SPACEDG: BENCHMARKING SPATIAL INTELLIGENCE UNDER VISUAL DEGRADATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have made rapid progress in spatial intelligence, yet existing spatial reasoning benchmarks largely assume pristine visual inputs and overlook the degradations that commonly occur in real-world deployment, such as motion blur, low light, adverse weather, lens distortion, and compression artifacts. This raises a fundamental question: how robust is the spatial intelligence of current MLLMs when visual observations are imperfect? To answer this question, we introduce **SpaceDG**, the first large-scale dataset for degradation-aware spatial understanding. It is constructed with a physically grounded degradation synthesis engine that embeds degradation formation process into 3D Gaussian Splatting (3DGS) rendering, enabling realistic simulation of nine degradation types. The resulting dataset contains approximately 1M QA pairs with over 160K images. We further introduce **SpaceDG-Bench**, a human-verified benchmark with 1,102 unique questions spanning 11 reasoning categories and 9 visual degradation types, yielding 10K VQA instances. Evaluating 25 open- and closed-source MLLMs reveals that visual degradations consistently and substantially impair spatial reasoning, exposing a critical robustness gap. Finally, we show that finetuning on SpaceDG markedly improves degradation robustness and can even surpass human performance under degraded conditions without any performance drop on clean images, highlighting the promise of degradation-aware training for robust spatial intelligence.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have achieved remarkable success in spatial intelligence, bridging the crucial gap between 2D visual recognition and 3D physical reasoning Liu et al. (2023); Wu et al. (2025); Black et al. (2026). As a fundamental capability of visual cognition Yang et al. (2024); Luo et al. (2025); Li et al. (2025b), spatial intelligence poses immense challenges to a model's ability to perceive, parse, and reason within the complex real world. To evaluate and advance this, researchers have proposed a myriad of benchmarks Yang et al. (2025c); Zhang et al. (2025); Yang et al. (2025b); Jia et al. (2026); Wang et al. (2026); Li et al. (2025a); Yang et al. (2026), upon which current state-of-the-art models Chen et al. (2024); Wu et al. (2025); Yang et al. (2025a;b); Cai et al. (2025a) demonstrate impressive spatial awareness, positioning them as the foundational brains for embodied agents and autonomous systems.

Existing spatial benchmarks predominantly evaluate MLLMs under a "perfect observation" assumption, using clean, high-resolution, and well-illuminated images. Yet in real-world embodied and autonomous systems, visual observations are produced by imperfect sensing pipelines, where degradations naturally arise during acquisition, transmission, and deployment. These degradations are not merely artificial corruptions, but common conditions faced by agents operating in physical and resource-constrained environments. They have been extensively studied in low-level vision and computational photography, spanning motion blur Su et al. (2017); Nah et al. (2017); Zhong et al. (2020; 2023b;a), low-resolution imaging Dong et al. (2015); Ledig et al. (2017); Wang et al. (2018); Lu et al. (2022), geometric distortion Liu et al. (2020); Zhong et al. (2021); Cao et al. (2022), low-light Chen et al. (2018); Niu et al. (2023a;b), and adverse weather He et al. (2010); Fu et al. (2017). Under such conditions, the robustness of spatial intelligence becomes a critical requirement, since spatial reasoning often depends on fine-grained geometric evidence, including object boundaries, relative positions, and multi-view consistency. Despite this rich literature on degradation and recent advances

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

in benchmarking general VLM robustness Tang et al. (2026); Saxena et al. (2026), how current MLLMs perform spatial reasoning under imperfect observations remains an open question.

To systematically answer this question, a suitable benchmark must satisfy three requirements: it should introduce realistic visual degradations, preserve the underlying 3D spatial structure, and support diverse spatial reasoning tasks with reliable ground truth. To fill this gap, we introduce SpaceDG and SpaceDG-Bench, the first large-scale VQA dataset and benchmark dedicated to degradation-aware spatial understanding, and conduct a comprehensive evaluation of current MLLMs under imperfect visual observations.

Specifically, we develop an automatic degradation data engine. First, we reconstruct multi-view images into geometrically accurate 3D Gaussian Splatting (3DGS) Kerbl et al. (2023) representations and pair them with auto-annotated spatial QAGao et al. (2026) . Second, on top of the reconstructed 3DGS, we design a physically realistic degradation synthesis pipeline that simulates nine representative degradations across four categories: (1) **optical and dynamic degradations**, including defocus, distortion and motion blur; (2) **meteorological degradations**, including haze and water droplets; (3) **photometric degradations**, including low light and over-exposure; and (4) **digital degradations**, including JPEG compression and low resolution. Each degradation is generated from underlying physical formation process, as shown on the left in Figure **??**.

Leveraging this engine, we construct SpaceDG, a large-scale dataset derived from nearly 1,000 scenes in ScanNet++ Yeshwanth et al. (2023). SpaceDG comprises approximately 1M QA pairs over more than 160K images and covers a diverse range of visual degradations. To establish a rigorous evaluation protocol, we further introduce SpaceDG-Bench, a manually curated and verified benchmark comprising 1K high-quality QA pairs. For comprehensive assessment, we systematically design 11 distinct question categories, encompassing camera-centric, object-centric and object-camera relation questions with single-view or multi-view images.

We conduct a comprehensive evaluation of 25 models and identify four key findings. **First**, visual degradations consistently impair spatial reasoning across all evaluated MLLMs, highlighting the need for degradation-aware spatial evaluation. **Second**, humans also suffer clear performance drops under degraded conditions. This suggests that the design of MLLMs should not simply imitate human perception, but should learn degradation-aware spatial knowledge to better handle diverse real-world visual inputs. **Third**, degradation-based SFT yields substantial improvements on both clean and degraded inputs, indicating that exposure to physically grounded degradations can enhance robust spatial understanding. **Finally**, we observe that visual degradations affect fine-grained object-level perception, such as object counting, more strongly than certain geometric reasoning tasks, such as camera-centric translation, revealing that detailed visual grounding is particularly sensitive to degraded visual evidence.

2 Related works

Spatial intelligence of MLLMs Recent advances in spatial MLLMs have expanded their capabilities from basic visual understanding Qwen Team (2026); Wang et al. (2025a); Team et al. (2025); Xiaomi (2025) to fine-grained spatial reasoning Yang et al. (2025b); Cai et al. (2025a); Wu et al. (2025); Yang et al. (2025a); Cheng et al. (2024); Daxberger et al. (2025) with large-scale spatial datasets. For example, Cambrian-S Yang et al. (2025b), VST Yang et al. (2025a), and SenseNova-SI Cai et al. (2025a) adopt VSI-590K, 4.1M samples, and SenseNova-SI-8M to boost spatial intelligence. To evaluate these models, researchers have developed various benchmarks Yang et al. (2024; 2025c); Zhang et al. (2025); Zhou et al. (2025); Wang et al. (2026). However, both models and benchmarks operate under a "perfect image assumption", where images are clear and well illuminated, failing to reflect physical constraints and visual imperfections in real-world deployment.

Robustness of MLLMs Against Visual Degradations In unconstrained physical environments, visual inputs inevitably suffer from degradations caused by dynamic motion, adverse weather, and sensor limitations. Such corruptions have been standardized in ImageNet-C Hendrycks & Dietterich (2019), and recent works have begun evaluating MLLM robustness against common image corruptions Cui et al. (2023); Saxena et al. (2026); Usama et al. (2025); Tang et al. (2026); Fan et al. (2025). However, existing studies mainly focus on semantic understanding, object recognition, or

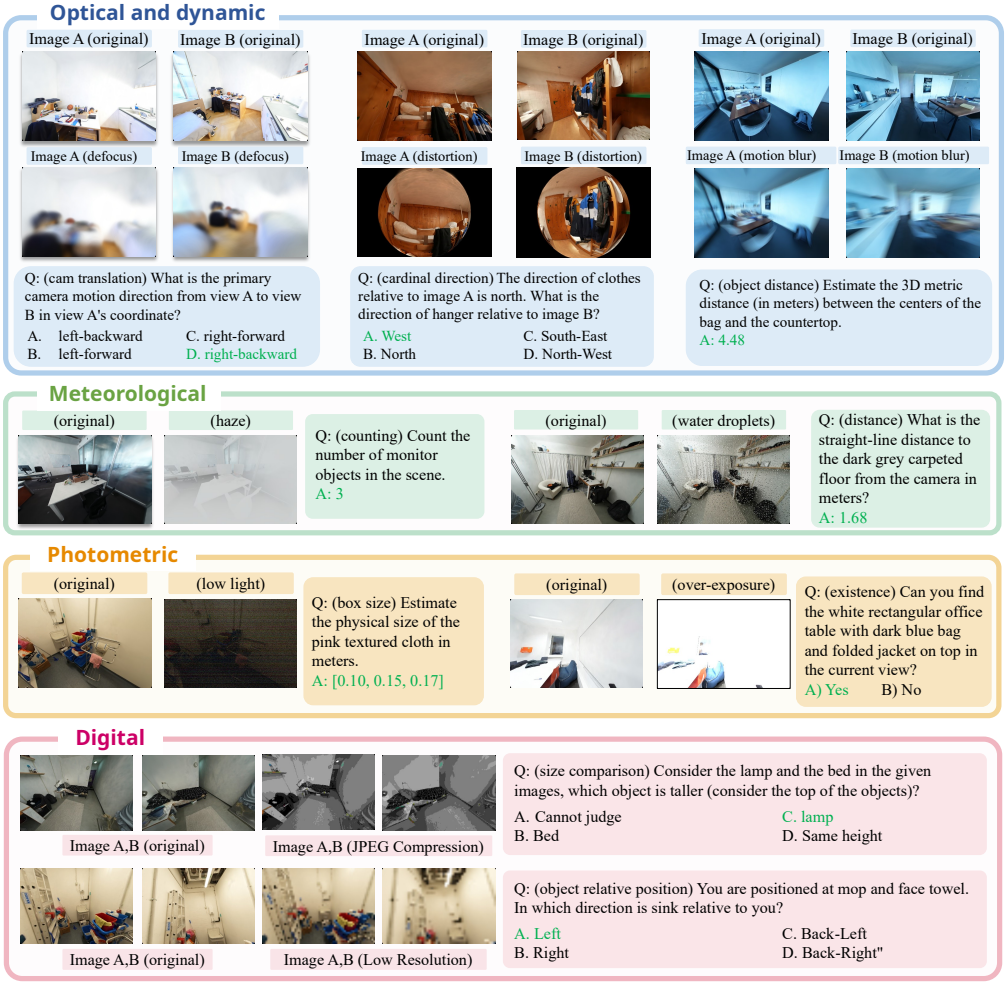


Figure 1: Four degradation categories and their specific degradation QA Examples. We show both clean and degraded views with the corresponding spatial question which are simplified version without detailed object descriptions. For full QA examples please refer to Appendix G.

basic visual reasoning Usama et al. (2025); Fan et al. (2025); Tang et al. (2026). The robustness of MLLMs under visual degradation for fine-grained spatial intelligence remains unclear.

3 SPACEDG

This section presents SpaceDG and SpaceDG-Bench, the first dataset and benchmark for spatial intelligence under visual degradations. We introduce our proposed data engine, starting with 3DGS-based scene representation and QA initialization, followed by a physically realistic degradation synthesis pipeline. Then we detail the constructed dataset, covering diverse spatial tasks, multiple viewpoints, and various visual degradations.

