

Affordance Benchmark for MLLMs

Junying Wang^{1,2}, Wenzhe Li², Yalun Wu³, Yingji Liang², Yijin Guo^{2,3}, Chunyi Li^{2,3}, Haodong Duan², Zicheng Zhang^{2,†}, Guangtao Zhai^{2,3,†}

¹Fudan University, ²Shanghai Artificial Intelligence Laboratory, ³Shanghai Jiao Tong University

[†]Corresponding author.

Project Page: <https://github.com/JunyingWang959/A4Bench> Team: <https://aiben.ch>

Abstract

Affordance theory suggests that environments inherently provide action possibilities shaping perception and behavior. While Multimodal Large Language Models (MLLMs) achieve strong performance in vision-language tasks, their ability to perceive affordance, which is crucial for intuitive and safe interactions, remains underexplored. To address this, we introduce **A4Bench**, a novel benchmark designed to evaluate the affordance perception abilities of MLLMs across two dimensions: 1) *Constitutive Affordance*, assessing understanding of inherent object properties through 1,282 question-answer pairs spanning nine sub-disciplines, and 2) *Transformative Affordance*, probing dynamic and contextual nuances (e.g., misleading, time-dependent, cultural, or individual-specific affordance) with 718 challenging question-answer pairs. We evaluate 17 MLLMs (nine proprietary and eight open-source) and compare them to human performance. Results show that proprietary models generally outperform open-source ones, yet all models perform far below humans, especially in transformative affordance. Furthermore, even top-performing models, such as Gemini-2.0-Pro (18.05% overall exact match accuracy), significantly lag behind human performance (best: 85.34%, worst: 81.25%). These findings highlight critical gaps in environmental understanding of MLLMs and provide a foundation for advancing AI systems toward more robust, context-aware interactions.

Introduction

‘What we perceive when we look at objects are their affordances, not their qualities’. This bold assertion by James J. Gibson encapsulates the essence of *affordance theory* (Gibson 2014), which argues that environments inherently provide organisms with a spectrum of action possibilities shaping perception and behavior. Unlike traditional theories that focus on physical attributes, affordance theory emphasizes the direct perception of *action possibilities* embedded in environmental features. For example, a door is not merely a static structure but something that affords multiple actions: it can be unlocked with a key, pushed open, or even kicked. As illustrated in Figure 1 (bottom left), affordances go beyond conventional functionality to encompass a wide range of potential interactions grounded in the material, shape, and context of an object. Thus, affordances form a crucial link between ecological features and behavioral responses, supporting a unified ecological approach to perception.

Why affordance is important?

Understanding affordance is vital for intelligent agents, whether biological or artificial, to **engage meaningfully with their environments**. For artificial intelligence systems navigating complex settings, understanding affordance ensures robust and intuitive interactions while enhancing safety. For example, a robot perceiving a surface as affording support can navigate terrain securely, while one identifying an object as graspable can manipulate it effectively. Moreover, effective affordance perception enables industrial robots to **execute tasks with greater efficiency** and allows rehabilitation robots to **enhance human-machine interaction** by adapting to user needs. The capacity to discern affordance, whether beneficial or harmful, underpins behaviors ranging from survival to complex social interactions. Gibson underscored this by noting how humans modify environments ‘to change what it affords’, emphasizing the profound link between affordance and intentional action.

What Affordance Perception Should MLLMs Possess?

Accurate perception of an object affordance enables recognition of its true utility, as Gibson asserts, ‘If the affordance of a thing are perceived correctly, we say that it looks like what it is’. Yet discerning genuine affordance often demands experiential learning, since a leaf appearing benign may conceal a nettle sting or a politician seeming helpful may mask deceptive demagoguery.

To evaluate this capability rigorously, we introduce a novel benchmark (**A4Bench**) to assess affordance perception across two primary dimensions. First, **constitutive affordance** examines how MLLMs apprehend inherent object and environmental properties such as shape, size, or material that determine whether a surface affords walking or an object affords grasping, using 1282 question-answer pairs across nine sub-disciplines. Second, **transformative affordance** probes comprehension of dynamic affordance, including **misleading affordance** (like a glass barrier appearing as open air but affording collision), **time-independent affordance** (e.g., fruit ripening changes its food value affordance, specified by color), **cross-cluture affordance** (e.g., a postbox affording letter-mailing only within a postal system), and **individual affordance** (e.g., a ledge being sit-onable depending on leg length).

A4Bench pioneers comprehensive affordance perception evaluation, highlighting contextual and dynamic nuances to

