasets from diverse domains and image sources. In particular, it standardizes training and evaluation by repurposing diverse datasets. Image-caption datasets (COCO [6], Fashion200k [66], and Visual News [67]) are adapted by treating captions as queries, while NIGHTS [68] addresses visual similarity in nighttime scenes. It also includes retrieval-based VQA datasets (InfoSeek [19], WebQA [86]), where documents are relevant if they contain the answer. These tasks are designed to evaluate performance under both missing-modality and fully multimodal scenarios. For example, Task #2 involves purely textual inputs on both the query and document sides, whereas Task #8 is fully multimodal, with both queries and documents containing visual and textual information. Other tasks, such as #3 and #7, introduce asymmetry by providing multimodal inputs on only one side – either the query or the document – testing the ability of the models to handle missing modalities. Detailed dataset splits are reported in Table XIII.

TABLE XII
SUMMARY OF THE M2KR BENCHMARK [18]. FOR EACH DATASET, WE REPORT THE NUMBER OF TRAINING, VALIDATION, AND TEST SAMPLES, ALONG WITH THE SIZE OF THE DOCUMENT POOL (SPLIT INTO TRAINING AND VALIDATION/TEST). PURPLE COLOR HIGHLIGHTS DATASETS AUGMENTED WITH DOCUMENT-SIDE IMAGES, WHILE GRAY DENOTES DATASETS EXCLUDED FROM BOTH TRAINING AND TEST. THE LAST ROW REFLECTS ONLY THE DATASETS USED IN OUR EXPERIMENTS.

		Query			Document
Task	**Dataset**	**# Train**	**# Val**	**# Test**	**# Pool**
$q^T \rightarrow d^T$	MSMARCO [62]	400k	6.9k	5.1k	8.8M/200k
$(q^V, q^T) \rightarrow d^T$	WIT [59]	2.8M	20.1k	5.1k	4.1M/40k
	IGLUE [60]	-	-	685	-/1k
	KVQA [61]	16k	13.4k	5.1k	16.3k/4.6k
	CC3M [7]	595k	-	-	595k/-
$(q^V, q^T) \rightarrow (d^V, d^T)$	OVEN [24]	339k	20k	5.1k	10k/3.1k
	OKVQA [25]	9k	5k	5k	110k/110k
	InfoSeek [19]	76k	-	4.7k	100k/100k
	E-VQA [20]	167k	9.8k	3.7k	50k/50k
2 tasks	8 datasets	4.8M	68.3k	29.4k	4.98M/308k

TABLE XIII
SUMMARY OF THE M-BEIR BENCHMARK [1]. FOR EACH DATASET, WE REPORT THE NUMBER OF TRAINING, VALIDATION, AND TEST SAMPLES, ALONG WITH THE SIZE OF THE DOCUMENT POOL.

		Query			Document
Task	**Dataset**	**# Train**	**# Val**	**# Test**	**# Pool**
#1: $q^T \rightarrow d^V$	VN [67]	99k	20k	20k	542k
	COCO [6]	100k	24.8k	24.8k	5k
	F200k [66]	15k	1.7k	1.7k	201k
#2: $q^T \rightarrow d^T$	WQA [86]	16k	1.7k	2.4k	544k
#3: $q^T \rightarrow (d^V, d^T)$	EDIS [64]	26k	3.2k	3.2k	1M
	WQA [86]	17k	1.7k	2.5k	403k
#4: $q^V \rightarrow d^T$	VN [67]	100k	20k	20k	537k
	COCO [6]	113k	5k	5k	25k
	F200k [66]	15k	4.8k	4.8k	61k
#5: $q^V \rightarrow d^V$	NIGHTS [68]	16k	2k	2k	40k
#6: $(q^V, q^T) \rightarrow d^T$	OVEN [24]	150k	50k	50k	676k
	InfoSeek [19]	141k	11k	11k	611k
#7: $(q^V, q^T) \rightarrow d^V$	FIQ [65]	16k	2k	6k	74k
	CIRR [29]	26k	2k	4k	21k
#8: $(q^V, q^T) \rightarrow (d^V, d^T)$	OVEN [24]	157k	14.7k	14.7k	335k
	InfoSeek [19]	143k	17.6k	17.6k	481k
8 tasks	10 datasets	1.1M	182k	190k	5.6M

Retrieval-Augmented VQA. In our experiments, we prompt the MLLM with the top-K documents retrieved using ReT-2. The prompt used to generate answers is defined as follows:

```
{Image} Given the context, answer the
question based on the image.
Question: {Question}
Context:
## {C1}
## {...}
## {CK}
If the context does not help with the
question, try to answer it anyway. Do not
generate anything but the short answer.
Short answer:
```

where $\texttt{C}_K$ is replaced with the text content of the top-K retrieved documents.

