External Memory Matters: Generalizable Object-Action Memory for Retrieval-Augmented Long-Term Video Understanding

Jisheng Dang[1,2,3], **Huicheng Zheng**[2*], **Xudong Wu**[4], **Jingmei Jiao**[5], **Bimei Wang**[6,3*], **Jun Yang**[5], **Bin Hu**[1], **Jianhuang Lai**[2] and **Tat Seng Chua**[3]
[1] School of Information Science and Engineering, Lanzhou University, China
[2] School of Computer Science and Engineering, Sun Yat-sen University, China
[3] School of Computing, National University of Singapore, Singapore
[4] School of Electronics and Information Technology, Sun Yat-sen University, China
[5] School of Electronic Information Engineering, Lanzhou Jiaotong University, China
[6] College of Cyber Security, Jinan University, China
dangjsh@mail2.sysu.edu.cn, zhenghch@mail2.sysu.edu.cn, wangbm@stu2021.jnu.edu.cn

Abstract

Long video understanding with Large Language Models (LLMs) enables the description of objects that are not explicitly present in the training data. However, continuous changes in known objects and the emergence of new ones require up-to-date knowledge of objects and their dynamics for effective understanding of the open world. To alleviate this, we propose an efficient **R**etrieval-**E**nhanced **V**ideo **U**nderstanding method, dubbed REVU, which leverages external knowledge to enhance the performance of open-world learning. First, REVU introduces an extensible external text-object memory with minimal text-visual mapping, involving static and dynamic multimodal information to help LLMs-based models align text and vision features. Second, REVU retrieves object information from external databases and dynamically integrates frame-specific data from videos, enabling effective knowledge aggregation to comprehend the open world. We conducted experiments on multiple benchmark datasets, and our model demonstrates strong adaptability to out-of-domain data without requiring additional fine-tuning or retraining. Experiments on benchmark video understanding datasets reveal that our model achieves state-of-the-art performance and robust generalization.

1 Introduction

Large Language Models (LLMs) excel in knowledge, context, multitasking, and high-quality text generation. Existing video content understanding tasks using LLMs [Chiang *et al.*, 2023; Touvron *et al.*, 2023] primarily rely on large pre-trained models trained on progressively larger datasets. These models are typically trained on vast datasets to acquire extensive knowledge and abilities, enabling them to handle various complex video understanding tasks, such as image or video classification [Li *et al.*, 2024; Ma *et al.*, 2024; Meng *et al.*, 2024; Wang *et al.*, 2024; Meng *et al.*, 2025], video object segmentation [Dang *et al.*, 2024a; Dang and Yang, 2021] and scene parsing [Dai *et al.*, 2024]. Despite strong performance in various tasks and benchmarks, their training data's timeliness limits their ability to process the latest knowledge, especially for new objects and dynamic long video content, resulting in poor performance in recognition and content generation.

One of the most significant challenges is that model performance often suffers when dealing with rare and ambiguous objects, as well as knowledge appearing frequently in very long videos. Such rare or uncommon objects or concepts are not adequately represented in traditional datasets, and thus may not be accurately recognized or understood by large models during training. In addition, long videos often contain complex scenes and dynamic backgrounds, which can cause existing models to fail to capture certain details or objects [Dang *et al.*, 2023a]. These issues typically require supplementation from external knowledge bases to help bridge the knowledge gap, significantly enhancing the model's understanding of video content.

Moreover, as new concepts continue to emerge, the task of understanding video content is also constantly evolving. The introduction of new elements and the dynamic nature of video content often lead to exponential increases in the computational costs of existing training models, posing more significant challenges for the training and application of large-scale language models [Liu *et al.*, 2024]. As the data volume grows, traditional training methods require substantial computing resources and face the challenge of delayed knowledge updates, making it difficult to keep pace with real-world changes on time. Therefore, continuously updating the model's object information and dynamic knowledge at a reasonable computational cost is crucial for achieving efficient video content understanding.

In this paper, instead of relying on larger datasets or more complex network architectures, we build a scalable external text-based object memory and propose an efficient retrieval-

* Corresponding author.

