ROBOSTRESSBENCH: A DIAGNOSTIC BENCHMARK FOR VISUAL STRESS IN EMBODIED SCENE UNDERSTANDING

Anonymous authors
Paper under double-blind review

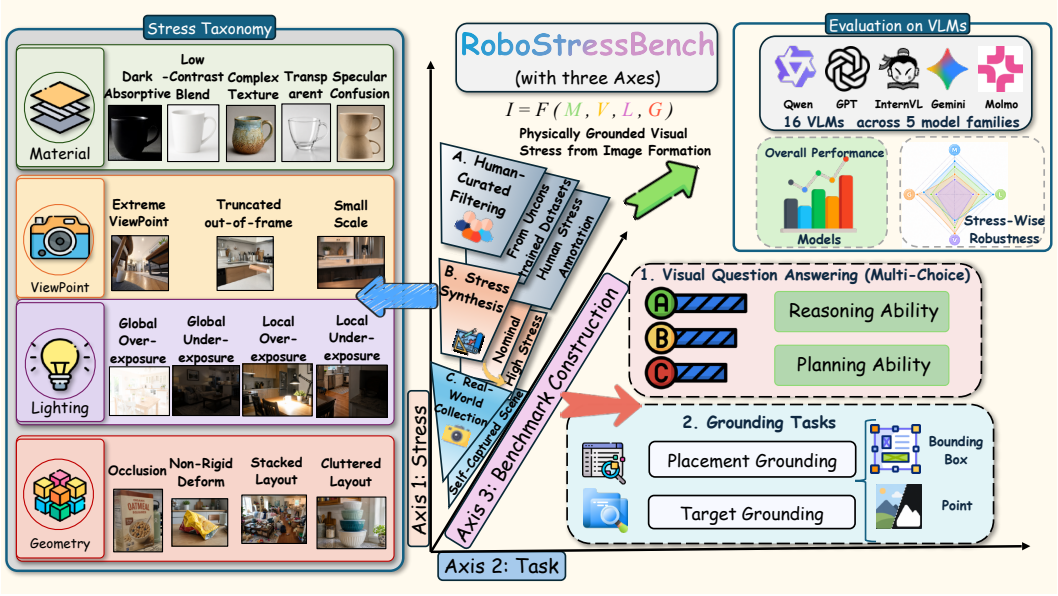

Figure 1: **Overview of RoboStressBench.** RoboStressBench evaluates VLM robustness under physical visual stress in embodied scenes. We organize visual stress according to four image-formation factors: Material, Viewpoint, Lighting, and Geometry. The benchmark is constructed from human-curated filtering, stress synthesis, and real-world collection, and supports task-aligned evaluation through multiple-choice visual question answering and grounding tasks. We further evaluate diverse VLM families to analyze overall performance and stress-wise robustness.

ABSTRACT

Vision-language models (VLMs) used in embodied systems must interpret scenes under challenging visual conditions, yet it remains difficult to distinguish failures caused by visual stress from those caused by the underlying task or scene. We introduce **RoboStressBench**, a diagnostic benchmark for evaluating VLM robustness to visually challenging conditions in robot-centric scenes. An image-formation-inspired taxonomy organizes visual stress into four dimensions—**Material**, **Viewpoint**, **Lighting**, and **Geometry**—and 16 fine-grained categories, with multi-label annotations capturing co-occurring visual conditions. RoboStressBench contains **7,183 examples** collected through naturally occurring stress cases, targeted stress synthesis, and additional real-world collection, and evaluates complementary capabilities including target and placement grounding, spatial reasoning, state understanding, and action-selection-style question answering. To complement broad stress coverage, we evaluate matched nominal–stressed scene–question pairs, measuring within-instance performance changes before and after targeted stress edits while holding the question and task definition fixed. Evaluating **16 VLMs** reveals task-dependent stress profiles that are obscured by aggregate scores, while matched comparisons show substantial performance

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.

degradation after targeted stress editing for several model groups. We additionally introduce **StressDART**, a test-time baseline that uses predicted stress categories to guide image rectification before reasoning. With Qwen3-VL-4B as the reasoner, StressDART improves performance without updating model parameters. Together, these evaluations provide a structured testbed for characterizing visual robustness and studying test-time interventions for embodied-relevant visual understanding.

1 INTRODUCTION

Vision-Language Models (VLMs) are increasingly used as visual reasoning components in embodied systems, supporting tasks such as object grounding, spatial understanding, state recognition, and action selection Liu et al. (2025b); Kim et al. (2024); Driess et al. (2023); Zitkovich et al. (2023). In these settings, the visual evidence available to a model is often far from ideal: objects may be transparent or reflective, partially occluded, observed from unusual viewpoints, poorly illuminated, or surrounded by clutter. Such conditions can weaken, distort, or obscure the evidence required to answer a question or localize an action-relevant region. We refer to these task-relevant visual challenges as *physical visual stress*. Robustness to such stress is therefore particularly important for embodied-relevant visual capabilities, including object grounding, spatial reasoning, state understanding, and action-related decision making.

Evaluating such robustness is challenging because low performance on a difficult image does not by itself reveal why the model failed. A model may fail because of the underlying scene or task, because task-relevant evidence is visually degraded, or because several challenging factors co-occur. General-purpose VLM benchmarks (Yu et al., 2023; Liu et al., 2024b; Li et al., 2023; Fu et al., 2023) provide broad coverage of multimodal capabilities, while recent robustness benchmarks Ishmam et al. (2025); Li et al. (2026); Saxena et al. (2026) explicitly study performance under visual degradation, resolution changes, and other distribution shifts, as summarized in Fig. 2. For robot-relevant visual understanding, our focus is on integrating three elements within a common evaluation framework: *(i)* interpretable annotations of task-relevant scene and observation conditions; *(ii)* evaluation across complementary capabilities such as grounding, spatial reasoning, and action selection; and *(iii)* matched comparisons that measure performance changes after targeted stress editing of the same scene–question instance. Together, these elements connect broad coverage of challenging observations with within-instance robustness analysis, making model performance more informative than an aggregate score alone.

To address these issues, we introduce **RoboStressBench**, a diagnostic benchmark for studying VLM robustness under visually challenging conditions in embodied scenes. Rather than treating visual difficulty as a single undifferentiated category, we use an image-formation-inspired taxonomy that organizes stress into four dimensions: **Material** (M), **Viewpoint** (V), **Lighting** (L), and **Geometry** (G), further divided into 16 fine-grained categories. These dimensions provide an interpretable vocabulary for task-relevant challenges associated with surface appearance, observation and framing, illumination, and scene structure. Multi-label annotations allow an example to represent several co-occurring visual conditions. RoboStressBench contains **7183 examples** constructed from three complementary sources: naturally occurring stress cases filtered from existing datasets, targeted stress synthesis from nominal images, and additional Internet-sourced and self-captured examples. The benchmark supports target and placement grounding, spatial reasoning, state understanding, and planning-style action selection.

RoboStressBench supports two complementary forms of analysis. Task- and stress-conditioned evaluation characterizes model performance across the heterogeneous benchmark, while matched nominal–stressed comparisons measure performance changes after targeted editing of the same task instance. The former provides broad coverage of challenging visual conditions; the latter reduces variation arising from differences in scene content and question difficulty. We evaluate **16 VLMs** spanning Qwen Qwen Team (2026a), InternVL Wang et al. (2025), Molmo Clark et al. (2026), Gemini Google DeepMind (2026), and GPT OpenAI (2026). As shown in Fig. 3, the evaluated models exhibit substantially different performance profiles across visual stress categories. Models that perform similarly in aggregate can still show markedly different failure patterns under specific visual conditions. On the matched subset, the evaluated Qwen model groups show substantial per-

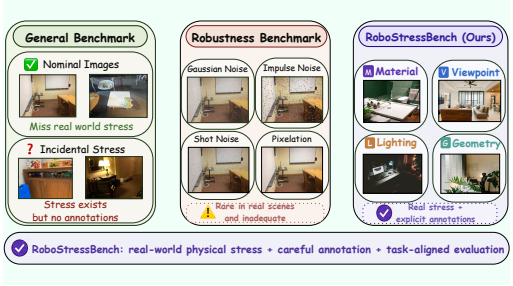

Figure 2: **Motivation for RoboStressBench.** RoboStressBench combines explicit stress annotations with robot-relevant visual tasks to support structured analysis of model performance under challenging observation conditions.

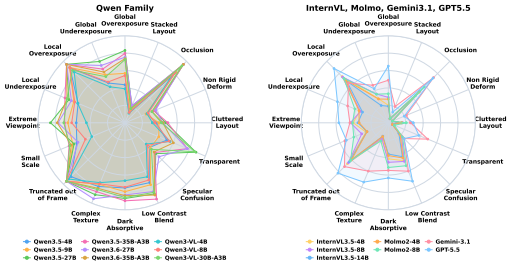

Figure 3: **RoboStressBench evaluation results.** We visualize the performance of all evaluated VLMs across RoboStressBench stress dimensions. Comprehensive numerical results are reported in Table 2.

formance drops after targeted stress editing (Table 1). These observations motivate reporting task- and stress-specific robustness in addition to a single overall score.

Beyond characterizing model failures, RoboStressBench also supports the evaluation of test-time interventions. We introduce **StressDART**, a modular baseline that uses the stress taxonomy to guide visual rectification: a VLM predicts a task-relevant stress category, an image editor produces a rectified observation, and a final reasoner receives the original and rectified images. Our experiments compare direct inference, rectified-only reasoning, and reasoning with both observations. The resulting performance gains demonstrate the utility of the evaluated rectification pipeline without updating model parameters.

In summary, our contributions are threefold:

• We introduce **RoboStressBench**, a **7,183-example** benchmark for embodied-relevant visual understanding. It combines naturally occurring, synthesized, and additionally collected stress cases across five task types, with multi-label annotations spanning four image-formation-inspired dimensions and 16 fine-grained categories.

