



Benchmarking large and small MLLMs

Xuelu Feng¹ · Yunsheng Li² · Dongdong Chen² · Mei Gao² · Mengchen Liu² · Junsong Yuan¹ · Chunming Qiao¹

Received: 14 April 2025 / Revised: 30 July 2025 / Accepted: 14 October 2025 / Published online: 30 October 2025
© The Author(s) 2025

Abstract

Large multimodal language models (MLLMs) such as GPT-4V and GPT-4o have achieved remarkable advancements in understanding and generating multimodal content, showcasing superior quality and capabilities across diverse tasks. However, their deployment faces significant challenges, including slow inference, high computational cost, and impracticality for on-device applications. In contrast, the emergence of small MLLMs, exemplified by the LLaVA-series models and Phi-3-Vision, offers promising alternatives with faster inference, reduced deployment costs, and the ability to handle domain-specific scenarios. Despite their growing presence, the capability boundaries between large and small MLLMs remain underexplored. In this work, we conduct a systematic and comprehensive evaluation to benchmark both small and large MLLMs, spanning general capabilities such as object recognition, temporal reasoning, and multimodal comprehension, as well as real-world applications in domains like industry and automotive. Our evaluation reveals that small MLLMs can achieve comparable performance to large models in specific scenarios but lag significantly in complex tasks requiring deeper reasoning or nuanced understanding. Furthermore, we identify common failure cases in both small and large MLLMs, highlighting domains where even state-of-the-art models struggle. We hope our findings will guide the research community in pushing the quality boundaries of MLLMs, advancing their usability and effectiveness across diverse applications.

Keywords Multi-modal large language model · Evaluation · General capabilities · Application

1 Introduction

1.1 Motivation and overview

The breakthroughs in large language models (LLMs) [1–4] have shown remarkable versatilities and capabilities across various domains and tasks. After that, the focus has shifted toward multimodal models that integrate vision and language, Multimodal Large Language Models (MLLMs) [5–13] marks a significant advancement in the ability to process and reason with inputs from different modalities, *e.g.*, text, image and video. These models demonstrate unparalleled quality in integrating visual and textual information, enabling breakthroughs in both general-purpose applications and advanced reasoning scenarios. However, their deployment is hindered by significant challenges. The massive computational resources required for training and inference result in slow processing speeds and high operational costs, making these models super expensive for many real-world applications. Additionally, their reliance on substantial hardware infrastructure prevents on-device

✉ Xuelu Feng
xuelufen@buffalo.edu

Yunsheng Li
yunshengli@microsoft.com

Dongdong Chen
dongdong.chen@microsoft.com

Mei Gao
xuemei.gao@microsoft.com

Mengchen Liu
mengcliu@microsoft.com

Junsong Yuan
jsyuan@buffalo.edu

Chunming Qiao
qiao@buffalo.edu

¹ Department of Computer Science and Engineering,
University at Buffalo, Buffalo, NY, USA

² Microsoft, Redmond, WA, USA

deployment, which is increasingly critical for applications demanding low latency, high efficiency, and privacy-preserving operations.

In response to these limitations, there has been a surge in the development of small multimodal language models, designed to overcome the deployment challenges of their larger counterparts. These models, exemplified by the LLaVA-series and Phi-3-Vision, prioritize efficiency and accessibility, which significantly reduces deployment costs and makes device deployment possible by reducing the model size. They can enable real-time applications and fostering new opportunities for domain-specific scenarios. For instance, small MLLMs can be fine-tuned for specialized fields such as healthcare, automotive, and industry, addressing niche requirements that large models cannot feasibly accommodate. Despite their reduced size and computational footprint, these models are beginning to demonstrate promising capabilities, further expanding the applicability of multimodal AI technologies.

To better understand these MLLMs' capability, many different benchmarks have been proposed [14–21]. However, previous evaluations have been narrow in scope, often focusing on limited task ranges or specific models:

- **Limited MLLM Comparison and Quantitative Analysis:** Some benchmarking methods [22, 23], while extensive in task diversity, focused primarily on assessing the capabilities of a single representative MLLM rather than conducting systematic comparisons across multiple models. Moreover, their evaluations were largely qualitative and based on limited-scale case studies, making it difficult to draw broader quantitative conclusions. In contrast, our benchmark explicitly emphasizes comparative analysis between large and small MLLMs across diverse tasks, enabling a more comprehensive understanding of their respective strengths and limitations.
- **Narrow Task Focus:** Most prior benchmarks have concentrated on a limited set of tasks, such as basic visual perception or caption generation, lacking the comprehensiveness. This focus on narrow tasks fails to capture the full range of multimodal interactions and reasoning that MLLMs are capable of, limiting our understanding of their true potential. For instance, benchmarks like BLINK [15] have primarily tested models on visual perception, not on comprehensive multimodal reasoning. Similarly, [24] evaluates MLLMs on a small number of specialized tasks across medical and industrial domains, while [25] focuses exclusively on chart-based reasoning. Although insightful, these studies remain limited to narrow application scenarios or input modalities and do not assess broader capabilities such as multi-image

understanding, temporal reasoning, or interleaved image-text comprehension.

- **Limited Input Variety:** A common limitation of existing benchmarks is their reliance on simple input formats like static image-text pairs. Real-world scenarios, however, often require models to handle more complex formats, such as multi-frame image sequences or interleaved image-text inputs. Existing methods such as MMBench [17] have primarily focused on simple multi-image and text formats, which do not reflect the intricacies involved in tasks requiring temporal understanding or dynamic multimodal reasoning.

Given these gaps, our key motivation is to develop a comprehensive assessment that systematically evaluating multiple MLLMs, exploring a broad range of tasks in multiple capabilities and covering diverse input types. Specifically, we include both large models like GPT-4V [26] and GPT-4o [27], and smaller models such as LLaVA-NeXT [28] and Phi-3-Vision [29]. This allows us to compare the performance of resource-heavy models with more efficient, smaller models to assess their suitability for real-world applications. Rather than focusing on narrow task sets, our approach spans multiple domains, including reasoning, temporal anticipation, spatial understanding, and dynamic task execution, reflecting the full complexity of multimodal interactions. We move beyond traditional image-text pairs and introduce a variety of input formats such as multi-frame sequences, interleaved image-text sequences, and more complex visual inputs. This helps to capture the challenges presented by real-world tasks, such as those found in industry and automation settings.

To systematically assess MLLMs' capabilities, we propose a comprehensive evaluation strategy that includes both general capability evaluation and real-world scenario-based task assessments. We categorize the input formats into four types to assess the models' core abilities, including single image-text pairs, multiple images with a single text prompt, multi-frame with text inputs and interleaved image-text inputs. Each task tests different capabilities such as object recognition, temporal reasoning, task execution, and multimodal comprehension. By systematically comparing the performance of both large and small MLLMs across these input formats, we provide insights into where smaller models can suffice and where large models still hold an edge. Beyond general capabilities, we apply the models to real-world scenarios, categorized by domain (*e.g.*, industry, automotive, customized captioner). Each task in these domains is evaluated for its reliance on specific capabilities, as illustrated in Fig. 1, showing the mapping between task types and general capabilities. For instance, defect detection in industry requires a strong combination of object

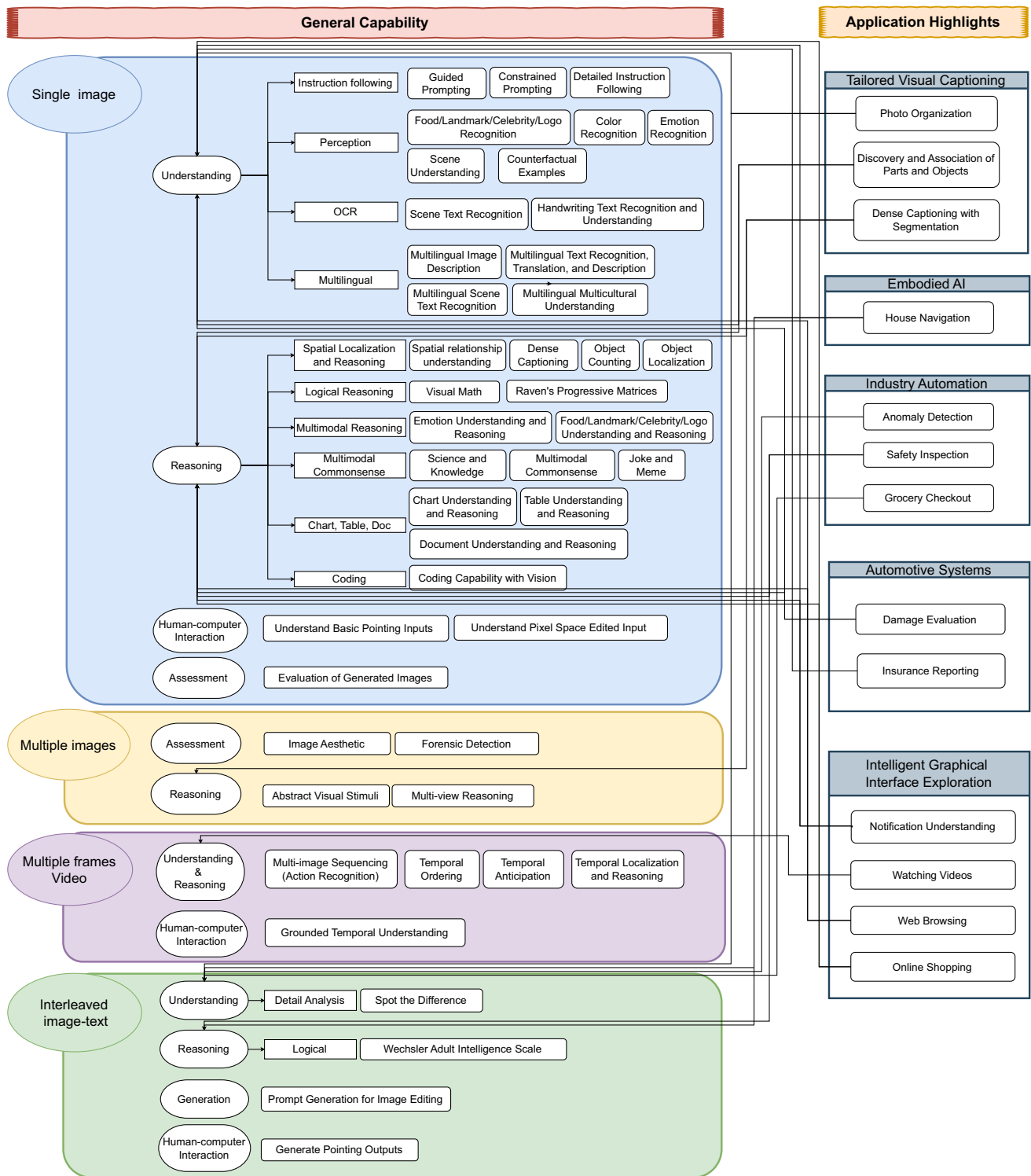


Fig. 1 Mapping of task types, input formats, and general capabilities across different domains. The figure illustrates the relationships between task categories—single image-text pair, multi-image with text, multi-frame with text, and interleaved image-text pair—and their corresponding general capabilities. Additionally, it highlights how

these tasks are applied across various real-world domains, such as industry, automotive, and customized captioning, providing a comprehensive overview of the model's applicability and performance across different input formats

recognition and anomaly detection, while grocery checkout systems hinge on object counting and contextual description generation.

Our contributions are as follows:

- **Broad Comparison of MLLMs:** We compare the performance of two representative proprietary large MLLMs (GPT-4V, GPT-4o) and two representative small MLLMs (LLaVA-NeXT, Phi-3-Vision), providing insights into where small models are viable alternatives and where large models excel.
- **A Comprehensive Evaluation for MLLMs:** We conduct a comprehensive evaluation for MLLMs across four input types, linking tasks to general capabilities, and testing their performance on a wider range of real-world scenarios than existing benchmarks.
- **Real-World Scenario Applications:** We go beyond simple tasks and evaluate how well MLLMs can be applied to real-world domains such as industry, automotive, and automated systems, providing valuable insights into their practical utility.
- **Highlighting Small Model Potential:** By comparing small and large MLLMs across these varied tasks, we identify the strengths and limitations of small models, contributing to the development of more efficient models for specific applications.

This work fills the gap left by existing benchmarks and preliminary explorations, offering a more comprehensive evaluation of MLLMs and their applicability to real-world tasks. To strengthen the validity of our conclusions, we employed three annotators to assess the accuracy of four MLLMs, determining which outputs were correct and which were incorrect. We further discuss model limitations and failure modes in Sect. 7, providing task-specific observations and hypothesized causes to complement the empirical results.

1.2 Model selection and input modality support

In this study, our primary objective is to compare the capabilities of large and small MLLMs across diverse tasks, as well as the impact of reasoning capability for completing diverse tasks. Accordingly, we selected two commonly used large-scale models (GPT-4V and GPT-4o) and two small-scale models (LLaVA-NeXT and Phi-3-Vision) as representative examples of their respective size categories. These models vary not only in size but also in their ability to perform multimodal reasoning, such as instruction following, logical inference, and temporal understanding. To systematically compare their performance across different capabilities, we construct a benchmark of 48 diverse tasks spanning perception, logical reasoning, multilingual understanding,

OCR, spatial relationship, etc. This design enables a controlled and comprehensive evaluation of how both model scale and reasoning capacity influence performance across a wide range of real-world scenarios.