3.1 DATA ENGINE

3D Data Collection SpaceDG builds on the automatic 3D data curation pipeline of Holi-Spatial Gao et al. (2026), which converts raw video streams into geometry-consistent 3D semantic scenes. For each video, we first estimate depth and camera-pose priors with DepthAnything-v3 Lin et al. (2025) and COLMAP Schönberger & Frahm (2016) to optimize a geometrically constrained 3DGS representation. This gives us a renderable scene with calibrated camera poses and dense depth, which is critical for producing degradation variants without changing the underlying spatial ground truth. We then apply

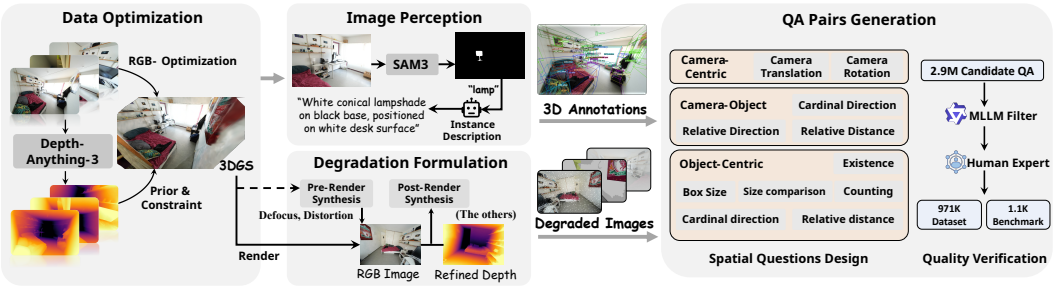


Figure 2: The SpaceDG data engine. Input multi-view RGB images are first reconstructed into geometry-consistent 3DGS on which degradations are formulated. Pre-rendered degradations include defocus and distortion, while post-rendered degradations contain the others. Then SAM3 masks are lifted to 3D instances with generated descriptions. QA pairs are generated with structured templates and automatically-calculated answers, followed by a two-stage MLLM-plus-human verification.

SAM3 Carion et al. (2026) to key frames to obtain per-view semantic masks. The masks are lifted and associated across views using the reconstructed depth, camera poses, and bounding-box IoU, yielding object-level 3D instances with 3D bounding boxes, visible-frame lists, and the highest-confidence view for each instance.

QA Pairs Generation We initialize spatial QA pairs directly from the reconstructed 3D scene information. First, for each 3D instance we generate a short, view-invariant language description by asking VLM for its appearance in its highest-confidence SAM3 mask image. These descriptions allow questions to refer to natural objects without adding artificial markers like boxes or points on evaluated imagesDeng et al. (2025). Second, we sample valid single-view and two-view observations using pairwise image covisibility and minimum baseline constraints, so that each question is both visually answerable and geometrically non-trivial. Following MapAnything Keetha et al. (2026), the covisibility score between two images is computed by reprojecting depth-supported 3D points from one calibrated view into the other and counting projections that pass a depth-based reprojection consistency check. Finally, we instantiate structured QA templates and compute answers from camera extrinsics, 3D box centers, object extents, and relative directions. This produces physically grounded answers for camera translation and rotation, object distance and direction, size comparison, and cross-view relational reasoning. Depending on the task, answers are represented as multiple-choice labels, binary decisions, or metric values. We provide QA examples in Figure 1, and detailed generation rules in Appendix D.

Degradation Synthesis Methods for compositing various degradations to RGB images have been thoroughly explored Nah et al. (2019); Wei et al. (2021); Wang et al. (2025b); Steinrucken (2017); Liao et al. (2024). To ensure physical realism and multi-view consistency, we further develop a physically grounded degradation pipeline that operates directly on 3DGS rendering process or linear light domain. We systematically inject 9 representative degradations across four categories: optical and dynamic degradations (defocus, distortion, motion blur), meteorological degradations (haze, water droplets), photometric degradations (low-light, over-exposure) and digital degradations (JPEG compression, low-resolution). As illustrated in Figure 2, all degradations are designed such that the underlying 3D spatial ground-truth remains invariant, ensuring accurate answers. We provide detailed formulations for each degradation process in Appendix C.

3.2 SpaceDG Dataset and SpaceDG-Bench

Statistics of SpaceDG Built upon the data engine, we construct SpaceDG dataset and SpaceDG-Bench. As shown in Table 1 and Figure 3, SpaceDG dataset contains 971,090 QA instances, covering 584 real indoor scenes with physically synthesized degraded images. Each sample is organized as image observations and spatial questions with corresponding answers derived from geometry-consistent 3D annotations. We further curate SpaceDG-Bench from 320 representative scenes that are disjoint from the SpaceDG training set, resulting in 1,102 manually verified questions (723 multi-view and 379 single-view). For each benchmark item, we render one clean condition (original) and nine

degraded conditions: defocus, distortion, haze, JPEG compression, low-light, low-resolution, motion blur, over-exposure, and water droplets. The benchmark is balanced at the image level, with 1,725 images per degraded condition, resulting in a benchmark with actual 9918 VQA pairs.

Statistic	SpaceDG Dataset	SpaceDG-Bench
Unique questions	971,090	1102
Unique images	162,071	15525
Number of degradations	9	9
Number of scenes	584	320
Multi-degradations per question	✗	✓
Single-view questions	276,542	379
Multi-view questions	694,548	723
Average image count per question	1.72	1.56
Final VQA pairs	971,090	9918

Table 1: Statistics of SpaceDG and SpaceDG-Bench. We report the number of unique questions, images, scenes, and degradation types, along with the breakdown of views.

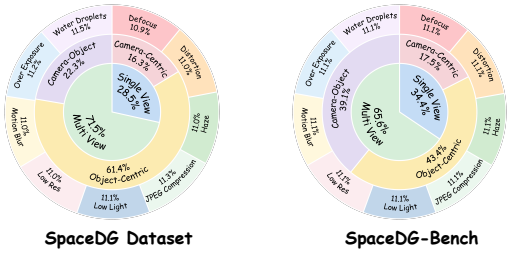


Figure 3: Distribution of SpaceDG and SpaceDG-Bench. Inner-to-outer rings show the proportion of QA pairs by view configuration, spatial task group, and degradation type.

Spatial Questions Design To guarantee a comprehensive assessment of spatial intelligence, we systematically design 11 distinct question categories categorized into single-view and multi-view settings. These tasks evaluate three fundamental aspects: (1) Camera-centric, requiring models to estimate camera translation distance and relative rotations (e.g., yaw, pitch, and roll) between viewpoints; (2) Object-centric, encompassing object counting, object direction, distance estimation, and fine-grained 3D spatial extents ($[w, l, h]$); and (3) Camera-object Relational, which evaluates inter-object and camera-object spatial relations like cross-view direction and relative positioning.

Quality Verification To ensure data quality, we employ a two-stage filtering pipeline combining a VLM-based agent with human review. In the first stage, Qwen3-VL-32B serves as an automated judge to eliminate ambiguous questions — any question description that could plausibly refer to multiple objects or be incorrect is discarded. In the second stage, a human expert manually screens the remaining QA pairs through a dedicated interface, resulting in approximately 2,000 candidate pairs. Finally, two experts independently review the candidate set and remove QA pairs with ambiguous descriptions, incorrect answers, or ill-formed options.

4 Experiments

4.1 Evaluation Setup

Baselines We systematically evaluate 25 models on SpaceDG-Bench, including proprietary models like GPT-5.4 OpenAI (2026), Gemini-3.1-Pro Google (2026b), Gemini-3.1-Flash-Lite Google (2026a), Claude-Sonnet-4.6 Anthropic (2026), open-source general models like Qwen3.5 Qwen Team (2026), InternVL3.5 Wang et al. (2025a), Kimi-VL Team et al. (2025), LLaVA-OneVision-1.5 An et al. (2025) and so on. We also evaluate domain-specific models like spatial-intelligence models Cai et al. (2025a); Yang et al. (2025a;b) and robotic brains Gong et al. (2026); Dang et al. (2026). For each model, we evaluate it on both clean images and 9 visual degradations using EASI Cai et al. (2025b) and VLMEvalKit Duan et al. (2024) under zero-shot settings. We also include two baselines: human-level assessment and non-image on GPT-5.4 and Qwen3-VL-8B-Instruct.

Evaluation Metrics To rigorously evaluate these heterogeneous answers, we adopt two metrics tailored to the output formats. For multiple-choice questions (MCQ) and binary-decision questions, we report Accuracy (Acc). For numerical answer (NA) questions requiring exact metric scalars (e.g., distances and sizes), we employ Mean Relative Accuracy (MRA) Yang et al. (2024) with confidence thresholds $\Theta = \{0.50, 0.55, \ldots, 0.95\}$. For list-type numerical answer (e.g., size estimation), we require the model to output answers in ascending order and calculate the metric by weighting each number.

Models	Clean Image	Degradation Types									Avg.
		Defocus	Distortion	Haze	JPEG-com.	Low-light	Low-res.	Motion-blur	Over-exp.	Water-droplets	Avg
Base											
Human Level	80.4	62.2	67.0	55.5	61.4	51.7	62.0	50.9	55.8	69.0	59.5
Non-Image *Qwen3-VL-8B-Instruct*	-	-	-	-	-	-	-	-	-	-	33.5
Non-Image *GPT-5.4*	-	-	-	-	-	-	-	-	-	-	35.1
Proprietary											
GPT-5.4	50.3	43.3	49.9	45.6	48.2	42.5	42.8	46.2	49.9	47.2	46.2
Gemini-3.1-Flash-Lite	56.9	44.2	55.7	46.1	52.8	43.1	48.0	46.5	52.3	50.4	48.8
Gemini-3.1-Pro	63.1	51.8	63.6	57.2	60.4	51.2	53.1	56.0	63.3	53.8	56.7
Claude-Sonnet-4.6	52.4	44.8	52.5	39.5	49.8	37.9	40.9	43.9	49.0	39.7	44.2
Grok-4.1-Fast	39.7	34.8	37.9	35.5	37.3	33.1	35.1	34.8	36.3	35.5	35.6
Qwen3.6-Plus	58.3	40.9	54.8	46.0	49.9	37.8	43.6	45.2	50.1	45.4	46.0
Open-source general model											
InternVL3-8B	42.5	37.0	41.9	37.7	41.5	38.2	41.4	39.5	42.4	40.9	40.1
InternVL3-5-38B	52.9	45.3	50.6	44.3	50.1	45.2	47.4	47.1	50.6	48.5	47.7
InternVL3-5-8B	46.7	39.2	45.2	38.6	44.1	36.3	40.8	41.2	44.6	43.0	41.4
Llava-OneVision-Qwen2-7b-SI	38.4	33.4	36.3	32.1	36.4	30.6	33.4	32.6	35.2	33.9	33.8
Gemma-4-26B-A4B-it	43.7	29.8	39.9	27.5	37.1	23.5	29.4	28.9	36.0	26.9	31.0
Llama-4-Maverick	41.1	31.6	39.4	34.3	37.4	30.5	33.8	29.9	35.8	35.7	34.3
Kimi-VL-A3B-Instruct	40.3	32.3	39.1	32.3	37.2	28.7	31.9	31.0	37.4	35.8	34.0
Qwen3-VL-4B-Instruct	48.5	37.2	44.1	37.5	43.1	34.8	38.5	38.1	43.7	40.4	39.7
Qwen3-VL-8B-Instruct	49.1	38.4	48.5	40.4	44.8	36.1	41.3	40.1	45.5	44.2	42.1
Qwen3-VL-32B-Instruct	55.0	43.8	53.9	41.9	49.2	36.8	42.3	43.6	49.4	45.9	45.2
Qwen3.5-4B	47.1	38.8	48.2	38.0	42.4	34.1	37.3	39.5	44.2	40.9	40.4
Qwen3.5-9B	49.3	38.4	46.7	38.3	45.4	37.3	41.0	40.6	46.8	41.2	41.7
Qwen3.5-27B	55.5	40.8	52.9	40.1	48.7	35.1	41.1	42.5	47.9	44.9	43.8
Qwen3.6-35B-A3B	53.9	40.3	50.9	38.1	47.3	33.4	39.8	41.4	47.5	42.6	42.4
Open-source spatial-intelligence model											
Cambrian-S-7B	28.4	25.7	28.1	26.3	28.7	27.2	25.5	25.4	27.5	27.9	26.9
VST-7B	46.8	40.0	46.4	39.0	42.3	35.5	40.7	41.7	44.6	39.4	41.1
SenseNova-SI-InternVL3-8B	57.9	51.2	57.2	50.7	54.2	49.7	53.5	53.4	56.5	53.7	53.3
Open-source robotic brain											
ACE-Brain-0-8B	50.2	43.1	48.9	45.1	46.7	41.2	42.7	43.8	47.5	46.6	45.1
RynnBrain-8B	51.7	45.6	50.5	41.3	47.1	41.0	43.7	44.9	47.5	45.3	45.2
Ours											
SpaceDG-SFT$_{InternVL-3.5-8B}$	70.9	64.6	67.6	62.2	67.5	**61.1**	**64.3**	64.4	68.5	66.7	65.2
SpaceDG-SFT$_{Qwen3-VL-8B-Instruct}$	**73.2**	**65.6**	**69.2**	**64.8**	**68.6**	59.2	63.9	**66.1**	**70.3**	**67.5**	**66.1**