• We combine task- and stress-conditioned evaluation on the heterogeneous benchmark with matched nominal–stressed comparisons. These complementary analyses characterize performance across diverse visual conditions and measure within-instance changes under targeted stress editing.

• We evaluate **16 VLMs** and introduce **StressDART** for stress-conditioned test-time rectification. Our experiments reveal task-dependent stress profiles and demonstrate performance gains from the evaluated rectification pipeline without model fine-tuning.

2 Related Work

Visual Robustness and Distribution Shift. Robustness evaluation in computer vision has long studied how model performance changes under perturbations and distribution shifts. Early work focused on adversarial perturbations (Szegedy et al., 2013; Goodfellow et al., 2014; Madry et al., 2017), while ImageNet-C/P (Hendrycks & Dietterich, 2019) established systematic evaluation under common non-adversarial corruptions. Subsequent work extended robustness analysis to natural distribution shifts and more realistic changes in scene appearance, including viewpoint, background, rendering style, and physically plausible corruptions (Barbu et al., 2019; Koh et al., 2021; Kar et al., 2022; Zhang et al., 2024). These benchmarks provide important tools for measuring visual robustness, but they primarily organize difficulty around corruption types or dataset shifts. RoboStressBench instead uses an image-formation-inspired taxonomy to organize task-relevant visual challenges by Material, Viewpoint, Lighting, and Geometry. We treat these dimensions as an interpretable annotation vocabulary rather than as independently controlled physical variables.

Robustness Evaluation for Multimodal Models. Recent work has extended robustness evaluation from conventional vision models to vision-language models and multimodal large language models.

The Visual Robustness Benchmark for VQA (Ishmam et al., 2025) studies multimodal question answering under realistic visual corruptions, Res-Bench (Li et al., 2026) evaluates robustness to dynamic input resolution, and VLM-RobustBench (Saxena et al., 2026) evaluates VLMs across a broad range of visual augmentations and task settings. Other benchmarks, including R-Bench (Li et al., 2025), Eva-VLA (Liu et al., 2025a), and DarkEQA (Park et al., 2025), further investigate multimodal robustness under real-world corruptions, physical variations, or challenging observation conditions. RoboStressBench complements these efforts by combining image-formation-inspired stress annotations, multiple robot-relevant visual tasks, and matched nominal–stressed evaluation. Its heterogeneous examples support task- and stress-conditioned performance analysis, while verified pairs enable within-instance comparisons under targeted stress editing. Our emphasis is on using these complementary views to characterize visual robustness in embodied-scene understanding.

Embodied Benchmarks for Vision-Language Models. Embodied VLM evaluation has expanded from situated question answering to grounding, spatial reasoning, affordance prediction, and visually guided action. OpenEQA Majumdar et al. (2024) and RoboVQA Sermanet et al. (2024) evaluate situated understanding over embodied observations and task context. RoboRefIt Lu et al. (2023) studies language-conditioned object grounding for manipulation, while RoboSpatial Song et al. (2025) and RefSpatial-Bench Zhou et al. (2025) focus on robot-centered spatial understanding and referring. More action-oriented benchmarks (Yuan et al., 2025; Hao et al., 2025; Yuan et al., 2026) evaluate decision-relevant visual outputs, and broader embodied evaluations (Chen et al., 2025; Qiu et al., 2024; Yang et al., 2025a;b) consider temporal reasoning, memory, planning, and agent-level capabilities. RoboStressBench complements these task-centric benchmarks by focusing on how visual observation conditions interact with grounding and reasoning performance in embodied scenes. Its scope is the robustness of embodied-relevant visual understanding rather than end-to-end closed-loop robot execution.

Table 1: **Matched nominal–stressed evaluation.** We compare performance before and after targeted stress editing of the same scene–question instances. $\Delta = \mathrm{Stress} - \mathrm{Nom.}$ is reported in percentage points; negative values indicate degradation.

Model group	Accuracy (%)		
	Nom.	Stress	Δ (pp)
Qwen3VL	51.0	35.5	-15.5
Qwen3.5	53.5	36.8	-16.7
Qwen3.6	64.3	40.1	-24.2
InternVL3.5	10.0	9.9	-0.1
Molmo2	12.2	11.5	-0.7

3 ROBOSTRESSBENCH

RoboStressBench evaluates VLM robustness under task-relevant visual challenges in embodied scenes. We first introduce an image-formation-inspired stress taxonomy (Sec. 3.1) and then describe benchmark construction (Sec. 3.2). Fig. 4 gives an overview of the benchmark.

3.1 STRESS TAXONOMY

Image-Formation Perspective. We motivate our taxonomy from the physical image-formation process. In physically based rendering, outgoing radiance can be written as

$$L_o(\mathbf{x}, \boldsymbol{\omega}_o) = L_e(\mathbf{x}, \boldsymbol{\omega}_o) + \int_\Omega f_s L_i \, |\mathbf{n}\cdot\boldsymbol{\omega}_i| \, d\boldsymbol{\omega}_i, \tag{1}$$

where surface scattering, illumination, viewpoint, and geometry jointly determine the visual evidence observed by a camera. Motivated by these factors, we use the conceptual abstraction $I = \mathcal{F}(M, V, L, G)$ and organize visual stress into four dimensions: **Material** (M), **Viewpoint** (V), **Lighting** (L), and **Geometry** (G). This abstraction motivates the taxonomy rather than assuming a calibrated or factorized physical model; multiple factors may co-occur in natural scenes.

RoboStressBench further divides these dimensions into 16 fine-grained categories. **Material** includes dark absorptive, low-contrast blend, complex texture, transparent, and specular confusion;

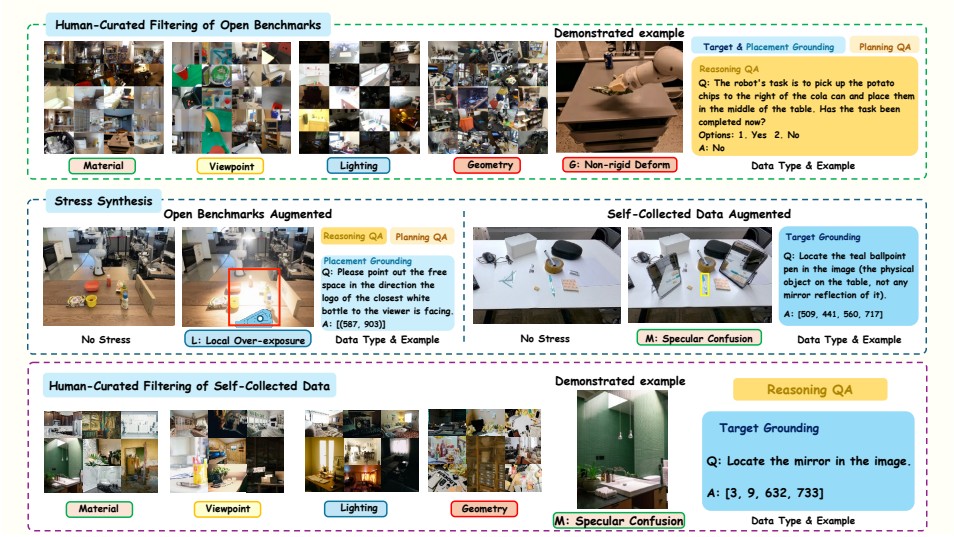

Figure 4: **Stress categories and curation pipeline.** Overview of the four stress dimensions and three data sources in RoboStressBench.

Viewpoint includes extreme viewpoint, truncated out-of-frame, and small scale; **Lighting** includes global/local overexposure and underexposure; and **Geometry** includes occlusion, non-rigid deformation, stacked layout, and cluttered layout. The benchmark supports multi-label annotation, and stress-wise results should therefore be interpreted as performance associated with annotated visual conditions rather than isolated causal effects.

3.2 Benchmark Construction

RoboStressBench contains **7,183 examples** from three complementary sources: 2,927 naturally stressed examples selected from existing robotics and spatial-reasoning datasets Du et al. (2024); Zhou et al. (2025); Tang et al. (2025); Song et al. (2025); Chen et al. (2025); Yuan et al. (2026; 2025); Lu et al. (2023), 2,596 targeted stress variants synthesized from nominal images, and 1,660 additional real images collected from Internet sources and self-captured scenes. Naturally occurring examples provide broad coverage, while targeted synthesis supports matched nominal–stressed comparisons. Prompt-based editing is used as a benchmark-construction tool rather than as calibrated physical control.

The benchmark covers five task types: *target grounding*, *placement grounding*, *spatial reasoning*, *state understanding*, and *planning-style action selection*. Original task annotations are reused when valid; otherwise, annotators create or revise the corresponding questions, answers, or grounding labels. For synthesized examples, annotations are retained only when task semantics and relevant geometry remain valid, and are otherwise re-annotated or excluded from matched analysis.

A verified subset forms matched nominal–stressed scene–question pairs for within-instance robustness evaluation. Each pair preserves the question and task definition, while the stressed observation is produced through a targeted visual edit. Performance is compared before and after editing to reduce variation due to differences in scene content and question difficulty.

Six trained annotators verify stress labels, task validity, and final annotations; ambiguous examples are revised or excluded. Further details on annotation, synthesis, and dataset statistics are provided in Appendix A.

