DarkQA: Benchmarking Vision-Language Models on Visual-Primitive Question Answering in Low-Light Indoor Scenes

Yohan Park[1], Hyunwoo Ha[2], Wonjun Jo[2], and Tae-Hyun Oh[1]

***Abstract*— Vision Language Models (VLMs) are increasingly adopted as central reasoning modules for embodied agents. Existing benchmarks evaluate their capabilities under ideal, well-lit conditions, yet robust 24/7 operation demands performance under a wide range of visual degradations, including low-light conditions at night or in dark environments–a core necessity that has been largely overlooked. To address this underexplored challenge, we present DarkQA, an open-source benchmark for evaluating perceptual primitives under multi-level low-light conditions in embodied scenarios. DarkQA evaluates single-view egocentric observations across controlled degradation levels, isolating low-light perceptual failures before they are entangled with complex embodied tasks. The benchmark contains 9.4K deterministically generated and verifiable question–image pairs spanning five visual-primitive families. A key design feature of DarkQA is its physical fidelity: visual degradations are modeled in linear RAW space, simulating physics-based illumination drop and sensor noise followed by an ISP-inspired rendering pipeline; we further validate the synthesis against real paired low-light camera data. We evaluate representative VLMs and Low-Light Image Enhancement (LLIE) preprocessing methods. Results show consistent VLM degradation under low illumination and sensor noise, while LLIE provides severity-dependent but unstable recovery. We demonstrate the utility of DarkQA by evaluating a wide range of state-of-the-art VLMs and Low-Light Image Enhancement (LLIE) models, and systematically reveal VLMs' limitations when operating under these challenging visual conditions. Our code and benchmark dataset will be released upon acceptance. Project website: `https://darkqa-benchmark.github.io`**

I. INTRODUCTION

Advances in vision-language models (VLMs) have significantly enhanced robotic perception and decision-making, supporting semantic scene understanding [1], spatial reasoning [2], and vision-language-action (VLA) policies [3], [4]. However, household robots are often intended for 24/7 operation, which means they will frequently encounter low-light scenarios, such as nighttime, entering dark rooms or power blackouts. Such low-light conditions are especially problematic because they globally weaken the visual-primitives—the visual evidence available to VLMs to tackle embodied tasks. In embodied pipelines, these front-end perception failures can affect downstream tasks (*e.g.*, memory update, spatial or affordance reasoning). Representative egocentric and embodied QA benchmarks are not designed to systematically evaluate VLMs under controlled low-light visual degradation [5], [6], [7], [8], [9], [10], [11], [12]. Illumination remains an orthogonal and under-controlled evaluation variable, even though robust perception under low illumination is not an edge case but a core necessity for real-world embodied deployment [13]. Accordingly, benchmarks that explicitly stress-test VLMs on visual-primitives under controlled low illumination in embodied scenarios are essential to quantify visual robustness before it is confounded with navigation, memory, or action failures. Nevertheless, acquiring large-scale, real-world low-light images with clean, paired annotations—ideally with corresponding well-lit reference views—is challenging and costly, which has hindered the construction of such benchmarks. As a result, existing benchmarks have largely overlooked systematic evaluation of VLM-based reasoning and perception under degraded illumination, limiting their ability to predict real-world robustness.

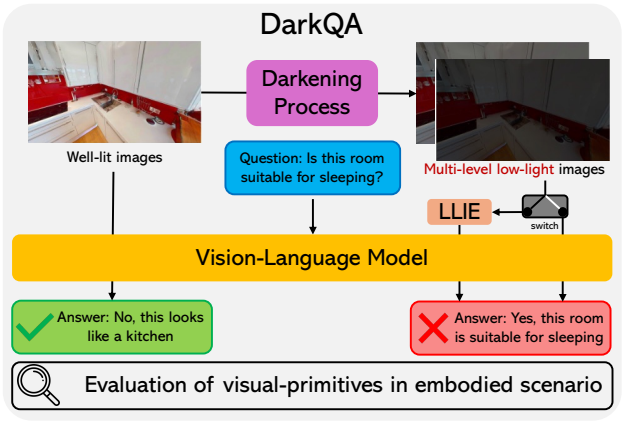


Fig. 1. **Illustration of the DarkQA benchmark.** We present DarkQA, a new benchmark that evaluates VLM robustness in visual-primitives under low-light conditions in embodied scenario. DarkQA assesses VLM performance under two distinct conditions: clean, well-lit inputs (L0) and a multi-level ladder of physics-based low-light images (L1-L5). Furthermore, the benchmark examines the effect of applying Low-Light Image Enhancement (LLIE) models as a pre-processing step.

To fill this evaluation void, we present DarkQA, an open-source benchmark for evaluating VLM robustness in visual-primitive question answering (QA) from egocentric indoor observations across controlled low-light levels, from moderate degradation to extreme stress-test regimes. The design of DarkQA is primarily grounded in a physically based formulation, where all visual degradations are modeled at the RAW sensor data level (or in linear RGB space). This follows the physics of illumination and sensor noise

Corresponding author: Tae-Hyun Oh.
[1]Yohan Park and Tae-Hyun Oh are with Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea (e-mail: john.a.park@kaist.ac.kr, taehyun.oh@kaist.ac.kr).
[2]Hyunwoo Ha and Wonjun Jo are with Pohang University of Science and Technology (POSTECH), Pohang 37673, South Korea (e-mail: hyunwooha@postech.ac.kr, jo1jun@postech.ac.kr).

