EMPOWERING MULTIMODAL LARGE LANGUAGE MODELS FOR RAW IMAGE UNDERSTANDING

Anonymous authors
Paper under double-blind review

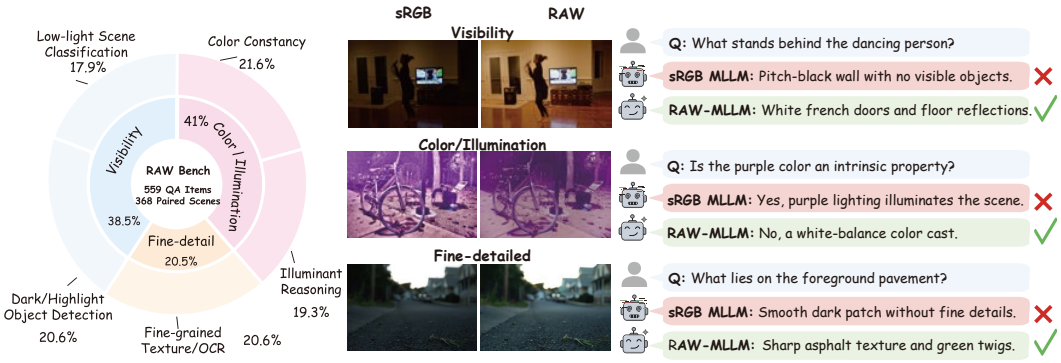

Figure 1: **Overview of RAW Bench and sensor-grounded reasoning.** (Left) The benchmark evaluates five RAW-sensitive dimensions across three evidence families. (Middle and Right) Matched sRGB/RAW observations show that ISP processing discards critical physical cues, causing standard MLLMs to fail, whereas our RAW-aware model reliably recovers obscured evidence.

ABSTRACT

In this paper, we enable multimodal large language models (MLLMs) to reason from RAW measurements by addressing a critical blind spot: existing benchmarks evaluate only post-ISP images, in which processing may weaken low-light, highlight, illumination, color, and texture cues. Current evaluations therefore cannot expose this information loss. We introduce RAW Bench, a diagnostic benchmark comprising 559 questions over 368 paired scenes across five RAW-sensitive dimensions. Each item is RAW-indispensable, requiring physical evidence preserved in RAW but weakened in the matched sRGB image. To achieve this criterion, RAW Bench uses matched RAW/sRGB inputs, hidden fact anchors, and separate scores for pairwise correctness, fact grounding, and reasoning quality, distinguishing evidence-grounded answers from semantic guesses. We further propose a dual-pathway RAW-aware adaptation framework that preserves the pretrained semantic interface while injecting compact RAW context into projected visual tokens, yielding consistent same-backbone gains across two MLLM families, particularly in low-light, dark/highlight, and illuminant reasoning. These results demonstrate the value of sensor-domain evidence for multimodal reasoning.

1 INTRODUCTION

Multimodal large language models (MLLMs) have achieved remarkable performance on a wide range of vision-language tasks, involving natural images (Liu et al., 2024a; Chen et al., 2024b; Deitke et al., 2025), documents (Mathew et al., 2021; Wu et al., 2024b; 2025), and visual question answering (Antol et al., 2015; Hudson & Manning, 2019; Marino et al., 2019). However, nearly all existing MLLMs reason from ISP-rendered sRGB images rather than the underlying sensor measurements. Cameras first capture RAW data, which is converted to sRGB through an image signal processing (ISP) pipeline before reaching the model. This conversion may attenuate evidence related to low-light visibility, dynamic range, illumination, color, and fine detail, thereby limiting the physical information available for downstream reasoning.

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.

This creates a *fundamental blind spot* for visual reasoning: neither existing models nor existing benchmarks account for ISP-induced information loss, even when reliable reasoning depends on physical evidence preserved in RAW measurements. This limitation directly impairs safety-critical applications such as autonomous driving and robotic navigation in extreme lighting. When vehicles face oncoming high beams on unlit roads, or when outdoor cameras monitor scenes under severe backlighting, display-oriented ISPs crush dark pedestrians into solid black or blow out surrounding structures into saturated white. While standard sRGB inputs permanently lose this visual evidence, the underlying linear sensor measurements preserve usable signals across these high-dynamic-range extremes. When physical cues are discarded at the sensor interface, stronger semantic reasoning alone cannot recover them. Addressing this limitation requires both RAW-aware models and evaluations that verify whether visual reasoning is grounded in sensor measurements.

Existing MLLM benchmarks (Li et al., 2024a; Ying et al., 2024; Fu et al., 2025) assess diverse visual capabilities, from perception (Gurari et al., 2018; Liu et al., 2024b) and mathematical reasoning (Lu et al., 2022; 2024) to hallucination analysis (Li et al., 2023b; Guan et al., 2024). However, they cannot determine whether predictions rely on RAW measurements. Consequently, a correct answer may stem from scene semantics (Agrawal et al., 2018), language priors (Chen et al., 2024a), or model scale (Chowdhery et al., 2023) rather than sensor evidence. Without paired RAW/sRGB observations, existing evaluations cannot isolate pre-ISP contributions or verify genuine grounding.

To address this evaluation gap, we introduce **RAW Bench**, a diagnostic benchmark where every question is *RAW-indispensable*: the conclusion relies on evidence recoverable from RAW but weakened or distorted in sRGB. RAW Bench contains 559 questions over 368 paired scenes spanning five dimensions (Fig. 1, left). As illustrated in Fig. 1 (middle and right), conventional MLLMs fail when ISP rendering discards decisive cues, whereas RAW measurements enable grounded reasoning. Its paired design controls scene and text inputs while varying the image representation, and hidden fact anchors enable separate evaluation of correctness, fact grounding, and reasoning quality. This design distinguishes evidence-grounded reasoning from semantic guesses.

Adapting pretrained MLLMs to RAW measurements requires bridging the mismatch between packed sensor inputs and display-domain vision encoders without altering the evidence of interest through a full ISP. We therefore propose a dual-pathway framework: a pseudo-sRGB pathway preserves the pretrained semantic interface, while a compact RAW context pathway injects sensor cues into projected visual tokens. Across InternVL3.5-8B and LLaVA-1.5-7B, this design improves overall RAW Bench performance by 7.63 and 6.58 points with gains across all five dimensions.

Our main contributions are summarized as follows: (1) We introduce RAW Bench, the first benchmark dedicated to evaluating visual reasoning grounded in RAW measurements through matched RAW/sRGB observations, hidden fact anchors, and decomposed diagnostic metrics; (2) We propose a dual-pathway RAW-aware adaptation framework that preserves the pretrained semantic interface through a pseudo-sRGB pathway while injecting compact RAW context into projected visual tokens; and (3) We provide the first systematic study of RAW-grounded reasoning in MLLMs, demonstrating consistent same-backbone gains across two model families and identifying the capabilities benefiting most from RAW measurements and the limits of global RAW context.

2 Related Work

2.1 MLLMs for sRGB Images

Existing MLLMs differ in their vision–language architectures and alignment strategies, yet they share a common assumption: *reasoning begins from an ISP-rendered sRGB image*. Large-scale image–text pretraining (Radford et al., 2021; Jia et al., 2021; Kim et al., 2021; Wang et al., 2022), followed by instruction tuning (Dai et al., 2023; Liu et al., 2023), lightweight adapters (Ye et al., 2023), or cross-modal alignment modules (Alayrac et al., 2022; Li et al., 2023a; Peng et al., 2024), has enabled remarkable progress in semantic perception and language-guided reasoning (Bai et al., 2023; Li et al., 2025a; Wang et al., 2024; 2025). However, these capabilities are learned entirely from rendered images rather than the underlying sensor measurements. During ISP, evidence related to low-light visibility and fine details (Chen et al., 2018; Brooks et al., 2019), dynamic range (Hasinoff et al., 2016), and scene color or illumination (Cheng et al., 2014; Hu et al., 2017; Schwartz et al., 2018) may be attenuated or discarded before reaching the model. Consequently, stronger vision–

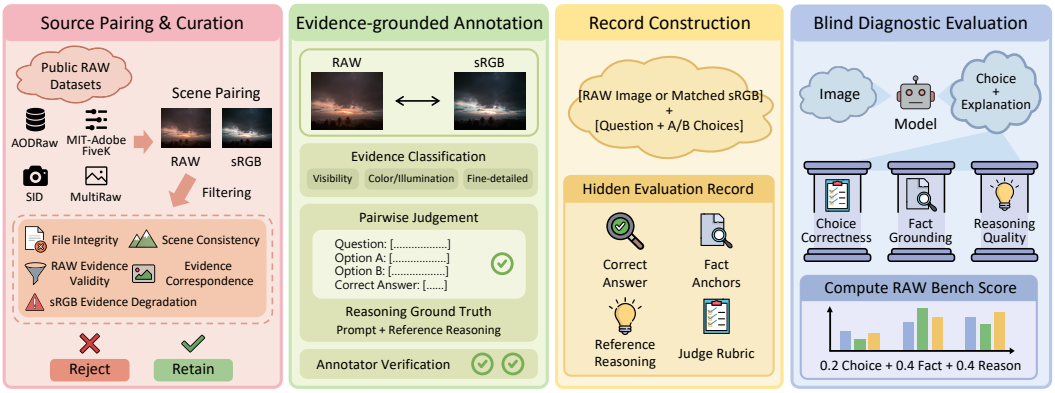

Figure 2: RAW Bench construction and evaluation protocol. Matched RAW/sRGB sources undergo RAW-indispensability verification and evidence-grounded annotation, followed by blind evaluation.

language alignment or larger language models cannot recover information that is absent from the input. In contrast, our framework augments pretrained MLLMs with RAW representations, enabling sensor-grounded reasoning while preserving existing semantic capabilities.

