Res-Bench: Benchmarking the Robustness of Multimodal Large Language Models to Dynamic Resolution Input

Chenxu Li[*1], **Zhicai Wang**[*1], **Yuan Sheng**[1], **Xingyu Zhu**[1], **Yanbin Hao**[2†], **Xiang Wang**[1]

[1] University of Science and Technology of China
[2] Hefei University of Technology
{cxli2024,wangzhic,yuansheng,xyzhuxyz}@mail.ustc.edu.cn,haoyanbin@hotmail.com,xiangwang1223@gmail.com

Abstract

Multimodal Large Language Models (MLLMs) increasingly support dynamic image resolutions. However, current evaluation paradigms primarily assess semantic performance, overlooking the critical question of resolution robustness - whether performance remains stable across varying input resolutions. To address this gap, we introduce **Res-Bench**, a comprehensive benchmark comprising 14,400 samples across 12 resolution levels and six core capability dimensions. We designed a novel evaluation framework that goes beyond traditional accuracy metrics to capture performance stability. This framework introduces multiple robustness metrics: Spearman's correlation for assessing resolution-performance trends, and Absolute/Relative Continuous Error (ACE/RCE) for measuring performance volatility. Using these metrics, we conducted a large-scale evaluation of leading MLLMs. Our analysis encompasses: (1) model-centric and task-centric robustness examination, (2) investigation of preprocessing strategies including padding and super-resolution, and (3) exploration of fine-tuning for stability enhancement.

Code — https://github.com/cxli2333/Res-Bench
Datasets —
https://huggingface.co/datasets/cxcx2333/ResBench
Extended version — https://arxiv.org/abs/2510.16926

Introduction

Visual perception and reasoning constitute fundamental components of human perceptual and cognitive capabilities. Notably, humans demonstrate remarkable robustness in processing and interpreting visual information across varying image resolutions. This is largely attributable to the adaptive capabilities of our visual system and the brain's cognitive completion mechanisms (Marr 2010). Specifically, we can adaptively focus on important regions of an image, and for low-resolution images, our brains can fill in missing details based on past experiences and prior knowledge. The quest to equip machines with such sophisticated perceptual abilities is a central theme in modern AI research. However, while multimodal large language models have demonstrated exceptional capabilities across various domains, it remains unclear how well MLLMs can achieve resolution robustness comparable to human visual systems.

Contemporary advanced MLLMs have incorporated sophisticated designed architectures to enable support for dynamic resolution input. These strategies generally fall into two categories: (1) **Native dynamic processing method**, such as the Vision Transformer (ViT) (Dosovitskiy et al. 2021) in Qwen2.5-VL (Bai et al. 2025) which employs Multimodal Rotary Position Embedding (MRoPE) and window attention to process images at their native resolution; and (2) **Patch-based method**, utilized by models like InternVL2.5 (Chen et al. 2024b), LLaVA-OneVision (Li et al. 2024b) and LLaVA-UHD (Guo et al. 2024), which segment high-resolution images into smaller sub-images and thumbnails for individual processing. Current research paradigms predominantly focus on benchmarking model performance at the semantic level, including visual perception (Fu et al. 2024), visual question answering (VQA) (Antol et al. 2015), visual reasoning (Lu et al. 2024), *etc.*. However, insufficient attention has been directed to the robustness discussion at the visual processing level. For each MLLM, three critical ingredients contribute to the resolution robustness: (1) the processing mechanism for dynamic resolution input entailed in the vision encoder (Liu et al. 2021), (2) the level of question spanning both perceptual and cognitive dimensions (Vaishnav and Tammet 2025), and (3) the distribution of training data (Bai et al. 2024).

To systematically investigate this gap, we introduce Res-Bench, a new benchmark specifically designed to evaluate the resolution robustness of MLLMs. Res-Bench is constructed based on 1,200 high-quality image-question pairs carefully selected by humans. For each image, we generate a series of lower-resolution versions via downsampling. The benchmark features a diverse set of questions across six primary categories and 15 fine-grained subcategories (see Section 3.2), covering a wide range of visual-linguistic tasks. We employ a suite of four metrics to comprehensively assess model behavior: overall performance is measured by accuracy, the performance trend by Spearman's correlation coefficient, and stability by Absolute/Relative Continuous Error (ACE/RCE), as detailed in Section 3.3.

We conduct extensive experiments on Res-Bench to eval-

[*]These authors contributed equally.
[†]Corresponding author

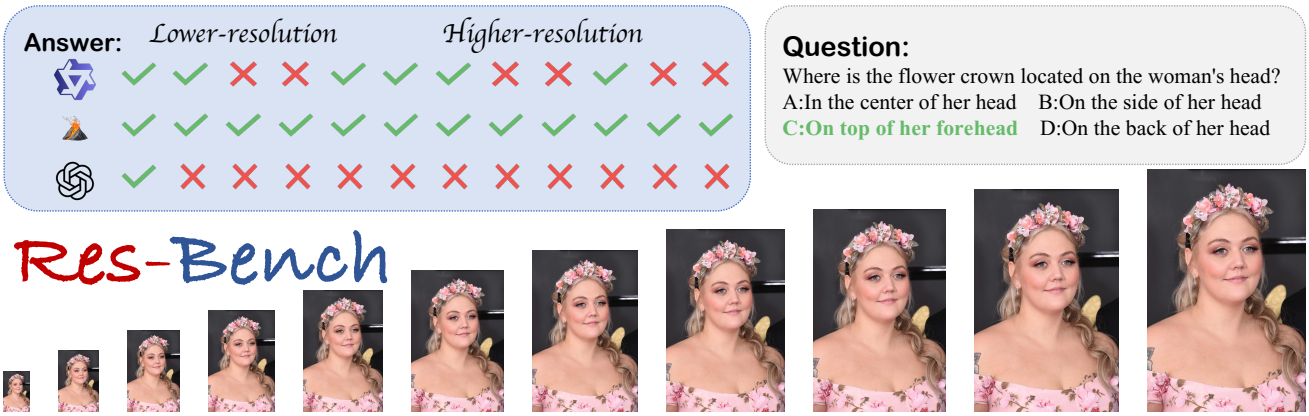

Figure 1: An Example of ResBench. There is a lack of resolution robustness in MLLMs, where models' performance on the given task fluctuates unpredictably as the input resolution increases.