APPENDIX B
THEORETICAL MOTIVATION OF TOKEN REDUCTION AND LAYER PRUNING

This appendix provides the full derivations and literature grounding for the token reduction and layer pruning design choices introduced in Sec. IV and empirically validated in Sec. V (cf. Table II and Table III).

A. Token Reduction and Rank Collapse

To understand why reducing the number of learnable tokens from $k = 32$ to $k = 1$ does not sacrifice representational capacity, we relate the internal state of the recurrent cell to the rank-collapse phenomenon described by Dong *et al.* [58] for pure self-attention networks (SANs). This analysis further accompanies the empirical rank-collapse measurements reported in Sec. V.

Reinstating, for this analysis, the general k-token notation used for ReT in Sec. III, let $\mathbf{H}_l \in \mathbb{R}^{k\times d}$ denote the stacked hidden state of the recurrent cell at layer l (which reduces to the single row vector $\boldsymbol{h}_l$ used elsewhere in the paper once $k = 1$). Following [58], $\mathbf{H}_l$ can be decomposed into a component shared across all k rows and a residual capturing their differences:

$$\mathbf{H}_l = \mathbf{1}\bar{\mathbf{h}}_l^\top + \mathbf{R}_l, \qquad \bar{\mathbf{h}}_l = \frac{1}{k}\mathbf{1}^\top\mathbf{H}_l, \qquad \mathbf{R}_l = \Pi\mathbf{H}_l, \quad (12)$$

where $\Pi = \mathbf{I}_k - \frac{1}{k}\mathbf{1}\mathbf{1}^\top$ is the centering matrix. The term $\mathbf{1}\bar{\mathbf{h}}_l^\top$ has rank at most one and identical rows, whereas $\mathbf{R}_l$ contains all the information that distinguishes one token from another; $\mathbf{R}_l = \mathbf{0}$ exactly when all k tokens have collapsed onto the same representation.

Dong *et al.* show that, in a pure SAN, repeated self-attention contracts $\mathbf{R}_l$ at a doubly-exponential rate with depth, and that this contraction is only counteracted by skip connections and MLPs, which slow but do not reverse it. The recurrent cell of ReT repeatedly mixes the same $k = 32$ latent rows of $\mathbf{H}_l$ across every processed layer of the backbone (*i.e.*, self-attention among the k tokens, followed by two cross-attention operations against $\mathbf{E}_l^T$ and $\mathbf{E}_l^V$, cf. Eq. (4)), so the number of recurrent steps plays a role analogous to the depth of a SAN with respect to the contraction of $\mathbf{R}_l$. Three structural properties of the cell reinforce this contraction:

- The cross-attention update in Eq. (4) uses $\mathbf{H}_l$ (or its normalized form $\hat{\boldsymbol{h}}_l$) to produce the queries. If two rows $\mathbf{h}_{l,i} \approx \mathbf{h}_{l,j}$, they yield near-identical query vectors, near-identical attention distributions over the shared context $\mathbf{E}_l^m$, and therefore near-identical outputs.
- The gates and the feed-forward network in Eq. (6)–(7) are shared across all k token positions. If two rows are identical, these shared, position-independent transformations map them to identical outputs.
- The residual path connecting $\mathbf{c}_l$ to $\boldsymbol{c}_{l+1}$ in Eq. (5) is modulated by the forget gate $\boldsymbol{f}_l$, which is confined to $[0, 1]$ (Eq. (6)). Because $\boldsymbol{f}_l$ is trained to actively suppress stale or irrelevant information, it weakens a residual connection that would otherwise counteract collapse. We verified this by removing $\boldsymbol{f}_l$ from Eq. (5), which restores an unmodulated residual path. This measurably increases token diversity, but it also removes the cell's ability to forget outdated content, which we found to hurt retrieval performance.

B. Layer Pruning and Recurrence Length

Our layer pruning strategy restricts the recurrent cell to attend one early, one middle, and one late layer of each backbone, rather than all L layers. This is motivated by two complementary mechanisms, one concerning the length of the recurrence, the other the statistical redundancy of its inputs.