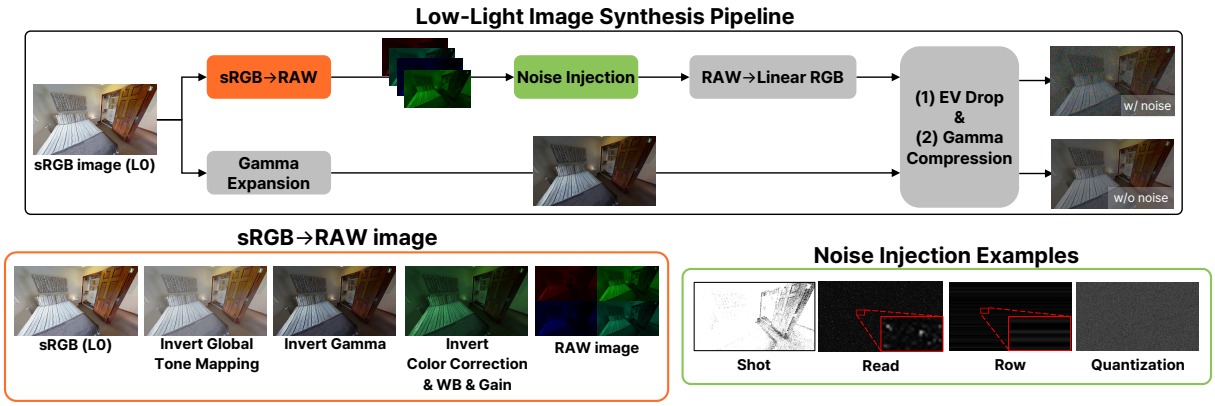


Fig. 2. **Low-light synthesis pipeline with disentangled illumination and noise factors.** To generate controlled low-light inputs for our benchmark, we adopt an ISP-inspired unprocessing and noise formulation from prior work [24], [25]. Crucially, we produce *paired* variants for each original image to disentangle failure sources in VLM-based QA: (a) a physics-based branch (top) that unprocesses sRGB to Bayer RAW, injects four noise components in RAW, and then applies EV drop and gamma compression; and (b) a noise-free branch (bottom) that applies the same EV drop in linear RGB without noise injection. This paired design enables separate evaluation of performance degradation due to illumination reduction versus sensor noise. The bottom-left panel summarizes the sRGB→RAW unprocessing steps, and the bottom-right panel visualizes the four noise components (shot, read, row-pattern, and quantization noise) as independent signals. The small red boxes in the read and row noise examples indicate zoomed-in crops for visualization.

to realistically simulate real-world Image Signal Processing (ISP) scenarios. Moreover, to ensure benchmark integrity and prevent potential data contamination [14], all Question Answering (QA) pairs are deterministically generated via rule-based procedure, rather than depending on commodity VLM services. QA generation results in a family of queries targeting perceptual primitives, including from simple object recognition (*e.g.*, "Is there a cushion in the image?") to affordance reasoning (*e.g.*, "I want to sleep, is this room suitable for this?").

DarkQA provides 9.4k question–image pairs, a standardized evaluation protocol, and a public codebase to reproduce our low-light degradation pipeline. Our DarkQA benchmarks a diverse set of vision–language models (VLMs), including both open- and closed-source systems [15], [16], [17], [18], [19]. We also evaluate four low-light image enhancement (LLIE) models [20] [21], [22], [23] as preprocessing baselines. Our evaluation yields two observations. First, all tested VLMs show a clear performance decline as the images degrade. Second, LLIE preprocessing is method- and severity-dependent: it can improve QA accuracy at some degradation levels, but is not uniformly beneficial and does not recover well-lit performance. Together, these results show that current VLMs remain brittle under low-light corruption, and that perceptual enhancement alone is insufficient as a general solution, motivating robustness-oriented evaluation and method development.

II. Related Work

A. Egocentric Question Answering Benchmarks

Egocentric question answering evaluates visual-language reasoning from first-person observations. Prior benchmarks study large-scale egocentric video understanding [5], episodic-memory QA [6], task-level reasoning [7], long-form video QA [8], and scene-text-aware assistance [9]. However, none of these benchmarks evaluates VLMs under controlled dark or low-light visual degradation. Unlike prior work [26], our benchmark evaluates VLM robustness under controlled, multi-level synthesized low-light degradation.

Embodied QA benchmarks such as EmbodiedQA [10], ScanQA [11], and OpenEQA [12] are closely related, as they evaluate agents that answer questions about embodied or 3D environments. Yet they focus on navigation, 3D scene QA, or environment-level memory rather than isolating low-light visual robustness. In contrast, *DarkQA* evaluates egocentric indoor QA under controlled low-light levels, directly assessing VLM robustness to degraded visual-primitives.

B. Handling Low-Light Images

Recent research addresses low-light visual perception in two directions. The first improves recognition under low illumination for task-specific vision problems, such as depth estimation, object detection, or pose estimation [27], [28], [29]. While effective, these studies do not evaluate VLM robustness for visual-primitive QA from egocentric indoor observations under controlled low-light degradation. The second direction is low-light image enhancement (LLIE), which aims to improve brightness, contrast, and detail visibility for human perception or downstream models. Representative LLIE models include DarkIR [20], RetinexFormer [21], ZeroDCE [22], and RUAS [23]. Although LLIE improves visual quality, its effect on VLM-based reasoning over low-light egocentric inputs remains underexplored. We therefore evaluate whether LLIE preprocessing helps VLMs answer visual-primitive questions in dark environments.

III. DarkQA: Dataset Construction

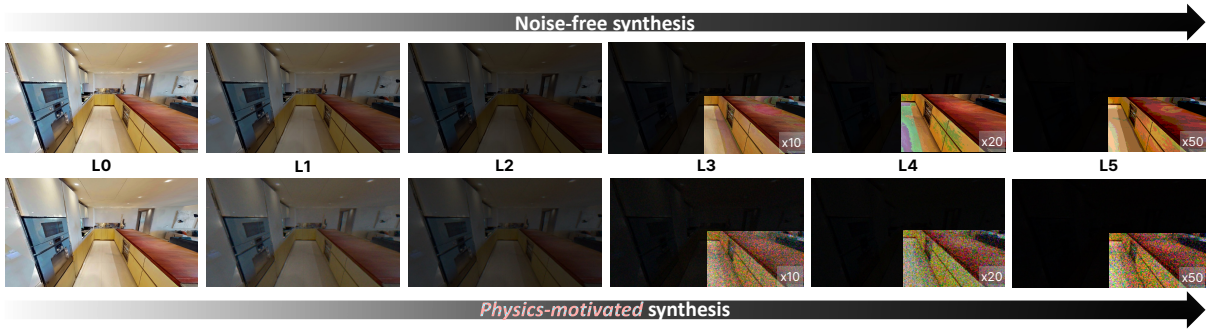


Fig. 3. **Example low-light image synthesization.** Synthesized low-light image examples across degradation levels L0–L5. The top row shows EV drop only, while the bottom row shows EV drop combined with noise injection. The lower-right insets show $1/4$-image crops with pixel intensities amplified for visibility; the numbers (×10, ×20, ×50) indicate the amplification factor.

