Divide, Conquer and Combine: A Training-Free Framework for High-Resolution Image Perception in Multimodal Large Language Models

Wenbin Wang[1*], Liang Ding[2*], Minyan Zeng[1], Xiabin Zhou[3], Li Shen[4], Yong Luo[1†], Wei Yu[1†], Dacheng Tao[5]

[1]School of Computer Science, National Engineering Research Center for Multimedia Software and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, China
[2]The University of Sydney, Australia
[3]Jiangsu University, China
[4]Shenzhen Campus of Sun Yat-sen University, China
[5]College of Computing and Data Science, Nanyang Technological University, Singapore
{wangwenbin97, luoyong, yuwei}@whu.edu.cn, {liangding.liam, minyanz977, xiabinzhou0625, mathshenli, dacheng.tao}@gmail.com

Abstract

Multimodal large language models (MLLMs) have experienced significant advancements recently, but still struggle to recognize and interpret intricate details in high-resolution (HR) images effectively. While state-of-the-art (SOTA) MLLMs claim to process images at 4K resolution, existing MLLM benchmarks only support up to 2K, leaving the capabilities of SOTA models on true HR images largely untested. Furthermore, existing methods for enhancing HR image perception in MLLMs rely on computationally expensive visual instruction tuning. To address these limitations, we introduce HR-Bench, the first deliberately designed benchmark to rigorously evaluate MLLM performance on 4K&8K images. Through extensive experiments, we demonstrate that while downsampling HR images leads to vision information loss, leveraging complementary modalities, *e.g.,* text, can effectively compensate for this loss. Building upon this insight, we propose Divide, Conquer and Combine, a novel training-free framework for enhancing MLLM perception of HR images. Our method follows a three-staged approach: 1) Divide: recursively partitioning the HR image into patches and merging similar patches to minimize computational overhead, 2) Conquer: leveraging the MLLM to generate accurate textual descriptions for each image patch, and 3) Combine: utilizing the generated text descriptions to enhance the MLLM's understanding of the overall HR image. Extensive experiments show that: 1) the SOTA MLLM achieves 63% accuracy, which is markedly lower than the 87% accuracy achieved by humans on HR-Bench; 2) our method brings consistent and significant improvements (a relative increase of +6% on HR-Bench and +8% on general multimodal benchmarks).

Code — https://github.com/DreamMr/HR-Bench

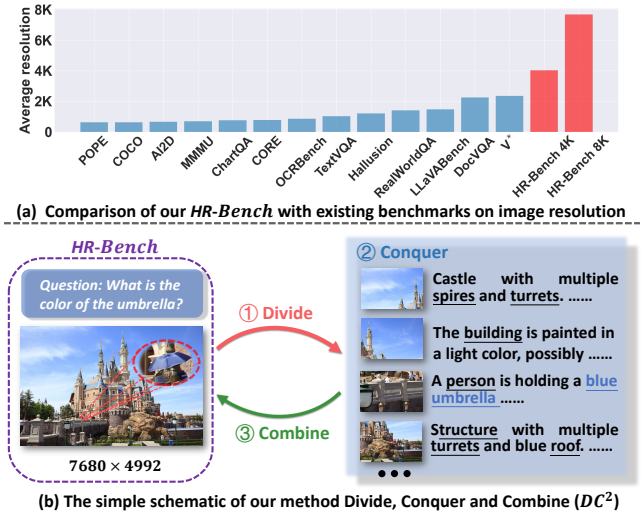

Figure 1: An overview of our work. (a) Existing multimodal benchmarks is only 2K resolution. To fill the gap, we introduce ***HR-Bench***, designed to evaluate MLLMs on high-resolution (HR) images up to 8K. (b) We propose a training-free framework – **D**ivide, **C**onquer and **C**ombine (**DC**^{2}), which recursively uses image patches to provide relevant text descriptions, helping existing MLLMs better perceive HR images.

Introduction

Recent advancements in multimodal LLMs (MLLMs) (Liu et al. 2024a; Dong et al. 2024a) have greatly enhanced their abilities in vision-language understanding, reasoning, and interaction (Xu et al. 2024). This progress is primarily due to the integration of visual signals into Large Language Models (LLMs), allowing them to perceive the world visually. A key component of this process is the visual encoding strategy. However, most current MLLMs (Liu et al. 2024a; Bai et al. 2023b) perceive images in a fixed resolution (*e.g.,* 336×336). This simplification often results in significant shape distortion and blurring of high-resolution (HR) image content, which hurts the performance of MLLMs. Given that real-world im-

*These authors contributed equally.
†Corresponding authors.

ages vary widely in resolution, this limitation poses substantial challenges for MLLMs across various applications (Tian et al. 2022, 2023; Wang et al. 2024).

To address this issue, recent studies improve MLLM's perceptual ability for HR image by carefully designing various strategies, which can be categorized into three types: 1) cropping-based methods (Chen et al. 2024c; Liu et al. 2024b; Li et al. 2024b), 2) HR visual encoder (Luo et al. 2024; Ge et al. 2024; Lu et al. 2024), and 3) visual search (Wu and Xie 2024). Although many advanced strategies have been proposed to enhance MLLM's perceptual ability for HR image, the current benchmark resolution is only up to 2K, as illustrated in Figure 1 (a). Meanwhile, the most advanced MLLMs are now capable of handling 4K HR images (Chen et al. 2024c; Dong et al. 2024b). This implies that **state-of-the-art (SOTA) MLLMs have not yet undergone rigorous validation on HR images**. Therefore, higher resolution benchmarks are needed in this field.

Firstly, to tackle the current lack of HR multimodal benchmarks, we introduce ***HR-Bench***. This benchmark is designed to evaluate the ability of MLLMs to perceive HR images. ***HR-Bench*** is available in two versions: ***HR-Bench 8K*** and ***HR-Bench 4K***. The ***HR-Bench 8K*** includes images with an average resolution of 8K, sourced from the open-source 8K resolution image dataset DIV8K (Gu et al. 2019) and Internet, with our manually annotated questions and answers. For ***HR-Bench 4K***, we manually annotate the coordinates of objects relevant to the questions within the 8K image and crop these images to 4K resolution. This benchmark aims to systematically evaluate the ability of MLLM to perceive HR images, thus paving the way for future research.