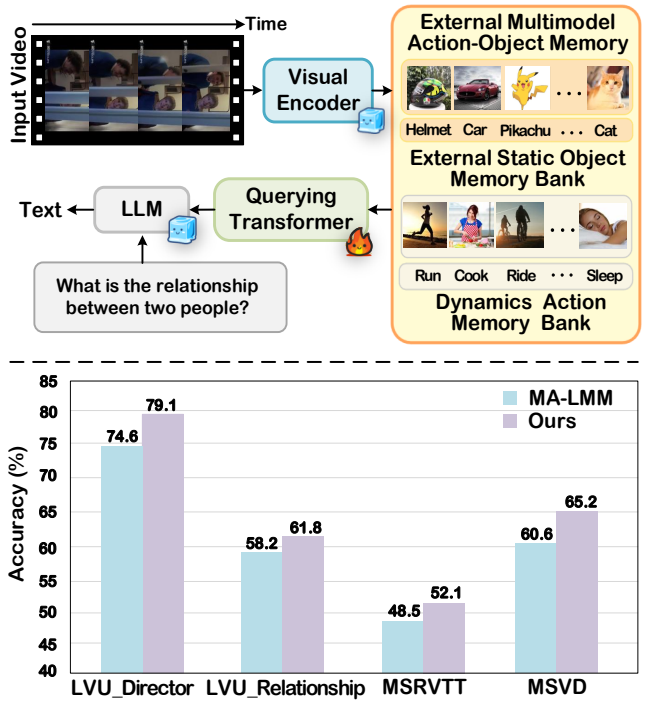

Figure 1: Top: Overview of REVU: An external text-object memory with static and dynamic multimodal information enhances text-vision alignment for open-world video understanding. REVU dynamically retrieves external memory and integrates it into videos for open-world video understanding. Bottom: Accuracy comparison across various video understanding benchmarks, demonstrating REVU's superior performance against MA-LMM.

augmented method to update this memory, addressing challenges in long-term video understanding. This external memory incorporates a minimal yet highly effective text-to-visual mapping mechanism that seamlessly integrates static and dynamic multimodal information, enabling LLM-based models to align textual and visual content more accurately and efficiently. Our method can access static object attributes, such as color, shape, and location, as well as dynamic video features including object motion trajectories, scene changes, and temporal coherence from the external memory. By leveraging these capabilities, the model effectively handles the complexity and dynamics of long videos, enhancing both efficiency and accuracy in long-term video understanding.

The key contribution of REVU is the construction of two key-value memories, similar to [Vo *et al.*, 2022; Dang *et al.*, 2023b]. The first key-value pair represents the features and names (semantic labels) of static objects, while the second captures the characteristics and names (semantic labels) of dynamic actions. Unlike the previous approaches [Vo *et al.*, 2022; Dang *et al.*, 2024b], which define dependent objects as keys, our method uses the visual characteristics of objects and actions as keys, taking advantage of the rich availability of object images and video data on the Internet.

In summary, our key contributions are as follows:

- We provide an extensible external object-action memory with minimal but useful text-visual mapping, which enables LLMs-based models to align textual and visual information to comprehend the open world.
- We propose a highly effective retrieval-augmented method for long-term video understanding, which can adaptively retrieve world knowledge, i.e., static object information from the external object-text memory and dynamic information from action-text memory.
- Our approach achieves state-of-the-art performance on various downstream video tasks, including long-term video understanding and video question answering. Furthermore, our method demonstrates strong generalization capabilities.

2 Related Work

2.1 Multimodal Large Language Models

Recent advancements in Multimodal Large Language Models (MLLMs) have significantly enhanced the integration of diverse modalities, including natural language processing and computer vision. By combining text, image and audio data, MLLMs can perform more complex tasks, including text and image generation, visual question answering, cross-modal retrieval and multimodal sentiment analysis. These models not only enhance unimodal performance but also advance cross-modal understanding and generation. For example, Contrastive Language-Image Pretraining (CLIP) [Radford *et al.*, 2021] improves zero-shot image classification and cross-modal tasks through contrastive learning on large-scale image-text pairs. The DALL-E series [Ramesh *et al.*, 2021] demonstrates the capability to generate high-quality images from textual descriptions, thereby expanding the application scope of generative models. Flamingo [Alayrac *et al.*, 2022] enhances multi-round dialogues and complex scene understanding, enabling more natural human-computer interactions. Recent models such as GPT-4 [Achiam *et al.*, 2023] and PaLM-E [Driess *et al.*, 2023] extend multimodal capabilities by incorporating additional modalities and optimizing architectures, further improving generalization and adaptability. These advancements not only push the boundaries of multimodal understanding and generation but also support practical deployments of intelligent systems.

2.2 Long-Term Video Understanding

Long-term video understanding aims to capture intricate spatiotemporal dynamics and semantic information over extended durations by analyzing prolonged video sequences. Recent advances in deep learning architectures and computational resources have significantly propelled progress in this field. For example, VideoMAE [Tong *et al.*, 2022] employs the Mask Autoencoder (MAE) framework to perform occlusion reconstruction on videos, thus significantly enhancing video representations and improving the performance of downstream tasks. The Action Graph Transformer (AGT) [Nawhal and Mori, 2021] effectively captures complex action relationships by introducing a graph-structured attention mechanism, thereby enabling accurate recognition and prediction of multi-level behavior sequences. VTimeLLM

