Myopia Rectification: KV Cache Pruning for MLLMs via Dynamic Attention Subsidy and Token Reclamation

Supplementary Material

1. Details of datasets

In Table A3, we provide a comprehensive taxonomy of Milebench, detailing its task composition along with the corresponding number of samples and metrics for each task. The dataset encompasses 6,440 multimodal multi-image samples derived from 21 existing or self-constructed datasets, averaging 15.2 images and 422.3 words per sample. It is divided into two primary subsets: **Realistic Evaluation** and **Diagnostic Evaluation**.

Realistic Evaluation component challenges MLLMs to perform tasks within extensive multimodal contexts, highlighting the models' capabilities in understanding and reasoning across prolonged interactions.

Diagnostic Evaluation requires MLLMs to extract relevant information from the provided context, emphasizing their proficiency in long-range information retrieval and the ability to filter out distractors. Figure A1 shows a example in the dataset.

Instruction: Your objective is the main goal. Evaluate your current environment and your past decisions, and decide your immediate course of action.

Question: Put a cup with a fork in the sink. Step Details: <image#1>Step#1: Move to the sink to the right of the fridge. <image#2>Step#2: Pick up the fork from the sink. <image#3>Step#3: Turn around to the counter behind you. <image#4>Step#4: Place the fork in the cup on the counter. <image#5>Current Step:

GroundTruth: Pick up the cup from the counter.

Figure A1. An illustration of a complicated multi-image reasoning sample from MileBench. Such tasks impose requirements on the model's long contextual understanding and comprehensive spatial awareness abilities.

2. Analysis on visual attention

Figure A2 illustrates the attention scores of visual tokens at different positions within the sequence, where a higher value indicates greater importance and a higher retention probability. It can be observed that later images in the sequence are assigned relatively low scores across various layers. These images may be improperly discarded during the pruning process. We consider this an unfair phenomenon and a fundamental cause of attention myopia.

Table A1. Comparison between head-wise and head-aggregate.

Method	T1	T4	S3	NH	IR
Head-wise	**39.9**	39.8	13.2	5.3	**4.9**
Head-aggregate	39.7	**39.9**	13.2	**5.5**	4.8

Table A2. Image caption task with LLaVA-Next-7B model under 40% cache budget.

Method	COCO	Nocaps	Flickr30K
FastV	93.62	82.15	61.88
Ours (α=0.5)	**95.33**	**84.28**	**63.49**
Ours (α=0.2)	94.11	82.80	63.22
Ours (α=0.8)	94.69	83.42	63.01

3. Head-wise information

We try to assign different γ and α for each attention head, but it do not lead to significant accuracy improvements (as in Table. A1). This indicates that the average head mechanism is sufficient to represent most of the information in attention.

4. Generalization of γ and α.

For γ, we observe that attention scores decay with position, resembling an exponential change. The parameter γ, serving as compensation (1.0 to 2.5), allows the derivative of this exponential function to smooth out within a specific range. For α, it is used to adjust the contribution of the cross-attention scores between text and image. We believe that a centered value (0.4 to 0.6) can maximize the gains from text guidance. Table A2 demonstrates the validity of α in single image tasks.

Table A3. Details of MileBench.

Category	Task	Dataset	Count	Metric
		Realistic Evaluation		
Temporal Multi-image	**Action Understanding and Prediction (T-1)**	Action Localization	200	Accuracy
		Action Prediction	200	Accuracy
		Action Sequence	200	Accuracy
	Object and Scene Understanding (T-2)	Object Existence	200	Accuracy
		Object Interaction	200	Accuracy
		Moving Attribute	200	Accuracy
		Object Shuffle	200	Accuracy
	Visual Navigation and Spatial Localization (T-3)	Egocentric Navigation	200	Accuracy
		Moving Direction	200	Accuracy
	Counterfactual Reasoning and State Change (T-4)	Counterfactual Inference	200	Accuracy
		State Change	200	Accuracy
		Character Order	200	Accuracy
		Scene Transition	200	Accuracy
Semantic Multi-image	**Knowledge Grounded QA (S-1)**	Webpage QA	200	Accuracy
		Textbook QA	200	Accuracy
		Complex Multimodal QA	200	Accuracy
		Long Text with Images QA	200	Accuracy
	Text-Rich Images QA (S-2)	Slide QA	200	Accuracy
		OCR QA	200	Accuracy
		Document QA	200	Accuracy
	Visual Relation Inference (S-3)	Visual Change Captioning	400	ROUGE-L
		Visual Relationship Expressing	200	ROUGE-L
	Dialogue (S-4)	Multimodal Dialogue	200	Accuracy
		Conversational Embodied Dialogue	200	ROUGE-L
	Space Understanding (S-5)	Space Understanding	200	Accuracy
		Diagnostic Evaluation		
Needle In A Haystack	**Text Needle (N-1)**	Text Needle In A Haystack	320	Accuracy
	Image Needle (N-2)	Image Needle In A Haystack	320	Accuracy
Image Retrieval	**Image Retrieval (I-1)**	Image Retrieval	600	Accuracy