(i) Recurrence Length and Memory Retention. The recurrent cell of ReT-2 follows an LSTM-style additive-forget update (Eq. (5)): at each of the processed layers, the previous candidate state is attenuated by the sigmoidal forget gate $\boldsymbol{f}_l \in [0, 1]$ and combined with fresh, layer-specific evidence. Unrolling Eq. (5) shows that the contribution of an early layer l_0 to the final state $\boldsymbol{h}_L$ is multiplicatively discounted by the product $\prod_{l=l_0}^{L-1} \boldsymbol{f}_l$ of every forget gate encountered afterwards. Because each factor lies in $(0, 1)$, this product can only remain non-negligible if the gates learn to stay close to 1 consistently across the entire remaining chain, which becomes harder as L grows, mirroring the classical difficulty of learning long-range dependencies with recurrent models [71], [72]. Gating, as introduced by LSTMs [21], still requires the gates to be correctly trained over the whole unrolled chain, which is easier over the shorter sequences used after layer pruning than over longer sequences.

(ii) Residual-stream Redundancy Across Depth. Both the visual and textual backbones used in ReT-2 are pre-trained Transformers, in which every layer updates the representation through an additive residual connection, $\mathbf{E}_{l+1}^m = \mathbf{E}_l^m + \Delta_l^m$. As the update Δ_l^m is small relative to $\mathbf{E}_l^m$ [73]–[75], consecutive layers end up representing highly similar information. Gromov *et al.* [74] show directly, via layer-wise cosine similarity, that deep and middle layers are often near-redundant with their immediate neighbors.

This redundancy has a direct consequence for the recurrent cell: the gates $\boldsymbol{f}_l$ and $\boldsymbol{i}_l^m$ at layer l are themselves computed from $\boldsymbol{z}_l^m$, which depends on $\mathbf{E}_l^m$ (Eq. (6)). If $\mathbf{E}_l^m \approx \mathbf{E}_{l+1}^m$, the gates computed at these two consecutive steps will also be near-identical – including any miscalibration. Concretely, if the cell under- or over-estimates the relevance of some cue at layer l it is likely to reproduce the same misjudgment at layer $l + 1$, since it is presented with virtually the same evidence. Two such consecutive steps do not behave as two independent opportunities to refine the internal state; they behave as one decision applied twice, consuming recurrence capacity without adding genuinely new information and, in the worst case, reinforcing a gating error.

Running the recurrent cell over all L layers therefore generates an error accumulation risk, which is lowered by selecting a subset of layers via layer pruning. Under this reading, layer pruning can be understood as a way of matching the recurrent cell's limited processing budget to the non-uniform rate at which the backbone actually accumulates new information across depth.

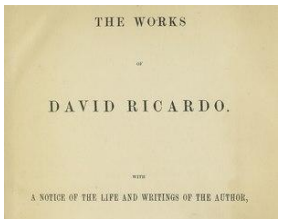
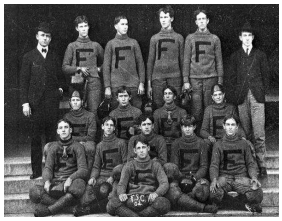
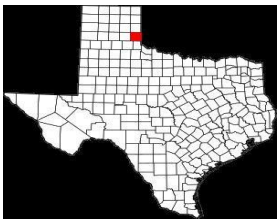
APPENDIX C
QUALITATIVE RESULTS

Qualitative Results on M2KR. In Fig. 6 and Fig. 7, we provide a qualitative comparison between PreFLMR, the original ReT, and our proposed ReT-2. In particular, to enable a direct comparison with PreFLMR, Fig. 6 focuses solely on M2KR datasets that do not include document images. As shown, ReT-2 consistently retrieves information that is more contextually relevant and detailed for the given queries. In contrast, Fig. 7 includes document images whenever available in the retrieved content[4]. These examples highlight that incorporating visual information from document images substantially enhances the ability of the model to answer queries accurately, effectively complementing textual content with visual context.

Qualitative Results on M-BEIR. Fig. 8 and Fig. 9 present qualitative results comparing UniIR and ReT-2 on two different tasks. In Task #3, we employ an example from the EDIS dataset where only the document side is multimodal, whereas in Task #8, where we use an example from the InfoSeek dataset, both the query and document sides are fully multimodal. The figures show the top-3 retrieved documents for each task, highlighting ReT-2 's ability to consistently retrieve the correct documents, while UniIR struggles to locate the relevant ones.

