M^4Bench: A Benchmark of Multi-domain Multi-granularity Multi-image Understanding for Multi-modal Large Language Models

Xiaojun Ye[1], **Guanbao Liang**[1], **Chun Wang**[1], **Liangcheng Li**[1], **Pengfei Ke**[2], **Rui Wang**[2], **Bingxin Jia**[2], **Gang Huang**[2], **Qiao Sun**[1] and **Sheng Zhou**[1*]

[1]Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems, Zhejiang University, Hangzhou, China

[2]Alibaba Group, Hangzhou, China

{yexiaojun,liangguanbao,zjuheadmaster,liangcheng_li,sunqiao_zju,zhousheng_zju}@zju.edu.cn, {kepengfei.kpf,wr246073,bingxin.jbx}@taobao.com, tengyuan.hg@alibaba-inc.com

Abstract

The increasing demands in analyzing complex associated scenes pose necessities to researching multi-image understanding abilities. Compared with understanding individual images, both the alignments and differences between images are essential aspects of understanding the intricate relationships for multi-image inference tasks. However, existing benchmarks face difficulties in addressing both of these aspects simultaneously, resulting in obstacles to modeling relationships under various granularities and domains of images. In this paper, we introduce M^4Bench to enhance the capability of aligning and distinguishing multi-images with multi-domain multi-granularity comparison. We carefully design five comparison tasks related to coarse and fine-grained granularities in single and multiple domains of images and evaluate them on 13 state-of-the-art multi-modal large language models with various sizes. Besides, we analyze the evaluation results and provide several observations and viewpoints for the multi-image understanding research. The data and evaluation code are available at https://github.com/eaglelab-zju/M4Bench.

1 Introduction

Multi-modal large language models (MLLMs) [Liu *et al.*, 2024a] have demonstrated outstanding performance in various vision-language tasks. Early attempts mainly focus on bridging the gap between text and image modalities in single-image scenarios, ranging from image captioning [Yuan *et al.*, 2024] and visual grounding [Xuan *et al.*, 2024] to VQA [Schwenk *et al.*, 2022]. Recently, increasing demands in analyzing complex associated scenes pose necessities to researching *multi-image understanding* abilities. For example, in automatic driving, models use multiple images to compare environmental changes [Naranjo *et al.*, 2005]. In video monitoring, models compare multiple temporal images to analyze human behavior [Ye *et al.*, 2024; Ye *et al.*, 2025].

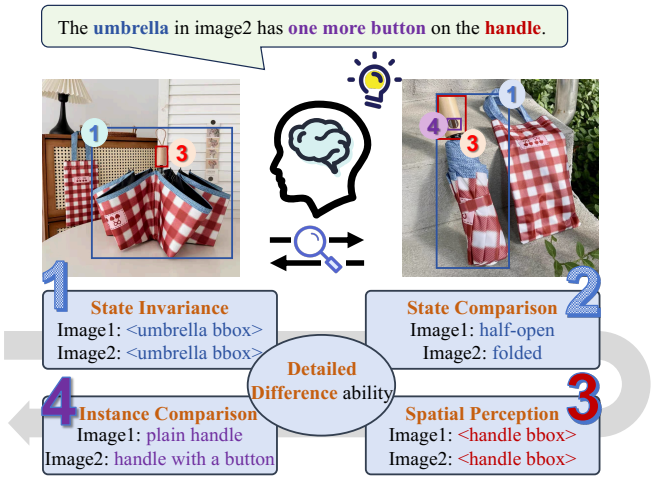


Figure 1: Illustration of five subtasks for imitating human thinking mode from M^4Bench. When people compare two images, they first focus on the coarse-grained object-level representation (in blue), then further attend to the part-level representation (in red), and finally perform a fine-grained feature visual comparison (in purple).

Compared with understanding individual images, the key to understanding multi-image lies in both the alignment and distinction between images [Wang *et al.*, 2024a]. However, existing benchmarks [Liu *et al.*, 2024c; Liu *et al.*, 2024b; Wu *et al.*, 2025] face difficulties in addressing both of these aspects simultaneously. On the one hand, some efforts [Wu *et al.*, 2025; Fu *et al.*, 2023; Wu *et al.*, 2024] have been directed towards enhancing the ability to identify differences between images, though the majority focus on a coarse-grained level. In practice, fine-grained differences, such as continuous action understanding [Lu *et al.*, 2024], are very common, which presents significant challenges for deploying existing MLLMs in such a scenario. On the other hand, some efforts [Fu *et al.*, 2025] have been made to enhance the ability to align parts between images. This typically requires a fine-grained alignment, yet the images in current benchmarks tend to originate from the same domain [Liu *et al.*, 2024b; Kil *et al.*, 2024]. In addtion, the input given to the model often consists of multiple images across different scenarios. For ex-

*Corresponding author.

ample, when comparing different design products, the model is required to overcome scenario difference and compare the detailed differences between products. Consequently, the challenge of simultaneously modeling relationships and differences across various granularities and domains has become critical, but remains largely underexplored.