uate their robustness to visual inputs. Additionally, for the obtained low-resolution images, we employ data preprocessing methods like white-padding and super-resolution, aiming to validate their effectiveness in both enhancing absolute performance and improving resolution robustness (Zhu et al. 2024a,b). The evaluation covers multiple proprietary models (GPT-4o((OpenAI 2024), and Gemini-1.5 Pro (Team 2024)) and open-source models (*e.g.*LLaVA-OneVision (Li et al. 2024b), Qwen2.5-VL (Bai et al. 2025), mPLUG-Owl3 (Ye et al. 2024), InternVL2.5(Chen et al. 2024b), *etc.*). We find there is a pervasive lack of resolution robustness. To understand the underlying factors contributing to this instability, we conducted an in-depth analysis of the experimental results, revealing several novel findings that pose new challenges for the development of MLLMs:

- **Architectural Trade-offs:** Native processing methods tend to achieve higher peak performance but are less robust to resolution changes. Conversely, patch-based methods demonstrate better robustness but at the cost of lower overall performance.
- **Task-Dependent Robustness:** Our analysis reveals that resolution robustness is highly task-dependent. While some tasks are largely immune to resolution degradation, others show a strong, positive correlation between input quality and performance.
- **Enhanced Visual Evidence via Preprocessing:** Information-free padding shows that performance peaks at a moderate input size, declining if the padding becomes excessive. Super-resolution provides a more significant boost by restoring visual information, outperforming padding.
- **Enhancing Robustness via Fine-tuning:** Fine-tuning on a resolution-balanced dataset can notably enhance a model's resolution robustness. The enhanced resolution robustness of the fine-tuned model shows generalization to out-of-distribution data.

Related Work

Multimodal Large Language Model. A fundamental challenge in Multimodal Large Language Models (MLLMs) (Bai et al. 2025; Li et al. 2024b; Team 2024) lies in effectively processing visual information, particularly at high resolutions where fine-grained details are critical. Early models addressed this by resizing images to a low, fixed resolution, a method prone to distorting critical details necessary for tasks like OCR or fine-grained perception (Radford et al. 2021; Bai et al. 2023). To address these issues, several common solutions have emerged: (1) Native resolution processing methods, such as those used in Qwen2.5-VL (Bai et al. 2025) and Kimi-VL (Team et al. 2025). (2) Patch-based method: employed by models like LLaVA-OneVision (Li et al. 2024b), InternVL2.5 (Chen et al. 2024b), and DeepSeek-VL2 (Wu et al. 2024). While both approaches handle high-resolution inputs, the trade-offs between them regarding resolution robustness remain largely (Zhu et al. 2025a, 2024c). This paper systematically investigates the impact of these distinct architectural choices on model stability.

MLLM Benchmarks. The evaluation of MLLMs has rapidly evolved, moving from single-task benchmarks like VQA (Antol et al. 2015; Bigham et al. 2010), caption (Lin et al. 2014; Plummer et al. 2015) to comprehensive, multi-faceted evaluation suites. Recent efforts have produced a rich ecosystem of benchmarks targeting a wide array of high-level capabilities, including: Visual perception (Fu et al. 2024), Chart reasoning (Masry et al. 2022, 2025; Wang et al. 2024c), Mathematical reasoning (Lu et al. 2024; Zhang et al. 2024), Spatial reasoning (Wang et al. 2024a; Li et al. 2025; Xu et al. 2025), Expert-level multimodal understanding (Yue et al. 2024, 2025), Optical Character Recognition (OCR) capabilities (Liu et al. 2024; Yang et al. 2024). Additionally, comprehensive benchmarks (Li et al. 2023; Chen et al. 2024a) have been developed to evaluate overall multimodal intelligence. While their source images may have varying native resolutions, these benchmarks ignore the impact of the resolution distribution.

Robustness Evaluation. Evaluating the robustness of

Benchmark	Dimension	Task	Robustness
MLLM-COMPBENCH	8	Multi-images MCQ	Relative
MMVU	12	Multi-images MCQ	Relative
RobustBench	15	Single-image MCQ	Absolute
NaturalBench	1	Single-image MCQ	Relative
Res-Bench (ours)	15	Single-image & Multi-images MCQ, VQA	Absolute, Relative

Table 1: Robustness Benchmarks and Res-Bench. The comparison highlights key differences in scope, including task diversity and the robustness evaluation methodology.

MLLMs—their ability to maintain stable performance against input perturbations—is a critical area of research. Existing work has primarily focused on several key dimensions: (1) robustness against semantic-level perturbations (Cao et al. 2024; Liu et al. 2025); (2) robustness against stylistic or natural variations in images (Cai et al. 2023; Li et al. 2024a); and (3) robustness against common image corruptions (Croce et al. 2021; Li et al. 2024c; Zhu et al. 2025b). A direct comparison of these approaches, summarized in Table 1, reveals a shared limitation: while they comprehensively examine model stability, they primarily test against content-level corruptions and adversarial inputs. The impact of fundamental visual properties like resolution on model performance remains a critical, under-explored area. We introduce **Res-Bench** to systematically fill this gap, providing the first dedicated benchmark for evaluating MLLM resolution robustness.

