Benchmarking Multimodal Large Language Models Against Image Corruptions

Xinkuan Qiu[1,3,5]*, Meina Kan[2], Yongbin Zhou[1,4], Shiguang Shan[2]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, 100085, China

[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

[3]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, 100049, China

[4]School of Cyber Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China

[5]State Key Laboratory of Cyberspace Security Defense, Beijing, 100085, China

qiuxinkuan@iie.ac.cn, {kanmeina, sgshan}@ict.ac.cn, zhouyongbin@njust.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) have made significant strides in visual and language tasks. However, despite their impressive performance on standard datasets, these models encounter considerable robustness challenges when processing corrupted images, raising concerns about their reliability in safety-critical applications. To address this issue, we introduce the MLLM-IC benchmark, specifically designed to assess the performance of MLLMs under image corruption scenarios. MLLM-IC offers a more comprehensive evaluation of corruption robustness, enabling a multi-dimensional assessment of various MLLM capabilities across a broad range of corruption types. It includes 40 distinct corruption types and 34 low-level multimodal capabilities, each organized into a three-level hierarchical structure. Notably, it is the first corruption robustness benchmark designed to facilitate the evaluation of fine-grained MLLM capabilities. We further evaluate several prominent MLLMs and derive valuable insights into their characteristics. We believe the MLLM-IC benchmark will provide crucial insights into the robustness of MLLMs in handling corrupted images and contribute to the development of more resilient MLLMs. Dataset and evaluation code are available at https://github.com/EdyQiu/MLLM-IC/

1. Introduction

In recent years, Multimodal Large Language Models (MLLMs) have achieved significant progress in visual and language tasks, such as image captioning, visual question answering and cross-modal retrieval, demonstrating broad applicability across various domains [15, 27, 34, 35, 37]. However, despite their impressive performance on standard, uncorrupted data, these models exhibit notable limitations in robustness when confronted with corrupted images [8, 24, 33, 40]. This issue raises concerns regarding their reliability in safety-critical applications. For instance, monitoring systems could be compromised by adverse weather conditions such as rain or snow, while autonomous driving systems could struggle with challenges like motion blur. Therefore, it is essential to evaluate the performance and analyze the robustness of MLLMs under image corruption scenarios from both theoretical and practical perspectives.

In this paper, we introduce the MLLM-IC (Multimodal Large Language Model-Image Corruption) benchmark, specifically designed to address the challenges associated with image corruption. MLLM-IC features carefully structured three-level hierarchies for both MLLM capabilities and image corruption types, enabling multi-dimensional assessments of robustness and facilitating a more granular evaluation of model performance.

We then evaluate the corruption robustness of MLLMs using the MLLM-IC benchmark, leading to four key findings: (1) MLLMs are significantly affected by image corruption; (2) Under image corruption, MLLMs exhibit lower absolute performance in capabilities such as spatial perception and knowledge reasoning, whereas demonstrate relatively higher performance in tasks like instance recognition, global context sensing, and logical reasoning; (3) MLLMs perform excellently in handling geometric transformations, color perturbations, occlusion, and weather effects, but struggle with blur and compression corruptions; (4) MLLMs exhibit varying sensitivity to increased corruption severity across corruption dimensions, whereas this trend is not observed across capability dimensions.

To the best of our knowledge, MLLM-IC is the most comprehensive benchmark for evaluating image corruption robustness. We hope that this benchmark, along with our findings, will contribute to a deeper understanding of

*This work was carried out during Xinkuan Qiu's visiting period at Institute of Computing Technology, Chinese Academy of Sciences.

Table 1. Comparison between existing image corruption robustness benchmarks and MLLM-IC. MLLM-IC demonstrates its superiority in terms of the number of question-answer pairs, model capability dimensions, corruption type dimensions, and severity levels.

Benchmark	# QA pairs	# Capability dimensions	# Corruption dimensions	# Severity levels
CIFAR10-C (2019) [9]	750k	1 (Classification)	15	5
ImageNet-C (2019) [9]	3750k	1 (Classification)	15	5
ImageNet-C-bar (2021) [26]	500k	1 (Classification)	10	1
Pascal-C (2019) [25]	369k	1 (Detection)	15	5
Cityscapes-C (2019) [25]	37.5k	1 (Detection)	15	5
COCO-C (2024) [18]	400k	1 (Detection)	16	5
BDD100k-C (2024) [18]	800k	1 (Detection)	16	5
Kamann2020 (2020) [11]	400k	1 (Segmentation)	19	5
Delva2024 (2024) [4]	400k	1 (Segmentation)	16	1
MSRVTT-P (2022) [30]	2770k	1 (Text-to-video retrieval)	18	5
MMC-Bench (2024) [39]	29k	1 (Image captioning)	29	1
R-Bench(2024) [14]	48.1k	3 (Multiple choice questions, visual question answering and captioning)	33	3
MMRobustness (2024) [28]	2380k	5 (Visual entailment, image captioning, visual reasoning, image-text retrieval) and text-to-image generation)	17	5
MLLM-IC (ours)	**4270**k	**34 (19 perception and 15 reasoning tasks)**	**40**	**5**

MLLM behavior in corrupted image scenarios and inspire further research in this area.

2. Related Work

Benchmarks for Multimodal Large Language Models Early evaluations were performed on well-established multimodal datasets, such as COCO Caption [3] and Science QA [23]. However, these benchmarks focus predominantly on basic model capabilities, failing to capture the full range of progress in MLLM capabilities. To address these issues, LAMM [38] and LLaVA-Bench [17] incorporate a broader spectrum of evaluation dimensions, encompassing a diverse range of tasks that better reflect the multifaceted nature of model performance. Building on these efforts, MME [6] and MMBench [19] introduce hierarchical frameworks to assess model capabilities, arranging evaluations into three levels: two high-level tasks, multiple mid-level tasks, and around 20 fine-grained low-level tasks. SeedBench [13] broadens the evaluation scope to cover more complex tasks, including the processing of interleaved image-text content. Collectively, these efforts underscore the growing demand for more nuanced frameworks that can effectively capture MLLMs' advanced capabilities.