2.2 From RAW Measurements to Semantic Reasoning

Traditionally, RAW research has exploited sensor measurements to recover images or estimate image-formation properties. Classical and learning-based ISP methods address demosaicing (Malvar et al., 2004), denoising (Gharbi et al., 2016; Plotz & Roth, 2017; Brooks et al., 2019), restoration (Zamir et al., 2020), and color reconstruction (Schwartz et al., 2018), while low-light (Chen et al., 2018; 2019), HDR (Hasinoff et al., 2016), and color-constancy approaches (Buchsbaum, 1980; Finlayson & Trezzi, 2004; Gehler et al., 2008; Cheng et al., 2014; Hu et al., 2017) leverage RAW measurements to preserve information about exposure, dynamic range, noise, and illumination. More recently, RAW representations have been adapted for high-level vision tasks, including image classification (Yoshimura et al., 2023), object detection (Wu et al., 2019; Li et al., 2025b), semantic segmentation (Li et al., 2024b), and general visual understanding (Huang et al., 2025; Cui et al., 2026b; Chen et al., 2026). Despite this progress, existing approaches remain focused on structured prediction, producing class labels, bounding boxes, or segmentation masks. They do not evaluate whether models justify decisions using evidence uniquely preserved in RAW.

2.3 Diagnostic Evaluation Beyond sRGB

Existing vision-language benchmarks evaluate broad perception and reasoning (Liu et al., 2024b; Li et al., 2024a; Yu et al., 2024; Yue et al., 2024; Lu et al., 2024), while diagnostic benchmarks probe specific failure modes such as hallucination (Li et al., 2023b; Guan et al., 2024), language priors (Chen et al., 2024a), OCR dependence (Singh et al., 2019; Liu et al., 2024c), and low-level visual sensitivity (Fu et al., 2024; Wu et al., 2024a). These benchmarks have substantially improved the evaluation of MLLMs, but they all assume reasoning from a single rendered image. As a result, they cannot determine whether predictions rely on sensor evidence or cues preserved post-ISP.

RAW Bench addresses this gap by treating the imaging interface as the experimental variable. Matched RAW/sRGB pairs keep the scene and question fixed while changing only the available visual evidence. Hidden fact anchors and decomposed diagnostic scores further distinguish evidence-grounded reasoning from semantic guessing.

3 RAW Bench

RAW Bench comprises *559* question instances over *368* matched RAW/sRGB scenes. As illustrated in Fig. 2, its design proceeds from source curation and indispensability filtering to evidence-grounded annotation and blind evaluation.

Visibility Evidence	Color & Illumination	Fine-Detail & Texture
		
Pairwise Judgment: Q: Why do the cliffside buildings appear nearly black? A: They are naturally dark, unreflective buildings under dim ambient light. B: They are light buildings underexposed because exposure favors the bright sunset and sky.	**Pairwise Judgment:** Q: Is the purple sky a natural scene property? A: No. Natural daylight and a blue sky are obscured by a global red/magenta color cast. B: Yes. Reddish illumination makes the sky and foliage purple.	**Pairwise Judgment:** Q: What lies directly below the "duxtop" brand name? A: Random noisy pixels with no meaningful structure. B: Degraded but structured text reading "Induction Cooking Expert."
Reasoning Prompt: Explain the silhouette using sunset backlighting, extreme dynamic range, and the camera exposure that prioritizes bright regions.	**Reasoning Prompt:** Explain the true colors of the sky, statue, foliage, and blossoms, and identify the source of the cast.	**Reasoning Prompt:** Describe how word shapes, relative lengths, and spacing separate faint printed text from random image noise.
Pairwise Answer: B	**Pairwise Answer: A**	**Pairwise Answer: B**
Reference Reasoning: White buildings face away from the sun; exposure for the sun crushes foreground shadow detail.	**Reference Reasoning:** A global white-balance or processing cast shifts every region magenta; the sky is blue, statue neutral, foliage green.	**Reference Reasoning:** Word-group spacing and lengths support the phrase despite severe noise.
Fact Anchors: low sun; white buildings; shadowed cliffside.	**Fact Anchors:** blue sky; stone statue; green foliage; white blossoms.	**Fact Anchors:** duxtop; induction cooktop; product box.

Figure 3: Representative RAW Bench examples across the three evidence families. RAW panels are display renderings of original sensor measurements for visual inspection. Questions, candidate statements, reference reasoning, and fact anchors are condensed from full evaluation records.

3.1 Task Definition and Benchmark Design

RAW Bench evaluates whether pre-ISP sensor measurements enable an MLLM to reach and justify conclusions that cannot be reliably supported by the corresponding sRGB image alone. Given a scene, a question, and two candidate statements, the model must select the statement supported by physical evidence and provide a concise explanation. We term this *sensor-grounded reasoning*.

To isolate the contribution of RAW measurements, the benchmark compares matched RAW and sRGB observations of the same scene under identical questions, choices, and response formats. This paired design rigorously controls for scene semantics, prompt phrasing, and language priors. An item is strictly *RAW-indispensable* if the correct conclusion is readily verifiable from RAW measurements but obscured or distorted in the matched sRGB view.

We adopt a pairwise judgment format because sensor-sensitive attributes (e.g., exposure, dynamic range, illumination) are continuous and admit multiple valid descriptions. Contrasting two physically distinct hypotheses constrains the search space without imposing rigid phrasing constraints. Crucially, requiring an accompanying explanation exposes whether the model grounds its decision in sensor evidence or succeeds via semantic guessing (Fig. 3).

3.2 Evidence-Grounded Construction

Source pairing and curation. RAW Bench is built from publicly available datasets that provide original RAW measurements together with corresponding rendered images or sufficient information to reconstruct matched sRGB views. We collect paired scenes from AODRaw (Li et al., 2025b), MIT-Adobe FiveK (Bychkovsky et al., 2011), See in the Dark (SID) (Chen et al., 2018), and MultiRAW (Li et al., 2024b). For each pair, we verify scene correspondence and sufficient alignment for controlled comparison, then inspect file integrity, usable exposure, scene visibility, and imaging-pipeline differences. Corrupted, severely misaligned, visually ambiguous, or otherwise unsuitable pairs are discarded before annotation.

Indispensability filtering. To guarantee benchmark rigor, human annotators inspect candidate pairs and filter out items answerable via scene semantics, common sense, or cues already discernible in sRGB. An item is retained only when the designated answer strictly requires physical evidence uniquely preserved in RAW. This ensures benchmark difficulty originates exclusively from ISP-induced information loss rather than perceptual ambiguity or linguistic tricks.

Evidence taxonomy. Retained instances are structured into five RAW-sensitive dimensions grouped under three evidence families: (1) *Visibility* (low-light visibility, dark/highlight recovery), targeting signals crushed by tone mapping or underexposure; (2) *Color and illumination* (color constancy, illuminant reasoning), probing intrinsic surface reflectance versus ambient casts; and (3) *Fine-detail*

(texture/OCR), assessing high-frequency structures degraded by aggressive ISP denoising. This taxonomy provides a structured basis for analyzing where RAW measurements contribute to reasoning and where models remain limited.

Annotation and hidden record construction. For each verified pair, Gemini-2.5-Flash generates an initial pairwise question, two candidate hypotheses, the correct answer, and an evidence explanation. Human experts manually inspect both image domains to refine the question, ensure unambiguous physical distinctions, and distill decisive observations into compact *fact anchors*. Each retained item is compiled into a hidden record containing the answer key, fact anchors, reference reasoning, and an explicit scoring rubric, which are strictly withheld from evaluated models.

3.3 Evaluation Protocol and Metrics

Blind evaluation. For each item, the evaluated model receives only the visual input, question, two candidate statements, and an instruction to provide a concise explanation. The correct option, fact anchors, reference reasoning, evidence dimension, and scoring rubric remain hidden. RAW-aware models are evaluated on the original RAW measurement, whereas conventional MLLMs receive the matched sRGB image; all textual inputs and response requirements are held fixed. This rigorous protocol prevents models from exploiting annotation metadata and enables a strictly controlled, unbiased comparison across heterogeneous imaging representations.

Decoupled diagnostic metrics. RAW Bench evaluates decision accuracy and evidence retrieval through decoupled metrics. *Pairwise correctness* $p_i \in \{0, 1\}$ is computed deterministically by exact label matching against the ground truth, requiring no model-based judging. In parallel, GPT-5.1 scores whether the explanation recovers the physical evidence defined by hidden anchors (*fact grounding* $f_i \in [0, 5]$) and whether this evidence logically supports the designated conclusion (*reasoning quality* $r_i \in [0, 5]$). During judging, visual inputs and candidate options are withheld from GPT-5.1: omitting sRGB images eliminates confounding from ISP rendering artifacts, while omitting candidate options prevents lexical shortcut matching. GPT-5.1 operates strictly as an evidence extractor evaluated against hidden physical anchors. For sample i, the sample score is

$$s_i = 0.2p_i + 0.4\frac{f_i}{5} + 0.4\frac{r_i}{5}. \tag{1}$$

This weighting assigns 80% of the score to physical evidence recovery and valid reasoning, penalizing unsupported guesses. We average s_i across samples and multiply by 100 to obtain overall and dimension-wise scores.

Human-LLM judge agreement. To verify the fidelity of GPT-5.1 scoring, two domain experts independently scored a stratified sample of 100 benchmark items across all five dimensions under identical fact anchors and rubrics. Human consensus exhibits strong agreement with GPT-5.1, achieving a Pearson correlation of $r = 0.945$ on the composite benchmark score (MAE $= 4.00/100$), $r = 0.907$ on fact grounding, and $r = 0.813$ on reasoning quality. Furthermore, GPT-5.1 achieves a 100% adjacent-score agreement (≤ 1 point difference on the 0–5 scale) with both human raters, confirming that anchor-based blind evaluation reliably reflects physical evidence recovery without evaluation hallucinations (see Appendix B.4 for details).

4 RAW-aware MLLM Adaptation

4.1 Framework Overview

We adapt an sRGB-pretrained MLLM to RAW measurements while preserving its established vision–language alignment. As shown in Fig. 4, our framework decouples visual processing into a pseudo-sRGB semantic pathway and a compact RAW context pathway. Let x^r denote a packed four-channel RAW input and q a text instruction. The base MLLM consists of a frozen vision encoder E_v, a multimodal projector P, and a language model L.