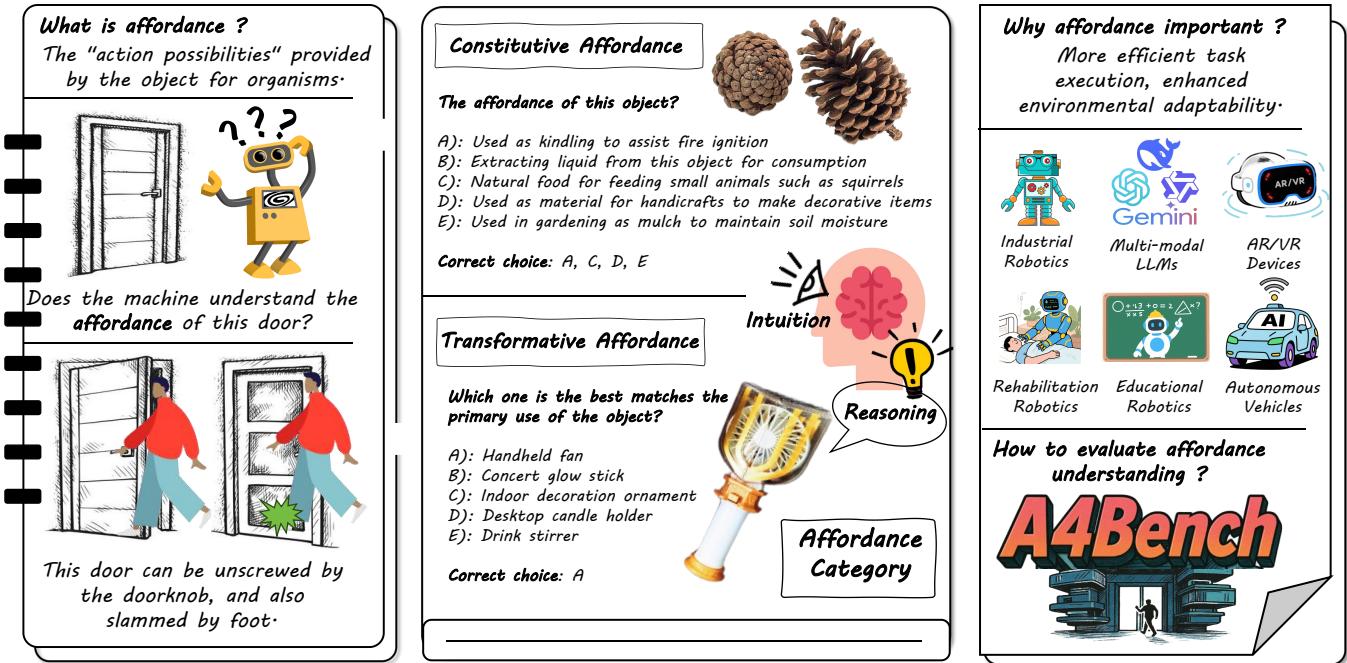


Figure 1: The motivation of the A4Bench. The affordance theory proposed by James J. Gibson (Gibson 2014) defines the action possibilities provided by the object for organisms. Evaluating the affordance perception abilities of MLLMs can help enable more efficient task execution and improved adaptability to diverse environments for AI systems.

reveal MLLM capabilities and steer future development. Testing 17 leading MLLMs, both open-source and closed-source, on **A4Bench** reveals that even the top-performing model significantly trails human performance, leading to a critical conclusion:

MLLMs are still poor at affordance perception.

MLLMs exhibit significant limitations in perceiving affordance with human-like proficiency, with marked disparities between open-source and closed-source variants. This deficiency spans constitutive and transformative affordance. Challenges arise from limited contextual understanding, notably in medical disciplines within constitutive affordance and agent-specific dimensions within transformative affordance. Conversely, human observers consistently excel across all dimensions, highlighting inadequate comprehension by MLLMs of object-human-environment interactions essential for robust affordance perception (poor at affordance perception).

Evaluating affordance perception in MLLMs shapes AI advancements by driving **A4Bench** development, as shown in Figure 1. And our contributions are summarized as:

- We extend affordance theory to the context of MLLMs and introduce a systematic evaluation framework for affordance that rigorously tests the understanding of MLLMs on understanding constitutive and transformative affordance.
- We propose **A4Bench**, the first comprehensive benchmark specifically designed to assess the affordance perception capabilities of MLLMs. It encompasses 2,000 multimodal question-answer pairs, covering both

Constitutive Affordance (1,282 pairs across nine sub-disciplines) and Transformative Affordance (718 pairs addressing misleading, time-dependent, cross-cultural, and individual-specific affordance).

- By evaluating 17 MLLMs (9 proprietary and 8 open-source) against human performance, we provide a detailed analysis of their affordance perception capabilities. Our findings reveal significant limitations in existing models, with even top performers lagging far behind human benchmarks, thus offering a roadmap for future improvements in context-aware AI systems.

Related Works

Multimodal Large Language Models. Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities in diverse vision-language tasks. Among them, diverse proprietary (*close-source*) models (OpenAI 2024; Anthropic 2024; Google 2024a; step-1o; BailingMM-Pro-0120; MUG-U-7B) and representative open-source models (Wu et al. 2024a; Team et al. 2025; Dai et al. 2023; Zhu et al. 2023; Liu et al. 2023, 2024a; Li et al. 2024b; Lu et al. 2024; Guo et al. 2025a; Dong et al. 2025; Yao et al. 2024; Dai et al. 2024; Agrawal et al. 2024; Liu et al. 2024d; Zhang et al. 2025a; Lin et al. 2023; Chen et al. 2025) have exhibited impressive superiority from embodied agents (Song et al. 2023) to other real-world applications (Huynh et al. 2025; Wang et al. 2025a; Jiao, Zhang, and Li 2025a; Jiao and Li 2025). However, whether these MLLMs are **masters at perceiving the affordance of object** is still questionable, which needs further investigation.

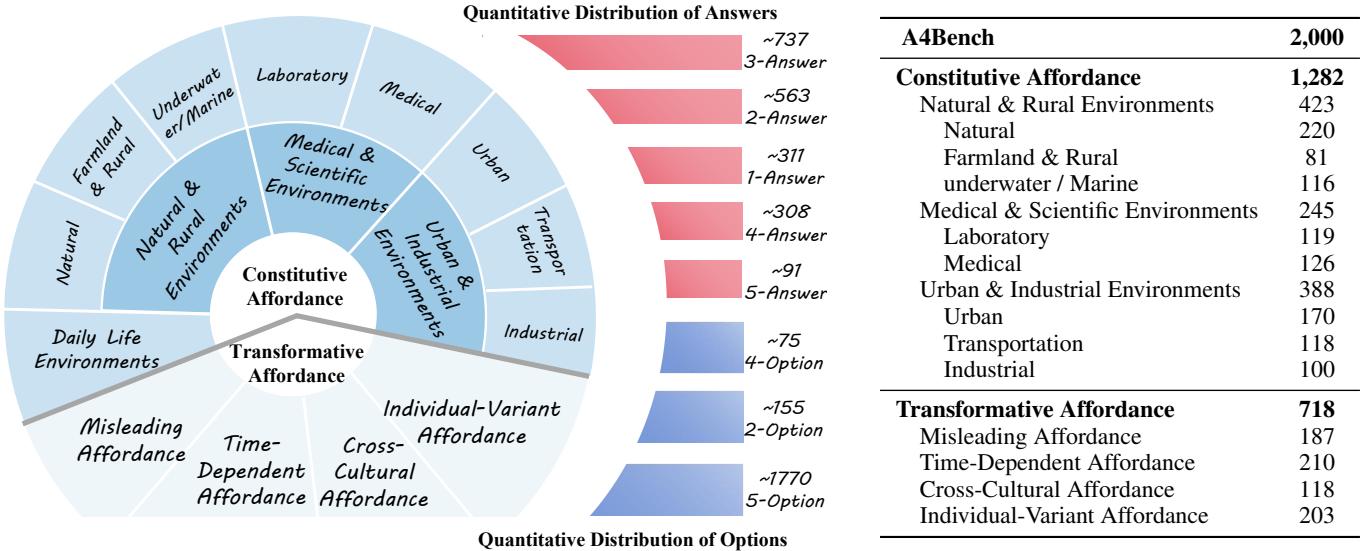


Figure 2: Structure and quantitative overview of **A4Bench**. The left panel presents the focused aspects of the benchmark, detailing the primary dimensions (Constitutive and Transformative Affordance) and their respective sub-dimensions. The middle panel depicts the distributions of answer counts and option counts. The right panel reports the number of multimodal question–answer pairs across each sub-dimension, providing a comprehensive overview of the dataset composition.