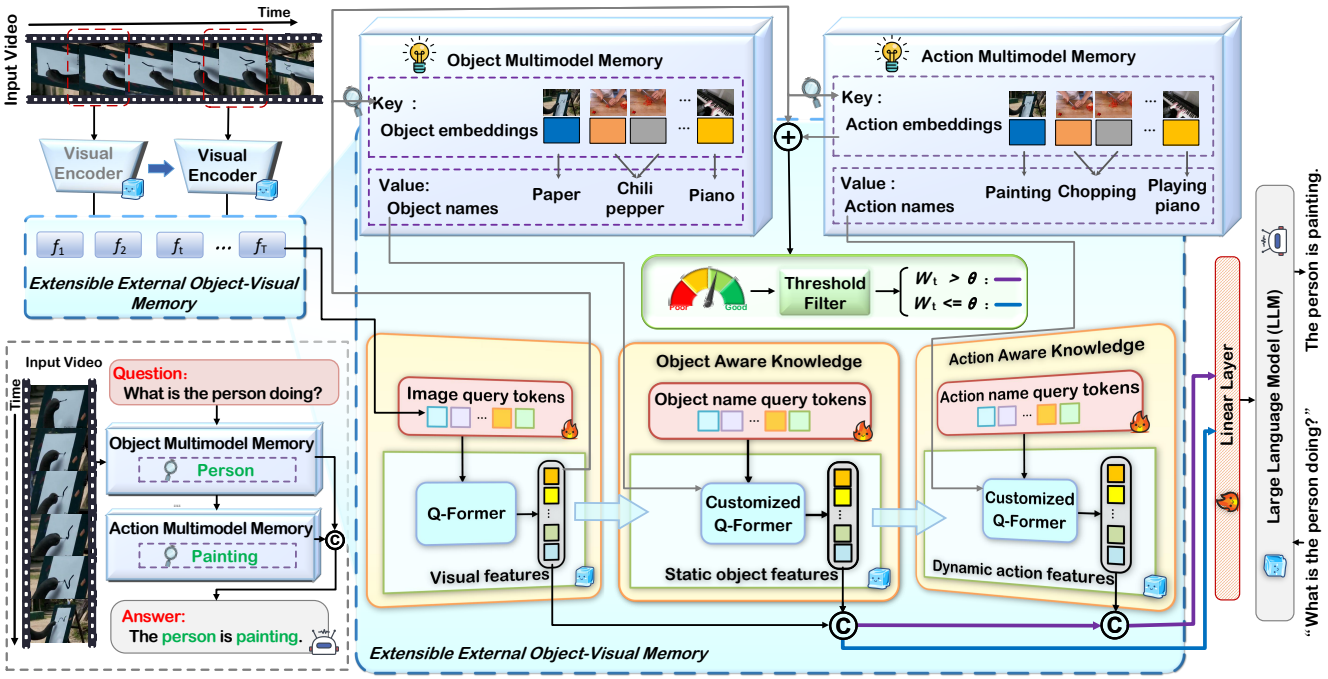

Figure 2: **Overview of the proposed REVU framework.** REVU mainly integrates an extensible object-visual memory and action multimodal memory. Visual encoders extract frame-level features from input videos. Object memory stores embeddings and names, while action memory stores embeddings and actions from the external world. Customized Q-Formers refine object and action features using trainable query tokens and threshold filtering. These refined features are dynamically combined with visual features and fed into an LLM for text decoding.

[Huang *et al.*, 2023] improves video understanding by improving temporal modeling, effectively capturing events, actions, and contextual changes over time.

2.3 Memory Models in Vision Tasks

Memory models in vision tasks aim to improve the storage, retrieval, and processing of complex visual information by leveraging memory mechanisms. Computer vision tasks such as image classification, object detection, and video understanding require efficient processing of large-scale visual data. To handle long videos, some works [Dang and Yang, 2022; Dang *et al.*, 2024c] employ 3D Convolutional Neural Networks (CNNs) to capture motion information of objects and persons, enabling the construction of long-term feature banks. However, traditional CNNs struggle to model long-term dependencies in visual data. To address this, researchers have developed memory models that enhance the ability to represent complex spatiotemporal information through external or internal memory mechanisms. DynMemNet [Kumar *et al.*, 2015] efficiently comprehends dynamic scenes in videos by introducing a dynamically updated memory mechanism. It excels in action recognition and event prediction tasks, enabling real-time updates to memory content and adaptation to complex visual environments. SelfMem [Cheng *et al.*, 2023] employs a self-supervised learning method to enhance the representational capacity of its memory module through pre-training on large-scale unlabeled video data. It excels in video classification and video question answering tasks, demonstrating the potential of self-supervised memory models to improve visual task performance.

3 Method

3.1 Overview of REVU

We propose an innovative video content understanding model based on retrieval-augmented LLM. As shown in Fig. 2, this model incorporates an extensible external text-object storage that seamlessly integrates static and dynamic multimodal information through an efficient text-visual mapping mechanism. This mechanism enhances the model's alignment capability, making the matching between visual information and textual descriptions more accurate, particularly in complex multimodal scenarios. To further enhance the model's performance in multimodal information fusion, we design two scalable external memory modules: one for static information and another for dynamic information. These memory modules enable the model to retrieve relevant information from external resources, thereby leveraging this data for reasoning and generating descriptions more accurately, particularly when handling complex scenes. Unlike traditional visual-name pair memory models, our external memory does not rely on a single visual-name pair for each object. Instead, it constructs a diverse memory by collecting visual features from multiple objects and actions. This approach significantly enhances the diversity and flexibility of the memory, allowing the model to maintain efficient comparative tension even when processing richer and more complex video scenes.