Benchmarks for Image Corruption Robustness Research on the robustness of visual perception models against image corruption originates from studying the response of traditional machine learning models to noisy data [21, 31]. This field gained substantial attention with the introduction of the ImageNet-Corruption (ImageNet-C) dataset [9], which applies 15 artificial corruptions to ImageNet images [5] and becomes a widely adopted image corruption benchmark. Building upon this, subsequent datasets have enabled the assessment of image corruption robustness within the domains of object detection [7] and segmentation [20] by applying similar corruptions to existing benchmarks, such as Pascal-C [25], Cityscapes-C [25], COCO-C [18], BDD100k-C [18], Kamann2020 [11], and Delva2024 [4]. In addition to these classic computer vision domains, MSRVTT-P [30] facilitates the evaluation of text-to-video retrieval tasks, while MMC-Bench [39] supports the assessment of image captioning tasks. R-Bench [14] incorporates three distinct task formats, while MMRobustness [28] evaluates five model capabilities.

Through a review of the literature on MLLM benchmarks and image corruption benchmarks, we observed that the former have progressively evaluated a broader range of model capabilities, but have largely overlooked the assessment of model robustness. Conversely, the latter have introduced a more diverse set of corruption types but remain limited in their evaluation of model capabilities. These limitations motivate the development of the MLLM-IC dataset. As shown in Table 1, our dataset surpasses existing image corruption benchmarks in terms of dataset size, number of model capability dimensions, number of corruption type di-

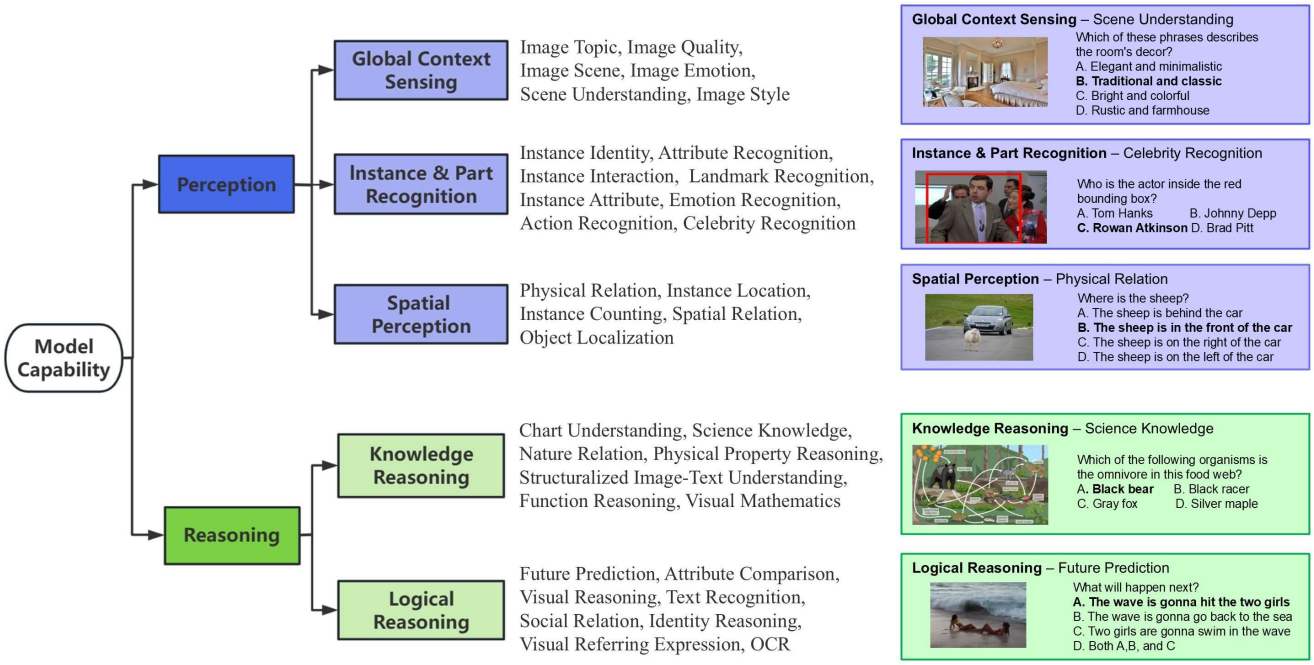

Figure 1. The three-level hierarchical structure of capability dimensions: 2 high-level tasks (Perception and Reasoning), 5 mid-level tasks (such as Global Context Sensing and Instance & Part Recognition), and 34 low-level tasks (such as Image Topic, Image Quality, and Image Scene). The images on the right illustrate one representative capability from each mid-level task, with correct answers highlighted in bold.

mensions, and number of corruption severity levels.

3. MLLM Image Corruption Dataset

To systematically and comprehensively evaluate the robustness of MLLMs in handling corrupted images, we propose the MLLM Image Corruption Benchmark (MLLM-IC). This benchmark employs carefully structured three-level hierarchies for both model capabilities and image corruption types, as shown in Figures 1 and 2. These design choices enable multi-dimensional robustness assessments and facilitate fine-grained performance evaluations. Additionally, to enable faster evaluation, we provide MLLM-IC-mini, a smaller evaluation subset, with details shown in Appendix A1. Dataset construction steps are described below.