Multimodal Benchmarks. Current benchmarks provide valuable insights, assessing capabilities ranging from bilingual perception and reasoning (Liu et al. 2024c), hierarchical comprehension (Li et al. 2024a; Wang et al. 2025b; Wen et al. 2025; Zhang et al. 2025c), expert-level multimodal tasks (Ying et al. 2024; Jiao, Zhang, and Li 2025b; Guo et al. 2025b; Wu et al. 2023, 2024b; Zhang et al. 2022), and complex embodied scenarios (Chen et al. 2024b), to critical nuances like visual dependency (Chen et al. 2024a), safety against adversarial inputs (Liu et al. 2024b; Wang, Zhang, and Yuan 2025; Zhang et al. 2025b), and understanding of AI-generated images (Zhang et al. 2024). Despite these efforts, there is still a gap in explicitly assessing the affordance perception abilities of MLLMs. Since understanding affordance is fundamental for any intelligent agent to interact effectively, meaningfully, and safely with its surroundings, **A4Bench** is developed to bridge this gap.

Constructing the A4Bench

Key Principles

Covering Constitutive and Transformative Affordance
The human brain, when visually recognizing action interactions, processes static forms alongside dynamic movements (Blake and Shiffrar 2007). Inspired by this, assessing MLLMs in perceiving object affordance involves *understanding inherent object properties and their empowering potential*. Assessing whether these models meet such criteria requires examining proficiency in (**constitutive affordance**) and (**transformative affordance**) perception. As shown in Figure 2, constitutive affordance covers a diverse range of scenarios, while transformative affordance includes misleading affordance, individual-variant affordance, time-dependent affordance, and cross-cultural affordance.

Guaranteeing Benchmarking Difficulty *Rigorous benchmarking effectively evaluates rapid advancements in MLLMs at the forefront of human knowledge* (Phan et al. 2025). To ensure the difficulty and quality of benchmark, diverse strategies are implemented: a quality control mechanism, a vision-language prompt removal approach, and an unspecified options and answers strategy. 1) **Quality control mechanism** employs a human-MLLM mixed-obfuscation adaptation process. This iterative method starts with human-generated problems, followed by alternating revisions between models and experts until the model fails to respond correctly, ensuring a challenging benchmark. 2) **Vision-language prompt removal approach** enhances affordance comprehension assessment by replacing explicit object names in images, questions, and options with ‘this object.’ This method replicates real-world multimodal perception environments critical for affordance evaluation while increasing difficulty for robust model assessment. 3) **Unspecified options and answers strategy** heightens challenge by concealing the exact number of correct answers from models and participants, thus reducing reliance on guessing while prioritizing deep comprehension. And the quantitative distributions of the answers and options are shown in the middle panel of Figure 2.

Focused Aspects

Evaluating affordance perception examines inherent properties such as structure and material alongside potential dynamic action interactions, encompassing constitutive and transformative affordance dimensions. And the representative examples of these multifaceted dimensions are clearly and visually illustrated in Figure 3.