Table 2: Quantitative comparison of models on SpaceDG-Bench. We evaluate proprietary, open-source general, spatial-intelligence, and robotic-brain models under the clean condition and nine visual degradations, together with human-level and non-image baselines.

4.2 Evaluation Results on SpaceDG-Bench

Visual degradation consistently impairs spatial reasoning. As shown in Table 2, all evaluated MLLMs achieve lower performance under degraded inputs than under clean images, demonstrating that spatial intelligence remains highly sensitive to realistic visual corruptions. For instance, Gemini-3.1-Pro performs best among tested proprietary models, achieving 63.1% on clean images but decreases to 56.7% on degraded images. Qwen3.6-Plus exhibits strong performance on clean images but suffers a severe performance decrease under degraded conditions, especially on defocus and low light, as shown in Figure 4. Open-source models exhibit a similar trend: for example, InternVL3.5-38B drops from 52.9% on clean images to 47.7% under degraded inputs, revealing a substantial performance gap between ideal and degraded visual conditions. Additionally, we report the performance of GPT-5.4 and Qwen-3-VL-8B-Instruct when provided with no input image, and both models perform significantly worse than their degraded-image counterparts, approaching random-guess level. This result confirms that our benchmark contains few exploitable language shortcuts, and that degraded images still retain rich visual information necessary for correct answers.

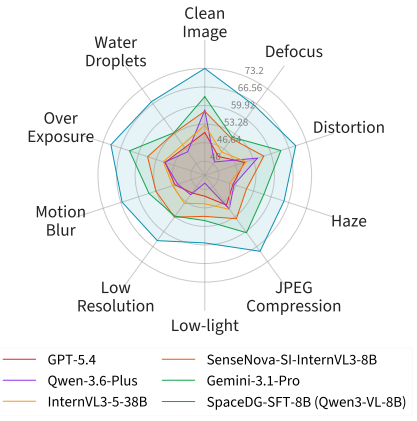


Figure 4: Per-degradation performance of representative models.

Humans struggle with extreme visual degradation. To establish a human reference baseline, we evaluate human performance on a 900-question subset of SpaceDG-Bench. As shown in Table 2,

humans achieve 80.4% accuracy on clean images, substantially outperforming all evaluated MLLMs. However, their performance drops by 20.9% under degraded conditions, indicating that severe visual corruptions can significantly impair fine-grained spatial judgment. These results suggest that spatial reasoning under degradation is challenging not only for current MLLMs but also for human observers, underscoring the need for degradation-aware training and evaluation protocols. Details of the human study are provided in Appendix B.5.

Degradation-aware SFT effectively improves the performance of MLLMs. We utilize the constructed SpaceDG to conduct supervised fine-tuning on Qwen-3-VL-8B-Instruct and InternVL-3.5-8B for 1 epoch with a batch size of 2048 using 8×H200 GPUs. Our SpaceDG-SFT-Qwen3 achieves substantial improvements over its base model across both clean and degraded conditions, rising from 49.1% to 73.2% on clean images and from 42.1% to 66.1% on degraded inputs. Notably, under degraded conditions, SpaceDG-SFT-Qwen3 surpasses the human reference performance of 59.5% by 6.6 percentage points. These results provide two key insights into degradation-aware spatial intelligence. First, supervised fine-tuning with degradation-augmented data substantially improves the spatial reasoning capability and robustness of MLLMs, suggesting that degradation-aware training is a practical path toward robust real-world spatial intelligence. Second, the gap between human and model performance under degraded conditions indicates that severe visual corruptions can also limit human spatial judgment, while models trained on large-scale degradation-aware data can learn to better exploit visual cues in challenging observations.

Models	**Degradation Types (Δ)**									**Avg. Δ**
	Defocus	Distortion	Haze	JPEG-com.	Low-light	Low-res.	Motion-blur	Over-exp.	Water-droplets	
Llava-OneVision-Qwen2-7b-SI	+0.4	-0.1	+1.5	+0.5	+1.6	-0.6	+0.4	+0.5	+1.1	+0.6
Qwen3-VL-4B-Instruct	+1.7	+1.3	+1.5	+0.9	+0.4	+1.3	+0.8	+2.5	+2.5	+1.4
Qwen3-VL-8B-Instruct	+2.4	+0.6	-1.0	+1.2	+1.1	-0.9	+0.4	+2.5	+0.6	+0.8
Qwen3-VL-32B-Instruct	+1.4	-0.1	+0.8	-0.3	+1.5	+1.4	+1.2	+0.6	-0.8	+0.6
Qwen3.5-9B	+0.6	+2.0	+0.7	+1.0	+2.3	-1.6	+1.1	+0.3	+0.2	+0.7
RynnBrain-8B	-1.8	+0.1	-0.1	+0.5	+0.0	-1.5	-0.5	+1.1	-1.3	-0.4
ACE-Brain-0-8B	-1.2	+0.3	-0.6	-0.3	+1.6	+0.1	+0.2	-0.0	-0.1	-0.0
VST-7B	-0.2	-1.8	-0.9	-0.1	+0.6	+0.1	-0.8	-1.3	-0.1	-0.5
SenseNova-SI-InternVL3-8B	-1.3	+0.4	-0.7	+0.2	-0.1	-0.9	-0.4	+0.1	+0.3	-0.3
SpaceDG-SFT-Qwen3-VL-8B	-1.1	-0.9	+2.5	-0.6	+0.4	+0.9	+0.6	-1.3	-0.5	-0.0

Table 3: Performance changes when degradation type and severity are explicitly provided by prompts. We use the degradation prompt template in Appendix B.6.

Spatial fine-tuning enhances the visual robustness of MLLMs, while reducing degradation comprehension capability. As shown in Table 2, spatially fine-tuned and robotic brain models exhibit a smaller performance drop when transitioning from clean to degraded inputs. On average, these models decline by 5.5%, compared to 7.6% for general models, indicating stronger inherent robustness to visual corruptions. However, Table 3 reveals an opposing trend in degradation comprehension capability: when the degradation type and severity are explicitly provided in the prompt, general-purpose models consistently benefit, achieving notable performance gains across all degradation categories. In contrast, spatially fine-tuned models show little to no improvement, with some even exhibiting a slight performance decrease. This suggests that spatial fine-tuning encourages models to develop degradation-agnostic visual representations, trading away sensitivity to image quality cues in favor of task-level robustness.

4.3 Degradation-wise Correlation Analysis

Metric Design To quantify the sensitivity of spatial reasoning performance to various image degradations, we adopt the absolute point-biserial Pearson correlation coefficient, $|r|$. For each analysis slice (e.g., answer format, task group, or specific question type) and degradation type, we construct paired observations across all evaluated models. Let the binary indicator $\mathbf{x} \in \{0, 1\}$ denote whether a score originates from a clean ($x = 0$) or degraded ($x = 1$) condition, and $\mathbf{y}$ represent the concatenated score vector across all M models from Table 2:

$$\mathbf{x} = [0, \ldots, 0, 1, \ldots, 1], \quad \mathbf{y} = [s_{\text{ori}}^{(1)}, \ldots, s_{\text{ori}}^{(M)}, s_{\text{deg}}^{(1)}, \ldots, s_{\text{deg}}^{(M)}]. \tag{1}$$

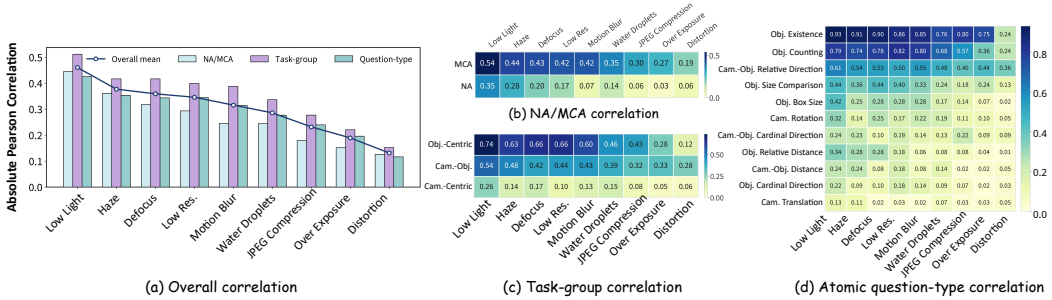


Figure 5: Degradation-wise correlation analysis. Sensitivity of spatial intelligence is measured by the absolute point-biserial correlation $|r|$. (a) Overall $|r|$ per degradation. (b) Breakdown by answer format. (c) Breakdown by task group. (d) Per-atomic-question correlation.

The correlation is computed as:

$$r = \mathrm{corr}(\mathbf{x}, \mathbf{y}) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}. \tag{2}$$

We report $|r|$ to reflect the magnitude of the degradation effect; a larger $|r|$ indicates a more significant score shift between clean and degraded conditions. The results are shown in Figure 5.