To avoid destructive full-ISP rendering while maintaining semantic compatibility, we construct a parameter-free three-channel pseudo-sRGB representation x^p for E_v. Concurrently, a lightweight RAW encoder E_r extracts acquisition-level physical context directly from x^r:

$$U = P(E_v(x^p)) \in \mathbb{R}^{N\times d}, \qquad c^r = E_r(x^r) \in \mathbb{R}^{128}, \tag{2}$$

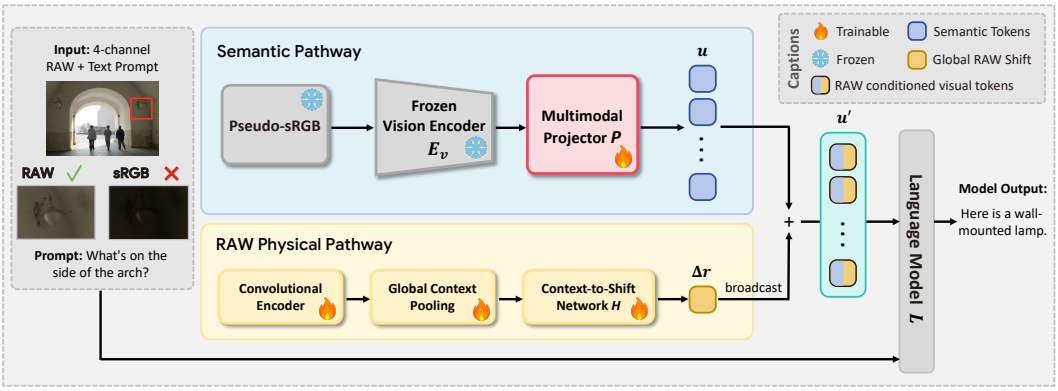

Figure 4: **RAW-aware MLLM adaptation.** The pseudo-sRGB pathway preserves the pretrained semantic interface, while the RAW context pathway injects a global residual into projected visual tokens before language reasoning.

where N is the token sequence length and d is the language model hidden dimension. A context mapper H transforms c^r into a token-space residual Δr, which is broadcast across all visual tokens:

$$\Delta r = H(c^r) \in \mathbb{R}^d, \qquad U' = U + \mathbf{1}_N(\Delta r)^\top. \tag{3}$$

The language model then generates responses conditioned on U' and q. Injecting RAW context after projection recalibrates visual features with capture-level cues while strictly preserving pretrained token length, ordering, and semantic structure.

4.2 Semantic Pathway via Pseudo-sRGB

Pretrained vision encoders expect standard three-channel inputs with natural scene statistics. However, passing RAW data through a full ISP pipeline introduces tone mapping and non-linear color clipping that permanently destroy physical evidence. To resolve this tension, we construct an ISP-free pseudo-sRGB representation x^p. Given $x^r = (R, G_1, G_2, B)$, we average the two green channels and concatenate the planes:

$$x^p = \mathrm{Concat}_{\mathrm{ch}}\left(R,\ \frac{G_1 + G_2}{2},\ B\right).$$

This transformation bridges the channel mismatch without demosaicing, white balancing, or tone mapping. It provides a structurally aligned semantic scaffold for E_v while leaving raw radiometric ratios intact for downstream reasoning.

4.3 RAW Context Encoding and Injection

While x^p preserves spatial structures, channel packing discards absolute sensor statistics. The RAW context pathway compensates for this by explicitly extracting exposure, dynamic range, illumination, and noise characteristics. As detailed in Appendix Tab. 7, E_r employs a convolutional stem, three residual stages, global average pooling, and a projection head to compress x^r into $c^r \in \mathbb{R}^{128}$.

To integrate these physical cues smoothly into language reasoning, the final linear layer of H is zero-initialized. Consequently, training begins strictly from the pretrained model baseline and gradually incorporates RAW-dependent corrections via Eq. 3. By applying this broadcast residual post-projection, we prevent perturbation of low-level visual filters in E_v while ensuring the language model receives globally calibrated sensor context.

4.4 Instruction Supervision and Training

We construct a separate RAW–text instruction corpus following the evidence taxonomy of RAW Bench. Unlike the manually verified benchmark items, this corpus provides scalable supervision for

identifying and explaining sensor-domain evidence and does not reuse benchmark annotations. During instruction construction, a controlled sRGB rendering is used only to identify cues weakened by the imaging pipeline. Gemini-2.5-Flash (Comanici et al., 2025) generates a pairwise judgment prompt, two candidate statements, and a concise evidence-based explanation. We discard instructions answerable through generic scene semantics, unavailable camera metadata, or recognition of the RAW format itself. During training, the model receives only the RAW input x^r and text prompt q; the paired rendering is never used as model input.

Let $a = (a_1, \ldots, a_T)$ denote the target response, including the selected statement and its explanation. We optimize

$$\mathcal{L} = -\sum_{t=1}^{T} \log p_\theta(a_t \mid a_{<t}, q, x^r)\,. \tag{4}$$

We use two-stage training to separate RAW representation learning from language adaptation. In Stage A, the vision encoder and language model remain frozen, while the RAW context pathway, context-mapping network, and multimodal projector are optimized to align RAW evidence with the pretrained visual-token space. In Stage B, these modules continue to be trained, and LoRA (Hu et al., 2022) is applied to selected language-model projections so that the model can express RAW-conditioned evidence through pairwise judgments and concise physical explanations.

5 Experiments

5.1 Experimental Setup

Evaluation setting. All models are evaluated on the same 559 RAW Bench questions using a fixed pairwise-judgment and concise-reasoning protocol. Conventional MLLMs receive matched sRGB images, whereas RAW-aware models receive the corresponding RAW measurements. Questions, candidate statements, option order, and response format are identical across input representations, and evaluated models receive neither evidence-dimension labels nor hidden evaluation records.

Models and scoring. We evaluate both open-source and closed-source MLLMs, together with RAW-aware variants built on InternVL3.5-8B (Wang et al., 2025) and LLaVA-1.5-7B (Liu et al., 2024a). We use deterministic decoding when supported and store all responses before evaluation. Pairwise correctness is computed directly, while GPT-5.1 assesses fact grounding and reasoning quality using hidden fact anchors and a fixed rubric. All scores are reported on a 0–100 scale over valid responses; a few commercial APIs suffer sporadic exceptions (< 10 items), causing slight offsets from exact weighted sums. We use *O*, *LL*, *D/H*, *CC*, *IR*, and *T/O* to denote overall, low-light, dark/highlight, color constancy, illuminant reasoning, and texture/OCR performance, respectively. Further implementation details are provided in the supplementary material.

Training details. The RAW-aware models are trained on 38K RAW–text instruction pairs from six RAW datasets: AODRaw (Li et al., 2025b), MIT-Adobe FiveK (Bychkovsky et al., 2011), PPR10K (Liang et al., 2021), See in the Dark (SID) (Chen et al., 2018), RAISE (Dang-Nguyen et al., 2015), and MultiRAW (Li et al., 2024b), with 2K additional pairs from the same six datasets reserved for validation. The training and validation splits are completely disjoint from RAW Bench at both the scene and source-file levels, including RAW files, paired sRGB images, questions, annotations, and hidden evaluation records, all of which are reserved exclusively for RAW Bench. Both training stages run for one epoch with a frozen vision encoder; further implementation details are provided in the supplementary material.

5.2 Benchmark and Baseline Results

We benchmark existing MLLMs on RAW Bench to evaluate their reliance on ISP-rendered representations versus true physical cues. As shown in Tab. 1, existing models exhibit pronounced intra-model variance across dimensions rather than uniform reasoning capabilities. For instance, while Qwen3-VL-8B (Bai et al., 2025) achieves the strongest color-constancy score (77.75), its performance drops markedly in underexposed shadow recovery (64.41 D/H). Conversely, MiniCPM-o-4.5 (Cui et al., 2026a) excels at high-frequency texture parsing (64.35 T/O) but degrades when separating illuminant casts from intrinsic reflectance (72.19 IR). Crucially, scaling parameters offers

Table 1: Performance of existing MLLMs across RAW Bench dimensions. Best results within each model group are shown in bold. InternVL3.5-8B and LLaVA-1.5-7B denote sRGB-adapted controls.

Open-source MLLMs Model	*O*	*LL*	*D/H*	*CC*	*IR*	*T/O*
Qwen3-VL-8B	**70.16**	**72.51**	**64.41**	**77.75**	**74.25**	62.13
MiniCPM-o-4.5	69.53	71.28	63.83	76.07	72.19	**64.35**
Ovis2.5-9B	66.48	69.52	61.36	71.44	70.30	60.17
InternVL3.5-38B	66.32	69.20	58.64	77.29	67.41	58.92
InternVL3.5-8B	62.80	66.44	57.01	69.62	64.89	56.28
LLaVA-OneVision-7B	60.94	60.44	55.62	66.71	62.89	58.82
Phi-3.5-Vision	59.63	61.40	56.17	66.41	61.56	52.59
Molmo-7B-D	59.46	60.64	53.81	68.23	61.33	53.11
Qwen2.5-VL-7B	54.69	54.68	50.33	59.41	58.22	50.78

Open-source MLLMs (cont.) Model	*O*	*LL*	*D/H*	*CC*	*IR*	*T/O*
LLaVA-1.5-7B	44.77	44.36	44.45	47.41	46.85	40.73
DeepSeek-VL2-Small	26.87	32.68	28.66	24.50	23.93	25.29
Closed-source MLLMs						
GPT-5-nano	**70.36**	**75.92**	60.21	**70.93**	**72.10**	**65.98**
GLM-4.5	66.95	62.23	67.58	69.45	69.70	62.12
GLM-4.6	65.50	71.28	71.48	67.11	68.78	56.90
GPT-4.1-nano	64.93	69.12	61.22	69.92	67.78	58.09
GLM-4.7	64.69	70.72	64.31	67.17	64.37	55.90
GPT-4o-mini	64.19	67.96	**71.92**	63.67	65.55	62.96

Table 2: Same-backbone comparisons on RAW Bench under matched adaptation. InternVL and LLaVA denote sRGB controls.