4 StressDART: A Test-Time Rectification Baseline

RoboStressBench provides a setting for evaluating whether test-time visual processing can improve performance under visual stress. We introduce **StressDART** (**D**etection **A**nd **R**ectification at **T**est

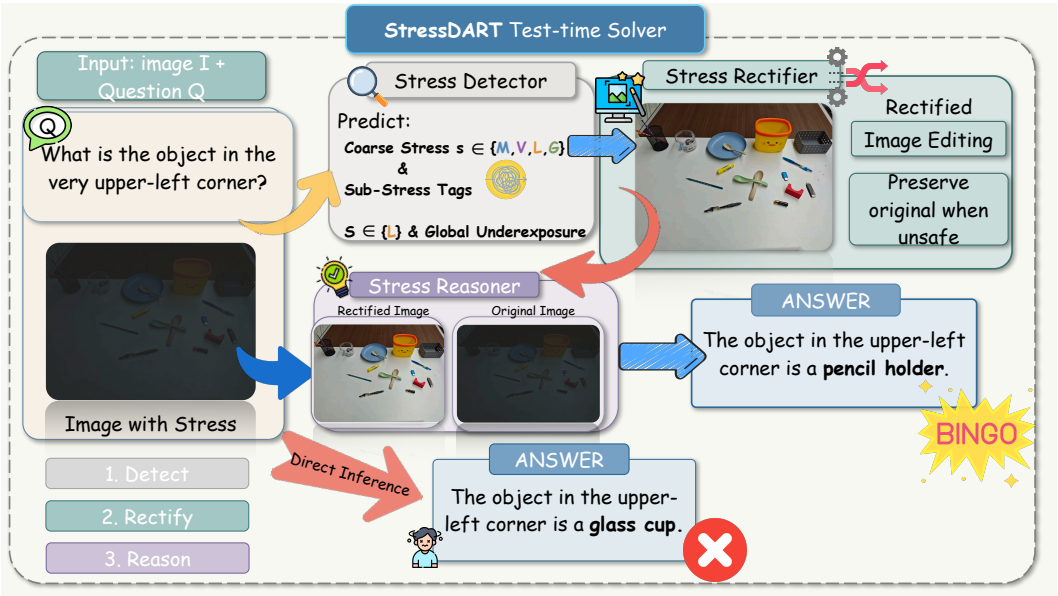

Figure 5: **Overview of StressDART.** Given an image and task query, StressDART first predicts a task-relevant visual stress category and uses it to condition an image-editing operation. The final reasoner then receives the original and rectified observations to produce an answer. StressDART is used as a test-time intervention baseline for analyzing whether stress-conditioned visual processing can improve robustness and when editing may alter task-relevant evidence.

time), a modular baseline that uses the benchmark taxonomy to condition image rectification before VLM reasoning. As illustrated in Fig. 5, the pipeline consists of stress prediction, stress-conditioned rectification, and final reasoning. All components are used without parameter updates.

Stress Prediction. Given an input image I and task query Q, a VLM-based detector predicts a coarse stress dimension and a fine-grained stress category from the taxonomy in Sec. 3.1. The prediction is conditioned on both the image and the task query because a visual condition may be relevant to one task while having little effect on another. We denote the resulting prediction by (s, c), where $s \in \{M, V, L, G\}$ is the coarse category and c is a fine-grained stress label. The baseline uses one predicted category to select the rectification instruction, even when the input contains multiple co-occurring visual stresses.

Stress-Conditioned Rectification. The predicted stress category is then used to construct a category-specific image-editing instruction. A visual editor processes the original image to produce a rectified observation $\tilde{I}$ intended to make task-relevant evidence more accessible. For example, the rectification instruction may emphasize poorly visible content under underexposure, reduce the effect of excessive brightness, or make a small task-relevant region easier to inspect. We implement rectification with Qwen-Image-Edit Wu et al. (2025). The edited observation is intended to make task-relevant visual cues easier to use, but generative editing can also modify those cues. We therefore treat rectification as an input transformation rather than as a guaranteed reconstruction of the original scene.

Reasoning with Original and Rectified Observations. The final VLM reasoner receives the original image I, the rectified image $\tilde{I}$, and the original task query Q, and produces the final prediction. Providing the original observation alongside the edited one gives the reasoner access to the unmodified visual evidence while allowing it to use potentially helpful information from the rectified view. We additionally evaluate a rectified-only variant to assess whether also providing the original observation improves the final score.

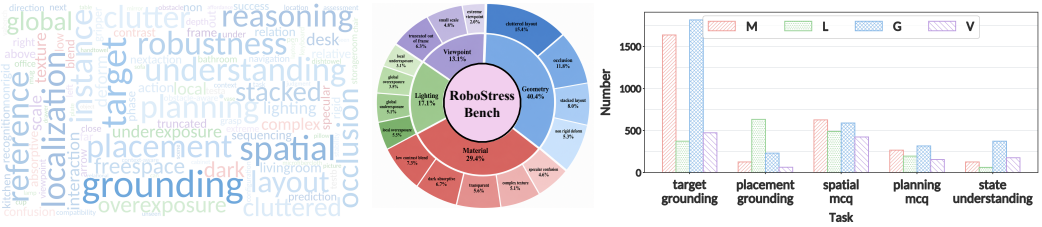

Figure 6: Overview of **RoboStressBench**'s statistical distributions. (***Left***) Word distribution of prompt suites; (***Middle***) Data distribution across 16 sub-stress types; and (***Right***) Data distribution across different tasks.

Our evaluation compares original-only inference, rectified-only reasoning, and reasoning with both observations using the same final reasoner. This comparison measures the performance of the complete rectification pipeline and the effect of additionally supplying the original visual input.

5 Experiments

We evaluate 16 VLMs on RoboStressBench, analyze task- and stress-conditioned performance, examine matched nominal–stressed pairs, and study StressDART as a test-time visual intervention.

5.1 Experimental Settings

Evaluation Protocol. For multiple-choice tasks, we report exact-match accuracy. For point grounding, a prediction is correct if it falls inside the target mask or valid placement region. For box grounding, we report mean localization accuracy (**mAcc**) over IoU thresholds from 0.50 to 0.95 with a step of 0.05. The grounding score in Table 2 is $\text{Grounding} = (\text{PointAcc} + \text{BoxmAcc})/2$. Detailed grounding metrics are reported in Table 4. Because stress annotations are multi-label, stress-wise scores are interpreted as conditional performance rather than isolated causal effects.

Models and Implementation. We evaluate Qwen3-VL Bai et al. (2025), Qwen3.5 Qwen Team (2026a), Qwen3.6 Qwen Team (2026b), InternVL3.5 Wang et al. (2025), Molmo2 Clark et al. (2026), Gemini-3.1 Google DeepMind (2026), and GPT-5.5 OpenAI (2026), covering 16 model variants in total. We use official inference pipelines with deterministic decoding and model-specific input/output adapters. For StressDART, Qwen3-VL-4B serves as the stress predictor and reasoner, and Qwen-Image-Edit Wu et al. (2025) performs visual rectification.

5.2 Overall Benchmark Results

Table 2 shows substantial variation across both models and tasks. Qwen3.5-27B achieves the highest grounding score (68.4%), GPT-5.5 performs best on spatial reasoning (80.3%), and Qwen3-VL-30B-A3B obtains the strongest planning result. Model size is also not uniformly predictive of performance: larger Qwen checkpoints often improve, whereas InternVL3.5 shows a non-monotonic trend. These results motivate task- and stress-specific analysis rather than relying only on aggregate performance.

5.3 Stress- and Task-wise Analysis

We next examine how model performance varies across the annotated stress dimensions and evaluation tasks. Fig. 8 summarizes category-conditioned performance for representative models. Because the underlying examples differ across stress categories and may carry multiple stress labels, this analysis characterizes performance associations rather than controlled causal effects.

The relative difficulty of the four stress dimensions varies substantially across task types. For grounding-oriented tasks, examples annotated with Geometry stress are often among the most challenging, particularly when occlusion, clutter, or complex object arrangements reduce the visibility

Model	Size	Overall	Stress Dimensions															Task Dimensions			
			Material					Viewpoint			Geometry				Lighting				Grounding	Reasoning		Planning
			Dark	L-Con	C-Tex	Tran.	Spec.	Extr.	Trun.	Small	Occl.	Non-R	Stack	Clust	G-Ovr	L-Ovr	G-Und	L-Und	Ground.	Spa.	Sta.	Plan
Qwen3VL	4B	57.0	50.6	49.8	53.7	44.0	32.4	38.1	62.0	36.6	53.4	27.9	16.8	30.6	34.5	57.7	38.4	45.6	59.4	65.2	49.4	53.8
Qwen3VL	8B	64.1	58.9	57.6	59.9	49.0	38.4	52.4	69.8	45.8	62.8	33.3	21.5	36.9	48.6	66.6	46.6	53.0	60.1	73.4	58.8	64.2
Qwen3VL	30B-A3B	**75.5**	**64.7**	65.0	63.3	**65.2**	42.2	57.1	**70.8**	**56.2**	67.9	36.3	25.2	41.6	58.2	64.5	50.3	**58.0**	61.8	71.9	**68.6**	**99.8**
Qwen3.5	4B	65.2	59.4	59.1	61.5	47.3	40.4	49.5	65.6	51.6	61.9	32.3	22.4	37.9	42.6	66.4	45.9	50.8	59.9	72.6	59.6	68.8
Qwen3.5	9B	65.3	61.5	60.4	60.3	41.6	40.8	54.7	69.5	54.2	63.1	31.4	23.4	39.7	50.0	66.6	51.6	53.6	62.9	73.7	58.9	65.5
Qwen3.5	27B	71.5	**65.3**	**66.0**	**66.4**	**65.9**	**46.3**	**64.6**	**73.2**	53.4	**68.9**	**38.2**	**27.6**	**44.4**	**63.5**	**71.3**	**60.1**	**56.7**	**68.4**	77.1	**63.3**	77.0
Qwen3.5	35B-A3B	**74.9**	**66.5**	**69.0**	**64.6**	**60.7**	**45.3**	56.6	**73.2**	**56.7**	**69.1**	**38.0**	27.1	**45.8**	**61.8**	**71.0**	**55.3**	**59.2**	**64.7**	**79.0**	62.9	**92.9**
Qwen3.6	27B	70.5	63.3	64.5	**68.8**	**60.7**	**49.1**	**60.4**	**73.2**	50.8	68.7	**39.2**	**28.3**	**42.5**	**59.6**	**70.6**	**57.6**	55.2	**67.9**	**78.3**	**67.1**	68.8
Qwen3.6	35B-A3B	**72.7**	63.7	**66.6**	62.6	51.8	44.6	**60.8**	**74.1**	**55.2**	**70.1**	36.4	25.4	**42.5**	58.8	**71.3**	53.4	**58.0**	64.3	78.1	**67.1**	**81.1**
InternVL3.5	4B	43.9	41.2	44.5	27.6	13.4	28.3	40.6	55.0	32.7	45.5	9.5	11.7	23.2	27.2	59.5	34.2	37.6	23.1	65.2	41.1	46.2
InternVL3.5	8B	46.0	43.6	45.1	26.5	13.6	28.1	42.5	54.3	36.1	46.9	8.6	12.4	24.3	33.5	59.1	37.8	37.0	24.1	67.9	41.5	50.3
InternVL3.5	14B	42.2	37.8	41.7	24.5	11.8	24.8	37.3	53.4	31.9	42.7	9.5	12.0	19.8	26.4	55.8	28.2	35.4	15.1	67.6	40.0	45.9
Molmo2	4B	41.8	39.0	42.6	24.9	13.6	26.5	36.3	51.6	31.7	46.1	9.7	13.3	23.6	32.1	56.7	33.8	38.2	19.6	62.9	40.3	44.5
Molmo2	8B	45.6	47.0	48.1	29.0	16.9	29.5	41.0	52.5	42.1	50.8	10.9	14.7	27.5	36.3	58.1	37.6	41.4	25.4	63.6	39.2	54.3
Gemini-3.1	–	54.9	48.8	51.8	51.8	45.7	31.1	42.9	57.5	48.5	58.7	28.8	**27.3**	31.1	45.6	54.1	44.3	46.6	51.8	70.0	41.6	56.3
GPT-5.5	–	59.3	53.5	58.3	58.8	39.1	31.1	50.0	62.8	47.4	58.3	26.9	21.8	33.1	54.4	65.9	41.8	53.4	60.8	**80.3**	38.9	57.0

