Appendix: Benchmarking Multimodal Large Language Models Against Image Corruptions

A. Details on the MLLM-IC Benchmark

A.1. MLLM-IC-mini

We introduce a streamlined evaluation subset by sampling 50 QA pairs per capability dimension, resulting in only 8.4% of the full set while preserving the hierarchical structure. Using 8×RTX 3080 GPUs, the evaluation time is reduced from 75 to 6.3 hrs for DeepSeek and from 60.7 to 5.1 hrs for LLaVA.

A.2. Definition of Mid-Level Capabilities

In constructing MLLM-IC, we streamlined redundant tasks and reorganized the mid-level structure to better align with the capabilities required by MLLMs. The mid-level capabilities are defined as follows:

Global Context Sensing This task evaluates the model's ability to interpret global information in an image, with a focus on the overall atmosphere or general layout of the scene.

Instance & Part Recognition This task evaluates the model's proficiency in identifying fine-grained details within an image, particularly in recognizing individual objects and their constituent parts.

Spatial Perception This task examines the model's ability to understand spatial relationships, including the physical positioning, orientation, and relative arrangement of objects within a given scene.

Knowledge Reasoning This task measures the model's capability to draw conclusions or make inferences by leveraging external background knowledge or domain-specific understanding.

Logical Reasoning This task evaluates the model's ability to perform logical deductions, predictions, and comparisons. Unlike knowledge reasoning, which relies on external information beyond the provided input, logical reasoning is grounded in the contextual data.

A.3. Synthetic vs real-world corruptions

Synthetic corruptions are easy to generate and allow for controlled, comparable evaluations, as all corrupted images are derived from a base dataset to the best of our knowledge. This design choice is consistent with all benchmarks listed in Table 1 of the main paper.

However, the gap between synthetic and real-world corruptions remains uncertain, largely due to the absence of a suitable real-world corruption dataset. As a preliminary study, we examine the extent to which synthetic corruptions resemble real-world cases. We applied severe-2 synthetic motion blur (SB) to 4,913 reference images (R) from RealBlur dataset and compared to real blurred (RB) reference images. We observe FID(RB, SB) = 15.9 is lower than FID(R, RB) = 22.9 which is close to FID(R, SB) = 26.8, indicating that synthetic corruptions provide a faithful approximation of real-world effects.

To improve the generalizability of MLLM-IC, we plan to incorporate a real-world validation set in future versions of the benchmark. Specifically, we collect images from denoising datasets (e.g., SIDD), deblurring datasets (e.g., RealBlur and RWBI), and weather-related datasets (e.g., RealRain, RSOD, and REVIDE). We then use GPT-4o to generate question-answer pairs following the MMBench/SEEDBench format, and subsequently refine incorrect responses with the assistance of human experts.

A.4. General and Specific Corruption Types

The mapping from general corruption types to specific corruption types is outlined below, followed by the design principles for severity levels.

Global-Level Corruptions These corruptions involve applying a uniform transformation to all pixels in the image. For instance, color channel corruptions are introduced by applying a consistent function specific to color channels across all pixels. Geometric transformations are performed by applying an identical transformation matrix to every pixel. Blurring is achieved through a uniform smoothing operation, such as convolutional kernels, which process pixel neighborhoods consistently throughout the image. Texture-changing corruptions typically involve gradient-based operations, such as edge detection or sharpening, which enhance contrast and emphasize structural features while potentially degrading fine texture details.

Regional-Level Corruptions These corruptions divide the image into multiple spatial regions, each undergoing a uniform transformation that may differ across regions. For example, occlusions are introduced by setting pixel values within specific regions to zero. Weather-based corruptions serve as a specialized form of occlusion, where environmental effects act as occlusion layers, often accompanied by blurring to simulate real-world conditions. Image compression reduces storage requirements by exploiting the high correlation between adjacent pixels in certain regions, thereby minimizing the number of bits needed for representation. Shape-altering transformations involve segmenting the image into regions and applying affine transformations selectively, resulting in localized distortions.