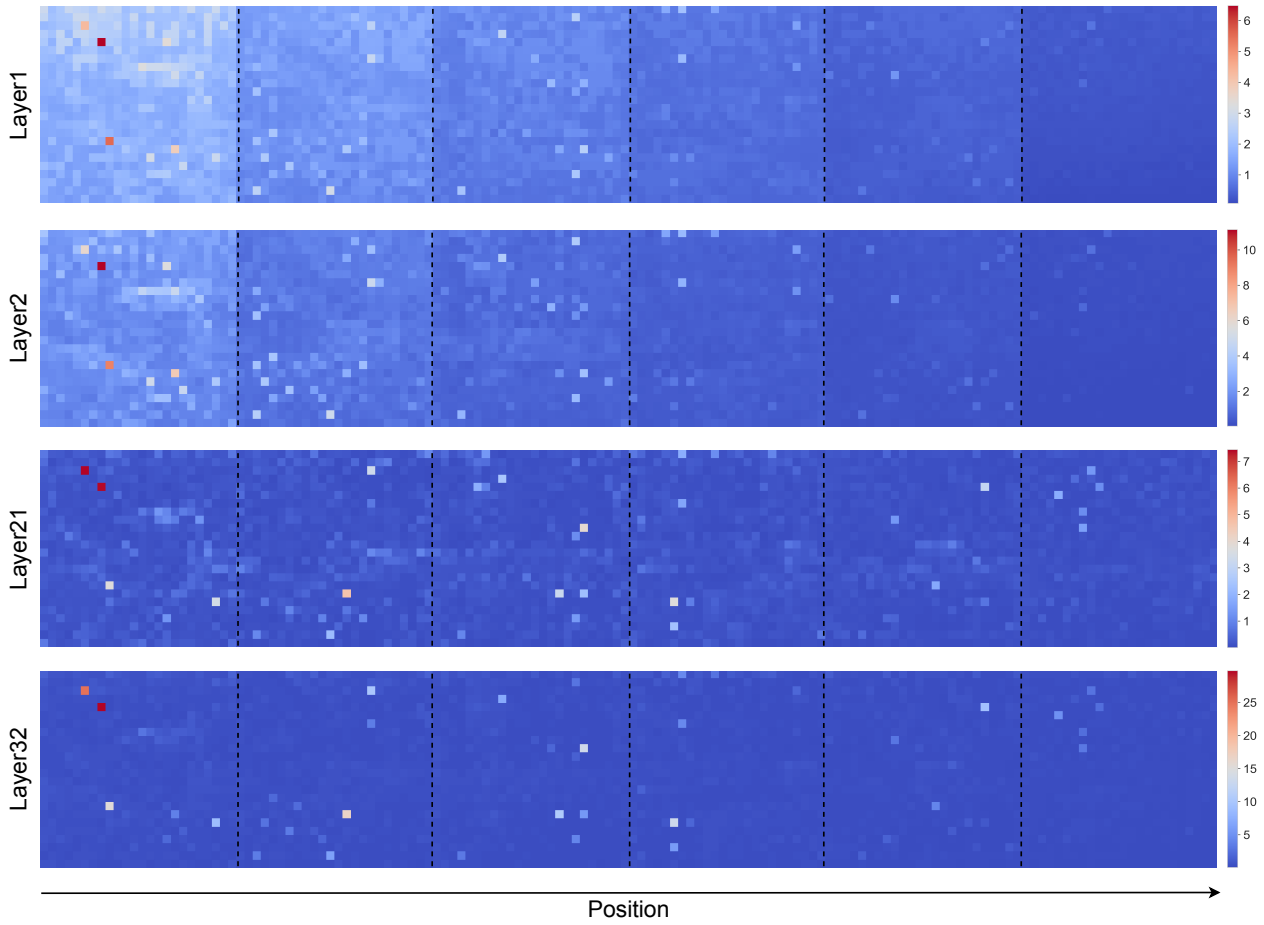

Figure A2. Visualization of attention score on each image.