Analysis Across all subfigures in Figure 5 we can observe that low-light and haze consistently induce the most pronounced performance drops across models, whereas over-exposure and distortion have comparatively weaker effects. Figure 5 (b) further shows that Multiple-Choice Answer (MCA) questions exhibit higher degradation correlation than Numerical Answer (NA) questions. The task-group analysis in Figure 5 (c) indicates that object-centric tasks are the most sensitive to visual degradations, while camera-centric tasks remain relatively robust, suggesting that localized object grounding is more severely disrupted than global scene-level perception. At the atomic question-type level, fine-grained semantic perception tasks, such as existence estimation and object counting, exhibit the highest correlation, whereas tasks requiring global understanding, such as camera translation, show the lowest correlation. These findings suggest that **visual degradations primarily impair MLLMs' fine-grained semantic perception**, thereby disproportionately affecting tasks that require detailed visual grounding.

5 Degradation-guided Spatial Reasoning

5.1 Two-stage Chain-of-Thought Reasoning

To investigate the impact of Chain-of-Thought (CoT) Wei et al. (2023) for visual degradations, we design a structured two-stage CoT prompt. The MLLM is first required to explicitly output the degradation type from the provided list with a short description, followed by a classic prompt for reasoning. We conduct the experiment on Gemini-3.1-Flash-Lite and provide the detailed prompt in Appendix B.7. As shown in Table 4, the model with CoT suffers from a 1.8% decrease, demonstrating the harmness of CoT for degradation-guided reasoning.

Model	Method	Performance
Gemini-3.1-Flash-Lite	w/o CoT	48.8
	with CoT	47.0

Table 4: CoT performance of Gemini-3.1-Flash-Lite on SpaceDG-Bench.

5.2 Error Analysis

As shown in Figure 6, we further systematically examine the reasoning processes of Gemini-3.1-Flash-Lite and categorize degradation-induced errors into four types: (1) **degradation attribution**

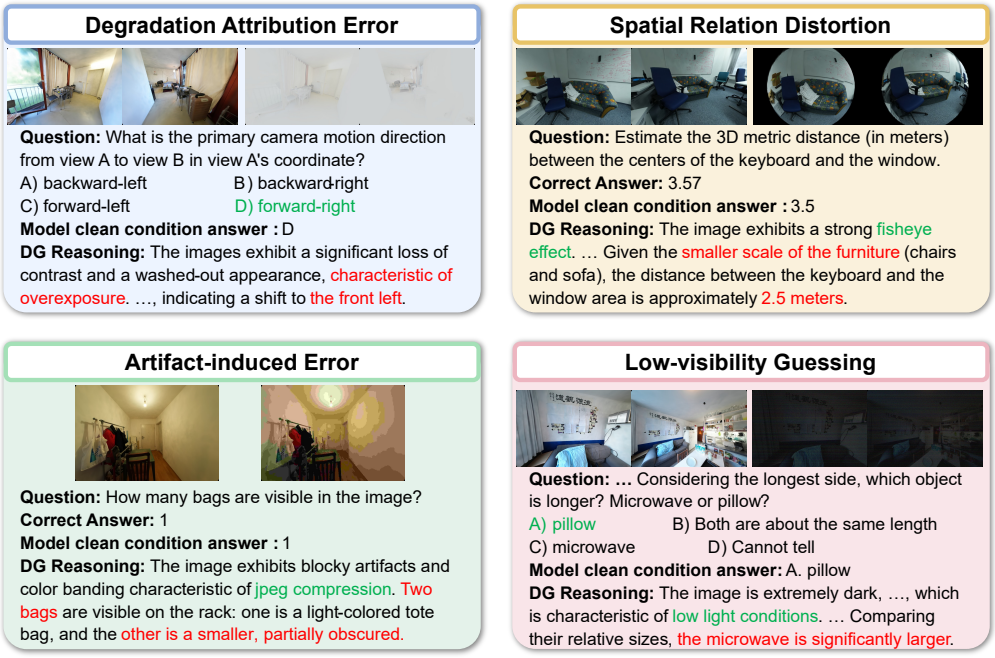


Figure 6: The four identified errors caused by visual degradations. We provide the correct answer, model answers under clean condition and reasoning processes under degraded condition.

errors, where the model misidentifies the underlying corruption type (e.g., mistaking haze for over-exposure), leading to an incorrect reasoning premise from the first step; (2) **spatial relation distortion**, where degraded visual cues bias orientation and relative-position judgments, causing systematic errors in directional and relational reasoning; (3) **artifact-induced errors**, where compression artifacts, blur, and low-resolution textures introduce spurious patterns that mislead object counting and numeric/metric estimation; and (4) **low-visibility guessing**, where the model acknowledges poor observability but still produces overconfident answers instead of performing calibration or giving a conservative answer. While these errors reflect different shortcomings of the model, a sample may contain multiple errors.

6 Conclusion

In this work, we study spatial intelligence of MLLMs under realistic visual degradations, a setting that is critical for real-world embodied and autonomous systems but largely overlooked by existing spatial reasoning benchmarks. We introduce SpaceDG, a large-scale dataset constructed with a physically grounded degradation synthesis engine built on 3DGS, and SpaceDG-Bench, a human-verified benchmark covering diverse spatial reasoning categories and nine representative degradation types. Through a comprehensive evaluation of 25 proprietary, open-source, spatially fine-tuned, and robotic-brain models, we show that visual degradations consistently impair spatial reasoning, revealing a substantial robustness gap between clean and imperfect visual conditions. Our analysis further shows that fine-grained object-level perception is particularly vulnerable to degradations, while certain global geometric reasoning tasks remain relatively more robust. Finally, we demonstrate that supervised fine-tuning on SpaceDG substantially improves both clean and degraded performance, suggesting that degradation-aware training is a promising direction for building robust spatially intelligent MLLMs. We hope SpaceDG and SpaceDG-Bench will facilitate future research on spatial reasoning beyond idealized visual inputs and encourage the development of models that can reason reliably under imperfect real-world observations.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writting and search literature. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training. In *arXiv*, 2025.

Anthropic. Introducing claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, February 2026.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2026. URL https://arxiv.org/abs/2410.24164.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. *arXiv preprint arXiv:2511.13719*, 2025a.

Zhongang Cai, Yubo Wang, Qingping Sun, Ruisi Wang, Chenyang Gu, Wanqi Yin, Zhiqian Lin, Zhitao Yang, Chen Wei, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Jiaqi Li, Xiangyu Fan, Hanming Deng, Lewei Lu, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Holistic evaluation of multimodal llms on spatial intelligence. *arXiv preprint arXiv:2508.13142*, 2025b.

Mingdeng Cao, Zhihang Zhong, Jiahao Wang, Yinqiang Zheng, and Yujiu Yang. Learning adaptive warping for real-world rolling shutter correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17785–17793, 2022.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 2024.

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3291–3300, 2018.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language models, 2024. URL https://arxiv.org/abs/2406.01584.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks, 2023. URL https://arxiv.org/abs/2312.03777.

Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, Jiangpin Liu, Yunxuan Mao, Zhikai Wang, Yuqian Yuan, et al. Rynnbrain: Open embodied foundation models. *arXiv preprint arXiv:2602.14979*, 2026.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and Peter Grasch. Mm-spatial: Exploring 3d spatial understanding in multimodal llms, 2025. URL https://arxiv.org/abs/2503.13111.

Nianchen Deng, Lixin Gu, Shenglong Ye, Yinan He, Zhe Chen, Songze Li, Haomin Wang, Xingguang Wei, Tianshuo Yang, Min Dou, et al. Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. *arXiv preprint arXiv:2506.18385*, 2025.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2): 295–307, 2015.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024.

Zhiyuan Fan, Yumeng Wang, Sandeep Polisetty, and Yi R. Fung. V^2r-bench: Holistically evaluating lvlm robustness to fundamental visual variations, 2025. URL https://arxiv.org/abs/2504.16727.

Xueyang Fu, Jiabin Huang, Xinghao Ding, Yinghao Liao, and John Paisley. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Transactions on Image Processing*, 26(6):2944–2956, 2017.

Yuanyuan Gao, Hao Li, Yifei Liu, Xinhao Ji, Yuning Gong, Yuanjun Liao, Fangfu Liu, Manyuan Zhang, Yuchen Yang, Dan Xu, Xue Yang, Huaxi Huang, Hongjie Zhang, Ziwei Liu, Xiao Sun, Dingwen Zhang, and Zhihang Zhong. Holi-spatial: Evolving video streams into holistic 3d spatial intelligence, 2026. URL https://arxiv.org/abs/2603.07660.

Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fqehqG4WvL.

Ziyang Gong, Zehang Luo, Anke Tang, Zhe Liu, Shi Fu, Zhi Hou, Ganlin Yang, Weiyun Wang, Xiaofeng Wang, Jianbo Liu, et al. Ace-brain-0: Spatial intelligence as a shared scaffold for universal embodiments. *arXiv preprint arXiv:2603.03198*, 2026.

Google. Gemini 3.1 flash-lite: Built for intelligence at scale. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-flash-lite/, March 2026a.

Google. Gemini 3.1 pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, Febrary 2026b.

Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. URL https://arxiv.org/abs/1903.12261.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models, 2026. URL https://arxiv.org/abs/2506.03135.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*. IEEE, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models, 2025a. URL https://arxiv.org/abs/2505.21500.

Yanbang Li, Ziyang Gong, Haoyang Li, Xiaoqi Huang, Haolan Kang, Guangping Bai, and Xianzheng Ma. Robotic visual instruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12155–12165, 2025b.

Zimu Liao, Siyan Chen, Rong Fu, Yi Wang, Zhongling Su, Hao Luo, Li Ma, Linning Xu, Bo Dai, Hengjie Li, Zhilin Pei, and Xingcheng Zhang. Fisheye-gs: Lightweight and extensible gaussian splatting module for fisheye cameras, 2024. URL https://arxiv.org/abs/2409.04751.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Peidong Liu, Zhaopeng Cui, Viktor Larsson, and Marc Pollefeys. Deep shutter unrolling network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5941–5949, 2020.

Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 457–466, 2022.

Gen Luo, Ganlin Yang, Ziyang Gong, Guanzhou Chen, Haonan Duan, Erfei Cui, Ronglei Tong, Zhi Hou, Tianyi Zhang, Zhe Chen, et al. Visual embodied brain: Let multimodal large language models see, think, and control in spaces. *arXiv preprint arXiv:2506.00123*, 2025.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3883–3891, 2017.

Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1996–2005, 2019. doi: 10.1109/CVPRW.2019.00251.

Muyao Niu, Zhuoxiao Li, Zhihang Zhong, and Yinqiang Zheng. Visibility constrained wide-band illumination spectrum design for seeing-in-the-dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13976–13985, 2023a.

Muyao Niu, Zhihang Zhong, and Yinqiang Zheng. Nir-assisted video enhancement via unpaired 24-hour data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10778–10788, 2023b.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, March 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Rohit Saxena, Alessandro Suglia, and Pasquale Minervini. Vlm-robustbench: A comprehensive benchmark for robustness of vision-language models, 2026. URL https://arxiv.org/abs/2603.06148.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Martijn Steinrucken. Heartfelt – Shadertoy. https://www.shadertoy.com/view/ltffzl, 2017. License: CC BY-NC-SA 3.0.

Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1279–1288, 2017.