3.1. Design of MLLM Capability Dimension

In line with the evolving trends of MLLM benchmarks, we have incorporated a diverse range of tasks to comprehensively capture the capabilities of MLLMs. To achieve this, we selected low-level tasks from two well-established MLLM benchmarks, namely MMBench and SeedBench. We then consolidated redundant tasks and reorganized the mid-level structure to better align with the capabilities required for MLLMs. The definitions of these mid-level tasks are listed in Appendix A2. Consequently, we propose a three-level hierarchy for capability dimensions, as shown in Figure 1. The high-level capability involves perception and reasoning; the mid-level capability includes global context sensing, instance & part recognition, spatial perception, knowledge reasoning, and logical reasoning; and the low-level involves 34 fine-grained capabilities.

3.2. Design of the Image Corruption Hierarchy

We begin by defining three operational levels of image corruption—global, regional, and pixel—based on transformation functions. Subsequently, nine general corruption types are assigned to these levels according to their shared formation mechanisms. Finally, we extend these general corruption types into specific variants using evidence from the imgaug library [1], resulting in a total of 40 specific corruption types. These 40 corruption types are further extended across five severity levels, yielding a total of 200 distinct corruption scenarios. Consequently, we propose a three-level hierarchy of corruption dimensions, as shown in Figure 2. The effectiveness of this hierarchy is further validated through feature visualization experiments. It is important to note that all corruptions in MLLM-IC are synthetic and we discuss the gap to real-world corruptions in Appendix A3.

To the best of our knowledge, MLLM-IC is the first image corruption benchmark to incorporate a meticulously structured hierarchy of corruption types, whereas previous studies typically enumerate these types without offering detailed explanations of their hierarchical relationships.

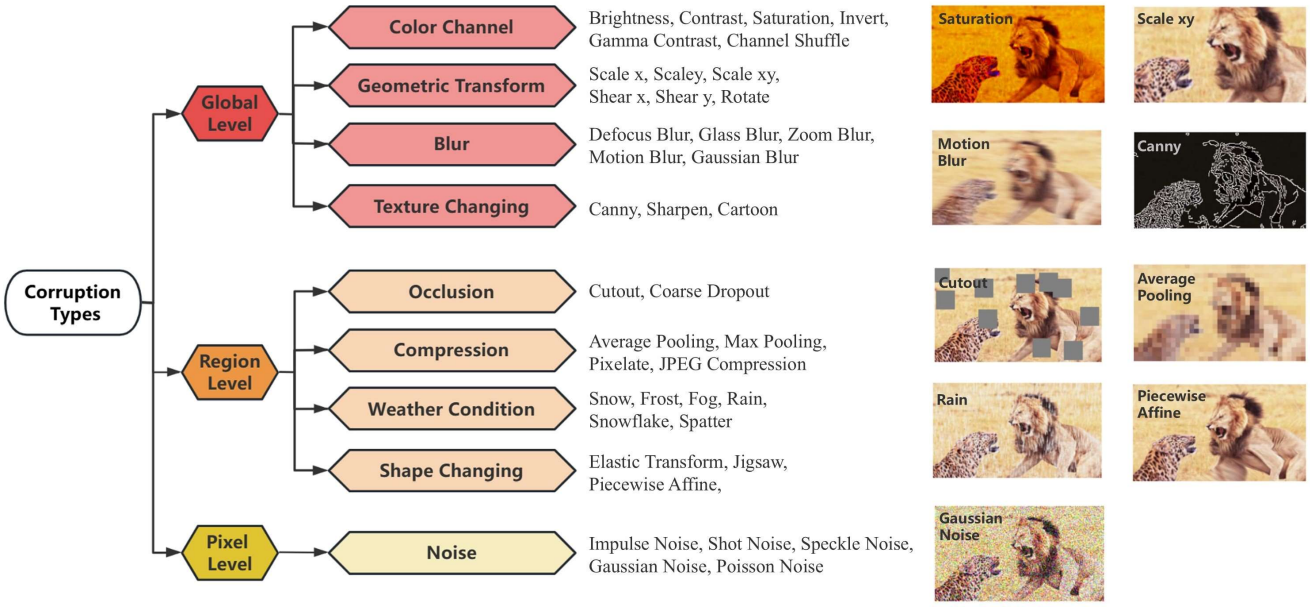

Figure 2. The three-level hierarchical structure of corruption dimensions: 3 operational levels (Global, Region, and Pixel levels), 9 general categories (such as Color Channel, Geometric Transformation and Blur) and 40 specific types (such as Brightness, Contrast and Saturation). The images on the right illustrate one representative corruption from each general category.

Operational Levels of Image Corruption Types Corruption can be conceptualized as a process in which specific transformation functions are applied to different regions of an image. Each region R_k consists of square blocks or irregular areas formed by adjacent pixels, and is associated with a transformation function F_k that is uniformly applied to all pixels within the region. Consequently, for a given input image I, the value of the corrupted image I' at a specific coordinate (x, y) is defined as:

$$I'(x, y) = F_k(I(x, y)), \quad \text{if } (x, y) \in R_k. \tag{1}$$

Based on this formula, the corruption process can be categorized into three levels, determined by the total number of regions K. For global-level corruption, the entire image is treated as a single region ($K = 1$), where all pixels are processed uniformly by function F_{Global}. E.g., scale-x is defined by $F_{\text{Global}}(I(x, y)) = I(sx, y)$, with s being a scaling. For regional-level corruption, the image is divided into $K > 1$ regions. Each region R_k is associated with a transformation function F_{Region_k}, which may be specific to individual regions or shared among multiple regions. E.g., average pooling is defined by $F_{\text{Region}_k} = \frac{1}{n^2} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} I(x+i, y+j)$, with n representing the size of region R_k. For pixel-level corruption, each pixel is treated as an independent region ($K =$ the number of pixels). Each pixel has its own transformation function $F_{\text{Pixel(x, y)}}$, enabling completely independent operations. E.g., Gaussian noise is defined by $F_{\text{Pixel(x, y)}} = I(x, y) + c(x, y)$, with c being a constant sampled from Gaussian distribution $\mathcal{N}(0, \sigma^2)$.