Table 2: **Evaluation results on RoboStressBench.** We report performance across 16 stress categories and four task dimensions. Overall is the macro-average of the four capability scores. Grounding averages Point-acc and box mAcc (IoU thresholds 0.50–0.95). The **best**, **second-best**, and **third-best** results in each column are highlighted.

of task-relevant regions. A similar pattern appears in spatial reasoning, where many models obtain lower scores on Geometry-annotated examples than on Material, Viewpoint, or Lighting subsets. In contrast, planning-style questions exhibit a different profile: several models remain comparatively stronger on Geometry examples but degrade more on Material or Viewpoint conditions. State Understanding also shows a distinct pattern, with some models exhibiting noticeable weakness under Lighting-annotated conditions.

These results indicate that visual stress interacts differently with different embodied-relevant capabilities. Models with similar aggregate performance can exhibit markedly different stress-wise profiles, motivating task- and stress-conditioned reporting in addition to a single benchmark score. At the same time, the category subsets differ in scene content, source, and co-occurring stress factors, so these observations should not be interpreted as isolated causal effects of Material, Viewpoint, Lighting, or Geometry. We therefore complement this analysis with matched nominal–stressed evaluation below.

5.4 Matched Nominal–Stressed Analysis

Category-conditioned results on the full benchmark provide broad coverage, but differences across subsets may also reflect scene and task difficulty. We therefore complement this analysis with matched nominal–stressed pairs, where the same scene–question instance is evaluated before and after targeted stress editing.

Table 1 reports family-level performance on the matched subset. Qwen3-VL, Qwen3.5, and Qwen3.6 exhibit accuracy drops of 15.5, 16.7, and 24.2 percentage points, respectively. These paired comparisons provide more direct evidence of stress-induced performance degradation because the underlying task instance is held fixed.

InternVL3.5 and Molmo2 have low nominal scores on this subset and show small absolute nominal–stressed differences. These differences should be interpreted together with their nominal performance rather than as evidence of greater robustness. Their small nominal–stress gaps should therefore not be interpreted as evidence of greater robustness.

5.5 StressDART Results

We finally evaluate whether test-time visual rectification can improve performance on RoboStressBench. Using Qwen3-VL-4B as the base reasoner, we compare direct inference on the original image with reasoning from the rectified image and from both the original and rectified observations.

As shown in Table 3, rectified-only reasoning increases the benchmark score from 57.0 to 63.9, a gain of 6.9 points over direct inference.

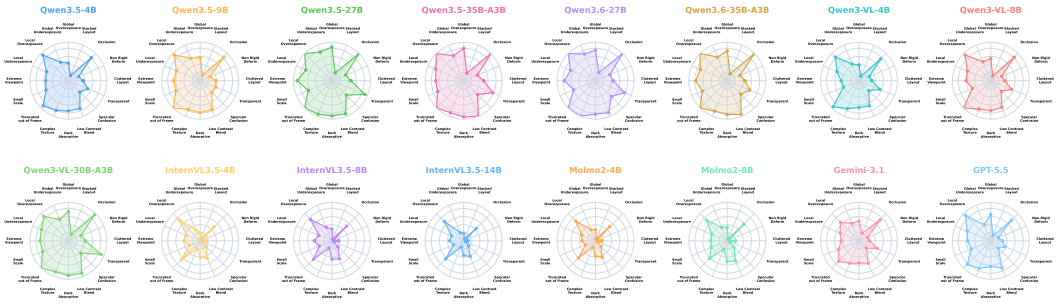

Figure 7: **Per-model dimension profiles on RoboStressBench.** Each panel shows one model's scores over the 16 dimensions; see Table 2 for the raw numbers.

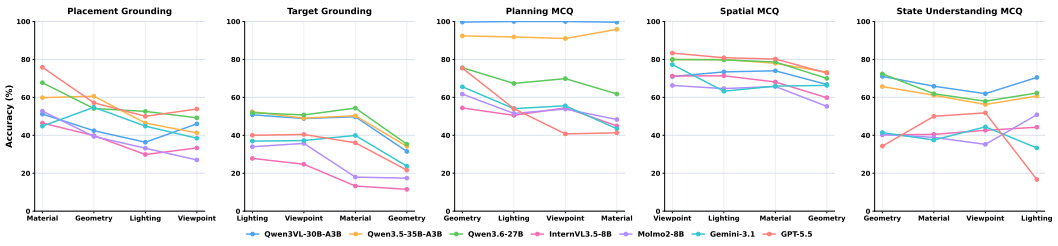

Figure 8: **Task- and stress-conditioned performance.** Model performance across Material, Viewpoint, Lighting, and Geometry annotations.

Providing both the original and rectified observations yields 64.0, corresponding to a 7.0-point improvement. The two rectification variants differ by only 0.1 points in aggregate, although this comparison does not establish that they make the same predictions on individual examples.

These results demonstrate performance gains from the complete StressDART pipeline with Qwen3-VL-4B, without model fine-tuning. The comparison evaluates the utility of the rectified observation and the choice of inputs to the final reasoner; it does not isolate the incremental benefit of stress prediction over generic image editing.

Table 3: **Effect of reasoner inputs in StressDART.** We compare direct inference, rectified-only reasoning, and reasoning with both observations using Qwen3-VL-4B as the final reasoner.

Method	Reasoner Input	Acc.
Qwen3-VL-4B	Original	57.0%
StressDART	Rectified only	63.9%
StressDART	Original + Rectified	64.0%

6 Conclusion

We introduced **RoboStressBench**, a **7,183-example** benchmark for evaluating VLM robustness under task-relevant visual stress in embodied-scene understanding. Its image-formation-inspired taxonomy organizes visual challenges into four dimensions and 16 fine-grained categories, supporting multi-label analysis across grounding, spatial reasoning, state understanding, and planning-style action selection.

Our evaluation of **16 VLMs** reveals task-dependent stress profiles that are obscured by aggregate scores. Matched nominal–stressed comparisons complement this broad coverage by measuring performance changes after targeted editing of the same task instances. We further introduced **StressDART**, demonstrating that the evaluated test-time rectification pipeline improves performance without parameter updates. Together, these results highlight the value of combining stress-conditioned reporting, within-instance comparison, and intervention evaluation to study visual robustness in embodied-relevant VLM capabilities.

AI ASSISTANCE STATEMENT

In this work, we used generative AI tools to draft sections of the paper, to aid or polish writing for readability, and for retrieval and discovery of related work. Separately, multiple VLMs are evaluated as experimental subjects in our evaluation protocol and are described in the paper; their use is part of the studied system rather than undisclosed author assistance. We have reviewed all AI-assisted text and literature suggestions: authors verified technical claims, notation, equations, and experimental descriptions against the methods and results, and checked suggested related work against the cited sources. The authors take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

Kaiyuan Chen, Shuangyu Xie, Zehan Ma, Pannag R. Sanketi, and Ken Goldberg. Robo2vlm: Improving visual question answering using large-scale robot manipulation data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 346–355, 2024.

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

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Google DeepMind. Gemini 3.1 pro model card, 2026. URL `https://deepmind.google/models/model-cards/gemini-3-1-pro/`. Accessed: 2026-04-30.

Google Gemini. Nano Banana Pro: Gemini ai image generator & photo editor. `https://gemini.google/us/overview/image-generation/`, 2026. Accessed: 2026-05-07.

Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang, Yanbiao Ma, Yunfeng Diao, Ziyu Jia, Wenbo Ding, Hangjun Ye, and Long Chen. Roboafford++: A generative ai-enhanced dataset for multimodal affordance learning in robotic manipulation and navigation. *arXiv preprint arXiv:2511.12436*, 2025.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Farhan Ishmam, Ishmam Tashdeed, Talukder Asir Saadat, Hamjajul Ashmafee, Abu Raihan Mostofa Kamal, and Azam Hossain. Visual robustness benchmark for visual question answering (vqa). In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 6623–6633, 2025.

Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18963–18974, 2022.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

Chenxu Li, Zhicai Wang, Yuan Sheng, Xingyu Zhu, Yanbin Hao, and Xiang Wang. Res-bench: Benchmarking the robustness of multimodal large language models to dynamic resolution input. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31545–31553, 2026.

Chunyi Li, Jianbo Zhang, Zicheng Zhang, Haoning Wu, Yuan Tian, Wei Sun, Guo Lu, Xiongkuo Min, Xiaohong Liu, Weisi Lin, et al. R-bench: Are your large multimodal model robust to real-world corruptions? *IEEE Journal of Selected Topics in Signal Processing*, 2025.

Hanqing Liu, Shouwei Ruan, Jiahuan Long, Junqi Wu, Jiacheng Hou, Huili Tang, Tingsong Jiang, Weien Zhou, and Wen Yao. Eva-vla: Evaluating vision-language-action models' robustness under real-world physical variations. *arXiv preprint arXiv:2509.18953*, 2025a.

Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu, et al. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. *arXiv preprint arXiv:2503.10631*, 2025b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Yuhao Lu, Yixuan Fan, Beixing Deng, Fangfu Liu, Yali Li, and Shengjin Wang. Vl-grasp: A 6-dof interactive grasp policy for language-oriented objects in cluttered indoor scenes. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 976–983, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16488–16498, 2024.

OpenAI. GPT-5.5 system card, 2026. URL https://openai.com/index/gpt-5-5-system-card/. Accessed: 2026-04-30.

Yohan Park, Hyunwoo Ha, Wonjun Jo, and Tae-Hyun Oh. Darkeqa: Benchmarking vision-language models for embodied question answering in low-light indoor environments. *arXiv preprint arXiv:2512.24985*, 2025.

Pexels. Free stock photos, royalty free stock images & copyright free pictures. https://www.pexels.com/, 2026. Accessed: 2026-05-03.

Lu Qiu, Yuying Ge, Yi Chen, Yixiao Ge, Ying Shan, and Xihui Liu. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *arXiv preprint arXiv:2412.04447*, 2024.

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

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Rohit Saxena, Alessandro Suglia, and Pasquale Minervini. Vlm-robustbench: A comprehensive benchmark for robustness of vision-language models. *arXiv preprint arXiv:2603.06148*, 2026.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 645–652. IEEE, 2024.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15768–15780, 2025.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Yingbo Tang, Lingfeng Zhang, Shuyi Zhang, Yinuo Zhao, and Xiaoshuai Hao. Roboafford: A dataset and benchmark for enhancing object and spatial affordance learning in robot manipulation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 12706–12713, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3758209. URL https://doi.org/10.1145/3746027.3758209.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025a.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 70576–70631. PMLR, 2025b.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction in robotics. In *Proceedings of The 8th Conference on Robot Learning*, volume 270, pp. 4005–4020. PMLR, 2025.

Yifu Yuan, Haiqin Cui, Yibin Chen, Zibin Dong, Fei Ni, Longxin Kou, Jinyi Liu, Pengyi Li, Yan Zheng, and Jianye Hao. From seeing to doing: Bridging reasoning and decision for robotic manipulation. In *International Conference on Learning Representations*, 2026.

Chenshuang Zhang, Fei Pan, Junmo Kim, In So Kweon, and Chengzhi Mao. Imagenet-d: Benchmarking neural network robustness on diffusion synthetic object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21752–21762, 2024.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and Shanghang Zhang. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

APPENDIX

A ROBOSTRESSBENCH DETAILS

A.1 DATA SOURCES

RoboStressBench is constructed from three types of source data: existing public benchmarks, Internet-sourced real-world images, and self-collected images. For existing public benchmarks, we use samples from EmbSpatial-Bench Du et al. (2024), RefSpatial-Bench Zhou et al. (2025), RoboAfford-Eval Tang et al. (2025), RoboSpatial-Home Song et al. (2025), Manipulation-VQA Chen et al. (2025), VABench-P Yuan et al. (2026), Where2Place Yuan et al. (2025), and RoboRefit Lu et al. (2023). We retain the license and usage terms of each original source. For Internet-sourced examples, we mainly collect images from Pexels Pexels (2026) and follow the Pexels License, which allows free use and modification while restricting redistribution as standalone stock content. Self-collected images are captured by ourselves in diverse physical environments and are used to cover naturally occurring stress cases that are difficult to obtain from public datasets alone.

A.2 ANNOTATION PROTOCOL

Six trained annotators with background knowledge in embodied AI and vision-language models perform the annotations. For examples from existing public benchmarks, annotators first identify images that already exhibit physical visual stress and assign both coarse stress dimensions and fine-grained stress tags according to the Material, Viewpoint, Lighting, and Geometry taxonomy. When the original task annotations remain valid, we directly reuse the original questions and answers after manual verification. For images that do not originally contain clear stress but are suitable for controlled augmentation, we generate stressed variants and then manually check whether the original questions can still be reused. If a question becomes ambiguous or no longer matches the edited image, annotators revise the question or answer accordingly.

For Internet-sourced and self-collected images, we use a vocabulary-driven annotation pipeline. We first define a fixed object and scene vocabulary to guide candidate collection. Then, GroundingDINO Liu et al. (2024a) and SAM Kirillov et al. (2023) are used to generate candidate object annotations, which are ranked by confidence. Annotators manually inspect the candidates, remove noisy or ambiguous cases, assign stress labels, and write task-specific questions and answers. This process ensures that each example is physically meaningful, visually grounded, and aligned with the intended evaluation task. The annotation interface is shown in Fig. 9.

A.3 GROUNDING ANNOTATION NORMALIZATION

For grounding tasks, we normalize all point and bounding-box annotations to a unified coordinate range of $[0, 1000]$. During evaluation, models are explicitly prompted to output grounding results in the same normalized coordinate system. For point-based grounding, the model outputs a point coordinate (x, y); for box-based grounding, it outputs a bounding box (x_1, y_1, x_2, y_2), where all coordinates are represented in the $[0, 1000]$ range. This normalization makes grounding evaluation independent of the original image resolution and aspect ratio. It also allows different VLMs to follow a unified output format when images have different sizes, avoiding ambiguity caused by pixel-coordinate conventions.

A.4 TARGETED STRESS SYNTHESIS

Targeted stress synthesis complements naturally occurring data by increasing coverage of stress categories that are rare or difficult to obtain from existing sources. Starting from a nominal image, we use Gemini-3-Pro-Image Google Gemini (2026) and Qwen-Image-Edit Wu et al. (2025) to generate a stressed variant. Each editing prompt specifies the target stress category, the intended visual modification, and preservation constraints for task-relevant scene content. The generated image and its annotations are then manually verified.

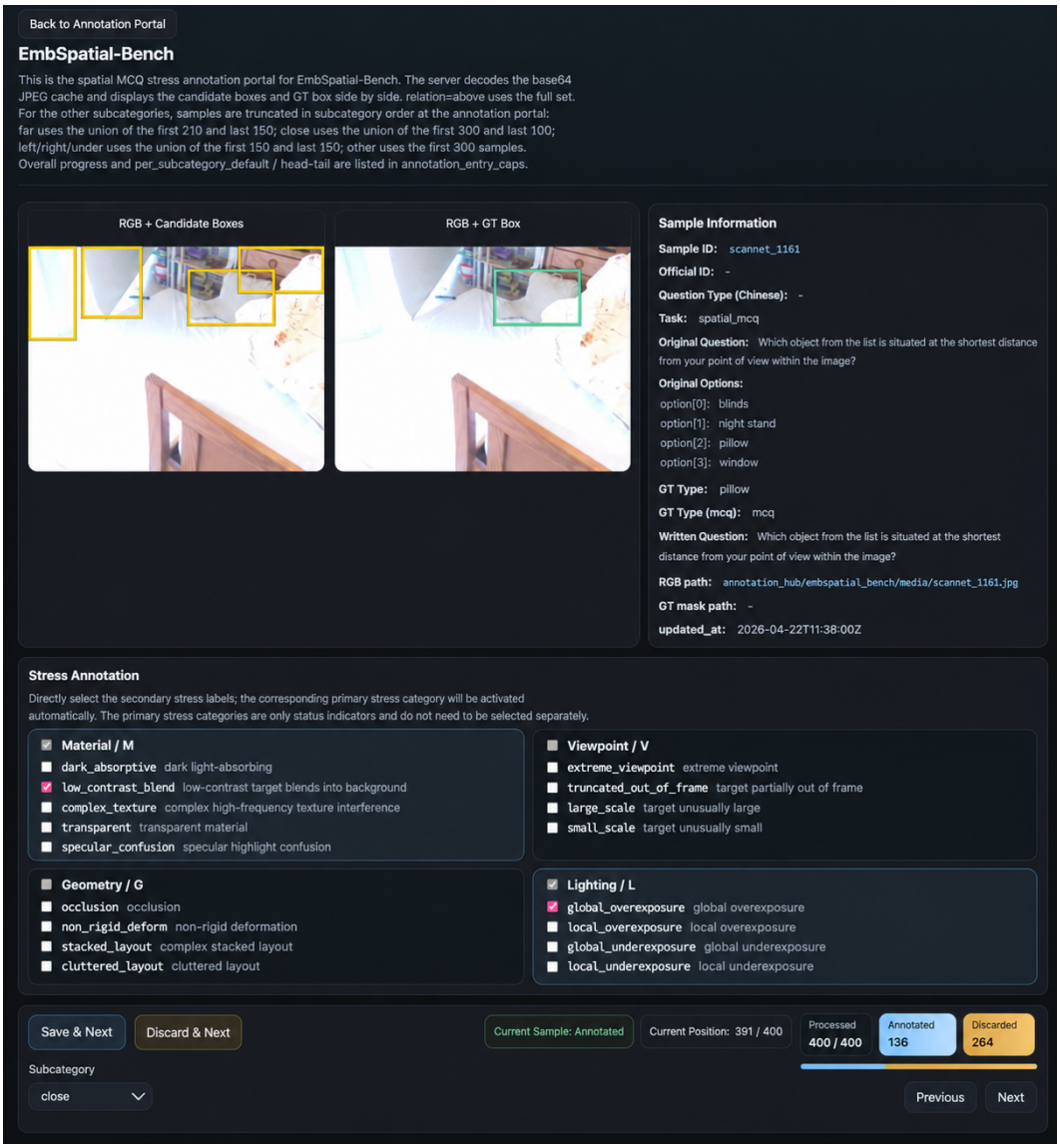

Figure 9: **Annotation interface for RoboStressBench.** Annotators inspect each image, assign coarse stress dimensions and fine-grained stress tags, verify or revise task questions and answers, and check grounding annotations when applicable.

