LumiBench: Benchmarking Illumination Robustness in Robotic Manipulation Policies

Abstract— **Illumination shifts can substantially degrade the performance of robotic manipulation policies, yet existing benchmarks typically treat illumination as a single perturbation factor, leaving its internal structure underexplored. We introduce LumiBench, an illumination-centric diagnostic benchmark that organizes lighting variations by their dominant physical mechanisms and evaluates them through paired nominal–perturbed rollouts. LumiBench further introduces within-condition severity sweeps to distinguish sensitivity to perturbation onset from degradation under increasing perturbation strength. We instantiate LumiBench in RoboTwin 2.0 and evaluate five representative policy families across ten manipulation tasks. Our results show that illumination robustness is strongly policy, task, and mechanism-dependent, while severity sweeps reveal progressive, onset-dominated, and strongly policy-specific response patterns. Controlled physical experiments further reproduce the strong Spot vulnerability observed in simulation, providing representative evidence that selected illumination sensitivities persist under real lighting changes.**

I. INTRODUCTION

Vision-based robotic manipulation policies have made substantial progress in increasingly diverse tasks and environments [1]–[7]. However, real-world deployment inevitably exposes robots to illumination conditions that differ from those encountered during training. Such variation extends well beyond changes in global brightness. A mobile manipulator may move from a bright to a dim zone (intensity), operate under light sources of different color temperature (chromaticity), work beside a task lamp that creates a localised highlight (spatial non-uniformity), observe the scene under flickering artificial lighting (temporal variation), or be overtaken by the shadow of a passing occluder (dynamic shadows). These illumination shifts can affect object appearance, local contrast, and color cues through different mechanisms, potentially leading to perceptual errors, incorrect actions, and ultimately task failure [8].

Prior work has established illumination variation as an important source of visual distribution shift for robotic manipulation policies [9]–[11]. However, these benchmarks primarily target broad environmental robustness, with lighting evaluated alongside factors such as object appearance, background, camera viewpoint, and distractors [12]. Two limitations remain. First, illumination is often instantiated through coarse or coupled lighting settings, making it difficult to attribute a measured performance drop to a specific illumination mechanism. Second, even when graded difficulty is considered, existing evaluations provide a limited characterization of how performance evolves as a single, physically interpretable illumination mechanism is progressively strengthened.

To address these limitations, we introduce LumiBench (Fig. 1), an illumination-centric diagnostic benchmark for robotic manipulation policies. LumiBench organizes illumination variations according to five dominant mechanisms—global intensity, chromaticity, spatial non-uniformity, temporal variation, and dynamic shadows—and implements them as controlled interventions that preserve task semantics, scene geometry, robot configuration, camera setup, and evaluation initialization. For representative mechanisms with interpretable control variables, LumiBench further evaluates policies over progressively increasing perturbation severity. This formulation moves beyond asking whether lighting affects robotic manipulation and instead investigates three questions: (1) how illumination robustness varies across perturbation mechanisms, policies, and tasks; (2) how policy performance evolves as perturbation severity increases; and (3) whether selected illumination vulnerabilities identified in simulation persist under qualitatively corresponding physical lighting conditions.

We instantiate LumiBench in RoboTwin 2.0 and evaluate five policies across ten manipulation tasks under one nominal condition and nine illumination perturbations, complemented by controlled physical experiments on a representative subset. Our results reveal strongly mechanism-dependent robustness patterns: FastWAM is far more vulnerable to Grating than to Strobe, whereas LingBot-VA exhibits the reverse ordering. Nominal competence alone does not reliably indicate illumination robustness: averaged across nine illumination perturbations, OpenVLA-OFT retains 88.3% of its Base success rate, compared with 46.1% for FastWAM despite FastWAM's higher nominal performance.

Our contributions are threefold:

1. We formulate illumination robustness as a controlled intervention problem and introduce a mechanism-structured evaluation space covering global intensity, chromaticity, spatial non-uniformity, temporal variation, and dynamic shadows, together with a composite illumination stress condition.

2. We introduce a paired severity-response evaluation protocol that separates sensitivity to perturbation onset from further degradation under increasing severity.

3. We instantiate LumiBench in RoboTwin 2.0 and evaluate five representative policy families across ten manipulation tasks, complemented by controlled physical experiments that reproduce selected simulated vulnerabilities, notably the strong sensitivity to localized Spot illumination.