Res-Bench

We present our dataset, Res-Bench, a comprehensive benchmark for evaluating the dynamic resolution robustness of MLLMs. This section details the construction of Res-Bench, beginning with our multi-stage data collection process in Section 3.1. In Section 3.2 we provides an overview of the dataset's composition and statistical properties. In Section 3.3, we introduce four benchmark-specific metrics designed to assess model robustness.

Data Collection Process

Res-Bench is constructed through a rigorous, multi-stage pipeline designed to ensure data quality, diversity, and relevance. The process begins with establishing strict collection standards, followed by stringent automated filtering and expert manual verification.

Data Collection Standards. We defined three core principles for sample selection: **(1) Visual Dependency.** When provided with only text input, the model fails to correctly answer the sample questions. **(2) Diversity.** The dataset must span a wide range of MLLM capabilities—from low-level perception (*e.g.*, text recognition, counting) to high-level reasoning (*e.g.*, spatial relationships, mathematics)—to enable a holistic evaluation. **(3) Pristine Image Quality.** Source images must be of high resolution and quality to serve as a reliable, artifact-free baseline. This is essential for systematically studying the effects of controlled resolution degradation via downsampling.

Preliminary Data Filtering. To establish a comprehensive and high-quality benchmark, we selected 13 existing benchmarks as data sources. These include benchmarks designed for general evaluation of MLLMs (*e.g.*, SEED (Li et al. 2023)) as well as specialized benchmarks for assessing specific capabilities of MLLMs (including AI2D (Kembhavi et al. 2016), HR-Bench (Wang et al. 2024b), NaturalBench (Li et al. 2024a), SpatialEval (Wang et al. 2024a), MMMU (Yue et al. 2024), MMMU-Pro (Yue et al. 2025), ChartQA-Pro (Masry et al. 2025), BLINK (Fu et al. 2024), MathVerse (Zhang et al. 2024), MathVista (Lu et al. 2024), OCRBench (Liu et al. 2024), CCOCR (Yang et al. 2024)). To ensure the data consists exclusively of high-quality images, we first conducted preliminary filtering by selecting all high-resolution images (with both width and height exceeding 1344 px) as the foundational dataset. Building upon this foundation, we further refined the dataset by leveraging powerful models to select samples that meet visual dependency criteria. Specifically, we employed GPT-4o (OpenAI 2024) and InternVL-2.5-72B (Chen et al. 2024b) as evaluators for this filtering process. We conducted text-only input evaluations, while employing circular evaluation to mitigate random biases. A sample was automatically discarded if both models could answer it correctly without the image, as this indicated weak visual dependency. After this filtering process, the number of candidate samples was reduced from 69840 to 2868.

Manual Verification. To further ensure data quality, we performed human expert review on the pre-filtered samples. The evaluation criteria remained consistent with the aforementioned standards. Additionally, we further optimized the data distribution to ensure better balance. To comprehensively evaluate the capabilities of large models across diverse question types and facilitate systematic assessment, we further refined the phrasing of questions for a subset of the data. For instance, we restructured the single-image multi-question format in NaturalBench (Li et al. 2024a) to a single-question multi-image paradigm, while also refining certain question formulations in CCOCR (Yang et al. 2024). We then systematically generated the full Res-Bench dataset by creating multiple resolution versions for each image, resulting in a total of 14,400 data instances for evaluation. Our data filtering pipeline is illustrated in Figure 2.

Comprehensive Dataset Analysis

Our benchmark comprehensively evaluates 6 core capability dimensions with 15 fine-grained sub-capabilities, compris-

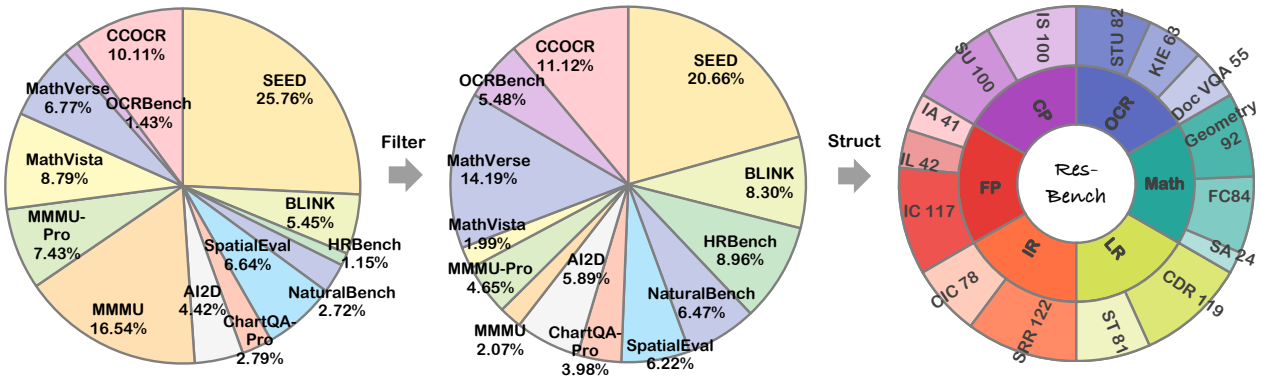

Figure 2: ResBench Dataset Construction Overview. The pie charts illustrate the proportional contribution of each source dataset before and after our filtering process. The final sunburst chart shows the composition of the selected data, organized by our six core capability dimensions and 15 sub-tasks. Tasks and sub-tasks are abbreviated by their initials in the figure.

ing 200 high-quality samples per dimension. The data encompasses diverse question formats including: Single-image multiple-choice questions (MCQ), Multi-image MCQ, VQA. We systematically generated 14,400 images across 12 resolution groups through controlled resizing of the 1,200 original samples, aligning each image to fixed target dimensions along its longer edge. Res-Bench includes six core capabilities (including Coarse Perception, Fine-grained Perception, Instance Reasoning, Logical Reasoning, Mathematical, OCR) and 15 sub-capabilities. The distribution across capability dimensions is as shown in Figure 2. More detailed capability definitions and cases can be seen in the appendix.