Model	*O*	*LL*	*D/H*	*CC*	*IR*	*T/O*
InternVL3.5-8B	62.80	66.44	57.01	69.62	64.89	56.28
+ Multi-exposure sRGB	65.64	70.28	66.82	68.86	62.59	59.90
RAW-InternVL	**70.43**	**77.09**	**67.58**	**72.17**	**74.78**	**61.36**
Gain vs. Base	+7.63	+10.65	+10.57	+2.54	+9.88	+5.07
LLaVA-1.5-7B	44.77	44.36	44.45	47.41	46.85	40.73
+ Multi-exposure sRGB	21.48	17.52	19.51	30.94	20.89	17.50
RAW-LLaVA	**51.36**	**49.32**	**48.87**	**55.77**	**56.56**	**46.09**
Gain vs. Base	+6.58	+4.96	+4.41	+8.36	+9.70	+5.35

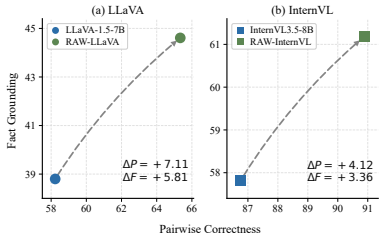

Figure 5: Pairwise correctness vs. fact-grounding gains across both model families.

no remedy for sensor-domain information loss: InternVL3.5-38B trails several 8B-scale architectures in aggregate score (66.32 vs. 70.16). These discrepancies confirm that benchmark difficulty does not stem from general visual-linguistic deficits, but reflects an inherent blind spot: when nonlinear ISP transforms discard radiometric extremes, stronger language priors or parameter scaling cannot reliably hallucinate the missing ground-truth photons.

Closed-source MLLMs demonstrate an identical architectural bottleneck. While GPT-4o-mini (Hurst et al., 2024) leads in highlight recovery (71.92 D/H), it underperforms on low-light scene discrimination (67.96 LL), where GPT-5-nano (Singh et al., 2025) dominates (75.92 LL). Because cross-model evaluations conflate encoder designs and training mixtures, isolating the specific utility of RAW inputs requires controlled same-backbone experimentation.

5.3 Controlled Evaluation of RAW Adaptation

We conduct controlled adaptation on InternVL3.5-8B and LLaVA-1.5-7B. While newer backbones such as Qwen3-VL-8B achieve competitive scores on sRGB images, their dynamic image tiling and patch-merging layers vary visual token counts across different resolutions and aspect ratios. We select InternVL and LLaVA to maintain strictly fixed token sequences, isolating the physical benefits of RAW context from token-budget and visual-capacity confounders.

For each backbone, the sRGB control and RAW-aware model use identical configurations, instruction data, and training budgets; only the visual input differs. As shown in Tab. 2, adapting to RAW yields consistent gains across both model families. RAW-InternVL improves the overall score by 7.63 points over InternVL3.5-8B, while RAW-LLaVA gains 6.58 points over LLaVA-1.5-7B, with improvements across all five dimensions. Fig. 5 confirms that these gains stem from simultaneous improvements in pairwise accuracy and fact grounding, showing that RAW inputs guide models toward physically verifiable evidence rather than semantic guesses.

Multi-exposure sRGB baseline. A natural alternative to RAW adaptation is exposure bracketing: rendering multiple sRGB views across exposure stops to recover dynamic range without architectural modifications. To test whether this post-ISP heuristic can substitute for sensor data, we evaluate off-the-shelf InternVL3.5-8B and LLaVA-1.5-7B on bracketed inputs (-2, 0, and $+2$ EV views concatenated sequentially). As shown in Tab. 2, multi-exposure heuristics reveal severe physical and architectural limits across both backbones. For InternVL3.5-8B, exposure scaling uncovers

Table 3: Component ablations of RAW-InternVL on RAW Bench. All variants include Stage A.

Variant	Configuration			RAW Bench Scores					
	Pseudo-sRGB	RAW Context	Stage B	*O*	*LL*	*D/H*	*CC*	*IR*	*T/O*
InternVL3.5-8B	×	×	√	62.80	66.44	57.01	69.62	64.89	56.28
Pseudo-sRGB adaptation	√	×	√	66.54	70.06	61.37	70.52	70.88	60.42
Stage A only	√	√	×	66.06	72.96	62.67	68.51	69.78	57.43
RAW-InternVL	√	√	√	**70.43**	**77.09**	**67.58**	**72.17**	**74.78**	**61.36**

shadow structures in low-light (70.28 vs. 66.44) and dark/highlight (66.82 vs. 57.01), but degrades color constancy (68.86 vs. 69.62) and illuminant reasoning (62.59 vs. 64.89), as tone mapping distorts radiometric color ratios across views. For LLaVA-1.5-7B, multi-exposure input triggers catastrophic reasoning collapse, plunging the overall score to 21.48 (vs. 44.77): tripling token count overwhelms single-image pretraining, causing format breakdown and empty explanations across 83.5% (467/559) of instances. In contrast, our RAW-aware models operate on a single compact RAW input, outperforming multi-exposure sRGB by +4.79 points on InternVL and +29.88 points on LLaVA while preserving standard single-image token efficiency and inference latency. These empirical results confirm that multi-exposure heuristics cannot replace direct sensor-level RAW modeling.

5.4 Ablation and Diagnostic Analysis

Component contributions. We analyze the pseudo-sRGB pathway, RAW context injection, and two-stage training using RAW-InternVL. As shown in Tab. 3, pseudo-sRGB adaptation improves the overall score from 62.80 for the sRGB-adapted control to 66.54, and improves all dimensions, indicating that RAW-derived sRGB can benefit the adapted backbone. Under the same two-stage training, adding the RAW context pathway and post-projector residual further raises the overall score by 3.89 points to 70.43, with particularly large gains of 7.03 points and 6.21 points on low-light and dark/highlight evidence, respectively. Compared with the Stage-A-only variant, completing Stage B adds 4.37 points overall and improves every dimension, with the largest gain on illuminant reasoning. These comparisons demonstrate complementary benefits from pseudo-sRGB adaptation, explicit RAW context injection, and Stage B instruction tuning.

Dimension-wise diagnosis. We examine where explicit RAW context provides the greatest benefit by comparing the full model with pseudo-sRGB adaptation alone. Across all five dimensions (detailed in Appendix Fig. 11), RAW context improves fact grounding, with the largest gains in low-light (+11.7), dark/highlight (+7.9), and texture/OCR (+6.4). The low-light and dark/highlight gains align with the global context pathway, which captures exposure, dynamic range, and illumination statistics that influence scene-level interpretation. Improvements are smaller for color constancy and illuminant reasoning, which inherently rely more on localized chromatic cues.

While global RAW conditioning successfully recalibrates semantic tokens, auditing our evaluation logs reveals distinct performance ceilings: correct binary choices are occasionally unaccompanied by fully grounded factual explanations, particularly in ultra-low-light and fine-detail regimes. To rigorously diagnose these physical boundaries, we present a comprehensive failure mode taxonomy and detailed case audit across representative instances in Appendix E.4.

6 Conclusion

We introduced RAW Bench, a diagnostic benchmark for sensor-grounded reasoning in MLLMs using matched RAW/sRGB observations, hidden fact anchors, and decomposed metrics. We also proposed a dual-pathway RAW-aware framework that preserves the pretrained semantic interface while injecting compact RAW context into projected visual tokens. Same-backbone evaluations across InternVL and LLaVA showed consistent gains across all five dimensions, while ablations revealed complementary benefits from pseudo-sRGB adaptation, explicit RAW context injection, and Stage B instruction tuning. These findings demonstrate the value of RAW measurements for multimodal reasoning and highlight the need for spatially resolved RAW features in fine-grained tasks. More broadly, our results show that the imaging interface determines the visual evidence available to an MLLM and therefore constitutes an integral component of multimodal reasoning.

AI USE STATEMENT

Generative AI tools were used to aid and polish manuscript writing, including language editing for clarity and organization, as well as to assist in drafting portions of the experimental and data-processing code. As part of the methodology reported in this paper, generative models were also used to draft candidate benchmark items and RAW–text instruction records and to score fact grounding and reasoning quality under fixed rubrics. Human annotators reviewed the matched RAW/sRGB evidence and revised or rejected generated benchmark items, and the authors thoroughly reviewed, verified, and tested all AI-assisted code, text, data records, prompts, and evaluation outputs. The authors take responsibility for the final content of this work, including all text, code, claims, and artifacts produced with AI assistance.

REPRODUCIBILITY STATEMENT

The main paper specifies the task, model architecture, training objective, evaluation protocol, and primary results. The appendix provides the benchmark composition, annotation and response-recovery procedures, judge inputs and rubric, RAW preprocessing, model implementation, instruction-corpus construction, training configuration, and additional fusion-depth analysis needed to reproduce the reported study.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, pp. 4971–4980, 2018.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, et al. VQA: Visual question answering. In *ICCV*, 2015.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, et al. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *CVPR*, 2019.

Gershon Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin Institute*, 310(1):1–26, 1980.

Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR*, 2011.

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018.

Chen Chen, Qifeng Chen, Minh N Do, and Vladlen Koltun. Seeing motion in the dark. In *ICCV*, 2019.

Kai Chen, Jin Xiao, Leheng Zhang, Kexuan Shi, and Shuhang Gu. Task-aware image signal processor for advanced visual perception. In *CVPR*, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, et al. Are we on the right way for evaluating large vision-language models? *NeurIPS*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024b.

Dongliang Cheng, Dilip K Prasad, and Michael S Brown. Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. *Journal of the Optical Society of America A*, 31(5):1049–1058, 2014.

Aakanksha Chowdhery, Sharan Narang, and Jacob Devlin. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24:1–113, 2023.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026a.

Ziteng Cui, Jianfei Yang, and Tatsuya Harada. Raw-adapter: Adapting pre-trained visual model to camera raw images and a benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026b.

Wenliang Dai, Junnan Li, Dongxu Li, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 2023.

Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: A raw images dataset for digital image forensics. In *MMSys*, 2015.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*, 2025.

Graham D Finlayson and Elisabetta Trezzi. Shades of gray and colour constancy. In *Color and imaging conference*, 2004.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *NeurIPS*, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, et al. BLINK: Multimodal large language models can see but not perceive. In *ECCV*, 2024.

