Non-Semantic Retrieval in Multimodal Long Contexts

Anonymous ACL submission

Abstract

Using a memory buffer to cache long contexts and retrieving relevant information based on current input to enhance the long-context capabilities of large language models has proven effective. However, the retrieval mechanisms used in existing approaches focus solely on the semantic alignment between the query and the memory or knowledge base, which is insufficient in multimodal contexts. To address this limitation, we introduce the problem of non-semantic retrieval and aim to improve the memory augmentation of multimodal large language models by tackling this issue. The introduction of visual modalities increases the complexity of the context structure, making it more common for queries in multimodal scenarios to require retrieval beyond simple semantic alignment. We refer to these as non-semantic queries. To solve this, we propose the Hybrid Semantic and Non-Semantic Retriever (HSNSR), capable of performing both semantic and non-semantic retrieval. We pre-train this retriever using a custom-built Semantic and Non-Semantic Retrieval Dataset (SNSR). For the pre-trained retriever, we designed a memory model fully decoupled from the multimodal large language model, called Universal Multi-modal Memory (UM-Mem). UM-Mem caches lengthy multimodal contexts into a memory buffer, and based on the current input, HSNSR retrieves relevant information from the memory. We validated the necessity of non-semantic retrieval in a zero-shot setting on the MM-NIAH benchmark, which focuses on long-context tasks. In two selected cross-modal long-context tasks involving both text and image data, we achieved approximately a 3% improvement in accuracy. This demonstrates the significance of the non-semantic retrieval problem and the potential of memory-augmented MLLM. We hope that this work inspires further research into MLLM and multimodal retrieval techniques.

1 Introduction

Multimodal large language models (MLLMs) have made significant strides recently. Following the promising visual understanding capabilities demonstrated by models like BLIP2, LLaVA, and Flamingo, newer models such as LLaVA-Next-Interleave (Li et al., 2024), Mantis (Jiang et al., 2024), and mPLUG-Owl3 (Ye et al., 2024) have successfully extended these capabilities to multi-image understanding by training with multi-image datasets. These MLLMs have achieved impressive results across a variety of tasks, including multi-image comprehension and video understanding. However, due to the inherent challenges of multimodal large language models, such as the excessively long token sequences resulting from image projection and the information loss caused by compressing visual features, these models tend to perform poorly when confronted with multimodal long-context tasks.

In this context, utilizing memory to cache long multimodal contexts presents a compelling approach to enhancing multimodal large language models. Caching lengthy contexts or knowledge bases and retrieving them on demand to improve outputs has proven effective in the realm of pure text through techniques like RAG (Lewis et al., 2020) and MemoryBank (Zhong et al., 2024). These methods not only enable LLMs to comprehend longer texts but also offer unique advantages in knowledge editing and operational efficiency.

When dealing with pure text modalities, the aforementioned techniques typically employ semantic retrieval to obtain the required context. However, this approach is insufficient in multimodal environments. We have observed that, in such contexts, semantic retrieval fails to meet the demands due to the inherently complex structure of multimodal contexts. For example, consider the current input: "What is the content of the image

before image A?<image A token>" Clearly, it is necessary to retrieve the memory of the image preceding image A to enhance the generation. However, a purely semantic retrieval approach would only be able to access the information related to image A itself, missing the crucial context needed for a complete understanding.

To address this issue, we propose the innovative concept of non-semantic retrieval, designed to handle cases where effective retrieval cannot rely on semantic features. We define non-semantic retrieval as the process in which a retriever, when enhancing the capabilities of a multimodal large language model through RAG or memory, retrieves target data from cached multimodal contexts by leveraging non-semantic features inherent to the multimodal context structure itself.

To tackle the challenge of non-semantic retrieval, we propose the Hybrid Semantic and Non-Semantic Retriever (HSNSR). HSNSR is designed to handle both semantic and non-semantic retrieval tasks. In addition to conventional semantic retrieval, it can also determine whether a query contains elements requiring non-semantic retrieval, capturing non-semantic information in cross-modal contexts. Specifically, HSNSR is built upon the CLIP (Radford et al., 2021) architecture and follows a dual-tower model structure, consisting of a memory encoder and a query encoder. Both encoders utilize transformers (Vaswani, 2017) to extract structural features from the memory and non-semantic retrieval features from the query. This allows the context's structural characteristics to be incorporated into the retrieval process. The query encoder evaluates the query text to determine the proportions of non-semantic and semantic retrieval needs. Based on this assessment, it performs a weighted fusion of semantic and non-semantic features, ultimately yielding retrieval results through a dot product, similar to the approach used in CLIP.