AND QA PAIR GENERATION

Our DarkQA is designed to evaluate VLMs' recognition of core visual primitives from a single image-question pair under controlled low-light conditions. However, acquiring real-world low-light images with clean, paired annotations is challenging. To address this, we synthesize low-light images from the well-established indoor scene dataset (*i.e.*, HM3D-Sem [30]). This section describes the low-light image synthesis for VLM input (Sec. III-A) and the visual-primitive QA dataset construction process (Sec. III-B). A key feature of our work is a dataset construction pipeline designed for high reproducibility and expandability.

A. Low-Light Image Synthesis for Benchmark Inputs

Low-light images suffer from two distinct physical degradations. First, the reduced photon count leads to a fundamental loss of signal, which we term illumination degradation (*i.e.*, exposure-value (EV) drop). Second, this weakened signal yields a low Signal-to-Noise Ratio (SNR), as sensor noise (*e.g.*, shot, read, pattern, and quantization noise) becomes dominant relative to the remaining signal [25]. To reproduce these conditions for benchmark inputs, we design a physics-based low-light synthesis pipeline. Specifically, across multiple degradation severities (L1–L5, increasing severity), we synthesize two *paired* low-light variants per original image: (i) A noise-free EV-drop variant and (ii) a physics-motivated variant with level-dependent sensor noise injection in the RAW domain, as in Fig. 3. This design enables disentangling the respective impacts of illumination degradation and sensor noise on perceptual peformance of VLMs.

1) Noise-free low-light image synthesis: Exposure-value (EV) drop is applied at linear RGB space after decoding sRGB images as shown in the lower branch of low-light image synthesis pipeline depicted in Fig. 2.

Decoding to linear RGB. Since EV is physically defined with respect to scene-linear irradiance, we conduct linearization before applying exposure scaling. Hence, we first approximate linearization using gamma expansion. Let x_{sRGB} represent a sRGB pixel value in an input image and x_{lin} its linear form. Following [31], [24], we compute

$$x_{\mathrm{lin}} = \left(\max(x_{\mathrm{sRGB}}, \epsilon)\right)^{2.2}, \tag{1}$$

where $\epsilon = 10^{-8}$ ensures numerical stability.

Exposure scaling. Next, let $\Delta\mathrm{EV}$ denote the absolute change in exposure value. Reducing the exposure by $\Delta\mathrm{EV}$ scales the x_{lin} by $2^{-\Delta\mathrm{EV}}$. The exposure-scaled pixel value is

$$x'_{\mathrm{lin}} = 2^{-\Delta\mathrm{EV}}\, x_{\mathrm{lin}}. \tag{2}$$

Re-encoding to sRGB. Finally, the exposure-scaled pixel value x'_{lin} is mapped back to sRGB via gamma encoding:

$$x'_{\mathrm{sRGB}} = (x'_{\mathrm{lin}})^{1/2.2}. \tag{3}$$

We standardize an degradation levels L1–L5 with $\Delta\mathrm{EV} \in \{2.0, 4.0, 6.0, 7.5, 9.0\}$, respectively (L0 is the original).

2) Physics-motivated low-light image synthesis: We synthesize realistic low-light images using a physics-based pipeline that combines ISP inversion/forward pass [24] and raw-domain noise modeling [25]. The process is shown in the upper branch of low-light image synthesis pipeline of Fig. 2. Realism of our physics-based low-light image synthesis pipeline is discussed in Sec. IV-D.

Unprocessing (sRGB → RAW). We first normalize an 8-bit sRGB image $\mathbf{I} \in \{0, \ldots, 255\}^{H\times W\times 3}$, where H and W denote the image height and width, respectively, to

$$\mathbf{I}_{\mathrm{sRGB}} = \frac{\mathbf{I}}{255} \in [0,1]^{H\times W\times 3}.$$

To obtain a camera-linear RAW image from $\mathbf{I}_{\mathrm{sRGB}}$, we invert the ISP following [24]. We denote the unprocessing operator by $u(\cdot)$, and express the resulting Bayer RAW mosaic as

$$\mathbf{B} = u(\mathbf{I}_{\mathrm{sRGB}}), \tag{4}$$

where $\mathbf{B} \in [0,1]^{\frac{H}{2}\times\frac{W}{2}\times 4}$. The unprocessing operator $u(\cdot)$ consists of five steps: (i) inverse tone mapping, (ii) gamma expansion, (iii) RGB→Camera color correction with sampled matrix $\mathbf{M}_{\mathrm{rgb}\to\mathrm{cam}}$, (iv) inversion of white-balance/brightness gains with highlight preservation, and (v) mosaic extraction into RGGB Bayer representation. This restores a scene-referred signal where noise statistics are defined with respect

to photon counts and sensor readout electronics, not post-ISP perceptual tone curves.

Noise formation in RAW. Following the physics-based formation model of [25], we inject four noise components into the camera-linear RAW signal. Let $\mathbf{B}$ denote the clean, mosaiced RAW image obtained from unprocessing. After converting $\mathbf{B}$ from normalized units to the sensor's ADU domain, we sample a system gain K log-uniformly from $[0.1, 6.0]$. The noisy RAW image is then expressed as

$$\mathbf{B}_{\text{noisy}} = \mathcal{N}_4 \circ \mathcal{N}_3 \circ \mathcal{N}_2 \circ \mathcal{N}_1(\mathbf{B}, K), \tag{5}$$

where $\mathcal{N}_i$ denotes the i-th noise operator mapping a Bayer RAW tensor and system gain K to a Bayer RAW tensor described below.