In this paper we categorize 48 general tasks into four categories based on four types of input: single image-text pair, multi-image with text, multi-frame with text, and interleaved image-text pairs (optionally with multiple images). Large MLLMs can process all the input format. In contrast, during our writing process, small MLLMs, such as LLaVA-NeXT and Phi-3-Vision, cannot support multi-image or interleaved image-text very well, so we mainly test their capability in handling single image-text pair. For each task, we borrow some samples existing in previous works (e.g., [22]) and also collect a lot of new samples from internet and benchmark datasets. For the corresponding text prompts. in most tasks that were studied in previous works, we follow the prompts provided in these works. However, for a few tasks where the results are unsatisfactory (e.g., no bounding box coordinate in the generated output), we further refine the prompts by ourselves. For newly added tasks, we create entirely new prompts.

1.3 Implementation details

We evaluate four multimodal large language models (MLLMs) in our benchmark: GPT-4o, GPT-4V, LLaVA-NeXT, and Phi-3-Vision. The exact model versions and experiment timeline are specified below to ensure transparency and reproducibility. We use `gpt-4o-2024-02-15-preview` and `gpt-4-vision-preview` (2024-02-15) via the OpenAI API, and all corresponding experiments were conducted between May and June 2024. For LLaVA-NeXT, we use the `v1.6` checkpoint released in June 2024. For Phi-3-Vision, we use the initial public release checkpoint available on HuggingFace, also from June 2024. The hyperparameters of the inference setting for each model are provided in Appendix.

For each input prompt, we generate a single output per MLLM to simulate typical usage scenarios and ensure consistent evaluation conditions. LLaVA-NeXT and Phi-3-Vision use deterministic decoding (`do_sample = False`), resulting in identical outputs across runs. GPT-4V is configured with a low temperature (`temperature = 0.2`) to minimize output variance, while GPT-4o uses slightly more exploratory settings (`temperature = 0.3`, `top_p = 0.2`) to better handle reasoning tasks. Nonetheless, only one forward pass is executed per prompt for all models. Each output is independently reviewed by three human annotators to determine whether it correctly answers the given question and fulfills the task requirements.

We also provide a task-wise summary of the prompting strategy and data sources in Appendix A.3. Representative prompt–input–output examples with annotations are included in Appendices B–F to illustrate prompt design, model responses, and evaluation. The full set of prompts, outputs, and metadata (model versions and inference settings) will be released in a public repository upon acceptance to ensure transparency and reproducibility.

2 Single image-text pair inputs

In this section, we focus on tasks that utilize single image-text pair inputs, a foundational format for many visual-linguistic applications. We evaluate four MLLMs (Multi-Modal Large Language Models) across four key capabilities: understanding, reasoning, assessment, and interaction. In Sect. 2.1, we investigate how both large and small MLLMs comprehend and interpret the single image through a text prompt. We begin by examining the models' ability to generate open-ended descriptions for general visual captioning, extract and analyze textual information, comprehend and generate descriptions in multilingual scenarios, and demonstrate coding proficiency with visual inputs. In Sect. 2.2, we explore the application of MLLMs in more advanced reasoning tasks, including spatial relationship analysis, logical thinking, and information extraction from sources such as scene text, tables, charts, and documents. Finally, in Sects. 2.3 and 2.4, we delve into the models' abilities to assess generated image quality and interact with various types of human prompts.

Across four core capabilities—understanding, reasoning, assessment, and interaction with human—the models show distinct performance profiles. GPT-4o consistently achieves the highest performance across all capabilities, excelling in complex tasks like logical reasoning, multilingual comprehension, and coding. GPT-4V follows closely, showing strong results in language generation, OCR, and spatial understanding. LLaVA-NeXT demonstrates effectiveness in specific recognition tasks, such as logo and emotion identification, while Phi-3-Vision displays competitive performance in targeted recognition tasks but faces challenges with higher-level reasoning and spatial tasks. Overall, GPT-4o and GPT-4V set the benchmark for versatility, with LLaVA-NeXT and Phi-3-Vision performing well in specialized areas.

2.1 Understanding

In the “Understanding” capability, we evaluate four models on their ability to follow instructions, perceive different aspects of input, perform Optical Character Recognition

(OCR) [30], and demonstrate multilingual multimodal and coding capabilities. This evaluation covers a total of 18 tasks, with detailed task definitions, datasets, and examples provided in Online Resource.

To generate natural language descriptions across various perception scenarios, GPT-4V and GPT-4o outperform smaller MLLMs, while LLaVA-NeXT and Phi-3-Vision show competitive results in some recognition tasks, *e.g.* logo and emotion recognition. In OCR and multilingual tasks, we observe that large MLLMs excel at extracting text from complex visual content. Both GPT-4V and GPT-4o demonstrate strong comprehension and accurate generation of descriptions in multiple languages, showcasing their versatility in handling diverse linguistic contexts. For the coding task, only GPT-4o successfully generates code across all examples. Regarding the quality of generated code, GPT models and Phi-3-Vision produce similar curves and figures. Overall, GPT-4o performs the best in understanding tasks (fifteen times in the first place), followed by GPT-4V (five times in the first place), then LLaVA-NeXT (twice in the first place), and finally Phi-3-Vision (zero time in the first place).

2.1.1 Instruction following

Guided prompting

Instructions offer an effective way to define and customize outputs for various vision-language tasks. Therefore, before delving into how large and small MLLMs can be employed to comprehend and interpret the visual world, we first assess their ability to understand and follow textual instructions [31, 32]. To this end, we evaluated the ability of four models to understand and follow textual instructions by assessing their performance on 15 image-text pairs with constraints such as sentence length, word choice, and perspective. Among the models, GPT-4o demonstrates the strongest instruction-following capability, achieving near-perfect accuracy, followed by GPT-4V with slightly lower but reliable performance. LLaVA-NeXT shows moderate performance, while Phi-3-Vision struggles significantly with this task. Detailed evaluation results and task-specific examples are provided in Online Resource.

Constrained prompting

To evaluate MLLMs' ability to follow structured instructions, we conduct a driver's license information extraction task that requires returning results in a specified JSON format. The results show that large models significantly outperform smaller models, demonstrating better accuracy and adherence to the specified format. While smaller models struggle with both the extraction task and format compliance, large models prove to be more robust and reliable.

Detailed task setups, individual model performances, and examples are provided in Online Resource.

Detailed prompting

To evaluate the ability of MLLMs to generate text outputs conditioned on specific emotional contexts, we test their performance on 11 images spanning natural scenes and human activities. Large models, particularly GPT-4o, demonstrate superior performance, achieving perfect accuracy and consistently aligning with the desired emotional context. Smaller models show more limited abilities, with varying degrees of difficulty in interpreting and responding to emotional prompts. Detailed quantitative results, qualitative evaluations, and examples are provided in Online Resource.

2.1.2 Perception

Color recognition

Color recognition is a foundational skill for MLLMs, as it plays a vital role in enabling accurate understanding and interpretation of visual content. Our evaluation on 77 image-text pairs reveals that GPT-4o achieves the highest accuracy, followed by GPT-4V, while smaller models like LLaVA-NeXT and Phi-3-Vision show limited capabilities. Despite their overall strength, even GPT-4o and GPT-4V encounter difficulties when tasked with recognizing regions described by both color and spatial position, particularly in complex scenarios involving overlapping areas or regions spanning multiple states or provinces. For example, prompts like "Identify the region colored orange located directly north of Guangdong" often lead to inaccuracies, as the models tend to prioritize spatial positioning over color information. These results suggest that while large models demonstrate stronger generalization abilities, further refinement is needed to improve their integration of spatial and color cues. Detailed task setups and performance analyses are included in Online Resource.

Food recognition

Recognizing food or dishes is a challenging task [33, 34] due to the vast variety of appearances and potential occlusions. In our experiments, GPT-4o achieves the highest accuracy, correctly recognizing 14 out of 15 dishes, followed by GPT-4V with 11 out of 15. Smaller models like LLaVA-NeXT and Phi-3-Vision demonstrate limited ability, scoring 4/15 and 2/15, respectively. Despite their struggles, GPT-4o and GPT-4V show reasonable success in identifying dish names, though GPT-4V encounters difficulty with Chinese cuisine samples. LLaVA-NeXT and Phi-3-Vision fail to provide specific dish names but can identify some ingredients, with Phi-3-Vision also struggling to recognize the type of meat in certain cases. Detailed results and task examples are provided in Online Resource.

Landmark recognition

Recognizing landmarks under varying conditions, such as changes in viewpoint, lighting, occlusions, and seasonal shifts, requires models to generalize effectively and handle a wide range of visual appearances [35, 36]. Our evaluation on 28 landmark images reveals that both GPT-4V and GPT-4o achieve an accuracy of 22/28, significantly outperforming smaller models, with LLaVA-NeXT scoring 15/28 and Phi-3-Vision achieving 9/28. Despite their strengths, none of the models recognize the six samples featuring Huangguoshu Waterfall and Jiuzhaigou, even when provided with diverse angles of these landmarks. Additionally, while large models and LLaVA-NeXT generate detailed descriptions for correctly recognized samples, Phi-3-Vision tends to produce brief and less descriptive outputs. Notably, LLaVA-NeXT fails to recognize natural landscapes and architecture from China, while Phi-3-Vision demonstrates an ability to identify prominent historical landmarks in Europe and America, such as the Eiffel Tower and the Colosseum. Detailed analyses and illustrative examples are provided in Online Resource.

Celebrity recognition

Recognizing human appearance presents a significant challenge due to its inherent variability [37]. Our evaluation on 15 celebrity images reveals that smaller models perform better in recognition tasks, with LLaVA-NeXT achieving the highest accuracy at 5/15, followed by Phi-3-Vision with 1.5/15 (where 0.5 indicates partial recognition). In contrast, GPT-4V and GPT-4o struggle to identify celebrities but excel at describing individual appearances and actions in the images. While LLaVA-NeXT successfully identifies celebrities from Western countries with a 5/9 accuracy, it fails to recognize Asian celebrities. Both GPT-4V and GPT-4o explicitly acknowledge their inability to identify individuals in most cases. Phi-3-Vision, limited by RAI's policy on human faces, performs poorly overall, providing relevant activity descriptions in only 5 samples. Detailed results and examples are provided in Online Resource.

Logo recognition

Logo recognition, especially in complex scenarios, is a challenging task that tests models' ability to handle occlusions, distortions, and low resolutions. In our evaluations, GPT-4o consistently outperforms the other models, achieving the highest accuracy in both controlled and in-the-wild scenarios. In the controlled setting with a combined image of multiple logos, GPT-4o achieves 4/5 accuracy, demonstrating strong capabilities in providing elaborate descriptions of logos, though it occasionally misidentifies parts of a logo as separate entities. GPT-4V, LLaVA-NeXT, and Phi-3-Vision show limited capabilities in recognizing logos, with GPT-4V identifying only one logo accurately, while LLaVA-NeXT and Phi-3-Vision provide partial or succinct information. In the more challenging in-the-wild scenario,

GPT-4o again performs the best with 15/18 correct identifications, while the smaller models show varying levels of improvement. Detailed results and task-specific examples are provided in Online Resource.

Counterfactual examples

Inspired by [38], which highlights the tendency of LLMs to hallucinate descriptions of nonexistent objects or activities, we evaluate four models on their ability to resist such hallucinations. Using 15 randomly selected examples from MSCOCO (val2014) [39], we craft prompts combining incorrect colors with shown objects or nonexistent objects with shown colors. GPT-4V performs the best, achieving 14/15 accuracy, followed by GPT-4o with 12/15. In contrast, LLaVA-NeXT and Phi-3-Vision struggle more, with accuracies of 8/15 and 7/15, respectively, often failing to distinguish between existent and nonexistent objects. Both large and small models make similar errors in some cases, such as misidentifying ham as tomato due to color similarity. Detailed task examples and analyses are included in Online Resource.

Scene understanding

Scene understanding, a fundamental task in computer vision [40, 41], involves recognizing and interpreting elements such as road signs, weather, and objects in a scene. We evaluate four MLLMs on 12 scene images using the query "Describe the image." The results reveal a clear performance gap between large and small models. GPT-4o and GPT-4V outperform smaller models, demonstrating strong capabilities in interpreting scenes and accurately describing key elements. In contrast, LLaVA-NeXT and Phi-3-Vision struggle to identify important features like signs and their meanings, often providing incomplete or incorrect outputs. Detailed results, examples, and analysis are provided in Online Resource.

Emotion recognition

Reading emotions from facial expressions is a critical aspect of emotional quotient testing. In our evaluation of 64 facial images with varying resolutions, GPT-4V, GPT-4o, and LLaVA-NeXT demonstrate strong capabilities in identifying and interpreting emotions, often providing well-reasoned explanations based on visual cues. However, all models struggle with ambiguous emotions, such as embarrassment, failing to explicitly identify them, though some responses hint at indirect associations. In contrast, Phi-3-Vision performs poorly due to RAI blocking, generating only four correct analyses, which relied on body language cues rather than facial expressions. Detailed analyses and examples are provided in Online Resource.

2.1.3 Optical character recognition

Scene text recognition

Reading and understanding scene text is a critical vision-language task [42, 43]. Our evaluation reveals that GPT-4o performs the best, achieving an accuracy of $[9 + 0.5]/12$, followed by GPT-4V with $[8 + 0.5]/12$. Smaller models, LLaVA-NeXT and Phi-3-Vision, achieve lower accuracies of $[4 + 3 \times 0.5]/12$ and $[6 + 0.5]/12$, respectively. While large models show higher performance, particularly in complex scenarios with handwritten and printed text, all models struggle with recognizing small, blurred, or occluded text. Detailed results and additional examples are provided in Online Resource.