To train HSNSR, we constructed the Non-Semantic Retrieval Dataset (SNSR), which encompasses both semantic and non-semantic retrieval scenarios. We collected various common non-semantic retrieval situations and employed a hybrid approach of manual curation and LLM generation to create the SNSR across multiple text, image, and video datasets.

To validate the effectiveness of non-semantic retrieval, we propose a lightweight, MLLM-agnostic memory model called Universal Multi-modal Memory (UM-Mem). This model utilizes memory to store lengthy multimodal contexts and, based on current input, employs HSNSR to retrieve the necessary memories from the memory bank. We evaluate our memory model and retriever using the MM-NIAH (Wang et al., 2024a) benchmark, which focuses on long-context tasks. We conduct experiments on two challenging multimodal tasks within MM-NIAH and perform ablation studies. The experimental results demonstrate that non-semantic retrieval significantly enhances the ability of multimodal large language models to handle multimodal long contexts effectively.

In summary, our main contributions are as follows:

1) We identified and defined the concept of non-semantic retrieval, constructing a dataset that offers a novel approach to memory augmentation for multimodal large language models.

2) We designed the Hybrid Semantic and Non-Semantic Retriever (HSNSR), which can distinguish the retrieval type within input queries, extracting non-semantic retrieval features from both memory and queries to facilitate both non-semantic and semantic retrieval.

3) We implemented a lightweight, model-agnostic multimodal memory module that enhances the ability of MLLMs to process multimodal long contexts, demonstrating the significance of non-semantic retrieval.

We plan to make our code publicly available upon acceptance of this paper, ensuring transparency and reproducibility of our results.

2 Related Work

Many excellent works have enhanced language models through the use of memory. MemoryBank (Zhong et al., 2024) stores conversations and employs a dual-tower dense retrieval model, similar to knowledge retrieval, for memory retrieval. LongMem (Wang et al., 2024b) retains the k and v values from a specific layer of the LLM as memory, storing them in chunks, and retrieves them by calculating the dot product between the attention queries of each token in the current input and the average pooled attention keys of the candidate chunks. In the multimodal domain, MA-LMM (He et al., 2024) has designed two long-term memory banks to store the original visual features and learned queries at each time step. At every time step, the input query interacts with the memory bank through cascaded Q-Former blocks.

Models such as eKNN-LM (Khandelwal et al., 2019), RAG (Guu et al., 2020) and RETRO (Borgeaud et al., 2022) demonstrate that retrieval-augmented frameworks are effective enhancement techniques for text-based language models. In the multimodal domain, visual-language models like CLIP (Radford et al., 2021) align visual data with text data. Models such as MuRAG (Chen et al., 2022), MARVEL (Lin et al., 2023), UVLR (Zhou et al., 2023), and RagLLaVA (Chen et al., 2024a) have achieved multimodal retrieval between visual and textual modalities. However, these models often extend the principles of traditional text retrieval, primarily focusing on the alignment of semantics across different modalities. This approach is insufficient when it comes to memory banks, as it does not adequately address the complexities inherent in multimodal contexts.

3 Non-Semantic Retrieval

3.1 Hybrid Semantic and Non-Semantic Retriever

HSNSR is a dual-tower model consisting of a memory encoder and a query encoder, as illustrated in Figure 1. In the chunked memory bank, the memory encoder uses CLIP to extract the semantic features of each memory chunk, where each chunk contains only one modality—either an image or text. We sequentially arrange the outputs from CLIP, referring to this as the memory semantic sequence M_s. Here, we utilize the unnormalized outputs from the CLIP encoder; unless otherwise specified, all CLIP outputs mentioned in Figure 1 and the following text refer to these unnormalized outputs. Subsequently, the memory encoder employs a transformer structure to extract the structural features of the memories from the memory semantic sequence. The output embedding sequence from the transformer is referred to as the memory non-semantic sequence M_n.