Evaluation Metrics

To quantify model robustness against resolution changes, we employ a suite of four complementary metrics.

Accuracy. We measure the model's fundamental capability on a given task at a specific resolution level. The accuracy at a resolution res , denoted asAcc_{res} is calculated as:

$$Acc_{res} = Score(GT, MLLM(I_{res})), \tag{1}$$

where $M(I_{res})$ is the model's prediction for an image I at resolution res G is the ground truth, and the $Score$ function returns 1 for a correct answer and 0 otherwise. To gauge the model's overall performance across all tested resolutions, we then compute the **Average Accuracy**:

$$Acc_{avg} = \frac{1}{N_{res}} \sum_{i=1}^{N_{res}} Acc_{res_i}, \tag{2}$$

where N_{res} is the total number of resolution levels (12 in Res-Bench), and res_i denotes the i-th resolution.

Spearman's Correlation Coefficient. While Acc_{avg} provides an overall score, it does not capture the relationship between resolution and performance. The Spearman's correlation coefficient assesses the ordinal relationship between two variables. We treat the rank of the resolution and the rank of the accuracy as the two variables to assess their correlation. The detailed definition is as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}, \tag{3}$$

where d_i denotes the rank difference, and n denotes the resolution-sample size, which is 12 in our Res-Bench.

Absolute Continuous Errors. When two models may have equal average accuracy and Spearman's correlation coefficient but differ in the fluctuation magnitude of their accuracy as resolution changes, it is necessary to quantify such fluctuations. We define Absolute Continuous Errors(ACE) as the sum of absolute differences in accuracy between adjacent resolutions:

$$ACE = \sum_{i=1}^{n-1} \left| Acc_{res_{i+1}} - Acc_{res_i} \right|, \tag{4}$$

where res_i means the i-th resolution.

Relative Continuous Errors. To enable fair comparison between models with different capability levels, RCE normalizes the fluctuation by the average accuracy:

$$RCE = \frac{ACE}{Acc_{avg}}. \tag{5}$$

RCE allows for a fairer comparison of robustness between models with different overall capability levels.

Experiment

In this section, we conduct a comprehensive experimental evaluation to assess the resolution robustness of state-of-the-art MLLMs using our proposed Res-Bench benchmark and metrics. We begin by detailing our experimental setup. Section 4.1 presents an analysis of the results from eight leading models on Res-Bench, providing a deeper, multi-faceted analysis of these results. Section 4.2 investigates the impact of data preprocessing techniques. Section 4.3 explores the potential of fine-tuning as a strategy to enhance robustness. Detailed evaluation criteria can be seen in Appendix.

Baseline. We use Res-Bench to evaluate several leading MLLMs. This includes two powerful proprietary models, GPT-4o (OpenAI 2024) and Gemini 1.5-Pro (Team 2024), and six powerful open-source models employing different dynamic resolution strategies. Qwen2.5-VL (Bai et al. 2025)

Model	Resolution of Input												$Acc_{avg.}\uparrow$	$\rho\uparrow$	ACE↓	RCE↓
	112	224	336	448	560	672	784	896	1008	1120	1232	1344				
Proprietary Model																
Gemini1.5-Pro	0.420	0.482	**0.517**	**0.571**	**0.584**	**0.590**	**0.592**	0.596	0.596	**0.591**	0.603	0.595	**0.561**	0.895	0.201	0.358
GPT-4o	**0.423**	0.474	0.498	0.510	0.522	0.534	0.548	0.548	0.564	0.557	0.560	0.567	0.526	0.972	0.157	0.299
Open-source Model																
Qwen2.5-VL†	0.362	0.445	0.504	0.532	0.565	0.575	0.582	**0.598**	**0.607**	0.586	**0.609**	**0.599**	0.547	0.951	0.299	0.547
Kimi-VL†	0.369	0.415	0.469	0.512	0.541	0.542	0.547	0.568	0.559	0.567	0.565	0.573	0.519	0.951	0.226	0.437
LLaVA-OV‡	0.420	**0.489**	0.505	0.515	0.529	0.525	0.532	0.537	0.541	0.555	0.542	0.538	0.519	0.944	0.159	0.306
InternVL-2.5‡	0.403	0.460	0.498	0.533	0.546	0.552	0.553	0.557	0.551	0.559	0.559	0.556	0.527	0.909	0.170	0.323
MiniCPM-o-2.6‡	0.363	0.436	0.472	0.503	0.534	0.539	0.540	0.551	0.563	0.572	0.572	0.585	0.510	**1.000**	0.222	0.428
mPLUG-Owl3	0.333	0.355	0.379	0.380	0.369	0.391	0.396	0.400	0.410	0.405	0.407	0.404	0.385	0.916	**0.109**	**0.282**

Table 2: Main Results. The central columns show model accuracy at each of the 12 specific resolutions. The final four columns present the overall average accuracy and three robustness metrics. † means the models directly process images at their native dynamic resolution. ‡ means the models process high-resolution images by dividing them into smaller patches or sub-images. The best results are highlighted in bold, respectively. ↑: higher is better. ↓: lower is better.