Each synthesis job is defined by an *edit profile*. An edit profile contains: (i) a nominal image and its original task annotation; (ii) a target stress category; (iii) an editing instruction describing the desired physical change; (iv) preservation constraints for task-relevant objects, layout, camera pose, lighting consistency, and photorealism; and (v) an annotation policy specifying whether the original label can be reused. For grounding tasks, we resize the nominal and stressed images to the same resolution and check whether the target region remains pixel-aligned. If alignment is preserved, we transfer the original normalized point or bounding box annotation. If the edit changes the target location, shape, visibility, or surrounding evidence, annotators re-label the stressed image. All generated samples are manually verified for three conditions: the intended stress is visually present, the task semantics remain valid, and the final annotation is correct.

Figures 10–13 illustrate representative synthesis protocols. We organize these examples by *how the edit is controlled*, rather than by enumerating every stress axis or sub-category. The selected cases cover three common control modes used throughout the pipeline: temporary spatial guides,



(a) (b) (c)

Figure 10: **Controlled synthesis with a temporary bounding-box guide.** (a) Nominal inputs with rasterized red rectangles used only as editing guides. The guides indicate regions whose target object, pose, and alignment should be preserved during editing. (b) Stressed outputs after adding surrounding clutter; cyan boxes visualize annotations that are reused only when post-edit alignment is verified. (c) Example edit prompt specifying the target stress, the protected region, and the requirement that the guide must not appear in the final image.



(a) (b) (c)

Figure 11: **Language-only synthesis for spatial and non-rigid stress.** (a) Nominal scene without rasterized guides. (b) Stressed output with an inserted deformable foreground object; boxes indicate annotator-verified regions after editing. (c) Example prompt that specifies the non-rigid object, its placement, and preservation constraints for other objects, camera pose, lighting, and photorealistic style.

language-only spatial edits, and appearance-factor edits. Across all modes, the principle is the same: introduce the intended visual stress, check task-relevant content, and determine annotation reuse through post-edit verification.

Region-guided preservation. Figure 10 illustrates a guide-based protocol for cases where the target annotation should remain stable while the surrounding scene becomes more challenging. We rasterize the original bounding box as a temporary editing guide on the reference image and instruct the editor to keep the object inside the guide fixed while adding clutter outside the guide. The guide is used only during synthesis and is never included in the final evaluation image. This protocol is useful when the desired stress affects the nearby spatial context, such as clutter or background complexity, rather than the target object itself. When the edited image remains aligned with the nominal image, the original grounding annotation and question can be reused.

Language-only spatial edits. Figure 11 shows a complementary protocol for edits that are easier to specify with natural language than with a rasterized guide. The prompt describes the inserted or modified object, its spatial placement, and the scene elements that must remain unchanged. This mode is useful for non-rigid geometry stress, deformable foreground insertions, and other cases where exact pixel preservation is not assumed. Because the inserted object may change the visible scene layout or object extent, annotators verify the edited content and update grounding labels whenever the target region, visibility, or spatial evidence changes.

Appearance-factor edits. Figures 12 and 13 illustrate controlled edits to appearance factors such as illumination and surface material. These edits do not aim to change the object layout; instead, they modify visual evidence by altering how the existing scene is seen. For lighting stress, the



Figure 12: **Appearance-factor synthesis via local overexposure.** (a) Nominal scene before editing. (b) Stressed output with a bright, localized overexposed region produced by a directional lighting change, while the overall scene structure and task-relevant objects are preserved where possible. (c) Example prompt describing the lighting manipulation and preservation constraints.

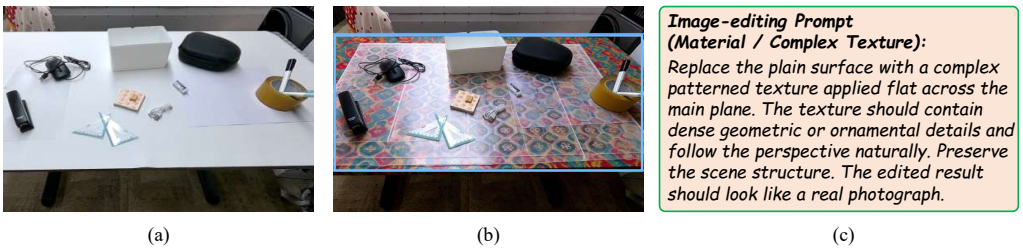

Figure 13: **Appearance-factor synthesis via complex texture.** (a) Nominal scene before editing. (b) Stressed output with dense texture or typographic patterns applied to a surface in a perspective- and lighting-consistent manner, while keeping the manipulation layout stable when possible. (c) Example prompt describing the material change and the requirement to preserve scene structure and photorealism.

prompt can introduce a concentrated highlight, glare region, shadow pocket, or exposure change while preserving global geometry and object identity. For material stress, the prompt can replace a plain surface with a dense texture, decal, or typographic pattern that follows the scene perspective and lighting. In both cases, annotators check that the edited appearance is physically plausible, that the intended stress category is satisfied, and that task-relevant evidence remains valid. Grounding annotations are reused only when the target remains aligned; otherwise, the stressed image is re-annotated.

Scope of examples. The examples above are representative synthesis profiles, not an exhaustive catalog of all fine-grained stress categories. Many sub-categories follow the same workflow with different prompt instantiations: specify the target stress, preserve task-relevant content, generate the stressed variant, and verify annotation validity. The selected examples cover the main control modes in our pipeline—region-guided preservation, language-only spatial editing, and appearance-factor editing—and therefore illustrate the full synthesis and quality-control procedure.

A.5 Dataset Statistics

RoboStressBench contains 7183 examples in total. Among them, 2927 examples are filtered from existing unconstrained datasets, 2596 examples are generated through controlled stress synthesis, and 1660 examples are collected from additional real-world sources, including Internet-sourced images and images captured by ourselves. This combination allows the benchmark to include both naturally occurring stress cases and controlled high-stress variants.

In terms of stress distribution, RoboStressBench includes 2785 Material examples, 1292 Viewpoint examples, 1753 Lighting examples, and 3327 Geometry examples. For Material stress, the dataset contains 711 dark absorptive, 761 low-contrast blend, 551 complex texture, 575 transparent, and

495 specular-confusion examples. For Viewpoint stress, it contains 212 extreme-viewpoint, 665 truncated-out-of-frame, and 496 small-scale examples. For Lighting stress, it contains 364 global-overexposure, 575 local-overexposure, 521 global-underexposure, and 319 local-underexposure examples. For Geometry stress, it contains 1205 occlusion, 579 non-rigid-deformation, 865 stacked-layout, and 1658 cluttered-layout examples. Note that a single example may be associated with multiple stress tags; consequently, the per-tag counts are reported independently and their sum may exceed the total number of examples.

RoboStressBench also covers multiple evaluation tasks. Specifically, it contains 949 placement-grounding examples, 3411 target-grounding examples, 1369 spatial-reasoning multiple-choice examples, 633 state-understanding multiple-choice examples, and 821 planning multiple-choice examples. These task types are designed to evaluate complementary embodied capabilities, including object localization, target grounding, spatial relation reasoning, object-state understanding, and high-level planning under physical visual stress. We provide representative examples from RoboStressBench in Fig. 14–Fig. 18. All examples and annotations of RoboStressBench are provided in the supplementary material.

B Additional Experimental Details

B.1 Detailed Grounding Results

RoboStressBench contains both point-based and box-based grounding tasks. For point-based grounding, a prediction is considered correct if the predicted point falls inside the ground-truth mask. For box-based grounding, we follow COCO-style IoU evaluation and report IoU@0.50, IoU@0.95, and the mean accuracy averaged over thresholds from 0.50 to 0.95 with a step size of 0.05. In Table 2, the grounding score is the equal-weight average of point-based grounding accuracy and box-based mean localization accuracy: $\mathrm{Grounding} = (\mathrm{PointAcc} + \mathrm{BoxmAcc})/2$. BoxmAcc averages localization accuracy over IoU thresholds from 0.50 to 0.95 with a step of 0.05. Table 4 reports the separate point-based results and the complete box-based grounding metrics.

B.2 Compute Resources

All experiments in RoboStressBench are conducted in an inference-only setting, without model fine-tuning or parameter updates. For open-source VLMs, we run evaluation on 8 NVIDIA H100 GPUs, each with 80 GB memory, using the official inference implementations of each model. All models are evaluated with deterministic greedy decoding, setting the maximum generation length to 64 new tokens and disabling sampling (temperature = 0.0, top-p = 1.0), as described in Sec. 5. The full open-source model evaluation takes approximately 48 GPU-hours in total.

For StressDART, the base reasoner is Qwen3-VL-4B Bai et al. (2025), and the Stress Rectifier is implemented with Qwen-Image-Edit Wu et al. (2025). This rectification step introduces additional test-time image-editing cost, requiring approximately 150 GPU-hours for the evaluated subset. Closed-source models, including Gemini-3.1 Google DeepMind (2026) and GPT-5.5 OpenAI (2026), are evaluated through official APIs and therefore do not consume local GPU resources.