Jiaqi Tang, Jianmin Chen, Wei Wei, Xiaogang Xu, Runtao Liu, Xiangyu Wu, Qipeng Xie, Jiafei Wu, Lei Zhang, and Qifeng Chen. Robust-r1: Degradation-aware reasoning for robust visual understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Muhammad Usama, Syeda Aishah Asim, Syed Bilal Ali, Syed Talal Wasim, and Umair Bin Mansoor. Analysing the robustness of vision-language-models to common corruptions, 2025. URL https://arxiv.org/abs/2504.13690.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Jiajun Wu, Li Fei-Fei, and Manling Li. Mindcube: Spatial mental modeling from limited views, 2026. URL https://arxiv.org/abs/2506.21458.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025a. URL https://arxiv.org/abs/2508.18265.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

Yujie Wang, Praneeth Chakravarthula, and Baoquan Chen. Dof-gs: Adjustable depth-of-field 3d gaussian splatting for refocusing, defocus rendering and blur removal. *The IEEE / CVF Computer Vision and Pattern Recognition Conference*, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8520–8537, 2021.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025.

LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. URL https://arxiv.org/abs/2506.03569.

Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*, 2024.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025a.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025b.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *ICLR*, 2025c.

Yuchen Yang, Yuqing Shao, Duxiu Huang, Linfeng Dong, Yifei Liu, Suixin Tang, Xiang Zhou, Yuanyuan Gao, Wei Wang, Yue Zhou, et al. Stepping vlms onto the court: Benchmarking spatial intelligence in sports. *arXiv preprint arXiv:2603.09896*, 2026.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Ziang Zhang, Zehan Wang, Guanghao Zhang, Weilong Dai, Yan Xia, Ziang Yan, Minjie Hong, and Zhou Zhao. Dsi-bench: A benchmark for dynamic spatial intelligence, 2025. URL `https://arxiv.org/abs/2510.18873`.

Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *European conference on computer vision*, pp. 191–207. Springer, 2020.

Zhihang Zhong, Yinqiang Zheng, and Imari Sato. Towards rolling shutter correction and deblurring in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9219–9228, 2021.

Zhihang Zhong, Mingdeng Cao, Xiang Ji, Yinqiang Zheng, and Imari Sato. Blur interpolation transformer for real-world motion from blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5713–5723, 2023a.

Zhihang Zhong, Ye Gao, Yinqiang Zheng, Bo Zheng, and Imari Sato. Real-world video deblurring: A benchmark dataset and an efficient recurrent neural network. *International Journal of Computer Vision*, 131(1):284–301, 2023b.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Eric Xin Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8600–8612, 2025.

A Analysis of the Synthetic Degradation Gap

To evaluate whether the synthetic degradations used in SpaceDG approximate real-world visual degradation, we compare degradation statistics between synthetic samples generated by SpaceDG and real degraded images from DPDD (defocus blur), RealBlur-J (motion blur), I-HAZE (haze), and LOLv2-real (low light). For each degradation type, we compute statistics on paired clean/degraded images and normalize them using the median and interquartile range (IQR) of the corresponding real distribution. We further measure the distribution discrepancy using the 1-Wasserstein distance:

$$\tilde{x} = \frac{x - \mathrm{median}(X_{\mathrm{real}})}{\mathrm{IQR}(X_{\mathrm{real}})}, \quad W_1(P, Q) = \int_0^1 |F_P^{-1}(u) - F_Q^{-1}(u)| du. \tag{3}$$

Table 5: Comparison between real and SpaceDG synthetic degradations. Statistics are computed on paired clean/degraded images. Lower W_1 indicates closer distributions.

Degradation	Statistic	Real median [Q1,Q3]	SpaceDG median [Q1,Q3]	$W_1 \downarrow$
Defocus blur	SSIM	0.820 [0.713,0.895]	0.807 [0.742,0.889]	0.145
Defocus blur	MAE	0.0377 [0.0261,0.0524]	0.0387 [0.0227,0.0677]	0.316
Motion blur	Gradient coherence	0.335 [0.192,0.462]	0.378 [0.261,0.542]	0.267
Haze	SSIM	0.714 [0.651,0.782]	0.705 [0.574,0.794]	0.165
Low light	SSIM	0.174 [0.107,0.288]	0.213 [0.181,0.248]	0.297
Low light	Gradient retention	0.327 [0.251,0.373]	0.347 [0.287,0.453]	0.504

As shown in Table 5, SpaceDG produces degradation statistics that are close to real-world samples across multiple degradation types. The relatively small standardized Wasserstein distances indicate that the proposed synthetic degradations capture important characteristics of real degradations, including structural distortion, blur effects, contrast reduction, and illumination attenuation. While synthetic degradations cannot fully reproduce all acquisition-specific factors, these results demonstrate that SpaceDG provides a practical approximation of real visual degradation patterns.

B Additional Experiments

Training Method	**Clean Image**	**Degradation Types**									**Avg.**
		Defocus	Distortion	Motion-blur	Haze	Water-droplets	Low-light	Over-exp.	JPEG-com.	Low-res.	
No-SFT	49.1	38.4	48.5	40.4	44.8	36.1	41.3	40.1	45.5	44.2	42.1
Full-SFT with degradations	73.2	65.6	69.2	66.1	64.8	67.5	59.2	70.3	68.6	63.9	66.1
Full-SFT with clean images	73.1	63.2	68.3	64.3	65.4	65.9	58.3	67.1	67.0	64.2	64.8
Held-out Degradation Types											
w/o Optical & Dynamic	Defocus, Distortion, Motion Blur	64.0	69.9	65.2							
w/o Meteorological	Haze, Water Droplets				64.3	66.02					
w/o Photometric	Low Light, Over Exposure						57.8	66.1			
w/o Digital	JPEG Compression, Low Resolution								67.4	63.8	

Table 6: Ablation study of SpaceDG SFT. Using Qwen3-VL-8B-Instruct as the base model, we compare the no-SFT baseline, full SFT with and without degraded images, and held-out variants that exclude each degradation category at a time.

B.1 Data Validation

Degradation-aware SFT enhances robust spatial reasoning. To isolate the effect of degradation augmentation from the general benefit of SFT, we compare degradation-augmented SFT against a clean-image SFT with the same training settings. As shown in Table 6, both achieve nearly identical performance on clean inputs (73.2% vs. 73.1%), indicating comparable spatial understanding ability under pristine conditions. However, under degraded inputs, degradation-augmented SFT improves the average performance from 64.6% to 66.1%, demonstrating that robustness gains from degradations. Further held-out experiments show that models trained without specific degradation categories still generalize well to unseen corruptions, substantially outperforming the no-SFT baseline. These results suggest that degradation-aware SFT encourages degradation-agnostic spatial reasoning strategies and is essential for robust real-world spatial intelligence.

B.2 Generalization to Outdoor Autonomous Driving Scenes

To further evaluate the cross-domain generalization ability of SpaceDG-SFT, we conduct experiments on Ego3D-Bench Gholami et al. (2026), which consists of real-world outdoor autonomous driving scenes collected from Waymo, nuScenes, and Argoverse. These scenes differ substantially from the indoor ScanNet++ training domain in terms of scene content, spatial scale, camera viewpoints, and visual complexity. We randomly sample 500 questions from Ego3D-Bench and evaluate both the original model and SpaceDG-SFT under clean and visually degraded inputs.

As shown in Table 7, SpaceDG-SFT consistently outperforms the base model under both clean and degraded conditions. Notably, SpaceDG-SFT improves the accuracy by 11.05% on clean inputs and 13.74% on degraded inputs, while maintaining a much smaller performance gap between clean and

degraded settings (0.86% vs. 3.55%). These results demonstrate that the spatial reasoning capability and degradation robustness learned from indoor scenes can generalize effectively to challenging outdoor environments with dynamic objects.

Model	Setting	Accuracy (%) ↑
Qwen3-VL-8B	Clean	45.97
Qwen3-VL-8B	Degraded	42.42
SpaceDG-SFT-Qwen3-VL-8B	Clean	**57.02**
SpaceDG-SFT-Qwen3-VL-8B	Degraded	**56.16**

Table 7: Generalization results on Ego3D-Bench outdoor autonomous driving scenes. SpaceDG-SFT is trained only on degraded indoor scenes but transfers effectively to outdoor environments.

B.3 Real-world Inspired Mixture of Degradations

Real-world images rarely suffer from a single isolated corruption; instead, multiple degradations often co-occur due to complex acquisition conditions. To evaluate robustness under such compound effects, we extend SpaceDG from single-degradation evaluation to a real-world inspired mixture protocol. We design six mixture recipes that reflect representative capture scenarios: night capture, hazy long-range observation, wet-lens motion, backlit dynamic scenes, motion-defocus, and compressed portrait sharing. In each recipe, we pick one primary degradation applied the same degradation settings with SpaceDG-Bench, while all auxiliary degradations are applied at easier severity level. The primary degradation for each recipe is highlighted in Table 8.

As shown in Table 8, compound degradations substantially challenge the base model, whose performance drops to an average accuracy of 37.0 across the six mixture settings. In contrast, SpaceDG-SFT-Qwen3-VL-8B-Instruct achieves consistently higher performance in all scenarios, achieving an average accuracy of 62.4. The gains are especially pronounced under hazy long-range observation, compressed portrait sharing, and motion-defocus, suggesting that SpaceDG training improves not only robustness to individual corruptions but also generalization to realistic combinations of multiple visual degradations. These results indicate that spatial reasoning remain sensitive to compound image degradation, while degradation-aware training can substantially enhance their reliability in real-world visual conditions.

Models	Night Capture	Hazy Long-range	Wet-lens Motion	Backlit Dynamics	Motion-Defocus	Compressed Portrait
	LL+MB+LR	**HZ**+LR+MB	**WD**+MB+LL	**OE**+MB	**MB**+DF+LR	**DF**+JPEG
Qwen3-VL-8B-Instruct	31.5	36.2	32.5	44.7	39.8	37.4
SpaceDG-SFT-Qwen3-VL-8B-Instruct	55.8	65.1	55.6	68.3	65.7	63.8

Table 8: Performance under six mixed-degradation settings. Each column corresponds to one real-world inspired mixture recipe. The degradation in bold is the primary corruption. “LL”, “MB”, “LR”, “HZ”, “WD”, “OE”, “DF”, and “JPEG” denote low-light, motion blur, low resolution, haze, water droplets, over-exposure, defocus, and JPEG compression, respectively.

B.4 Will supervised fine-tuning on degraded dataset affect the performance of general benchmarks?

To investigate whether fine-tuning on a degraded dataset adversely affects spatial capabilities on clean images, we evaluate SpaceDG-SFT-8B on two general benchmarks: MMSI-Bench and MindCube. As shown in Table 9, SpaceDG-SFT-8B achieves scores of 30.0 and 37.0, respectively. These results outperform several baselines of comparable scale, including Qwen3-VL-8B and SpaceI-SFT-7B, and remain competitive with stronger models such as Intern3-VL-8B and VST-SFT-7B. These findings suggest that fine-tuning on the degraded dataset does not significantly compromise the model’s general capabilities.

Table 9: Extra comparison on general benchmarks MMSI-Bench and MindCube.