Similarly to the memory, we apply the same chunking strategy to each query and obtain an embedding sequence through CLIP, referred to as the query semantic sequence Q_s. For non-semantic retrieval, the query encoder uses BERT (Devlin et al., 2018) to encode the text modality of the query (as non-semantic retrieval can only occur in the text). The text embeddings are projected into the dimensions of the query semantic sequence, and then we use a transformer that receives both the query semantic sequence and the projected BERT output to extract non-semantic retrieval features. In the first layer of the transformer, both the query semantic sequence and the projected BERT output participate in the multi-head attention computation. The former provides the Q (query), while the latter supplies the K (key) and V (value). The output from the transformer in the query encoder is referred to as the query non-semantic sequence Q_n.

To determine whether the query contains non-semantic elements, we utilize the CLS vector e_{cls} output from BERT and apply a feed-forward network (FFN) to obtain the value g, which quantifies the proportion of semantic and non-semantic queries in the input. g value, we perform a weighted summation of two pairs of vector sequences with the same shape: the memory semantic sequence and memory non-semantic sequence, as well as the query semantic sequence and query non-semantic sequence. Each resulting vector is then normalized to produce the final memory embedding M_e and query embedding Q_e. The weighted summation can be represented by the following formula:

$$M_{un} = gM_s + (1-g)M_n \tag{1}$$

$$Q_{un} = gQ_s + (1-g)Q_n \tag{2}$$

M_{un} and Q_{un} are the unnormalized outputs of M_e and Q_e, respectively. Clearly, when g is equal to 1, HSNSR effectively functions as a pure semantic retriever.

The training of HSNSR differs from that of most dual-tower models. Since the retrieval targets are within the sequence, this task is treated as a special multi-label classification problem. Similar to CLIP, we compute the dot product between the vectors of the memory embeddings and query embeddings. We then extract the memory chunks corresponding to the Top-N values in the dot product matrix as the query results. However, since non-semantic queries depend on structural information, the query results cannot be treated independently of the sequence. Therefore, during training, we compute the binary cross-entropy (BCE) loss by comparing the dot product matrix output from HSNSR with the ground truth, which is represented as a 0-1 matrix. Considering the imbalance between positive and negative samples in long-context scenarios, where negative samples greatly outnumber positive samples, we enhance the positive sample loss with a weighting mechanism to balance the influence of