Handwriting text recognition and understanding

Handwriting recognition poses significant challenges for MLLMs due to the variability in writing styles, speeds, and individual differences, as well as external factors like image quality, noise, and text angles. In our evaluation of 14 handwriting images from paper sheets and whiteboards, the large MLLMs perform better than smaller models, though overall accuracy remains low. GPT-4o achieves the highest score with 5/14, followed by GPT-4V with 3/14. Among the smaller models, Phi-3-Vision recognizes only one image correctly, while LLaVA-NeXT fails to provide any correct answers. Detailed results and task-specific examples are included in Online Resource.

2.1.4 Multilingual multimodal

Multilingual image description

Multilingual capability is crucial for ensuring accessibility to worldwide users. To evaluate this, we test four MLLMs using natural images without scene text. Models are prompted to "Describe the image" in traditional Chinese, French, and Czech, and alternatively in English with a specified output language. GPT-4V excels with 100% accuracy in both scenarios. GPT-4o and LLaVA-NeXT perform well, failing only in one case each—GPT-4o mistakes a chick for a duckling, and LLaVA-NeXT repeats accurate descriptions. In contrast, Phi-3-Vision fails to generate outputs in the specified languages in both scenarios. Additionally, when prompted in Spanish to describe images in 20 different languages, only GPT-4o successfully processes both input and output in different languages. These findings underscore the superior multilingual capabilities of large models, particularly GPT-4V and GPT-4o, while smaller models exhibit significant limitations. Detailed results and further explorations are provided in Online Resource.

Multilingual text recognition, translation, and description

To evaluate the performance of MLLMs in more challenging scenarios involving images with multilingual text, we test their ability to recognize and translate text in 19 images containing text in different languages. Large MLLMs outperform smaller ones, with GPT-4o achieving

the highest accuracy at $[15 + 0.5]/19$ and GPT-4V scoring 10/19. In contrast, smaller models struggle, with LLaVA-NeXT achieving $[2 + 2 \times 0.5]/19$ and Phi-3-Vision scoring $[4 \times 0.5]/19$. Notably, GPT-4o and GPT-4V demonstrate the ability to sequentially extract text and then translate it, though with varying success rates. In a separate test requiring translation into 20 languages, GPT-4o is the only model capable of effectively completing the task, while the other models produce outputs solely in English. Detailed analyses and additional examples are included in Online Resource.

Multilingual scene text recognition

To evaluate multilingual scene text recognition, we test four MLLMs on images containing scene text in various languages, such as signs, tombstones, subway signs, and murals. Using the prompt "What are all the scene texts in the image?", the large models outperform smaller ones. GPT-4o achieves the highest accuracy at $[20 + 3 \times 0.5]/28$, followed by GPT-4V with 14/28. The smaller models, LLaVA-NeXT and Phi-3-Vision, score $[7 + 0.5]/28$ and $[4 + 4 \times 0.5]/28$, respectively. Despite their strengths, all models struggle to recognize and interpret minority languages like Arabic and Egyptian. GPT-4o demonstrates notable capabilities by identifying some minority languages and generating correct text for them, while LLaVA-NeXT produces repetitive answers in 7 cases. Detailed analyses and examples are provided in Online Resource.

Multilingual multicultural understanding

To evaluate MLLMs' multicultural understanding [44], we test their ability to describe wedding scenarios across diverse cultures using multilingual prompts. Both GPT-4o and GPT-4V demonstrate a strong grasp of cultural nuances, consistently generating reasonable multilingual descriptions for the provided wedding images. In contrast, LLaVA-NeXT struggles, achieving an accuracy of 2/9, often failing to generate the appropriate local language and sometimes producing incorrect or irrelevant outputs, such as repeating sentences or generating text unrelated to the image. Phi-3 Vision, limited by RAI's policy of blocking human faces, is unable to produce results. Detailed analyses and examples are provided in Online Resource.

2.1.5 Coding

Coding capability with vision

We further explore the models' coding abilities, particularly in writing Python code. To do this, we test them on 22 chart images from the web and the ChartInsights [16] test set, consisting of 7 images without metadata and 15 with metadata. Among the charts with metadata, GPT-4o demonstrates the best performance with an accuracy of 13/15, while LLaVA-NeXT shows the lowest accuracy at 2/15, including instances of generating non-runnable code.

GPT-4V scores 8/15, slightly behind Phi-3-Vision, which achieves 12/15. For charts without metadata, GPT-4o produces the most adaptable outputs, generating layouts closely resembling the originals that can be easily modified. In contrast, Phi-3-Vision occasionally fails to generate Python code, instead providing chart descriptions. In most cases, GPT-4o and Phi-3-Vision demonstrate better color accuracy in their outputs, while LLaVA-NeXT is the least consistent. Detailed task results and examples are included in Online Resource.

2.2 Reasoning

In this subsection, we evaluate the reasoning capabilities of MLLMs across 13 tasks, encompassing spatial location, logical thinking, reasoning from charts, tables, and documents, multimodal reasoning, knowledge, and common-sense. Among these, GPT-4o consistently demonstrates superior performance, particularly in logical reasoning and quantitative tasks involving charts and tables. It excels in document reasoning by accurately summarizing research papers and posters and extracting essential information from images containing extensive text. However, spatial location remains challenging for all models, with both large and small MLLMs struggling to identify objects specified in text prompts due to the inherent complexities of object localization in computer vision and the preprocessing strategies employed by these models. Advanced prompting techniques are required to improve performance in crowded or complex environments. Detailed findings for each task are provided in Online Resource.

2.2.1 Spatial location

Spatial relationship understanding

Understanding spatial relationships between humans and objects is a critical aspect of visual intelligence, with applications in areas such as robotics [45, 46]. Both GPT-4o and GPT-4V demonstrate superior capabilities in this task, achieving 7/10 accuracy. They successfully identify spatial relationships, such as between a frisbee and a man or a man and a car, and account for the impact of camera perspectives on perceived object sizes. However, they occasionally struggle, as in GPT-4V's failure to fully describe the relationship between a man and a dog. In contrast, smaller MLLMs like LLaVA-NeXT and Phi-3-Vision show significant limitations. LLaVA-NeXT often misinterprets actions, such as mistaking a person throwing a frisbee for catching it, and struggles with size comparisons and perspective understanding. Similarly, Phi-3-Vision frequently misidentifies positions and makes inaccurate size comparisons. These findings underscore the gap between large and small

MLLMs in understanding complex spatial relationships. Detailed analyses and examples are provided in Online Resource.

Object counting

Object counting, a foundational task in computer vision with applications in crowd analysis, traffic monitoring, inventory management, and wildlife population estimation [47–50], is critical for downstream tasks like density estimation and anomaly detection. To evaluate object counting capabilities, we test four models on 50 images from MSCOCO (val2014) [39] and FSC-147 [51], using the prompt "Count the number of X in the image." GPT-4o consistently outperforms other models, demonstrating higher accuracy and fewer large deviations from the ground truth. Notably, GPT-4o achieves 70.0% accuracy and an absolute error of 43 on MSCOCO (val2014). However, challenges such as cluttered scenes and occlusions cause errors, particularly in FSC-147's complex scenarios involving large object quantities. Smaller models, such as LLaVA-NeXT and Phi-3-Vision, perform significantly worse, frequently making errors in cluttered or diverse object categories. Detailed results and additional findings, such as models correcting inaccuracies in prompts, are discussed in Online Resource.

Object localization

Object localization [39, 52, 53] is a fundamental challenge in computer vision, requiring accurate identification of object positions using bounding box coordinates. To evaluate this task, we select 50 well-annotated images from MSCOCO (val2014) [39] and utilize the prompt "Localize X in the image using bounding boxes (x1, y1, width, height)." Initial experiments reveal that GPT-4V cannot generate bounding box coordinates with simple prompts. However, by employing a step-by-step prompt strategy, results improve significantly. The evaluation demonstrates that GPT-4o achieves the highest performance on mean precision, mean recall, and F1 score, while GPT-4V records the highest mean average precision (mAP). These results underscore the strengths of large models in object localization tasks, though their effectiveness varies depending on the evaluation metric. Detailed findings and examples are provided in Online Resource.

Dense captioning

Dense captioning [54], which involves generating natural descriptions for salient regions in an image, remains a highly challenging task in vision-language systems. To evaluate the capabilities of GPT-4V and other models in this domain, we use an instructional prompt. The results indicate significant limitations across all models, with none able to answer all questions correctly. GPT-4o achieves the highest accuracy in recognizing image resolution, correctly identifying 4 out of 12 cases, while the other models score zero. Additionally,

all models struggle with providing detailed captions, such as identifying individuals' names, positions, or professions, and none can accurately locate people within the images. These findings highlight the gap in current MLLMs' dense captioning capabilities. Detailed analyses and examples are provided in Online Resource.

2.2.2 Logical thinking

Visual math

Mathematics plays a vital role in intelligence testing, evaluating a model's capability for numerical reasoning and problem-solving. We assess four MLLMs using 15 mathematical problems ranging from basic arithmetic to advanced Olympic-level challenges. The results highlight the superior performance of GPT-4o, which achieves 11 full and 2 partial correct answers. GPT-4V follows with 8 full and 2 partial correct answers, demonstrating competency but with notable gaps in precision for complex problems. In contrast, smaller models perform significantly worse: Phi-3-Vision achieves 2 full and 1 partial correct answers, while LLaVA-NeXT manages only 1 correct answer. These findings emphasize the advanced reasoning abilities of large models, particularly GPT-4o, in solving complex mathematical tasks, while smaller models struggle with abstraction and logical problem decomposition. Detailed analyses and examples are provided in Online Resource.

Raven's progressive matrices

Raven's Progressive Matrices (RPM) [55], a widely used non-verbal IQ test, assesses abstract reasoning and problem-solving capabilities by minimizing cultural and linguistic biases [56–58]. To evaluate MLLMs, we present the entire question page as a single image, replicating the format humans encounter during IQ tests and ensuring compatibility with both large and small models. The results indicate that RPM remains a significant challenge for MLLMs. GPT-4V demonstrates the highest capability, answering 5 out of 15 samples correctly, followed closely by GPT-4o with 4 correct answers. Despite their relative success, both models struggle to fully process and integrate the complex visual-spatial patterns required to identify the missing image in 2-by-2 or 3-by-3 matrices. Conversely, smaller models like LLaVA-NeXT and Phi-3-Vision fail to solve any of the samples, highlighting fundamental limitations in their ability to recognize patterns and engage in abstract reasoning. Detailed findings and examples are provided in Online Resource.

2.2.3 Chart, table and document

Chart understanding and reasoning

We explore the performance of both large and small MLLMs in chart understanding and reasoning using the ChartInsights [16] dataset, which consists of 400 unique chart images and 17,552 ChartQA samples across 7 chart types and 10 low-level data analysis tasks. Our results indicate that large MLLMs significantly outperform smaller models. GPT-4o achieved the highest performance across all models, with an average accuracy of 65.28% and the best results in six out of ten tasks. Phi-3-Vision showed notable strengths in specific areas, excelling in three tasks, particularly in logical reasoning, though it did not match the overall performance of GPT-4V and GPT-4o. In more complex scenarios involving multiple chart types, GPT-4o maintained its top position, with a 41.43% accuracy, while smaller models like LLaVA-NeXT struggled, achieving only 23.93% accuracy. Phi-3-Vision, however, demonstrated comparable performance to large models in these complex cases, suggesting it has strong generalization capabilities. Overall, the results emphasize the superior performance of large MLLMs, particularly GPT-4o, in chart reasoning tasks, with Phi-3-Vision showing potential in certain areas. Detailed task-wise results and additional evaluations are provided in Online Resource.

Table understanding and reasoning

In our evaluation of table understanding and reasoning, we observe that large MLLMs significantly outperform small MLLMs. Among the models, GPT-4o achieved the highest performance, with strong generalization capabilities and superior results in 6 out of 10 tasks, particularly in logical reasoning tasks involving table figures. GPT-4V followed closely, performing best in 4 out of 10 tasks. In contrast, LLaVA-NeXT exhibited the weakest performance at 26.47%. When examining the impact of different textual prompt strategies, Phi-3-Vision excelled with the Yes-or-No prompt, achieving the highest accuracy of 72.50%, while GPT-4o performed well across various prompt types, including Fill-in-the-Blank, Multiple-Choice, and Error Correction. Both large and small models showed better performance with Fill-in-the-Blank, Multiple-Choice, and Yes-or-No prompts compared to Error Correction. The small models struggled with Fill-in-the-Blank and Error Correction prompts, achieving zero accuracy. Other details can be found in Online Resource.

Document understanding and reasoning

For document understanding across various types of documents, GPT-4o consistently outperforms the other models. Specifically, in the exam paper scenario, GPT-4o provides accurate answers to all questions, while GPT-4V offers a problem-solving approach but does not directly answer the questions. LLaVA-NeXT and Phi-3-Vision struggle, with LLaVA-NeXT failing to provide correct answers and Phi-3-Vision giving one incorrect answer and missing another

entirely. In the poster understanding task, GPT-4o successfully extracts and responds to the prompt, while GPT-4V and LLaVA-NeXT only extract text without addressing the question. Phi-3-Vision shows mixed results, succeeding in one case but missing another. Regarding floor plan analysis, GPT-4o and GPT-4V both achieve 2 out of 4 correct answers, with GPT-4o providing more accurate descriptions in the remaining cases. In research paper understanding, both GPT-4o and GPT-4V demonstrate a strong ability to describe the papers and identify key contributions, although they occasionally make mistakes, such as incorrect dataset names. Overall, GPT-4o proves to be the best performer across these tasks. More detailed findings and task breakdowns are available in Online Resource.