To achieve this, our method leverages a frozen, pre-trained

visual model and adds several trainable layers of LLMs on top, ensuring low computational costs during training. Specifically, we first align visual features extracted from video content with dynamic and static feature embeddings stored in external memory and then retrieve the corresponding object and action names. These retrieved features provide critical context to guide the LLM in generating more accurate descriptions. Additionally, we introduce an attention fusion module that effectively filters out irrelevant retrievals, automatically focusing on the most relevant object and action features to optimize the generation process. After fusion, the integrated visual and object name features form the final prompt, which is input to the LLM for description generation. This design allows the model to efficiently handle long video understanding tasks while significantly lowering training and inference costs without compromising performance.

3.2 External Multimodel Object-Action Memory

External Memory Source. To construct the Extensible External Text-Visual Memory, we first collect object-name and action-name pairs from multiple open standard datasets (e.g., LVIS, kinetics700) and gather visual information from the internet. Each collected object-name pair and action-name pair represents the diversity of an object or action in the visual space. Each video clip is processed by a trained visual encoder (e.g., ViT) to generate high-dimensional visual embeddings. These visual embeddings are stored as keys in external memory, with the corresponding object or action names stored as values. This design allows the memory to store not only the visual features of objects or actions but also their associated semantic labels (i.e., names). During inference, visual embeddings are matched with object or action features in external memory to retrieve the corresponding labels.

Key-Value Mapping. For each video, we first use a frozen Visual Encoder to process the video frames sequentially, extracting both spatial and temporal features. After encoding each frame, the generated feature vector captures the spatial information of that frame while also embedding temporal changes, which aids in understanding the dynamic processes in the video. These encoded features are stored as keys in external memory, with the corresponding object/action names for each frame stored as values. Specifically, we constructed a Visual Name Memory that stores both the visual features of each object or action in the video and their corresponding semantic labels (object names or action names). This design allows the memory system to accurately retrieve relevant object or action information, providing robust support for subsequent model inference. To enhance retrieval efficiency, we utilize FAISS [Douze *et al.*, 2024] to build an index and perform fast vector searches by measuring the similarity between visual features.

Dynamic Retrieval. We employ fast and slow query methods to independently retrieve Object Multimodal Memory and Action Multimodal Memory, respectively. Given the significant differences in semantic and dynamic features between objects and actions, their retrieval requirements and optimization strategies are also distinctly different.

In the object multimodal memory, we focus on understanding static objects and scenes. Since objects have stable spatial distributions in video frames, their features primarily involve visual information such as shape, color, and texture. To efficiently retrieve relevant object information, we adopt a combination of fast and slow queries. The multimodal information of objects includes static features and potential detail changes. The fast query method quickly retrieves and calculates similarity, while the slow query refines the results by matching object names to textual information and formatting it according to the model's requirements. This approach enables rapid retrieval of object names and their multimodal features from external memory.

In contrast, the Action Multimodal Memory focuses on dynamic changes and action sequences in videos. Actions are typically composed of consecutive frames, whose features include not only visual information but also temporal dynamics. Therefore, action retrieval is more complex than object retrieval, as it requires considering temporal feature changes. Moreover, not all videos contain obvious actions. To reduce unnecessary computation and avoid introducing erroneous information, we use the fast query method for preliminary action retrieval and similarity calculation. We set a default threshold: if the similarity value exceeds the threshold, the video is considered to contain action features, prompting the slow queries to extract relevant action information from Action Multimodal Memory. If the similarity is below the threshold, the video is assumed to lack action features, and the Action Multimodal Memory is skipped.

Fast and slow queries complement each other to enhance both retrieval efficiency and accuracy. Fast queries quickly locate relevant objects or actions, while slow queries provide precise details and semantic alignment. This decoupled design allows better handling of static and dynamic information in videos, ensuring effective utilization of multimodal object and action features for video understanding. Additionally, the query strategy can be dynamically adjusted based on specific requirements, further improving overall performance.

3.3 Retrieval-Augmented Visual Encoding

Visual Encoder. Previous video content understanding methods typically extracted multimodal features through frozen visual encoders, converting them into a unified representation. A Memory Bank stored historical information to support dynamic queries, while Q-former generated task-specific representations by querying relevant data from the Memory Bank, enabling efficient fusion of multimodal information. However, although the memory bank is highly effective in storing and querying historical data, its capacity is limited and cannot extend beyond the knowledge of the current input. As a result, the model's performance may suffer when dealing with rare, ambiguous, or neglected objects in widely used datasets. Therefore, without an external memory base, the performance of existing models remains constrained.