To bridge this research gap, in this paper, we introduce M^4Bench, a benchmark dedicated to simultaneously evaluating the capability of aligning and distinguishing **m**ulti-images with **m**ulti-domain **m**ulti-granularity comparison for **m**ultimodal language models. Humans naturally excel at synthesizing information from multiple image sources [Wang *et al.*, 2024a]. Following human-like thinking mode, M^4Bench is structured along three dimensions, encompassing a total of five multi-image subtasks. Our benchmark concatenates two images horizontally as the visual input, employing multi-granularity visual prompts as contextual cues and response options to minimize the model's dependence on linguistic prior knowledge. In comparison tasks, humans typically first identify the main objects in each image, which tests *Multi-domain Coarse-grained Comparison* ability. Once these high-level objects are determined, they employ a top-down processing approach [Noudoost *et al.*, 2010] to guide their attention towards more detailed features, which assesses *multi-domain Fine-grained Comparison* ability. *Single-domain Fine-grained Comparison* ability is a prerequisite for this assessment. Correspondingly, as illustrated in Figure 1, State Invariance (SI) first examines the models' ability to correctly recognize objects in different states. After correct recognition, State Comparison (SC) requires the model to distinguish between different physical states of the same category object. Detailed Difference (DD) evaluates the models' fine-grained perception ability. Finally, multi-domain tasks Spatial Perception (SP) and Instance Comparison (IC) test whether models can overcome scenario differences and focus on key fine-grained features. With the above design, M^4Bench evaluates the comparison capabilities of models at different stages.

We evaluated 13 MLLMs with various sizes (i.e., 2B, 4B, 7B) on M^4Bench. Our results reveal that existing MLLMs struggle in fine-grained comparison tasks, especially in multi-domain scenario. Besides, we also designed the caption test experiment to prove that our M^4Bench previously overcomes the conflation of multi-image comparison with language reasoning ability in previous benchmarks. We perform further analysis of error cases, providing insight for future MLLM improvements. Our contributions are summarized as follows:

- We present M^4Bench for evaluating the multi-granularity, multi-domain comparison capability of MLLMs. Following human thinking mode, M^4Bench provides multi-grained visual prompts as contextual cues and response options, which can reliably evaluates the comparison capabilities of models at different stages.
- We designed four automatic pipelines and employed manual annotation to construct M^4Bench, which addresses a gap in the community for multi-domain fine-grained comparison datasets.
- The evaluation on M^4Bench reveals that current MLLMs perform poorly in multi-granularity, multi-domain comparison tasks. We also provide further analyses and insights for future improvement.

2 Related Work

2.1 Multi-Image mid-scale Multimodal Models

Multimodal Large Language Models (MLLMs) refer to models capable of processing multiple forms of input (e.g., text, images, and videos) and performing joint reasoning [Yin *et al.*, 2023]. Most open-source autoregressive vision-language models [Liu *et al.*, 2023; Zhang *et al.*, 2023; Li *et al.*, 2023c] are primarily focused on processing a single image, and exhibit limitations when handling more complex tasks that require comparison across multiple images. As research progresses, an increasing number of models [Awadalla *et al.*, 2023; Li *et al.*, 2023a; Sun *et al.*, 2024] have begun to support multi-image input. OpenFlamingo [Awadalla *et al.*, 2023], for instance, leverages the Perceiver Resampler and cross-attention layers to achieve in-context learning capabilities. However, in low-data scenarios, this approach struggles with performance efficiency and computational cost. InternVL 1.5 [Chen *et al.*, 2024] and Qwen2-VL [Wang *et al.*, 2024b] introduce dynamic resolution support, which dynamically transforms multi-image input into a variable number of visual tokens, optimizing MLLM computational efficiency in multi-image scenarios. Qwen2-VL also employs a Multimodal Rotary Positional Embedding (M-RoPE) technique to enhance cross-image contextual reasoning. MiniCPM [Hu *et al.*, 2024] further improves efficiency by optimizing the density of visual tokens, allowing the model to handle multiple images without sacrificing performance. The LLaVA-NEXT series [Li *et al.*, 2024b; Li *et al.*, 2024a], while inheriting LLaVA's minimalist design, enhances multi-image task processing through a high-quality multi-image instruction dataset. Despite these advancements, current multi-image models still exhibit shortcomings in multi-domain comparison tasks.

2.2 Multi-Image Multimodal Benchmarks

Most benchmarks [Liu *et al.*, 2025; Yu *et al.*, 2023] primarily focus on single-image evaluation, and often overlook multi-image perception and reasoning abilities, which hold even greater practical value. Some recent studies exploring multi-image scenarios, but they mainly focus on coarse-grained multi-image perception and lack object-centric comparison. Q-Bench [Wu *et al.*, 2023] is proposed to evaluate the low-level perception of MLLMs by comparing multiple image inputs. Memontos [Wang *et al.*, 2024c] evaluates the temporal understanding of the image sequences. The DEMON benchmark [Li *et al.*, 2023b] also present several subset of questions requiring multi-image reasoning, where the main focus is on evaluating the demonstrative instruction following abilities of MLLMs. MIBench [Liu *et al.*, 2024b] and CompBench [Kil *et al.*, 2024] has expanded the evaluation dimensions and sample numbers for multi-modal tasks, but each subtask is limited to either coarse-grained or fine-grained capabilities separately. BLINK [Fu *et al.*, 2025] evaluates the core visual perception ability based on classic computer vision tasks, but by directly drawing the options on the images,

Subtask	#Samples	Domain	Granularity	collection
SI	2K	multi	object	pipeline
SC	0.5K	multi	object	pipeline
DD	1K	single	object/part	pipeline
SP	2K	multi	part	pipeline
IC	0.2K	multi	part	manual

Table 1: the Statistics of M^4Bench. SI means State Invariance, SC means State Comparison, DD means Detailed Difference, SP means Spatial Perception, IC means Instance Comparison.