Secondly, we conduct a series of experiments on the ***HR-Bench*** to explore the effects of image resolution on MLLMs. We select SOTA MLLMs (Chen et al. 2024c; Liu et al. 2024a; Bai et al. 2023b) to evaluate their performance across varying image resolutions (*e.g.,* 1K, 2K and 8K resolution). The experimental results indicate that downsampling HR images to a lower, fixed resolution leads to a significant *loss of visual information*. This degradation increases the uncertainty in the model's output, making them more prone to errors. *Notably, the integration of information from other modalities (e.g., text), proves effective in mitigating the adverse effects of lost visual information.*

Finally, we combine what we have learned above to design a new training-free framework which we call ①**D**ivide, ②**C**onquer and ③**C**ombine (**DC**^{2}). Our **DC**^{2} processes HR images by breaking them down into smaller, manageable image patches and using their accurate text descriptions to enhance MLLMs perception, as shown in Figure 1 (b). Specifically, ① we divide an HR image into smaller patches recursively until they match the resolution of the pretrained visual encoder (*e.g.,* 336 $\times$ 336). To enhance computational efficiency, similar patches are merged. In the conquer stage②, we use MLLM to generate text descriptions for each image patch. During the combine stage③, we aggregate the text descriptions and filter out hallucinations caused by the dividing stage. Directly using all text descriptions can hurt performance due to excessive input length. Inspired by Wu and Xie (2024), we introduce a visual memory $\mathcal{M}$ to store objects which appear in the text description, and coordinates of image patches. In the inference stage, we use the user prompt to interact with $\mathcal{M}$, enabling MLLM to generate more precise text descriptions. Experiments demonstrate that our **DC**^{2} significantly improves performance on HR image benchmarks and outperforms existing methods on general multimodal benchmarks.

Our contributions are summarized as follows:

- We introduce ***HR-Bench*** to systematically evaluate the perception ability of MLLMs in HR images. To the best of our knowledge, we are the first to propose an 8K image resolution benchmark for MLLMs.
- Based on our ***HR-Bench***, we explore the impact of image resolution on MLLMs. we find that downsampling HR images reduces visual information, increasing uncertainty and errors in model outputs. Fortunately, adding proper textual information can effectively restore these lost information.
- Given our observation, we propose a training-free framework **DC**^{2} to effectively enhance the MLLM's perceive ability on HR images. Experimental results on our ***HR-Bench*** and general multimodal benchmarks using several advanced MLLMs, show that our approach brings consistent and significant improvements (up to +12.0% accuracy).

Preliminaries and Related Work

MLLMs generally include a **Visual Encoder** (Radford et al. 2021) for extracting visual features and a **Large Language Model (LLM)** (Touvron et al. 2023a,b; Bai et al. 2023a; Cai et al. 2024; GLM et al. 2024) for decoding text sequences. Both the visual encoder and LLM are usually initialized from pre-trained models. The vision and language modalities can be connected by **Multimodal Connector** (*e.g.,* MLP). MLLMs generate sentences in an auto-regressive manner, predicting the probability distribution of the next token progressively. To maintain consistency with the image resolution used during visual encoder pre-training, MLLMs typically resize the image to a fixed resolution (*e.g.,* 336 $\times$ 336 in LLaVA) before extracting visual features through the visual encoder. However, this simplification often results in significant shape distortion and blurring of HR image content. To address this issue, current solutions can be divided into **1) cropping-based methods**, **2) incorporating HR visual encoder methods**, and **3) visual search methods**.

Cropping-Based Methods. The representative cropping-based methods for HR MLLMs are introduced in LLaVA-NeXT (Liu et al. 2024b) and InternVL-v1.5 (Chen et al. 2024c), which partition an image into several patches, each encoded separately by ViT (Dosovitskiy et al. 2021) and subsequently concatenated for LLM processing. Several methods have adopted cropping to scale up resolution (Chen et al. 2024a; Zhang et al. 2024; Liu et al. 2024c).

HR Visual Encoder. Incorporating a HR visual encoder for HR image understanding does not substantially increase the number of visual tokens. Vary (Wei et al. 2023) and Deepseek-VL (Lu et al. 2024) harness SAM (Kirillov et al.

2023) as a HR visual encoder to boost ViT's capabilities. Similarly, MiniGemini-HD (Li et al. 2024a), LLaVA-HR (Luo et al. 2024) and ConvLLaVA (Ge et al. 2024) leverage ConvNeXt (Liu et al. 2022) to handle HR images, utilizing cross-attention or adapter to extract visual features.

Visual Search. Inspired by key elements in the human visual search process, Wu and Xie (2024) introduce SEAL, a meta-architecture for MLLMs. SEAL is designed to actively reason about and seek out necessary visual information, a crucial capability for vision-intensive multimodal tasks, especially when dealing with HR images.

Despite numerous advanced strategies proposed to enhance MLLMs' perceptual ability for HR images, the image resolution of current benchmarks (Wu and Xie 2024; Mathew, Karatzas, and Jawahar 2021; Masry et al. 2022; Yue et al. 2024; Kembhavi et al. 2016; Yifan et al. 2023; Yuan et al. 2023; Fu et al. 2023; Yu et al. 2024; Han et al. 2023; Chen et al. 2024b) remains capped at 2K. In contrast, the latest MLLMs now handle 4K HR images (Chen et al. 2024c; Dong et al. 2024b). This discrepancy indicates a pressing need for higher resolution MLLM benchmarks. Additionally, the factors influencing perceptual ability of MLLMs for HR images have not been thoroughly investigated.

How does Image Resolution Affect MLLMs?