Pixel-Level Corruptions These corruptions are applied independently to each pixel. For example, noise corruptions introduce random, independent variations to each pixel's intensity, altering pixel values in a stochastic manner.

Five Severity Levels The severity levels of each corruption type are determined based on the following principles: (1) For parameters with a fixed range, the range is divided into five discrete levels, with corresponding values assigned to each level. (2) For parameters with an open range, the highest severity level is set based on the maximum value at which the image remains recognizable to humans. The remaining levels are then defined using the first criterion.

A.5. Validation Experiment on ImageNet

To further validate our findings, we replicate the experiment in Section 3.2 of the main paper on a 500-image subset of the ImageNet dataset. As illustrated in Figure 1, the results align with those from the CIFAR dataset, reinforcing the consistency of our observations.

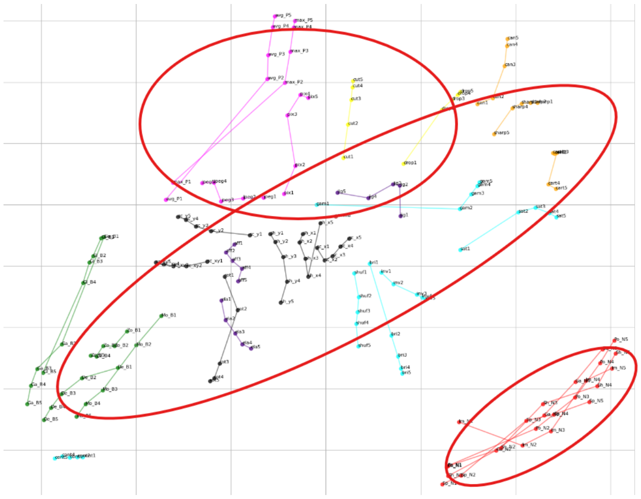

Figure 1. t-SNE visualization of the mean feature representations for 200 corruption types on ImageNet subset.

A.6. Visualization of Corruption Types

Figure 3 illustrates the 40 specific corruption types designed in the MLLM-IC benchmark.

B. Details on the MLLM Evaluation

B.1. Experimental Settings

In Section 4 of the main paper, we evaluate the performance of several MLLMs using the proposed benchmark. To ensure fairness and consistency, all models are evaluated using their officially recommended parameter configurations. Table 1 provides detailed information about the models employed in our experiments, including their versions and checkpoints. These MLLMs are primarily sourced from GitHub repositories and downloaded from HuggingFace.

Table 1. Summary of models used in the evaluation

Models	Versions and Checkpoints
LLaVA-1.5	liuhaotian/llava-v1.5-7b
HoneyBee	Honeybee-C-7B-M256
Inf-MLLM	mightyzau/InfMLLM_7B_Chat
Transcore-M	PCIResearch/TransCore-M
MiniGPT-4	Vicuna-V0-7B
mPLUG-Owl2	MAGAer13/mplug-owl2-llama2-7b
Otter	luodian/OTTER-Image-MPT7B
DeepSeek-VL	deepseek-ai/deepseek-vl-7b-chat
Qwen2.5-VL	Qwen/Qwen2.5-VL-7B-Instruct
GPT-4o	gpt-4o
Gemini 2.0	gemini-2.0-flash

B.2. Evaluation on MLLM-IC-mini

Evaluation on MLLM-IC-mini is presented in Table 2, providing a compact yet informative summary of model performance. The results include overall accuracy as well as fine-grained evaluations along both the capability and corruption dimensions. The overall robustness trend remains consistent as the full-scale dataset. This subset enables efficient comparison across models while preserving the hierarchical structure of the full benchmark, thus supporting rapid diagnosis of robustness characteristics.

B.3. Multi-dimensional Performance Heatmap

Figure 9 in the main paper illustrates DeepSeek-VL's performance across three dimensions, providing a diagnostic assessment of its robustness to image corruption. This heatmap serves as a predictive tool for estimating model performance in specific application scenarios. For example, in the context of photographing sports events and capturing individual players, the model's capability to recognize instances under blurred conditions is particularly highlighted.