General and Specific Corruption Types Based on their formation mechanisms, the 40 specific corruption types, selected from the imgaug library, are grouped into nine general categories within the proposed three-level framework, as illustrated in Figure 2. For each of the 40 specific corruption types, we define five severity levels, resulting in a total of 200 distinct corruption scenarios. Detailed design for this part is shown in Appendix A4.

Validation of the Corruption Dimension Hierarchy To validate the hierarchy presented in Figure 2, we conduct a feature visualization experiment to examine the relationships among different corruption types in feature space.

First, we select a subset of 1,000 images from the CIFAR-10 dataset as the base dataset and apply all 200 corruption scenarios to generate 200 corrupted datasets. Next, we extract features from all images and compute the mean feature representation for each of the 200 corrupted datasets. Following the methodology of Mintun et al. [26], we use WideResNet as the feature extractor, which is trained on the respective datasets to ensure effective feature representation. Subsequently, we perform dimensionality reduction on these 200 mean features using t-SNE to explore the structural relationships among corruption types.

Figure 3 presents the t-SNE visualization results for the CIFAR-10 dataset. Each point represents one of the 200 corruption scenarios, with lines connecting the five severity levels within the same specific corruption type. Each color corresponds to one of nine general corruption categories defined in Figure 2. "Weather Condition" are excluded due to

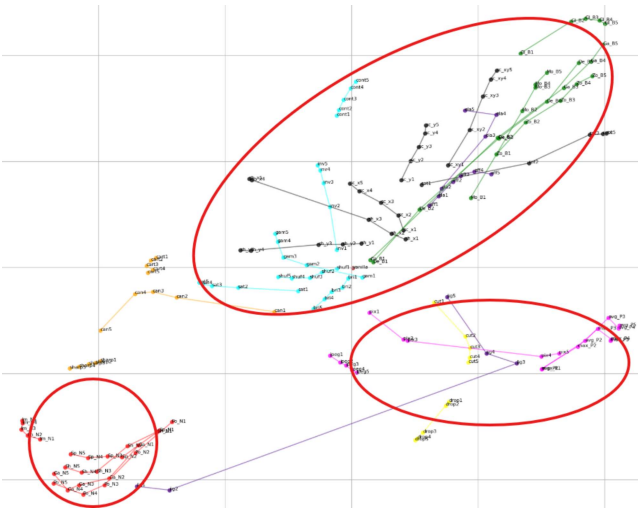

Figure 3. t-SNE visualization of the mean feature representations for 200 corruption types.

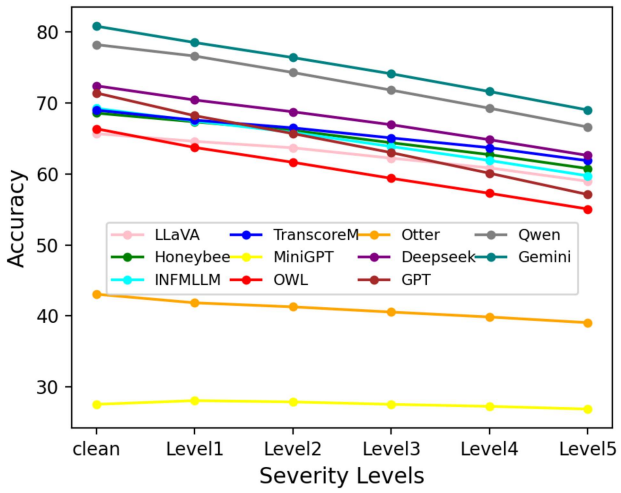

Figure 4. Performance of eleven MLLMs on clean and corrupted images in MLLM-IC benchmark.

their inherent complexity from non-synthetic nature.

Figure 3 illustrates our findings: (1) Corruptions within the same operational levels are generally clustered together, as highlighted by the red circles. (2) Corruptions within the same general category tend to group together (indicated by the points of the same color), reflecting their shared characteristics. (3) Corruptions with higher severity levels are typically positioned farther from the central point, which represents the uncorrupted case. This experiment is repeated using a subset from ImageNet, detailed in Appendix A5. These observations provide strong support for the rationale behind the hierarchical organization of corruption types in the proposed benchmark.

3.3. Corrupted Image Generation

Corresponding to the task design described in Section 3.1, a total of 4,329 question-answer pairs were selected from MMBench and 17,023 from SeedBench. All 40 corruption types with five severity levels introduced in Section 3.2 were applied to these data sources, forming the MLLM-IC benchmark. The 40 corruption types are implemented using the imgaug library, which provides a systematic way to simulate real-world distortions. The parameters for each severity level are determined based on range discretization for bounded values and human perceptual thresholding for unbounded ones. Visual illustrations are provided in Figure 5 and more details are provided in Appendix A6.

4. Evaluating and Analyzing the Robustness of MLLMs under Image Corruption

This section presents a systematic evaluation of contemporary MLLMs through our hierarchical benchmark. Based

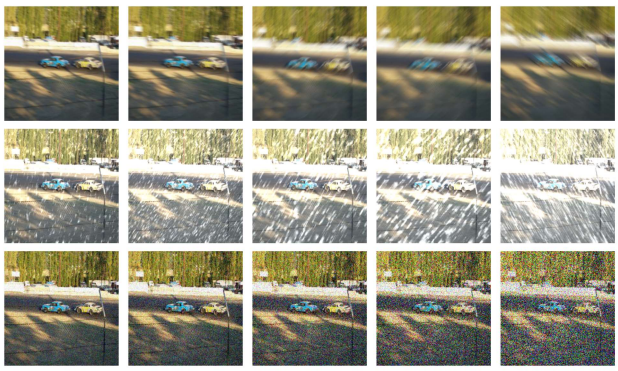

Figure 5. Visualization of images corrupted with five severity levels: motion blur (top), snow (mid), and Gaussian noise (bottom).

on the experimental results, we analyze the performance characteristics of these models and offer detailed insights into their strengths and limitations in different scenarios. Our investigation is structured along three analytical axes: severity, capability, and corruption dimension. These investigations highlight the advantages of our hierarchical evaluation design for comprehensive robustness assessment, serving as a practical guide for benchmark utilization.