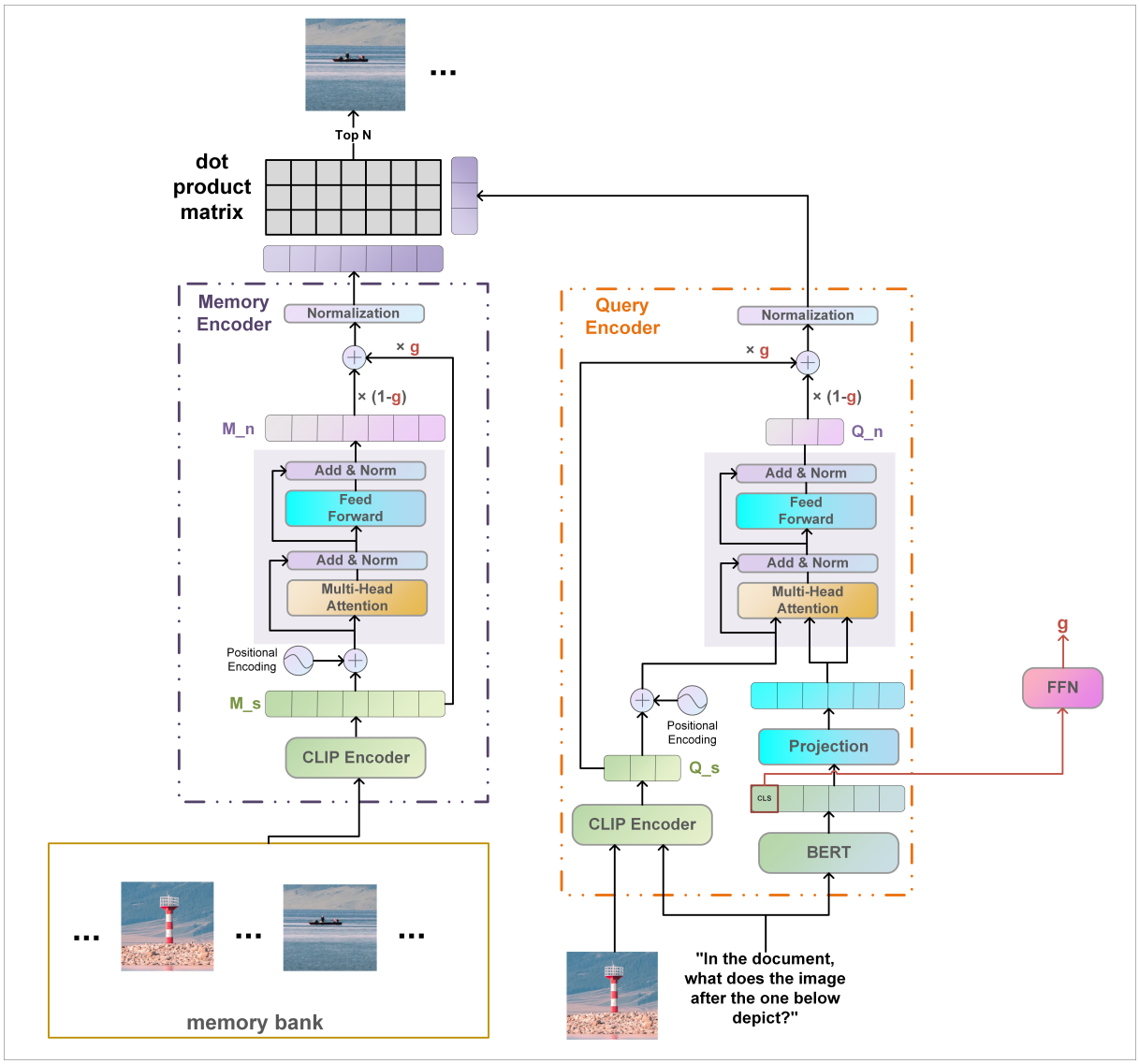

Figure 1: HSNSR is a dual-tower retrieval model that consists of a memory encoder and a query encoder. The memory encoder encodes the partitioned memory bank, while the query encoder encodes the query. The outputs of the two encoders are used to compute dot products, and the TopN values in the dot product matrix correspond to the original memory data, which are the retrieval results.

positive and negative examples on the model. The loss function can be expressed as follows

$$\mathrm{BCE}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\Big[\alpha y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\Big] \quad (3)$$

Here, y represents the correct answer, $\hat{y}$ corresponds to the dot product result, and α is the weight for positive examples.

During training, we freeze all the weights of the CLIP encoders and BERT, allowing only the transformer structure in the memory encoder, as well as the transformer, projection layer, and feed-forward network (FFN) in the query encoder, to be trained.

3.2 Semantic and Non-Semantic Retrieval Dataset

The SNSR dataset consists of three target modalities, encompassing a total of 26 multimodal semantic and non-semantic retrieval scenarios, with approximately 40,000 data entries. Each entry includes a context chunk memory and a query, with the memory containing up to 2,000 chunks, corresponding to about 40,000 words or 2,000 video frames. We pre-trained HSNSR on this dataset. Based on existing long-context multimodal bench-

marks and practical experience, we collected and organized several common scenarios for multi-modal non-semantic retrieval. These scenarios are categorized into two types: The first category is retrieval without semantic involvement: this includes image indices (such as a single image or multiple images specified by indices, e.g., the 3rd image or images/documents/conversations between the 3rd and 8th images), vague locations (e.g., beginning, end, etc.), content proportions (retrieving a portion of all content based on a ratio, e.g., the first half of a video or the last third of a document), specific video timestamps, and time ranges. The second category is retrieval with semantic involvement, where the query contains a segment of semantic information that acts as an anchor point. The content of the query is then determined by comparison with this anchor, for example: "What happened before/after/near a segment of video content featuring a car driving?" Although this type of retrieval requires semantic information, the semantics of the retrieval target are not present in the query itself.

We provide contextual content based on the datasets BookCorpus (Yao and Huang, 2018), Wikipedia (Xu and Lapata, 2019), OpenWebText2 (Liu, 2019), MultiWOZ (Wu, 2019), COCO (Lin et al., 2014), and shareGPT4V (Chen et al., 2023). Queries are generated using a large language model (LLM) to ensure diversity, tailored to the retrieval scenarios and randomly selected retrieval targets.