, Kimi-VL-Instruct (Team et al. 2025) directly process native high-resolution images, while InternVL-2.5 (Chen et al. 2024b), LLaVA-OneVision (Li et al. 2024b) , MiniCPM-o-2.6 (Hu et al. 2024) use patch-based method. We also tested model that do not support dynamic resolution for single images, mPLUG-Owl3 (Ye et al. 2024).

Experiment Analysis

Main Result Analysis. The overall performance of the evaluated models on Res-Bench is presented in Table 2. Our findings indicate that Res-Bench poses a significant challenge to current state-of-the-art MLLMs. The top-performing models, Gemini-1.5 Pro and Qwen2.5-VL, achieve an average accuracy(Acc_{avg}) below 60%. A central finding is the pervasive lack of resolution robustness. Most models exhibit unstable performance across different resolutions. This is highlighted by the metrics for Qwen2.5-VL, which, despite its high accuracy, shows the most fluctuation with the worst ACE and RCE scores among all models. In stark contrast, MiniCPM-o-2.6 demonstrates a perfectly monotonic performance trend ($\rho = 1$), where accuracy strictly increases with resolution, though its absolute accuracy is not the highest. This reveals a clear and critical trade-off between peak performance and robustness.

Model-centric Analysis. We now dissect the results from a model-centric perspective, focusing on how architectural choices influence resolution robustness: proprietary models: A distinct trade-off is also visible between the two leading proprietary models. Gemini-1.5 Pro achieves one of the highest accuracy scores but demonstrates moderate robustness. Conversely, GPT-4o, while having a slightly lower average accuracy, exhibits superior stability with more favorable robustness metrics (lower ACE/RCE). The open-source models reveal a clear divergence based on their high-resolution processing strategy: **(1) Native dynamic processing models:** These models, which process images at their native resolution, generally achieve higher peak accuracy, particularly on high-resolution inputs. However, they are highly sensitive to resolution degradation, resulting in poor performance at lower resolutions and, consequently, high fluctuation scores (ACE/RCE). **(2) Patch-based processing models:** These models, which rely on image patches and thumbnails, demonstrate superior robustness. Their performance is more stable across resolutions, and they perform comparatively better on low-resolution inputs. The trade-off is a lower performance ceiling, as their accuracy does not scale as effectively with increasing resolution. **(3) Fix-resolution models:** mPLUG-Owl3, which lacks a dynamic resolution mechanism, lags significantly in overall accuracy. While its ACE and RCE scores appear low, we posit this is likely an artifact of its compressed performance range rather than an indicator of true robustness.

Task-centric Analysis. In Figure 3, analyzing performance by capability dimension reveals that sensitivity to resolution varies significantly across tasks, as clearly indicated by the linear regression trends for each task category. For Coarse Perception (CP) tasks, the regression line demonstrates an almost flat slope ($y = 0.000017x + 0.7214$), indicating that these tasks are highly robust to resolution loss. Models maintain strong performance even at very low resolutions (e.g., 112px), and further increasing resolution yields negligible improvements. This suggests that holistic understanding requires minimal visual detail. For Fine-grained Perception (FP), Instance Reasoning (IR), Logical Reasoning (LR), and Mathematical Reasoning (Math), the regression lines exhibit weak positive slopes . This indicates only a marginal correlation between resolution and accuracy ($\rho \approx 0.5$). Furthermore, significant accuracy fluctuations between adjacent resolution steps are observed, highlighting the instability of model performance in these tasks. Merely increasing resolution does not guarantee better or more stable reasoning for these categories. For OCR tasks, the regression slope is considerably steeper ($y = 0.000372x + 0.2610$), reflecting a strong dependency on resolution. A distinct threshold effect is observed: performance degrades severely at lower resolutions, but once a minimal clarity threshold is reached, accuracy improves significantly with higher resolutions before plateauing. This suggests that OCR is highly resolution-sensitive but becomes robust once sufficient clarity is achieved. The linear regression trends across all tasks

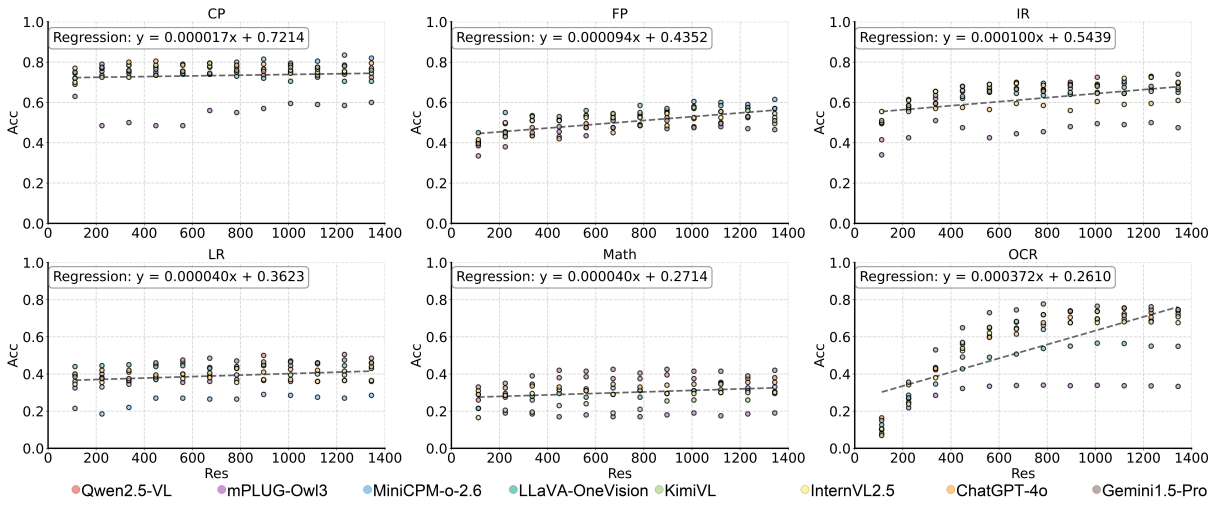

Figure 3: Task-level Results. We show the result with different input resolution across six core tasks. To better show the difference of resolution robustness among tasks, we show the results of linear regression on all data points in each subplot. CP: Coarse-grained Perception, FP: Fine-grained Perception, IR: Instance Reasoning, LR: Logical Reasoning, Math: Mathematical Reasoning, OCR: Optical Character Recognition.

provide a key quantitative perspective on resolution robustness, reinforcing the task-dependent nature of model sensitivity to resolution changes.