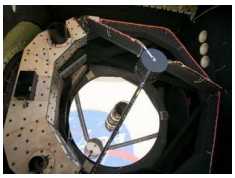
Qualitative Results on Retrieval-Augmented VQA. Qualitative results on sample image-question pairs from InfoSeek and Encyclopedic-VQA are shown in Fig. 10 and Fig. 11, comparing answers generated by augmenting Qwen2.5-VL with context retrieved by different multimodal retrieval models, including PreFLMR, ReT, and ReT-2. The results demonstrate that ReT-2 consistently retrieves more accurate and relevant documents, enabling better responses to specific multimodal questions and outperforming the other approaches.

[4]Note that both ReT and ReT-2 employ datasets augmented with document images when available. For space constraints, in the reported qualitative results, we only include document images retrieved by ReT-2.

WIT		IGLUE	
			
Could you elucidate the document associated with this image?	*Identify the document that this image pertains to.*	*Please give information on the document that goes with this image.*	*Could you elucidate the document associated with this image?*
PreFLMR [18]: title: Lilith, The Legend of the First Woman hierarchical section title: Lilith, The Legend of the First Woman caption [...]	**PreFLMR [18]**: title: Pomona-Pitzer Sagehens / History caption reference description: Members of the Pomona football team in 1907 caption [...]	**PreFLMR [18]**: title: Yoakum County, Texas hierarchical section title: Yoakum County, Texas caption reference description: Location [...]	**PreFLMR [18]**: title: Tomaszów Mazowiecki section title: Sulejowski Reservoir hierarchical section title: Tomaszów Mazowiecki [...]
ReT [2]: title: The Marble Faun hierarchical section title: The Marble Faun caption reference description: First edition title page caption [...]	**ReT [2]**: title: 1900 Western University of Pennsylvania football team caption attribution description: English: The 1900 Pittsburgh [...]	**ReT [2]**: title: Camp County, Texas hierarchical section title: Camp County, Texas caption reference description: Location within the [...]	**ReT [2]**: title: Tomaszów Mazowiecki section title: Sulejowski Reservoir hierarchical section title: Tomaszów Mazowiecki [...]
ReT-2 (Ours): title: David Ricardo section title: Publications hierarchical section title: David Ricardo / Publications caption [...]	**ReT-2 (Ours):** title: List of Florida State University athletes caption reference description: Florida State's first football team, "The Eleven" [...]	**ReT-2 (Ours):** title: Collingsworth County, Texas hierarchical section title: Collingsworth County, Texas caption reference description: [...]	**ReT-2 (Ours):** title: Yelagin Island section title: Current use hierarchical section title: Yelagin Island / Current use caption reference description [...]

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 color is the dog in the image?*	*What type of train is seen in the image?*
PreFLMR [18]: This is an image of Byrne at the Sydney film premiere of I Give It a Year in 2013. Rose Byrne went to Australian Theatre for [...]	**PreFLMR [18]**: This is an image of Tõnis Lukas on Estonian Science Communication Conference 2016. Tõnis Lukas went to [...]	**PreFLMR [18]**: The dog in the image is white.	**PreFLMR [18]**: The train shown in the image is a passenger train.
ReT [2]: This is an image of Brie at the 2009 Los Angeles Film Festival. Alison Brie went to California Institute of the Arts, Royal [...]	**ReT [2]**: This is an image of Jean-Luc Warsmann (2016). Jean-Luc Warsmann went to Sciences Po, date of birth is 1965-10-22 [...]	**ReT [2]**: The dog in the image is brown, with some black markings as well.	**ReT [2]**: The train shown in the image is a passenger train.
ReT-2 (Ours): This is an image of Lucas at the 2011 WonderCon. Isabel Lucas date of birth is 1985-01-29, knows English, is a actor, film [...]	**ReT-2 (Ours):** This is an image of Rui Tavares (2013). Rui Tavares date of birth is 1972-07-29, is a member of LIVRE (political party), [...]	**ReT-2 (Ours):** The dog in the image is black.	**ReT-2 (Ours):** A Plaza Santa Fe passenger train is seen in the image.

Fig. 6. Qualitative results on the M2KR benchmark [18], for datasets that do not include document images.