HR-Bench

To systematically evaluate the effect of image resolution on MLLMs, we need a benchmark with sufficiently high resolution. We analyze the image resolution of 21 commonly used multimodal benchmarks and find that the benchmark with the highest image resolution is $\mathbf{V}^*$, which is only 2246×1582. The average resolution of the 21 multimodal benchmarks is 530×518. However, it is known that the current SOTA MLLMs (Chen et al. 2024c; Dong et al. 2024b) are capable of handling 4K images. Thus, we introduce the ***HR-Bench*** using 200 8K resolution images from DIV8K (Gu et al. 2019) and the Internet. Our ***HR-Bench*** has significantly high resolution than other MLLM benchmarks – $4\times$ more than $\mathbf{V}^*$.

HR-Bench* Curation.** ***HR-Bench consists two sub-tasks: ***Fine-grained Single-instance Perception (FSP)*** and ***Fine-grained Cross-instance Perception (FCP)***. The ***FSP*** task includes 100 samples, challenging the MLLM to identify specific attributes such as color and material of an object. Similarity, the ***FCP*** task also comprises 100 samples but focuses on assessing the MLLM's ability to determine the relative positions between objects in an image. Both the images and questions are meticulously selected and crafted by human annotators to ensure it is challenging to "guess" the correct answer without accurately grounding the relevant objects in the image. In addition, the 8K images are cropped around the objects in question to produce 4K images. For clarity, 8K resolution images are termed ***HR-Bench 8K***, while the 4K resolution images are referred to as ***HR-Bench 4K***.

Evaluation of Protocol. To quantitatively compare MLLMs on our ***HR-Bench***, we create multiple choice options for each question. Recognizing that MLLMs can be sensitive to the order of options in multiple choice questions, we use a more robust evaluation strategy called **Cyclic Permutation** (Zheng et al. 2023). In particular, each question is presented to an MLLM N times, with N being the number of choices. Each time, the order of options are rotated to form a new prompt for the MLLMs. After completing N passes, we calculate and report the average accuracy, ensuring a more reliable assessment.

Pilot Experiments

Despite the numerous advanced strategies proposed to handle HR images (Chen et al. 2024c; Dong et al. 2024b), the impact of image resolution on MLLMs remains underexplored. Here, we raise two questions:

- ***How does image resolution affect MLLMs?***
- ***How can we use answer to the above question to improve on prior methods?***

To answer these questions, we perform experiment on the existing SOTA MLLMs, selecting four widely used models: LLaVA-v1.5 7B & 13B (Liu et al. 2024a), Qwen-VL-Chat (Bai et al. 2023b) and InternVL-v1.5-Chat (Chen et al. 2024c). These models cover various dimensions, including model scales, types of multimodal connectors, and types of visual encoders, enabling a more comprehensive analysis.

How does image resolution affect MLLMs? We conduct experiments on our ***HR-Bench***. We manually annotate the coordinates of relevant objects in each sample, and then crop the images centered on these coordinates to obtain images with different resolutions. During the cropping process, we maintain the original image aspect ratio. We use the following metrics to assess the impact of resolution for MLLMs: (1) **Accuracy** and (2) **Uncertainty Score**, which measures the model's confidence in generating the next token (Zhou et al. 2024). A higher uncertainty score indicates greater uncertainty about the output, suggesting that the model's outputs are more likely to be inaccurate.

Figure 2 (a) and (b) illustrate that, across all models, **the accuracy significantly decreases and uncertainty score increases as the image resolution grows.** This can be intuitively explained by the significant loss of visual information during the downsampling of HR images. To prove this, Figure 2 (c) shows examples of images resized from various resolutions to 336. We observe that resizing from HR to low-resolution results in blurriness and a loss of detailed visual information.

Finding 1: Downsampling higher-resolution images to a fixed resolution leads to greater visual information loss, increasing model output uncertainty, thereby causing output errors.

Can we use language modality information to compensate for the missing visual information? To answer this question, we design two experiments: 1) We manually provide rich text descriptions of the images in our ***HR-Bench 8K***, detailing the attributes of the objects (*e.g.,* color) and the relative positions between them in the image ("**T**"). We **DO NOT** directly provide the answer to the question. 2) We extract the key image region, which the MLLM can rely on to

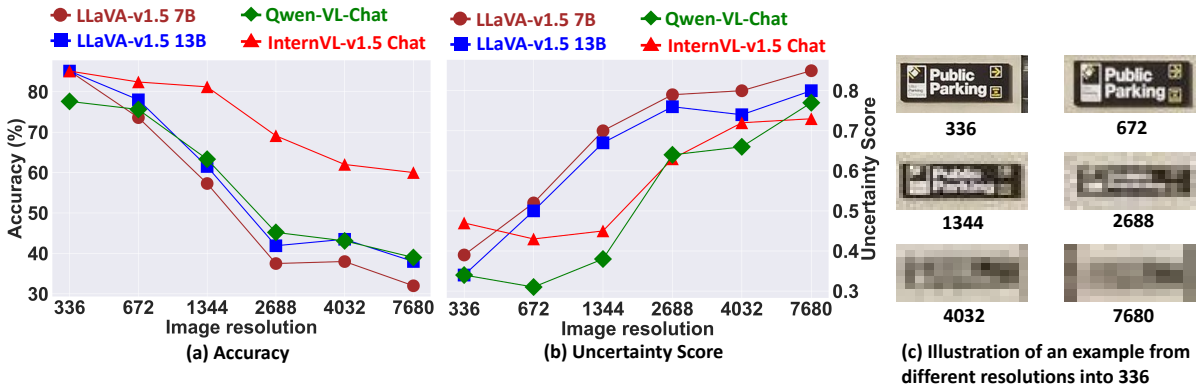

Figure 2: Experimental results for accuracy and uncertainty scores under different image resolutions. We illustrate the accuracy (a) and uncertainty score (b) on four models with different image resolutions. Additionally, we visualize an example that is resized from different resolutions into 336 (c).

generate correct answers, and replace the HR input to prevent visual information loss during downsampling ("**P**"). As shown in Figure 3, we find that 1) by introducing rich text descriptions, the performance is significantly improved on our ***HR-Bench 8K***; and 2) incorporating text descriptions can achieve performance comparable to preserving key regions of the image.