(1) Photon shot noise. Photon arrival is discrete and stochastic. For each pixel, the number of photoelectrons N follows $N \sim \text{Poisson}(\lambda)$ where λ is proportional to scene irradiance. To simulate extreme low-light capture, we apply an ISO amplification ratio $r \in [100, 300]$: (i) reduce the signal by r (low-light capture), (ii) add Poisson noise, (iii) amplify back by r using sensor gain. This preserves the characteristic of low-photon-count statistics while allowing the final output brightness to be controlled independently via the EV drop.

(2) Read noise. Readout electronics introduce an additive noise term N_{read}. We model it using a Tukey–λ distribution with a channel-wise DC offset (color bias). The scale parameter σ_{TL} grows log-linearly with the system gain K:

$$\log \sigma_{\text{TL}} = a_{\text{TL}} \log K + b_{\text{TL}} + \epsilon,$$

capturing the heavy-tailed distribution observed under extreme low-light [25].

(3) Row noise. Line-wise variations in the readout circuitry produce banding artifacts. Each row i receives a shared offset $n_r^{(i)} \sim \mathcal{N}(0, \sigma_r^2)$, where σ_r also scales log-linearly with K.

(4) Quantization noise. Analog-to-digital conversion introduces rounding error N_q modeled as $N_q \sim \mathcal{U}(-0.5, 0.5)$, where $\mathcal{U}$ represents a uniform distribution on $[-0.5, 0.5]$, assuming a standard unit (1 ADU) quantization step.

Simplified ISP (RAW → sRGB). Converting RAW to sRGB is an inverse operation of unprocessing: (i) white balance with sampled gains, (ii) bilinear demosaicing from RGGB Bayer to RGB, (iii) color correction using $\mathbf{M}_{\text{cam}\to\text{rgb}}$, (iv) EV drop by ΔEV in linear space (multiplying intensities by $2^{-\Delta\text{EV}}$) to match the target degradation levels L1–L5, (v) gamma compression, and (vi) quantization to 8-bit sRGB.

B. Dataset Construction

We build the dataset for evaluation upon a representative subset of 52 scenes from HM3D-Sem [30], selected for diversity and semantic richness. For each scene, we record a human-demonstrated navigation trajectory that systematically explores the environment to maximize spatial coverage. To generate the ground-truth QA pairs, we uniformly subsample the trajectory and select keyframes at a fixed time interval (*e.g.*, one frame every 2,s), rendering their geometric and semantic modalities (*e.g.*, RGB, depth, segmentation). We

Algorithm 1 Deterministic procedure for QA generation

Require: Scene set $\mathcal{S}$; frames $\mathcal{F}_s$ for each $s \in \mathcal{S}$
Ensure: QA pairs $\mathcal{Q}$ with ground-truth answers

Definitions:
$\Omega_f = \{1, \dots, W\} \times \{1, \dots, H\}$: Pixel grid of frame f
$M_i^f \in \{0,1\}^{H \times W}$: Mask for segment i in frame f
$\mathcal{A}_i \in \mathbb{R}^d$: Attribute vector for segment i (semantic class, color, depth, area, bbox)
$\Phi_f = \{\mathcal{A}_i\}_{i=1}^{N_f}$: Frame statistics (all segment attributes)
r_f: Room type label for frame f
$\mathcal{C}_f \subseteq \{1, 2, 3, 4, 5\}$: Viable question families of frame f

Generate QA from Frames
$\mathcal{Q} \leftarrow \emptyset$
for $f \in \bigcup_{s \in \mathcal{S}} \mathcal{F}_s$ **do** ▷ Process each frame exactly once
  — 1. Extract Statistics
  Load $I^f_{\text{RGB}}, I^f_{\text{depth}}, I^f_{\text{sem}}, I^f_{\text{over}}$
  for $i \in \text{Segments}(I^f_{\text{over}})$ **do**
    $M_i^f(x, y) \leftarrow \mathbf{1}[(x, y) \in \Omega_f \wedge I^f_{\text{over}}(x, y) = i]$
    $\mathcal{A}_i \leftarrow \text{ComputeStats}(M_i^f, I^f_{\text{RGB}}, I^f_{\text{depth}}, I^f_{\text{sem}})$
  end for
  $\Phi_f \leftarrow \{\mathcal{A}_i : \forall i\}$ ▷ Collect stats for frame f
  — 2. Generate QA
  $r_f \leftarrow \text{ClassifyRoom}(\Phi_f)$
  $\mathcal{C}_f \leftarrow \text{Survey}(\Phi_f, r_f)$ ▷ Find viable question types
  for $k \in \mathcal{C}_f$ **do** ▷ Generate all viable questions
    $(\text{q}, a) \leftarrow \text{Rule}_k(\Phi_f, r_f)$
    $\mathcal{Q} \leftarrow \mathcal{Q} \cup (\text{q}, a)$
  end for
end for
return $\mathcal{Q}$

then use Algorithm 1 as deterministic procedure to automatically generate QA pairs from the common interface ($I^f_{\text{RGB}} \in \mathbb{R}^{H \times W \times 3}, I^f_{\text{depth}} \in \mathbb{R}^{H \times W}, I^f_{\text{sem}} \in \mathbb{R}^{H \times W}, I^f_{\text{over}} \in \mathbb{R}^{H \times W}$) (RGB image, depth map, semantic label map, and over-segmentation map, respectively.), which can be applied to any dataset that can be mapped to the interface for each frame. This approach ensures each question has a single verifiable answer by filtering ambiguities (*e.g.*, tiny objects), requires no manual QA annotation beyond the source annotations, and avoids potential data contamination by not using commodity VLM services. This entire process is fully reproducible.