2.2.4 Multimodal reasoning

Emotion understanding and reasoning

In our assessment of visual sentiment analysis, we observe that the two large models and LLaVA-NeXT perform exceptionally well, each achieving a perfect accuracy of 100%. These models demonstrate strong abilities in interpreting human emotions evoked by visual content, effectively recognizing sentiments like contentment, anger, awe, and fear. This performance highlights their capability to understand both the semantic content and image style. In contrast, Phi-3-Vision achieves a lower accuracy of 9/15, with its success primarily in natural scenes and outdoor events, where more visual cues are available to infer emotions. Detailed results and visual examples can be found in Online Resource.

Food/landmark/celebrity/logo understanding and reasoning

This task evaluates the multimodal reasoning capabilities of MLLMs by challenging their ability to integrate cultural knowledge, textual comprehension, and world knowledge across diverse categories, including food, celebrities, landmarks, and logos. Among the four models, GPT-4o demonstrates the strongest performance, correctly answering 25 out of 35 questions, showcasing its advanced reasoning and knowledge integration capabilities. GPT-4V and LLaVA-NeXT follow, each with 17 correct answers, reflecting moderate competency, while Phi-3-Vision lags behind with only 6 correct responses. These results highlight significant disparities in multimodal reasoning, with GPT-4o excelling in understanding complex visual and contextual information. Additional details and task-specific findings are provided in Online Resource.

2.2.5 Multimodal knowledge and commonsense

Science and knowledge

We evaluate the reasoning capabilities of four MLLMs in tasks requiring scientific knowledge across topics such as geography, physics, chemistry, biology, and earth science. Among the models, GPT-4V and GPT-4o each achieve 4 correct answers, while LLaVA-NeXT and Phi-3-Vision demonstrate moderate performance with 3 fully correct answers and one partially correct response each. Both GPT models excel in chemistry-related questions, with GPT-4o also performing well in biology tasks, but they struggle with geography. The smaller models show general competence in science reasoning tasks but often miss finer details. Notably, all models successfully adopt a tutorial approach when presented with specific instructional prompts. Additional details and examples are provided in Online Resource.

Joke and Meme

Understanding jokes and memes is a challenging yet essential task for MLLMs, as it requires advanced comprehension of cultural references, visual-text interactions, and nuanced humor. Among the four models evaluated, GPT-4o and GPT-4V demonstrate strong performance, failing in only one of the 13 examples and effectively capturing the humor and underlying messages. In contrast, the smaller models, LLaVA-NeXT and Phi-3-Vision, struggle significantly, achieving accuracy rates of 6/13 and 4/13, respectively. These results highlight the superior capability of larger models in understanding humor and cultural nuances compared to their smaller counterparts. Additional details and specific examples are provided in Online Resource.

Multimodal commonsense

We evaluate the multimodal commonsense reasoning abilities of four MLLMs in two distinct scenarios: reasoning with bounding boxes as visual prompts and inferring actions directly from visual content. In the bounding box reasoning task, where models are provided with labeled images and bounding boxes, GPT-4V achieves the highest performance, with an accuracy of $[10 + 0.5]/12$. GPT-4o follows closely with $[8 + 2 \times 0.5]/12$, with deductions for missing actions and incorrect references to the [person n]. LLaVA-NeXT and Phi-3-Vision show moderate performance, with LLaVA-NeXT scoring $[3 + 6 \times 0.5]/12$ and Phi-3-Vision scoring $[6 + 2 \times 0.5]/12$. In the action inference task, GPT-4V and GPT-4o excel with perfect scores of 3/3, demonstrating strong reasoning from visual cues. LLaVA-NeXT scores 2.5/3 due to a repeated answer, and Phi-3-Vision also achieves 3/3 but with fewer details in its responses. These results highlight the superior reasoning capabilities of the large models, with smaller models facing challenges in both recognizing and inferring actions from visual content. More detailed task breakdowns and performance metrics are provided in Online Resource.

2.3 Assessment

Except “undersatnding” and “reasoning”, we draw connections to another prominent area of multimodal assessment research: visual synthesis. By examining image generation, we explore how MLLMs can contribute to this field through comprehensive evaluation. We find that GPT-4o achieves the best performance, though the gap in capability compared to the former two is not as significant.

2.3.1 Evaluation of generated images

Building on the findings from Sect. 2.2.4, we extend the evaluation of image assessment to include smaller models, *i.e.*, LLaVA-NeXT and Phi-3-Vision, focusing on their ability to assess alignment between generated images and text-to-image generation prompts. In this task, both GPT-4o and GPT-4V outperform the smaller models, achieving scores of 12/26 and 10/26, respectively, compared to 7/26 for LLaVA-NeXT and 8/26 for Phi-3-Vision. GPT-4o demonstrates the best overall performance, accurately identifying high-quality matches with prompts. Notably, all models provide explanations for their ratings, offering insights that could be utilized to improve the image generation process. Additional details, including evaluation criteria and specific examples, are available in Online Resource.

2.4 Human-computer interaction

Pointing to a specific spatial location is a critical capability in human-computer interaction with multimodal systems, particularly in visually grounded dialogues. We conduct experiments using both basic pointing and pixel-based pointing as input methods. The results show that GPT-4V and GPT-4o demonstrate a better understanding of visual pointers directly drawn on images compared to smaller MLLMs. Notably, GPT-4o achieves 100

2.4.1 Understand basic pointing inputs

In our evaluation of MLLMs’ ability to understand grounded spatial inputs, we find that larger models, such as GPT-4o and GPT-4V, exhibit superior performance compared to smaller models like LLaVA-NeXT and Phi-3-Vision. When using overlaid visual markers (*e.g.*, circles, boxes, and irregular shapes) to identify objects within complex images, GPT-4o achieves the highest accuracy with 6/8 correct cases, followed by GPT-4V with 5/8. In contrast, LLaVA-NeXT and Phi-3-Vision struggle significantly, with only 1/8 accurate cases each. These results underscore the larger models’ advanced visual reasoning capabilities and ability to interpret visual markers in densely populated images.

For text-based region coordinates, none of the models accurately localize objects, confirming their limitations in handling coordinate-based localization tasks. This highlights a critical gap in their ability to translate numerical spatial information into accurate object identification, a capability that requires further development. Additional details and experimental setups are provided in Online Resource.

2.4.2 Understand pixel space edited input

We evaluate the capabilities of large and small MLLMs in leveraging pixel-space edited prompting across scenarios such as math problems, charts, and document understanding. GPT-4o demonstrates consistently strong performance, achieving perfect accuracy in both math problems and chart/document understanding tasks, highlighting its advanced visual reasoning capabilities. GPT-4V performs well but with occasional errors, while smaller models like LLaVA-NeXT and Phi-3-Vision show significant limitations, struggling to interpret visual prompts effectively. These results emphasize the superior ability of larger models to handle complex visual reasoning tasks. Further details and experimental setups are provided in Online Resource.

3 Multi-image with text inputs

In this section, we assess the performance of large MLLMs across two critical capabilities: reasoning and assessment, when processing multiple images. Smaller models are excluded due to their limitations in handling multi-image inputs. In Sect. 3.1, we examine the models' ability to interpret abstract visual stimuli, relationships between images, and symbolic reasoning, which are key components of human intelligence. In Sect. 3.2, we explore their capacity to analyze visual sentiment and detect manipulated or fake content, essential skills in addressing modern digital challenges. GPT-4o demonstrates consistently higher accuracy than GPT-4V in all tested scenarios, reflecting its superior ability to reason over abstract inputs and perform complex assessments. These findings highlight the advantages of larger models in tackling sophisticated multimodal tasks.

3.1 Reasoning

This subsection evaluates the ability of large MLLMs to perform complex reasoning tasks, including abstract visual stimuli interpretation and multi-view reasoning. For abstract visual stimuli interpretation, GPT-4o excels at recognizing patterns and geometric relationships in tasks like tangrams and ASCII art, outperforming GPT-4V, which shows notable

weaknesses in spatial reasoning. Smaller models struggle with these abstract tasks.

In multi-view reasoning, both GPT-4o and GPT-4V demonstrate moderate spatial reasoning abilities, correctly identifying motion direction in some cases, but showing significant challenges with 3D spatial transformations and consistent biases, especially in predicting leftward motion. These results highlight the strengths and limitations of large MLLMs in visual reasoning tasks.

3.1.1 Abstract visual stimuli

Humans excel at interpreting complex and abstract visual patterns, and this section evaluates the capability of large MLLMs to perform similar tasks through tangram and ASCII art analysis. GPT-4o demonstrates superior abstract reasoning, achieving significantly higher accuracy than GPT-4V in both tasks, reflecting its stronger ability to recognize patterns, geometric relationships, and abstract representations. In contrast, GPT-4V shows notable weaknesses in spatial reasoning and interpreting unconventional visual formats. The smaller MLLMs struggle significantly, failing to generalize from conventional image tasks to abstract visual reasoning challenges. Detailed analyses and qualitative observations are presented in the Online Resource.

3.1.2 Multi-view reasoning

Multi-view reasoning assesses the spatial reasoning capabilities of MLLMs by evaluating their ability to deduce relative camera motion from two images of the same object taken from different viewpoints. Using 20 samples from the BLINK dataset [15], both GPT-4o and GPT-4V correctly identify the motion direction in 7 cases, demonstrating moderate spatial reasoning abilities. However, they struggle significantly with leftward motion, and GPT-4V exhibits a consistent bias, predicting rightward motion in all cases. These results reveal notable gaps in their understanding of 3D spatial transformations. Additional analysis and observations are provided in the Online Resource.

3.2 Assessment

This subsection evaluates the ability of large MLLMs, specifically GPT-4V and GPT-4o, in tasks related to image aesthetics and forensic detection. In the assessment of image aesthetics, both models demonstrate strong performance in aligning with human subjective judgments, offering detailed explanations that reflect their understanding of concepts like beauty and design based on societal standards. However, GPT-4o shows a slight edge in this area.

For forensic detection, which is essential for identifying AI-generated content, GPT-4o outperforms GPT-4V, correctly identifying more real images (15 out of 20) compared to GPT-4V's 9 correct classifications. This highlights GPT-4o's stronger forensic detection abilities, although there is still room for improvement across different models. These findings underscore the models' capabilities in both subjective evaluation and distinguishing real from fake content.

3.2.1 Image aesthetics

Building on the evaluation of visual sentiment in Sect. 2.2.4, we further assess the two large models, GPT-4V and GPT-4o, on their ability to align with human subjective judgments, such as image aesthetics. Both models demonstrate a strong capacity to evaluate aesthetics based on societal standards, offering detailed explanations for their ratings. These results highlight their effectiveness in understanding subjective concepts like beauty and design. Further details are provided in the Online Resource.

3.2.2 Forensic detection

Recent advancements in generative AI have raised concerns about its potential misuse, necessitating reliable methods for detecting fake content. To evaluate the capabilities of Multimodal Large Language Models (MLLMs) in this domain, we assess their performance in distinguishing real images from AI-generated ones using datasets sourced from BLINK [15]. GPT-4o outperforms GPT-4V, correctly identifying 15 out of 20 real images compared to GPT-4V's 9 correct classifications. These findings highlight GPT-4o's stronger forensic detection abilities, though further refinement is needed to ensure robust performance across various MLLMs. Additional analysis and details are provided in the Online Resource.

4 Multi-frame with text inputs

In this section, we evaluate the ability of large MLLMs, specifically GPT-4o and GPT-4V, to understand temporal and video content by processing multiple frames as input. This evaluation is crucial, as real-world events unfold over time, requiring AI systems to interpret dynamic processes accurately. In Sect. 4.2, we examine key capabilities such as temporal anticipation, temporal ordering, temporal localization, and temporal reasoning. These capabilities serve as important indicators of a model's ability to interpret event sequences, forecast future events, and analyze activities across time, even when limited to static image inputs. GPT-4o consistently outperforms GPT-4V in these tasks,

although both large models demonstrate a comparable level of understanding for temporal sequences and video content.

Section 4.3 focuses on grounded temporal understanding, further investigating how these models interact with temporal inputs to better understand real-world events. The ability to analyze these temporal sequences is a key aspect of improving the versatility and practical applications of these models. In contrast, smaller models like LLaVA-NeXT and Phi-3-Vision struggle significantly with tasks requiring multi-frame input, demonstrating the limitations of their architectures in handling complex temporal analysis.

In summary, GPT-4o excels across temporal reasoning tasks, while GPT-4V shows promise in some areas but occasionally lags behind GPT-4o. The smaller models, LLaVA-NeXT and Phi-3-Vision, show considerable difficulty with multi-frame inputs and temporal reasoning tasks. Further details and analysis are provided in the Online Resource.

4.1 Multi-image sequencing understanding

We demonstrate the capability of MLLMs in understanding multi-image sequencing through the "Action Recognition" task. To assess the performance of smaller MLLMs, we utilize samples featuring either action sequences or multiple frames within a single image. Our observations reveal that the two smaller models perform significantly worse compared to the larger models.

4.1.1 Action recognition

We assess the performance of small models in action recognition by analyzing ten samples containing either action sequences or multiple frames within a single image. GPT-4o leads with the highest accuracy of 8/10, followed by GPT-4V at 6/10. LLaVA-NeXT and Phi-3-Vision score lower, with LLaVA-NeXT achieving 3.5/10 and Phi-3-Vision scoring 1.5/10. Despite these differences, all models demonstrate strength in describing background details and human features. Further details of the evaluation and specific failure cases are provided in the Online Resource.

4.2 Video understanding

In video understanding, only the large MLLMs, GPT-4o and GPT-4V, are capable of processing image sequences. Evaluating their performance on three video tasks—temporal ordering, temporal anticipation, and temporal localization and reasoning—we find that GPT-4o consistently outperforms GPT-4V, ranking first in all tasks. GPT-4V ranks first in one task, matching GPT-4o's performance in that instance.