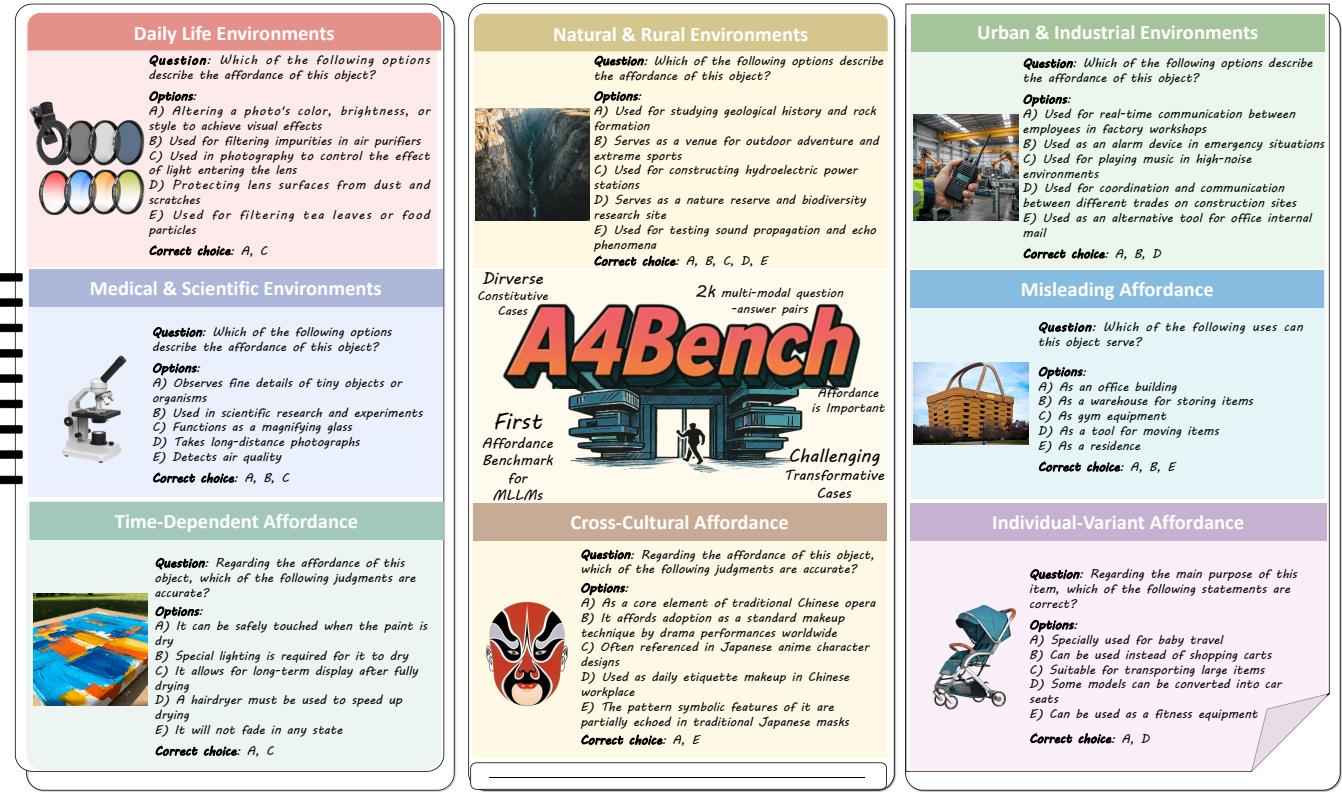


Figure 3: Typical samples from the **A4Bench**. Each sample is accompanied by a image-question-answer pair. **A4Bench** evaluates models across diverse disciplines (Constitutive Affordance) and challenging dimensions (Transformative Affordance), ensuring a comprehensive evaluation of the affordance perception capabilities.

Constitutive Affordance Perception We propose *A4Bench*, a robust benchmark designed to assess affordance perception in multimodal large language models. To evaluate basic affordance comprehension, the **Constitutive Affordance** component includes 1282 multimodal question-answer pairs across four disciplines and nine sub-disciplines, emphasizing inherent object properties such as shape, size, and material that determine walkability or graspability. These pairs probe static environmental cues critical for accurate affordance detection. The disciplines **Daily Life Environments** (226 pairs), **Natural and Rural Environments** (423 pairs), **Urban and Industrial Environments** (388 pairs), and **Medical and Scientific Environments** (245 pairs) comprise 64.1% of the benchmark. And the details of sub-discipline are illustrated in Figure 2.

Transformative Affordance Perception. To evaluate comprehension of dynamic object potential in MLLMs, the **Transformative Affordance** component of *A4Bench* employs 718 challenging multimodal question-answer pairs designed to probe complex context-sensitive affordance perception across diverse real-world scenarios. This component rigorously tests the ability of MLLMs to interpret nuanced interactions beyond static properties, ensuring robust evaluation of adaptive affordance understanding critical for practical applications. These pairs, categorized into four distinct aspects, address high-difficulty dynamic interactions:

1) **Misleading Affordance Perception** examines visually deceptive objects such as pillows resembling fire hydrants or phone cases mimicking slippers, testing perceptual accuracy under visual ambiguity in everyday contexts. 2) **Time-Dependent Affordance Perception** investigates affordance evolving over time, exemplified by a concrete surface transitioning from impassable to walkable after six months of curing or fruit ripening to edibility, reflecting temporal dynamics in natural processes. 3) **Cross-Cultural Affordance Perception** analyzes variations due to cultural contexts, where a thumbs-up gesture conveys approval in some cultures but signals taxi-hailing in others, highlighting cultural specificity in global interactions. 4) **Individual-Variant Affordance Perception** explores affordance differing by individual, such as a stroller providing seating for infants but not adults, emphasizing tailored interaction potential across populations with distinct needs.

Question Collection

Question Type In the **A4Bench**, two distinct question formats are employed: *Yes-or-No* and *What* questions. *Yes-or-No* questions (7.8%) evaluate fundamental judgment capabilities in MLLMs, while *What* questions (92.2%) demand deeper affordance comprehension due to their inherent complexity, facilitating a comprehensive and robust assessment of nuanced conceptual understanding.

Table 1: Benchmark results on the **A4Bench**, displaying *Exact match Accuracy*, with the best performance marked in **bold** and the second-best underlined for both proprietary and open-source MLLMs. The *Overall score of Constitutive Affordance* is the weighted average of Daily, Natural, Medical, and Urban disciplines. The *Overall score of Transformative Affordance* averages Time-Independent, Cross-Culture, and Individual dimensions. The *final Overall score* encompasses all pairs.

Categories	Constitutive Affordance					Transformative Affordance				Overall↑
	Daily↑	Natural↑	Medical↑	Urban↑	Overall↑	Misleading↑	Time↑	Culture↑	Individual↑	
HUMAN (BEST)	83.63%	87.35%	83.39%	87.50%	85.99%	87.17%	83.67%	84.75%	81.66%	83.14%
HUMAN (WORST)	79.20%	85.71%	78.83%	81.55%	81.88%	83.96%	81.43%	78.81%	76.03%	78.78%
Proprietary MLLMs:										
o3	14.64%	11.45%	10.28%	11.37%	11.76%	63.64%	10.95%	10.17%	7.39%	9.42%
CHATGPT-4O	16.41%	10.45%	9.63%	10.98%	11.52%	67.38%	13.33%	<u>16.10%</u>	7.39%	<u>11.68%</u>
GPT-4.1	18.18%	10.45%	9.95%	9.99%	11.57%	<u>68.98%</u>	8.10%	10.17%	2.96%	6.59%
GPT-4.1-MINI	12.87%	10.86%	9.45%	9.79%	10.59%	63.10%	7.24%	14.58%	2.81%	7.18%
GPT-4O	19.95%	10.45%	8.98%	9.39%	11.49%	65.24%	9.52%	9.32%	4.93%	7.72%
CLAUDE-3.5-SONNET	14.19%	<u>13.71%</u>	9.63%	10.38%	11.92%	61.50%	7.62%	14.41%	6.40%	8.66%
CLAUDE-3.7-SONNET	17.29%	17.80%	<u>11.26%</u>	<u>12.96%</u>	14.86%	45.99%	10.48%	24.58%	12.32%	14.31%
GEMINI-2.0-FLASH	17.29%	10.04%	9.95%	9.59%	11.15%	68.45%	5.71%	11.02%	5.91%	6.97%
GEMINI-2.0-PRO	18.26%	13.71%	14.19%	14.35%	14.82%	67.38%	9.52%	13.56%	4.43%	8.47%
Open-source MLLMs:										
DEEPSEEK-VL2	9.33%	9.63%	8.65%	8.79%	9.11%	69.50%	8.10%	12.71%	3.97%	7.54%
DEEPSEEK-VL2-SMALL	9.77%	8.82%	8.65%	8.79%	8.95%	65.78%	6.19%	11.86%	2.46%	6.03%
DEEPSEEK-VL2-TINY	9.77%	8.82%	8.65%	8.79%	8.95%	65.78%	6.13%	11.80%	2.43%	5.98%
INTERNVL3-14B	8.44%	8.41%	9.30%	9.19%	8.84%	60.43%	<u>11.43%</u>	15.25%	2.96%	9.04%
INTERNVL3-38B	9.33%	9.63%	9.30%	8.60%	9.17%	62.03%	7.62%	13.56%	2.96%	7.16%
MPLUG-OWL3-7B	10.21%	9.63%	8.33%	9.79%	9.54%	57.75%	7.62%	11.02%	0.99%	5.84%
QWEN2.5-VL-32B	8.88%	10.04%	9.30%	8.40%	9.15%	66.31%	7.14%	8.47%	1.97%	5.46%
QWEN2.5-VL-72B	10.65%	13.31%	9.95%	9.98%	11.10%	64.71%	8.10%	9.32%	5.42%	7.34%
Random guess	1.33%	1.55%	1.22%	2.60%	1.79%	44.92%	1.42%	0.85%	0.05%	0.77%
										5.55%