Experimental Settings To evaluate the robustness of MLLMs against image corruption, we test eleven high-performance models from well-established MLLM benchmarks [6, 13, 17, 19, 38]: LLaVA-1.5 [16], HoneyBee [2], Inf-MLLM [41], Transcore-M [29], MiniGPT-4 [42], mPLUG-Owl2 [36], Otter [12], DeepSeek-VL [22], Qwen2.5-VL [32], GPT-4o [10] and Gemini 2.0 using the MLLM-IC benchmark. Details of model configurations are provided in Appendix B1. To support efficient evaluation, results on the MLLM-IC-mini are reported in Appendix B2.

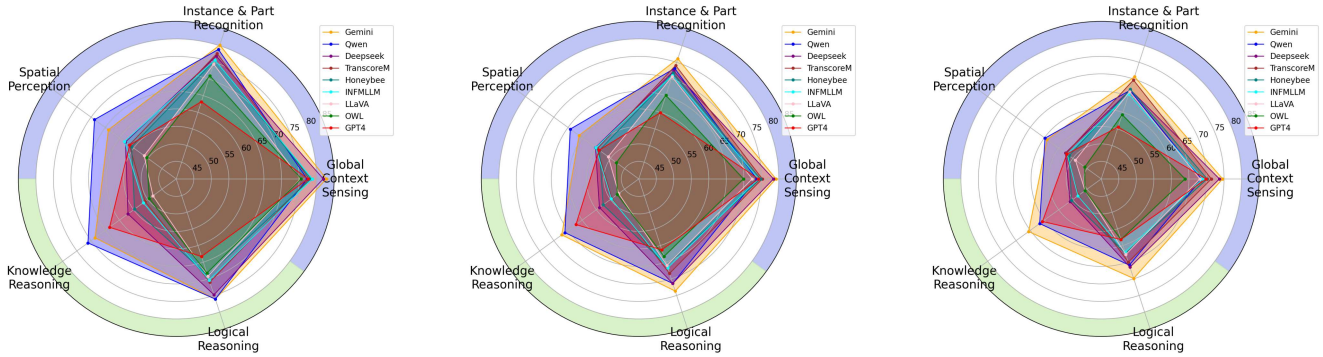

Figure 6. Radar charts depicting MLLM performance across five mid-level capability dimensions at severity levels 1, 3, and 5. The color of the outer ring indicates the high-level capabilities: purple for perception and green for reasoning.

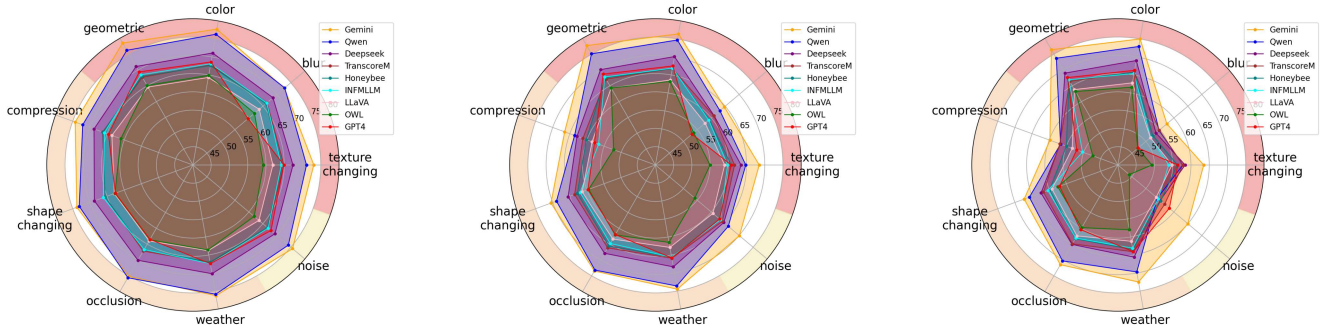

Figure 7. Radar charts depicting the performance of MLLMs across nine general corruption dimensions at severity levels 1, 3, and 5. The color of the outer ring indicates the operational level of corruptions: red for global, orange for regional, and yellow for pixel-level.

4.1. Performance in Severity Dimensions

Figure 4 presents the overall results for eleven MLLMs on the MLLM-IC benchmark across five severity levels. The results are averaged over 40 corruption dimensions and 34 capability dimensions. As shown, the accuracy of MLLMs experiences notable degradation relative to the uncorrupted baseline. For example, DeepSeek-VL exhibits an average accuracy drop of 5.8%, and becomes 9.8% under the most severe corruption; Gemini 2.0 outperforms all other models on average, achieving the highest accuracy of 73.9%, followed by Qwen2.5-VL with an accuracy of 71.6%. This superiority is consistent across all severity levels; GPT-4o shows the highest degradation, with accuracy declining by 14.3% in the most challenging conditions. Additionally, the gradual and linear decline in accuracy further supports the effectiveness of the severity level design in the proposed benchmark. MiniGPT-4 deviates from this trend by maintaining a consistently low performance (also reported by MMBench) regardless of corruption severity, since its near-random baseline limits the degradation under corruption.