4.2.1 Temporal ordering

Temporal ordering is a crucial task for evaluating MLLMs' temporal commonsense capabilities. In this task, GPT-4o significantly outperforms GPT-4V, achieving 8/12 accuracy compared to GPT-4V's 2/12. This highlights GPT-4o's stronger ability to understand the logical flow of events and construct meaningful narratives from unordered images. The task involves sequential steps in recipes with 3–15 images. Further details are provided in the Online Resource.

4.2.2 Temporal anticipation

We demonstrate GPT-4V's and GPT-4o's ability to anticipate future events based on initial frames. In our evaluation, both models accurately predict subsequent steps in 8 cases, covering scenarios such as ball sports and cooking recipes. The models show a strong understanding of current progress and effectively anticipate future actions, showcasing their proficiency in interpreting and predicting complex, multi-step processes. These results highlight the models' capabilities in both short-term and long-term temporal anticipation. Further details are provided in Online Resource.

4.2.3 Temporal localization and reasoning

We assess the temporal localization and reasoning abilities of GPT-4o and GPT-4V using 8 cases involving both long-term and short-term sequences. GPT-4o demonstrates superior performance, accurately identifying specific events in 7 out of 8 cases, such as pinpointing when a player strikes a ball or jumps over a bar. In contrast, GPT-4V struggles significantly, failing to correctly identify the frame in 7 of the 8 cases. These findings highlight GPT-4o's advanced reasoning capabilities in temporal localization, while GPT-4V's performance suggests limitations in tracking fast-moving actions across frames. More detailed analysis can be found in the Online Resource.

4.3 Human–computer interaction

4.3.1 Grounded temporal understanding

We assess the grounded temporal understanding abilities of GPT-4o and GPT-4V by providing them with 10 sets of video frames featuring specific regions of interest marked by colored circles. GPT-4o outperforms GPT-4V, correctly identifying temporal context in 8 out of 10 cases, demonstrating a more refined understanding of interactions across frames. GPT-4V achieves a lower accuracy of 5 out of 10, often struggling to identify actions associated with specific individuals, particularly in visually cluttered

scenarios. These results suggest that GPT-4o exhibits stronger grounded temporal understanding, while GPT-4V faces challenges in complex background contexts. Detailed findings can be found in the Online Resource.

5 Interleaved image-text inputs

In this section, we evaluate MLLMs using interleaved image-text pairs to test their interpretative and reasoning capabilities across diverse scenarios, including understanding detailed information, logical reasoning, and prompt generation. While GPT-4o demonstrates strong performance in abstract reasoning tasks, such as IQ test questions, it struggles with more complex spatial reasoning tasks, including associating coordinates with visual content and generating accurate visual pointing outputs. Similarly, GPT-4V exhibits limitations in spatial localization, highlighting the need for further model refinement and improved prompt designs. These findings suggest that while MLLMs show promise in abstract reasoning, enhancements in handling fine-grained spatial relationships are essential. Additional details are provided in the Online Resource.

5.1 Understanding

In this subsection, we assess the ability of MLLMs to interpret and understand complex visual stimuli, which forms the foundation for tasks requiring perception. While GPT-4o and GPT-4V excel at interpreting images with clear instructions, they face challenges with more abstract and subtle differences, such as in the "Spot the Difference" task. Smaller models like LLaVA-NeXT and Phi-3-Vision struggle more with both abstract reasoning and perception tasks, highlighting the superior capabilities of larger models in understanding complex visual patterns and relationships. This sets the stage for reasoning, where we explore deeper cognitive tasks.

5.1.1 Detail analysis

Spot the difference

In this section, we evaluate the performance of MLLMs in a "Spot the Difference" task, testing their ability to detect and explain differences between pairs of visually similar images. In the first experiment, models were prompted to compare two images and identify five differences. GPT-4o performed the best, with accuracy ranging from 0.5/5 to 4/5, though it struggled to identify all five differences in any of the 10 samples. While it could detect some differences, it often failed to explain them accurately. GPT-4V achieved accuracy between 0.5/5 and 3.5/5, showing limited success

in identifying differences. LLaVA-NeXT and Phi-3-Vision performed significantly worse, scoring between 0/5 and 2.5/5, with difficulties in identifying subtle changes such as missing objects and a tendency to only recognize color differences.

In the second experiment, we tested two variations of the task. In the first, where the prompt did not specify the number of differences, performance improved in only 1 out of 5 cases. In the second variation, specifying that five differences needed to be identified, GPT-4o and GPT-4V showed improved performance in 3 out of 5 cases. However, performance dropped in the remaining cases, suggesting that clear input prompts can enhance performance, but ambiguity in the task setup negatively impacts outcomes. These results indicate that while large MLLMs demonstrate some capability in identifying image differences, they face challenges with complex or subtle discrepancies and perform best when prompts are precise and explicit. Further details and analysis are provided in Online Resource.

5.2 Reasoning

Building on the understanding capabilities evaluated earlier, this subsection shifts focus to the models' abstract reasoning abilities, where tasks like the WAIS test the logical thinking and spatial reasoning of the models. Here, GPT-4o and GPT-4V perform well, though their difficulty with spatial reasoning tasks emphasizes that even large models face challenges with more complex, multi-dimensional problems. Smaller models such as LLaVA-NeXT and Phi-3-Vision show clear weaknesses, underscoring the need for further advancements in spatial and abstract reasoning. The generation subsection builds on this, exploring how well these models generate solutions to tasks using prompts.

5.2.1 Logical thinking

Wechsler adult intelligence scale

We evaluate the performance of four MLLMs in abstract reasoning tasks using the Wechsler Adult Intelligence Scale (WAIS) [59], a widely recognized benchmark for IQ testing. GPT-4V and GPT-4o demonstrate strong performance, correctly answering 9 out of 10 questions. In contrast, LLaVA-NeXT and Phi-3-Vision significantly underperform, each scoring only 1 correct answer out of 6. Despite the overall success of the larger models, both GPT-4V and GPT-4o struggle with tasks that involve spatial decomposition, such as identifying the correct split figure, indicating limitations in their ability to perform abstract reasoning tasks that require understanding the relationship between the whole and its parts. These findings highlight the advantages of larger models in complex reasoning tasks, but also point to

areas where further improvements are needed. Additional details and in-depth analysis can be found in the Online Resource.

5.3 Generation

Following the evaluation of both understanding and reasoning, we focus on how well MLLMs generate solutions, specifically in tasks requiring image editing and prompt optimization. While GPT-4o and GPT-4V demonstrate strong capabilities in generating and refining image-editing prompts, smaller models like LLaVA-NeXT and Phi-3-Vision struggle to meet the task requirements effectively. These findings further highlight the practical strengths of larger models in handling more complex tasks that require a combination of understanding, reasoning, and generation skills.

5.3.1 Prompt generation for image editing

In this section, we evaluate the ability of four MLLMs to assist in image editing by refining or generating prompts tailored for the task. The large models, GPT-4V and GPT-4o, demonstrate strong capabilities in optimizing image editing prompts, leading to more visually appealing final outcomes. They are able to refine the text prompt iteratively, considering both the original and edited images, and thereby improving the quality of the edited content. However, LLaVA-NeXT and Phi-3-Vision struggle with prompt optimization, often generating overly long prompts, contrary to the instruction to keep them concise. These findings highlight the advantages of larger models in providing more accurate and effective assistance for image editing tasks.

We also explore how these models can enhance image editing through iterative refinement of the editing prompt. GPT-4V and GPT-4o excel in this task, allowing for continuous adjustments to improve the image's quality. This process offers users greater control and creative flexibility in editing tasks. However, LLaVA-NeXT and Phi-3-Vision have limitations in prompt optimization and struggle to follow the task's specific requirements. This reinforces the conclusion that larger models offer more effective solutions for complex image editing processes. Further details on the experiment and model performance can be found in the Online Resource.

5.4 Human-computer interaction

5.4.1 Generate pointing outputs

Building on previous work in human-interaction applications, we assess the performance of GPT-4V and GPT-4o

in generating visual pointing outputs. The models are tested on predicting region coordinates based on text descriptions and reference images. Our findings show that both models exhibit limited accuracy in localizing objects based on the provided prompts. Despite performing well in other tasks, they demonstrate only a coarse understanding of spatial locations, and their generated outputs often do not align precisely with the expected coordinates. This suggests that while the models are capable of performing basic spatial reasoning, further refinement is needed to enhance their accuracy in interpreting spatial relationships. Further details and comprehensive analysis of the experiment and model performance are provided in the Online Resource.

6 Application

In this section, we examine a broad array of high-impact applications facilitated by the advanced capabilities of both large and small Multi-Modal Language Models (MLLMs). These applications showcase how MLLMs can address multi-faceted challenges in various domains, enhancing user interaction, understanding, and decision-making. Through our evaluation of different real-world scenarios, we observe that GPT-4V achieves remarkable accuracy, with 100% performance in 7 out of 11 tasks. This highlights its potential for effectively integrating visual, linguistic, and contextual information to solve complex problems. Additionally, Phi-3-Vision excels in safety inspection tasks, while LLaVA-Vision demonstrates strong performance in online shopping applications when aided by GPT-4V's text prompts. Overall, this section underscores the versatility and adaptability of MLLMs, emphasizing their significant role in transforming AI applications across diverse, real-world contexts. Further details on the models' performance and specific tasks can be found in the Online Resource.

6.1 Tailored visual captioning

6.1.1 Photo organization

In this scenario, we assess the ability of large MLLMs, specifically GPT-4o and GPT-4V, to enhance a family photo album by generating personalized captions that identify each family member within the images. GPT-4V demonstrates the capacity to accurately recognize individuals, pets, and even objects in some cases, achieving 4 out of 8 correct captions. However, when family members are missing from the photo, GPT-4V often fails to identify or match names, reflecting its limitations in dealing with incomplete data. GPT-4o, on the other hand, generates descriptions using key-value pairs, detailing family members' features

and positions in the photo. While GPT-4o produces more consistent captions, it lacks the accuracy in naming family members compared to GPT-4V. These results emphasize the strengths and weaknesses of each model in providing personalized, detailed photo captions. Further details on the models' performance can be found in the Online Resource.

6.1.2 Discovery and association of parts and objects

In this task, we assess the ability of large MLLMs, particularly GPT-4o and GPT-4V, to discover and associate object parts in complex images. The models were tested on two types of input: one where they localize an object part based on its semantic meaning, and another where they associate parts segmented by SAM [60]. While both models demonstrate some success, their performance is limited. For semantic localization, GPT-4V and GPT-4o show significant struggles, with GPT-4V failing to provide accurate coordinates in some cases and GPT-4o exhibiting zero mIoU in several others. On the other hand, in associating segmented object parts, GPT-4o outperforms GPT-4V, achieving 4 out of 6 correct associations. These findings suggest that while both models can handle simpler cases involving part-to-whole associations, their ability to perform accurate object localization and handle more complex spatial relationships is limited. Further details on the experiment and model performance can be found in the Online Resource.

6.1.3 Dense captioning with segmentation

Dense captioning is a key application for assessing a model's compositional understanding and contextual awareness by utilizing powerful segmentation models [60, 61]. In this task, we evaluate the ability of GPT-4V and GPT-4o to generate dense captions by incorporating object cut-outs created using SAM [60]. The models were tasked with generating detailed descriptions for object parts while referencing the context image. Both GPT-4V and GPT-4o successfully recognize and describe the object parts, generating intricate captions that reference the surrounding context. However, GPT-4o demonstrates superior performance by providing more precise and accurate descriptions, effectively integrating semantic information from the entire image. For example, GPT-4o successfully identifies cucumber and tomato parts as components of a camera, while GPT-4V fails to make this distinction. These findings highlight GPT-4o's ability to synthesize contextual information more effectively than GPT-4V. Further details on the experiment and model performance can be found in the Online Resource.

6.2 Industry automation

6.2.1 Anomaly detection

Anomaly detection is a critical component of manufacturing processes aimed at ensuring product quality. Timely identification and remediation of faults or defects are crucial for minimizing operational and quality-related costs. This application focuses on MLLMs' visual reasoning abilities, specifically its spatial reasoning and pattern recognition capabilities. We assess the performance of MLLMs in the context of anomaly detection, specifically in identifying defects in manufacturing processes using the MVTec AD dataset [62]. The results show that while large MLLMs, such as GPT-4o and GPT-4V, outperform smaller models, they still fall short of achieving perfect accuracy. GPT-4o leads with an accuracy of 5/6, while GPT-4V follows with 4/6. Smaller models, LLaVA-NeXT and Phi-3-Vision, struggle with anomaly detection, achieving only 1.5/6 and 0.5/6, respectively. Incorporating a reference image of a defect-free product in the input helps improve the performance of the large models, as evidenced by the increase in GPT-4V's accuracy. However, this approach does not benefit the smaller models. These findings highlight the potential of large MLLMs in industrial anomaly detection, but also emphasize the need for further improvements to achieve more consistent and accurate defect identification. Further details on the experiment and model performance can be found in the Online Resource.

6.2.2 Safety inspection

We explore the role of Personal Protective Equipment (PPE) counting in enhancing safety inspections, particularly within high-risk environments like construction sites. Proper use of PPE, such as helmets, is critical to ensuring worker safety. In this evaluation, we examine the capabilities of MLLMs in detecting and quantifying instances of safety violations, focusing on helmet usage by motorcycle drivers. The results demonstrate that Phi-3-Vision achieves perfect accuracy in detecting and counting individuals wearing helmets. In contrast, GPT-4V, GPT-4o, and LLaVA-NeXT all achieve a lower accuracy of 2/5, with successes primarily in simpler cases involving fewer people and fewer occlusions. To address these challenges, we introduced a two-step approach involving a person detector followed by the large MLLMs, which significantly improved GPT-4o's performance to 3/3. However, GPT-4V still struggled in one instance, highlighting areas for further improvement. Further details on the experiment and model performance can be found in the Online Resource.