Algorithm 1 operates in two stages: frame-statistics extraction (Stage 1) and QA generation (Stage 2). In Stage 1, we cache the frame statistics Φ_f required for Stage 2. Each frame f is represented as a quadruple $f = (I^f_{\text{RGB}}, I^f_{\text{depth}}, I^f_{\text{sem}}, I^f_{\text{over}})$. For each segment mask M_i^f obtained from I^f_{over}, $\text{ComputeStats}(M_i^f, I^f_{\text{RGB}}, I^f_{\text{depth}}, I^f_{\text{sem}})$ extracts an attribute vector $\mathcal{A}_i$ summarizing the segment's information. The collection of all segment attributes forms the frame statistics $\Phi_f = \{\mathcal{A}_i\}_{i=1}^{N_f}$.

In Stage 2, $\text{ClassifyRoom}(\Phi_f)$ assigns a room label r_f using deterministic object-signature rules, such as beds for

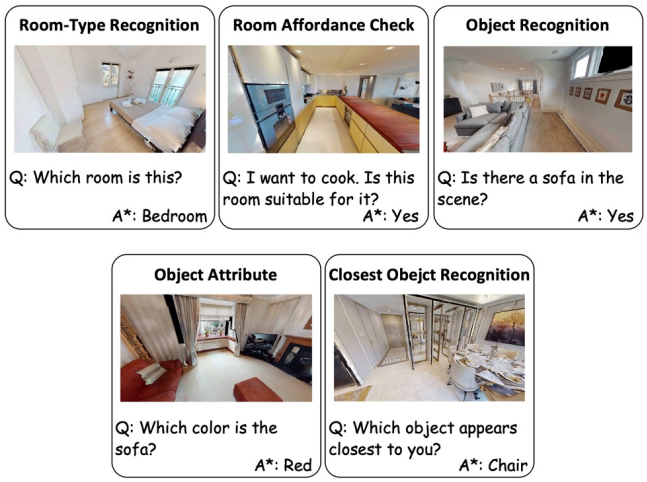


Fig. 4. **Question family of our DarkQA benchmark.** Five DarkQA question categories with examples. DarkQA encompasses questions asking room-type recognition, room affordance check, object recognition, object attribute.

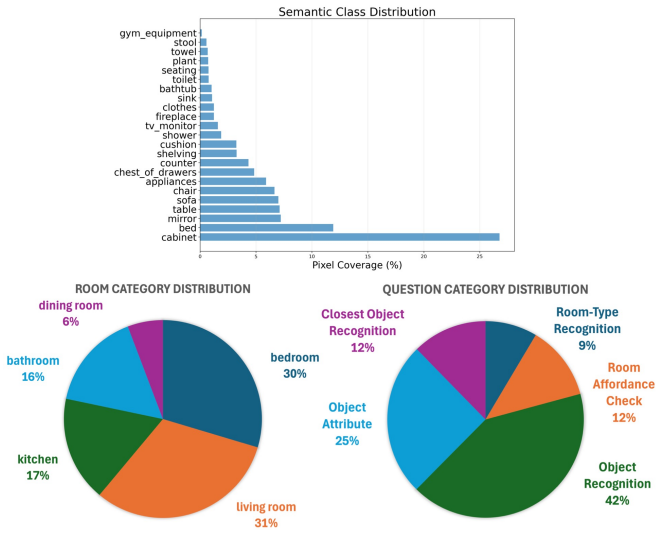


Fig. 5. **Statistics of our DarkQA benchmark.** Dataset statistics, including semantic-class coverage, room-category distribution, and question-category distribution.

bedrooms, toilets or showers for bathrooms, etc. Given Φ_f and r_f, $\mathrm{Survey}(\Phi_f, r_f)$ selects the viable question families $\mathcal{C}_f$, and each $\mathrm{Rule}_k : (\Phi_f, r_f) \mapsto (q, a)$ instantiates the corresponding question template and ground-truth answer, where k indexes the five families in Fig. 4. A rule is applied only when its required evidence is available and unambiguous in Φ_f. For example, consider the "Closest Object Recognition" question in Fig. 4. Object-level statistics are first extracted. The QA generation pipeline validates two conditions: (i) at least two non-structural, non-quasi-2D object instances with valid depth measurements exist, and (ii) the depth gap between top-two closest objects exceeds a minimum threshold to ensure perceptual validity. If satisfied, the closest object is determined as the ground-truth answer. In this example, "chair" is identified as the closest object.

This pipeline generates five question families targeting visual-primitives for embodied operation: *Room-Type Recognition*, *Room Affordance Check*, *Object Recognition*, *Object Attribute*, and *Closest Object Recognition*. The examples for each family are provided in Fig. 4. These categories are chosen to cover atomic visual evidence commonly required for embodied operation. Evaluating these primitives separately isolates low-light perceptual failures before they are entangled with complex downstream embodied tasks.

C. Dataset Statistics

Our DarkQA comprises 52 scenes selected from HM3D-Sem, yielding 3,911 frames at 1440×2560 resolution with $\sim$9.4K QA pairs. Fig. 5 shows that the dataset exhibits semantic class and room category distributions that are representative of typical residential environments. The semantic annotation covers 23 non-structural object classes, with the most prevalent being cabinet, bed, mirror, and table taking up about 53%. Room category distribution reflects the natural spatial composition of household scenes. The question distribution across the five question families shows moderate imbalance, with frequencies determined by the geometric and semantic constraints of our rule-based QA generation pipeline and subsequent validation through human sanity checks to ensure answer correctness.