Furthermore, sensitivity heatmaps allow for analyzing a model's responsiveness to specific tasks under various corruption types. Figure 2 presents the sensitivity heatmap for DeepSeek-VL, computed as the percentage performance drop from severity level 1 to severity level 5. From this figure, we observe that blur significantly affects accuracy in in-

Table 2. Main tables reproduced on the streamlined subset. The overall robustness trend remains consistent, with red for weakness and green for strength.

Overall	LLa	Honey	Inf	Trans	Mini	Owl	Otter	Deep	**GPT**	**Qwen**	**Gemini**	Avg
	VA	Bee	mllm	coreM	GPT			Seek				
Clean	66.6	70.1	67.7	70.2	27.7	65.9	44.9	73.1	70.5	78.7	79.7	65.0
Corrupt	62.7	65.3	63.2	66.3	28.3	58.9	41.6	67.2	62.2	71.9	73.2	60.1
Capability												
Spatial-P	55.9	58.4	55.3	56.9	25.2	50.5	32.5	59.1	57.2	67.1	64.8	53.0
Knowlg-R	48.7	55.5	50.6	51.4	27.7	48.9	42.5	57.4	65.9	69.6	71.2	53.6
Logical-R	66.1	65.2	64.9	69.5	30.1	61.0	45.6	69.1	61.3	71.5	73.7	61.6
Instance-R	67.6	69.3	68.5	74.0	25.5	62.4	38.0	69.4	57.8	73.8	73.4	61.8
Global-S	72.5	76.8	74.8	76.3	32.5	70.3	47.8	79.7	75.7	76.4	80.8	69.4
Corruption												
Blur	58.1	60.4	58.4	62.0	27.2	54.0	40.4	61.7	53.6	65.3	65.3	55.1
Compress	59.2	61.4	57.3	63.2	26.5	53.3	41.8	63.7	58.6	65.7	67.4	56.2
Texture-C	59.0	61.5	58.8	62.6	26.2	54.5	40.2	63.6	61.3	66.8	69.0	56.7
Noise	60.2	63.0	61.2	64.2	29.5	53.5	41.6	63.4	63.9	66.9	71.1	58.0
Shape-C	62.2	63.9	63.0	65.3	27.3	59.7	40.4	66.5	60.7	71.9	71.6	59.3
Occlusion	64.6	67.4	64.7	67.3	28.6	61.6	40.4	69.0	63.0	75.5	74.6	61.5
Weather	63.9	67.0	65.2	67.2	28.9	60.9	40.5	69.3	66.8	75.2	75.6	61.9
Color	64.9	68.4	66.1	67.9	28.7	63.1	42.3	71.0	68.4	77.1	77.4	63.2
Geometric	65.4	69.2	67.2	68.5	29.3	64.1	42.3	71.2	69.6	77.2	78.7	63.9

stance and part recognition tasks, leading to a performance drop of 0.25, while its impact on tasks requiring spatial perception is comparatively lower, with a drop of 0.17. Furthermore, geometric transformations have minimal influence on tasks involving global context sensing. While these findings provide valuable insights, the underlying mechanisms remain unexplored and warrant further investigation.