OVEN		InfoSeek	
what kind of plant is this?	*which type of item is depicted in the image?*	*Who is the creator of this object?*	*What is this building dedicated to?*
PreFLMR [18]: Hippeastrum is a genus of about 90 species and over 600 hybrids and cultivars of perennial herbaceous bulbous plants.	**PreFLMR [18]**: Handball (also known as team handball, European handball or Olympic handball) is a team sport in which two teams [...]	**PreFLMR [18]**: Optical axis of the Schmidt design creates a Schmidt-Newtonian telescope. The addition of a convex secondary mirror to [...]	**PreFLMR [18]**: [...] the reign of the Grand Prince Vsevolod the Big Nest of Vladimir-Suzdal to the honour of Saint Demetrius of Thessaloniki.
ReT [2]: The **poinsettia** (or "Euphorbia pulcherrima") is a commercially important plant species of the diverse spurge family [...]	**ReT [2]**: Netball is a ball sport played by two teams of seven players, usually on an indoor court, and is predominantly played by women.	**ReT [2]**: Manageable at large focal ratios – most Schiefspieglers use f/15 or longer, which tends to restrict useful observation to the moon [...]	**ReT [2]**: [...] dedicated to **St Peter of Moscow**, was long regarded as a typical monument of the Naryshkin style and dated to 1692.
ReT-2 (Ours): The **poinsettia** (or "Euphorbia pulcherrima") is a commercially important plant species of the diverse spurge family [...]	**ReT-2 (Ours):** A **basket** is a container that is traditionally constructed from stiff fibers and can be made from a range of materials [...]	**ReT-2 (Ours):** [...] as the Gregorian telescope. **Isaac Newton** has been generally credited with building the first reflecting telescope in 1668.	**ReT-2 (Ours):** [...] dedicated to **St Peter of Moscow**, was long regarded as a typical monument of the Naryshkin style and dated to 1692.
			

E-VQA		OKVQA	
In which part of the world does this animal live?	*When was this bridge built?*	*What model of computer is this?*	*Where would one find these animals?*
PreFLMR [18]: Henricia leviuscula, the "Pacific blood star", it is a species of sea star found along the Pacific coast of North America.	**PreFLMR [18]**: The bridge was built about **1160 AD** and a bridge chapel was built dedicated to Thomas Becket in 1235 on [...]	**PreFLMR [18]**: A Compact **Macintosh** (or Compact **Mac**) is an all-in-one Apple **Mac** computer with a display integrated in the [...]	**PreFLMR [18]**: The Sudanian Savanna is a broad belt of tropical savanna that runs east and west across the **African continent**, from the [...]
ReT [2]: Evasterias troschelii is a species of starfish in the family Asteriidae. Its common names include the mottled star [...]	**ReT [2]**: The Ripon Canal is located in North Yorkshire, England. It was built by the canal engineer William Jessop to link the city of Ripon [...]	**ReT [2]**: The PC-D and PC-X were personal computers sold by Siemens between 1982 (PC-X)/1984 (PC-D) and 1986. The PC-D was [...]	**ReT [2]**: Assassination attempt by one of Loveless' henchwomen, West mistakes a female guest for a disguised Gordon resulting in [...]
ReT-2 (Ours): Patiriella regularis, or **New Zealand** common cushion star, is a sea star of the family Asterinidae, native to **New Zealand**.	**ReT-2 (Ours):** The bridge was built about **1160 AD** and a bridge chapel was built dedicated to Thomas Becket in 1235 on [...]	**ReT-2 (Ours):** Prior versions of the **Mac** mini were much more difficult to open. Some **Mac** mini owners used a putty knife or a pizza [...]	**ReT-2 (Ours):** Shaba National Reserve was the setting for the book and film "Born Free", for the film "Out of **Africa**" and for [...]

Fig. 7. Qualitative results on the M2KR benchmark [18], for datasets that include document images. We highlight the reference answer in bold font whenever it is found in the retrieved text. For ReT-2, we also add the document image attached to the text.

Task #3: $q^T \rightarrow (d^T, d^V)$

Find a news image that matches the provided caption.

Having finished with the likes of Mercedes driver Lewis Hamilton, the world's press were treated to a rare display of candor from F1 team principals.

UniIR [1]:	The Mercedes mechanics, as if impervious to the heat of the Lombardy sunshine, moved quickly on Friday, their urgent actions resembling those of.	Mercedes are to investigate why Lewis Hamilton, the fastest Formula One driver of the modern era, is now not even the quickest man in their team.	Lewis Hamilton has not only brought the best out of Mercedes this year but also given fresh impetus to his team mate, Nico Rosberg.
ReT-2 (Ours):	The second of the two press conferences held at a Formula One race weekend tends to be the drier.	Lewis Hamilton is banking on a return to one of his favorite tracks this weekend to kickstart his faltering season.	Lewis Hamilton says Mercedes can manage without Ross Brawn, the most successful team principal of his generation, because he believes no team depends.