3.3 UM-Mem

UM-Mem is a streamlined memory model designed for multi-modal large language models (MLLMs) to handle image-text cross-dialogues. It employs HSNSR as the retrieval mechanism and is completely decoupled from the MLLM, allowing it to be used across different language models without requiring additional training. The memory bank of UM-Mem stores historical dialogue transcripts in chunks, with each chunk containing either a single text line or an image. Additionally, excessively long contexts can be directly input into the memory bank for subsequent dialogues. Unlike RAG, UM-Mem is not trained with any specific model, which may result in its performance being weaker than that of the RAG family on particular tasks. However, this characteristic enables it to validate the general effectiveness of non-semantic retrieval.

4 Experiments

4.1 Implementation Details

We trained our retriever on the SNSR dataset, utilizing CLIP-ViT-B/32 as the semantic encoder. The transformers in both encoders are single-layer structures. We prepend a two-digit value to the head of each CLIP encoded vector to distinguish its modality. The version of BERT we used is BERT-large-uncased.

We conducted 0-shot testing of InternVL-1.5 enhanced by the pre-trained UM-Mem on two image-text cross tasks in MM-NIAH (Wang et al., 2024a) to validate the effectiveness of the non-semantic retrieval knowledge learned by HSNSR. MM-NIAH is a benchmark specifically designed to evaluate the ability of existing multi-modal language models (MLMs) to understand long multi-modal documents. The benchmark includes three types of assessment tasks: multi-modal retrieval, counting, and reasoning, comprising approximately 12,000 samples in total. MM-NIAH extends the classic "needle in a haystack" (NIAH) tasks from the text domain into the multimodal domain, introducing a range of challenging tasks. For the multimodal haystack, the maximum number of tokens is limited to 72k, with up to 36 images. We chose the image-based counting and reasoning tasks for evaluation. In the reasoning task, the model is asked to answer a series of questions about the needles in the images. In the counting task, MM-NIAH requires the model to list the number of needles in each statement or image, rather than simply outputting the total count of needles. This task places high demands on the model's ability to comprehend long contexts, while being less reliant on pure semantic retrieval. We set the retrieval Top-N to 10.

We utilized UM-Mem to cache the context and performed retrieval based on the questions, feeding the retrieved results into the InternVL-1.5 (Chen et al., 2024b) model for evaluation. Table 1 shows a comparison of the performance of InternVL-1.5 with and without non-semantic retrieval enhancement on both tasks.

4.2 Experimental Results

As shown in Table 1, UM-Mem provides a performance boost to InternVL-1.5 in both tasks, with an overall improvement exceeding 3%. In the counting task, UM-Mem successfully mitigates the performance degradation that typically occurs as

Task	Model	1k	2k	4k	8k	12k	16k	24k	32k	40k	48k	64k	Overall
Counting Image Needle	InternVL-1.5	30.6	16.6	6.1	0.7	0.5	0.3	0.0	0.0	0.0	0.0	0.0	5.0
	InternVL-1.5+**RAG**	44.8	21.8	4.9	1.8	0.6	0.2	0.0	0.0	0.0	0.0	0.0	6.7
	InternVL-1.5+**UM-Mem**	35.1	27.0	15.9	6.4	2.7	1.2	0.0	0.0	0.0	0.0	0.0	**8.1**
Reasoning Image Needle	InternVL-1.5	49.2	52.8	49.5	50.1	51.3	48.5	53.2	48.9	44.4	51.1	52.2	50.1
	InternVL-1.5+**UM-Mem**	51.8	50.1	53.2	53.0	52.9	56.0	58.2	51.2	53.2	52.8	53.1	**53.2**

Table 1: The accuracy rates of the two tasks on MM-NIAH at different context lengths.

context length increases, compared to the original model. We also compared InternVL-1.5+RAG, and the results indicate that UM-Mem achieves higher accuracy than RAG, which relies solely on semantic retrieval. This strongly demonstrates the effectiveness and necessity of non-semantic retrieval.

In the reasoning task, as context length increases, UM-Mem maintains stable performance, outperforming the base model across most context lengths.

As multimodal large language models (MLLMs) are increasingly required to handle longer and more complex inputs, non-semantic retrieval becomes essential for improving their performance, enabling them to process and reason about multimodal data more effectively.

4.3 Ablation Studies

To further validate the effectiveness of non-semantic retrieval, we conducted ablation experiments. Specifically, we fixed the g value of HSNSR to 1, effectively disabling non-semantic retrieval. We continued testing on the counting and reasoning tasks involving images in MM-NIAH, and the results are presented in Table 2.