6.2.3 Grocery checkout

Self-checkout machines are now widely used in supermarket chain stores such as Wegmans and Walmart, offering a quicker alternative to traditional cashier-assisted checkout and reducing staff workload. Despite their benefits, these machines often pose challenges for customers, who must locate product barcodes or manually input item codes for fresh produce like apples. This process can be particularly frustrating for first-time users or those unfamiliar with the system.

The task highlights the models' proficiency in perception and reasoning, including handling visual and textual inputs as well as spatial relationships. In the baseline evaluation, none of the models were able to accurately identify all the items in a shopping basket. GPT-4o achieved the highest accuracy with 1/5, followed by GPT-4V with 2/5, Phi-3-Vision with 1/5, and LLaVA-NeXT with 0/5. However, when prompted with additional catalog images of the grocery items from a retail website, GPT-4o's performance improved to 4/5, while GPT-4V remained at 2/5. The enhancement did not benefit the smaller models, as they could not process multiple images as input. These findings suggest that while large MLLMs show potential for improving the self-checkout experience, challenges remain in managing multiple images and spatial reasoning tasks. Further details on the experiment and model performance can be found in the Online Resource.

6.3 Automotive systems

6.3.1 Damage evaluation

This application evaluates the performance of large and small MLLMs in understanding and reasoning about car damage through image analysis. The task assesses the models' capabilities in perception and spatial/multimodal reasoning by prompting them to describe the damage in a car image. Among the models tested, GPT-4o achieves the highest performance, correctly identifying and localizing all instances of damage in 13 out of 13 cases, followed by GPT-4V with 11 out of 13. LLaVA-NeXT and Phi-3-Vision perform moderately well, each achieving 10 out of 13 correct responses. GPT-4o excels in providing comprehensive and structured descriptions, while GPT-4V shows variability in damage localization. The smaller models, while successful in some instances, occasionally miss critical details or misinterpret the damage location due to factors such as light reflections or incomplete views. These findings demonstrate that large MLLMs, particularly GPT-4o, perform exceptionally well in identifying and describing car damage, whereas smaller models still face challenges in accurately interpreting visual

content. Further details on the experiment and model performance are provided in Online Resource.

6.3.2 Insurance reporting

Building on the success in damage evaluation, we adjust our prompt to instruct both large and small MLLMs to identify the make, model, and license plate of the damaged vehicle depicted in the image and return the extracted information in JSON format. This application evaluates the MLLMs' abilities in understanding (OCR, perception, and instruction following). In this evaluation, GPT-4o outperforms the other models, achieving the lowest number of N/A answers (31) and incorrect responses (5). While none of the models successfully predicted the estimated repair cost, they were able to return relevant vehicle details in most cases. GPT-4V produced 47 N/A answers and 6 incorrect responses, while LLaVA-NeXT and Phi-3-Vision generated 36 and 27 N/A answers, respectively, and produced more incorrect responses. These results demonstrate that GPT-4o excels in understanding and processing vehicle-related information from images, even under constraints such as occlusion. Further details on the experiment and model performance can be found in Online Resource.

6.4 Embodied AI

In this part, we explore the exciting applications and implications of GPT-4V and GPT-4o for embodied agent, focusing on how these models can bridge the gap between multimodal understanding of static inputs and physical interaction with dynamic environments. To provide a concrete example, we consider a scenario where GPT-4o and GPT-4V take on the role of a home robot.

6.5 House navigation

In this setting, we examine how these models navigate task-oriented challenges throughout the house, showcasing their ability to comprehend and effectively interact with their surroundings to accomplish specific tasks. Using a virtual house tour to simulate real-world scenarios, we evaluate how these models perform in task-oriented navigation. Both GPT-4V and GPT-4o demonstrate effective performance, successfully completing the assigned chores. GPT-4o, however, delivers more succinct and efficient responses, showcasing its stronger capability in handling interactive tasks in dynamic environments. These findings underscore the potential for these models to support embodied agents in real-world applications, though further improvements in task execution and decision-making are necessary for optimal performance.

6.6 Intelligent graphical interface exploration

6.6.1 Notification understanding

Notification handling is a straightforward scenario requiring an AI agent to navigate and interpret content. This task evaluates the ability of four MLLMs to handle notifications across different devices, focusing on their comprehension, multimodal reasoning, and decision-making skills. GPT-4o stands out as the best performer, achieving 100% accuracy, while GPT-4V follows closely behind with 8 out of 9 correct responses. LLaVA-NeXT performs moderately with 4 out of 9, and Phi-3-Vision shows significant limitations, achieving only 1 out of 9. GPT-4o excels in accurately locating the app associated with the notification, while Phi-3-Vision struggles, particularly with determining the correct next step. These results underscore GPT-4o's robustness in handling notifications, whereas smaller models have considerable difficulty in navigating this task. Further details on the experiment and model performance can be found in the Online Resource.

6.6.2 Web browsing

In this work, we evaluate the performance of four MLLMs—GPT-4V, GPT-4o, LLaVA-NeXT, and Phi-3-Vision—on a task-oriented GUI navigation task. Given a screenshot, a list of possible actions, and an end goal, the models are tasked with predicting the next steps to complete tasks such as installing Anaconda. GPT-4V and GPT-4o successfully navigate the GUI to retrieve the required instructions, demonstrating their potential for automating the process without human intervention. In contrast, LLaVA-NeXT and Phi-3-Vision struggle to generate the precise next steps, highlighting a significant performance gap. Detailed comparisons and the experimental setup are provided in Online Resource.

6.6.3 Online shopping

Navigating a smartphone GUI for online shopping is a promising application for evaluating MLLMs' capabilities in spatial reasoning and action planning within user interfaces (UIs). In this task, we assess the performance of four models—GPT-4V, GPT-4o, LLaVA-NeXT, and Phi-3-Vision—by providing them with a screenshot of the current phone screen, a list of possible actions, and a prompt to predict the subsequent steps required to shop for an item within a specified budget range. The evaluation shows that GPT-4V and GPT-4o can successfully navigate the GUI to complete the shopping task. Additionally, LLaVA-NeXT, when guided by GPT-4V's prompts, is able to complete

the task, whereas Phi-3-Vision struggles to do so. Detailed experimental results and analysis are provided in the Online Resource.

6.6.4 Watching videos

Videos are a vital source of online information, and evaluating large MLLMs' capabilities in understanding video content involves assessing their temporal reasoning and multi-frame understanding. To test these abilities, we collect eight different image sequences from YouTube and evaluate the performance of GPT-4V and GPT-4o in generating insightful descriptions of the video content. As shown in Sect. 4.1, both models perform well in describing video content when aided by text descriptions in the images. However, they both struggle to identify humorous elements in certain cases, even with textual assistance. Additionally, GPT-4o outperforms GPT-4V in handling videos with complex content, such as cooking videos, by successfully generating descriptions for sequences with around 15 images, where GPT-4V fails to produce any output. Detailed findings and analysis are provided in the Online Resource.

7 Statistical analysis of model capabilities

To provide a comprehensive overview of model capabilities, we summarize the performance of four multimodal large language models across all evaluated tasks in Tables 1 and 2. These tasks are grouped into two categories: general capabilities, which cover foundational skills such as visual grounding, color recognition, and constrained prompting, and application tasks, which assess models in more practical and diverse scenarios, including document understanding, anomaly detection, and temporal reasoning. For clarity, we adopt a symbolic representation in the tables, where the number of "+" signs indicates relative performance within the full accuracy distribution, with more "+" denoting stronger performance. Specifically, one "+" denotes an accuracy between 20% and 40%, with each additional "+" indicating an increase of 20% accuracy. A "-" symbol is assigned to the lower-performing model(s) on each task, ranging from 0% to 20% which means that the model happened to get a few accurate results by chance. This representation highlights both model strengths and limitations in a compact form, facilitating direct comparison across a wide range of tasks.

Additionally, we also statistic the distribution of each level that four models perform in general capability tasks and application tasks. As shown in Fig. 2, GPT-4o consistently demonstrates superior performance, with over 50% of its application task outputs achieving the highest accuracy

level ("++"). In contrast, LLaVA-NeXT and Phi-3-Vision tend to produce a larger proportion of lower-accuracy outputs ("-" and "--"), particularly in application scenarios. The distribution on general capabilities reveals that GPT-4V performs more evenly across mid- to high-accuracy levels, while LLaVA-NeXT and Phi-3-Vision still show significant limitations in higher-level performance.

We observe from Tables 1 and 2 that none of the four MLLMs demonstrate reliable object localization capabilities which need to output the bounding box coordinate of referred object in the prompt. This limitation is evident in the poor performance on general tasks such as "Object Localization" and "Dense Captioning", as well as application-level tasks that require spatial grounding, including "Spot the Difference" and "Generate Pointing Outputs". A major contributing factor is that most current MLLMs are optimized for high-level visual-language understanding rather than fine-grained spatial reasoning. In particular, these models are typically optimized on image-caption pairs and lack region-level supervision or dedicated localization objectives. As a result, when presented with locate-style prompts—where the model is asked to identify a referred object in a complex scene—they often fail to resolve the correspondence between the textual reference and the correct visual region. This may be due to limitations in spatial representation, insufficient alignment between attention and object-level features, or the absence of specialized modules designed for grounding.

Besides, two large MLLMs do not perform well in "Celebrity Recognition" because these models might not explicitly be trained for facial identity matching, and lack dedicated face recognition modules. Unlike specialized face recognition systems that rely on finely tuned embeddings and large-scale labeled identity datasets, current MLLMs primarily focus on high-level image-text understanding and often treat faces as generic visual patterns without associating them with specific identities. In addition, safety considerations and dataset limitations further restrict their ability to recognize and name public figures. OpenAI's system may still apply intentional obfuscation or filtering to certain identity-related outputs to align with its responsible AI policies.

Grouping the results by both input modalities and high-level capability categories (Tables 3, 4, 5 and 6) provides a comprehensive view of model behavior across a wide range of tasks. We observe that large models, especially GPT-4o, consistently outperform smaller models in cognitively demanding categories such as instruction following, logical reasoning, multilingual understanding, temporal reasoning, and spatial localization. These models exhibit strong generalization across input types, particularly in fine-grained perception and multimodal reasoning tasks. In contrast, small models like LLaVA-NeXT and Phi-3-Vision perform

Table 1 Performance of four models on **General Capabilities** tasks

Task	GPT-4o	GPT-4V	LLaVA-NeXT	Phi-3-Vision
<i>Single Image-text Pair Inputs</i>				
Guided Prompting	+ + + +	+ + +	+ +	+
Constrained Prompting	+ +	+ +	- -	+
Detailed Prompting	+ + + + +	+ + +	+ +	- -
Color Recognition	+ + +	+	+	+
Food Recognition	+ + + +	+ + +	+	-
Landmark Recognition	+ + +	+ + +	+ +	+
Celebrity Recognition	- -	- -	+	-
Logo Recognition	+ + + +	+	+ +	+ + +
Counterfactual Examples	+ + + +	+ + + +	+ +	+ +
Scene Understanding	+ + + +	+ + + +	+ +	+
Emotion Recognition	+ + + +	+ + + +	+ + + +	-
Scene Text Recognition	+ + +	+ + +	+ +	+ +
Handwriting Text Recognition and Understanding	+	+	- -	-
Multilingual Image Description	+ + + +	+ + + +	+ +	- -
Multilingual Text Recognition, Translation, and Description	+ + + +	+ +	-	-
Multilingual Scene Text Recognition	+ + +	+ +	+	+
Multilingual Multicultural Understanding	+ + + + +	+ + + + +	+	- -
Coding Capability with Vision	+ + + +	+ +	-	+ + + +
Spatial relationship understanding	+ + +	+ + +	+	+
Object Counting	+ +	+	+	+
Object Localization	-	-	-	-
Dense Captioning	-	- -	- -	- -
Visual Math	+ + + +	+ +	- -	- -
Raven's Progressive Matrices	+	+	-	-
Chart Understanding and Reasoning	+ +	+ +	+	+ +
Table Understanding and Reasoning	+ +	+ +	+	+
Document Understanding	+ + + +	+ +	- -	-
Emotion Understanding	+ + + + +	+ + + + +	+ + + + +	+ +
Food/Landmark/Celebrity/Logo Understanding	+ + +	+ +	+ +	-
Science and Knowledge	+ +	+ +	+ +	+ +
Joke and Meme	+ + + +	+ + + +	+ +	+
Multimodal Commonsense	+ + +	+ + + +	+ +	+ +
Evaluation of Generated Images	+ +	+	+	+
Understand Basic Pointing Inputs	+ + +	+ + +	-	-
Understand Pixel Space Edited Input	+ + + + +	+ +	- -	-
<i>Multi-image with Text Inputs</i>				
Abstract Visual Stimuli	+ + +	+	- -	-
Multi-view Reasoning	+	+		
Image Aesthetics	+ + + + +	+ + + + +		
Forensic Detection	+ + +	+ +		
<i>Multi-frame with Text Inputs</i>				
Action Recognition	+ + + +	+ +	+	-
Temporal Ordering	+ + +	-		
Temporal Anticipation	+ + + + +	+ + + + +		
Temporal Localization and Reasoning	+ + + +	-		
Grounded Temporal Understanding	+ + + +	+ +		
<i>Interleaved Image-text Inputs</i>				
Spot the Difference	-	-	-	-
Wechsler Adult Intelligence Scale	+ + + +	+ + + +	-	-
Prompt Generation for Image Editing	+ + + + +	+ + + + +	+ + + + +	+ + + + +
Generate Pointing Outputs	-	-		

A higher number of “+” symbols indicates better accuracy, while “-” represents lower accuracy levels. Specifically, “+ + + + +” denotes perfect performance (100% accuracy), and “- -” marks failure cases with 0% accuracy

Table 2 Performance of four models on **Application** tasks

Task	GPT-4o	GPT-4V	LLaVA-NeXT	Phi-3-Vision
Photo	–	+ +		
Orga-niza-tion				
Dis-covery and Asso-ciation of Parts and Objects	+ +	–		
Dense Cap-tioning with Seg-menta-tion	+ + + + + + + +			
Anom-aly Detec-tion	+ + + +	+ + +	+	–
Safety Inspec-tion	+ + + + + +		+	+ + + + +
Gro-cery Check-out	+ + + +	+ +	– –	+
Dam-age Evalu-ation	+ + + + + + + +		+ + +	+ + +
Insur-ance Report-ing	–	–	–	–
House Navi-gation	+ + + + + + + + + +			
Notifi-cation Under-stand-ing	+ + + + + + + +			
Web Brows-ing	+ + + + + + + + + +			–
Online Shop-ping	+ + + + + + + + + +			–
Watch-ing Videos	+ +	+		

A higher number of “+” symbols indicates better accuracy, while “–” represents lower accuracy levels. Specifically, “+ + + + +” denotes perfect performance (100% accuracy), and “– –” marks failure cases with 0% accuracy

comparably only in basic perception-oriented tasks (e.g., color recognition, OCR), but show clear limitations in tasks requiring explicit spatial grounding, multi-image comparison, or sequential understanding, where performance drops significantly.