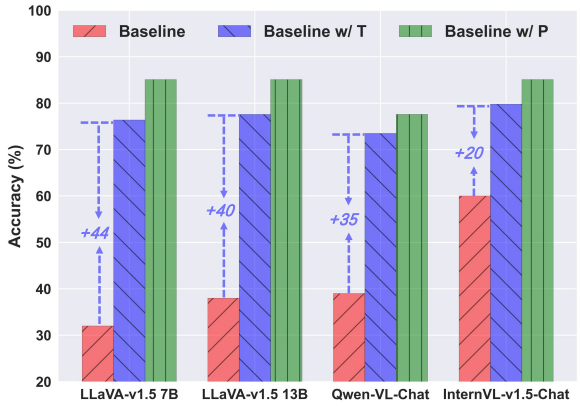

Figure 3: The effect of incorporating rich text description on model performance. "**T**" represents text descriptions. "**P**" represents key image regions.

Finding 2: The visual information loss due to downsampling in HR images can be compensated by relevant textual information.

Methodology

Overviews. Based on the aforementioned findings, we propose a novel training-free framework — ① **D**ivide, ② **C**onquer and ③ **C**ombine (**DC**^{2}) (see Figure 4). *The design principle of our method is to use the accurate text descriptions of image patches to help MLLM better perceive HR image.* To achieve this, we first recursively split an image into image patches until they reach the resolution defined by the pretrained vision encoder (*e.g.,* 336×336), merging similar patches for efficiency (**Divide**). Next, we utilize MLLM to generate text description for each image patch and extract objects mentioned in the text descriptions (**Conquer**). Finally, we filter out hallucinated objects resulting from image division and store the coordinates of the image patches which objects appear (**Combine**). During the inference stage, we retrieve the related image patches according to the user prompt to provide accurate text descriptions.

Dividing: Image Division

The goal of the **Divide** stage is to decompose the image into the resolution defined by pretrained visual encoder (*e.g.,* 336×336), avoiding excessive visual information loss due to downsampling. However, we find that decomposing HR images into an excessive number of image patches disrupts object integrity, hindering the acquisition of global image information. Inspired by CNN (LeCun et al. 1989; He et al. 2016), we recursively decompose the image, dividing it into four equal parts until the resolution defined by pretrained vision encoder is reached, thereby reducing the loss of global information. As shown in Figure 4, the entire process can be visualized as a tree-like structure.

Specifically, given an image v_l, we crop v_l into four patches. Mathematically, this operation can be described as:

$$\{\overline{v}^i\}_{i=1}^4 = \mathcal{F}_{crop}(v_l), \tag{1}$$

where l represents the indices of the current recursive layer and i represents the $i-$th image patch. $\mathcal{F}_{crop}(\cdot)$ is the cropping function used to split the image into four image patches.

However, it is not efficient to perform recursion for each image patch. In fact, visual signals have high redundancy (Bolya et al. 2023). To optimize computational efficiency, we merge (*i.e.,* by averaging) the image patches with similarity greater than θ by performing hierarchical clustering (HC) on the image patches. This can be formulated as follows:

$$[C_1, ..., C_k] = HC(\{\overline{v}^i\}_{i=1}^4, \theta), \tag{2}$$

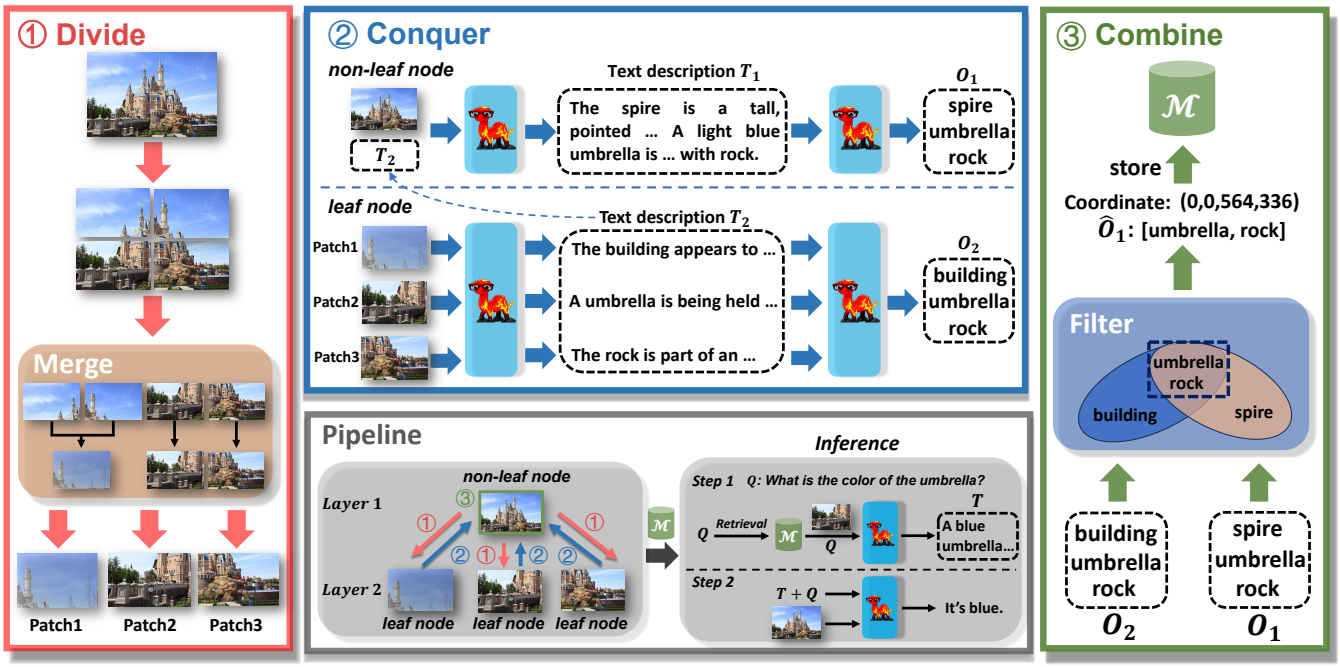

Figure 4: Detailed illustration of our proposed schema **DC**^{2} with a running example. ① We divide the image into four image patches and then merge the patches that have a high degree of similarity. ② We use MLLMs to generate text descriptions and object information from image. ③ We filter out uncertainty objects and then store the coordinates of the actually existing objects.

$$v_{l+1}^{i} = \frac{1}{|C_i|} \sum_{\overline{v} \in C_i} \overline{v}, \tag{3}$$

where the k represents the number of clusters and C_i represents the $i-$th cluster. Thus, the output of **Divide** stage is $\{v_{l+1}^{i}\}_{i=1}^{k}$ in current recursive layer. Then, the v_{l+1}^{i} serves as the input to the next recursive layer.