Peter Vincent Gehler, Carsten Rother, Andrew Blake, Tom Minka, and Toby Sharp. Bayesian color constancy revisited. In *CVPR*, 2008.

Michaël Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand. Deep joint demosaicking and denoising. *ACM Transactions on Graphics*, 35(6):1–12, 2016.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024.

Danna Gurari, Qing Li, Abigale J Stangl, et al. VizWiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018.

Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics*, 35(6):1–12, 2016.

Edward J Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Yuanming Hu, Baoyuan Wang, and Stephen Lin. Fc4: Fully convolutional color constancy with confidence-weighted pooling. In *CVPR*, 2017.

Wenjun Huang, Ziteng Cui, Yinqiang Zheng, et al. Dr. RAW: Towards general high-level vision from RAW with efficient task conditioning. In *NeurIPS*, 2025.

Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.

Bo Li, Yuanhan Zhang, Dong Guo, et al. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, 2024a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023b.

Zhihao Li, Ming Lu, Xu Zhang, Xin Feng, M Salman Asif, and Zhan Ma. Efficient visual computing with camera raw snapshots. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (7):4684–4701, 2024b.

Zhong-Yu Li, Xin Jin, Bo-Yuan Sun, Chun-Le Guo, and Ming-Ming Cheng. Towards raw object detection in diverse conditions. In *CVPR*, 2025b.

Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *CVPR*, 2021.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, et al. MMBench: Is your multi-modal model an all-around player? In *ECCV*, 2024b.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024c.

Pan Lu, Swaroop Mishra, Tanglin Xia, et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022.

Pan Lu, Hritik Bansal, Tony Xia, et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

H.S. Malvar, Li wei He, and R. Cutler. High-quality linear interpolation for demosaicing of bayer-patterned color images. In *ICASSP*, 2004.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. DocVQA: A dataset for VQA on document images. In *WACV*, 2021.

Zhiliang Peng, Wenhui Wang, Li Dong, et al. Grounding multimodal large language models to the world. In *ICLR*, 2024.

Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *CVPR*, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Eli Schwartz, Raja Giryes, and Alex M Bronstein. Deepisp: Toward learning an end-to-end image processing pipeline. *IEEE Transactions on Image Processing*, 28(2):912–923, 2018.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, et al. Towards VQA models that can read. In *CVPR*, 2019.

Peng Wang, An Yang, Rui Men, et al. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022.

Weihan Wang, Qingsong Lv, Wenmeng Yu, et al. CogVLM: Visual expert for pretrained language models. *NeurIPS*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, 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, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Chyuan-Tyng Wu, Leo F. Isikdogan, Sushma Rao, Bhavin Nayak, Timo Gerasimow, Aleksandar Sutic, Liron Ain-kedem, and Gilad Michael. Visionisp: Repurposing the image signal processor for computer vision applications. In *ICIP*, 2019. doi: 10.1109/ICIP.2019.8803607.

Haoning Wu, Zicheng Zhang, Erli Zhang, et al. Q-Bench: A benchmark for general-purpose foundation models on low-level vision. In *ICLR*, 2024a.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024b.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Kaining Ying, Fanqing Meng, Jin Wang, et al. MMT-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In *ICML*, 2024.

Masakazu Yoshimura, Junji Otsuka, Atsushi Irie, and Takeshi Ohashi. Rawgment: Noise-accounted raw augmentation enables recognition in a wide variety of environments. In *CVPR*, 2023.

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. In *ICML*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *CVPR*, 2020.

A RAW Bench Construction and Composition

A.1 Source and Dimension Coverage

RAW Bench contains 559 questions over 368 unique scenes drawn from AODRaw, MIT-Adobe FiveK, See in the Dark (SID), and MultiRAW, as summarized in Fig. 6. These sources collectively cover underexposure, dynamic-range loss, illumination shifts, color casts, and degraded fine detail.

Among the 368 scenes, 177 contribute one question, while 191 contribute two complementary physical judgments. Specifically, 93 scenes pair low-light classification with dark/highlight assessment,

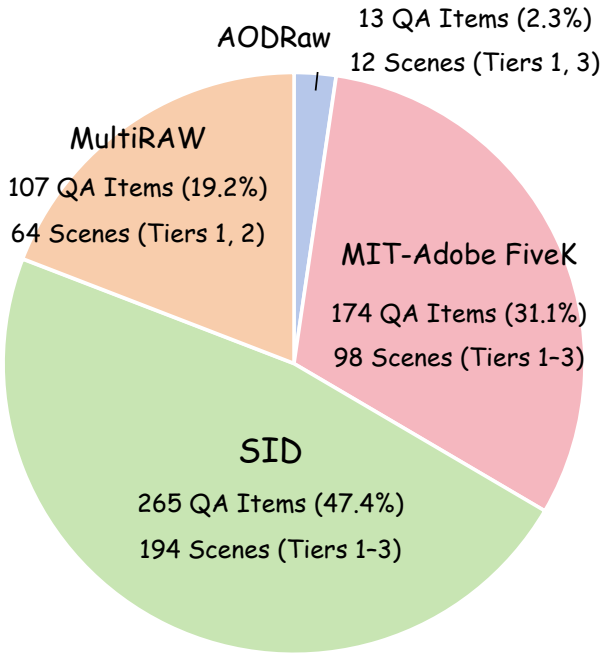

Figure 6: Source composition of RAW Bench across four public RAW datasets. Each segment reports the number and proportion of questions, together with the corresponding scene count.

Table 4: Definitions and item counts for RAW Bench.

Dimension	**Decisive evidence**	**Items**
Low-light	Genuine low-light conditions versus normally illuminated scenes rendered with misleading exposure or contrast.	100
Dark/Highlight	Scene content preserved in underexposed shadows or saturated highlights.	115
Color Constancy	Intrinsic object color versus capture- or rendering-induced color casts.	121
Illuminant Reasoning	Color patterns attributable to the illuminant, white balance, or ambient lighting.	108
Texture/OCR	Small text, material structure, edges, or patterns degraded by noise, blur, or image processing.	115
	Total	**559**

and 98 pair color constancy with illuminant reasoning. These paired questions probe distinct physical cues while holding the captured scene fixed.

The resulting benchmark contains 215 visibility questions, 229 color-and-illumination questions, and 115 fine-detail questions. Tab. 4 defines the five evaluation dimensions and reports their item counts. Each record includes an `evidence_family` field identifying one of the three broader evidence families and a `dimension` field specifying the physical cue required by the item, enabling dimension-wise analysis.

A.2 Record Structure and Balance

Each RAW Bench record includes a model-facing pairwise task, hidden evidence fields for diagnostic scoring, and metadata for analysis. The evaluated model receives only the question, two candidate

statements, and a request for a concise explanation. The correct option, fact anchors, scoring focus, and reference reasoning remain hidden during inference, while the `source`, `evidence_family`, and `dimension` fields are used only for post-hoc aggregation and analysis.

Across the 559 records, RAW Bench contains 1,657 fact anchors, averaging 2.96 per item. Specifically, 540 items contain three anchors, 18 contain two, and one contains a single decisive anchor. Answer positions are approximately balanced, with 302 items assigned to option A and 257 to option B. Option order is fixed across models so that matched comparisons use identical linguistic inputs.

A.3 Evidence-Grounded Item Construction

Item construction follows a progressive evidence chain. We first verify that the RAW and sRGB observations correspond to the same scene and identify a meaningful discrepancy between the two representations. This discrepancy is then converted into a pairwise reasoning question whose designated answer depends on evidence preserved in RAW but attenuated, obscured, or distorted in sRGB. This procedure ensures that item difficulty arises from differences in the imaging representation rather than artificial linguistic complexity.

Gemini-2.5-Flash generates an initial draft containing a question, two physically distinct candidate statements, a designated answer, and a concise evidence-based explanation. Human annotators subsequently inspect the matched RAW/sRGB observations and revise or reject the draft according to the verification and retention criteria described in the following subsection. For each retained item, the decisive physical observations are encoded as compact fact anchors that connect the designated answer to explicit sensor-domain evidence and support evidence-centered evaluation without requiring exact lexical overlap.

A.4 Human Verification and Retention Criteria

Human annotators inspect each matched RAW/sRGB pair and revise or reject the generated draft. An item is retained only if it satisfies six criteria: (1) the RAW and sRGB observations depict the same scene; (2) the two candidate statements express clearly distinguishable physical interpretations; (3) the RAW measurement provides sufficient evidence for the designated answer; (4) the decisive cue is attenuated, obscured, or distorted in the matched sRGB image; (5) the question and candidate statements do not introduce representation-specific hints; and (6) the hidden fact anchors adequately cover the evidence required for evaluation.

For each retained item, annotators convert the decisive observations into compact fact anchors and refine the reference reasoning. The anchors specify the required physical evidence without prescribing exact wording, allowing semantically equivalent explanations to receive credit. The scoring rubric further evaluates factual accuracy, awareness of imaging artifacts, causal consistency, and unsupported speculation. Items are rejected when the answer can be inferred from scene semantics, language priors, or cues already visible in sRGB. They are also rejected when the RAW evidence does not support an unambiguous conclusion.

B Blind Diagnostic Evaluation

B.1 Evaluation Inputs and Field Visibility

RAW Bench separates model inference, blind judging, and post-score analysis. During inference, conventional MLLMs receive the matched sRGB image, whereas RAW-aware models receive the corresponding RAW measurement. All non-visual inputs are held fixed, including the scene, question, candidate statements, option order, and response format. The evaluated model has no access to the correct option, fact anchors, reference reasoning, scoring focus, evidence-family labels, or dimension labels.

As detailed in Tab. 5, the judge receives only the saved model response, the question, hidden fact anchors, and the scoring focus. It does not receive the visual input, correct option, candidate statements, reference reasoning, or evidence labels. The correct option is used separately by the deterministic scorer to compute pairwise correctness, while evidence-family and dimension metadata are introduced only during post-score aggregation.

Table 5: Information during model inference, blind judging, and post-score analysis. The evaluated model receives either the RAW measurement or its matched sRGB image.