IV. EXPERIMENTS

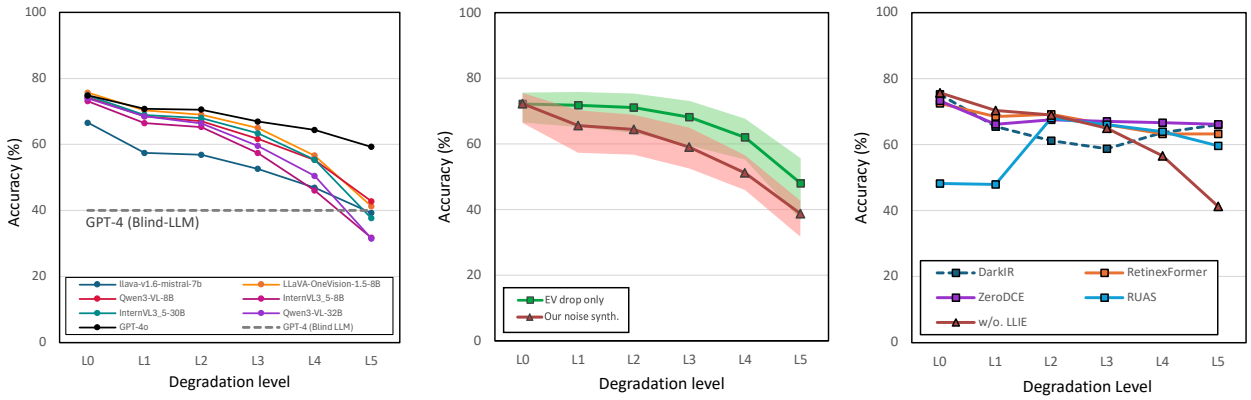
In this section, we describe our experimental settings and provide quantitative evaluation results of various VLMs on DarkQA, along with VLMs performance on DarkQA, validation of physics-motivated low-light synthesis pipeline, and open-ended question anwering evaluation.

A. Experimental Setup

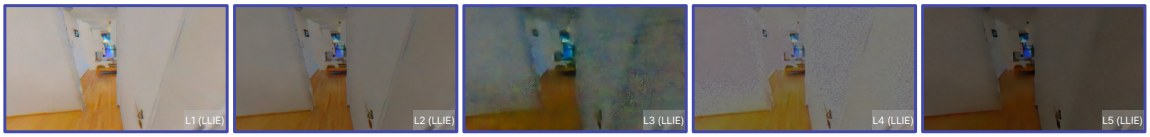
We evaluate DarkQA on both VLMs and text-only LLMs (blind LLMs). For each keyframe and degradation condition, we present a single question together with a fixed, small set of candidate answers (room-type labels, object classes, color names, or a candidate list for closest objects). VLMs receive the image and the question–choice template, whereas blind LLMs see only the textual question and choices. Each question is thus cast as a multiple-choice problem, and models are instructed to output exactly one answer from the choices. This constrains the response space, avoids ambiguities in free-form generation, and enables exact-match scoring.

B. Baseline Models

Blind LLMs. We set the scenario of blind agents that produces an answer based on the question that requires visual information to answer [12]. Even though our DarkQA focuses on the VLM's behavior according to illumination change and noise injection, we use the result of blind LLMs to catch the possible bias of our dataset while also testing how well the questions may be answered with an assumption of indoor environments. For the LLM choice, we report the results of GPT-4 [32] and LLaMA-3.1-8B [33].



(a) Model-wise comparison (b) Impact of noise modeling (c) Impact of LLIE pre-processing



(d) Image samples enhanced with LLIE model

Fig. 6. **Summary of the evaluation results on our DarkQA.** *Degradation level* indicates the severity of low-light corruption: $L0$ corresponds to the original (well-lit) input, and higher levels ($L1 \rightarrow L5$) denote progressively darker (lower-illumination) inputs. We evaluate a range of open-source VLMs (LLaVA [15], [16], InternVL [17], and Qwen-VL [18] series, 7B–32B). The shaded regions in (b) denote the minimum–maximum accuracy across models at each degradation level. **(a) Model-wise comparison. (b) Impact of noise modeling. (c) Impact of LLIE pre-processing. (d) Image samples enhanced by the DarkIR [20] model.** We include GPT-4 as a Blind-LLM baseline (evaluated without vision; gray dashed line) and GPT-4o [19] as an upper-bound reference (black line).

VLMs. We evaluate a range of VLMs across different parameter scales. For 7–8B models, we report results for LLaVA-1.6-7B [15], LLaVA-OneVision-8B [16], InternVL3.5-8B [17], and Qwen3-VL-8B [18]. For larger-scale models ($\geq$ 30B), we additionally evaluate InternVL3.5-30B[17] and Qwen3-VL-32B [18] using the same respective series. Finally, we include GPT-4o [19] as an upper bound.

LLIE model. For VLM evaluation on LLIE-enhanced low-light images, we use DarkIR [20], RetinexFormer [21], ZeroDCE [22], and RUAS [23].

C. VLMs performance on DarkQA

Model-specific accuracy. Fig. 6-(a) provides a detailed comparison of the performance trends across individual VLMs under noisy inputs without LLIE preprocessing. While the specific degradation curves vary slightly across each models, the overall trend is a largely similar decline in accuracy as low-light conditions intensify. Although the commodity service GPT-4o consistently demonstrates the highest performance, it also shows performance degradation under low-light conditions. Furthermore, we observe an interesting point: at the most severe low-light level (L5), some VLMs achieve accuracy lower than that of GPT-4 (Blind-LLM baseline), which operates solely on textual input without any visual information. This indicates that for images under extreme degradation, the models are unable to effectively utilize these visual information, leading to a poorer understanding of semantic information compared to relying purely on language priors.

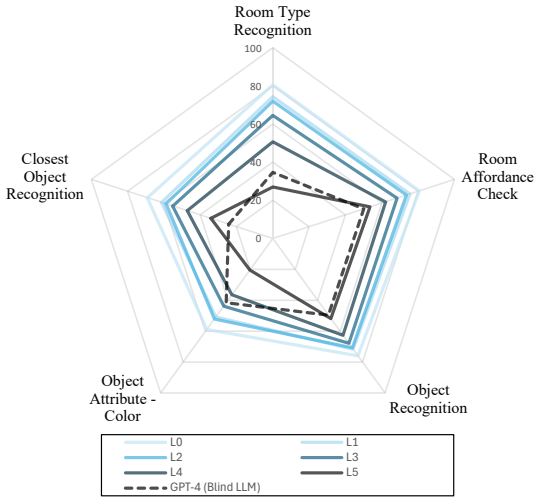


Fig. 7. **Question-wise accuracy.** We plot VLM accuracy across different question types under increasing low-light degradation, where darker lines indicate more severe degradation and the gray dashed line denotes the GPT-4 Blind-LLM baseline. We observe significant drops in "Room Type Recognition" and "Object Attribute – Color," where VLM performance falls below the GPT-4 Blind-LLM baseline.