Figure 6 illustrates the performance profiles across five mid-level capability dimensions under varying severity levels. MiniGPT-4 and Otter were excluded due to low accuracies to avoid scale compression and visual clutter. The three radar charts exhibit similar configurations with progressively decreasing scale magnitudes, indicating consistent relative superiority of MLLMs under corruption conditions. Figure 7 presents the performance profiles across nine general corruption types under varying severity levels. In contrast to the charts in capability dimension, these charts reveal significant structural divergence as severity increases. At level 1, the near-circular configurations suggest uniform performance across corruption types. However, higher severity levels expose model-specific sensitivity patterns, revealing that models have different sensitivities for various corruption dimensions: GPT-4o is clearly vulnerable to blur, while mPLUG-Owl shows pronounced degradation under noise; although Qwen2.5-VL initially performs comparably to Gemini 2.0, it becomes increasingly susceptible to multiple corruptions as severity increases.

4.2. Performance in Capability Dimensions

We provide a detailed analysis of the performance characteristics of MLLMs across various capability dimensions, averaged over 40 corruption types and five severity levels. As shown in Table 2, MLLMs generally exhibit robust performance in tasks related to global context sensing, instance

Table 2. Accuracies of eleven MLLMs in capability dimensions. The best result is highlighted and the second-best result is underlined.

Capability	LLaVA	Honey Bee	Inf mllm	Trans coreM	Mini GPT	Owl	Otter	Deep Seek	GPT	Qwen	Gem ini	Avg
Spatial Perception	50.4	53.6	55.1	54.5	23.6	48.0	31.7	55.3	54.7	**64.3**	62.8	50.4
Knowledge Reasoning	47.0	52.6	49.8	49.5	25.2	47.6	37.5	54.0	63.2	66.3	**68.1**	51.0
Logical Reasoning	66.1	65.8	66.4	68.1	28.8	63.3	44.7	70.9	62.3	71.1	**74.4**	62.0
Instance Recognition	70.5	71.5	70.8	73.8	26.4	65.1	41.7	72.0	59.3	72.9	**75.1**	63.6
Global Context Sensing	72.0	74.5	74.0	74.9	33.2	69.8	44.2	78.0	74.4	73.2	**79.5**	68.0

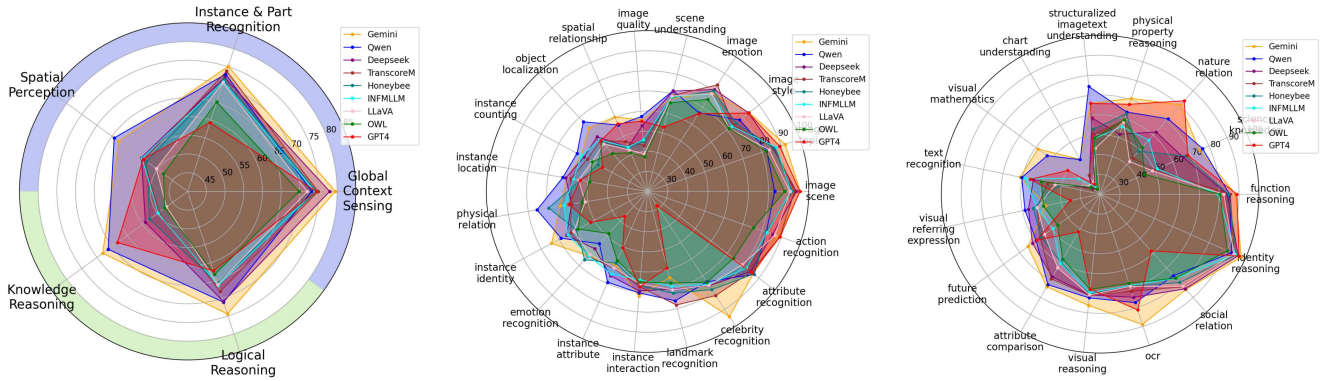

Figure 8. Radar charts showing the performance of nine MLLMs across hierarchical capability dimensions. From left to right: (a) performance across five mid-level capabilities, (b) performance across 19 low-level perceptual capabilities, and (c) performance across 15 low-level reasoning capabilities.

recognition, and logical reasoning under image corruption. In contrast, MLLM performances are much inferior in tasks requiring spatial perception and knowledge reasoning. Regarding the performance of individual models, none of the evaluated models demonstrate superior performance across all tasks. Gemini 2.0 achieves the best results in global context sensing, instance recognition, and two reasoning tasks, while Qwen2.5-VL exhibits the highest robustness in spatial perception. Notably, GPT-4o exhibits distinct robustness characteristics, particularly showing reduced accuracy in instance recognition tasks. This can be attributed to its conservative fallback behavior (e.g., defaulting to "don't know") in scenarios such as celebrity identification.

Figure 8 presents the radar charts corresponding to the data in Table 2. The leftmost chart directly reflects the data from Table 2, while the remaining two charts show the performance breakdown for low-level tasks in perception and reasoning dimensions respectively. These figures provide additional insights and further validate our conclusions. Furthermore, different models show strengths in distinct low-level capability dimensions, in contrast to the results in Table 2, where Gemini 2.0 consistently dominates across all mid-level capability dimensions. This observation underscores the importance of incorporating fine-grained capability dimensions into robustness benchmarks.

4.3. Performance Across Corruption Dimensions

We provide a comprehensive analysis of MLLMs' performance across various corruption dimensions, averaged over five severity levels. As shown in Table 3, MLLMs demonstrate strong performance when exposed to image corruptions such as geometric transformations, color perturbations, weather effects, and occlusions. However, performance deteriorates notably under corruptions involving blur and compression. When evaluating individual models, Gemini 2.0 consistently outperforms all other models across almost every corruption type.

Figure 9 presents the corresponding radar charts for the results in Table 3. The leftmost chart is directly derived from Table 3, while the remaining three charts break down performance for specific corruption types at the global, regional, and pixel levels. These figures provide further insights and reinforce the conclusions drawn from the data.