Information	Model Input	Judge Input	Post-Score Record
Visual input	✓	×	×
Question	✓	✓	×
Candidate statements	✓	×	×
Correct option	×	×	✓
Model-selected option	Generated	✓	✓
Model explanation	Generated	✓	✓
Fact anchors	×	✓	×
Scoring focus	×	✓	×
Reference reasoning	×	×	✓
Common-mistake annotations	×	×	✓
Evidence-family and dimension labels	×	×	✓

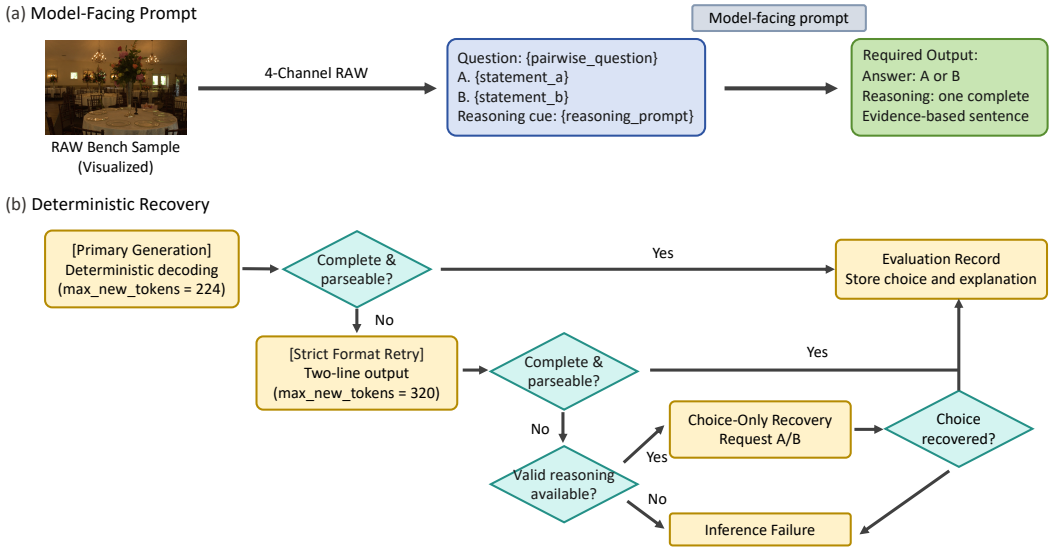

Figure 7: Model-facing prompt and deterministic response-recovery procedure. A strict-format retry handles malformed outputs, while choice-only recovery preserves valid reasoning when the selected option is unparseable.

B.2 Prompting and Response Recovery

All evaluated models receive the same semantic prompt fields and output requirements. As illustrated in Fig. 7, the prompt presents the question, two candidate statements, and a reasoning request, and instructs the model to return a discrete choice followed by a concise evidence-based explanation. Generation uses deterministic decoding with a maximum of 224 new tokens. If the response is malformed or truncated, we retry with a stricter two-line format and a maximum of 320 new tokens. When the explanation is valid but the selected option remains unparseable, we preserve the explanation and issue a final choice-only query requesting either A or B. Responses that remain incomplete after these steps are marked as inference failures and excluded from judging. All generation attempts and recovery outcomes are stored in the inference record.

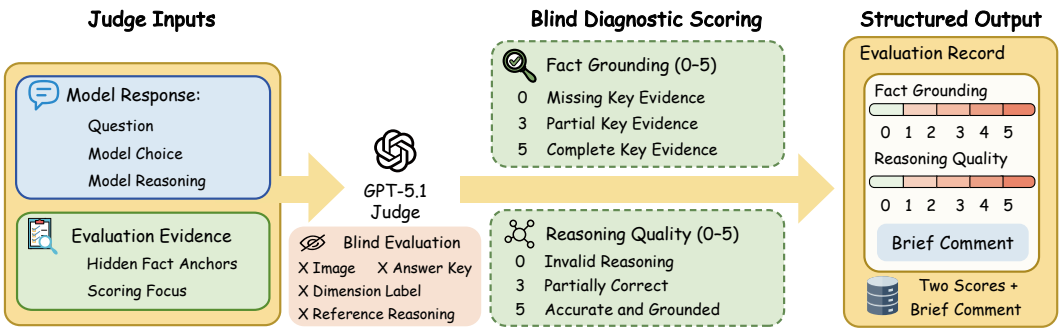

Figure 8: Blind diagnostic evaluation protocol. GPT-5.1 scores fact grounding and reasoning quality from the model response, hidden fact anchors, and scoring focus, without access to the image, answer key, dimension labels, or reference reasoning.

B.3 Judge Protocol and Score Computation

The judge evaluates each saved model response using only the question, model-selected option, model explanation, hidden fact anchors, and scoring focus (Fig. 8). It does not receive the visual input, correct option, candidate statements, reference reasoning, evidence-family labels, or dimension labels. This protocol prevents the judge from inheriting rendering artifacts present in sRGB images, while excluding candidate statements prevents superficial word-matching between options and explanations. We use GPT-5.1 with temperature zero and require structured JSON output for deterministic parsing and aggregation.

The judge assigns two scores on a 0–5 scale. Fact grounding measures whether the explanation recovers the decisive physical observations specified by the hidden anchors. Reasoning quality assesses whether those observations accurately and causally support the valid physical conclusion. Under the scoring rubric, explanations that cite relevant cues but draw an invalid physical conclusion receive strict penalties in reasoning quality. This decoupled design separates deterministic choice verification from semantic evidence extraction, ensuring that final scores reflect verified physical evidence rather than unsupported textual generation.

For item i, let $p_i \in \{0, 1\}$ denote pairwise correctness, and let $f_i, r_i \in \{0, \ldots, 5\}$ denote fact grounding and reasoning quality, respectively. The item score is

$$s_i = 0.2p_i + 0.4\frac{f_i}{5} + 0.4\frac{r_i}{5}.$$

The overall RAW Bench score is the mean of s_i across all items, with the same computation applied within each evidence dimension. This weighting assigns 80% of the score to evidence recovery and reasoning, limiting the contribution of unsupported correct guesses.

B.4 Human–LLM Judge Agreement Study

To rigorously assess the objectivity and consistency of GPT-5.1 blind scoring, we conduct a double-blind human agreement study on a stratified subset of 100 RAW Bench samples. The evaluation subset is balanced across all five evidence dimensions (20 instances per dimension) and covers all three difficulty tiers. Two independent human domain experts scored the blinded model explanations following the identical fact anchors and scoring focus used by GPT-5.1. Crucially, human raters were also blinded to the input images and candidate statement options, ensuring an identical evaluation information flow to GPT-5.1.

As detailed in Tab. 6 and Fig. 9, the empirical evaluation yields three critical observations:

1. **High Human–LLM Alignment**: The composite benchmark score evaluated by GPT-5.1 achieves a Pearson correlation of $r = 0.945$ and a Spearman rank correlation of $\rho = 0.891$ against human consensus, with a minimal MAE of 4.00 points across the 100-point scale.
2. **Elimination of Assessment Hallucinations**: When inspecting discrete scores, both human experts reach a 100.0% adjacent match ($\leq$ 1 grade margin) with GPT-5.1 on both fact

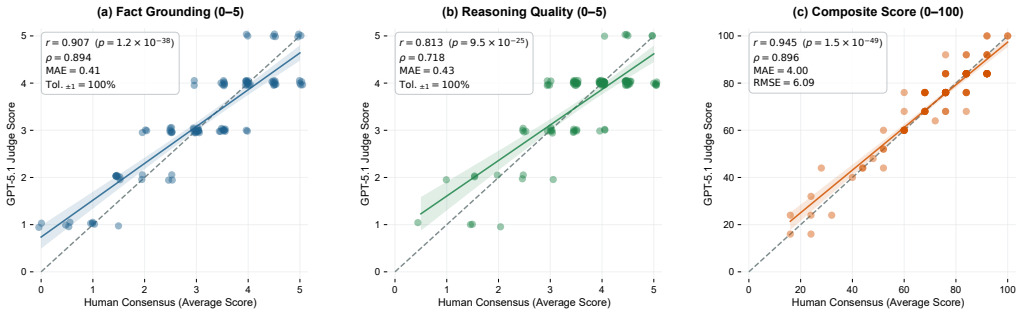

Figure 9: **Scatter distributions of Human Consensus vs. GPT-5.1 Judge on 100 stratified validation instances.** The dashed diagonal line indicates ideal agreement ($y = x$). (a) Fact grounding exhibits a Pearson correlation of $r = 0.907$ with an average discrepancy of 0.41 points. (b) Reasoning quality attains $r = 0.813$ with zero catastrophic outliers (100% within ± 1 grade). (c) The composite benchmark score achieves $r = 0.945$ with an MAE of only 4.00 on a 100-point scale, confirming that anchor-guided GPT evaluation closely mirrors expert human consensus.

Table 6: Quantitative agreement statistics between human domain experts and GPT-5.1 on 100 stratified RAW Bench instances. Adjacent Match denotes the percentage of score pairs differing by at most ≤ 1 point on the 0–5 discrete scale.

Evaluation Pair	**Metric Target**	**Pearson** r	**Spearman** ρ	**Kappa** κ_w	**MAE**	**Exact (%)**	**Adjacent (%)**
Human 1 vs. GPT-5.1	Fact Grounding ($0 \sim 5$)	0.814	0.808	0.796	0.60	40.0%	100.0%
	Reasoning Quality ($0 \sim 5$)	0.690	0.576	0.676	0.60	40.0%	100.0%
	Composite ($0 \sim 100$)	0.873	0.775	–	5.76	–	–
Human 2 vs. GPT-5.1	Fact Grounding ($0 \sim 5$)	0.844	0.809	0.811	0.60	40.0%	100.0%
	Reasoning Quality ($0 \sim 5$)	0.713	0.638	0.690	0.60	40.0%	100.0%
	Composite ($0 \sim 100$)	0.896	0.852	–	5.12	–	–
Inter-Human (H1 vs. H2)	Fact Grounding ($0 \sim 5$)	0.673	0.647	0.668	0.88	31.0%	81.0%
	Reasoning Quality ($0 \sim 5$)	0.490	0.367	0.489	0.84	33.0%	83.0%
	Composite ($0 \sim 100$)	0.751	0.639	–	9.60	–	–
Human Avg vs. GPT-5.1 (Consensus Upper Bound)	Fact Grounding ($0 \sim 5$)	**0.907**	**0.894**	–	**0.41**	–	–
	Reasoning Quality ($0 \sim 5$)	**0.813**	**0.718**	–	**0.43**	–	–
	Composite ($0 \sim 100$)	**0.945**	**0.891**	–	**4.00**	–	–

Dimension-wise Breakdown	**Sample Count** (n)	**Composite** (r)	**Fact Grounding** (r)	**Reasoning Quality** (r)
Color Constancy	20	0.944	0.903	0.735
Fine-grained Texture/OCR	20	0.978	0.941	0.872
Illuminant Reasoning	20	0.888	0.894	0.828
Dark/Highlight Object Detection	20	0.952	0.912	0.877
Low-light Scene Classification	20	0.851	0.897	0.486

grounding and reasoning quality, demonstrating that the LLM judge reliably extracts factual overlap without generating spurious extreme scores.