In particular, temporal reasoning tasks reveal a marked scale gap: large models can effectively model temporal dependencies and multi-frame interactions, whereas small models often fail to capture such patterns. Similarly, structured inputs like charts, tables, and documents highlight the advantages of larger models with better visual-text alignment and memory capacity. Conversely, all models—including GPT-4o—struggle with interleaved reasoning tasks such as Spot the Difference, indicating an ongoing challenge in fine-grained visual grounding and alignment under complex context switching.

These findings underscore that model scale and reasoning ability are key determinants of performance across multimodal tasks. While lightweight models retain utility in focused recognition scenarios, advancing multimodal understanding will require further progress in temporal modeling, grounding, and multi-view integration.

Overall, GPT-4o consistently outperforms the other models across both categories, demonstrating strong generalization and robustness. GPT-4V also performs competitively in most tasks, while LLaVA-NeXT and Phi-3-Vision show relatively limited capability, particularly on tasks involving multilingual reasoning, visual mathematics, or structured document input. These observations reinforce the gap between high-capacity proprietary models and lighter open-source alternatives, especially in complex multimodal reasoning scenarios.

8 Conclusion

In this paper, we conduct a comprehensive benchmarking of large and small Multi-Modal Language Models (MLLMs) across a diverse set of core tasks using various input types, ranging from single image-text pairs to interleaved image-text pairs. Through this evaluation, we reveal each model’s strengths and limitations, offering insights into their applicability in real-world scenarios and highlighting areas for further development.

- *Single image-text pair* GPT-4o demonstrates consistently superior performance in tasks involving single image-text pairs, excelling particularly in complex capabilities such as logical reasoning, multilingual comprehension, and code generation. GPT-4V closely follows, with notable strengths in OCR, language generation, and spatial understanding. In contrast, smaller models like

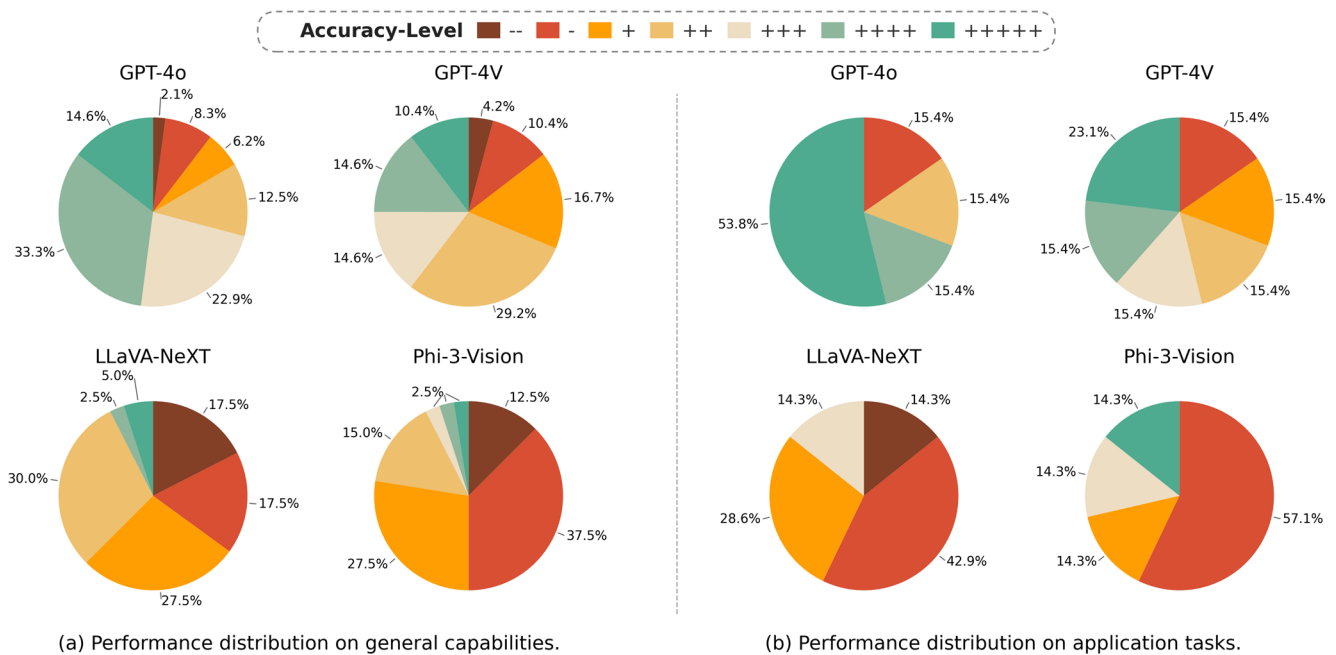


Fig. 2 Comparison of accuracy-level distributions across four multimodal models

LLaVA-NeXT and Phi-3-Vision perform well in specific recognition tasks, such as logo and emotion identification, yet face challenges in advanced reasoning and spatial tasks. These findings emphasize the scalability advantages of larger models for more cognitively demanding applications.

- *Multi-image with text input* For tasks requiring multi-image processing, GPT-4o outperforms GPT-4V, particularly in abstract reasoning, visual sentiment analysis, and manipulation detection. Smaller MLLMs lack multi-image processing capabilities, underscoring the necessity of large models for complex visual inputs. GPT-4o's consistent accuracy in these tasks demonstrates its robustness in handling integrated, cross-image analyses.
- *Multi-frame with text input* In sequence-based tasks such as action recognition and video understanding, only the large MLLMs, GPT-4o and GPT-4V, effectively process multi-frame inputs. GPT-4o excels in temporal understanding, consistently achieving top performance in tasks involving temporal anticipation, ordering, localization and reasoning, showcasing a human-like capacity for interpreting dynamic sequences. While GPT-4V occasionally achieves comparable results, smaller models like LLaVA-NeXT and Phi-3-Vision have difficulty in handling sequence inputs, limiting their suitability for temporal tasks.
- *Interleaved Image-text pair* Interleaved image-text inputs present distinct challenges, especially in spatial reasoning and association. GPT-4o leads in performance,

though it struggles with fine-grained spatial localization, particularly in tasks requiring accurate coordinate alignment with visual content. Both GPT-4o and GPT-4V show limitations in generating accurate visual pointing outputs, suggesting a need for advancements in prompt design and model architecture to enhance spatial reasoning capabilities. Despite these challenges, GPT-4o demonstrates strong abstract reasoning in specific IQ tasks, highlighting its potential in complex problem-solving scenarios.

- *Application scenarios* The models exhibit notable potential in real-world applications, showing versatility across a variety of tasks. GPT-4V achieves high accuracy in multiple qualitative tasks, illustrating its adaptability in user-centered applications. Phi-3-Vision performs well in safety inspection tasks, while LLaVA-NeXT excels in online shopping contexts when paired with GPT-4V's text prompts. These application scenarios underscore the practical impact of MLLMs, revealing their potential for user engagement, informed decision-making, and task-specific multimodal solutions.

In summary, we select representative large and small Multimodal Language Models (MLLMs) across diverse input modalities and high-level capability categories, revealing systematic differences in their performance. We find that both large and small models can handle basic perception-oriented tasks, such as logo recognition, color classification, and OCR, reasonably well—indicating that such tasks do not require strong reasoning or high model capacity.

Table 3 Capability-level performance summary for **Single image-text** inputs

Top-level Taxonomy	High-level Category	Included Tasks	#Task Number	GPT-4o	GPT-4V	LLaVA-NeXT	Phi-3-Vision
Understanding	Instruction following	Guided Prompting; Constrained Prompting; Detailed Instruction Following	3	0.839	0.623	0.382	0.193
	Perception	Food, Landmark, Celebrity, Logo Recognition; Color Recognition; Scene Understanding; Counterfactual Examples	7	0.734	0.615	0.503	0.288
	OCR	Scene Text Recognition; Handwriting Text Recognition	2	0.574	0.506	0.229	0.307
	Multilingual	Multilingual Image Description; Multilingual Text Recognition, Translation, and Description; Multilingual Scene Text Recognition; Multilingual Multicultural Understanding	4	0.877	0.737	0.297	0.080
Reasoning	Spatial Localization & Reasoning	Spatial Relationship; Object Counting; Object Localization; Dense Captioning	4	0.358	0.276	0.191	0.144
	Logical Reasoning	Visual Math; Raven's Progressive Matrices	2	0.533	0.467	0.033	0.083
	Multimodal Reasoning	Emotion Understanding and Reasoning; Food/Landmark/Celebrity/Logo Understanding and Reasoning	2	0.857	0.743	0.743	0.386
	Multimodal Commonsense	Science and Knowledge; Abstract Visual Stimuli; Joke and Meme	3	0.748	0.790	0.463	0.440
	Chart, Table, Document Understanding & Reasoning	Chart Understanding & Reasoning; Table Understanding & Reasoning; Document Understanding & Reasoning	3	0.655	0.522	0.187	0.296
	Coding	Coding Capability with Vision	1	0.867	0.533	0.133	0.800
Human-computer Interaction	Interaction-driven Visual Understanding & Reasoning	Understand Basic Pointing Inputs; Understand Pixel Space Edited Input	2	0.875	0.590	0.063	0.118
Assessment	Image Quality & Authenticity Comparison	Evaluation of Generated Images	1	0.4615	0.3846	0.2692	0.3077

Values indicate the average accuracy (higher is better)

Table 4 Capability-level performance summary for **Multi-image** inputs

Top-level taxonomy	High-level category	Included tasks	#Task number	GPT-4o	GPT-4V	LLaVA-NeXT	Phi-3-Vision
Assessment	Image Quality & Authenticity Comparison	Image Aesthetic; Forensic Detection	2	0.875	0.725	–	–
Reasoning	Visual Logical Reasoning	Abstract Visual Stimuli; Multi-view Reasoning	2	0.545	0.351	0.042	0.042

Values indicate the average accuracy (higher is better)

However, large models—particularly GPT-4o—exhibit clear advantages in cognitively demanding categories, including instruction following, multilingual understanding, logical reasoning, and temporal and spatial reasoning, where small models like LLaVA-NeXT and Phi-3-Vision

consistently underperform. GPT-4V also shows strong results in structured tasks such as document analysis and chart/table reasoning. In contrast, small models tend to struggle with multi-image, multi-frame, and interaction-driven tasks, which require more advanced reasoning and

Table 5 Capability-level performance summary for **Multi-frame** inputs

Top-level Taxonomy	High-level Category	Included Tasks	#Task Number	GPT-4o	GPT-4V	LLaVA-NeXT	Phi-3-Vision
Understanding & Reasoning	Temporal & Sequential Reasoning	Multi-image Sequencing; Temporal Ordering, Temporal Anticipation, Temporal Localization and Reasoning	4	0.835	0.473	0.087	0.037
Human-computer Interaction	Interaction-driven Visual Understanding & Reasoning	Grounded Temporal Understanding	1	0.800	0.500	–	–

Values indicate the average accuracy (higher is better)

Table 6 Capability-level performance summary for **Interleaved image-text** inputs

Top-level Taxonomy	High-level Category	Included Tasks	#Task Number	GPT-4o	GPT-4V	LLaVA-NeXT	Phi-3-Vision
Understanding Reasoning	Detail Analysis	Spot the Difference	1	0	0	0	0
	Logical	Wechsler Adult Intelligence Scale	1	0.900	0.900	0.167	0.167
Generation	Editing-oriented Prompt Generation	Prompt Generation for Image Editing	1	1	1	1	1
Human-computer Interaction	Interaction-driven Visual Understanding & Reasoning	Generate Pointing Outputs	1	0	0	–	–

Values indicate the average accuracy (higher is better)

memory. Furthermore, tasks involving high-level multimodal reasoning, such as abstract visual understanding and emotional inference, consistently favor large models regardless of input type. These results underscore the importance of both model scale and reasoning capability in achieving strong multimodal performance, while also highlighting the potential utility of small models in lightweight or recognition-focused applications.