Human Expert Annotation A team of 60 human annotators, categorized by professional expertise into 10 senior experts, 20 junior researchers, and 30 general researchers, develops questions for **A4Bench**. Senior experts formulate transformative questions leveraging their extensive experience, while junior and general researchers focus on foundational questions, ensuring a comprehensive benchmark. The annotation process, conducted in controlled laboratory and online settings for consistency, involves sourcing or generating relevant images, designing precise questions, and defining their content and structure using specialized knowledge. Each question undergoes rigorous review by at least five additional expert annotators, whose critical feedback ensures accuracy, clarity, and alignment with the benchmark objectives, safeguarding the integrity and utility of the **A4Bench**.

Question Response Specifically, the example input query to MLLMs can be exemplified as:

- #User: Which of the following options can describe the affordance of this object as shown in the image?
A. This object can be used as a tool for determining directional position
B. It can be used as a tool for measuring precise time
C. Essential navigational equipment for wilderness exploration
D. Used as a tool for drawing accurate circles in drafting
E. A small device that operates using Earth's magnetic field

Answer directly using the letters in the given options.

During evaluation, answer candidates and correct responses are randomized to ensure impartiality. Given that MLLM responses may vary in format, we implement a prompt-human progressive choice evaluation technique to rigorously validate the accuracy of the responses.

Experiment Results

Benchmark Candidates

In A4Bench, **9 proprietary MLLMs** (*closed-source*) and **8 open-source MLLMs** are chosen for benchmarking. The **Proprietary MLLMs** (*closed-source*) include o3 (2025-04-16) (OpenAI 2025a), ChatGPT-4O (latest) (OpenAI 2024), GPT-4.1 (2025-04-14) (OpenAI 2025b), GPT-4.1-Mini (2025-04-14) (OpenAI 2025b), GPT-4O (2024-11-20) (Achiam et al. 2023), Claude-3.5-Sonnet (2024-10-22) (Anthropic 2024), Claude-3.7-Sonnet (2025-02-19) (Anthropic 2025), Gemini-2.0-Flash (Google 2024b), Gemini-2.0-Pro (exp-02-05) (Georgiev et al. 2024), which are all up-to-date and popular MLLMs. The **open-source MLLMs** cover DeepSeek-VL2 (Wu et al. 2024c), DeepSeek-VL2-Small (Wu et al. 2024c), DeepSeek-VL2-Tiny (Wu et al. 2024c), InternVL3-14B (Zhu et al. 2025), Intern-VL3-38B (Zhu et al. 2025), MpLUG-OWL3-7B (2024-11-01) (Ye et al. 2024), Qwen2.5-VL-32B (Instruct) (Bai et al. 2025), Qwen2.5-VL-72B (Instruct) (Bai et al. 2025), which are all

