AVAM: a Universal Training-free Adaptive Visual Anchoring Embedded into Multimodal Large Language Model for Multi-image Question Answering

Supplementary Material

A. Benchmark details

A.1. MuirBench

MuirBench is designed for multi-image understanding, comprising $11,264$ images and $2,600$ multiple-choice questions, with an average of 4.3 images per instance. MuirBench evaluates models across 12 key tasks, each representing 2.5% to 17.8% of the dataset:

- Geographic Understanding (GU): Reasoning over maps and geographic features (e.g., *"Among these map images, which one depicts overlapping geographic regions like* $< img1 >$*?"*).
- Counting (C): Quantifying specific objects across multiple images (e.g., *"How many vases have a painted design all over in the images?"*).
- Action Understanding (AU): Matching sequential images to actions (e.g., *"What is the action displayed in the video?"*).
- Visual Grounding (VG): Locating specific objects and extracting relevant information (e.g., *"This is the McDonald's my sister bought* $< img1 >$*. This is the McDonald's $1 $2 $3 Dollar Menu* $< img2 >$*. Could you please tell me how much my sister spent on this McDonald's?"*).
- Image-Text Matching (ITM): Associating text snippets with corresponding images (e.g., *"Which images has* 1 *apple and* 5 *bananas?"*).
- Ordering (O): Arranging images based on textual descriptions (e.g., *"The baby attempts to take off the clothes. What is the correct order of images according to the given context?"*).
- Scene Understanding (SU): Analyzing multi-view scenes from surveillance images (e.g., *"What's the color of the car parked behind the black van in the given images?"*).
- Difference Spotting (DS): Identifying differences between images (e.g., *"Can you determine which slide serves a different function compared to the others?"*).
- Cartoon Understanding (CU): Interpreting stories conveyed in cartoon images (e.g., *"What is the main content of this comic strip?"*).
- Diagram Understanding (DU): Extracting information from diagram images (e.g., *"Which object is below the bed?"*).
- Attribute Similarity (AS): Identifying a specific attribute across multiple images (e.g., *"Which of the following images shares the same scene with* $< img1 >$ *but contains the object potted plant?"*).
- Visual Retrieval (VR): Identifying images containing the same building (e.g., *"Can you find the images containing the same building as in* $< img1 >$*?"*).

A.2. MIBench

In MIBench, multi-image inputs are categorized into three scenarios: Multi-Image Instruction (MII), Multimodal Knowledge-Seeking (MKS), and Multimodal In-Context Learning (MIC).

- MII involves perception, comparison, and reasoning across multiple images (e.g., *"Do the two images show the same number of cats?"*).
- MKS assesses an MLLM's ability to retrieve relevant information from external knowledge provided in an interleaved image-text format. Unlike MII, MKS questions may focus on a single image or be independent of visual content.
- MIC evaluates MLLMs' ability to answer visual questions with the aid of multimodal demonstrations (i.e., examples).

A.2.1. Multi-Image Instruction (MII)

Based on the semantic types of instructions, MII is further divided into 5 tasks:

- General Comparison (GC): Evaluates the model's understanding of individual images, including aspects like scene, attributes, and location, and its ability to compare these images (e.g., *"Can the given sentence accurately illustrate what's in these two images? Two dogs are lying in the grass in each of the images."*).
- Subtle Difference (SD): Assesses fine-grained perception to detect minor differences between similar images (e.g., *"What are the differences between image 1 and image 2?"*).
- Visual Referring (VR): Tests the model's ability to understand object relationships based on referring expressions (e.g., *"Based on image 1, what is the relationship between image 2 and image 3?"*).
- Temporal Reasoning (TR): Measures comprehension of temporal relationships in consecutive images (e.g., *"What action do these images show?"*).
- Logical Reasoning (LR): Requires causal reasoning about objects or events depicted in images (e.g., *"Why did the boy in black extended his hands after the boy in white extended his hands?"*).

A.2.2. Multimodal Knowledge-Seeking (MKS)

Based on the form of external knowledge, MKS is divided into 4 tasks:

- Fine-grained Visual Recognition (FVR): Evaluates the model's ability to recognize objects in a query image using multiple reference images, requiring an understanding of image-label correspondence and similarity linking (e.g., *"Look at the dog pictures presented above and tell me which type of dog is represented in this image."*).
- Text-Rich Images (TRI) VQA: Assesses the model's ability to extract relevant information from text - heavy images, which represent a common real-world scenario involving tasks like reading slides and documents (e.g., *" What is the population of the country where the cabinet is named 'Kabinet Kerja'?"*).
- Vision-linked Textual Knowledge (VTK): Tests the model's ability to link query images with relevant external knowledge (e.g., Wikipedia) and extract useful information from corresponding text (e.g., *"Which city or region does this building locate in?"*).
- Text-linked Visual Knowledge (TVK): Evaluates the model's capability to answer text-only questions about the visual attributes of specific objects when given interleaved image-text knowledge (e.g., *"At the victory ceremony for Boxing at the 2018 Summer Youth Olympics how many medalists were holding their hand over their heart?"*).