	VP	G	S	BC
Q-Bench	✗	image	multi	independent
MMRA	✗	object	multi	independent
BLINK	✗	part	multi	independent
MIBench	✗	object	both	independent
CompBench	✗	object	both	independent
M^4Bench	✓	part	both	thinking mode

Table 2: Comparison of M^4Bench with existing benchmarks. VP means visual prompts, G means visual feature granularity, S means scenario, "both" means the benchmark encompasses both single-domain and multi-domain scenario, BC means benchmark construction methodology."independent" denotes that previous benchmarks typically design subtasks independent of each other, yielding results that only demonstrate a model's accuracy on a specific type of question-answering task. While M^4Bench incorporates subtasks ranging from coarse-grained to fine-grained, encompassing both single-domain and multi-domain scenarios, which enables a multi-stage evaluation of MLLMs' comparison abilities.

it indirectly compromises the reliability of evaluating models' multi-image comparison ability. Additionally, BLINK's images are entirely sourced from existing datasets, failing to address the community's lack of object-centric comparison datasets. In this paper, we propose a brand-new "visual-centric" multi-image comparison benchmark that provides visual object-level and part-level visual prompts, evaluating multi-granularity perception ability.

3 M^4Bench

3.1 Overview of M^4Bench

The M^4Bench includes three dimension of single and multi domain scenarios along with five vision comparison tasks as illustrated in Figure 2:

- **Multi-domain Coarse-grained scenario** aims to recognize objects in different structural forms and distinguish the physical states of objects.
- **Single-domain Fine-grained scenario** aims to identify fine-grained difference by detecting subtle changes between images.
- **Multi-domain Fine-grained scenario** aims to compare fine-grained features of the interest object under diverse environments.

When given two images, coarse-grained scenarios require distinguishing features at the image or object level, while fine-grained scenarios focus on the comparison of subtle local parts of objects. Furthermore, multi-domain scenarios have diverse prior environments of two images, such as image attribute settings, background complexity, arrangement of objects, lighting, and shadows, which are challenges in distinguishing the invariance of objects in two images.

The subtasks consist of intra-class image pairs and multi-grained visual prompts collected via four automatic pipelines and manual annotation. Table 1 shows detailed settings.

In addition, we compare our M^4Bench with previous benchmarks from four aspects: with or without visual prompts, visual features granularity, scenario, and benchmark construction methodology. Table 2 shows the details.

3.2 Dataset Collection Process

Multi-domain Coarse-grained

State Invariance (SI). The ability to correctly detect objects in different states is a prerequisite for the multi-image comparison task. However, when the state of an object changes, the object embeddings generated by current MLLMs also change. This task helps us evaluate the robustness of the model in generating discriminative features at the object level. We design the task with each sample containing two images and a yes-no multiple-choice question to judge the difference. We collect positive samples from the ObjectsWithStateChange dataset [Sarkar and Kak, 2024] to test whether the model can correctly classify objects with different structural forms as the same object. Negative samples are constructed from the Grocery Store dataset [Klasson *et al.*, 2019] to test the model's ability to distinguish between similar but different objects.

State Comparison (SC). Understanding object states is essential for common sense physical reasoning. We design a subtask to evaluate MLLM's ability to encode the objects' physical states, focusing on the comparison of diverse states of objects. Our data is sourced from the ChangeIt dataset [Souček *et al.*, 2022], which features videos of various objects undergoing state changes in different scenarios, along with annotations for actions and state localization. Our pipeline automatically capture the initial and end state keyframes from each video, and then pair these state keyframes of objects in the same category into image pairs with corresponding state pairs as answers.

Single-domain Fine-grained

Detailed Difference (DD). Given that objects can be futher subdivided into diverse parts, reasoning at a coarse-grained level is insufficient for subtle comparison tasks, and part-level features should be considered to support the detailed difference. Therefore, we consider datasets that provide image pairs with similar layouts but subtle changes. To generalize diverse vision scenarios, we provide both synthetic (DD-SI) and natural (DD-NI) multi-image pairs from MagicBrush [Zhang *et al.*, 2024] and Spot-the-diff [Jhamtani and Berg-Kirkpatrick, 2018] respectively. The former consists of triplets of source image, mask images and target image, while the latter provides image pairs in several outdoor scenes.

We also provide an automatic pipeline for constructing the detailed difference task with multiple-choice questions.