C Limitations

RoboStressBench is designed as a diagnostic benchmark for physical visual stress in embodied scenes, but it still has several limitations. First, although our Material–Viewpoint–Lighting–Geometry taxonomy is physically grounded and interpretable, it is not intended to exhaust all possible sources of visual difficulty in real-world embodied environments. Although RoboStressBench supports multi-label stress annotation, stress axes are not perfectly orthogonal in real scenes; factors such as viewpoint and geometry or lighting and material appearance can still be entangled, making fine-grained attribution challenging.

Second, our dataset construction combines human-curated filtering, controlled stress synthesis, and additional real-world collection. While this design balances realism, diversity, and controllability, it may still introduce source bias from the datasets and scenes we sample, as well as artifacts from generative editing for synthesized stress cases. We reduce this risk through manual verification



Figure 14: **Representative examples from RoboStressBench.**We show several physically stressed examples with their questions, answers, and stress annotations.

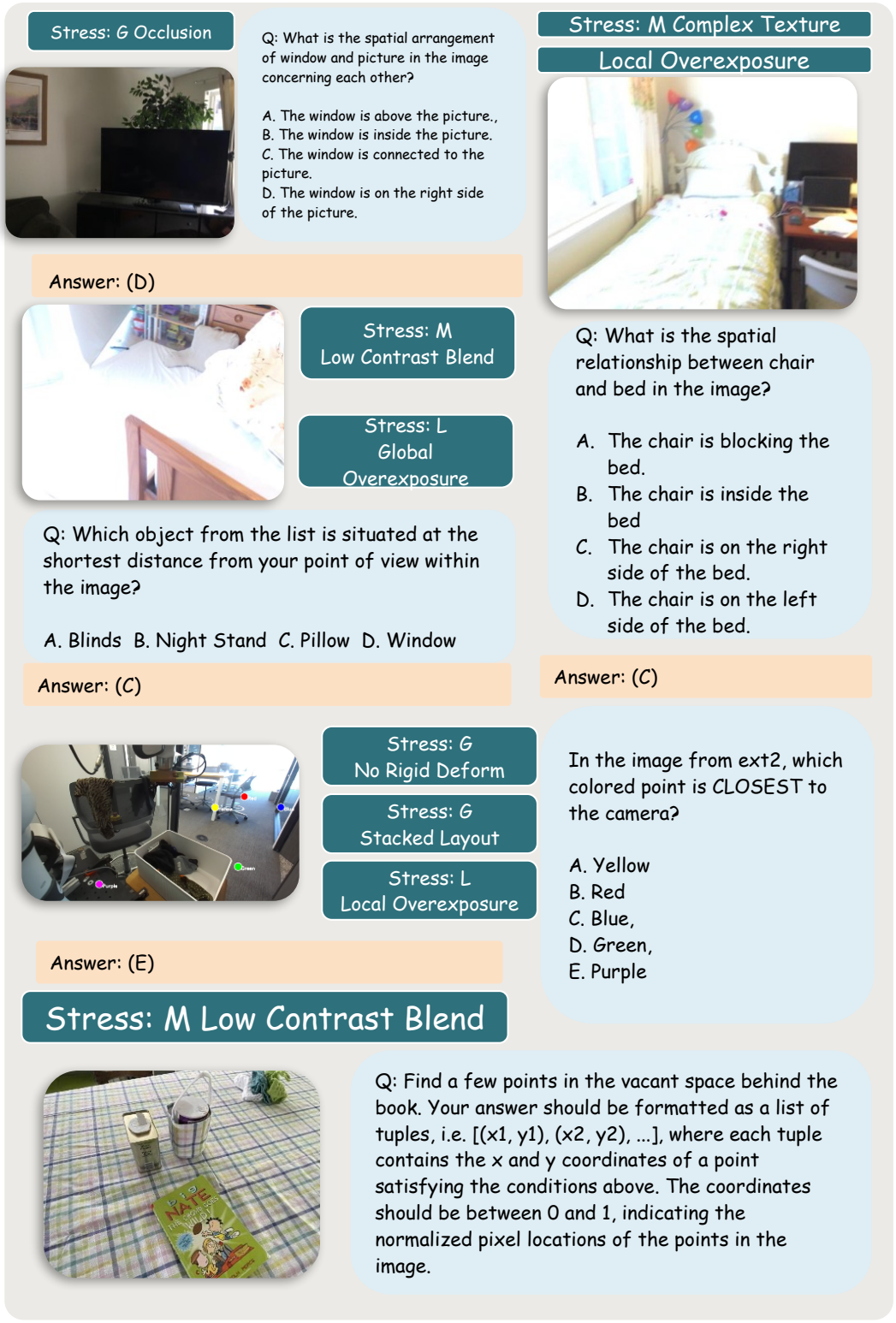

Figure 15: **Representative examples from RoboStressBench.**We show several physically stressed examples with their questions, answers, and stress annotations.



Figure 16: **Representative examples from RoboStressBench.**We show several physically stressed examples with their questions, answers, and stress annotations.

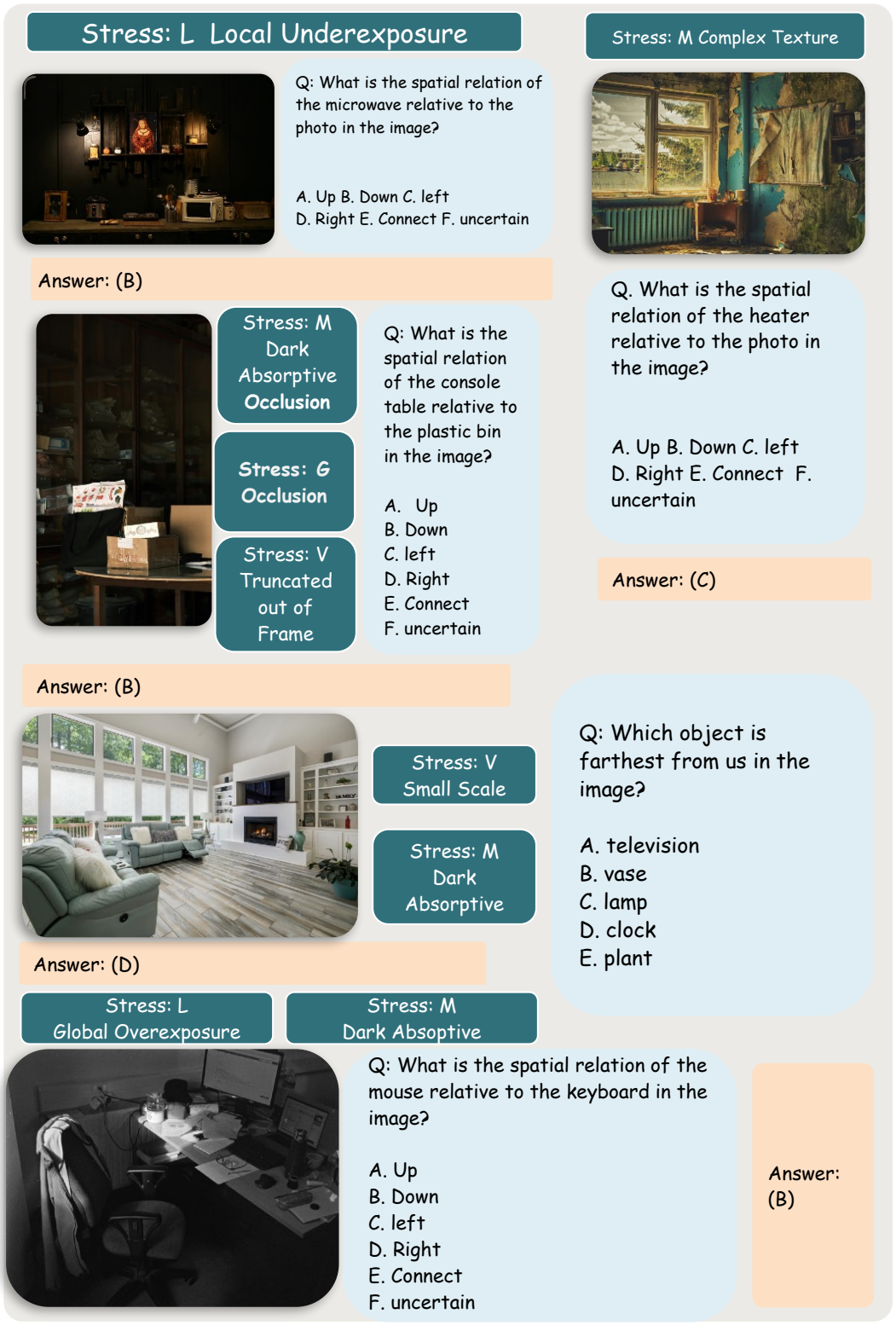

Figure 17: **Representative examples from RoboStressBench.**We show several physically stressed examples with their questions, answers, and stress annotations.