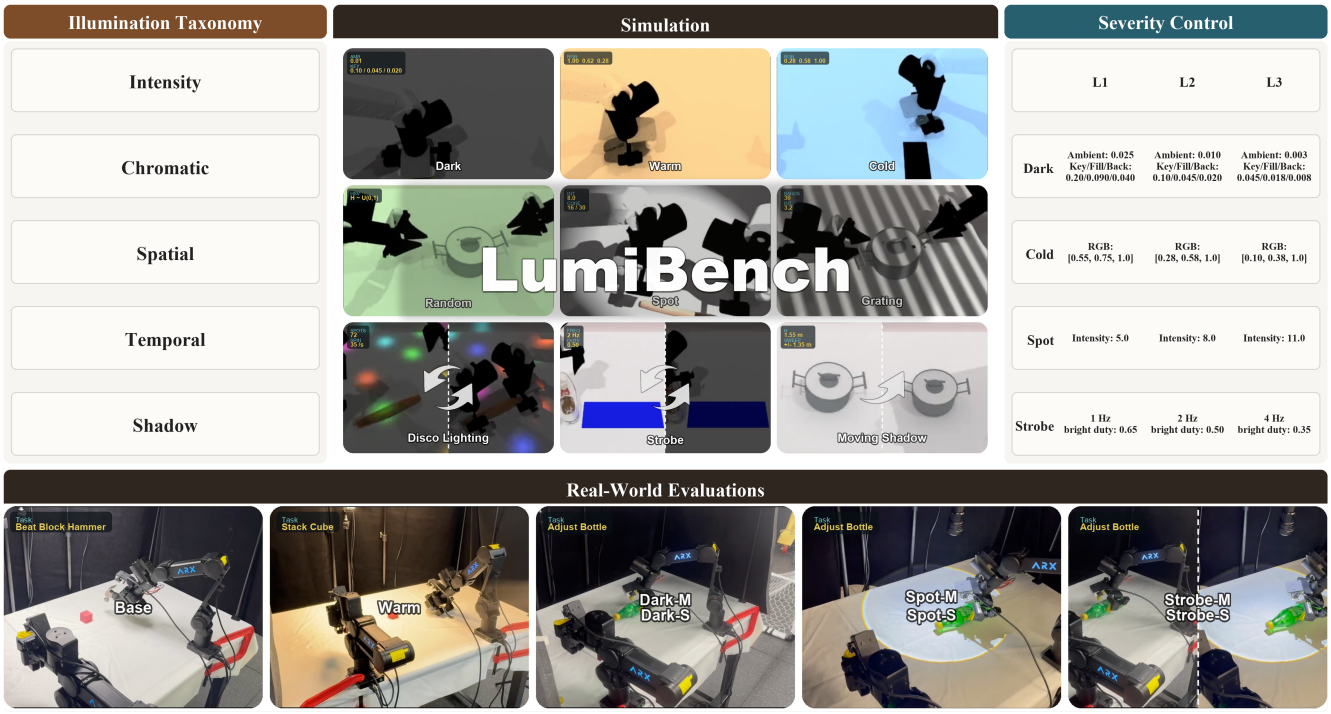

Fig. 1. **LumiBench overview.** LumiBench organizes illumination perturbations into five mechanisms—intensity, chromaticity, spatial non-uniformity, temporal variation, and dynamic shadows—with *Disco Lighting* included as a composite stress condition. The figure presents the illumination taxonomy (left), representative simulated perturbations (center), and physically interpretable control parameters defining three ordered severity levels from mild to severe for Dark, Cold, Spot, and Strobe (right). We evaluate policies under nine simulated perturbations and complement simulation with controlled physical experiments on a real ARX platform, including a localized physical implementation of Strobe (*bottom*).

II. Related Work

A. Benchmarks for Robotic Manipulation

RLBench [13] and Meta-World [14] established early multi-task and meta-RL suites, and later benchmarks turned to lifelong learning, real-world evaluation, reproducibility, and compositional generalization through LIBERO [15], TOTO [16], SceneReplica [17], and GEMBench [18]. RoboTwin [19] and its successor RoboTwin 2.0 [20] introduced generative digital-twin data generation with structured domain randomization including lighting, while RoboDojo [21] unifies simulation-and-real evaluation. Generalist policies such as Octo [22], VIMA [23], and Dream-Trajectory [24] have scaled alongside large datasets such as Open X-Embodiment [25], BridgeData V2 [26], and DROID [27], which expose policies to illumination diversity only incidentally through uncontrolled collection sites.

These works primarily target task generalization, data scale, or broad environmental diversity, treating illumination as fixed, randomized, or one factor among many.

B. Robustness Evaluation under Distribution Shift

Recent work questions whether nominal success reflects true robustness. Xie et al. [11] analyze the generalization gap of imitation-learning policies across controlled visual factors including lighting, and RoboNVS [28] studies multi-view demonstration synthesis under camera shifts. THE COLOSSEUM [9] evaluates 20 tasks across 14 perturbation axes and finds lighting among the most damaging, with degradation correlating between simulation and physical evaluation. LIBERO-Plus [10] extends LIBERO [15] with seven perturbation factors and five difficulty levels, though its lighting factor only varies intensity, direction, and color, with difficulty defined by aggregate performance rather than controlled severity. RoboTwin 2.0-Plus [12] similarly evaluates VLA and world-action models across seven perturbation dimensions for broad rather than illumination-specific robustness. Image-space corruptions [8] offer a cheaper approximation but fail to capture scene-consistent shading or geometry-coupled dynamics.

C. Illumination Variation and Physical Evaluation

Domain randomization [29] showed that randomizing visual factors in simulation aids sim-to-real transfer, and RoboTwin 2.0 [20] incorporates such randomization over clutter, textures, and lighting. At the image level, training with diverse corruption augmentations [8], [30] improves perturbation robustness, motivating our policy-level evaluation. VISER [31] studies the sim-to-real gap through visual realism, highlighting lighting and material properties for geometric reasoning, ...while TOTO [16], THE COLOSSEUM [9], and RoboDojo [21] provide physical evaluation under environmental variation, and SIMPLER [32] shows that visually aligned simulation can track real-world policy sensitivity to distribution shifts. Across these works, lighting mainly serves as training diversity or one component of sim-to-real evaluation rather than the primary object of study.