A.2.3. Multimodal In-Context Learning (MIC)

In-context learning allows LLMs to improve performance when provided with a series of demonstrations. Recent studies introduce a more fine-grained assessment by dividing MIC into 4 distinct tasks:

- Close-ended VQA: Requires the model to select answers from a predefined set provided via multimodal demos, assessing its ability to learn image-label mappings.
- Open-ended VQA: Evaluates the model's ability to infer task patterns from demos when answers fall outside the predefined set.
- Hallucination Mitigation: Investigates the impact of MIC on hallucination.
- Demo-based Task Learning: Tests the model's ability to rapidly adapt to new tasks with few-shot demonstrations by removing explicit task instructions and presenting demos in a structured format (e.g., *"rabbit: 3"*).

A.3. Mantis-Instruct

Mantis-Instruct, the first multi-image instruction-tuning dataset, comprising 721K instances across 14 subsets, designed to cover all essential multi-image skills.

10 subsets are sourced from existing datasets:

- Reasoning: NLVR2, IconQA.
- Comparison: DreamSim, Birds-to-Words.
- Temporal Understanding: NExT-QA, STAR.

4 newly curated subsets:

- Coreference Resolution: LLaVA-665k-multi, LRV-multi.
- Expanded Reasoning: Contrast-Caption, Multi-VQA.

MLLM	Accuracy			
	Vanilla	Replaced visual inputs		
		1	2	3
Mantis-8B	28.6	62.8	74.2	98.9
Qwen-VL-9.6B	32.2	37.8	48.1	79.2

Table 5. Accuracy of MLLMs on the FVR task under different numbers of original visual input replacements, where "Vanilla" denotes 0 replaced visual inputs, and columns 1–3 represent 1, 2, and 3 replaced visual inputs, respectively.

To enhance instruction formatting, interleaving image placeholders are inserted into text based on various heuristics.

B. The impact of black image injection on distribution shift

We added redundant black blank images to the visual input to explore the impact of visual redundancy on MLLMs in the MVQA task. However, this impact might also stem from the distribution shift of visual input, i.e., the pure black images themselves could affect the output responses of MLLMs. Thus, we designed an exploratory experiment to demonstrate the true cause of the performance drop in Fig. 2(a). As shown in Tab. 5, we replaced the original visual inputs (except the correct answer) in the FVR task with 1-3 pure black images, respectively. Notably, the accuracy of MLLMs in question answering gradually increased as more visual inputs were substituted. This improvement occurs because the reduction of potentially confusing images, coupled with the MLLMs' ability to recognize these blank black images, facilitates the selection of the correct answers. This experiment serves to further validate the conclusion drawn in Sec 3.2.

C. Model details

- LLaVA-v1.5-7B [29]: CLIP ViT-L/14 [35] serves as the vision encoder and Vicuna-v1.5-7B [52] as the LLM.
- DeepSeek-vl-7B [33]: SAM-B [21] & SigLIP-L [47] serve as the vision encoder and DeepSeek-7B [5] as the LLM.
- Mantis-8B [20]: SigLIP SoViT-400M/14 * serves as the vision encoder and Llama3-8B [14] as the LLM.
- InternVL2-8B [11]: InternViT-300M-448px [12] serves as the vision encoder and InterLM2.5-7B [6] as the LLM.
- Qwen-VL-9.6B [4]: CLIP ViT-G/14 [35] serves as the vision encoder and Qwen-7B [3] as the LLM.
- Idefics2-8B [22]: SigLIP-L [47] serves as the vision encoder and Mistral-7B [19] as the LLM.

