Scaling the Long Video Understanding of Multimodal Large Language Models via Visual Memory Mechanism

Supplementary Material

7. Quantitative analysis

7.1. Comparison with Non-LLaVA Models

Tab. 7 evaluates the performance gains from incorporating FlexMem with more non-LLaVA models. The consistent improvements further highlight the effectiveness of FlexMem.

Table 7. Comparison with two non-LLaVA base models.

Method	LLM	LVBench	LongVideoBench	MLVU
Qwen2.5-VL	7B	45.3	56.0	70.2
+ FlexMem	**7B**	**52.3**	**64.3**	**75.0**
InternVL3.5	8B	43.8	62.0	70.6
+ FlexMem	**8B**	**51.6**	**65.7**	**71.9**

7.2. Impact of Layer Selection in MemIndex

Tab. 8 examines the effectiveness of our proposed optimization-based approach for the layer selection in MemIndex. Results show that optimization-based method achieves the best performance by identifying and selecting the most important cache layers for relevance computation.

Table 8. Ablation study on layer selection of memory indexing.

Layer Selection	LVBench	MLVU			
		Single	Multi	Holistic	M-avg
Eq. 10-based‡	**45.7**	**77.1**	**53.1**	**77.5**	**72.1**
Last three layers	43.8	76.6	52.0	77.1	71.5
Equal Interval (4,15,26)	45.0	76.8	52.5	77.3	71.7