It is evident that after disabling non-semantic retrieval, the model's performance declined to varying degrees. Since the reasoning task is more reliant on semantic retrieval, its performance decreased less. We use 16k as a boundary point. For context lengths ranging from 1 to 16k (excluding 16k), the average accuracy of responses under pure semantic retrieval is 52.7, which is higher than HSNSR's 52.2. However, for context lengths greater than or equal to 16k, HSNSR achieves an average response accuracy of 54.08, surpassing the 53.36 accuracy of pure semantic retrieval. This indicates that as the context increases, its structure becomes more complex, and at this stage, non-semantic retrieval is better able to enhance model performance. In the more challenging counting task, semantic retrieval alone caused a rapid performance drop when the context length reached 8k. This strongly demonstrates that non-semantic retrieval can effectively enhance the model's ability to understand long multimodal contexts. It also highlights the necessity of addressing the non-semantic retrieval problem, which deserves further research and exploration.

5 Conclusion

In this paper, we introduced a novel retrieval problem in the domain of long-context tasks for multimodal large models, which we term non-semantic retrieval. This approach leverages the structural information within multi-modal long contexts to perform retrieval. We designed HSNSR, which builds upon semantic retrieval by further extracting structural features from the context and identifying non-semantic retrieval features within queries. This allows for the simultaneous execution of both semantic and non-semantic retrieval tasks.

We also collected and organized various non-semantic retrieval scenarios to construct the SNSR dataset, which is used to train the retrieval model. Finally, we developed UM-Mem, a memory model that is completely decoupled from MLLMs, utilizing a chunked memory bank to store long image-text cross contexts. We tested our UM-Mem on the challenging multi-modal long-context benchmark MM-NIAH, demonstrating gains and conducting ablation studies that effectively validate the existence of the non-semantic retrieval problem and the necessity of addressing it.

We believe that our work opens new research di-

Task	Model	1k	2k	4k	8k	12k	16k	24k	32k	40k	48k	64k	Overall
Counting Image Needle	InternVL-1.5+**CLIP**	35.1	21.3	15.9	0.8	0.3	0.1	0.0	0.0	0.0	0.0	0.0	7.7
	InternVL-1.5+**UM-Mem**	35.1	27.0	15.9	6.4	2.7	1.2	0.0	0.0	0.0	0.0	0.0	**8.1**
Reasoning Image Needle	InternVL-1.5+**CLIP**	53.5	52.8	53.1	52.5	51.6	53.0	53.6	54.7	56.8	48.3	53.8	53.0
	InternVL-1.5+**UM-Mem**	51.8	50.1	53.2	53.0	52.9	56.0	58.2	51.2	53.2	52.8	53.1	53.2

Table 2: Ablation Study Results.

rections in the fields of long-context tasks for multi-modal large language models, retrieval-augmented generation (RAG), and memory in multi-modal large language models. We hope it serves as a foundation for future research.

6 Limitations

In the collection of the SNSR dataset, all non-semantic retrieval scenarios were manually curated and collected, which means that the categories included cannot cover all task scenarios comprehensively. Furthermore, in the design of UM-Mem, we only considered one type of non-semantic retrieval related to contextual structure. However, in certain specific tasks, such as human dialogue tasks with MLLMs where temporal information, speakers, or details like patients and symptoms in medical assistant scenarios may also facilitate non-semantic retrieval for relevant memories. Due to the lack of effective evaluation tasks, UM-Mem focuses solely on contextual structure and has omitted the collection of additional information that could enhance its retrieval capabilities.

References

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

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

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. 2024a. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36:22820–22840.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. 2024a. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2024b. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36.

Chien-Sheng Wu. 2019. Learning to memorize in neural task-oriented dialogue systems. *arXiv preprint arXiv:1905.07687*.

Yumo Xu and Mirella Lapata. 2019. Weakly supervised domain detection. *Transactions of the Association for Computational Linguistics*, 7:581–596.

Wenlin Yao and Ruihong Huang. 2018. Temporal event knowledge acquisition via identifying narratives. *arXiv preprint arXiv:1805.10956*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. 2023. Unlock multi-modal capability of dense retrieval via visual module plugin. *arXiv preprint arXiv:2310.14037*.