MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding
Supplementary Material

Bo He[1,2] Hengduo Li[2] Young Kyun Jang[2] Menglin Jia[2] Xuefei Cao[2]
Ashish Shah[2] Abhinav Shrivastava[1] Ser-Nam Lim[2]

[1]University of Maryland, College Park [2]Meta [3]University of Central Florida

https://boheumd.github.io/MA-LMM/

We report additional ablation experiments in Sec. 1. We also present more qualitative results on the video captioning tasks in Sec. 2. And in Sec. 3, we show more dataset-specific implementation details and hyper-parameters. Finally, we discuss some limitations and future works in Sec. 4.

1. Additional Experiments

Memory bank compression at different spatial levels. In Table 10, we show comparison results of compressing the memory bank at different spatial levels (frame-level vs. token-level) on the LVU [1], Breakfast [2] and COIN [3] datasets. For the frame-level compression, we calculate the cosine similarity between adjacent frame features and average the frame-level features with the highest similarity. For the token-level compression, the cosine similarity is calculated between tokens at the same spatial location across the entire temporal axis, given that each frame-level feature contains multiple tokens at different spatial locations. The results indicate that token-level compression consistently surpasses frame-level compression in performance. Particularly, on the Breakfast dataset, the token-level surpasses the frame-level by 6.5% in top-1 accuracy. This superiority can be attributed to the importance of recognizing the object type of breakfast in videos. And token-level compression can help preserve much more fine-grained spatial information and details.

Inference time of different input frames In Figure 6, the inference time of MA-LMM increases linearly with respect to the frame lengths, due to its auto-regressive design of processing video frames sequentially. In contrast, directly concatenating frame-level features takes much longer time and higher GPU memory consumption, since it needs to process all video frames simultaneously.

Table 10. Memory bank compression at different spatial levels.

Spatial Level	LVU	Breakfast	COIN
Frame-level	61.8	86.5	91.1
Token-level	**63.0**	**93.0**	**93.2**

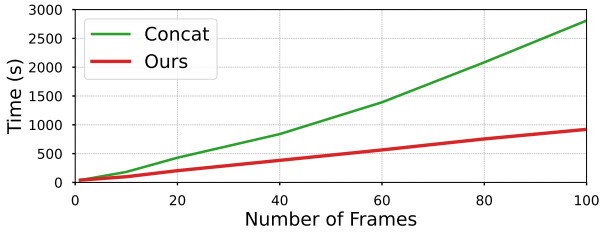

Figure 6. Inference time vs. input frame length.

2. More Qualitative Results

Our model's enhanced capabilities in video captioning are further showcased through additional visualization results in Figure 7. Here, our MA-LMM significantly outperforms Video-LLaMA [4] in generating detailed and accurate sentence descriptions. For instance, in the first video, our model precisely describes the action as "remove the onion rings and place them on the paper towel," capturing the entire action steps, while Video-LLaMA's description lacks this completeness, notably missing the crucial action of removing the onion rings. In the second video example, our model distinguishes itself by accurately identifying subtle details such as specific ingredients: chili powder, salt, and garlic powder, which Video-LLaMA overlooks. This highlights the enhanced capability of our MA-LMM in recognizing and describing fine-grained details.

3. Experiment Details

We show the details of hyper-parameters in the following table for different tasks and datasets. For all the experi-

Ground-Truth: *remove the onions and place on paper towel*
Video-LLaMA: *fry the onion rings in oil*
Ours: *remove the onion rings from the oil and place them on a paper towel*

Ground-Truth: *add garlic powder chili powder paprika salt cayenne pepper buffalo wing sauce to the wings and mix*
Video-LLaMA: *coat the chicken wings with the sauce*
Ours: *add chili powder salt garlic powder onion powder and paprika to the chicken and mix*

Figure 7. Visualization results on the video captioning task.

ments, we use a cosine learning rate decay. Table 11 shows the hyper-parameters for the long-term video understanding task. For the LVU dataset, we follow the same practice in [5, 6], we sample 100 frames of 1 fps for each video clip. For the Breakfast [2] and COIN [3], we uniformly sample 100 frames from the whole video. Table 12 shows the hyper-parameters on the MSRVTT-QA [7], MSVD-QA [7], and ActivityNet-QA [8] datasets for the video question answering task while Table 13 presents the hyperparameters on the MSRVTT [9], MSVD [10], YouCook2 [11] datasets for the video captioning tasks.

4. Limitation and Future Work

Since our model takes in video frames in an online manner, leading to reduced GPU memory usage, but at the cost of increased video processing time. This trade-off becomes particularly noticeable with extremely long videos, where processing times can become significantly prolonged. To mitigate this issue, we suggest a hierarchical method to process extremely long-term video sequences. This strategy involves dividing extensive videos into smaller segments and then processing each segment sequentially in an auto-regressive fashion as we present in the main paper. Then we can employ additional video modeling techniques to model inter-segment relationships. This method aims to strike a balance between memory efficiency and processing speed, making it a practical solution for long-term video understanding.

For the future work, there are several potential aspects to further enhance the model's capabilities. First, replacing the existing image-based visual encoder with a video or clip-based encoder can naturally enhance the model's ability to capture short-term video dynamics. This provides a better representation of the video's temporal dynamics. Second, the model's overall performance in understanding videos can substantially benefit from the pre-training stage on large-scale video-text datasets. This approach is a common practice in existing research and has proven effective in enhancing generalization capabilities. Finally, the flexibility inherent in our model's architecture allows for the incorporation of a more advanced LLM as the language decoder. This integration offers a clear opportunity for boosting the final performance, making our model more effective in interpreting and responding to complex video content.

Table 11. Hyperparameters of different datasets on the long-term video understanding task.

Dataset	**LVU**	**Breakfast**	**COIN**
LLM		Vicuna-7B	
Epochs		20	
Learning rate		1e-4	
Batch size		64	
AdamW β		(0.9, 0.999)	
Weight decay		0.05	
Image resolution		224	
Beam size		5	
Frame length		100	
Memory bank length		20	
Prompt	"What is the {task} of the movie?"	"What type of breakfast is shown in the video?"	"What is the activity in the video?"

Table 12. Hyperparameters of different datasets on the video question answering task.

Dataset	**MSRVTT**	**MSVD**	**ActivityNet**
LLM	Vicuna-7B		
Epochs	5		
Learning rate	1e-4		
Batch size	128		
AdamW β	(0.9, 0.999)		
Weight decay	0.05		
Image resolution	224		
Beam size	5		
Frame length	20		
Memory bank length	10		
Prompt	"Question: {} Short Answer:"		

Table 13. Hyperparameters of different datasets on the video captioning task.

Dataset	**MSRVTT**	**MSVD**	**YouCook2**
LLM	Vicuna-7B		
Epochs	10		
Learning rate	1e-5	1e-5	1e-4
Batch size	128		
AdamW β	(0.9, 0.999)		
Weight decay	0.05		
Beam size	5		
Image resolution	224		
Frame length	80		
Memory bank length	40		
Prompt	"what does the video describe?"		

References

[1] 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. 1

[2] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 1, 2

[3] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 1, 2

[4] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 1

[5] 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. 2

[6] Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6387–6397, 2023. 2

[7] 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. 2

[8] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019. 2

[9] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2

[10] David L 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-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011. 2

[11] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2