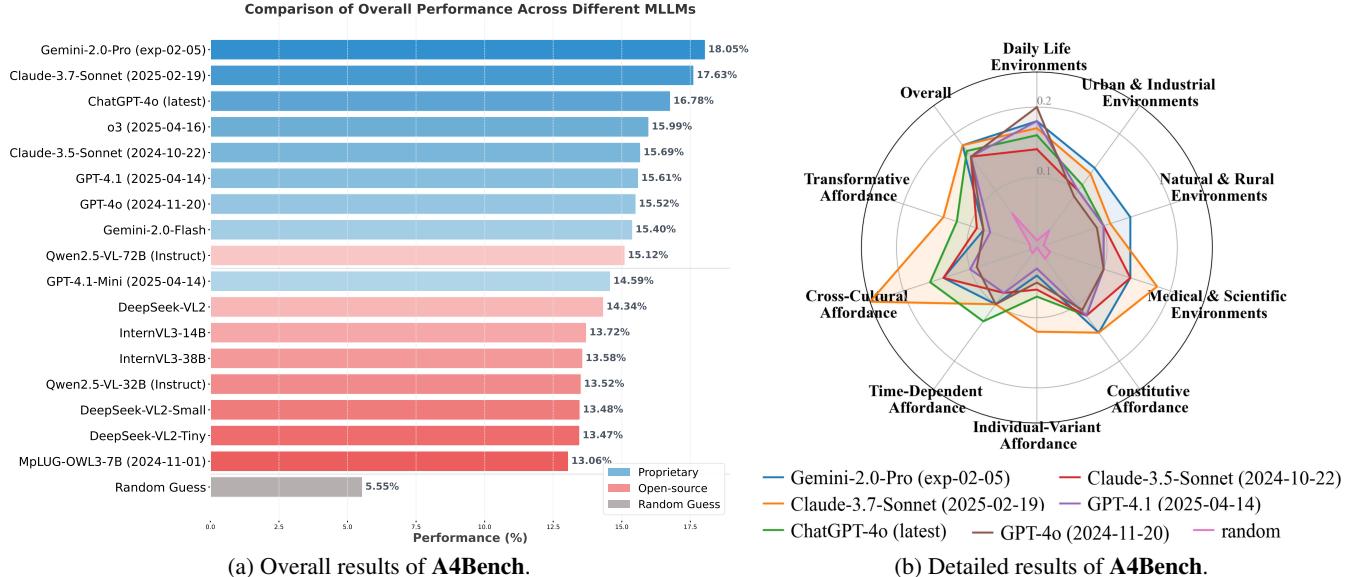


Figure 4: A Quick Look of the A4Bench Outcomes. (a) showcases a comparative analysis of the overall match score between 17 selected MLLMs (both closed-source and open-source), and random guess. (b) displays a radar chart that details the match score performance of the top-7 MLLMs across diverse disciplines (Constitutive Affordance) and challenging dimensions (Transformative Affordance excluding Misleading dimensions) of A4Bench.

competitive MLLMs. All MLLMs are tested with zero-shot setting. And the instruction prompts (e.g., concept and format prompts in this study) may vary slightly across different MLLMs, tailored to their respective official configurations.

Human Performance

To assess human performance on **A4Bench**, a user study is conducted in a controlled laboratory environment with carefully selected participants. Participants are initially familiarized with the task structure through exposure to representative examples, ensuring comprehension of the evaluation format. They subsequently provide responses to questions presented in **A4Bench**. To maintain experimental conditions analogous to those of MLLMs, question order is randomized, and participants receive only the concept prompt, format prompt, questions, and answer options, without supplementary information. Both the *best* and *worst* performance outcomes are documented for comparative analysis. This rigorous methodology ensures reliable and results. Statistical measures are applied to evaluate response consistency.

Findings of A4Bench

Proprietary MLLMs vs. Open-Source MLLMs vs. Human A concise overview of the **A4Bench** results is provided in Figure 4, which offering several key insights: 1) Proprietary MLLMs consistently outperform their open-source counterparts, with models like Gemini-2.0-Pro achieving an overall score of 18.05%, surpassing the best open-source model, Qwen2.5-VL-72B, at 15.12%. This gap underscores the advantage proprietary models have, likely due to access to more extensive and diverse training datasets. Although the performance margin is not always vast and the

first open-source models demonstrate competitiveness, proprietary systems currently maintain an advantage in overall affordance understanding as gauged by this benchmark. 2) All evaluated MLLMs show significantly limited capabilities in affordance perception, which is substantially below human-level performance. The top-performing MLLM, Gemini-2.0-Pro, lags behind the human best by 14.62%, and even the human worst at 81.25% surpasses the best MLLMs by 3.39%. This stark disparity highlights that current MLLMs, despite their advancements, struggle to replicate human-level understanding of affordance-related tasks, necessitating further development in this area. 3) All models outperform the random guess baseline, indicating some level of learned capacity, although insufficient for complex affordance perception and understanding.

As shown in Table 2, we further report comprehensive performance metrics beyond exact match accuracy, including Precision (Streiner and Norman 2006), Recall (Davis and Goadrich 2006), F1-Score, Partial Credit, AUC-ROC, and Hamming Loss, to provide a more nuanced evaluation. The results reveal several key findings: 1) Human performance substantially surpasses all MLLMs across every metric. 2) Proprietary MLLMs consistently outperform open-source counterparts, with Gemini-2.0-Pro leading among models. 3) Despite moderate Precision values, low Recall severely constrains the F1-Score, indicating difficulty in consistently identifying affordance. 4) Random guess achieves surprisingly high baseline Precision and Recall, but its extremely low Exact Match demonstrates poor practical utility. In general, these findings highlight the urgent need to advance the capacity of MLLMs to perceive context-sensitive and dynamic affordance.

Table 2: Comprehensive performance comparison of MLLMs and humans on A4Bench across multiple evaluation metrics.

Model	Precision↑	Recall↑	F1-Score↑	Exact Match↑	Partial Credit↑	AUC-ROC↑	Hamming Loss↓
MLLM (MLLM)							
Human (Best)	93.03%	94.16%	92.90%	85.34%	88.44%	90.81%	7.19%
Human (Worst)	89.73%	91.71%	89.68%	81.25%	84.23%	88.72%	9.73%
Proprietary MLLMs:							
O3	59.19%	35.36%	37.40%	15.99%	36.73%	61.24%	42.43%
CHATGPT-4O	59.17%	36.62%	38.32%	16.78%	37.36%	62.56%	42.25%
CLAUDE-3.7-SONNET	61.36%	37.74%	40.22%	17.63%	38.09%	62.56%	40.11%
GEMINI-2.0-PRO	68.27%	39.03%	41.91%	18.05%	40.64%	58.52%	39.72%
Open-source MLLMs:							
DEEPEEK-VL2	57.15%	34.57%	36.70%	14.34%	32.57%	59.69%	44.00%
INTERNVL3-38B	56.35%	32.95%	35.97%	13.58%	28.95%	56.74%	45.30%
MPLUG-OWL3-7B	55.60%	32.91%	35.83%	13.06%	27.91%	55.84%	44.38%
QWEN2.5-VL-72B	59.70%	33.42%	37.47%	15.12%	34.06%	60.58%	42.69%
<i>Random guess</i>	55.20%	57.81%	53.16%	5.55%	23.57%	51.63%	48.36%

Findings of Constitutive Affordance In the constitutive affordance category, which evaluates the ability to perceive affordance in static contexts, performance varies significantly across disciplines. 1) Proprietary models demonstrate a clear edge, with ChatGPT-4o and GPT-4-1 scoring 16.41% and 18.18% in the **Daily discipline**, respectively, indicating robust performance in understanding affordance in everyday scenarios. Similarly, in the **Natural discipline**, both models achieve 10.45%, reflecting an ability to handle natural environments effectively. However, their performance dips in more specialized disciplines, such as **Medical** (9.63% for ChatGPT-4o) and **Urban** (10.98% for GPT-4.1), suggesting limitations in domain-specific contextual understanding. 2) In contrast, open-source models such as DeepSeek-VL2 and InternVL3-38B demonstrate consistent yet comparatively lower performance across these disciplines, with scores ranging from 8.65% to 9.63% and 8.60% to 9.30%, respectively. This uniformity suggests a balanced but less specialized capability relative to proprietary models. 3) All MLLMs demonstrate suboptimal performance relative to human capabilities, exposing a substantial gap in interpreting constitutive affordance with the depth, nuance, and precision inherent in human understanding.