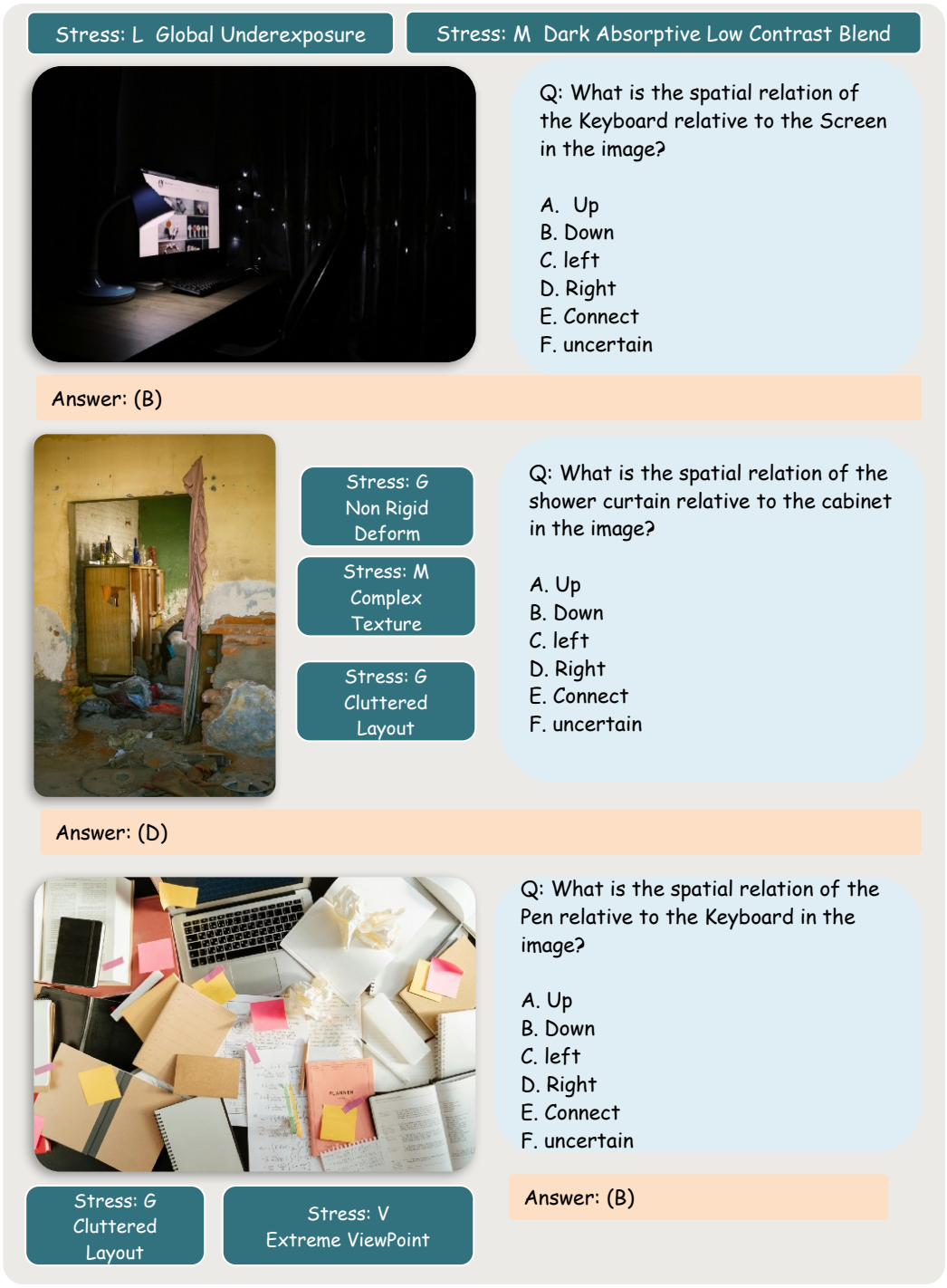

Figure 18: **Representative examples from RoboStressBench.**We show several physically stressed examples with their questions, answers, and stress annotations.

Table 4: **Detailed grounding results.** We report point-based grounding accuracy, box localization accuracy at IoU thresholds of 0.50 and 0.95, and mean localization accuracy over thresholds from 0.50 to 0.95 in increments of 0.05.

Model	Size	Point-acc	IoU@0.50	IoU@0.95	mAcc
Qwen3VL	4B	50.1	80.4	26.6	68.6
Qwen3VL	8B	54.4	79.7	24.7	65.8
Qwen3VL	30B-A3B	53.5	82.0	30.4	70.0
Qwen3.5	4B	52.0	82.0	25.7	67.7
Qwen3.5	9B	56.9	**82.6**	25.5	68.8
Qwen3.5	27B	**64.3**	**83.1**	**34.3**	**72.4**
Qwen3.5	35B-A3B	58.9	82.0	**32.1**	**70.5**
Qwen3.6	27B	**63.3**	**83.3**	**34.5**	**72.4**
Qwen3.6	35B-A3B	59.9	80.6	28.0	68.7
InternVL3.5	4B	37.7	18.8	0.5	8.4
InternVL3.5	8B	37.1	27.6	0.5	11.0
InternVL3.5	14B	28.0	7.9	0.0	2.2
Molmo2	4B	38.2	3.4	0.0	0.9
Molmo2	8B	49.5	4.4	0.1	1.3
Gemini-3.1	–	58.3	60.3	18.8	45.3
GPT-5.5	–	**60.4**	80.3	15.0	61.2

and re-annotation when necessary, but synthetic examples cannot fully replace naturally occurring physical stress.

Third, the current benchmark focuses on image-based VQA and grounding tasks. These tasks capture important perception, spatial reasoning, and planning-related abilities, but they do not fully evaluate closed-loop embodied behavior, long-horizon interaction, or temporal robustness in dynamic scenes. Extending RoboStressBench to video observations, multi-view interaction, and real robot execution would provide a more complete picture of embodied robustness.

Finally, StressDART is an initial test-time intervention rather than a fully optimized robustness framework. Its results show that explicit stress diagnosis and targeted rectification can substantially improve performance. An improvement in aggregate performance does not establish that rectification preserves task-relevant evidence in every example. Further analysis is needed to quantify correct-to-incorrect prediction changes and distinguish errors in stress prediction from changes introduced by image editing. Future work should investigate more reliable stress detectors, content-preserving rectification methods, and reasoning strategies that can better decide when to trust the original image, the rectified image, or both. We hope these limitations will motivate future research on more realistic, temporally grounded, and action-aware robustness evaluation for embodied VLMs.

D Broader Impacts

RoboStressBench aims to support the development of more reliable VLMs for embodied AI by exposing failures under physically plausible visual stress. Such evaluation can benefit robotics systems that must operate in challenging real-world environments, including low illumination, occlusion, reflective materials, unusual viewpoints, and cluttered scenes. By providing stress annotations and task-level evaluation, RoboStressBench can help researchers diagnose when VLM perception is unreliable and develop targeted robustness improvements before deployment.

At the same time, the benchmark and StressDART inherit broader risks associated with VLM-based embodied systems. Models may still hallucinate answers, mislocalize objects, or overestimate their confidence under severe visual ambiguity. When integrated into robotic pipelines, such errors may lead to unsafe manipulation, navigation, or planning decisions, especially in high-stakes environments. StressDART can improve robustness through test-time rectification, but visual editing may also alter task-relevant evidence if applied incorrectly, so it should not be treated as a substitute for calibrated uncertainty estimation or safety checks.

We believe that releasing RoboStressBench to the research community can have positive impact by enabling more systematic evaluation of perception robustness under realistic physical stress. Open access to the benchmark, annotations, and evaluation protocol can facilitate reproducible comparison, encourage stress-aware model development, and support safer embodied AI systems across robotic platforms such as mobile robots, robotic arms, and humanoids.

E Impact Mitigation Measures

RoboStressBench is an evaluation benchmark, not a deployed embodied agent, but we still consider the possible risks related to data release and benchmark usage. We will document the dataset sources, stress taxonomy, annotation process, task formats, evaluation metrics, and inference settings to make the benchmark transparent and reproducible. When releasing the data, we will follow the licenses of the original sources, remove or exclude images with personally identifiable or sensitive information, and clearly indicate which samples are real and which are synthesized. We will also release the benchmark data, annotation schema, evaluation scripts, and usage instructions as soon as possible to support responsible use by the community.

We will also provide clear guidance on what the benchmark should and should not be used for. RoboStressBench is intended to help researchers diagnose how VLMs fail under realistic physical visual stress, rather than to support surveillance, biometric identification, or high-stakes automated decisions. Similarly, StressDART should be viewed as an exploratory test-time strategy, not as a complete safety mechanism, since visual editing may sometimes change important visual cues. Therefore, we encourage users to report both successful and failed cases, keep the original image available during reasoning, and treat benchmark results as evidence for improving model robustness rather than as proof that a model is ready for real-world deployment.

F Licenses

RoboStressBench is constructed from existing public benchmarks, Pexels-sourced real-world images, and controlled stress synthesis. We retain the license and usage terms of each original data source. Our annotations, metadata, and benchmark construction code may be released under our chosen research license, while images and derived visual assets remain subject to the licenses or terms of their corresponding source data.

1. **Existing public benchmarks.** RoboStressBench uses samples from EmbSpatial-Bench Du et al. (2024), released under CC BY 4.0; RefSpatial-Bench Zhou et al. (2025), released under Apache 2.0; RoboAfford-Eval Tang et al. (2025), released under CC BY 4.0; RoboSpatial-Home Song et al. (2025), released under Apache 2.0; ManipulationVQA Chen et al. (2025), released under Apache 2.0; VABench-P Yuan et al. (2026), released under Apache 2.0; and Where2Place Yuan et al. (2025), released under Apache 2.0. RoboRefit Lu et al. (2023) is distributed via the official VL-Grasp repository without an explicit dataset license; we use it for non-commercial academic research only, consistent with common practice for unlicensed academic datasets.
2. **Pexels-sourced real-world images.** The dataset contains images sourced from Pexels Pexels (2026). Under the Pexels License, content is free to use and modify for commercial or non-commercial purposes without required attribution. The terms explicitly prohibit redistributing or selling the photos on other stock photo or wallpaper platforms. We release these images exclusively as part of an academic benchmark dataset, which strictly complies with these terms. Users of our benchmark are also subject to the original Pexels License.

3. **Controlled stress synthesis.** Some controlled stress samples are synthesized from existing benchmark images, such as lighting-stress variants generated from public benchmark sources. These derived samples inherit the license and usage constraints of their underlying source datasets and are not relicensed independently. Synthesis based on proprietary in-house data uses raw images provided by an industrial partner. We plan to release these specific derived samples under a research-only, non-commercial license to protect the proprietary nature of the original assets.