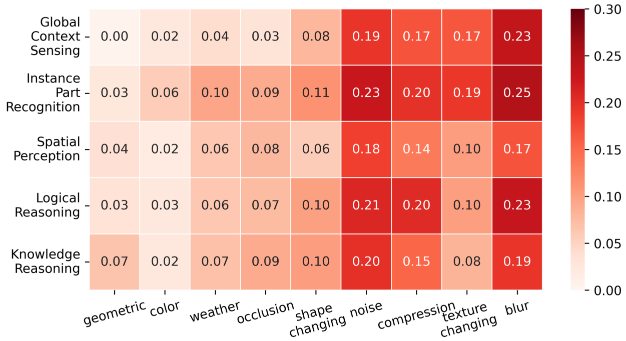

Figure 2. Sensitivity heatmap of DeepSeek-VL

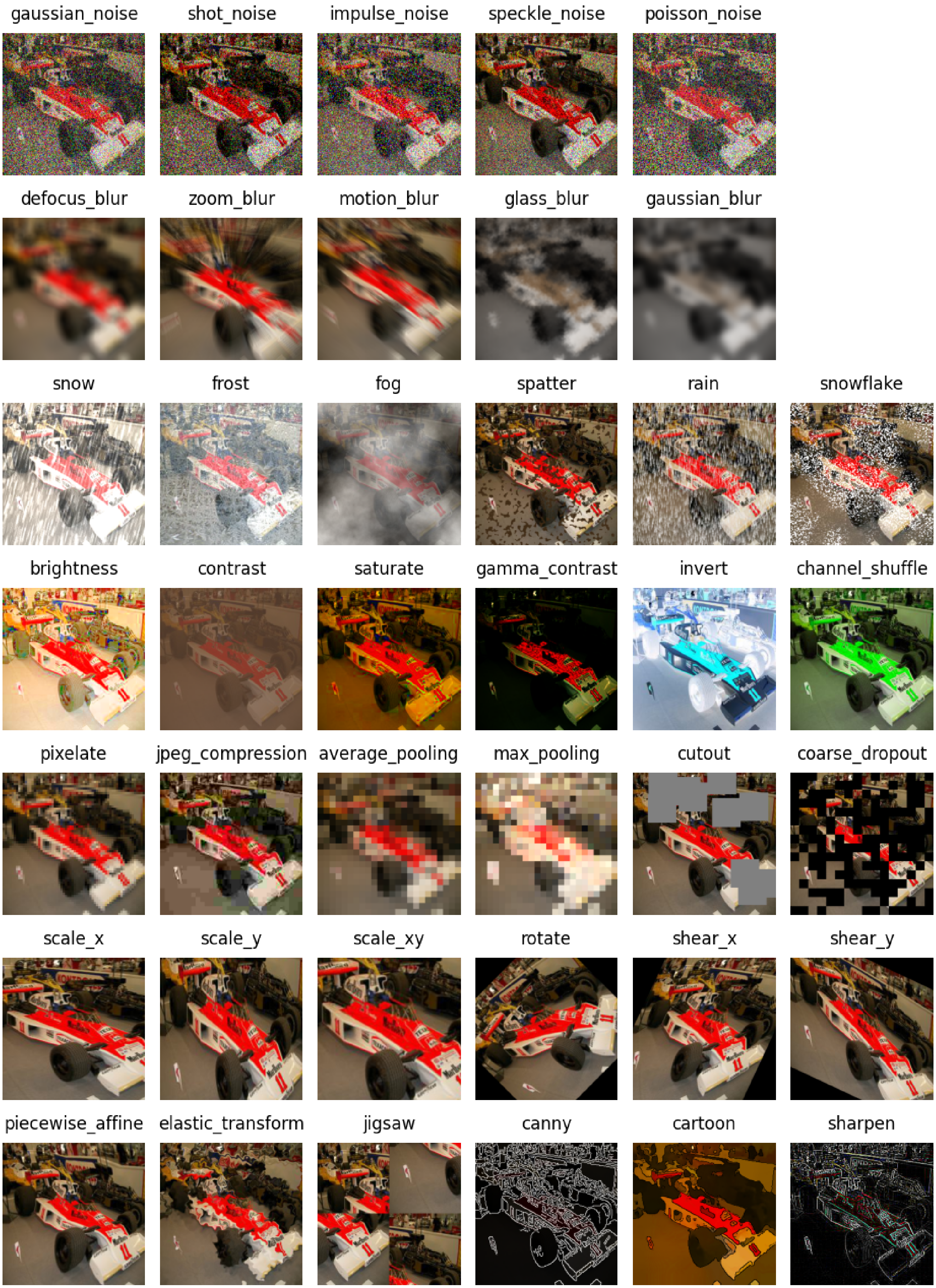

Figure 3. Visualization of the 40 specific corruption dimensions in the MLLM-IC benchmark.