Conquering: Local Image Perception

In the **Conquer** stage, image patches obtained in dividing process are used to generate text description and extract objects information by the MLLM. Specifically, given an image v_l, we firstly utilize MLLM to generate text description T_l. Then, we identify the main objects O_l mentioned in the generated text description T_l. We denote the image patch which does not branch out to any other image patches as leaf node, while others are called non-leaf nodes. For a leaf node, we directly use MLLM to generate text description T_l. For non-leaf node, we concatenate the text descriptions from image patches $\{v_{l+1}^{i}\}_{i=1}^{k}$ (*i.e.*, T_{l+1}) to generate text description for image v_l. This is formulated as follows:

$$\begin{cases} T_l, O_l = \mathcal{F}_{leaf}(v_l) & \text{if } v_l \text{ is a leaf node} \\ T_l, O_l = \mathcal{F}_{non-leaf}(v_l, T_{l+1}) & \text{otherwise} \end{cases}, \tag{4}$$

where the $\mathcal{F}_{leaf}(\cdot)$ is used to generate the text description T_l and extract objects O_l for leaf nodes while $\mathcal{F}_{non-leaf}(\cdot)$ is used for non-leaf nodes.

Combining: Global Fusion

In the **Combine** stage, we aggregate the information from image patches. Actually, image division disrupts the integrity of objects leading to output with object hallucination. Therefore, we need to **filter** out object hallucination caused by image division. Additionally, using text descriptions of all image patches can result in excessively long input text, which hurt performance during inference. Inspired by Wu and Xie (2024), we introduce visual memory $\mathcal{M}$. We obtain the coordinates of the image patch where each object is located and **store** them in visual memory $\mathcal{M}$. During inference, we retrieve the image patches containing the objects mentioned in the user prompt and generate text descriptions. We use (x, y, w, h) to represent the coordinate of image patch (*i.e.*, bounding box). The x and y represent the coordinate of the left and top in the global image. The w and h represent the width and height of the image patch respectively.

Filter. One straightforward approach is to calculate the uncertainty score (Zhou et al. 2024) by the probability of autoregressive decoding for each object. However, this method tends to be inefficient for filtering out hallucinated objects. Indeed, for a real existing object, it will be found by the MLLM in successive recursive layers. Based on this, we take the intersection of O_l and O_{l+1}, considering the objects $\widehat{O}_l$ in both sets to be actually existing. This can be formulated as follows:

$$\widehat{O}_l = O_l \cap O_{l+1}. \tag{5}$$

Storing in the Visual Memory $\mathcal{M}$. After obtaining the actually existing objects $\widehat{O}_l$ and the coordinates of image patches, we store them in the visual memory $\mathcal{M}$. Two issues arise during storage: 1) overlapping image patches for the same object, and 2) coordinate representation of merged image patches. To address overlapping image patches, we apply Non-Maximum Suppression (NMS) to retain the patch that best represents the object. For coordinate representation, we save the coordinates before merging the image patches.

Inference Details

In the inference stage, we utilize the user prompt to interact with visual memory $\mathcal{M}$. Specifically, given a user prompt Q, we use a textual retriever (Izacard et al. 2022) to retrieve related objects with confidence levels exceed α. Subsequently, we obtain the image patches for the retrieved objects to allow MLLM to generate accurate text descriptions T. Finally, we concatenate the accurate text descriptions T with user prompt Q and utilize the MLLM to generate the final response.

Experiments

Evaluation on *HR-Bench 8K*

Overall performance. As shown in Table 1, the most proficient open-source MLLM, InternVL2-llama3-76B (Chen et al. 2023), achieves accuracy of 61.4% on ***HR-Bench 8K***. Even the most advanced models, Gemini 1.5 Flash (Reid et al. 2024), GPT4o (Achiam et al. 2023), QWen-VL-max (Bai et al. 2023b) achieve accuracies of 62.8%, 55.5% and 52.5% on ***HR-Bench 8K***. The results demonstrate that existing MLLMs still have a significant gap compared to humans in their perception of HR images.

Our DC^2 brings consistent improvements on *HR-Bench 8K*. We observe that our **DC**^{2} achieves consistent and significant improvement across four models and two sub-tasks. Our **DC**^{2} brings a maximum of 5.7% and 3.0% accuracy improvement on ***FSP*** and ***FCP*** respectively. Additionally, InternVL-v1.5 with our **DC**^{2} surpasses the current SOTA Gemini 1.5 Flash in the ***FSP*** sub-task, achieving an accuracy of 75.0%. The results show that our method has a clear advantage with HR images.