Analysis of Preprocessing Strategies

To better understand the performance degradation on low-resolution images, we conducted experiments to disentangle two factors: reduced information content and shorter visual token length. We investigate this through two augmentation strategies: white-padding and super-resolution (SR). The example of processing strategies are shown in Figure 4.

Strategy	Padding		SR	
Final Res	224	448	224	448
224	0.445		0.445	
336	$0.447_{(+0.002)}$			
448	$0.448_{(+0.003)}$	0.532	$0.473_{(+0.028)}$	0.532
672	$0.453_{(+0.008)}$	$0.550_{(+0.018)}$		
784	$0.447_{(+0.002)}$	$0.544_{(+0.012)}$		
896	$0.438_{(-0.007)}$	$0.553_{(+0.021)}$	$0.478_{(+0.032)}$	$0.558_{(+0.026)}$
1008	$0.443_{(-0.002)}$	$0.538_{(+0.006)}$		
1232	$0.434_{(-0.011)}$	$0.541_{(+0.009)}$		
1344	$0.440_{(-0.005)}$	$0.525_{(-0.007)}$		

Table 3: Impact of Image Padding and Super-Resolution. Columns denote the initial image resolution (224px or 448px), while rows indicate the final resolution after processing. The baseline performance for each unprocessed initial image is shown on the diagonal where the initial and final resolutions are identical. SR: Super-Resolution.

Isolating the Effect of Token Length via Padding. A high-resolution image, when processed by a ViT, typically results in a longer sequence of visual tokens than a low-resolution one. To isolate the effect of this token length, we designed an experiment where information content was held constant while token length was increased. We took low-resolution images (specifically, those at 224px and 448px) and padded them with zero-value pixels to match the dimensions of various higher resolution targets.

For both the 224px and 448px source images, padding to a moderately higher resolution leads to a performance improvement, even though no new visual information was added. However, this trend reverses when padding to the higher resolutions, where we observe a slight decrease in accuracy. It may be because of the higher visual/textual token length ratio, which increases the model's attention to visual input, making the model more reliant on visual input rather than relying on prior knowledge. We hypothesize that while a longer token sequence can be beneficial, an excessive number of non-information padded tokens can act as noise, diluting the model's attention on the actual visual content and impairing its performance.

Restoring Information Content via Super-Resolution. We investigated whether algorithmically restoring visual details could recover lost performance more effectively. We employed DiffIR (Xia et al. 2023), a classical diffusion-based SR model, to upscale low-resolution images (224px and 448px) to higher resolutions (448px and 896px). Unlike padding, SR aims to intelligently "fill in" missing information, enhancing image clarity. As shown in Table 3, SR-enhanced images significantly outperform their original low-resolution counterparts and also surpass the performance of the simple padded images.

Discussion. Our experiments reveal two key findings: (1) Token sequence length affects performance - moderate padding is beneficial but excessive padding is detrimental; (2) Visual information quality is crucial - while SR techniques improve performance significantly, they cannot fully match the quality of original high-resolution images.

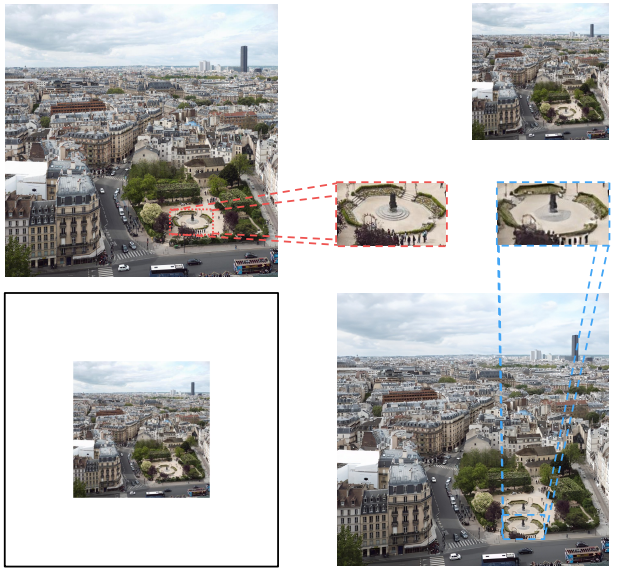

Figure 4: Example of Preprocessing. (a) Image padding by adding non-imformation pixels around images. (b) Image super-resolution using off-the-shelf SR models.

Enhancing Robustness via Fine-tuning

To explore whether resolution robustness can be learned, we investigated if fine-tuning on a mixed-resolution dataset could improve a model's stability. We fine-tuned the Qwen2.5-VL-3B model specifically on spatial reasoning tasks. The detailed experimental settings and hyperparameters are in the appendix.