Findings of Transformative Affordance In the transformative affordance category, which evaluates the understanding of dynamic and contextually shifting affordance. The overall score for this category is calculated as a weighted average of the Time-Independent, Cross-Culture, and Individual dimensions. The exclusion of the Misleading dimension arises because its random guess rate is notably high, necessitating the subtraction of this baseline to accurately reflect the model's true performance in that dimension. Performance disparities across dimensions emerge, as depicted in Table 1. Specifically, 1) proprietary models like Gemini-2.0-Pro lead with a **Time-Dependent** score of 15.2% and a **Culture score** of 13.56%, demonstrating some proficiency in handling temporal transformations and cultural nuances. However, they struggle significantly in the **Misleading** dimension (67.38%) and the **Individual** di-

mension (4.43%), indicating challenges in detecting deceptive affordance and interpreting individual-specific contexts. Similarly, Claude-3.5-Sonnet achieves a balanced performance with 14.86% overall but shows weaknesses in **Individual** at 6.40%. 2) Open-source models, such as Qwen2.5-VL-72B, display more uniform but lower scores across these dimensions, ranging from 8.10% in **Misleading** to 9.32% in **Culture**, with a particularly poor **Individual** score of 5.42%. This suggests that open-source models lack the depth required for nuanced transformative affordance perception. 3) In comparison, human performance remains consistently high, ranging from 83.67% (Time-Dependent) to 84.75% (Culture), highlighting a significant gap. The underperformance of MLLMs in transformative affordance tasks, especially in dimensions requiring complex reasoning like **Misleading** and **Individual**, highlights the imperative for enhanced training strategies to improve their capacity to address dynamic and subjective affordance scenarios.

Conclusion

This paper introduces **A4Bench**, a novel and comprehensive benchmark featuring 2,000 multimodal question-answer pairs, specifically designed to systematically assess the affordance perception capabilities of MLLMs across two critical dimensions: constitutive affordance and transformative affordance. Our comprehensive evaluation of 17 MLLMs compared to human performance reveals substantial limitations in current model capabilities. All models significantly underperform relative to human understanding, particularly in the nuanced domain of transformative affordance. These findings reveal critical challenges for MLLMs in grasping contextual and dynamic affordances. A4Bench addresses this gap by providing a rigorous diagnostic framework to identify weaknesses in perception and reasoning. Furthermore, it offers a standardized benchmark to foster fair comparison, drive innovation, and guide the development of safer, more context-aware multimodal AI systems.

References

- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Agrawal, P.; Antoniak, S.; Hanna, E. B.; Bout, B.; Chaplot, D.; Chudnovsky, J.; Costa, D.; De Monicault, B.; Garg, S.; Gervet, T.; et al. 2024. Pixtral 12B. *arXiv preprint arXiv:2410.07073*.
- Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. <https://www.anthropic.com/news/clause-3-family>.
- Anthropic. 2025. Claude 3.7 Sonnet. <https://www.anthropic.com/news/clause-3-7-sonnet>.
- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; et al. 2025. Qwen2.5-VL Technical Report. *arXiv:2502.13923*.
- BailingMM-Pro-0120. 2025. [Https://github.com/wwbin2017/bailing/](https://github.com/wwbin2017/bailing/).
- Blake, R.; and Shiffrar, M. 2007. Perception of human motion. *Annu. Rev. Psychol.*, 58(1): 47–73.
- Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; and Zhao, F. 2024a. Are We on the Right Way for Evaluating Large Vision-Language Models? In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 27056–27087. Curran Associates, Inc.
- Chen, L.; Zhang, Y.; Ren, S.; Zhao, H.; Cai, Z.; Wang, Y.; Wang, P.; Meng, X.; Liu, T.; and Chang, B. 2024b. PCA-Bench: Evaluating Multimodal Large Language Models in Perception-Cognition-Action Chain. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 1086–1104. Bangkok, Thailand: Association for Computational Linguistics.
- Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. *arXiv preprint arXiv:2501.17811*.
- Dai, W.; Lee, N.; Wang, B.; Yang, Z.; Liu, Z.; Barker, J.; Rintamaki, T.; Shoeybi, M.; Catanzaro, B.; and Ping, W. 2024. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*.
- Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500*. Originally announced May 2023.
- Davis, J.; and Goadrich, M. 2006. The relationship between Precision-Recall and ROC curves. ICML ’06. New York, NY, USA: Association for Computing Machinery. ISBN 1595933832.
- Dong, H.; Kang, Z.; Yin, W.; Liang, X.; Feng, C.; and Ran, J. 2025. Scalable vision language model training via high quality data curation. *arXiv preprint arXiv:2501.05952*.
- Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context. *arXiv preprint arXiv:2403.05530*.
- Gibson, J. J. 2014. *The ecological approach to visual perception: classic edition*. Psychology press.
- Google. 2024a. Gemini. <https://gemini.google.com>. Large language model.
- Google. 2024b. Introducing Gemini 2.0: Our New AI Model for the Agentic Era. <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/>.
- Guo, D.; Wu, F.; Zhu, F.; Leng, F.; Shi, G.; Chen, H.; Fan, H.; Wang, J.; Jiang, J.; Wang, J.; et al. 2025a. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.
- Guo, Y.; Ji, K.; Zhu, X.; Wang, J.; Wen, F.; Li, C.; Zhang, Z.; and Zhai, G. 2025b. Human-Centric Evaluation for Foundation Models. *arXiv:2506.01793*.
- Huynh, N. D.; Bouadjenek, M. R.; Aryal, S.; Razzak, I.; and Hacid, H. 2025. Visual Question Answering: From Early Developments to Recent Advances – A Survey. *arXiv preprint arXiv:2501.03939*. Cs.CV, cs.MM.
- Jiao, Z.; and Li, X. 2025. An End-to-End Deep Graph Clustering via Online Mutual Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2): 3847–3854.
- Jiao, Z.; Zhang, H.; and Li, X. 2025a. Cnn2gnn: How to bridge cnn with gnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Jiao, Z.; Zhang, H.; and Li, X. 2025b. Deep Graph Multi-View Representation Learning With Self-Augmented View Fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 1–12.
- Li, B.; Ge, Y.; Ge, Y.; Wang, G.; Wang, R.; Zhang, R.; and Shan, Y. 2024a. SEED-Bench: Benchmarking Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13299–13308.
- Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024b. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Lin, J.; Yin, H.; Ping, W.; Lu, Y.; Molchanov, P.; Tao, A.; Mao, H.; Kautz, J.; Shoeybi, M.; and Han, S. 2023. VILA: On Pre-training for Visual Language Models. *arXiv:2312.07533*.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306. IEEE. CVPR 2024 Highlight.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, 26924–26958. Curran Associates, Inc. NeurIPS 2023 Oral Presentation.
- Liu, X.; Zhu, Y.; Gu, J.; Lan, Y.; Yang, C.; and Qiao, Y. 2024b. Mm-safetybench: A benchmark for safety evaluation