Method	*HR-Bench 8K*		
	***FSP*↑**	***FCP*↑**	***Avg.*↑**
Human	**94.0**	**79.5**	**86.8**
Random Guess	25.0	25.0	25.0
Open-source MLLMs			
InternVL-2-llama3-76B	69.0	**53.8**	**61.4**
InternVL-1.5-26B	**69.3**	46.5	57.9
Xcomposer2-4kHD-7B	55.3	47.3	51.3
LLaVA-1.6-34B	44.5	50.3	47.4
LLaVA-HR-X-13B	49.5	44.3	46.9
Commercial chatbot systems			
Gemini 1.5 Flash	**69.2**	**56.7**	**62.8**
GPT4o	62.0	49.0	55.5
QWen-VL-max	54.0	51.0	52.5
w/ our DC^2			
InternVL-1.5 26B	**75.0**	**47.5**	**61.3**
$\Delta(\uparrow)$	+5.7	+1.0	+3.4
LLaVA-v1.6 7B	40.5	45.0	42.3
$\Delta(\uparrow)$	+3.3	+0.8	+2.1
LLaVA-v1.5 13B	40.0	41.0	40.5
$\Delta(\uparrow)$	+2.5	+3.0	+2.7
Yi-VL 6B	39.0	41.0	40.0
$\Delta(\uparrow)$	+0.5	+1.7	+1.1

Table 1: Results of different models on ***HR-Bench 8K***. The best performance in each task is in-bold. The "$\Delta(\uparrow)$" represents the performance gains of our **DC**^{2} against the baselines. Due to space limitations, only the results of the top 5 open-source MLLMs and the top 3 commercial chatbot systems are presented here.

General Multimodal Benchmarks Evaluation

To verify that our **DC**^{2} is not only applicable to HR images, we also conduct experiments on general MLLM benchmarks. As shown in Table 2, our **DC**^{2} not only brings up to a 12% improvement in accuracy on 2K resolution MLLM benchmark ***V**** (Wu and Xie 2024) but also shows significant improvements in object hallucination evaluation POPE (Yifan et al. 2023) and comprehensive multimodal benchmark MME (Fu et al. 2023).

Ablation Study

In Table 3, we explore different modules, including merge, retrieval, filter, visual memory and recursive crop. The merge module causes a minor 0.5% performance drop but enhances inference efficiency by reducing image patches. Excluding visual memory necessitates generating text descriptions for image patches during inference, leading to longer inputs and a 2.6% performance drop. Removing the filter compromises object integrity in images, causing incorrect text descriptions and a 4.6% performance decrease. Omitting recursive cropping severely impacts object integrity and increases input text length, resulting in a substantial 10.2% performance decline.

Trade-off Between Performance And Computational Cost

Researchers may have concerns regarding the efficiency of **DC**^{2}. To address this, we present Table 4, which illustrates the relationship between throughput and accuracy under various θ values (used to merge image patches), comparing these results with the SOTA method visual search (Wu and Xie 2024). As depicted, a decrease in θ leads to a continuous increase in accuracy on the ***HR-Bench 8K***, accompanied by a decrease in efficiency. Notably, **DC**^{2} achieves higher accuracy than visual search for equivalent throughput, indicating a reasonable trade-off between performance and efficiency.

When and Why Does Our Method Work?

Reviewing the design principles of **DC**^{2}: using text descriptions of image patches to help MLLM better perceive HR image. To explore the underlying mechanism of **DC**^{2}, we perform experiments that help address the following questions:

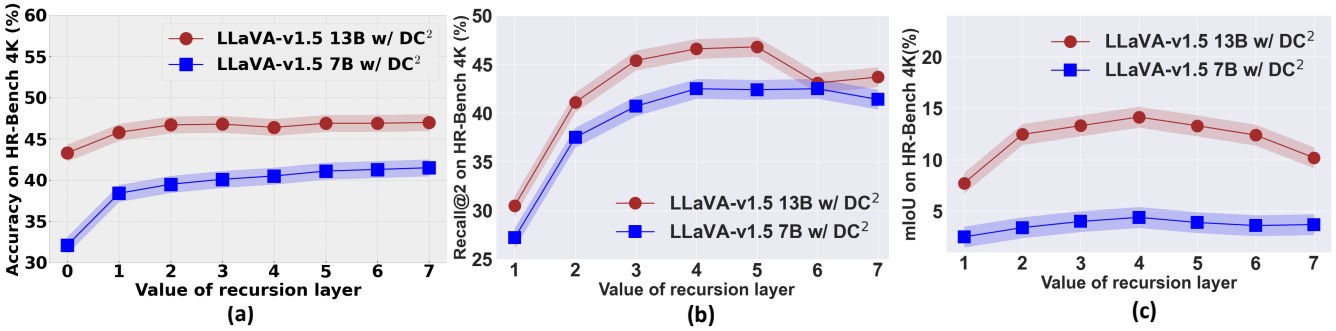

Figure 5: Effect of recursion layers on ***HR-Bench 4K***. (a) Overall performance, (b) Recall@2, (c) mIoU scores.

Method	V^* ↑	*POPE*↑	*MME*↑
Yi-VL-6B	40.9	83.1	1902.7
+DC^2	**46.2**	**83.2**	**1918.2**
LLaVA-v1.5-7B	46.2	85.6	1755.9
+DC^2	**57.3**	**86.8**	**1778.7**
LLaVA-v1.5 13B	42.7	85.5	1773.6
+DC^2	**54.7**	**86.5**	**1779.1**

Table 2: Evaluation on broader range of general multimoadal benchmarks. **DC^2** can also bring significant improvements on general multimodal benchmarks. The results are measured by VLMEVALKIT (Duan et al. 2024).

Method	V^*↑	*HR-Bench 8K*↑	*Avg.*↑
DC^2	57.3	**39.5**	48.4
w/o merge	**58.3**	39.5	**48.9**
w/o visual memory	55.6	36.0	45.8
w/o filter	52.6	35.0	43.8
w/o recursive crop	43.9	32.5	38.2

Table 3: Ablation studies of our **DC^2**. We conduct experiments on V^* and ***HR-Bench 8K*** using LLaVA-v1.5 7B.