3. **Consensus Centroid Property**: Human-to-GPT agreement significantly surpasses inter-human agreement ($r = 0.945$ vs. $r = 0.751$). In line with standard psychometric evaluation findings, anchor-conditioned prompting positions GPT-5.1 as an objective rater that averages out individual human variance, confirming the reliability of RAW Bench diagnostic metrics.

C RAW-AWARE MODEL IMPLEMENTATION

C.1 RAW INPUT PREPARATION

Camera-specific RAW formats (e.g., DNG, ARW, RAF) are proprietary container files that package sensor readouts with vendor metadata. Because these formats differ in header structures, bit depths, and Bayer pattern layouts across camera manufacturers, multimodal training and evaluation require a unified sensor representation. Following standard computational photography protocols (Chen et al., 2018; Li et al., 2025b), we unpack native Color Filter Array (CFA) mosaics into standardized four-channel tensors, $x^r = (R, G_1, G_2, B)$. This preprocessing eliminates vendor-specific container discrepancies and directly exposes linear sensor measurements for downstream reasoning. Source-provided RAW arrays from existing datasets are mapped to this identical four-channel interface using their respective sensor specifications.

All arrays are converted to floating point and normalized to $[0, 1]$ according to their recorded or source-specific bit depth. Values are then clipped to the valid range and resized to 448×448 using bilinear interpolation. The resulting four-channel tensor is used by the RAW context pathway, while the pseudo-sRGB semantic input is derived from the same standardized measurement.

This preprocessing pipeline applies only channel packing, normalization, clipping, and resizing. It does not perform demosaicing, denoising, tone mapping, white-balance correction, or other ISP operations, thereby preserving the sensor-domain evidence used by the model.

C.2 PSEUDO-SRGB SEMANTIC INTERFACE

Given a packed RAW tensor $x^r = (R, G_1, G_2, B)$, we average the two green channels and concatenate the resulting R, G, and B planes along the channel dimension:

$$x^p = \mathrm{Concat}_{\mathrm{ch}}\left(R, \frac{G_1 + G_2}{2}, B\right).$$

This operation converts the four-channel RAW tensor into a three-channel pseudo-sRGB representation compatible with the pretrained vision encoder. The conversion uses only channel averaging and concatenation, without demosaicing, denoising, tone mapping, white-balance correction, or other ISP operations.

The resulting representation preserves scene layout and coarse chromatic structure, providing a semantic scaffold for producing language-aligned visual tokens. Because x^p and x^r are derived from the same standardized measurement, the pseudo-sRGB pathway supplies the pretrained semantic interface, while the RAW context pathway separately retains capture-level information.

C.3 RAW CONTEXT ENCODER AND RESIDUAL INJECTION

The RAW context encoder extracts compact capture-level information from the standardized four-channel input x^r. As summarized in Tab. 7, it consists of a convolutional stem followed by three progressively downsampled residual stages, global average pooling, and a two-layer projection head. Each convolution is followed by GroupNorm and SiLU activation, and each residual block contains two 3×3 convolutions. The encoder produces a 128-dimensional context vector,

$$c^r = E_r(x^r) \in \mathbb{R}^{128}.$$

A context-mapping network transforms c^r into a residual in the language-model token space:

$$\Delta r = W_2\ \mathrm{GELU}(W_1\ \mathrm{LN}(c^r)), \qquad U' = U + \mathbf{1}_N(\Delta r)^\top,$$

where $U \in \mathbb{R}^{N \times d}$ denotes the projected visual tokens, N is the number of visual tokens, and d is the language-model hidden dimension. The residual $\Delta r \in \mathbb{R}^d$ is broadcast across all visual tokens without changing their number or ordering.

We inject the residual after multimodal projection, allowing RAW-dependent context to recalibrate language-aligned visual tokens while preserving the pretrained semantic pathway. The final mapping layer W_2 is zero-initialized, so training begins with the pretrained pseudo-sRGB behavior and gradually learns RAW-dependent corrections.

Table 7: Architecture of the RAW context encoder. The base channel width is $b = 32$, and the encoder produces a 128-dimensional context vector.

Block	Kernel / Stride	Channels	Operation
Stem	$5 \times 5 \,/\, 2$	$4 \to b$	Convolution followed by a residual block
Stage 1	$3 \times 3 \,/\, 2$	$b \to 2b$	Downsampling convolution followed by a residual block
Stage 2	$3 \times 3 \,/\, 2$	$2b \to 4b$	Downsampling convolution followed by a residual block
Stage 3	$3 \times 3 \,/\, 2$	$4b \to 8b$	Downsampling convolution followed by a residual block
Pooling	–	$8b$	Global average pooling
Projection head	–	$8b \to 128$	Two-layer linear projection

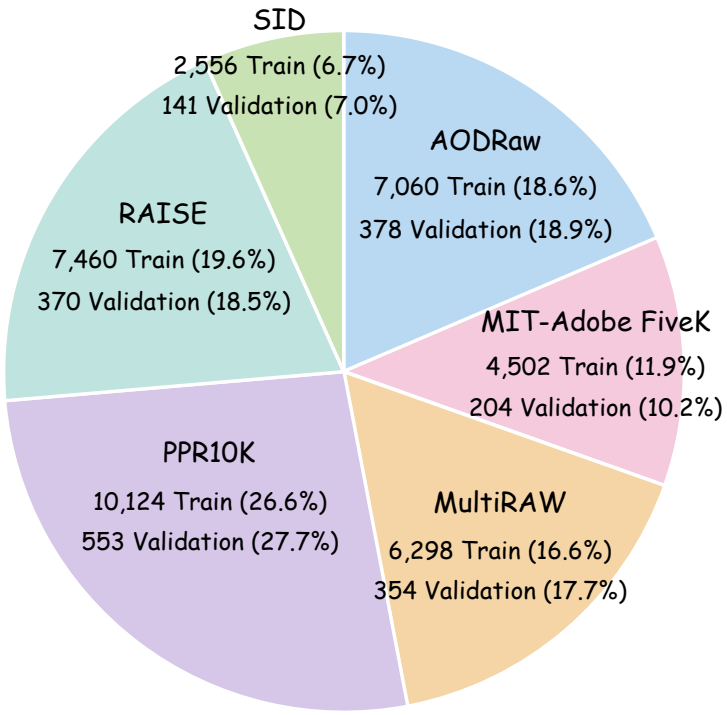

Figure 10: Source composition of the RAW–text instruction corpus across the training and held-out splits.

D Training and Implementation Details

D.1 Instruction Corpus

We construct a separate RAW–text instruction corpus to train the model to identify and explain sensor-domain evidence. The corpus contains 38,000 training records and 2,000 held-out records drawn from six RAW sources, as shown in Fig. 10. The training/validation counts are 7,060/378 for AODRaw, 4,502/204 for MIT-Adobe FiveK, 10,124/553 for PPR10K, 2,556/141 for SID, 7,460/370 for RAISE, and 6,298/354 for MultiRAW. Both splits are fully disjoint from RAW Bench at the scene and source-file levels, including RAW files, paired sRGB images, questions, annotations, and hidden evaluation records, all of which are reserved exclusively for RAW Bench.

Table 8: Training configuration for the two-stage optimization of RAW-InternVL. Stage A aligns RAW context with the pretrained visual-token space, while Stage B introduces language-side LoRA adaptation.

Setting	**Stage A**	**Stage B**
Initialization	InternVL3.5-8B	Stage-A checkpoint
Training epochs	1	1
Learning rate	1×10^{-4}	5×10^{-5}
Trainable modules	RAW context encoder, residual mapper, and multimodal projector	Stage-A modules and LoRA adapters
Frozen modules	Vision encoder and language model	Vision encoder and base language model
LoRA target modules	–	`q_proj`, `v_proj`, and `o_proj`
LoRA Settings $(r, \alpha, \text{dropout})$	–	$(16, 32, 0.05)$
RAW resolution	448×448	448×448
Maximum text length	1,536 tokens	1,536 tokens
Numerical precision	bfloat16	bfloat16
Attention implementation	SDPA	SDPA
Per-device batch size	1	1
Gradient accumulation steps	16	16
Learning-rate schedule	Cosine decay with 3% warmup	Cosine decay with 3% warmup
Weight decay	0	0
Random seed/data seed	42 / 42	42 / 42
Checkpoint retention	One checkpoint at the end of the stage	One checkpoint at the end of the stage

Gemini-2.5-Flash generates all 40,000 instruction records and rewrites 19,853 of them to improve linguistic clarity and diversity. Each record contains a RAW input, a pairwise question, two candidate statements, a designated answer, and a concise evidence-based explanation. During training, the data loader reconstructs the prompt from the question, candidate statements, and reasoning request, while the language-model loss is applied only to the selected option and target explanation.

We further categorize the instructions by their degree of evidence binding. Strong-binding items explicitly connect the designated conclusion to RAW-sensitive evidence, whereas weak-binding items provide broader sensor-domain supervision. The training split contains 33,410 strong-binding and 4,590 weak-binding items, while the held-out split contains 1,784 and 216, respectively. We assign weights of 1.0 and 0.2 to strong- and weak-binding items, respectively, and normalize the weights to preserve a unit mean over the dataset. This scheme emphasizes evidence-bearing supervision while retaining contextual diversity.