Retrieval-Augmented Q-Former. The visual encoder generates contextual features from the video, which may lack background knowledge or additional commonsense information about objects and their relationships. For instance, while the visual encoder may identify a person in a scene, it might fail to grasp the nuances of their actions or the broader context, such as their role in a social interaction or activity. This lim-

itation can lead to hallucinations or incorrect interpretations, particularly when encountering rare objects or actions not included in the training data. Leveraging external memory, without needing to increase the size of the core model, could flexibly inject diverse and representative knowledge. By doing this, the model could mitigate the hallucination caused by videos with rarely presented objects or actions, increasing the realibilty of the final output.

To address these limitations, we propose integrating external memory into the model without increasing its core size. By utilizing external memory, we inject diverse, representative information about objects and actions, allowing the model to incorporate knowledge not explicitly present in the video content but essential for understanding the broader context. By utilizing external memory, the model can mitigate hallucinations that arise when videos contain rare or unseen objects and actions. For instance, if an uncommon object appears in the video that the core visual encoder has not encountered, the external memory supplements the encoder's interpretation, providing the necessary context. This enhances the model's ability to understand rare or ambiguous scenes, improving the reliability and robustness of the final output, and ensuring correct interpretation of unfamiliar elements.

Specifically, we use two additional Q-formers that independently process retrieved results R_{action} and R_{object}. We employ two additional Q-formers that independently process the retrieved results for objects and actions, denoted as R_{object} and R_{action}. These Q-formers are essential for integrating external knowledge with the video's visual features. Given the encoded visual feature V_{vid} extracted from the video frames by the Visual Encoder, we retrieve the corresponding action and object information from external memory, denoted as:

$$R_{\text{object}} = \text{Retrieve}(\text{objects}, V_{\text{vid}}), \tag{1}$$

$$R_{\text{action}} = \text{Retrieve}(\text{actions}, V_{\text{vid}}). \tag{2}$$

Each Q-former independently processes the retrieved results using stacked cross-attention and self-attention mechanisms to generate two distinct contextual representations object contextual representation C_{object} and action contextual representation C_{action} as follows:

$$C_{\text{action}} = \text{Q-former}_{\text{action}}(R_{\text{action}}, V_{\text{vid}}), \tag{3}$$

$$C_{\text{object}} = \text{Q-former}_{\text{object}}(R_{\text{object}}, V_{\text{vid}}). \tag{4}$$

Each Q-former comprises L stacked layers of cross-attention and self-attention mechanisms, allowing the model to refine and integrate retrieved knowledge with visual features. The cross-attention mechanism aligns external knowledge (object and action names) with the video's visual content, while self-attention captures long-range dependencies within the retrieved knowledge, ensuring coherent representations.

After processing by the Q-formers, the two contextual representations C_{object} and C_{action} are fused into a unified representation. This final representation captures both dynamic actions and static objects, providing a comprehensive understanding of the video content.

$$C_{\text{final}} = \text{Fusion}(C_{\text{object}}, C_{\text{action}}). \tag{5}$$

The final representation C_{final} is passed to the LLM for caption generation or other downstream tasks. This ensures that the output is informed by both the visual features and the externally retrieved knowledge, enabling the model to generate more reliable and accurate interpretations, particularly for rare or complex scenes. By incorporating external knowledge in this structured manner, we overcome the limitations of the Visual Encoder, ensuring that the model remains robust and accurate even with rare or unseen objects and actions.

Dynamic Skipping. The introduction of two additional Q-forms increases computational complexity. To balance computational load with performance, we designed a dynamic retrieval module that selects query strategies based on video content characteristics, reducing unnecessary computation.

3.4 Text Decoding

During training, we process video frames autoregressively, with the Q-Former outputting a representation containing all historical information at the final time step, which is then fed into the LLM. This approach reduces the number of input text tags, addressing the context length limitation of the LLM and lowering GPU memory requirements. By compressing the dynamic information of the video, the Q-Former provides a concise representation that helps the LLM better understand video semantics. The model is supervised with a cross-entropy loss function, aiming to generate text descriptions for the video. Only the parameters of the Q-Former are updated, while the visual encoder and language model remain frozen. The LLM leverages the self-attention mechanism to process video features and generate accurate text, effectively addressing the context length limitation and enhancing model performance.

4 Experiments

4.1 Tasks and Datasets

We conducted experiments on long video understanding tasks using three widely used datasets: LVU [Wu and Krahenbuhl, 2021], MSRVTT [Xu *et al.*, 2017], and MSVD [Chen and Dolan, 2011], with top-1 classification accuracy as the primary metric. LVU is a large-scale, multimodal video dataset covering diverse domains with complex scenes, objects, and activities, making it suitable for various video analysis tasks. MSVD contains 120,000 sentences describing over 2,000 video clips, collected via Mechanical Turk. MSRVTT includes 10,000 clips from 20 categories, each annotated with 20 captions.