III. METHODOLOGY

A. Overview

We study the robustness of robotic manipulation policies under illumination shifts. Rather than treating illumination as a collection of unrelated visual corruptions, we formulate illumination robustness evaluation as a controlled intervention problem: given the same task semantics, scene configuration, robot state, camera setup, and policy checkpoint, how does task performance change when only the illumination condition is modified?

To this end, we introduce **LumiBench**, which organizes illumination variation into five primary mechanisms: **intensity**, **chromaticity**, **spatial non-uniformity**, **temporal variation**, and **dynamic shadows**. We additionally include **Disco Lighting** as a composite stress condition combining chromatic, spatial, and temporal variation.

LumiBench consists of three complementary evaluation components:

1) **Broad illumination evaluation**, which evaluates policies under one nominal condition and nine illumination perturbations spanning the five primary mechanisms, with Disco Lighting included as a composite stress condition;
2) **Severity-controlled evaluation**, which evaluates one representative perturbation from four primary mechanisms over three ordered severity levels; and
3) **Controlled physical evaluation**, which tests whether selected illumination-sensitivity patterns identified in simulation persist under qualitatively corresponding physical lighting conditions.

Throughout the simulation benchmark, task semantics, scene geometry, object initialization, robot state, camera configuration, policy checkpoint, and task-initialization seed are held fixed across illumination conditions. In the physical evaluation, the task protocol and camera poses are fixed, while exact object placement may vary slightly due to manual resetting.

B. Broad Illumination Evaluation

Real-world illumination can vary in intensity, chromaticity, spatial distribution, temporal behavior, and shadow structure. LumiBench therefore organizes illumination perturbations according to these five primary mechanisms and instantiates representative scene-level lighting interventions for each.

All perturbations are implemented by modifying scene-level lighting rather than post-processing camera observations. We instantiate LumiBench in RoboTwin 2.0 using a modular SAPIEN-based [33] illumination backend that controls ambient illumination, directional and spot lights, light color, source geometry, shadow casting, and temporal modulation.

The broad evaluation contains ten conditions: **Base**, **Dark**, **Warm**, **Cold**, **Random**, **Spot**, **Grating**, **Disco Lighting**, **Strobe**, and **Moving Shadow**. **Base** serves as the nominal reference, while the remaining nine conditions introduce controlled illumination shifts. The representative illumination conditions and control parameters are summarized in Fig. 1.

For a policy π, task T, and task initialization s_0, we construct paired evaluation environments

$$\mathcal{E}_{\mathrm{Base}}(T, s_0) \quad \text{and} \quad \mathcal{E}_l(T, s_0), \tag{1}$$

which differ only in illumination condition l. Base and perturbed rollouts use identical task-initialization seeds. For stochastic illumination conditions, lighting randomness is controlled by a separate fixed seed schedule so that task pairing is preserved across conditions.

a) Broad robustness metrics.: For policy p, task t, and illumination condition l, task success rate is defined as

$$SR_{p,t,l} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left[\mathrm{success}_i\right]. \tag{2}$$

Because perturbed success rate alone may conflate illumination robustness with nominal task competence, we additionally report the illumination-induced performance change

$$\Delta SR_{p,t,l} = SR_{p,t,l} - SR_{p,t,\mathrm{Base}}. \tag{3}$$

Negative values indicate degradation relative to nominal illumination.

We additionally report policy-level normalized robustness. Let

$$\bar{SR}_{p,l} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} SR_{p,t,l} \tag{4}$$

denote the task-averaged success rate of policy p under condition l. We define

$$R_{p,l} = \frac{\bar{SR}_{p,l}}{\bar{SR}_{p,\mathrm{Base}}}. \tag{5}$$

$R_{p,l}$ measures the fraction of nominal task-averaged performance retained under an illumination shift.

We further summarize policy-level mean retention across the nine illumination perturbations as

$$R_p = \frac{1}{|\mathcal{L}_{\mathrm{pert}}|} \sum_{l \in \mathcal{L}_{\mathrm{pert}}} R_{p,l}, \tag{6}$$

where $\mathcal{L}_{\mathrm{pert}}$ denotes the set of non-Base illumination conditions.

Policy–task pairs with near-zero Base success are interpreted cautiously, since a small ΔSR in this regime may reflect a floor effect rather than genuine illumination robustness.

C. Severity-Controlled Evaluation

Evaluation under a single perturbed condition does not reveal how policy robustness evolves as the illumination shift becomes stronger. We therefore extend the paired evaluation protocol with three ordered severity levels, denoted as L_1, L_2, and L_3, corresponding to mild, moderate, and severe perturbations.

Rather than defining severity for every lighting condition, we restrict severity sweeps to four representative perturbations with clear and physically interpretable ordered controls:

$$\mathcal{L}_{\mathrm{sev}} = \{\mathrm{Dark}, \mathrm{Cold}, \mathrm{Spot}, \mathrm{Strobe}\}. \tag{7}$$

Dark, Cold, and Spot use predominantly one-dimensional controls for intensity, chromaticity, and spatial non-uniformity, respectively. Strobe uses an ordered temporal schedule that jointly varies switching frequency and dark-phase exposure. Other conditions involve less interpretable coupling or stochastic variation and are therefore evaluated only at canonical settings.

a) Dark.: For Dark, severity is controlled by progressively reducing global illumination while keeping the illumination geometry and color fixed. For L_1, L_2, and L_3, the (ambient, key, fill, back) intensities are set to $(0.025, 0.20, 0.090, 0.040)$, $(0.010, 0.10, 0.045, 0.020)$, and $(0.003, 0.045, 0.018, 0.008)$, respectively.