1. Does increasing the number of image patches improve performance? We illustrate the relationship between the number of recursion layers and accuracy on ***HR-Bench 4K*** using LLaVA-v1.5 7B & 13B. Figure 5 (a) shows that 1) increasing the number of layers significantly improves accuracy; 2) however, as the recursion layers increase, the performance improvement gradually slows down. We observe that as the recursion layers increase, the performance of ***FSP*** improves significantly, but ***FCP*** appears to even slightly decline. ***More image patches reduce visual information loss, benefiting the FSP task.***

2. Can DC^2 provide precise text descriptions to compensate for the absence of visual information? To demonstrate that our **DC^2** can provide precise information about objects, we employ the widely used evaluation metric Recall@2 to assess the performance of retrieving pertinent objects from visual memory $\mathcal{M}$. Additionally, we utilize mIoU to provide a more precise quantification of the overlap between the predicted bounding boxes derived from our **DC^2** and ground truth. As shown in Figure 5, the results show that 1) within a reasonable range of recursion layers, increasing the depth of recursion layers can yield more accurate object location information. 2) The more accurate the object location information provided by MLLM, the higher the accuracy on ***HR-Bench 4K***. ***DC^2 can determine the position of objects in the image, thereby providing more accurate text description to compensate for missing visual information.***

Method	*Throughput*↑	*Acc.*↑
DC^2 *w/o merge*	2.8	**39.5**
DC^2($\theta = 0.1$)	3.1	**39.5**
DC^2($\theta = 0.2$)	4.6	36.5
DC^2($\theta = 0.3$)	**5.0**	35.5
Visual Search	4.6	35.6

Table 4: Performance and inference efficiency. We illustrate the correlation between throughput (samples per minute) and the accuracy of the LLaVA-v1.5 7B enhanced with the proposed **DC^2** across varying θ values on the ***HR-Bench 8K***. Additionally, we also compare with SOTA method Visual Search, which is also used for HR images.

Conclusion

In this paper, we propose an 8K image resolution benchmark, namely ***HR-Bench*** and introduce a training-free framework — **D**ivide, **C**onquer and **C**ombine (**DC^2**). We systematically evaluate open-source and commercial models on ***HR-Bench***. From the results, we mainly conclude that: (1) MLLMs currently fall significantly short of humans in perceiving HR images; (2) current MLLMs lose a significant amount of visual information when resizing HR images to low-resolution, but this loss can be compensated for with text information; (3) our **DC^2** improves the current MLLMs' ability to perceive HR images. In the future, we will explore advanced token compression technologies, such as token merging-based methods for more efficient processing of images at any resolution, which could further enhance the MLLM's ability for high-resolution perception.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China (No. 2022YFF0712300), the National Natural Science Foundation of China (Grant No. U23A20318, 62276195 and 62376200), the Science and Technology Major Project of Hubei Province (Grant No. 2024BAB046) and the Innovative Research Group Project of Hubei Province (Grant No. 2024AFA017). Dr Tao's research is partially supported by NTU RSR and Start Up Grants. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023a. Qwen technical report. *arXiv preprint*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023b. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint*.

Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2023. Token Merging: Your ViT but Faster. In *ICLR*.

Cai, Z.; Cao, M.; Chen, H.; Chen, K.; Chen, K.; Chen, X.; Chen, X.; Chen, Z.; Chen, Z.; Chu, P.; Dong, X.; Duan, H.; Fan, Q.; Fei, Z.; Gao, Y.; Ge, J.; Gu, C.; Gu, Y.; Gui, T.; Guo, A.; Guo, Q.; He, C.; Hu, Y.; Huang, T.; Jiang, T.; Jiao, P.; Jin, Z.; Lei, Z.; Li, J.; Li, J.; Li, L.; Li, S.; Li, W.; Li, Y.; Liu, H.; Liu, J.; Hong, J.; Liu, K.; Liu, K.; Liu, X.; Lv, C.; Lv, H.; Lv, K.; Ma, L.; Ma, R.; Ma, Z.; Ning, W.; Ouyang, L.; Qiu, J.; Qu, Y.; Shang, F.; Shao, Y.; Song, D.; Song, Z.; Sui, Z.; Sun, P.; Sun, Y.; Tang, H.; Wang, B.; Wang, G.; Wang, J.; Wang, J.; Wang, R.; Wang, Y.; Wang, Z.; Wei, X.; Weng, Q.; Wu, F.; Xiong, Y.; Xu, C.; Xu, R.; Yan, H.; Yan, Y.; Yang, X.; Ye, H.; Ying, H.; Yu, J.; Yu, J.; Zang, Y.; Zhang, C.; Zhang, L.; Zhang, P.; Zhang, P.; Zhang, R.; Zhang, S.; Zhang, S.; Zhang, W.; Zhang, W.; Zhang, X.; Zhang, X.; Zhao, H.; Zhao, Q.; Zhao, X.; Zhou, F.; Zhou, Z.; Zhuo, J.; Zou, Y.; Qiu, X.; Qiao, Y.; and Lin, D. 2024. InternLM2 Technical Report.

Chen, K.; Thapa, R.; Chalamala, R.; Athiwaratkun, B.; Song, S. L.; and Zou, J. 2024a. Dragonfly: Multi-Resolution Zoom Supercharges Large Visual-Language Model. *arXiv preprint*.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. 2024b. Are We on the Right Way for Evaluating Large Vision-Language Models? *arXiv preprint*.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; Li, B.; Luo, P.; Lu, T.; Qiao, Y.; and Dai, J. 2023. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint*.

Dong, X.; Zhang, P.; Zang, Y.; Cao, Y.; Wang, B.; Ouyang, L.; Wei, X.; Zhang, S.; Duan, H.; Cao, M.; Zhang, W.; Li, Y.; Yan, H.; Gao, Y.; Zhang, X.; Li, W.; Li, J.; Chen, K.; He, C.; Zhang, X.; Qiao, Y.; Lin, D.; and Wang, J. 2024a. InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model. *arXiv preprint*.

Dong, X.; Zhang, P.; Zang, Y.; Cao, Y.; Wang, B.; Ouyang, L.; Zhang, S.; Duan, H.; Zhang, W.; Li, Y.; Yan, H.; Gao, Y.; Chen, Z.; Zhang, X.; Li, W.; Li, J.; Wang, W.; Chen, K.; He, C.; Zhang, X.; Dai, J.; Qiao, Y.; Lin, D.; and Wang, J. 2024b. InternLM-XComposer2-4KHD: A Pioneering Large Vision-Language Model Handling Resolutions from 336 Pixels to 4K HD. *arXiv preprint*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Duan, H.; Yang, J.; Qiao, Y.; Fang, X.; Chen, L.; Liu, Y.; Dong, X.; Zang, Y.; Zhang, P.; Wang, J.; et al. 2024. VLMEvalKit: An Open-Source Toolkit for Evaluating Large Multi-Modality Models. *arXiv preprint*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint*.