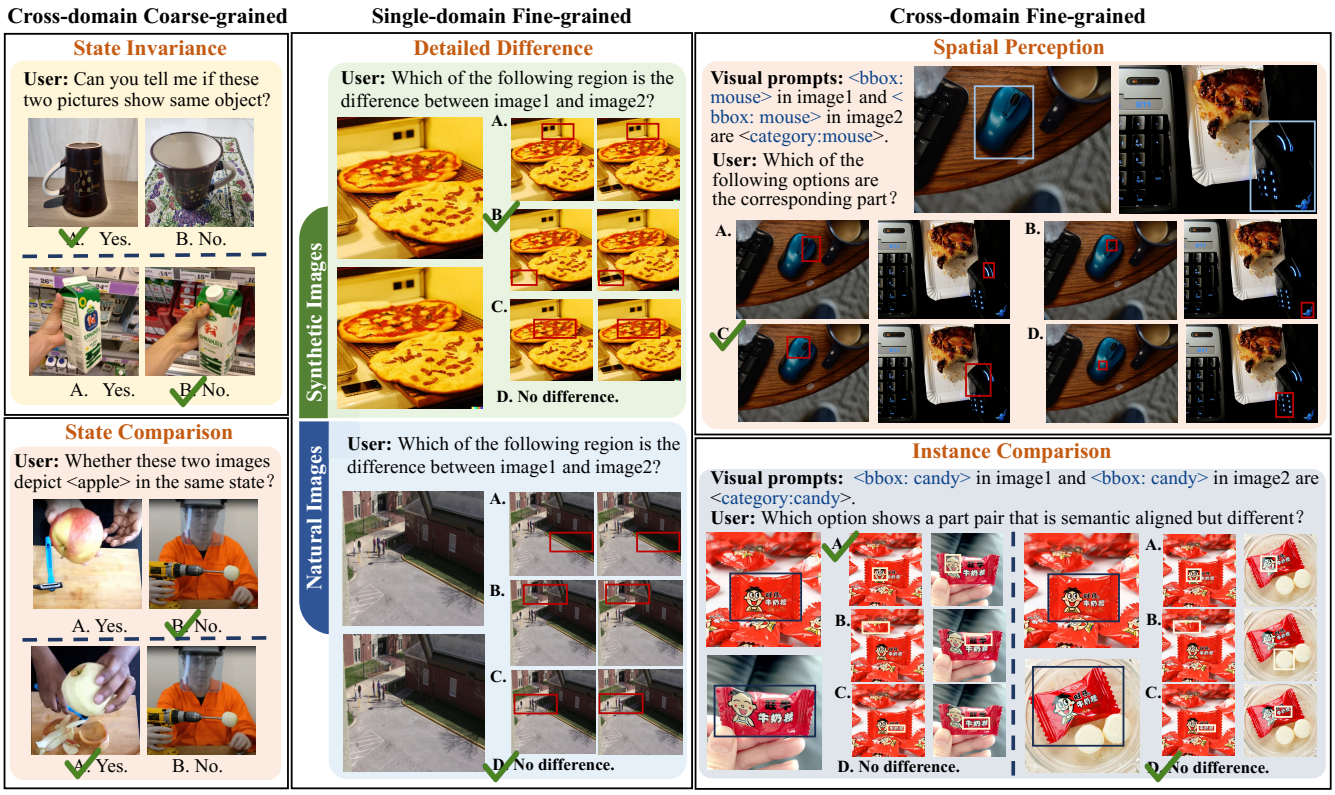


Figure 2: **Overview of the M^4Bench benchmark.** M^4Bench offers diverse triplets comprising two images from single and multi scenario, comparision question, and and answer based multi-grained key visual feature.

Our options consist of three parts: **(1) Generating bounding boxes of the difference region.** In synthetic images, the bounding boxes are derived from the masked area provided by mask image. In natural images, they are calculated by the surrounding bounding box of the difference cluster. **(2) Designing distractors.** We design an automatic pipeline to randomly generate non-overlapping bounding boxes in non-different areas of the image. **(3) Collecting image pairs with the same images.** We collect the same images and label the answer with "no difference" as negative samples.

Multi-domain Fine-grained

Spatial Perception (SP). Other than capturing coarse-grained global features and fine-grained local features, MLLMs should establish connections between them to learn the hierarchical layout correspondences in the image pairs to gain spatial perception. Therefore, we derive our samples from the PACO-LVIS dataset [Ramanathan *et al.*, 2023], which provides object and part maskon images in the COCO dataset [Lin *et al.*, 2014]. We select the images in which objects contain relative spatial parts for the spatial perception task, such as symmetrical components like a mouse with left and right buttons and subassembly parts like bottles with necks, shoulders and bottoms. We retrieve from the PACO-LVIS based on the object category and filter out image pairs containing four or more partial components with multiple choice questions designed. In the option design, we provide bounding boxes of the potential interest regions, with the correct answer setting the corresponding part and the other options randomly matching unaligned parts as distractors.

Instance Comparison (IC). We hope that the models will overcome the semantic ambiguity caused by changing vision scenarios when comparing multiple images. Thus, they are suggested to not only focus on the corresponding parts of the interest instances (in SP) but also distinguish the part-level differences. Therefore, we propose a challenging VQA task of Instance Comparison to address the lack of comparable image pairs of different designs for the same instances. Furthermore, we annotate a test set that contains images of the same product in different designs across diverse scenarios, along with multi-grained visual prompts, on the AIData platform.

Instance Comparison in deed is a two-stage task: object-level positioning and part-level comparision, so we provide two settings: with or without visual prompts. Visual prompts tell the models the bounding boxes on the interest objects, which facilitates the completion of the initial localization phase, subsequently allowing for the assessment of models' comparative capabilities in the second phase.