b) Cold.: For Cold, severity is controlled using three progressively stronger blue/cyan illumination chromaticities. The RGB ratios for L_1, L_2, and L_3 are set to $[0.55, 0.75, 1.0]$, $[0.28, 0.58, 1.0]$, and $[0.10, 0.38, 1.0]$, respectively. Each chromaticity is luminance-matched using Rec. 709 luminance normalization so that chromaticity remains the dominant source of variation across severity levels.

c) Spot.: For Spot, severity is controlled by increasing the spotlight intensity while keeping its position, orientation, color, cone angles, and background illumination fixed. The spotlight intensities for L_1, L_2, and L_3 are set to 5.0, 8.0, and 11.0 simulator units, respectively. Increasing spotlight intensity produces progressively stronger local contrast between the illuminated region and its surroundings.

d) Strobe.: Strobe illumination alternates between fixed bright and dark illumination states using a periodic square-wave pattern. Across severity levels, the bright and dark intensities, light geometry, color, and shadow configuration are kept fixed.

Severity is controlled through two temporal parameters: the switching frequency f and the bright-phase duty cycle D. Higher frequency produces more frequent illumination transitions, while a smaller duty cycle increases the fraction of time spent in the dark phase. We define

$$\begin{aligned}(f, D)_{L_1} &= (1\,\text{Hz},\ 0.65),\\ (f, D)_{L_2} &= (2\,\text{Hz},\ 0.50),\\ (f, D)_{L_3} &= (4\,\text{Hz},\ 0.35).\end{aligned} \tag{8}$$

Thus, increasing severity jointly increases the temporal switching rate and dark-phase exposure. We treat L as an ordinal severity variable and do not attribute the resulting performance change to either temporal factor in isolation.

e) Severity evaluation.: For each perturbation, we report the success rate at Base (L_0) and at the three ordered severity levels (L_1–L_3). These measurements characterize the severity-response pattern directly and distinguish sensitivity to the initial introduction of a perturbation from any additional degradation as its severity increases. We do not assume a linear or strictly monotonic relationship between perturbation strength and policy performance. Severity levels are ordered within each perturbation and are not assumed to be quantitatively comparable across different illumination mechanisms.

D. Controlled Real-World Evaluation

Large-scale simulation provides reproducible and tightly controlled evaluation, but robotic policies ultimately operate under physical lighting conditions. We therefore complement the simulation benchmark with controlled real-world experiments using selected physical illumination conditions.

The physical evaluation considers **Base**, **Warm**, **Dark**, **Spot**, and **Strobe**. Spot is produced using a fixed localized directional light, whereas the physical Strobe condition is implemented by intermittently switching the same localized light on and off. This implementation differs from the simulated Strobe perturbation, which modulates global scene illumination. Consequently, the physical Strobe condition couples spatial localization with temporal switching and does not constitute a mechanism-matched realization of the simulated global Strobe perturbation. We therefore treat it as a controlled physical stress condition rather than as a direct physical counterpart of simulated Strobe.

Camera pose and task configuration are kept fixed across conditions, and automatic exposure and white balance are disabled to prevent camera-side adaptation from compensating for the intended illumination changes.

For Dark and Spot, the moderate and severe physical settings are designed as qualitative counterparts of the simulated L_2 and L_3 conditions, respectively. We do not require numerical equivalence between simulator lighting parameters and physical lamp parameters.

For illumination mechanisms with qualitatively corresponding simulated and physical conditions, we compare illumination-induced performance changes for matched policy–task configurations:

$$\begin{aligned}\Delta SR^{\text{sim}}_{p,t,l} &= SR^{\text{sim}}_{p,t,l} - SR^{\text{sim}}_{p,t,\text{Base}},\\ \Delta SR^{\text{real}}_{p,t,l} &= SR^{\text{real}}_{p,t,l} - SR^{\text{real}}_{p,t,\text{Base}}.\end{aligned} \tag{9}$$

In the present study, this correspondence analysis is applied to Dark and Spot. Physical Strobe is reported separately because its spatially localized implementation does not have a mechanism-matched simulated counterpart.

IV. Experiments

We organize our experiments around three questions: (1) how illumination robustness varies across perturbation mechanisms, policies, and tasks; (2) how policy performance changes as illumination severity increases; and (3) whether selected illumination vulnerabilities identified in simulation persist on a physical robot.

A. Experimental Setup

a) Simulation.: We evaluate five policies on ten RoboTwin 2.0 tasks (five single-arm and five bimanual). Each policy is jointly fine-tuned across the ten tasks using 50 demonstrations per task collected under Base illumination with lighting randomization disabled, yielding one multi-task checkpoint per policy. No policy is adapted to the test illumination conditions. We evaluate each policy–task pair under Base and nine illumination perturbations, using 100 episodes per configuration.

b) Real-world setup.: We evaluate $\pi_{0.5}$ and FastWAM on Adjust Bottle, Beat Block Hammer, and Stack Cube using an ARX X5 robotic platform equipped with three cameras. Stack Cube is a physical-only task without a simulated counterpart and is therefore excluded from the simulation–real correspondence analysis. Camera poses and the task reset protocol remain fixed across conditions. We evaluate Base, Warm, Dark, Spot, and Strobe illumination, with moderate and severe settings for Dark, Spot, and Strobe. Strobe is produced by intermittently switching the same localized directional light used for Spot. Automatic exposure and white balance are disabled during evaluation. We conduct 16 trials per policy–task–condition configuration. FastWAM uses a three-task mixed checkpoint covering all three physical evaluation tasks, whereas $\pi_{0.5}$ uses a separate task-specific checkpoint for each task. This choice follows our preliminary empirical observations that FastWAM achieves substantially better performance when jointly fine-tuned on multiple tasks than when trained separately on individual tasks.

c) Metrics.: We report success rate (SR), signed success-rate change relative to Base (ΔSR, in percentage points), and policy-level mean retention (R_p).