- of multimodal large language models. In *European Conference on Computer Vision*, 386–403. Springer.
- Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, 216–233. Springer.
- Liu, Y.; Zhao, Z.; Zhuang, Z.; Tian, L.; Zhou, X.; and Zhou, J. 2024d. Points: Improving your vision-language model with affordable strategies. *arXiv preprint arXiv:2409.04828*.
- Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Ye, H.-J. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*.
- MUG-U-7B. 2025. MUG-U. [Https://github.com/Shopee-MUG/MUG-U/](https://github.com/Shopee-MUG/MUG-U/).
- OpenAI. 2024. Hello GPT-4o. <https://openai.com/index/hello-gpt-4o/>.
- OpenAI. 2025a. Introducing OpenAI o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>.
- OpenAI. 2025b. OpenAI Models Documentation. <https://platform.openai.com/docs/models>.
- Phan, L.; Gatti, A.; Han, Z.; Li, N.; Hu, J.; Zhang, H.; et al. 2025. Humanity’s Last Exam. *arXiv:2501.14249*.
- Song, C. H.; Wu, J.; Washington, C.; Sadler, B. M.; Chao, W.-L.; and Su, Y. 2023. LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- step-1o. 2024. [Http://www.eecs.harvard.edu/mdw/proj-codeblue/](http://www.eecs.harvard.edu/mdw/proj-codeblue/).
- Streiner, D. L.; and Norman, G. R. 2006. “Precision” and “Accuracy”: Two Terms That Are Neither. *Journal of Clinical Epidemiology*, 59(4): 327–330.
- Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; Wang, B.; Chen, C.; Zhang, C.; Du, C.; Wei, C.; et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.
- Wang, J.; Zhang, H.; Wang, H.; and Yuan, Y. 2025a. Graph Convolutional Network With Self-Augmented Weights for Semi-Supervised Multi-View Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7): 12257–12270.
- Wang, J.; Zhang, H.; and Yuan, Y. 2025. Adv-CPG: A Customized Portrait Generation Framework with Facial Adversarial Attacks. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 21001–21010.
- Wang, J.; Zhang, Z.; Guo, Y.; Wen, F.; Shen, Y.; Liang, Y.; Wu, Y.; Li, W.; Li, C.; Chen, Z.; Jia, Q.; and Zhai, G. 2025b. The Ever-Evolving Science Exam. *arXiv:2507.16514*.
- Wen, F.; Guo, Y.; Wang, J.; Xiao, J.; Zhou, Y.; Li, C.; Zhang, Z.; and Zhai, G. 2025. Improve MLLM Benchmark Efficiency through Interview. *arXiv:2506.00883*.
- Wu, C.; Chen, X.; Wu, Z.; Ma, Y.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; Ruan, C.; et al. 2024a. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*.
- Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Li, C.; Sun, W.; Yan, Q.; Zhai, G.; and Lin, W. 2024b. Q-Bench: A Benchmark for General-Purpose Foundation Models on Low-level Vision. *arXiv:2309.14181*.
- Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; Yan, Q.; Min, X.; Zhai, G.; and Lin, W. 2023. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. *arXiv:2312.17090*.
- Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; et al. 2024c. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. *arXiv:2412.10302*.
- Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*.
- Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models. *arXiv:2408.04840*.
- Ying, K.; Meng, F.; Wang, J.; Li, Z.; Lin, H.; Yang, Y.; Zhang, H.; et al. 2024. MMT-bench: a comprehensive multi-modal benchmark for evaluating large vision-language models towards multitask AGI. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- Zhang, B.; Li, S.; Tian, R.; Yang, Y.; Tang, J.; Zhou, J.; and Ma, L. 2025a. Flash-VL 2B: Optimizing Vision-Language Model Performance for Ultra-Low Latency and High Throughput. *arXiv preprint arXiv:2505.09498*.
- Zhang, Z.; Su, S.; Zhu, Y.; Yan, Q.; Sun, J.; and Zhang, Y. 2024. A-Bench: Are LMMs Masters at Evaluating AI-Generated Images? *arXiv preprint arXiv:2406.03070*.
- Zhang, Z.; Sun, W.; Min, X.; Wang, T.; Lu, W.; and Zhai, G. 2022. No-Reference Quality Assessment for 3D Colored Point Cloud and Mesh Models. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7618–7631.
- Zhang, Z.; Wang, J.; Guo, Y.; Wen, F.; Chen, Z.; Wang, H.; Li, W.; Sun, L.; Zhou, Y.; Zhang, J.; Yan, B.; Jia, Z.; Xiao, J.; Tian, Y.; Zhu, X.; Zhang, K.; Li, C.; Liu, X.; Min, X.; Jia, Q.; and Zhai, G. 2025b. AIBench: Towards Trustworthy Evaluation Under The 45° Law. <https://aiben.ch/>.
- Zhang, Z.; Zhao, X.; Fang, X.; Li, C.; Liu, X.; Min, X.; Duan, H.; Chen, K.; and Zhai, G. 2025c. Redundancy Principles for MLLMs Benchmarks. *arXiv preprint arXiv:2501.13953*.
- Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*. Originally announced April 2023.
- Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv:2504.10479*.