To contextualize the severe-degradation results, we additionally compute a random-choice baseline on the full DarkQA evaluation set. The overall chance accuracy is 33.35%, due to the mixed candidate-set sizes across question families.

Thus, model accuracies around 30–35% at L4/L5 indicate near chance-level performance, and should not be interpreted as evidence of successful visual understanding.

Impact of illumination drop and sensor noise. To understand the robustness of VLMs against visual illumination degradation, we first observe their performance under two types of low-light simulation: (1) pure EV drop and (2) physics-motivated noise modeling. As shown in Fig. 6-(b), both degradations consistently lead to a significant decrease in VLM accuracy. Notably, the introduction of sensor noise compounds this decline, resulting in a more pronounced performance drop compared to pure EV reduction. This confirms that VLMs are indeed highly sensitive to such visual degradation, with noise being a critical factor.

Effectiveness of low-light image enhancement (LLIE) pre-processing. Given the observed performance degradation, we investigate whether pre-processing low-light images with a state-of-the-art Low-Light Image Enhancement (LLIE) model [20] can mitigate these issues. We apply LLIE models to the noise-added low-light images before feeding them into the VLMs. As illustrated in Fig. 6-(c), this approach yields mixed results. While we observe a significant accuracy improvement at more severe low-light levels (L4 and L5), performance decreases at moderate levels (L1–L3). This unstable behavior highlights the challenge of reliably enhancing low-light images across different levels of degradation. While current LLIE models enhance perceptual quality, the results suggest that current LLIE models may be biased to certain degradation levels as in Fig. 6-(d).

Question-wise analysis. To gain a more granular understanding of the performance decline, we further analyze the accuracy degradation across different question types, as shown in Fig. 7. While most categories exhibit a steady decline, we observe a critical phenomenon in two specific types: "Room Type Recognition" and "Object Attribute - Color". For these categories, the VLM accuracy drops below that of the GPT-4 (Blind-LLM) baseline at severe degradation levels (L5 for the former, and L4 and L5 for the latter). The fact that this effect is particularly pronounced for the "Color" category strongly suggests that VLMs struggle to extract or preserve essential visual semantic information, such as color, when processing heavily dark images. Interestingly, this observation is analogous to the behavior of the human vision in dark scenes, where the visual primarily relies on rod cells that are sensitive to luminance because color-sensitive cone cells function much less effectively.

VLMs' Hallucination and Bias. We find interesting results on Object Recognition errors under severe low-light degradation. Interestingly, L5 does not increase object-presence false positives; instead, the model becomes more conservative, with a false-positive rate of only 4.96% and a false-negative rate of 89.36%. In other words, the dominant failure is missing present objects, rather than hallucinating absent ones.

For non-binary questions, severe darkness also induces answer collapse: at raw L5, 67.0% of Room Type predictions are "living room" although it accounts for only 32.4% of ground-truth room labels, and 87.7% of Color predictions become "black" although black accounts for only 15.0% of ground-truth colors. These analyses are based on LLaVA-OneVision-1.5-8B performance on DarkQA.

TABLE I
QUANTITATIVE REALISM VALIDATION OF LOW-LIGHT SYNTHESIS VARIANTS ON PAIRED REAL NORMAL-/LOW-LIGHT IMAGES FROM LLRGBD-REAL [34].

Method	PSNR ↑	SSIM ↑	D_{KL} ↓
Naïve EV-only baseline	25.39	0.757	14.8
Ours	26.06	0.785	**3.3**

TABLE II
OPEN-ENDED OPENEQA EM-EQA EVALUATION UNDER OUR LOW-LIGHT SYNTHESIS PIPELINE, WITH A HUMAN VISIBILITY STUDY ON DEGRADED OBSERVATIONS FROM 32 HM3D SCENES. SCORES ARE LLM-MATCH (%).

Method	**L0**	**L2**	**L4**
GPT-4o	75.0	72.4 -2.6	66.8 -8.3
Human	85.1	74.2 -10.9	50.7 -34.4

D. Realism of physics-motivated low-light synthesis pipeline

To validate whether our phyiscs-motivated synthetic degradations reflect real low-light camera observations, we conduct a quantitative paired-image validation using all paired normal-/low-light images in the validation split of LLRGBD-real [34], which contains real indoor images captured by an Intel RealSense D435i RGB-D camera. For each normal-light image, we synthesize its low-light counterpart using either a naïve EV-drop-only baseline or our physics-motivated pipeline with RAW-domain sensor-noise injection, while keeping the EV reduction identical, and compare both against the corresponding real low-light image. As shown in Table I, our pipeline improves image similarity from 25.39/0.757 to 26.06/0.785 in PSNR/SSIM and, more importantly, reduces the Kullback–Leibler divergence between real and synthesized noise-residual distributions from 14.8 to 3.3, following the residual-distribution validation protocol used in real-noise synthesis [35]. These results indicate that the proposed RAW-domain noise model better matches real low-light camera statistics than simple brightness reduction, thereby mitigating the real-to-sim concern while preserving the controllability required for systematic benchmark construction.

E. Open-Ended Embodied Question Answering Evaluation

We further validate whether the observed low-light vulnerability extends beyond our visual-primitive QA setting at embodied question answering using the HM3D subset of OpenEQA EM-EQA [12]. We keep the original scenes, trajectories, viewpoints, questions, and answers fixed, apply our physics-based low-light synthesis only to the RGB observations in episodic memory, and evaluate GPT-4o with the original LLM-Match protocol.