3.3 Avoiding Data Flaws

We construct test data of M^4Bench by utilizing the validation or test sets from existing datasets. Furthermore, we combine automated filtering and manual verification to ensure the quality and reliability of the test data. We discard samples that can still be answered correctly without visual input. This

Models	Multi-domain Coarse-grained		Single-domain Fine-grained		Multi-domain Fine-grained			
	SI	SC	DD-SI	DD-NI	SP-none	SP-vp	IC-none†	IC-vp†
Random Choice	50.0	50.0	25.0	25.0	25.0	25.0	25.0	25.0
Open-source MLLMs								
InternVL2-4B	51.6	64.4	14.1	3.0	27.8	28.7	17.0	22.6
InternVL2-8B	73.1	60.1	14.5	13.5	**33.9**	37.4	33.0	36.5
InternVL2.5-4B	70.3	65.4	8.1	6.5	31.7	30.9	26.5	34.4
InternVL2.5-8B	73.1	65.4	4.7	20.0	30.4	35.7	31.7	36.5
Qwen2VL-2B	64.4	54.3	39.3	21.3	28.3	25.2	34.4	40.0
Qwen2VL-7B	72.1	68.8	60.3	15.7	**33.9**	33.5	**47.0**	**42.2**
MiniCPM-V-2.6-8B	76.7	59.6	-	-	-	-	-	-
LLaVA-OneVision-7B	80.8	67.8	-	-	-	-	-	-
DeepSeek-VL2-tiny	47.0	56.7	69.2	28.7	26.1	24.4	19.1	23.9
DeepSeek-VL2-small	75.8	**71.2**	78.6	**65.2**	27.4	23.9	41.7	26.5
Closed-source MLLMs								
Qwen-VL-Max	84.5	64.9	**91.9**	16.1	21.7	33.5	30.0	27.8
Gemini-Pro	**92.7**	60.6	40.2	26.1	10.4	24.3	17.8	22.6
GPT-4o	88.6	66.8	41.0	24.8	14.8	**39.6**	21.3	19.1

Table 3: Evaluation results of M^4Bench on both open-source and closed-source MLLMs. † denotes manual annotation. '-' denotes the model does not support the subtask. **Bold** denotes the state-of-the-art model on the subtask.

avoids the overestimation of model performance due to the textual bias of the questions and options. To improve the quality of the generated samples, we apply manual verification after automated filtering. The process is conducted by three human annotators recruited within the team. They delete inferior image pairs such as the images are blurry and the main subject is relatively small.

4 Experiments

4.1 Models

We evaluate 13 MLLMs which can be categorized into two groups: (1) closed-source API models, including GPT-4o, Gemini-Pro and Qwen-VL-Max; (2) open-source models, including DeepSeek-VL2 (tiny and small), Qwen2VL (2B and 7B), MiniCPM-V-2.6 (8B), InternVL (v2, v2.5, model size 4B and 8B) and LLaVA-OneVision (7B). All these models inherently support multiple image inputs.

4.2 Evaluation Setup

To ensure reproducibility, we follow the approach used in VLMEvalKit [Contributors, 2023], extracting answers from the free-form outputs of MLLMs. Specifically, for multiple-choice questions, we obtain the options through predefined rules and the assistance of GPT-3.5-turbo, and finally report the results using accuracy (ACC). In addition, to solve the position bias [Liu *et al.*, 2025], we shuffle the options randomly to balance the distribution of correct answers among A, B, C, and D to mitigate the position bias.

We conducted Multi-domain Coarse-grained testing on all models. However, it is noteworthy that some of the models were not pre-trained on input data in the visual prompts format. They exhibited a negative effect in the Single-domain Fine-grained and Multi-domain Fine-grained tests, i.e., their performance decreased as the number of demos increased. Therefore, we only present the evaluation results of models that support visual prompt input in these two dimension tests.

4.3 Main Results

Overall performance. Table 3 shows the evaluation results on our M^4Bench, where "Random Choice" implies that the correct answer probability is 50% for yes-no questions and 25% for four-choice questions. From the table, the average accuracy across all tasks for MLLMs is 39.6%, which is 8.3% higher than random choice (31.3%). Even the top-performing models in our benchmark, the open-source DeepSeek-VL2-small and the closed-source Qwen-VL-Max, achieved only 51.3% and 46.3% accuracy, respectively. In general, the multi-domain coarse-grained tasks (68.3%) are significantly simpler compared to single-domain fine-grained tasks (31.9%) and multi-domain fine-grained tasks (29.0%), with the latter being particularly more challenging.

Performance limitations of models on M^4Bench. Figure 3 shows that MLLMs generally perform better on multi-domain coarse-grained tasks (SI and SC) in the upper right half, while struggling with multi-domain fine-grained tasks (IC and SP) in the left half. This performance disparity reveals the varying levels of challenge in multi-image understanding tasks. Although Qwen-VL-Max achieves the best performance on the DD-SI task, surpassing DeepSeek-VL2-small and demonstrating the potential of closed-source MLLMs on synthetic images, there are still notable limitations in current MLLMs. These limitations can be attributed to two main factors. First, current MLLMs primarily adopt data-driven optimization approaches that process mul-

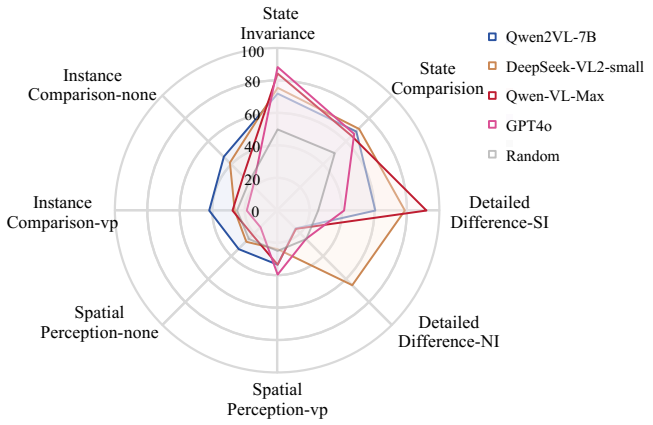


Figure 3: Accuracies of MLLMs on M^4Bench among different tasks. Here is a demonstration of the performance of two of the best open-source and two of the best closed-source MLLMs.