B. Robustness under Diverse Illumination

Table I reports success rates averaged over the ten tasks, and Fig. 3 summarizes the corresponding illumination-induced changes under the canonical benchmark settings: perturbations ranked by mean drop in (a) and per-policy ΔSR in (b). Task-level Base rates are shown in Fig. 2.

Under these canonical settings, Grating, Disco, and Spot produce by far the largest average drops, whereas uniform changes in intensity or colour are less damaging on average and Moving Shadow is essentially harmless. The heatmap, however, shows that each policy has a distinct vulnerability profile rather than a scaled version of this ranking. $\pi_{0.5}$ is nearly invariant to everything except spatial non-uniformity; LingBot-VA tolerates all static conditions but collapses under the two with a temporal component (Strobe and Disco); RDT is affected mainly by spatial and composite conditions; OpenVLA-OFT shows broad but mild sensitivity; and FastWAM degrades under almost every mechanism. These profiles are not nested, so a ranking obtained under one mechanism does not predict the ranking under another, and no single perturbation can stand in for illumination robustness in general.

Nominal competence does not reliably indicate illumination robustness. $\pi_{0.5}$ has both the highest Base SR and the highest mean retention (90.4%), but FastWAM, second in Base SR, retains only 46.1%, whereas OpenVLA-OFT starts lower and retains 88.3%. The two largest single-condition drops in the benchmark occur for FastWAM and LingBot-VA. However, the present evaluation does not isolate the effects of policy architecture, pretraining data, or fine-tuning procedure, and these differences should therefore not be attributed to any single design choice.

Finally, sensitivity depends on the task as well. Cold illumination sharply degrades Blocks Ranking RGB for

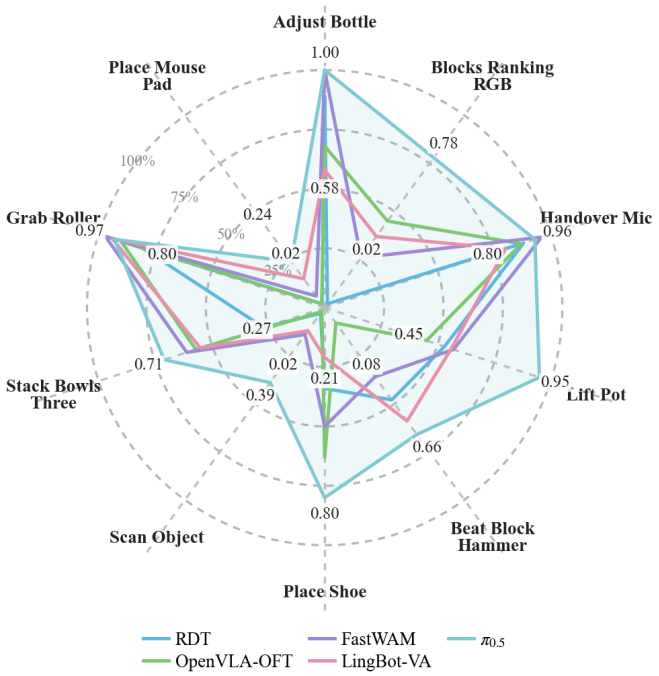

Fig. 2. Base success rates of the five policies across ten RoboTwin tasks. For each task, the outer and inner numbers denote the highest and lowest success rates among the five policies.

OpenVLA-OFT and FastWAM while leaving $\pi_{0.5}$ largely unaffected, consistent with differing reliance on colour cues. For $\pi_{0.5}$ itself, Grating is catastrophic on Adjust Bottle yet benign on Handover Mic. Illumination robustness is thus a joint property of policy, task, and perturbation mechanism, not a scalar attribute of a policy.

Fig. 4 provides a qualitative example on the Adjust Bottle task. Under matched initial conditions, Base and Disco Lighting produce markedly different end-effector trajectories, illustrating how illumination changes can alter manipulation behavior.

C. Robustness across Severity Levels

We next evaluate Dark, Cold, Spot, and Strobe at three ordered severity levels. Dark progressively reduces global illumination, Cold strengthens the blue/cyan chromatic shift, and Spot increases localized illumination contrast. For Strobe, severity is an ordered perturbation schedule that jointly increases switching frequency and dark-phase exposure; we therefore do not attribute its effect to either temporal factor in isolation.

Fig. 5 reveals three qualitatively distinct severity-response regimes. First, **progressive degradation** appears under Dark and Cold, where performance generally decreases as severity increases, with FastWAM showing the strongest decline. Second, **onset-dominated degradation** appears under Spot: most policies incur a substantial loss from Base to L_1, followed by comparatively limited additional degradation from L_1 to L_3. Third, Strobe exhibits strongly **policy-specific temporal sensitivity**; LingBot-VA degrades sharply at L_1 and deteriorates further toward L_3, whereas the other

TABLE I

AVERAGE SUCCESS RATE (%) ACROSS TEN ROBOTWIN TASKS. MEAN RETENTION R_p DENOTES THE AVERAGE FRACTION OF BASE PERFORMANCE RETAINED ACROSS THE NINE ILLUMINATION PERTURBATIONS. HIGHER IS BETTER.