This experiment is intended as a complementary validation. DarkQA is designed as a controlled visual-primitive QA benchmark, but OpenEQA allows us to check whether the same low-light degradation trend also appears in an open-ended episodic-memory QA format without predefined answer choices. Since all non-illumination factors are kept fixed, the consistent drop in LLM-Match suggests that low-light degradation affects not only our controlled visual-primitive QA setting, but also broader embodied QA settings built on episodic visual observations.

Table II shows that GPT-4o drops from 75.0% at L0 to 66.8% at L4, while human performance from 32 HM3D scenes drops from 85.1% to 50.7%. This result supports our main finding: low-light degradation harms not only visual-primitive QA, but also open-ended embodied QA.

V. CONCLUSION

We introduce DarkQA, a new benchmark designed to address an overlooked and critical regime in VLM evaluation: the lack of systematic analysis for embodied scenario in low-light conditions. Using a physically-grounded low-light image synthesis pipeline, we create a reproducible benchmark to measure VLM robustness against realistic visual degradations. Our findings reveal that current VLMs are brittle in the dark, and that seemingly straightforward solutions like LLIE pre-processing can yield unstable results. While our benchmark reveal the vulnerabilities of VLMs to low-light conditions, a detailed failure analysis remains a valuable direction.

REFERENCES

[1] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser, *et al.*, "Openscene: 3d scene understanding with open vocabularies," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 815–824, 2023.

[2] J. Yang, S. Yang, A. W. Gupta, R. Han, L. Fei-Fei, and S. Xie, "Thinking in space: How multimodal large language models see, remember, and recall spaces," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, 2025.

[3] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. 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.

[4] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, *et al.*, "Palm-e: An embodied multimodal language model," 2023.

[5] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

[6] S. Datta, S. Dharur, V. Cartillier, R. Desai, M. Khanna, D. Batra, and D. Parikh, "Episodic memory question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19119–19128, 2022.

[7] B. Jia, T. Lei, S.-C. Zhu, and S. Huang, "Egotaskqa: Understanding human tasks in egocentric videos," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3343–3360, 2022.

[8] K. Mangalam, R. Akshulakov, and J. Malik, "Egoschema: A diagnostic benchmark for very long-form video language understanding," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46212–46244, 2023.

[9] S. Zhou, J. Xiao, Q. Li, Y. Li, X. Yang, D. Guo, M. Wang, T.-S. Chua, and A. Yao, "Egotextvqa: Towards egocentric scene-text aware video question answering," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3363–3373, 2025.

[10] A. Das, D. Gordon, C. Divi, D. Batra, G. Gkioxari, and D. Parikh, "Embodied question answering," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2054–2063, 2018.

[11] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe, "Scanqa: 3d question answering for spatial scene understanding," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19129–19139, 2022.

[12] A. Majumdar, A. Ajay, X. Zhang, P. Putta, S. Yenamandra, M. Henaff, S. Silwal, P. Mcvay, O. Maksymets, S. Arnaud, *et al.*, "Openeqa: Embodied question answering in the era of foundation models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16488–16498, 2024.

[13] I. Keller and K. S. Lohan, "On the illumination influence for object learning on robot companions," *Frontiers in Robotics and AI*, vol. 6, p. 154, 2020.

[14] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, "Ai models collapse when trained on recursively generated data," *Nature*, vol. 631, no. 8022, pp. 755–759, 2024.

[15] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

[16] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[17] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao, *et al.*, "Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency," *arXiv preprint arXiv:2508.18265*, 2025.

[18] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[19] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[20] D. Feijoo, J. C. Benito, A. Garcia, and M. V. Conde, "Darkir: Robust low-light image restoration," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10879–10889, 2025.

[21] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage retinex-based transformer for low-light image enhancement," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12504–12513, 2023.

[22] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1780–1789, 2020.

[23] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10561–10570, 2021.

[24] T. Brooks, B. Mildenhall, T. Xue, J. Chen, D. Sharlet, and J. T. Barron, "Unprocessing images for learned raw denoising," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11036–11045, 2019.

[25] K. Wei, Y. Fu, Y. Zheng, and J. Yang, "Physics-based noise modeling for extreme low-light photography," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8520–8537, 2021.

[26] D. Zhang, Y. Fu, R. Yang, Y. Miao, T. Qian, X. Zheng, G. Sun, A. Chhatkuli, X. Huang, Y.-G. Jiang, *et al.*, "Egonight: Towards egocentric vision understanding at night with a challenging benchmark," *arXiv preprint arXiv:2510.06218*, 2025.

[27] K. Wang, Z. Zhang, Z. Yan, X. Li, B. Xu, J. Li, and J. Yang, "Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark," in *IEEE International Conference on Computer Vision*, pp. 16055–16064, 2021.

[28] S. Lee, J. Rim, B. Jeong, G. Kim, B. Woo, H. Lee, S. Cho, and S. Kwak, "Human pose estimation in extremely low-light conditions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 704–714, 2023.

[29] Y. Sasagawa and H. Nagahara, "Yolo in the dark-domain adaptation method for merging multiple models," in *European Conference on Computer Vision*, pp. 345–359, Springer, 2020.

[30] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. M. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, M. Savva, Y. Zhao, and D. Batra, "Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied

AI," in *Advances in Neural Information Processing Systems Datasets and Benchmarks*, 2021.

[31] T. Plotz and S. Roth, "Benchmarking denoising algorithms with real photographs," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1586–1595, 2017.

[32] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[33] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, *et al.*, "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.

[34] N. Zhang, F. Nex, N. Kerle, and G. Vosselman, "Lisu: Low-light indoor scene understanding with joint learning of reflectance restoration," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 183, pp. 470–481, 2022.

[35] Z. Fu, Y. Yang, X. Tu, Y. Huang, X. Ding, and K.-K. Ma, "srgb real noise synthesizing with neighboring correlation-aware noise model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1683–1691, 2023.