4.2 Implementation Details

Our framework is built upon InstructBLIP [Dai *et al.*, 2024] using methods outlined in [He *et al.*, 2024]. A 3×3 convolutional layer acts as a projector, and the pre-trained ViT-G/14 [Dosovitskiy, 2020] image encoder from EVA-CLIP [Fang *et al.*, 2023] is employed to create spatiotemporal memory. For fair comparison, we use the same backbone architecture as baseline [He *et al.*, 2024]. The Q-Former weights from InstructBLIP are used for query mechanism, while Vicuna-7B [Chiang *et al.*, 2023] serves as the large-scale language model. Training was performed on four NVIDIA 4090 GPUs

Method	Relation	Director	Year	Scene	Avg
Performer [Choromanski *et al.*, 2020]	50.0	58.9	41.3	60.5	52.7
Orthoformer [Patrick *et al.*, 2021]	50.0	55.1	43.4	66.3	53.7
VideoBERT [Sun *et al.*, 2019]	52.8	47.3	36.1	54.9	47.8
Obj_T4mer [Wu and Krahenbuhl, 2021]	54.8	47.7	37.8	52.9	48.3
VIS4mer [Islam and Bertasius, 2022]	57.1	62.6	44.8	67.4	58.0
LST [Islam and Bertasius, 2022]	52.5	56.1	39.2	62.8	52.7
MA-LMM [He *et al.*, 2024]	58.2	74.6	51.9	80.3	66.3
Ours	**60.4**	**79.1**	**52.6**	**80.5**	**68.1**

Table 1: Comparative results with state of the art methods on the LVU dataset.

Model	MSRVTT	MSVD
SINGULARITY [Lei *et al.*, 2022]	43.5	-
GiT [Lei *et al.*, 2022]	43.2	56.8
mPLUG-2 [Xu *et al.*, 2023]	48.0	58.1
UMT-L [Li *et al.*, 2023]	47.1	55.2
Valley [Luo *et al.*, 2023]	<u>51.1</u>	60.5
MA-LMM [He *et al.*, 2024]	48.5	<u>60.6</u>
Ours	**52.1**	**65.2**

Table 2: Comparison with state-of-the-art methods on video question answering. **Bold** and <u>underline</u> denote top-1 and top-2 results.

Dataset Ablation	Baseline	Ours
Splitting training set (70% train-30% val)	56.0	58.9
Splitting training set (50% train-50% val)	46.2	48.0

Table 3: Dataset ablation results on LVU training set for relationship task. Our method demonstrates stronger generalization ability.

using cosine learning rate decay. We sample 100 frames per video clip from the LVU dataset at 1 FPS.

4.3 Main Results

Long-Term Video Recognition on the LVU Dataset. As shown in Table 1, we compare our method with state-of-the-art approaches on the LVU [Wu and Krahenbuhl, 2021] dataset across five categories: Relationship, Director, Year, Scene, and Overall Average (Avg). Our method achieves the highest average score of 68.1, outperforming others in the 'Scene' (80.5) and 'Director' (79.1) categories. It also surpasses most competitors in the 'Relationship' category (60.4) and remains competitive in the 'Year' category (52.6). Our method also outperforms Performer [Choromanski *et al.*, 2020] (52.7) and Orthofomer [Patrick *et al.*, 2021] (53.7). While 'MA-LMM' [He *et al.*, 2024] scores slightly higher in 'Year' (74.6), it falls short in 'Scene' and 'Director', resulting in lower overall performance. These results highlight the effectiveness of our method in leveraging spatial and temporal cues for video understanding, delivering balanced performance across all categories. Its strong performance in the 'Scene' and 'Director' categories underscores its ability to handle complex video content, making it well-suited for large-scale multimodal video analysis.

Video Question Answering on the MSVD Dataset. Table 2 highlights the performance of our proposed method on the Video Question Answering task, specifically on the MSVD [Chen and Dolan, 2011] dataset. The results demonstrate the significant advantages of our method over existing state-of-the-art approaches. Our method achieves an accuracy of 65.2%, significantly outperforming another method and showcasing a substantial performance improvement. In comparison, the second-best method, the baseline MA-LMM, achieves an accuracy of 60.6%, indicating a notable improvement of 4.6% by our approach. Furthermore, compared to VideoCoCa, our method achieves an 8.3% improvement, reflecting its superior cross-modal understanding and robust question answering capabilities. The figure also includes results from several prominent approaches introduced in recent years, ranging from *JustAsk* (2021) to the latest *Valley* (2023). Our method not only leads to accuracy but also demonstrates stronger robustness and generalization in long-term video question answering tasks.

Video Question Answering on the MSRVTT Dataset. Table 2 illustrates the performance of our proposed method on the video question answering task, specifically evaluated on the MSRVTT dataset, and compares it with various state-of-the-art methods. The results clearly demonstrate the significant accuracy advantage of our approach over existing methods. Our method achieves a top accuracy of 52.1% on the MSRVTT dataset, outperforming all current methods and showcasing strong performance improvements. The second-best method, *Valley*, achieves an accuracy of 51.1%, which is 1.0% lower than our approach. Compared to *MA-LMM*, which achieves 48.5%, our method improves by a notable 3.6%, further highlighting its superior capability in video understanding tasks. These results establish our approach as a new benchmark for video question answering, reflecting its robust and effective design.