Policy	Base	Dark	Warm	Cold	Random	Spot	Grating	Disco	Strobe	Moving Shadow	Mean Ret. (%)
RDT [34]	42.6	38.3	41.3	41.6	41.7	27.6	18.2	18.9	41.7	42.3	81.3
OpenVLA-OFT [35]	46.7	40.8	46.6	35.6	47.7	38.6	40.3	31.0	43.7	47.0	88.3
FastWAM [36]	54.3	12.3	28.1	27.3	37.8	17.5	0.7	2.4	44.7	54.4	46.1
$\pi_{0.5}$ [37]	73.9	72.9	73.5	72.8	71.9	49.4	43.4	70.7	74.1	72.7	90.4
LingBot-VA [38]	48.7	45.4	45.2	50.7	45.7	40.8	35.1	23.0	13.5	47.0	79.0

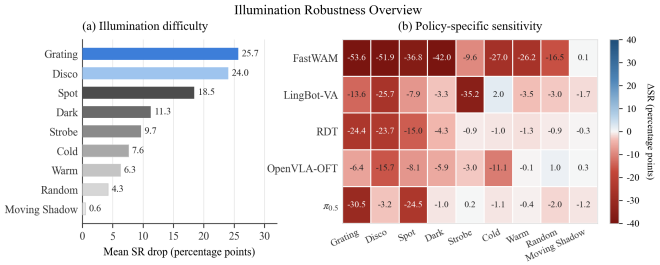

Fig. 3. **Illumination robustness overview.** (a) Mean success-rate drop by perturbation under the canonical benchmark settings, averaged across policies. (b) Per-policy, per-condition ΔSR relative to Base. Spatial and composite perturbations (Grating, Disco, Spot) are most damaging on average, while sensitivity rankings vary substantially across policies (e.g., FastWAM vs. LingBot-VA).

Fig. 4. End-effector trajectories on the Adjust Bottle task under Base (left) and Disco Lighting (right). Colors indicate normalized progress along each projected path.

policies show substantially smaller changes.

These curves distinguish sensitivity to perturbation onset from stability under severity escalation. For example, $\pi_{0.5}$ changes relatively little from L_1 to L_3 under Spot despite retaining a large performance gap from Base. Local reversals between adjacent severity levels occur for several policies. Given the finite number of closed-loop rollouts, we do not interpret these small reversals as evidence that stronger perturbations systematically improve performance; they may instead reflect sampling variability or nonlinear interactions between illumination and scene appearance. Overall, severity-response profiles are both policy- and mechanism-dependent and cannot be summarized by a single monotonic degradation model.

D. Real-World Evaluation

We finally examine whether illumination-induced failures observed in simulation also arise on a physical robot. Table III reports real-world results for $\pi_{0.5}$ and FastWAM on three representative tasks. Each policy–task–condition configuration is evaluated over 16 trials.

TABLE II

MEAN SUCCESS RATE (%) ACROSS FIVE POLICIES AND TEN TASKS UNDER BASE ILLUMINATION (L_0) AND INCREASING PERTURBATION SEVERITY. L_2 CORRESPONDS TO THE CANONICAL CONDITION IN TABLE I.

Perturbation	L_0 (Base)	L_1	L_2	L_3
Dark	53.2	46.6	41.9	38.3
Cold	53.2	49.5	45.6	44.4
Spot	53.2	35.6	34.8	34.2
Strobe	53.2	44.1	43.5	41.1

The trials reveal substantial illumination sensitivity on the physical platform. Under Base illumination, $\pi_{0.5}$ achieves success rates of 62.5–100.0% across the three tasks, and Warm illumination causes only moderate changes. In contrast, localized illumination, whether steady (Spot) or intermittent (Strobe), produces severe degradation. Spot reduces $\pi_{0.5}$ to at most 6.25% under the moderate setting and to 0% under the severe setting, while Strobe yields similarly low success rates. Dark exhibits stronger task dependence: $\pi_{0.5}$ remains relatively stable on Adjust Bottle under Dark-M, but degrades sharply on Beat Block Hammer and Stack Cube.

FastWAM shows an even sharper failure on Adjust Bottle: it succeeds in all Base trials and retains 87.5% SR under Warm illumination, but fails all trials under Dark, Spot, and Strobe. Its uniformly zero SR on Stack Cube should not be interpreted as illumination-induced degradation because the three-task mixed checkpoint already achieves 0% task success under Base. Qualitative inspection shows that the policy occasionally grasps a block but either fails to acquire it reliably or cannot complete the required stacking sequence. This result therefore reflects insufficient nominal task competence for the full success criterion, preventing a meaningful robustness comparison on this policy–task pair.

Table IV shows that Spot exhibits the clearest simulation–real correspondence among the evaluated perturbations. Spot reduces performance in all four valid policy–task pairs in both domains, with the chosen physical settings generally producing larger drops. Dark also causes substantial degradation in the physical experiments, although its magnitude is more task-dependent and differs from the simulated results.