4.4. Multi-dimensional Performance Heatmap

The proposed benchmark facilitates a thorough evaluation of model robustness across various dimensions, including severity, capability, and corruption. For instance, the 3D heatmap presented in Figure 10 illustrates DeepSeek-VL's performance across these three dimensions, serving as a diagnostic assessment of its robustness to image corruption. Further information is provided in Appendix B3.

Table 3. Accuracies of eleven MLLMs in corruption dimensions. The best result is highlighted and the second-best result is underlined.

Corruption	LLaVA	HoneyBee	Inf mllm	TranscoreM	MiniGPT	Owl	Otter	DeepSeek	GPT	Qwen	Gemini	Avg
Blur	57.7	59.3	59.1	60.7	26.4	54.3	39.1	61.0	53.0	<u>63.6</u>	**64.3**	54.6
Compression	58.5	60.5	57.6	61.5	26.3	53.5	40.1	62.9	57.8	<u>64.1</u>	**66.3**	55.5
Texture Changing	58.5	60.5	59.3	61.4	26.9	54.6	38.8	63.1	60.5	<u>64.6</u>	**67.8**	56.1
Noise	59.3	61.6	61.5	62.5	28.2	53.3	40.2	62.9	62.8	<u>64.8</u>	**69.6**	57.2
Shape Changing	61.2	62.8	62.8	63.6	27.1	59.9	38.6	65.6	59.7	<u>69.2</u>	**70.2**	58.4
Occlusion	63.3	65.2	64.9	66.1	27.0	61.9	38.8	67.7	61.8	**73.0**	<u>72.9</u>	60.5
Weather	62.7	65.4	65.4	65.7	28.0	61.0	39.7	68.0	65.8	<u>73.1</u>	**74.0**	60.9
Color	63.6	66.7	66.8	66.8	27.8	63.3	41.3	70.1	67.3	<u>74.6</u>	**76.2**	62.4
Geometric	64.2	67.4	67.7	67.0	28.1	64.2	40.8	70.1	68.6	<u>75.0</u>	**77.2**	62.9

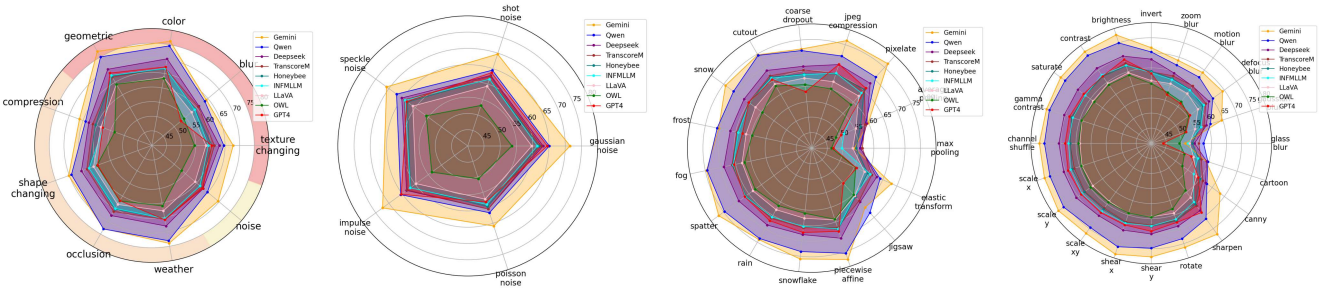

Figure 9. Radar charts depicting the performance of nine MLLMs across hierarchical corruption dimensions. From left to right: (a) performance in nine general corruption types, (b) performance in 20 specific global-level corruptions, (c) performance in 15 specific regional-level corruptions, and (d) performance in five specific pixel-level corruptions.

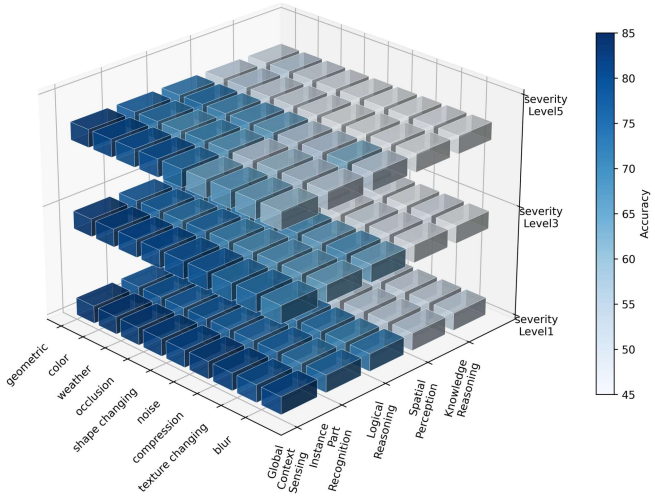

Figure 10. Multi-dimensional performance heatmap for DeepSeek.

4.5. Discussion

We evaluate the robustness of eleven MLLMs under image corruption and identify three patterns.

Performance in Capability Dimensions Under image corruption, MLLMs exhibit lower absolute robustness in capabilities such as spatial perception and knowledge reasoning, whereas demonstrate relatively higher robustness in tasks like instance recognition, global context sensing, and logical reasoning;

Performance in Corruption Dimensions MLLMs show excellent performance in handling geometric transformations, color perturbations, occlusions, and weather effects. In contrast, they exhibit more substantial performance degradation under blur and compression corruptions.

Sensitivity to Corruption Severity Different MLLMs display varying levels of sensitivity to increasing corruption severity within corruption dimensions. However, this trend is not observed within capability dimensions.

5. Conclusions

In this study, we introduce a comprehensive benchmark to assess the performance of MLLMs and analyze their behavior under various image corruption scenarios. Using this benchmark, we evaluate eleven SOTA MLLMs, providing a multi-dimensional analysis of their strengths and limitations. MLLM-IC is presented as a valuable tool for robustness evaluation and hope our findings provide meaningful insights into how MLLMs respond to image corruption.