Model	MMSI-Bench	MindCube
VST-SFT-3B	30.2	35.9
Cambrian-S-3B	25.2	32.5
VST-SFT-7B	32.0	39.7
Cambrian-S-7B	25.8	39.6
SpaceI-SFT-7B	27.4	37.9
Intern3-VL-8B	28.0	41.5
Spatial-MLLM	27.0	32.1
Qwen3-VL-8B	31.1	29.4
SpaceDG-SFT-Qwen3-VL-8B-Instruct	31.3	37.0

B.5 Human-level Assessment

We conduct human-level assessment on SpaceDG-Bench-900, a 900-question subset of SpaceDG-Bench. To validate the reliability of this subset, we compare the performance of 13 representative models on both the full SpaceDG-Bench (9,918 samples) and this 900-sample subset (100 questions per degradation). As detailed in Table 10, the models exhibit highly consistent performance across both evaluation sets, yielding an average absolute performance difference of merely 0.83%. Specifically, the maximum performance gap is only 1.8% (Gemini-3.1-Pro), and the minimum discrepancy is as low as 0.3% (GPT-5.4). This negligible variance empirically demonstrates that SpaceDG-Bench-900 preserves the data distribution and task difficulty of the full evaluation benchmark.

During the evaluation, we divide human annotators into two groups and assess their performance separately under clean and degraded conditions. On clean images, humans achieve an overall accuracy of 80.4%, with 90.2% accuracy on multiple-choice answer (MCA) questions, 61.2% on numerical-answer (NA) questions, and 48.1% on list-type NA questions. This breakdown suggests that humans perform well on general spatial reasoning questions but struggle with tasks requiring precise metric estimation. Human performance under each degradation type is summarized in Table 2.

Table 10: Model Performance on SpaceDG-Bench and SpaceDG-Bench-900. We present this result to demonstrate the reliability of human benchmarks.

Model	SpaceDG-Bench	SpaceDG-Bench-900
Number of samples	9918	900
GPT-5.4	46.2	46.5
Claude-Sonnet-4.6	44.2	43.4
Gemini-3.1-Pro	56.7	54.9
Gemini-3.1-Flash-Lite	48.8	48.3
Qwen3-VL-4B-Instruct	39.7	38.7
Qwen3-VL-8B-Instruct	42.1	42.9
InternVL3-5-8B	41.4	42.2
InternVL3-5-38B	47.7	48.8
Qwen3.5-4B	40.4	39.7
Qwen3.5-9B	41.7	42.5
Qwen3.6-35B-A3B	42.4	41.7
SenseNova-SI-InternVL3-8B	53.3	52.8
ACE-Brain-0-8B	45.1	44.1

B.6 Degradation Prompt Template

We use the prompt template shown in Figure 7 to explicitly provide the MLLM with degradation-aware information during evaluation. Specifically, for each degraded input, the prompt augments the original spatial question with the degradation type and its corresponding severity, represented by the rendering parameter range. The parameter ranges used to generate SpaceDG-Bench are summarized

Degradation	Parameter ranges
defocus	aperture $\in [10.0, 15.0]$, depth $\in [1.0, 8.0]$
distortion	k1 $\in [-0.24, -0.23]$, k2 $\in [0.0001, 0.0003]$, k3 $\in [0.0001, 0.0002]$, k4 $\in [0.0000, 0.0001]$, max_theta $= 1.5$
haze	density $\in [3.5, 6.0]$
jpeg_compression	quality $\in [2, 5]$
low_light	exposure $\in [0.003, 0.005]$
low_res	scale $\in [0.02, 0.05]$
motion_blur	trans $\in [0.2, 0.35]$, rot $\in [0.06, 0.12]$, sub_steps $\in [80, 80]$
over_exposure	exposure $\in [7.0, 10.0]$
water_droplets	scale $\in [2.5, 4.0]$, radius $\in [0.25, 0.75]$, strength $\in [0.3, 0.5]$, blur_sigma $\in [2.0, 2.5]$, blur_kernel $= 9$

Table 11: Parameter ranges for each degradation setting in SpaceDG-Bench.

in Table 11, and correspond to the degradation formulations introduced in Section C, covering optical and dynamic, meteorological, photometric, and digital degradation processes. Reporting these settings makes the degradation-guided evaluation protocol and benchmark rendering configuration explicit and reproducible.

Degradation Prompt Template

USER:
f"{question}\n"
f"Note: The provided images have a {degradation_type} degradation with parameters range: {param_range}\n"

Figure 7: Two-stage prompt template for degradation-guided spatial reasoning.

B.7 Degradation-guided Chain-of-Thought

We provide the two-stage prompt template used to elicit the reasoning capability of Gemini-3.1-Flash-Lite in Figure 8. The prompt first asks the model to identify the degradation type from a predefined set, and then perform step-by-step spatial reasoning based on the observed degraded images. We further report the degradation recognition accuracy in Table 12. Although Gemini-3.1-Flash-Lite accurately recognizes most degradation types, its performance is substantially lower on haze and low resolution, indicating that degradation attribution itself remains challenging. Such attribution errors propagate to the subsequent reasoning stage and lead to incorrect spatial conclusions.

Chain-of-Thought Prompt Template

USER:
"The images may exhibit at most one of the following degradations: [defocus, distortion, haze, jpeg compression, low light , low resolution, motion blur, over exposure, water droplets].\n"
"Output format (strict order):\n"
"1) First discribe the degradation using ONE sentence, then output exactly one line: <degradation>LABEL</degradation> where LABEL is one of: [defocus, distortion, haze, jpeg_compression, low_light, low res, motion_blur, over_exposure, water_droplets].\n"
"2) Then briefly think step by step about the user's visual/spatial question.\n"
" Finally output the task answer inside <answer>...</answer> (use the format required by the user question).\n"
f"Question: {question}"

Figure 8: Two-stage prompt template for degradation-guided spatial reasoning.

C Detailed Degradation Synthesis Pipeline

C.1 Optical and Dynamic Degradations

These degradations are strictly coupled with camera physics and motion. By implementing them internally within the 3DGS rasterizer, we ensure strict geometric consistency across multiple views.

Defocus. We model the depth-of-field effect caused by a finite camera aperture using the thin-lens approximation. The Circle of Confusion (CoC) radius r_{CoC} is computed based on the rendered view

Table 12: Accuracy of Gemini-3.1-Flash-Lite for reconizing each degradation.

Models	Accuracy of Degradation Recognition								
	Defocus	Distortion	Haze	JPEG-com.	Low-light	Low-res.	Motion-blur	Over-exp.	Water-droplets
Gemini-3.1-Flash-Lite	96.6	99.9	20.4	98.0	100.0	46.9	99.7	86.2	100.0

depth d, the focus depth f, and the aperture size a:

$$r_{\text{CoC}} = a\frac{|d-f|}{d} \tag{4}$$

To simulate this directly within the 3DGS pipeline, this variance is added to the 2D projected covariance matrix of each Gaussian:

$$\tilde{\mathbf{\Sigma}}_{2D} = \mathbf{\Sigma}_{2D} + r_{\text{CoC}}^2\mathbf{I} \tag{5}$$

Furthermore, an opacity compensation term $\alpha_{\text{comp}} = \sqrt{\det(\mathbf{\Sigma}_{2D})/\det(\tilde{\mathbf{\Sigma}}_{2D})}$ is applied to strictly ensure energy conservation during the differentiable rasterization process.

Distortion. Real-world wide-angle or fisheye lenses introduce significant non-linear geometric warping. We modify the standard pinhole projection model in the CUDA rasterizer using an equidistant polynomial model. The distorted angle θ_d is given by:

$$\theta_d = \theta\left(1 + \sum_{i=1}^{4} k_i\theta^{2i}\right) \tag{6}$$

where k_i represents the radial distortion coefficients. The Jacobian matrix $\mathbf{J}_{\text{fisheye}}$ is then recomputed via automatic differentiation to accurately project the 3D covariance into the distorted screen space.

Motion Blur. Caused by camera movement during exposure, motion blur is simulated via the continuous time integration of linear light. We interpolate the camera poses using Spherical Linear Interpolation (Slerp) and accumulate frames over N sub-steps:

$$I_{\text{blur}} = \frac{1}{N}\sum_{i=0}^{N-1} I_{\text{lin}}(\mathbf{T}(t_i)) \tag{7}$$

where $\mathbf{T}(t_i)$ denotes the camera extrinsic matrix at time t_i and $N = 80$. This approach guarantees highly realistic directional and rotational blur that strictly adheres to the 3D scene geometry.

C.2 Meteorological Degradations

These effects depend heavily on the continuous spatial depth of the scene, which is natively provided by our 3DGS representations.

Haze. We simulate atmospheric scattering using the classic Koschmieder's law. Utilizing the accurate depth map $d(x)$ rendered directly from the 3DGS model, the degraded image intensity $I(x)$ at pixel x is formulated as:

$$I(x) = J(x)e^{-\beta d(x)} + A\left(1 - e^{-\beta d(x)}\right) \tag{8}$$

where $J(x)$ is the original scene radiance, β is the scattering coefficient determining the haze density, and A is the global atmospheric light. This model naturally enforces depth-dependent visibility decay.

Water Droplets. To simulate droplets on the camera lens, we generate a procedural multi-layer height map $h(x, y)$ to derive pixel-wise surface normals $\hat{\mathbf{n}} = (n_x, n_y, n_z)$. These normals are utilized to compute localized refraction offsets based on simplified Snell's law:

$$\Delta u \propto n_x, \quad \Delta v \propto n_y \tag{9}$$

These offsets are combined with local optical blurring and Phong specular highlights to comprehensively mimic the complex optical behavior of water droplets interacting with the scene's light field.

C.3 PHOTOMETRIC DEGRADATIONS

These simulate real-world illumination changes and sensor imperfections. Crucially, these operations are performed in the linear light domain $I_{\text{lin}} = I_{\text{sRGB}}^{\gamma}$ before final image encoding.

Low-light. We first scale the scene illumination by an exposure coefficient $e \ll 1$. To simulate the degraded signal-to-noise ratio (SNR) in low-light environments, we inject a physics-based sensor noise model. The noisy observation I_{noisy} incorporates both photon shot noise (modeled as a Poisson distribution) and read noise (modeled as a Gaussian/Tukey-Lambda distribution):

$$I_{\text{noisy}} \sim \mathcal{P}\left(\frac{e \cdot I_{\text{lin}}}{k}\right) \cdot k + \mathcal{N}(0, \sigma_{\text{read}}^2) \tag{10}$$

where k is the system gain and σ_{read} represents the standard deviation of the electronic read noise.

Over-exposure. To simulate sensor saturation caused by intense light sources or prolonged exposure, we apply a large exposure gain $e \gg 1$ and inject standard sensor noise. The values are strictly clipped to the sensor's maximum capacity:

$$I_{\text{clip}} = \max(\min(I_{\text{noisy}}, 1), 0) \tag{11}$$

This clipping process is applied prior to Gamma encoding, accurately replicating the irreversible loss of high-frequency textures and geometric details in saturated regions (e.g., near windows or light bulbs).

C.4 DIGITAL DEGRADATIONS

These degradations model common artifacts introduced during post-capture signal processing, storage, and transmission phases. Unlike physical degradations, they operate directly on the encoded 2D image matrix.

JPEG Compression: We explicitly apply Discrete Cosine Transform (DCT) block quantization controlled by a quality factor q. This transformation discards high-frequency coefficients, intentionally introducing the ringing and blocking artifacts typical of low-bandwidth network transmission.

Low-resolution: We simulate limited sensor resolution or aggressive downsampling by reducing the image resolution of cameras with a scale factor s. The image is then upsampled to original resolution. This systematically truncates high-frequency spatial details while maintaining the original image dimensions for the MLLM input format.