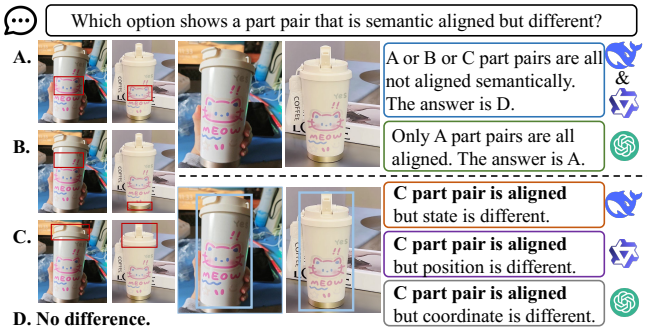


Figure 4: Case analysis of the effects of irrelevant factors in cross-scenario and visual prompts.

tiple images independently and concatenate them at the token level, potentially constraining their ability to perform multi-domain and multi-granularity visual comparisons, which are crucial for M^4Bench tasks. Second, most MLLMs rely on LLMs for reasoning, projecting visual information into textual space before analysis, which may result in the loss of essential fine-grained visual details.

4.4 Single-domain vs. Multi-domain

In single-domain scenarios, the input images have highly consistent visual scenes, such as similar backgrounds, lighting conditions, object sizes, and visual angles. Models with fine-grained perception capabilities can extract reliable features for difference comparison. In contrast, multi-domain scenarios introduce irrelevant parts(i.e. noise and biases), so we design SP and IC to examine the model's ability to overcome irrelevant factors and compare the corresponding and opposite details between two images. Experimental results show that the average value of SOTA in fine-grained multi-domain scenarios decreases by 37.88% compared to the average value of SOTA in fine-grained single-domain scenarios. The challenge of multi-domain tasks lies in the fact that multi-domain can divide the semantic image regions into positively correlated (**corresponding** parts), negatively correlated (**"aligned but different"** parts) and irrelevant parts. The irrelevant regions of interest come from prior knowledge of the options. As illustrated in Figure 4, IC requires the model to overcome the confusion caused by biases and noise, such as differences in background environment, and identify the differences of details in the interest regions. Changes in the visual angle and specific coordinates of the bounding box are irrelevant parts with instances of interest, but these changes may affect the model's judgment. From the above comparisons, it can be concluded that current MLLMs may have lost their sensitivity to fine-grained differences in the transition between different scenarios due to a lack of effective multi-domain adaptation strategies, resulting in a decrease in accuracy.

4.5 Coarse-grained vs. Fine-grained

In coarse-grained tasks, encoding global features is sufficient to answer questions on object categories and attributes, while there still exists flaws in answering questions on diverse states of objects, excluding the close-source model Gemini-Pro. In the construction of fine-grained scenarios, more part-level relations between the images can be excavated such as texture-feature and attribute-association relationships. As shown in the case of IC in Figure 2, to answer the question, the models need to pay attention to the hot kid's emoticon on the candy package, requiring part-level representation. As shown in Table 3, the average accuracy of multi-domain fine-grained scenarios compared to multi-domain Coarse-grained scenarios decreases by 39.3%, which indicates that current MLLMs are inadequate in fine-grained perception tasks.

4.6 Effect of Model Scaling on Performance

In the analysis of the performance of open-source MLLMs on the M^4Bench benchmark, we observe that increasing the model parameter scale does not uniformly lead to performance improvements across all tasks. Fine-grained tasks require modeling more complex features, and thus increasing the parameter scale can enhance the model's performance on single-domain tasks. Notably, model scaling has negligible impact on the performance of InternVL2 series on DD-SI, InternVL2.5 on SI, SC, and IC-vp, Qwen2VL on IC-vp, and DeepSeek series on SP-none. This indicates that further improvements may require architectural modifications or different strategies beyond merely improving the parameters scale. Increasing the parameter scale for test tasks with significantly different data distributions from the training set can lead to performance degradation. As shown in Table 3. InternVL2.5 on the DD-SI task drops from 8.1% to 4.7%, and Qwen2VL on the DD-SI task drops from 21.3% to 15.7%. The above analysis indicates that achieving an optimal balance across multiple subtasks when scaling up model parameters represents a promising direction for future research.