Generalization. Table 3 evaluates the generalization capability of our method on the LVU dataset's Relationship task with varying train-validation splits. With a 50%-50% split, our method outperforms the baseline, demonstrating strong generalization. For the 70%-30% split, it achieves Top-1 accuracy of 58.87%, showcasing its ability to effectively utilize training data while maintaining robust performance. The results confirm the adaptability of our method to different data distributions and highlight its generalization effectiveness.

4.4 Visualization

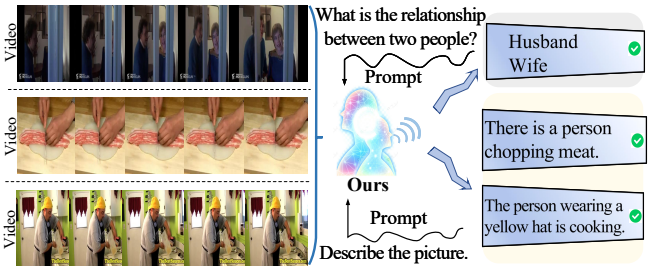

Figure 3: Visualization results of our method on long-term video recognition task on LVU dataset.

Figure 3 shows the experimental results of our method on six long-term video recognition tasks on the LUV dataset, each representing varying levels of video comprehension. The tasks span recognizing relationships between people and identifying actions, with each experiment using real movie scenes. These scenes were carefully chosen to maintain high semantic relevance and exhibit diversity in visual style, lighting and frame rate. The figure shows input videos for different tasks and the corresponding model outputs. In the first task, the model successfully identifies the relationship between two people (husband and wife), while in the second task, it accurately recognizes actions (e.g., 'chopping meat' and 'cooking'). The results demonstrate that our method excels across various video recognition tasks, showcasing its ability to handle diverse and complex visual inputs.

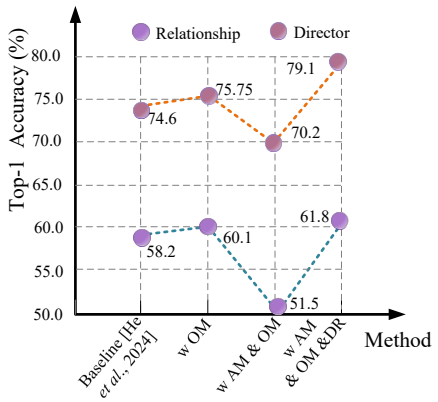

Figure 4: Ablation of the external multimodal memory on the LVU dataset for relationship and director task. Here, OM, AM, and DR denote object memory, action memory, and dynamic retrieval.

In Figure 4, we conduct an ablation study on the LVU dataset to verify the impact of external multimodal memory components including Object Memory (OM), Action Memory (AM), and Dynamic Retrieval (DR) on the baseline model's performance in the Relationship and Director tasks. Adding OM improves task accuracy, but combining OM and AM leads to performance loss without Dynamic Retrieval. Integrating all three components achieves the best performance, with accuracies of 61.8% (Relationship) and 79.1% (Director), highlighting the role of DR in complementing OM and AM for better long-term video understanding.

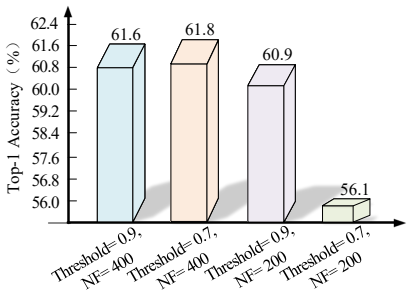

Figure 5: Ablation of external multimodal memory and dynamic retrieval. NF denotes the number of frames in action-text memory bank.

Figure 5 presents an ablation study evaluating the effect of threshold values and the number of frames (NF) in the action-text memory bank on Top-1 accuracy. Results indicate that lower thresholds consistently improve Top-1 accuracy over higher thresholds at a fixed large NF. For instance, with NF=400, reducing the threshold from 0.9 to 0.7 increases Top-1 accuracy from 61.6% to 61.8%. Increasing NF from 200 to 400 improves performance across all thresholds, highlighting the importance of a larger memory bank for better action-text alignment. These findings underscore the critical role of dynamic retrieval strategies and memory size in optimizing model performance for long video understanding tasks.

5 Conclusion

In this work, we introduce an efficient retrieval-enhanced method for long-term video understanding, addressing challenges of dynamic object evolution and the emergence of new objects in open-world scenarios. By leveraging a lightweight and easily trainable model, our approach integrates static object information from external memory with dynamic contextual cues from videos, enhancing LLMs with world knowledge. Through seamless alignment of retrieved textual data with static and dynamic features, our method achieves robust generalization across domains without requiring additional fine-tuning or retraining.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 61976231, Grant 61972435, Grant U20A20185, and Grant U1611461; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515012853, Grant 2022B1515020103, and Grant 2019A1515011869; and in part by the Shenzhen Science and Technology Program under Grant RCYX20200714114641140.