The physical Strobe condition is implemented by intermittently switching the same localized light used for Spot. Its performance is comparable to steady Spot across the evaluated configurations. At the present sample size, and given the

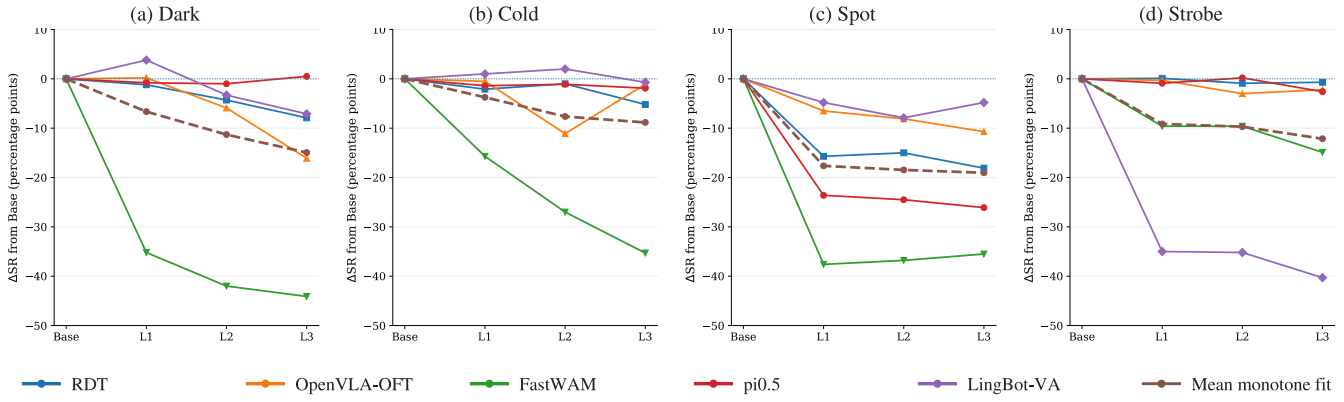

Fig. 5. Success-rate changes relative to Base (in percentage points), averaged across ten tasks, under increasing illumination severity for Dark, Cold, Spot, and Strobe.

TABLE III

REAL-WORLD SUCCESSFUL TRIALS OUT OF 16 UNDER CONTROLLED PHYSICAL ILLUMINATION. M AND S DENOTE MODERATE AND SEVERE SETTINGS, RESPECTIVELY. THE STROBE CONDITION IS IMPLEMENTED BY INTERMITTENTLY SWITCHING THE SAME LOCALIZED DIRECTIONAL LIGHT USED FOR SPOT.

Policy	Task	Base	Warm	Dark-M	Dark-S	Spot-M	Spot-S	Strobe-M	Strobe-S
$\pi_{0.5}$	Adjust Bottle	16/16	14/16	14/16	13/16	1/16	0/16	4/16	0/16
	Beat Block Hammer	14/16	15/16	1/16	0/16	0/16	0/16	1/16	0/16
	Stack Cube	10/16	12/16	8/16	0/16	0/16	0/16	0/16	0/16
FastWAM	Adjust Bottle	16/16	14/16	0/16	0/16	0/16	0/16	0/16	0/16
	Beat Block Hammer	16/16	16/16	6/16	7/16	0/16	0/16	0/16	0/16
	Stack Cube	0/16	0/16	0/16	0/16	0/16	0/16	0/16	0/16

TABLE IV

SIMULATION–REAL COMPARISON OF ILLUMINATION-INDUCED SUCCESS-RATE CHANGES (ΔSR, PERCENTAGE POINTS).

		Dark		Spot		Strobe*
Policy	Task	Sim	Real	Sim	Real	Real
$\pi_{0.5}$	Adjust Bottle	-2.0	-12.5	-62.0	-93.8	-75.0
	Beat Block	-11.0	-81.3	-15.0	-87.5	-81.3
FastWAM	Adjust Bottle	-70.0	-100.0	-73.0	-100.0	-100.0
	Beat Block	-34.0	-62.5	-36.0	-100.0	-100.0

Sim: L_2 for Dark and Spot; Real: moderate settings.
*Localized Strobe has no mechanism-matched simulated counterpart.
Stack Cube is a physical-only task with no simulated counterpart and is therefore omitted.

near-floor performance under Spot for several configurations, we observe no measurable additional degradation attributable to temporal switching. The physical study therefore does not isolate the temporal illumination mechanism, and the simulated global Strobe has no mechanism-matched physical counterpart in this evaluation.

Overall, the physical experiments provide representative evidence that selected illumination vulnerabilities identified in simulation can persist under real lighting changes. Because the physical study contains only 16 trials per configuration, small differences between nearby success rates should not be over-interpreted. We therefore regard the physical study as representative validation rather than a comprehensive simulation-to-real evaluation.

V. CONCLUSION AND LIMITATIONS

We introduced LumiBench, an illumination-centric diagnostic benchmark that organizes scene-level lighting shifts by their dominant physical mechanisms. Paired nominal–perturbed evaluation isolates illumination-induced changes, while severity sweeps characterize how policy performance evolves as representative perturbations strengthen.

Across five policies and ten manipulation tasks, illumination robustness depends strongly on the interaction among policy, task, and perturbation mechanism. Spatially structured and composite conditions produce the largest aggregate degradation, while severity sweeps reveal distinct response regimes, including progressive, onset-dominated, and policy-specific degradation. These findings demonstrate that neither nominal performance nor evaluation under a single illumination shift is sufficient to characterize illumination robustness.