4.7 Visual Prompts Analysis

Some studies show that humans tend to adopt the top-down processing approach when comparing multiple images [Gilbert and Li, 2013; Oliva *et al.*, 2003]. This means that the human visual system first locates the main objects in

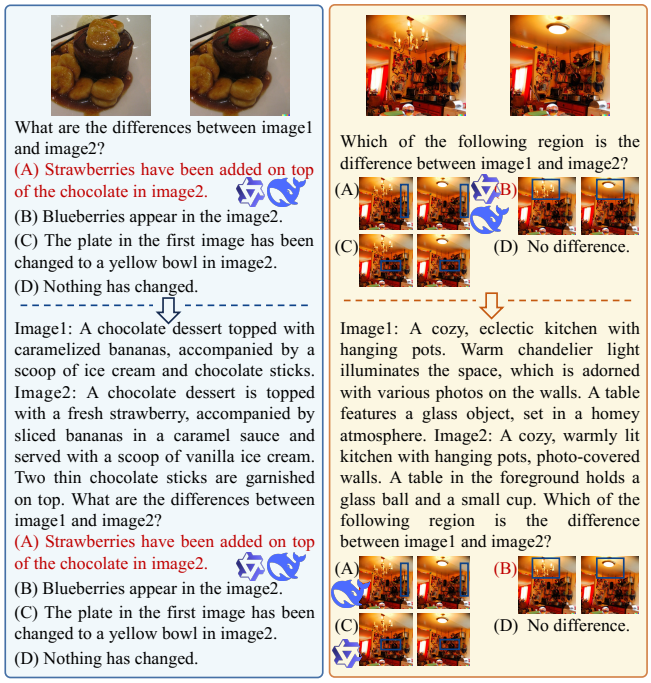
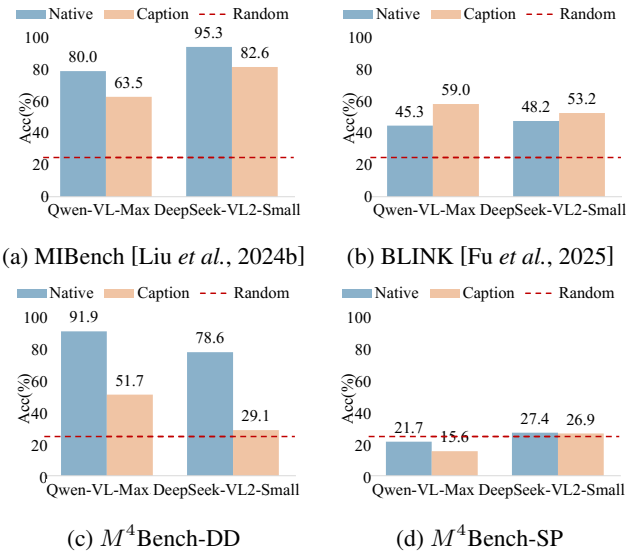


Figure 5: A qualitative case study comparing the performance of MIBench's Subtle Difference and our M^4Bench's Detailed Difference on the *native test* and the *caption test*.Red option indicates the ground truth.



(a) MIBench [Liu *et al.*, 2024b] (b) BLINK [Fu *et al.*, 2025]

(c) M^4Bench-DD (d) M^4Bench-SP

Figure 6: The comparison of experimental results between our M^4Bench and previous multi-image benchmarks on the *native test* (in green) and *caption test* (in pink).

images, and then compares the specific details of those objects. To mimic human cognition, we provided object bounding boxes as visual prompts (vp) and without visual prompts (none) settings for the SP and IC tasks. We found that SP-vp had a 4.6% increase in average accuracy compared to SP-none, which claims that the SP task requires the model to extract geometry layouts for building pixel-level semantic correspondences. Notably, providing visual prompts did not improve the average accuracy for the IC task. DeepSeek-VL2-small is the model most significantly affected, with an accuracy drop of 15.2%. Meanwhile, we analyzed four models, DeepSeek-VL2-small, Qwen2VL-7B, Qwen-VL-Max, and GPT-4o, where the accuracy decreased after incorporating vp. When analyzing error cases, we found that the drop in accuracy mainly stemmed from a decrease in the probability of the models answering D. Figure 4 depicts the details. Without vp, the models considered A, B, and C as semantically misaligned parts and therefore chose D. However, the provided bounding boxes label the regions of interest and help establish spatial perception, testing the models' ability to distinguish irrelevant factors in multi-domain tasks(§ 4.4).

4.8 M^4Bench is more robust against "single-image" solutions

Popular multi-image benchmarks like MIBench [Liu *et al.*, 2024b] inadvertently encourage "single-image" models that exploit language biases. We demonstrate this by conducting *native test* and *caption test* on the previous benchmark and our M^4Bench's subtasks, respectively. The *caption test* involves using image captions instead of the actual images as input. We use GPT-4 to generate dense captions for each image. Figure 6 presents a comparative analysis of the impact of the *caption test* on the performance of the previous multi-image benchmarks and our M^4Bench. On MIBench, the *caption test* led to accuracy declines of only 16.5% and 12.8% for Qwen-VL-Max and DeepSeek-VL2-small, respectively. On BLINK, Qwen-VL-Max and DeepSeek-VL2-small achieved accuracy increases of 13.7% and 5.0% respectively. However, on our M^4Bench DD-SI task, compared to *native test*, the *caption test* led to a 40.2% drop in accuracy for Qwen-VL-Max and a 49.5% drop for DeepSeek-VL2-small. On the M^4Bench SP task, the *caption test* even caused Qwen-VL-Max's accuracy to be 9.4% lower than random choice. Figure 5 confirms that M^4Bench's design prevents models from solving multi-image comparison problems through a "single-image" approach, establishing it as a more vision-centric benchmark for reliable evaluation of MLLMs.