Ge, C.; Cheng, S.; Wang, Z.; Yuan, J.; Gao, Y.; Song, J.; Song, S.; Huang, G.; and Zheng, B. 2024. ConvLLaVA: Hierarchical Backbones as Visual Encoder for Large Multimodal Models. *arXiv preprint*.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; Yu, H.; Wang, H.; Sun, J.; Zhang, J.; Cheng, J.; Gui, J.; Tang, J.; Zhang, J.; Li, J.; Zhao, L.; Wu, L.; Zhong, L.; Liu, M.; Huang, M.; Zhang, P.; Zheng, Q.; Lu, R.; Duan, S.; Zhang, S.; Cao, S.; Yang, S.; Tam, W. L.; Zhao, W.; Liu, X.; Xia, X.; Zhang, X.; Gu, X.; Lv, X.; Liu, X.; Liu, X.; Yang, X.; Song, X.; Zhang, X.; An, Y.; Xu, Y.; Niu, Y.; Yang, Y.; Li, Y.; Bai, Y.; Dong, Y.; Qi, Z.; Wang, Z.; Yang, Z.; Du, Z.; Hou, Z.; and Wang, Z. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools.

Gu, S.; Lugmayr, A.; Danelljan, M.; Fritsche, M.; Lamour, J.; and Timofte, R. 2019. Div8k: Diverse 8k resolution image dataset. In *ICCVW*.

Han, X.; You, Q.; Liu, Y.; Chen, W.; Zheng, H.; Mrini, K.; Lin, X.; Wang, Y.; Zhai, B.; Yuan, J.; Wang, H.; and Yang, H. 2023. CORE-MM: Complex Open-Ended Reasoning Evaluation For Multi-Modal Large Language Models.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *TMLR*.

Kembhavi, A.; Salvato, M.; Kolve, E.; Seo, M.; Hajishirzi, H.; and Farhadi, A. 2016. A Diagram is Worth a Dozen Images. In *ECCV*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *ICCV*.

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1.

Li, Y.; Zhang, Y.; Wang, C.; Zhong, Z.; Chen, Y.; Chu, R.; Liu, S.; and Jia, J. 2024a. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint*.

Li, Z.; Yang, B.; Liu, Q.; Ma, Z.; Zhang, S.; Yang, J.; Sun, Y.; Liu, Y.; and Bai, X. 2024b. Monkey: Image resolution and text label are important things for large multi-modal models. In *CVPR*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *CVPR*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; You, Q.; Han, X.; Wang, Y.; Zhai, B.; Liu, Y.; Tao, Y.; Huang, H.; He, R.; and Yang, H. 2024c. InfiMM-HD: A Leap Forward in High-Resolution Multimodal Understanding. *arXiv preprint*.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *CVPR*.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; Sun, Y.; Deng, C.; Xu, H.; Xie, Z.; and Ruan, C. 2024. DeepSeek-VL: Towards Real-World Vision-Language Understanding.

Luo, G.; Zhou, Y.; Zhang, Y.; Zheng, X.; Sun, X.; and Ji, R. 2024. Feast Your Eyes: Mixture-of-Resolution Adaptation for Multimodal Large Language Models. *arXiv preprint*.

Masry, A.; Long, D.; Tan, J. Q.; Joty, S.; and Hoque, E. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *ACL*.

Mathew, M.; Karatzas, D.; and Jawahar, C. 2021. Docvqa: A dataset for vqa on document images. In *WACV*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillicrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*.

Tian, Z.; Wen, Z.; Wu, Z.; Song, Y.; Tang, J.; Li, D.; and Zhang, N. L. 2022. Emotion-Aware Multimodal Pre-training for Image-Grounded Emotional Response Generation. In *Database Systems for Advanced Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Proceedings, Part III*, 3–19. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-00128-4.

Tian, Z.; Xie, Z.; Lin, F.; and Song, Y. 2023. A Multi-view Meta-learning Approach for Multi-modal Response Generation. In *Proceedings of the ACM Web Conference 2023*, WWW '23, 1938–1947. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394161.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*.

Wang, W.; Ding, L.; Shen, L.; Luo, Y.; Hu, H.; and Tao, D. 2024. WisdoM: Improving Multimodal Sentiment Analysis by Fusing Contextual World Knowledge. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 2282–2291. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Wei, H.; Kong, L.; Chen, J.; Zhao, L.; Ge, Z.; Yang, J.; Sun, J.; Han, C.; and Zhang, X. 2023. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint*.

Wu, P.; and Xie, S. 2024. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs. In *CVPR*.

Xu, R.; Yao, Y.; Guo, Z.; Cui, J.; Ni, Z.; Ge, C.; Chua, T.-S.; Liu, Z.; Sun, M.; and Huang, G. 2024. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint*.

Yifan, L.; Yifan, D.; Kun, Z.; Jinpeng, W.; Xin, Z.; and Ji-Rong, W. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *EMNLP*.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *ICML*.

Yuan, L.; Haodong, D.; Yuanhan, Z.; Bo, L.; Songyang, Z.; Wangbo, Z.; Yike, Y.; Jiaqi, W.; Conghui, H.; Liu, Z.; Kai, C.; and Dahua, L. 2023. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint*.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; Wei, C.; Yu, B.; Yuan, R.; Sun, R.; Yin, M.; Zheng, B.; Yang, Z.; Liu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen, W. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *CVPR*.

Zhang, Y.-F.; Wen, Q.; Fu, C.; Wang, X.; Zhang, Z.; Wang, L.; and Jin, R. 2024. Beyond LLaVA-HD: Diving into High-Resolution Large Multimodal Models. *arXiv preprint*.

Zheng, C.; Zhou, H.; Meng, F.; Zhou, J.; and Huang, M. 2023. Large language models are not robust multiple choice selectors. In *ICLR*.

Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 2024. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. In *ICLR*.