*https://huggingface.co/google/siglip-so400m-patch14-384

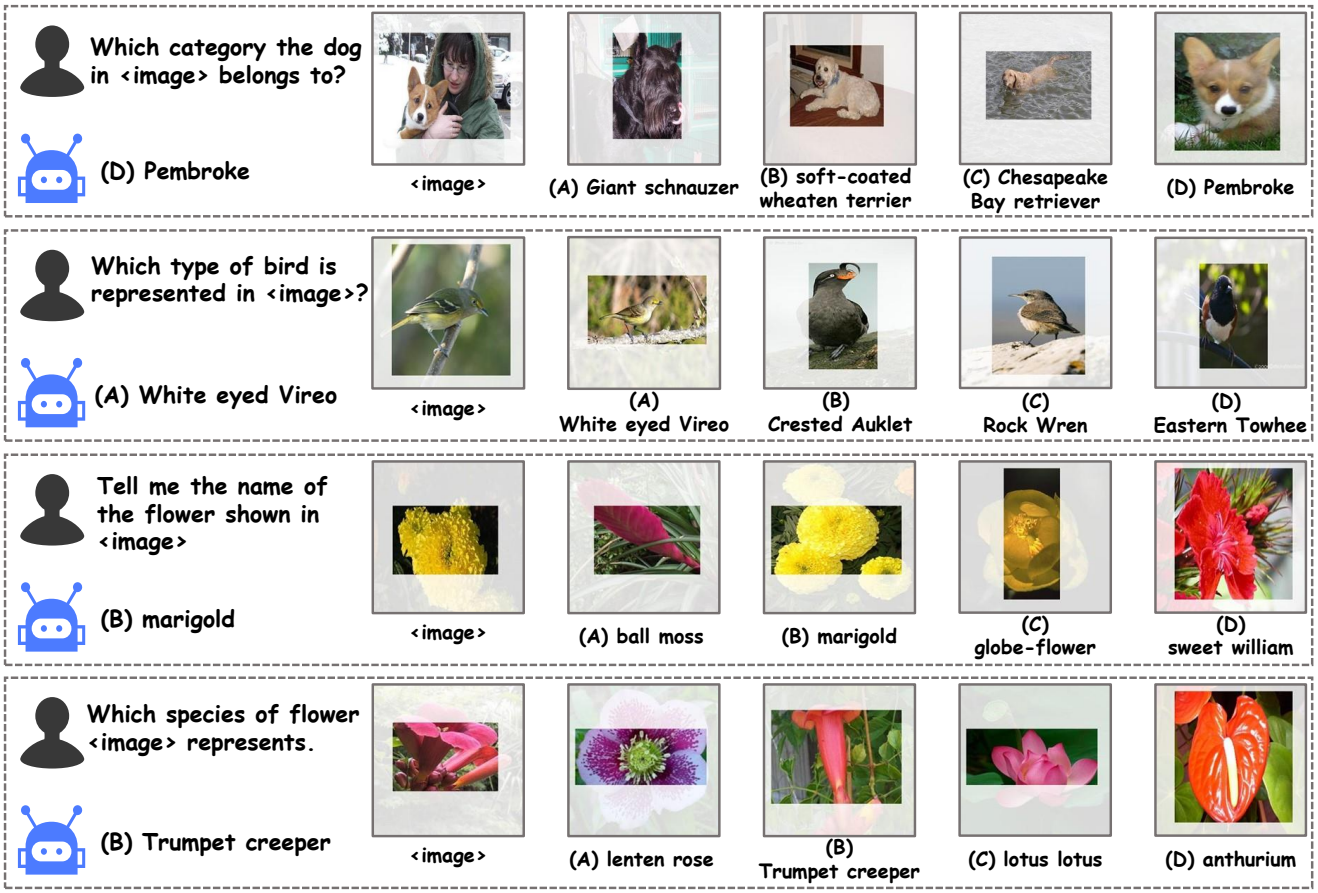

Figure 5. More Qualitative visualization results.

- LLaVA-OV-Qwen2-7B [23]: SigLIP SoViT-400M/14 * serves as the vision encoder and Qwen2-7B[39] as the LLM.

D. Ablation of text selection strategy

Strategy	FVR task of MIBench	
	Accuracy	Compression ratio
Vanilla	29.2	0.0%
Question-based	35.8	54.2%
Caption-based	**36.8**	57.6%

Table 6. Comparison of different text selection strategies. "Vanilla" denotes the original LLaVA-v1.5-7B, while "Question-based" and "Caption-based" represent the results of obtaining response maps using question texts and corresponding image captions, respectively.

Captions offer object-centric textual descriptions, which are typically more precise than question texts, and facilitate the accurate extraction of critical visual tokens. However, such captions are not always available. In the visual anchoring process, the response map is calculated exclusively from the image and question text when captions are absent. We compare caption-based and question-based strategies on the FVR task as shown in Tab. 6, where the latter replaces captions with questions to extract critical visual regions. Although question-based strategies use coarser text than captions, they effectively mitigate the submergence of critical visual tokens by visual redundancy. This approach does exhibit a slight performance decline, as question texts have less direct relevance to visual inputs than captions, leading to less accurate anchoring.