D DETAILED QA INITIALIZATION PIPELINE

The QA initialization from reconstructed scenes and instances follows Holi-Spatial Gao et al. (2026), and we also introducing additional single-view questions, including bbox extent, single-view object distance, object counting and existence.

Instance descriptions. For each reconstructed 3D instance, we select its highest-confidence SAM3 mask and overlay the mask contour on the corresponding RGB frame. A VLM is prompted to produce a concise description based on intrinsic, view-stable cues, such as color, material, texture, subtype, text markings, and distinctive structural details. View-dependent phrases are explicitly prohibited, since they would become invalid under a different camera pose. The generated description is stored with the 3D instance and later used as a natural-language reference in QA templates.

View sampling. For multi-view questions, we sample image pairs from the scene covisibility matrix while enforcing non-trivial camera motion. The matrix stores pairwise covisibility between every two images in a scene. We compute it in the same manner as MapAnything Keetha et al. (2026): depth pixels in a source view are lifted to 3D, reprojected into a target view using the calibrated poses, and counted as covisible only when the target-view depth agrees with the expected reprojected depth under a depth-association threshold. The final score is the normalized number of consistent

reprojected pixels. Camera-centric translation questions use a minimum translation baseline, rotation questions use a minimum relative rotation, and object-centric questions require the referenced instance or instances to be visible in the required view(s). For relational questions involving three objects, we additionally require the union of the two views to contain at least three distinct instances and avoid cases where all objects are simultaneously visible in a single image, preventing the task from collapsing into a single-view problem.

Question families. SpaceDG instantiates camera-centric, object-centric, and camera-object relational templates. Camera-centric templates cover dominant translation direction, metric translation distance, thresholded translation decisions, and relative rotation. Object-centric templates cover object depth, view-relative direction, inter-object 3D distance, and size or height comparison. Cross-view relational templates ask models to transfer an assumed direction or infer relative position across two views. All ground-truth answers are computed from calibrated camera extrinsics and 3D instance annotations rather than inferred from image pixels.

Option construction and boundary control. We use task-specific rules to construct multiple-choice options while avoiding geometrically ambiguous negatives. For camera- and object-direction questions, a secondary axis is included only when its magnitude is sufficiently large relative to the dominant axis ($0.5774\times$, corresponding to a 30° angular ratio), which prevents weak off-axis components from changing the textual direction label. For 8-way relative-position questions, the horizontal plane is divided into eight 45° sectors. If the ground-truth yaw falls within 3° of a sector boundary, the adjacent sector on the boundary side is forbidden as a distractor, so a near-boundary ground truth is not paired with an almost-correct neighboring option. Direction questions whose horizontal displacement is too small, or whose vertical component dominates the horizontal displacement, are discarded.

Geometric ambiguity filters. We discard spatial configurations that make the intended relation ill-defined. For triplet-based relation questions, any triplet with intersecting 3D bounding boxes is removed before constructing the local coordinate frame. The local frame is defined by an anchor object and a reference object; if the two centers are too close or the forward direction is nearly collinear with the world-up reference, the sample is skipped. Camera rotation questions are also filtered with a grey-zone rule: relative rotations below 5° are discarded, and cases near the boundary between "single-dominant" and "dual-dominant" rotation are skipped using a 0.1 buffer around the component-ratio threshold. For thresholded camera-translation questions, the sampled decision threshold is forced away from near equality by moving factors in $(0.85, 1.15)$ to the boundary, reducing accidental ambiguity between "yes" and "no". For size-comparison questions, objects are labeled as the same length or height only when their computed values differ by less than 10^{-2}.

Ambiguity filtering. Language descriptions can be ambiguous when multiple instances of the same category appear in one image. We therefore build an image-level index over same-label instances. If two same-label instances share the same description, the corresponding QA is discarded. Otherwise, a VLM judge is asked whether the target description uniquely identifies exactly one instance among same-category alternatives; uncertain cases are conservatively dropped. This automated filtering is followed by the manual verification procedure described in Section 3.2.

E Quality Verification

E.1 MLLM Filter Prompt Template

The MLLM filter prompt template is shown in Figure 9. We provide the MLLM filter with: clean image, segmentation category and generated description. The prompt requires the MLLM to answer with a key word "KEEP" or "DROP" according to observed ambiguity.

E.2 Human Review Interface

We design a benchmark editor interface to support manual verification of SpaceDG-Bench. As shown in the Figure 10, each sample is presented with its rendered image views, the corresponding spatial question, the ground-truth answer, and metadata such as task group and question type. Human

MLLM Filter Prompt Template

SYSTEM:
"You are a strict dataset QA filter.\n"
USER:
"You are given an image and multiple language descriptions of objects of the SAME category in this image.\n"
"Task: Decide if the TARGET description can uniquely identify exactly one object among OTHER objects of that category in the image.\n"
"Be conservative: if there is any plausible ambiguity, answer DROP.\n\n"
f"Scene: {ref.scene_id}\n"
f"Image: {ref.image_name}\n"
f"Category(label): {ref.label}\n\n"
f"TARGET description:\n{ref.description}\n\n"
"OTHER descriptions (same label in same image):\n"
f"{others_str}\n\n"
"Output JSON on a single line with keys:\n"
" - decision: "KEEP" or "DROP" \n
" - reason: short English reason\n "
"Do NOT output any other text."

Figure 9: MLLM filter prompt template.

reviewers can navigate through samples, filter cases by task or keyword, and directly edit the question or answer when ambiguity, formatting issues, or incorrect labels are observed. All modifications are saved back to the new benchmark file, enabling an efficient and traceable review process for improving the quality and consistency of the final evaluation set.

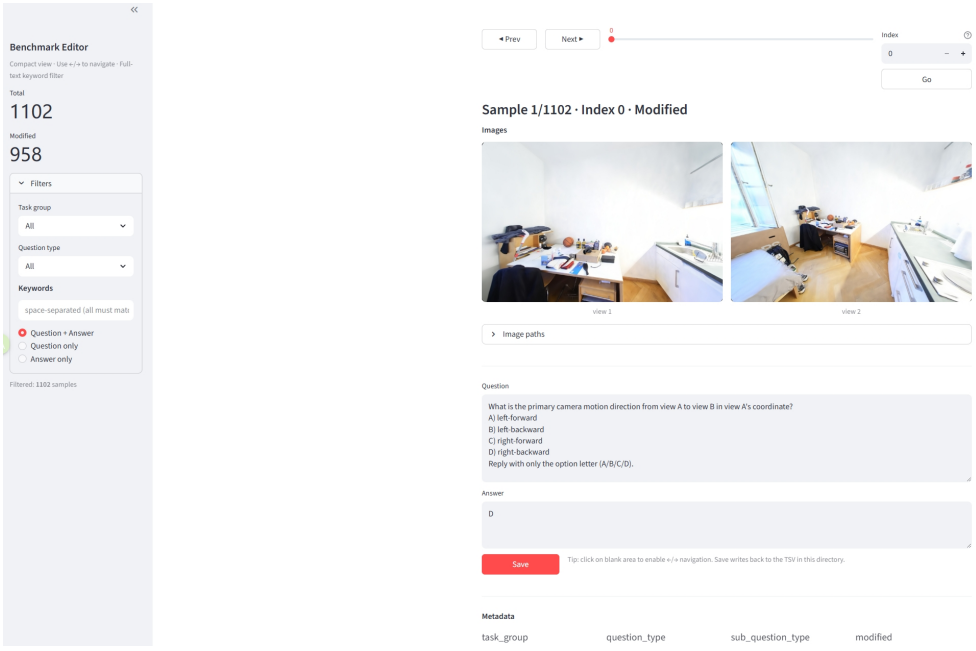


Figure 10: Human review interface for SpaceDG-Bench.

F Question Templates

In this section, we provide detailed tamplates of 11 question types, covering camera-centric, object-centric and camera-object questions.

F.1 CAMERA TRANSLATION (QUESTION_TYPE=CAMERA_TRANSLATION)

- **MCQ (main direction).** What is the primary camera motion direction from view A to view B in view A's coordinate?
  <options: A/B/C/D>
  Reply with only the option letter (A/B/C/D).
- **Numeric (meters).** What is the camera translation distance from view A to view B (meters)?
- **Binary (threshold).** Based on the images, decide whether the camera translation from view A to view B exceeds <THRESHOLD> meters. Return ONLY one token: 'yes' or 'no'. Output format: <answer>yes</answer> or <answer>no</answer> (no extra text).

F.2 CAMERA ROTATION (QUESTION_TYPE=CAMERA_ROTATION)

- **MCQ (single-axis dominant).** Given view A and view B, consider the relative rotation from A to B expressed in view A's camera frame. Which SINGLE rotation direction is the most prominent?
  (This question focuses on the dominant axis among <axis1> and <axis2>.)
  <options: A/B/C/D>
- **MCQ (two-axis).** Given view A and view B, consider the relative rotation from A to B expressed in view A's camera frame. Which option best describes the rotation direction using TWO components: <axis1> and <axis2>?
  <options: A/B/C/D>

F.3 CAMERA-OBJECT RELATIVE DISTANCE (QUESTION_TYPE=CAMERA_OBJECT_DISTANCE_ESTIMATION)

- What is the straight-line distance to the {target_obj} from the camera in meters?
- How far is the {target_obj} from the current viewpoint?
- Estimate the physical distance between the camera and the {target_obj}.
- Can you estimate the straight-line distance to the {target_obj} from the current viewpoint?
- **Two-view marked target.** <marker text about the target in image A>
  Locate the same physical object in image B. Estimate the 3D metric distance (in meters) from the camera position of image B (camera center) to the "<label>" (to the object surface/center point).
  This is NOT pixel distance. Return only one number in meters (e.g., 0.7). Output format: <answer>NUMBER</answer>.

F.4 CAMERA-OBJECT RELATIVE DIRECTION (QUESTION_TYPE=CAMERA_OBJECT_RELATIVE_DIRECTION)

- What is the relative position of the {target_obj_B} with respect to the {target_obj_A} in this view?
- Where is the {target_obj_B} located relative to the {target_obj_A}?
- Describe the spatial relationship between the {target_obj_B} and the {target_obj_A}.
- In which direction is the {target_obj_B} compared to the {target_obj_A}?
- **Two-view direction MCQ.** <marker text about the target in image A>
  Which direction is the "<label>" relative to you when taking image B?
  <options: A/B/C/D>
- **Two-view direction MCQ.** <marker text about the target in image A>
  When you were taking the photo in Image B, where is the <label> area relative to you?
  <options: A/B/C/D>

- **Two-view relpos MCQ.** <marker text for A/B/C across image A and/or B>
  You are positioned at <labelA> and face <labelB>. In which direction is <labelC> relative to you?
  <options: A/B/C/D>

F.5 CAMERA-OBJECT CARDINAL DIRECTION (QUESTION_TYPE=CROSS_VIEW_CARDINAL_DIRECTION)

- **Two-view assumed direction.** <marker text for object 1 in image A and object 2 in image B>
  The direction of <label1> relative to image A is <assumed_dir>. What is the direction of <label2> relative to image B?
  <options: A/B/C/D>

F.6 OBJECT-OBJECT DISTANCE (QUESTION_TYPE=INTER_OBJECT_DISTANCE)

- **Two-view numeric.** <marker text for object 1 in image A and object 2 in image B>
  Estimate the 3D metric distance (in meters) between the centers of these two physical objects.
  Return only one number in meters (e.g., 1.2). Output format: <answer>NUMBER</answer>.