Task	ACE↓		*Acc* ↑	
	w/o FT	FT	w/o FT	FT
CP	0.369	$0.433_{+0.064}$	0.678	$0.670_{-0.008}$
FP	0.578	$0.546_{-0.043}$	0.510	$0.495_{-0.015}$
IR	0.539	$0.478_{-0.061}$	0.454	$0.607_{+0.153}$
LR	0.309	$0.281_{-0.028}$	0.469	$0.445_{-0.024}$
Math	0.658	$0.576_{-0.082}$	0.266	$0.295_{+0.029}$
OCR	1.230	$1.281_{+0.051}$	0.540	$0.540_{+0.000}$

Table 4: The Impact of Fine-tuning. The table compares the metrics of the model before (w/o FT) and after (FT) being fine-tuned on spatial reasoning tasks.

Results and Analysis. The results, detailed in Table 4, show that this strategy was highly effective. The fine-tuned 3B model not only became more accurate but also significantly more robust in its target domain. On the 'Instance Reasoning' dimension, the model's performance improved dramatically. Accuracy surged from 0.454 to 0.607, while the RCE dropped sharply from 0.539 to 0.478. This indicates a clear, simultaneous improvement in both capability and stability. The fine-tuning experiment also demonstrated the model's generalization ability. Overall average accuracy increased from 0.486 to 0.508, while both ACE (0.224 to 0.220) and RCE (0.461 to 0.433) decreased, confirming a general enhancement of robustness. There is also a Robustness-Performance Trade-off. Interestingly, on some related but out-of-domain tasks like 'Fine-grained Perception', we observed a slight decrease in absolute accuracy. However, even in this case, the model's robustness improved, as evidenced by a lower RCE. This suggests that the model learned a more stable general representation, even if it came at the cost of peak performance on certain tasks it was not explicitly trained on. this experiment demonstrates that mixed-resolution fine-tuning is a promising and effective strategy for directly enhancing the resolution robustness of MLLMs. It teaches the model to handle visual inputs more consistently across quality levels.

Limitations and Future Work

While this work provides the first systematic evaluation of MLLM resolution robustness, we acknowledge several limitations that open avenues for future research:

Limited Range of Resolution Levels. Our existing Res-Bench supports a maximum resolution of approximately 1.5K and does not cover increasingly prevalent ultra-high-resolution scenarios such as 4K and 8K. Evaluating models on such inputs can reveal new challenges related to computational efficiency, architectural bottlenecks, and high information density processing.

Narrow Exploration of Preprocessing Techniques. The research on performance recovery is limited to the DiffIR diffusion-based super-resolution method. Future work can compare various super-resolution techniques such as GAN-based and Transformer-based models.

Conclusion

In this paper, we introduce Res-Bench, a new benchmark designed to evaluate the critical yet overlooked dimension of MLLM resolution robustness. Constructed from 1,200 curated samples across 6 core capabilities and 12 resolution levels, Res-Bench is paired with novel metrics like ACE/RCE to quantify stability beyond simple accuracy. Our comprehensive evaluation reveals a pervasive lack of stability in state-of-the-art models. Our analysis identified key factors influencing this behavior, including a clear architectural trade-off between performance and robustness, and the impact of input token length. Crucially, we also demonstrated that robustness can be explicitly learned through mixed-resolution fine-tuning, offering a direct path for improvement. We believe Res-Bench provides a vital tool for the community, pushing the development of MLLMs towards the more reliable and adaptive intelligence characteristic of human perception.

Acknowledgments

This research is supported by the National Science and Technology Major Project (2023ZD0121102), National Natural Science Foundation of China (No. 62572449 and No. 62472393).

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bai, T.; Liang, H.; Wan, B.; Xu, Y.; Li, X.; Li, S.; Yang, L.; Li, B.; Wang, Y.; Cui, B.; Huang, P.; Shan, J.; He, C.; Yuan, B.; and Zhang, W. 2024. A Survey of Multimodal Large Language Model from A Data-centric Perspective. arXiv:2405.16640.

Bigham, J. P.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R. C.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 333–342.

Cai, R.; Song, Z.; Guan, D.; Chen, Z.; Luo, X.; Yi, C.; and Kot, A. 2023. BenchLMM: Benchmarking Cross-style Visual Capability of Large Multimodal Models. arXiv:2312.02896.

Cao, T.; Trinh, M.-H.; Deng, A.; Nguyen, Q.-N.; Duong, K.; Cheung, N.-M.; and Hooi, B. 2024. Are Anomaly Scores Telling the Whole Story? A Benchmark for Multi-level Anomaly Detection. *arXiv preprint arXiv:2411.14515*.

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, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Croce, F.; Andriushchenko, M.; Sehwag, V.; Debenedetti, E.; Flammarion, N.; Chiang, M.; Mittal, P.; and Hein, M. 2021. RobustBench: a standardized adversarial robustness benchmark. arXiv:2010.09670.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N. A.; Ma, W.-C.; and Krishna, R. 2024. BLINK: Multimodal Large Language Models Can See but Not Perceive. arXiv:2404.12390.

Guo, Z.; Xu, R.; Yao, Y.; Cui, J.; Ni, Z.; Ge, C.; Chua, T.-S.; Liu, Z.; and Huang, G. 2024. LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images. In *ECCV*.

Hu, S.; Tu, Y.; Han, X.; Cui, G.; He, C.; Zhao, W.; Long, X.; Zheng, Z.; Fang, Y.; Huang, Y.; et al. 2024. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies. In *First Conference on Language Modeling*.

Kembhavi, A.; Salvato, M.; Kolve, E.; Seo, M.; Hajishirzi, H.; and Farhadi, A. 2016. A Diagram Is Worth A Dozen Images. arXiv:1603.07396.

Li, B.; Lin, Z.; Peng, W.; Nyandwi, J. d. D.; Jiang, D.; Ma, Z.; Khanuja, S.; Krishna, R.; Neubig, G.; and Ramanan, D. 2024a. NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples. 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, 17044–17068. Curran Associates, Inc.

Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; and Shan, Y. 2023. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. arXiv:2307.16125.

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*.

Li, C.; Wu, W.; Zhang, H.; Xia, Y.; Mao, S.; Dong, L.; Vulić, I.; and Wei, F. 2025. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*.

Li, C.; Zhang, J.; Zhang, Z.; Wu, H.; Tian, Y.; Sun, W.; Lu, G.; Liu, X.; Min, X.; Lin, W.; and Zhai, G. 2024c. R-Bench: Are your Large Multimodal Model Robust to Real-world Corruptions? arXiv:2410.05474.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, Y.; Li, Z.; Huang, M.; Yang, B.; Yu, W.; Li, C.; Yin, X.-C.; Liu, C.-L.; Jin, L.; and Bai, X. 2024. OCRBench: on the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12).

Liu, Y.; Liang, Z.; Wang, Y.; He, M.; Li, J.; and Zhao, B. 2025. Seeing clearly, answering incorrectly: A multimodal robustness benchmark for evaluating mllms on leading questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2024.

MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. arXiv:2310.02255.

Marr, D. 2010. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Masry, A.; Islam, M. S.; Ahmed, M.; Bajaj, A.; Kabir, F.; Kartha, A.; Laskar, M. T. R.; Rahman, M.; Rahman, S.; Shahmohammadi, M.; Thakkar, M.; Parvez, M. R.; Hoque, E.; and Joty, S. 2025. ChartQAPro: A More Diverse and Challenging Benchmark for Chart Question Answering. arXiv:2504.05506.

Masry, A.; Long, D. X.; Tan, J. Q.; Joty, S.; and Hoque, E. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 2263–2279. Dublin, Ireland: Association for Computational Linguistics.

OpenAI. 2024. Hello ChatGPT-4o. https://openai.com/index/hello-gpt-4o/.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Team, G. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530.

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*.

Vaishnav, M.; and Tammet, T. 2025. A Cognitive Paradigm Approach to Probe the Perception-Reasoning Interface in VLMs. arXiv:2501.13620.

Wang, J.; Ming, Y.; Shi, Z.; Vineet, V.; Wang, X.; Li, Y.; and Joshi, N. 2024a. Is A Picture Worth A Thousand Words? Delving Into Spatial Reasoning for Vision Language Models. arXiv:2406.14852.

Wang, W.; Ding, L.; Zeng, M.; Zhou, X.; Shen, L.; Luo, Y.; and Tao, D. 2024b. Divide, Conquer and Combine: A Training-Free Framework for High-Resolution Image Perception in Multimodal Large Language Models. arXiv:2408.15556.

Wang, Z.; Xia, M.; He, L.; Chen, H.; Liu, Y.; Zhu, R.; Liang, K.; Wu, X.; Liu, H.; Malladi, S.; et al. 2024c. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37: 113569–113697.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; and Van Gool, L. 2023. Diffir: Efficient diffusion model for image restoration. *ICCV*.

Xu, W.; Lyu, D.; Wang, W.; Feng, J.; Gao, C.; and Li, Y. 2025. Defining and Evaluating Visual Language Models' Basic Spatial Abilities: A Perspective from Psychometrics. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11571–11590. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Yang, Z.; Tang, J.; Li, Z.; Wang, P.; Wan, J.; Zhong, H.; Liu, X.; Yang, M.; Wang, P.; Bai, S.; Jin, L.; and Lin, J. 2024. CC-OCR: A Comprehensive and Challenging OCR Benchmark for Evaluating Large Multimodal Models in Literacy. arXiv:2412.02210.

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 preprint arXiv:2408.04840*.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; Wei, C.; Yu, B.; Yuan, R.; Sun, R.; Yin, M.; Zheng, B.; Yang, Z.; Liu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen, W. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. arXiv:2311.16502.

Yue, X.; Zheng, T.; Ni, Y.; Wang, Y.; Zhang, K.; Tong, S.; Sun, Y.; Yu, B.; Zhang, G.; Sun, H.; Su, Y.; Chen, W.; and Neubig, G. 2025. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. arXiv:2409.02813.

Zhang, R.; Jiang, D.; Zhang, Y.; Lin, H.; Guo, Z.; Qiu, P.; Zhou, A.; Lu, P.; Chang, K.-W.; Gao, P.; and Li, H. 2024. MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? arXiv:2403.14624.

Zhu, X.; Wang, S.; Lu, J.; Hao, Y.; Liu, H.; and He, X. 2024a. Boosting Few-Shot Learning via Attentive Feature Regularization. In *AAAI*, 7793–7801. AAAI Press.

Zhu, X.; Wang, S.; Zhu, B.; Li, M.; Li, Y.; Fang, J.; Wang, Z.; Wang, D.; and Zhang, H. 2025a. Dynamic Multimodal Prototype Learning in Vision-Language Models. *CoRR*, abs/2507.03657.

Zhu, X.; Zhu, B.; Tan, Y.; Wang, S.; Hao, Y.; and Zhang, H. 2024b. Enhancing Zero-Shot Vision Models by Label-Free Prompt Distribution Learning and Bias Correcting. In *NeurIPS*.

Zhu, X.; Zhu, B.; Tan, Y.; Wang, S.; Hao, Y.; and Zhang, H. 2024c. Selective Vision-Language Subspace Projection for Few-shot CLIP. In *ACM Multimedia*, 3848–3857. ACM.

Zhu, X.; Zhu, B.; Wang, S.; Zhao, K.; and Zhang, H. 2025b. Enhancing CLIP Robustness via Cross-Modality Alignment. *arXiv preprint arXiv:2510.24038*.