Contribution Statement

Bimei Wang and Huicheng Zheng contributed equally to the correspondence.

References

[Achiam *et al.*, 2023] OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, and Florencia Leoni Aleman et al. Gpt-4 technical report. 2023.

[Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022.

[Chen and Dolan, 2011] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.

[Cheng *et al.*, 2023] Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. Lift yourself up: Retrieval-augmented text generation with self memory. *ArXiv*, abs/2305.02437, 2023.

[Chiang *et al.*, 2023] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[Choromanski *et al.*, 2020] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[Dai *et al.*, 2024] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[Dang and Yang, 2021] Jisheng Dang and Jun Yang. Higcnn: Hierarchical interleaved group convolutional neural networks for point clouds analysis. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2825–2829, 2021.

[Dang and Yang, 2022] Jisheng Dang and Jun Yang. Lhphgcnn: Lightweight hierarchical parallel heterogeneous group convolutional neural networks for point cloud scene prediction. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):18903–18915, 2022.

[Dang *et al.*, 2023a] Jisheng Dang, Huicheng Zheng, Jinming Lai, Xu Yan, and Yulan Guo. Efficient and robust video object segmentation through isogenous memory sampling and frame relation mining. *IEEE Transactions on Image Processing*, 32:3924–3938, 2023.

[Dang *et al.*, 2023b] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, and Yulan Guo. Unified spatio-temporal dynamic routing for efficient video object segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(5):4512–4526, 2023.

[Dang *et al.*, 2024a] Jisheng Dang, Huicheng Zheng, Bimei Wang, Longguang Wang, and Yulan Guo. Temporo-spatial parallel sparse memory networks for efficient video object segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):17291–17304, 2024.

[Dang *et al.*, 2024b] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, Longguang Wang, and Yulan Guo. Beyond appearance: Multi-frame spatio-temporal context memory networks for efficient and robust video object segmentation. *IEEE Transactions on Image Processing*, 2024.

[Dang *et al.*, 2024c] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, Longguang Wang, Qingyong Hu, and Yulan Guo. Adaptive sparse memory networks for efficient and robust video object segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Douze *et al.*, 2024] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazar'e, Maria Lomeli, Lucas Hosseini, and Herv'e J'egou. The faiss library. *ArXiv*, abs/2401.08281, 2024.

[Driess *et al.*, 2023] Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023.

[Fang *et al.*, 2023] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.

[He *et al.*, 2024] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 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, 2024.

[Huang *et al.*, 2023] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14271–14280, 2023.

[Islam and Bertasius, 2022] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022.

[Kumar *et al.*, 2015] Ankith Jain Rakesh Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*, 2015.

[Lei *et al.*, 2022] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022.

[Li *et al.*, 2023] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19948–19960, 2023.

[Li *et al.*, 2024] Xiangxian Li, Yuze Zheng, Haokai Ma, Zhuang Qi, Xiangxu Meng, and Lei Meng. Cross-modal learning using privileged information for long-tailed image classification. *Computational Visual Media*, 10(5):981–992, 2024.

[Liu *et al.*, 2024] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models. *ArXiv*, abs/2402.17177, 2024.

[Luo *et al.*, 2023] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.

[Ma *et al.*, 2024] Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Jie Zhou. Triple sequence learning for cross-domain recommendation. *ACM Transactions on Information Systems*, 42(4):1–29, 2024.

[Meng *et al.*, 2024] Lei Meng, Zhuang Qi, Lei Wu, Xiaoyu Du, Zhaochuan Li, Lizhen Cui, and Xiangxu Meng. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Meng *et al.*, 2025] Lei Meng, Xiangxian Li, Xiaoshuo Yan, Haokai Ma, Zhuang Qi, Wei Wu, and Xiangxu Meng. Causal inference over visual-semantic-aligned graph for image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19449–19457, 2025.

[Nawhal and Mori, 2021] Megha Nawhal and Greg Mori. Activity graph transformer for temporal action localization. *ArXiv*, abs/2101.08540, 2021.

[Patrick *et al.*, 2021] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[Ramesh *et al.*, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021.

[Sun *et al.*, 2019] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.

[Tong *et al.*, 2022] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *ArXiv*, abs/2203.12602, 2022.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Vo *et al.*, 2022] Duc Minh Vo, Hong Chen, Akihiro Sugimoto, and Hideki Nakayama. Noc-rek: Novel object captioning with retrieved vocabulary from external knowledge. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17979–17987, 2022.

[Wang *et al.*, 2024] Yuqing Wang, Lei Meng, Haokai Ma, Yuqing Wang, Haibei Huang, and Xiangxu Meng. Modeling event-level causal representation for video classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3936–3944, 2024.

[Wu and Krahenbuhl, 2021] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021.

[Xu *et al.*, 2017] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[Xu *et al.*, 2023] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*, pages 38728–38748. PMLR, 2023.