These findings suggest that while large MLLMs generally achieve superior performance across a broad range of tasks, smaller models can be effective in specialized applications, serving as resource-efficient alternatives. This study provides a foundational reference for future research, offering a comprehensive analysis of MLLM capabilities and identifying pathways for optimization. Future work could build on these insights to explore new applications, refine model architectures, and enhance our understanding of how multimodal systems address complex, real-world challenges.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00138-025-01762-0>.

Acknowledgments This work is supported in part by NSF CNS-2120369 and the Institute of Education Sciences (IES) through Award 2229873 (AI Institute for Transforming Education for Children with Speech and Language Processing Challenges).

Author Contributions Xuelu Feng was responsible for collecting the evaluation data, conducting model testing and benchmarking, performing statistical analysis, and drafting the initial version of the manuscript. Yunsheng Li proposed the core research direction, contributed to experiment design, and provided extensive feedback and revisions

to the manuscript throughout its development. Dongdong Chen contributed experimental ideas, especially regarding the benchmarking methodology, and was actively involved in both writing and revising the manuscript. Mei Gao designed and provided the prompts used in various evaluation tasks, and contributed to the refinement of prompt-based experiments. Mengchen Liu suggested key experimental setups, curated domain-specific testing tasks, and assisted in the evaluation and analysis of results. Junsong Yuan offered high-level guidance on the structure, focus, and positioning of the paper, and provided in-depth feedback to improve the clarity and impact of the manuscript. Chunming Qiao supervised the overall project, provided conceptual guidance, and contributed significantly to the refinement and revision of the manuscript. All authors reviewed and approved the final version of the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no Conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S.: Scaling instruction-finetuned language models. *J. Mach. Learn. Res.* **25**(70), 1–53 (2024)
2. OpenAI, R.: Gpt-4 technical report. View in Article **2**(5). [arxiv:2303.08774](https://arxiv.org/abs/2303.08774) (2023)
3. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: open and efficient foundation language models. [arXiv:2302.13971](https://arxiv.org/abs/2302.13971) (2023)
4. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S.: Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.* **24**(240), 1–113 (2023)
5. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: a frontier large vision-language model with versatile abilities. [arXiv:2308.12966](https://arxiv.org/abs/2308.12966) (2023)
6. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning, pp. 19730–19742 (2023). PMLR
7. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: grounding multimodal large language models to the world. [arXiv:2306.14824](https://arxiv.org/abs/2306.14824) (2023)
8. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: improving large multi-modal models with better captions. [arXiv:2311.12793](https://arxiv.org/abs/2311.12793) (2023)
9. Gupta, T., Kamath, A., Kembhavi, A., Hoiem, D.: Towards general purpose vision systems: an end-to-end task-agnostic vision-language architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16399–16409 (2022)
10. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. [arXiv:2312.11805](https://arxiv.org/abs/2312.11805) (2023)
11. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: enhancing vision-language understanding with advanced large language models. [arXiv:2304.10592](https://arxiv.org/abs/2304.10592) (2023)
12. Zhang, P., Dong, X., Wang, B., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., Duan, H., Zhang, S., Ding, S., et al.: Internlm-xcomposer: a vision-language large model for advanced text-image comprehension and composition. [arXiv:2309.15112](https://arxiv.org/abs/2309.15112) (2023)
13. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. [arXiv:2304.14178](https://arxiv.org/abs/2304.14178) (2023)
14. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13299–13308 (2024)
15. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.-C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. [arXiv:2404.12390](https://arxiv.org/abs/2404.12390) (2024)
16. Wu, Y., Yan, L., Luo, Y., Wang, Y., Tang, N.: Evaluating task-based effectiveness of mllms on charts. [arXiv:2405.07001](https://arxiv.org/abs/2405.07001) (2024)
17. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: is your multi-modal model an all-around player? [arXiv:2307.06281](https://arxiv.org/abs/2307.06281) (2023)
18. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: evaluating large multimodal models for integrated capabilities. [arXiv:2308.02490](https://arxiv.org/abs/2308.02490) (2023)
19. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14375–14385 (2024)
20. Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Huang, X., Wang, Z., Sheng, L., Bai, L., et al.: Lamm: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Adv. Neural Inf. Process. Syst.* **36**, 26650–26685 (2024)
21. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. [arXiv:2306.09265](https://arxiv.org/abs/2306.09265) (2023)
22. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v (ision). **9**(1), 1. [arXiv:2309.17421](https://arxiv.org/abs/2309.17421) (2023)
23. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al.: Sparks of artificial general intelligence: Early experiments with GPT-4. [arXiv:2303.12712](https://arxiv.org/abs/2303.12712) (2023)
24. Jiang, Y., Yan, X., Ji, G.-P., Fu, K., Sun, M., Xiong, H., Fan, D.-P., Khan, F.S.: Effectiveness assessment of recent large vision-language models. *Vis. Intell.* **2**(1), 17 (2024)
25. Islam, M.S., Rahman, R., Masry, A., Laskar, M.T.R., Nayeem, M.T., Hoque, E.: Are large vision language models up to the challenge of chart comprehension and reasoning? An extensive investigation into the capabilities and limitations of LVLMS. [arXiv:2406.00257](https://arxiv.org/abs/2406.00257) (2024)
26. OpenAI: Gpt-4v(ision) system card. Technical report, OpenAI (2023)
27. OpenAI: Gpt-4o: Multimodal large language model by openai. Technical report, OpenAI (2024). Accessed: 30 Sep 2024. <https://openai.com/research/gpt-4o>
28. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: improved reasoning, OCR, and world knowledge (2024)
29. Abidin, M., Jacobs, S.A., Awan, A.A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al.: Phi-3 technical report: A highly capable language model locally on your phone. [arXiv:2404.14219](https://arxiv.org/abs/2404.14219) (2024)
30. Liu, Y., Li, Z., Yang, B., Li, C., Yin, X., Liu, C.-l., Jin, L., Bai, X.: On the hidden mystery of OCR in large multimodal models. [arXiv:2305.07895](https://arxiv.org/abs/2305.07895) (2023)
31. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A.: Training language models to follow instructions with human feedback. *Adv. Neural. Inf. Process. Syst.* **35**, 27730–27744 (2022)
32. Mishra, S., Khashabi, D., Baral, C., Hajishirzi, H.: Cross-task generalization via natural language crowdsourcing instructions. [arXiv:2104.08773](https://arxiv.org/abs/2104.08773) (2021)
33. Min, W., Wang, Z., Liu, Y., Luo, M., Kang, L., Wei, X., Wei, X., Jiang, S.: Large scale visual food recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(8), 9932–9949 (2023)
34. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101—mining discriminative components with random forests. In: Computer vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13, pp. 446–461. Springer (2014)
35. Zheng, Y.-T., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T.-S., Neven, H.: Tour the world: building a web-scale landmark recognition engine. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1085–1092. IEEE (2009)
36. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building Rome in a day. *Commun. ACM* **54**(10), 105–112 (2011)
37. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: Computer

- Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14, pp. 87–102. Springer (2016)
38. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. In: The Twelfth International Conference on Learning Representations (2023)
 39. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13, pp. 740–755 (2014). Springer
 40. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
 41. Zhou, B., Khosla, A., Lapedriza, A., Torralba, A., Oliva, A.: Places: An image database for deep scene understanding. [arXiv:1610.02055](https://arxiv.org/abs/1610.02055) (2016)
 42. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4291–4301 (2019)
 43. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vlbart: Pre-training of generic visual-linguistic representations. [arXiv:1908.08530](https://arxiv.org/abs/1908.08530) (2019)
 44. Liu, F., Bugliarello, E., Ponti, E.M., Reddy, S., Collier, N., Elliott, D.: Visually grounded reasoning across languages and cultures. [arXiv:2109.13238](https://arxiv.org/abs/2109.13238) (2021)
 45. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2901–2910 (2017)
 46. Bagherinezhad, H., Hajishirzi, H., Choi, Y., Farhadi, A.: Are elephants bigger than butterflies? Reasoning about sizes of objects. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30 (2016)
 47. Kang, D., Ma, Z., Chan, A.B.: Beyond counting: comparisons of density maps for crowd analysis tasks—counting, detection, and tracking. *IEEE Trans. Circuits Syst. Video Technol.* **29**(5), 1408–1422 (2018)
 48. Kutlimuratov, A., Khamzaev, J., Kuchkorov, T., Anwar, M.S., Choi, A.: Applying enhanced real-time monitoring and counting method for effective traffic management in tashkent. *Sensors* **23**(11), 5007 (2023)
 49. DeHoratius, N., Holzapfel, A., Kuhn, H., Mersereau, A.J., Sternbeck, M.: Evaluating count prioritization procedures for improving inventory accuracy in retail stores. *Manuf. Service Oper. Manag.* **25**(1), 288–306 (2023)
 50. Delplanque, A., Foucher, S., Théau, J., Bussière, E., Vermeulen, C., Lejeune, P.: From crowd to herd counting: how to precisely detect and count African mammals using aerial imagery and deep learning? *ISPRS J. Photogramm. Remote. Sens.* **197**, 167–180 (2023)
 51. Ranjan, V., Sharma, U., Nguyen, T., Hoai, M.: Learning to count everything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3394–3403 (2021)
 52. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)
 53. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
 54. Johnson, J., Karpathy, A., Fei-Fei, L.: Denscap: fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4565–4574 (2016)
 55. Raven, J.: Raven progressive matrices. In: Handbook of Nonverbal Assessment, pp. 223–237. Springer, Berlin (2003)
 56. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B.: Language is not all you need: Aligning perception with language models. *Adv. Neural. Inf. Process. Syst.* **36**, 72096–72109 (2023)
 57. Zhang, C., Gao, F., Jia, B., Zhu, Y., Zhu, S.-C.: Raven: A dataset for relational and analogical visual reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5317–5327 (2019)
 58. Barrett, D., Hill, F., Santoro, A., Morcos, A., Lillicrap, T.: Measuring abstract reasoning in neural networks. In: International Conference on Machine Learning, pp. 511–520 (2018). PMLR
 59. Wechsler, D.: Wais-r: Manual: Wechsler adult intelligence scale-revised. (No Title) (1981)
 60. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
 61. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. *Adv. Neural Inf. Process. Syst.* **36**, 19769–19782 (2024)
 62. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: MVTec AD—a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9592–9600 (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Xuelu Feng received the B.E. degree in Internet of Things Engineering from Sichuan Agricultural University, China, in 2020, and the M.E. degree at School of Software Engineering, Xi'an Jiaotong University, Xi'an, China, in 2023. She is currently pursuing the Ph.D. degree at department of Computer Science and Engineering, University at Buffalo, Buffalo, New York State, USA. Her research interests include computer vision and deep learning.

Yunsheng Li received his Ph.D. degree from the University of California at San Diego in 2021. He is currently a research in Microsoft Cloud AI, Redmond. His research interests include in vision language model, segmentation and domain adaptation.

Dongdong Chen is a principal research manager of Microsoft GenAI. He received his PhD degree under the joint phd program between University of Science and Technology of China and MSRA. His main research areas include the design and application of large-scale foundation cognitive models and AIGC models, AI security, visual representation learning, and neural network architecture design.

Mei Gao Ph.D., is a Member of Technical Staff at Microsoft, specializing in data development and model training for large language and vision-language models, including the Phi and Florence series. She earned her Ph.D. in Mechanical Engineering and M.Eng. in Electrical, Computer, and Systems Engineering from Rensselaer Polytechnic Institute. Her current work advances multimodal AI and foundation model development for large-scale applications.

Mengchen Liu received the BS degree in electronics engineering and the PhD degree in computer science from Tsinghua University. He is a senior researcher with Meta. His research interests include explainable AI and computer vision. He has served as a PC member and reviewer for various conferences and journals.

Junsong Yuan Junsong Yuan is Professor and Director of Visual Computing Lab at Department of Computer Science and Engineering (CSE), State University of New York at Buffalo, USA. Before joining SUNY Buffalo, he was Associate Professor (2015-2018) and Nanyang Assistant Professor (2009-2015) at Nanyang Technological University (NTU), Singapore. He obtained his Ph.D. from Northwestern University in 2009 (advised by Prof. Ying Wu), M.Eng. from National University of Singapore in 2005, and B.Eng. from Huazhong University of Science Technology (HUST) in 2002. He received Chancellor's Award for Excellence in Scholarship and Creative Activities from SUNY, Nanyang Assistant Professorship from NTU, Outstanding EECS Ph.D. Thesis award from Northwestern University, and Best Paper Award from IEEE Trans. on Multimedia. He serves as Senior Area Editor of Journal of Visual Communication and Image Representation (JVCI), Associate Editor of IEEE Trans. on Pattern Analysis and Machine Intelligence (T-PAMI), IEEE Trans. on Image Processing (T-IP), IEEE Trans. on Circuits and Systems for Video Technology (T-CSVT), and Machine Vision and Applications (MVA). He also serves as General/Program Co-chair of ICME and Area Chair for CVPR, ICCV, ECCV, ACM MM, etc. He was elected Faculty Senator at Both SUNY Buffalo and NTU. He is a Fellow of IEEE and IAPR.

Chunming Qiao is a SUNY distinguished professor and also the current chair of the Computer Science and Engineering Department, University at Buffalo, Buffalo, NY. He was elected to IEEE fellow for his contributions to optical and wireless network architectures and protocols. His current focus is on connected and autonomous vehicles. He has published extensively with an h-index of more than 69. Two of his papers have received the best paper award from IEEE and Joint ACM/IEEE venues. He also has seven US patents and served as a consultant for several IT and Telecommunications companies since 2000. His research has been funded by a dozen major IT and telecommunications companies including Cisco and Google, and more than a dozen NSF grants.