Acknowledgements

This work is partially supported by National Key R&D Program of China (No.2022YFB3103800), National Natural Science Foundation of China (No.62495082 and No.U2336205), and Strategic Priority Research Program of the Chinese Academy of Sciences (No.XDB0680202).

References

[1] Alexander B.Jung, Wada Kentaro, and Crall Jon et al. imgaug. *https: //github.com/aleju/imgaug*, 2020. 3

[2] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827, 2024. 5

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[4] Yusuf Dalva, Hamza Pehlivan, Said Fahri Altındiş, and Aysegul Dundar. Benchmarking the robustness of instance segmentation models. *IEEE Transactions on Neural Networks and Learning Systems*, pages 17021–17035, 2024. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[6] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Jinrui Yang Xu Lin, Xiawu Zheng, Ke Li, and Xing Sun et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2, 5

[7] Ross Girshick, Jeff Donahue, Trevor Darrell, , and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 142–158, 2015. 2

[8] Zhongyi Han, Guanglin Zhou, Rundong He, Jindong Wang, Tailin Wu, Yilong Yin, Salman Khan, Lina Yao, Tongliang Liu, and Kun Zhang. How well does gpt-4v (ision) adapt to distribution shifts? a preliminary investigation. *ICLR Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024. 1

[9] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations*, 2019. 2

[10] Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, and A. J. Ostrow et al. Gpt-4o system card., 2024. 5

[11] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models with respect to common corruptions. *International Journal of Computer Vision*, pages 462–483, 2021. 2

[12] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 5

[13] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 2, 5

[14] Chunyi Li, Jianbo Zhang, Zicheng Zhang, Haoning Wu, Yuan Tian, Wei Sun, and Guo Lu et al. R-bench: Are your large multimodal model robust to real-world corruptions? *arXiv preprint arXiv:2410.05474*, 2024. 2

[15] Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, Wei An-Hou, Li Ming, Tianyang Wang, Ziqian Bi, and Ming Liu. A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284*, 2024. 1

[16] Haotian Liu, Chunyuan Li, Yuheng Li, , and Yong Jae Lee. Improved baselines with visual instruction tuning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 5

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 2024. 2, 5

[18] Jiawei Liu, Zhijie Wang, Lei Ma, Chunrong Fang, Tongtong Bai, Xufan Zhang, Jia Liu, and Zhenyu Chen. Benchmarking object detection robustness against real-world corruptions. *International Journal of Computer Vision*, pages 1–19, 2024. 2

[19] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, and Yike Yuan et al. Mmbench: Is your multi-modal model an all-around player? *European Conference on Computer Vision*, pages 216–233, 2025. 2, 5

[20] Jonathan Long, Evan Shelhamer, , and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 2

[21] Dengsheng Lu and Qihao Weng. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, pages 823–870, 2007. 2

[22] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, and Jingxiang Sun et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 5

[23] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, pages 2507–2521, 2022. 2

[24] Hashmat Shadab Malik, Fahad Shamshad, Muzammal Naseer, Karthik Nandakumar, Fahad Khan, and Salman Khan. Robust-llava: On the effectiveness of large-scale robust image encoders for multi-modal large language models. *arXiv preprint arXiv:2502.01576*, 2025. 1

[25] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker,

Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 2

[26] Eric Mintun, Alexander Kirillov, , and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. *Advances in Neural Information Processing Systems.*, pages 3571–3583, 2021. 2, 4

[27] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, and Lama Ahmad et al. Gpt-4 technical report. *https://arxiv.org/abs/2303.08774*, 2024. 1

[28] Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. Benchmarking robustness of multimodal image-text models under distribution shift. *Journal of Data-centric Machine Learning Research*, 2023. 2

[29] PCI research and davidlight2018. Transcore-m. *https://github.com/PCIResearch/TransCore-M*, 2023. 5

[30] Madeline Schiappa, Shruti Vyas, Hamid Palangi, Yogesh Rawat, and Vibhav Vineet. Robustness analysis of video-language models against visual and language perturbations. *Advances in Neural Information Processing Systems*, pages 34405–34420, 2022. 2

[31] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, pages 222–245, 2013. 2

[32] Qwen Team. Qwen2.5-vl, 2025. 5

[33] Aayush Atul Verma, Amir Saeidi, Shamanthak Hegde, Ajay Therala, Fenil Denish Bardoliya, Nagaraju Machavarapu, Shri Ajay Kumar Ravindhiran, Srija Malyala, Agneet Chatterjee, Yezhou Yang, and Chitta Baral. Evaluating multimodal large language models across distribution shifts and augmentations. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 5314–5324, 2024. 1

[34] Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, and Mengyuan Liu et al. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv preprint arXiv:2408.01319*, 2024. 1

[35] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *https://arxiv.org/abs/2309.17421*, 2024. 1

[36] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. 5

[37] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), 2024. 1

[38] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, and Xiaoshui Huang et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 2024. 2, 5

[39] Jiawei Zhang, Tianyu Pang, Chao Du, Yi Ren, Bo Li, and Min Lin. Benchmarking large multimodal models against common corruptions. *arXiv preprint arXiv:2401.11943.*, 2024. 2

[40] Xingxuan Zhang, Jiansheng Li, Wenjing Chu, Junjia Hai, Renzhe Xu, Yuqing Yang, Shikai Guan, Jiazheng Xu, and Peng Cui. On the out-of-distribution generalization of multimodal large language models. *arXiv preprint arXiv:2402.06599*, 2024. 1

[41] Qiang Zhou, Zhibin Wang, Wei Chu, Yinghui Xu, Hao Li, and Yuan Qi. Infmllm: A unified framework for visual-language tasks. *arXiv preprint arXiv:2311.06791*, 2023. 5

[42] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 5