5 Conclusion

In this paper, we introduce M^4Bench, a benchmark dedicated to evaluate the capability of aligning and distinguishing multi-images with multi-domain multi-granularity comparison for MLLMs. M^4Bench constructs a hierarchical evaluation that includes five subtasks from coarse-grained to fine-grained, encompassing both single and multi domain scenarios. We evaluate 13 popular MLLMs on M^4Bench and measure the effect of converting images to dense captions. The experimental results shows multi-domain multi-grained comparison tasks pose significant challenges for current MLLMs. Even the SOTA models only achieve around 50% accuracy on M^4Bench. We offers detailed analyses and insights for future advancements in the multi-image tasks and the annotated data is publicly available to facilitate further research.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Grant No.62372408).

References

[Awadalla *et al.*, 2023] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[Chen *et al.*, 2024] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[Contributors, 2023] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. *GitHub repository*, 2023.

[Fu *et al.*, 2023] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

[Fu *et al.*, 2025] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2025.

[Gilbert and Li, 2013] Charles D Gilbert and Wu Li. Top-down influences on visual processing. *Nature reviews neuroscience*, 14(5):350–363, 2013.

[Hu *et al.*, 2024] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

[Jhamtani and Berg-Kirkpatrick, 2018] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*, 2018.

[Kil *et al.*, 2024] Jihyung Kil, Zheda Mai, Justin Lee, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, Arpita Chowdhury, and Wei-Lun Chao. Compbench: A comparative reasoning benchmark for multimodal llms. *arXiv preprint arXiv:2407.16837*, 2024.

[Klasson *et al.*, 2019] Marcus Klasson, Cheng Zhang, and Hedvig Kjellström. A hierarchical grocery store image dataset with visual and semantic labels. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 491–500. IEEE, 2019.

[Li *et al.*, 2023a] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023.

[Li *et al.*, 2023b] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*, 2023.

[Li *et al.*, 2023c] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[Li *et al.*, 2024a] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[Li *et al.*, 2024b] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[Liu *et al.*, 2024a] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[Liu *et al.*, 2024b] Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, et al. Mibench: Evaluating multimodal large language models over multiple images. *arXiv preprint arXiv:2407.15272*, 2024.

[Liu *et al.*, 2024c] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*, 2024.

[Liu *et al.*, 2025] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2025.

[Lu *et al.*, 2024] Quanfeng Lu, Wenqi Shao, Zitao Liu, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang,

Kaipeng Zhang, Yu Qiao, and Ping Luo. Gui odyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. *arXiv preprint arXiv:2406.08451*, 2024.

[Naranjo *et al.*, 2005] José Eugenio Naranjo, Carlos González, Ricardo García, Teresa de Pedro, and Rodolfo E Haber. Power-steering control architecture for automatic driving. *Ieee transactions on intelligent transportation systems*, 6(4):406–415, 2005.

[Noudoost *et al.*, 2010] Behrad Noudoost, Mindy H Chang, Nicholas A Steinmetz, and Tirin Moore. Top-down control of visual attention. *Current opinion in neurobiology*, 20(2):183–190, 2010.

[Oliva *et al.*, 2003] Aude Oliva, Antonio Torralba, Monica S Castelhano, and John M Henderson. Top-down control of visual attention in object detection. In *Proceedings 2003 international conference on image processing (Cat. No. 03CH37429)*, volume 1, pages I–253. IEEE, 2003.

[Ramanathan *et al.*, 2023] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023.

[Sarkar and Kak, 2024] Rohan Sarkar and Avinash Kak. Learning state-invariant representations of objects from image collections with state, pose, and viewpoint changes. *arXiv preprint arXiv:2404.06470*, 2024.

[Schwenk *et al.*, 2022] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022.

[Souček *et al.*, 2022] Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13956–13966, 2022.

[Sun *et al.*, 2024] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.

[Wang *et al.*, 2024a] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024.

[Wang *et al.*, 2024b] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[Wang *et al.*, 2024c] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*, 2024.

[Wu *et al.*, 2023] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023.

[Wu *et al.*, 2024] Siwei Wu, Kang Zhu, Yu Bai, Yiming Liang, Yizhi Li, Haoning Wu, JH Liu, Ruibo Liu, Xingwei Qu, Xuxin Cheng, et al. Mmra: A benchmark for evaluating multi-granularity and multi-image relational association capabilities in large visual language models. *arXiv preprint arXiv:2407.17379*, 2024.

[Wu *et al.*, 2025] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, et al. Towards open-ended visual quality comparison. In *European Conference on Computer Vision*, pages 360–377. Springer, 2025.

[Xuan *et al.*, 2024] Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. Pink: Unveiling the power of referential comprehension for multi-modal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13838–13848, 2024.

[Ye *et al.*, 2024] Xiaojun Ye, Junhao Chen, Xiang Li, et al. Mmad: Multi-modal movie audio description. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11415–11428, 2024.

[Ye *et al.*, 2025] Xiaojun Ye, Chun Wang, Yiren Song, et al. Focusedad: Character-centric movie audio description. *arXiv preprint arXiv:2504.12157*, 2025.

[Yin *et al.*, 2023] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[Yu *et al.*, 2023] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[Yuan *et al.*, 2024] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28202–28211, 2024.

[Zhang *et al.*, 2023] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[Zhang *et al.*, 2024] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024.