F.7 OBJECT-OBJECT CARDINAL DIRECTION (QUESTION_TYPE=OBJECT_PROXY_CARDINAL_DIRECTION)

- **Proxy-frame MCQ.** <marker text for A/B/C across image A and/or B>
  The direction of <labelA> relative to <labelB> is <assumed_dir>. What is the direction of <labelC> relative to <labelB>?
  <options: A/B/C/D>

F.8 OBJECT SIZE COMPARISON (QUESTION_TYPE=OBJECT_SIZE_COMPARISON)

- **MCQ (length).** <marker text for object 1 in image A and object 2 in image B>
  Which is longer (consider the longest side of the object)? <label1> or <label2>?
  <options: A/B/C/D>
- **MCQ (height).** <marker text for object 1 in image A and object 2 in image B>
  Which object is taller (consider the top of the objects)? <label1> or <label2>?
  <options: A/B/C/D>

F.9 OBJECT BOUNDING-BOX SIZE ESTIMATION (QUESTION_TYPE=OBJECT_BOUNDING_SIZE_ESTIMATION)

- What are the 3D physical dimensions of the {target_obj}? Please answer in the format [shortest edge, middle edge, longest edge] in meters (e.g. [0.10, 0.20, 0.30]).
- Estimate the physical size of the {target_obj} in meters, and respond as [shortest edge, middle edge, longest edge] (e.g. [0.10, 0.20, 0.30]).
- Could you provide the three edge lengths of the {target_obj} in the format [shortest, middle, longest] (meters, e.g. [0.10, 0.20, 0.30])?
- What is the bounding box extent of the {target_obj}? Reply as [shortest edge, middle edge, longest edge] in meters (e.g. [0.10, 0.20, 0.30]).

F.10 Object Existence Estimation (question_type=object_existence_estimation)

- Is there a {target_obj} visible in this image?
- Can you find the {target_obj} in the current view?
- Does the image contain the {target_obj}?
- Check if the {target_obj} is present in this picture.

F.11 Object Counting (question_type=object_counting)

- How many <label>s are visible in this image?
- Count the number of <label> objects in the scene.
- What is the total count of <label>s shown?
- Tell me how many <label>s exist in the current view.

G Additional QA Examples

As shown in Figure [11–19], we present more error examples of SpaceDG and the reasoning process generated by Gemini-3.1-Flash-Lite. For each example, we provide the complete question, ground truth, model's clean condition answer, and the reasoning process. We label the error type if the model responses a wrong answer.

H Limitations

Despite its large-scale and physically grounded design, SpaceDG has several limitations. First, since the current dataset is built upon ScanNet++, it is primarily restricted to indoor environments. Second, to preserve the physical realism and geometric consistency of synthesized observations, our degradation engine currently supports only nine representative degradation types, leaving other real-world degradations, such as rain, snow, lens flare, rolling-shutter artifacts, and more complex compound corruptions, for future exploration. Nevertheless, SpaceDG provides a systematic and controllable framework for studying degradation-aware spatial intelligence, enabling reliable evaluation and training of MLLMs under realistic visual degradations with accurate 3D spatial ground truth.

Distortion, Cam.-Obj. Distance



Question: In image A, the object described as "black modern chair with angular star-shaped base, positioned at desk" is the chair. Locate the same physical object in image B. Estimate the 3D metric distance (in meters) from the camera position of image B (camera center) to the "chair" (to the object surface point). This is NOT pixel distance to the image center. Return only one number in meters (e.g., 0.7). Output format: <answer>NUMBER</answer>
Ground Truth: 2.4
Model clean condition answer : 2.5
DG Reasoning: The image exhibits a strong fisheye effect, which is a form of geometric distortion.\n<degradation>distortion</degradation>. In image B, the chair is positioned at the desk in the corner of the room. Given the image B is close to the chair, the distance from the camera (which is positioned near the center of the room) to the chair is approximately 1.5 meters.

Spatial Relation Distortion

Figure 11: Complete QA example of distortion.

Defocus, Cam.-Rotation



Question: Given view A and view B, consider the relative rotation from A to B expressed in view A's camera frame. Which SINGLE rotation direction is the most prominent?

A. turn right B. turn left
C. clockwise roll D. counter-clockwise roll

Ground Truth: A. turn right

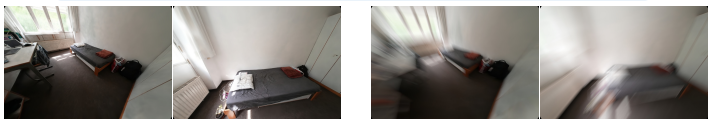
Model clean condition answer: A

DG Reasoning: The image exhibits a significant lack of focus, resulting in blurred edges and details throughout the scene.\n<degradation>defocus</degradation>. Comparing view A and view B, the camera shifts its perspective from a direct view of the shelving unit in A to a view where the shelving unit is pushed toward the right side of the frame in B. This lateral movement indicates that the camera has rotated to the left.

Spatial Relation Distortion + Low-visibility Guessing

Figure 12: Complete QA example of defocus.

Motion Blur, Obj. Size Comparison



Question: In image A, the object described as “light grey rectangular pillow with subtle horizontal stitching pattern” is the pillow. In image B, the object described as “dark grey fabric bed with brown pillow, positioned next to white wardrobe with wooden trim” is the bed. Which object is taller (consider the top of the objects)? pillow or bed?

A. Pillow B. bed C. Cannot be determined. D. Both are about the same height.

Ground Truth: A. pillow

Model clean condition answer : A

DG Reasoning: The image exhibits a strong radial blur effect, characteristic of motion blur. <degradation> motion_blur </degradation>. In the provided images, the bed is a piece of furniture that stands significantly higher off the ground than the pillow, which is resting on top of the bed. Therefore, the bed is taller.

Low-visibility Guessing

Figure 13: Complete QA example of motion blur.

Water Droplets, Cam.-Obj. Relative Direction



Question: In image B, the object described as "blue and black soft-sided bag, featuring a visible zipper and external pocket, resting on a table surface" is the bag. In image B, the object described as "wooden door with vertical grain texture, rectangular glass window, and metallic handle on right side" is the door. In image A, the object described as "rectangular window with dark grey frame, featuring visible red markings along its edges, set within a white wall" is the window. You are positioned at bag and face door. In which direction is window relative to you?
A. Back-Left B. Front-Right C. Right D. Front-Left"
Ground Truth: A. (Back-Left)
Model clean condition answer : A
DG Reasoning: The image displays numerous small, spherical droplets covering the surface, which is characteristic of water droplets. <degradation> water_droplets </degradation>. In image B, the bag is on the left side of the frame, and the wooden door is on the left as well, but the perspective shows the door is further back. In image A, the window is located on the right side of the frame. Since you are at the bag facing the door, the window is positioned to your right.

Spatial Relation Distortion + Low-visibility Guessing

Figure 14: Complete QA example of water droplets.

Haze, Camera Translation



Question: Based on the images, decide whether the camera translation from view A to view B exceeds 0.83 meters. Compute the translation distance as the Euclidean norm of the translation vector between the two camera centers (in meters). Return ONLY one token: 'yes' or 'no'. Outputformat: <answer>yes</answer> or <answer>no</answer>.
Ground Truth: no
Model clean condition answer : no
DG Reasoning: The image exhibits a significant loss of contrast and a washed-out appearance, which is characteristic of overexposure. <degradation> over_exposure </degradation>. The camera moves from a position closer to the kitchen counter (view A) to a position further back and slightly to the right (view B). Given the scale of the kitchen furniture and the change in perspective, the translation distance is clearly greater than 0.83 meters.

Degradation Attribution Error

Figure 15: Complete QA example of haze.

Low Light, Existence




Question: Does the image contain the white ceiling with rectangular fluorescent light fixture and circular smoke detector, smooth surface with slight irregular edge contour? Reply with only the option letter (A/B).

A) Yes B) No

Ground Truth: B

Model clean condition answer : B

DG Reasoning: The image is extremely dark, making it difficult to discern details, which is characteristic of a low-light condition. <degradation> low_light </degradation>. Despite the severe underexposure, the faint outlines of a ceiling structure, a rectangular light fixture, and a circular smoke detector can be identified in the upper portion of the frame. These features match the description provided.

Low-visibility Guessing

Figure 16: Complete QA example of low light.

Over Exposure, Cam.-Obj. Cardinal Direction




Question: In image A, the object described as "cardboard box with red outline, rectangular shape, flat panels, reinforced edges, standing upright, no visible text or logos" is the box. In image B, the object described as "white grid-patterned paper with rectangular cells, mounted on a surface, likely a planning or scheduling board" is the paper. The direction of box relative to image A is north. What is the direction of paper relative to image B?

A. North-East B. South-West C. East D. South

Ground Truth: B. South-West

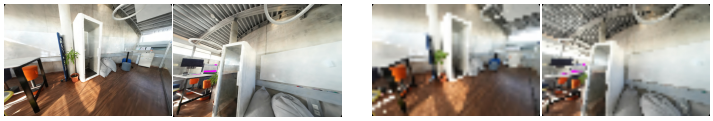
Model clean condition answer : B

DG Reasoning: The image is heavily washed out with bright light, indicating an over-exposure issue. <degradation> over_exposure </degradation>. In image B, the white grid-patterned paper is mounted on the wall, positioned towards the upper-left area of the frame relative to the door handle and the foreground object. Given the perspective of the room, this location corresponds to a North-West or North-East orientation depending on the specific framing, but among the provided choices, it is located in the upper portion of the wall, in North-East of the image B.

Spatial Relation Distortion

Figure 17: Complete QA example of over exposure.

Low Res., Obj. Cardinal Direction



Question: In image A, the object described as "light grey fabric beanbag next to another plain grey beanbag" is the pillow. In image B, the object described as "light grey fabric beanbag next to another plain grey beanbag" is the pillow. In image B, the object described as "white glass door featuring a vertical handle and transparent panel" is the door. In image A, the object described as "white rectangular pape placed on a shelf, featuring printed" is the paper. The direction of pillow relative to door is east. What is the direction of paper relative to door?
A. North B. North-West C. East D. West
Ground Truth: C. East
Model clean condition answer : C
DG Reasoning: The image exhibits a significant lack of sharpness across the entire scene, characteristic of a defocus effect. <degradation>defocus</degradation>. In the provided images, the door is the prominent white structure in the center. The paper is located on a shelf to the left of this door structure. Therefore, relative to the door, the paper is positioned to the west.

Artifact-induced + Degradation Attribution Error

Figure 18: Complete QA example of low resolution.

JPEG, Counting



Question: What is the total count of jackets shown? Answer with only a single integer number.
Ground Truth: 1
Model clean condition answer : 1
DG Reasoning: The image exhibits significant blocky artifacts and color banding characteristic of low-quality image compression. <degradation> jpeg_compression </degradation>. Scanning the room, there is a black jacket draped over the back of the office chair and another black jacket or piece of clothing resting on the chair/desk area near the guitar. Counting these items, there are two jackets visible.

Artifact-induced Error

Figure 19: Complete QA example of JPEG compression.