Controlled physical experiments further reproduce pronounced Spot sensitivity and task-dependent degradation under Dark conditions, providing representative evidence that illumination vulnerabilities observed in simulation can persist on real systems. Because the physical Strobe implementation couples localized spatial illumination with temporal switching, it does not isolate an additional temporal effect beyond Spot. Future work will extend LumiBench through

mechanism-isolated physical validation, additional simulators and robotic platforms, and evaluation of interactions between illumination and other visual shifts.

REFERENCES

[1] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, 2025.

[2] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[3] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[4] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[5] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[6] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[7] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[8] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[9] W. Pumacay, I. Singh, J. Duan, R. Krishna, J. Thomason, and D. Fox, "The colosseum: A benchmark for evaluating generalization for robotic manipulation," *arXiv preprint arXiv:2402.08191*, 2024.

[10] S. Fei, S. Wang, J. Shi, Z. Dai, J. Cai, P. Qian, L. Ji, X. He, S. Zhang, Z. Fei *et al.*, "Libero-plus: A progressive robustness benchmark for visual-language-action models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[11] A. Xie, L. Lee, T. Xiao, and C. Finn, "Decomposing the generalization gap in imitation learning for visual robotic manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[12] Z. Zhang, Z. Li, B. Rahmati, R. H. Yang, Y. Ma, A. Rasouli, S. Pakdamansavoji, Y. Wu, L. Zhang, T. Cao *et al.*, "Do world action models generalize better than vlas? a robustness study," *arXiv preprint arXiv:2603.22078*, 2026.

[13] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, 2020.

[14] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Conference on robot learning*, 2020.

[15] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *Advances in Neural Information Processing Systems*, 2023.

[16] G. Zhou, V. Dean, M. K. Srirama, A. Rajeswaran, J. Pari, K. Hatch, A. Jain, T. Yu, P. Abbeel, L. Pinto *et al.*, "Train offline, test online: A real robot learning benchmark," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[17] N. Khargonkar, S. H. Allu, Y. Lu, B. Prabhakaran, Y. Xiang *et al.*, "Scenereplica: Benchmarking real-world robot manipulation by creating replicable scenes," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[18] R. Garcia, S. Chen, and C. Schmid, "Towards generalizable vision-language robotic manipulation: A benchmark and llm-guided 3d policy," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[19] Y. Mu, T. Chen, Z. Chen, S. Peng, Z. Lan, Z. Gao, Z. Liang, Q. Yu, Y. Zou, M. Xu *et al.*, "Robotwin: Dual-arm robot benchmark with generative digital twins," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[20] T. Chen, Z. Chen, B. Chen, Z. Cai, Y. Liu, Z. Li, Q. Liang, X. Lin, Y. Ge, Z. Gu *et al.*, "Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation," *arXiv preprint arXiv:2506.18088*, 2025.

[21] T. Chen, Y. Chen, Z. Li, J. Tang, K. Su, H. Lu, W. Wan, B. Chen, S. Liu, H. Yan *et al.*, "Robodojo: A unified sim-and-real benchmark for comprehensive evaluation of generalist robot manipulation policies," *arXiv preprint arXiv:2607.04434*, 2026.

[22] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[23] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "Vima: General robot manipulation with multimodal prompts," 2023.

[24] Z. Yang, W. Zhang, X. Chen, W. Song, X. Wang, Y. Kang, W. Chen, L. Wang, R. Xu, and X. Chu, "Dreamtrajectory: Trajectory-guided action generation with world model alignment for mobile manipulation," *arXiv preprint arXiv:2608.01381*, 2026.

[25] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[26] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on robot learning*, 2023.

[27] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.

[28] B. Cai, Q. Liang, J. Li, S. Weng, Z. Zhang, T. Lin, X. Chen, W. Zhang, J. Mao, W. Xu *et al.*, "Beyond viewpoint generalization: What multi-view demonstrations offer and how to synthesize them for robot manipulation?" *arXiv preprint arXiv:2603.26757*, 2026.

[29] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 2017.

[30] E. Rusak, L. Schott, R. S. Zimmermann, J. Bitterwolf, O. Bringmann, M. Bethge, and W. Brendel, "A simple way to make neural networks robust against diverse image corruptions," in *European conference on computer vision*, 2020.

[31] Y. Zhu, Z. Wang, J. Yang, J. Xie, J. Yu, J. Gu, and B. Wang, "Toward visually realistic simulation: A benchmark for evaluating robot manipulation in simulation," *arXiv preprint arXiv:2605.06311*, 2026.

[32] X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani *et al.*, "Evaluating real-world robot manipulation policies in simulation," *arXiv preprint arXiv:2405.05941*, 2024.

[33] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang *et al.*, "Sapien: A simulated part-based interactive environment," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[34] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "Rdt-1b: a diffusion foundation model for bimanual manipulation," in *International Conference on Learning Representations*, 2025.

[35] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: Optimizing speed and success," *arXiv preprint arXiv:2502.19645*, 2025.

[36] T. Yuan, Z. Dong, Y. Liu, and H. Zhao, "Fast-wam: Do world action models need test-time future imagination?" *arXiv preprint arXiv:2603.16666*, 2026.

[37] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[38] L. Li, Q. Zhang, Y. Luo, S. Yang, R. Wang, F. Han, M. Yu, Z. Gao, N. Xue, X. Zhu *et al.*, "Causal world modeling for robot control," *arXiv preprint arXiv:2601.21998*, 2026.