Fig. 8. Qualitative results on the Task #3 of the M-BEIR benchmark [1], using an example from the EDIS subset [64]. Ground-truth image-text documents have a red frame around the image.

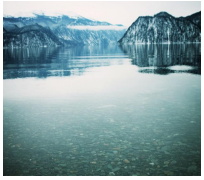
Task #8: $(q^T, q^V) \rightarrow (d^T, d^V)$

I want to address the query about this picture. Please pull up a relevant Wikipedia section and image. What is this building dedicated to?

UniIR [1]:	Church of Annunciation of Virgin Mary (Hrodna). but allowed nuns to stay and live there. Also, Benedictine nuns from Nyasvizh, Dominicans from [...]	Cathedral of St. Nicholas (Mozhaysk). The construction started in 1802 and finished only in 1814.The almost finished cathedral was badly damaged in 1812.	Church of Our Lady of the Assumption, Irkutsk. The Church of Our Lady of the Assumption also called the "Polish Church", it is a Catholic church in [...]
ReT-2 (Ours):	Transfiguration Church in Kovalyovo. Krasnorechyev. ## Architecture. The church is constructed in brick, and has one dome. It has a single apse and four square columns. This design is [...]	Vysokopetrovsky Monastery. in the Naryshkin Baroque style of architecture associated with their name. In the mid-18th century, several subsidiary structures were added, possibly based [...]	Church of Our Lady of the Sign, Verkhoturye. Church of Our Lady of the Sign - is an Orthodox church in Verkhoturye, Sverdlovsk oblast.The building was granted the status of regional [...]

Fig. 9. Qualitative results on the Task #8 of the M-BEIR benchmark [1], using an example from the InfoSeek subset [19]. Ground-truth image-text documents have a red frame around the image.

Q: What is the area in square kilometer of this lake?



PreFLMR [18]: 25.2 km² ✗
ReT [2]: Not enough information ✗
ReT-2 (Ours): 233 ✓

Q: In which year did this building come into service?

PreFLMR [18]: 1917 ✗
ReT [2]: 1914 ✗
ReT-2 (Ours): 1772 ✓

Q: What is the location of this garden?

PreFLMR [18]: Washington Park, Denver ✗
ReT [2]: United States Botanic Garden ✗
ReT-2 (Ours): National Mall ✓

Q: What is the height of this bridge in meter?

PreFLMR [18]: 70 ✗
ReT [2]: 95 ✗
ReT-2 (Ours): 150 ✓

Q: In which year did this building officially open?

PreFLMR [18]: 1934 ✗
ReT [2]: 1841 ✗
ReT-2 (Ours): 1898 ✓

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

PreFLMR [18]: Kamianets-Podilsky ✗
ReT [2]: Kazan Kremlin ✗
ReT-2 (Ours): Minsk ✓

Fig. 10. Sample results for the knowledge-intensive VQA task on the validation split of InfoSeek, augmenting Qwen2.5-VL with context retrieved by different multimodal retrieval models.

Q: What gender was the deity worshiped at this temple in its earliest phase?

PreFLMR [18]: Male ✗
ReT [2]: Male ✗
ReT-2 (Ours): Female ✓

Q: When did the San Diego savings bank leave this building?

PreFLMR [18]: March 18, 1994 ✗
ReT [2]: The San Diego savings bank left the building in 1930 ✗
ReT-2 (Ours): 1912 ✓

Q: What was the building at this canal converted into in 2012?

PreFLMR [18]: A museum ✗
ReT [2]: A museum and an heritage center ✗
ReT-2 (Ours): Four residential apartments ✓

Q: When was this church established?

PreFLMR [18]: 1986 ✗
ReT [2]: Not enough information ✗
ReT-2 (Ours): 1606 ✓

Q: Is the pale-billed woodpecker a large or small bird?

PreFLMR [18]: The pale-billed woodpecker is a large bird ✓
ReT [2]: The pale-billed woodpecker is a small bird ✗
ReT-2 (Ours): Large ✓

Q: What color is this plant flowers range from yellow to?

PreFLMR [18]: Pink or lilac-red ✗
ReT [2]: Lilac ✗
ReT-2 (Ours): Green ✓

Fig. 11. Sample results for the knowledge-intensive VQA task on the test split of Encyclopedic-VQA, augmenting Qwen2.5-VL with context retrieved by different multimodal retrieval models.