D.2 Two-Stage Optimization

We adopt a two-stage optimization strategy to separate RAW representation alignment from language adaptation. As detailed in Table 8, Stage A keeps the pretrained vision encoder and language model frozen while optimizing the RAW context encoder, residual mapping network, and multimodal projector. This stage aligns capture-level RAW context with the pretrained visual-token space without disrupting the backbone's existing semantic representations.

In Stage B, the visual-side modules continue to be optimized, and LoRA adapters are introduced into the language model's `q_proj`, `v_proj`, and `o_proj` modules. This stage adapts the language

Table 9: Effect of RAW-context injection depth on RAW Bench performance. Internal-layer variants report the mean and sample standard deviation over three runs. The post-projector result corresponds to RAW-InternVL reported in the main paper. Bold indicates the best internal-layer result in each column; bold values in the post-projector row indicate performance exceeding all internal-layer variants.

Injection Point	O	LL	D/H	CC	IR	T/O
After block 6	69.05 $\pm$0.16	**74.95 $\pm$0.09**	**67.14 $\pm$0.08**	69.92 $\pm$0.19	70.93 $\pm$0.19	63.17 $\pm$0.26
After block 12	**69.19 $\pm$0.15**	73.69 $\pm$0.12	66.79 $\pm$0.12	**71.19 $\pm$0.12**	71.26 $\pm$0.22	**63.62 $\pm$0.17**
After block 24	68.84 $\pm$0.20	72.91 $\pm$0.19	66.64 $\pm$0.24	70.51 $\pm$0.19	**72.01 $\pm$0.19**	62.81 $\pm$0.19
After projector	**70.43**	**77.09**	**67.58**	**72.17**	**74.78**	61.36

model to express RAW-conditioned evidence through pairwise decisions and concise physical explanations while keeping the language-model weights frozen.

D.3 Training Configuration

Tab. 8 summarizes the training settings for both optimization stages. Stage A is initialized from InternVL3.5-8B, while Stage B resumes from the Stage-A checkpoint. Each stage is trained for one epoch using learning rates of 1×10^{-4} and 5×10^{-5}, respectively. Both stages use 448×448 RAW inputs, a maximum text length of 1,536 tokens, bfloat16 precision, and scaled dot-product attention. We use a per-device batch size of one with 16-step gradient accumulation, a cosine learning-rate schedule with 3% warmup, and zero weight decay. In Stage B, LoRA is applied to the language model's `q_proj`, `v_proj`, and `o_proj` modules with rank 16, scaling factor 32, and dropout 0.05. The random and data seeds are fixed to 42, and one checkpoint is retained at the end of each stage.

E Additional Experimental Analysis

E.1 Fusion-Depth Analysis

We examine how the location of RAW context injection affects sensor-grounded reasoning. All variants use the same RAW encoder, residual formulation, instruction corpus, two-stage optimization, prompting protocol, and scoring procedure; only the fusion location is changed. InternVL3.5-8B contains 24 vision-transformer blocks; we select blocks 6, 12, and 24 to represent early, intermediate, and terminal visual stages, comparing them with our post-projector design.

As shown in Tab. 9, the preferred fusion depth varies across evidence dimensions. Injection after block 6 performs best among the internal-layer variants on low-light and dark/highlight evidence, while block 12 achieves the strongest internal-layer overall score and leads on color constancy and texture/OCR. Injection after block 24 performs best among the internal variants on illuminant reasoning. This pattern is consistent with a progression from local photometric and structural cues in earlier layers toward more globally integrated scene context in later layers.

Post-projector injection achieves the highest overall score and performs best on low-light, dark/highlight, color constancy, and illuminant reasoning. At this stage, the vision encoder has completed semantic encoding, while the projected visual tokens can be recalibrated using RAW context before language reasoning. Block-12 injection remains stronger on texture/OCR, suggesting that fine-grained spatial evidence may benefit from fusion before semantic compression is complete. These results support post-projector injection for global RAW context while indicating a complementary role for intermediate-layer fusion in fine-detail tasks.

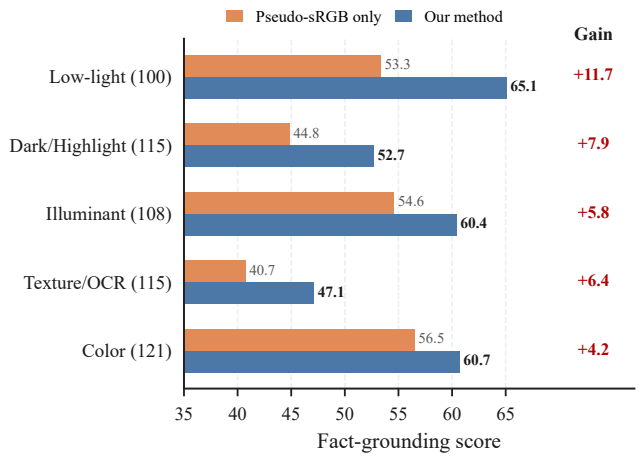

Figure 11: Dimension-wise fact-grounding comparison of Pseudo-sRGB only and RAW-InternVL.

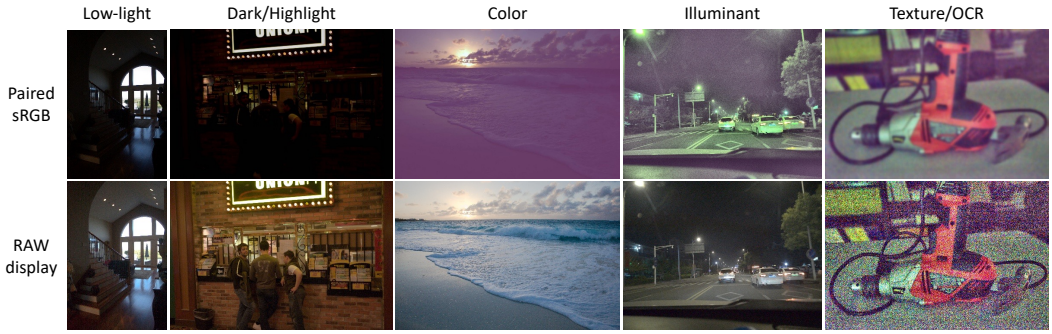

Figure 12: Representative matched sRGB/RAW scenes across the five RAW Bench dimensions. RAW images are rendered only for visualization and are not used as model inputs.

E.2 Dimension-Wise Fact-Grounding Diagnosis

We provide the detailed dimension-wise fact-grounding comparison between the pseudo-sRGB baseline and our full RAW-InternVL model across all five evidence categories. As visualized in Fig. 11, injecting explicit RAW context yields substantial gains across all dimensions, with the most pronounced improvements observed in low-light (+11.7) and dark/highlight (+7.9) scenarios.

E.3 Representative Examples Across Evidence Dimensions

Fig. 12 presents representative matched sRGB/RAW scenes across the five RAW Bench dimensions: low-light visibility, dark/highlight evidence, color constancy, illuminant reasoning, and texture/OCR. Each column shows the same captured scene under both representations, illustrating physical cues preserved in RAW but weakened, obscured, or distorted in sRGB, including shadow structure, saturated detail, intrinsic color, illumination patterns, and fine text or texture.

The displayed RAW images are used only for visual inspection. DNG files are rendered using camera white balance and global automatic brightness in `rawpy`, while the AODRaw array is normalized by its global 0.5th and 99.5th channel percentiles and displayed with a $1/2.2$ gamma correction. Training and evaluation instead use the standardized four-channel RAW tensors defined in the previous section, without these display renderings.

Table 10: Taxonomy and dimension-wise distribution of failure modes in RAW-InternVL on RAW Bench.

Failure Mode	**Physical Root Cause**	**Share (%)**
Sub-pixel OCR Degradation	Global residual broadcasting fails to reconstruct high-frequency character strokes smoothed during spatial downsampling.	26.42%
Dark/Highlight Ambiguity	Severe dynamic range compression or deep shadows obscure intrinsic surface properties and material reflectance.	23.83%
Color Constancy Spuriousness	Global processing casts or illuminant shifts are misattributed to intrinsic object color or masked by color noise.	22.80%
Low-light / Illuminant Blur	Ultra-low-light shot noise or complex optical reflections confound fine-grained causal reasoning.	26.94%

E.4 Failure Mode Analysis and Error Taxonomy

To establish rigorous diagnostic transparency and expose the physical boundaries of RAW-conditioned reasoning, we perform a fine-grained error analysis on failure instances produced by RAW-InternVL. Out of 559 evaluation items, 193 instances (34.53%) trigger our diagnostic screening conditions (defined as option error $p_i = 0$, fact grounding $f_i < 2.5/5.0$, or composite score $s_i < 0.5$). However, manual auditing reveals that this set predominantly reflects strict judge criteria and partial explanations rather than blind guesses; strictly incorrect option selections ($p_i = 0$) occur in only 51 items (9.12%). A key empirical insight is revealed: *correct binary choices are frequently unaccompanied by fully specific, detail-grounded explanations.*

Dimension-wise error distribution. As structured in Tab. 10, failures concentrate in *Texture/OCR* (44.35% dimension error rate) and *Dark/Highlight* (40.00% dimension error rate). Synthesizing predicted rationales with reference reasoning, we identify three primary error patterns:

1. Generalized descriptions substituting for specific evidence. A frequent pattern involves the model invoking generic capture descriptors (e.g., "low-light", "underexposed") without isolating the local decisive cue. For instance, in traffic signal evaluation, the model correctly deduces operational status from spatial priors but omits explicit countdown digits (e.g., "53"). This accounts for the 142 items where selection was correct ($p_i = 1$) yet fact grounding remained partial.

2. Conflating capture conditions with intrinsic object properties. The model occasionally recognizes illumination artifacts (such as deep shadows or global white-balance casts) but fails to decouple them from surface reflectance, leading to contradictions between its perceptual premise and final conclusion (e.g., attributing deep shadow attenuation to dark intrinsic material paint).

3. Misinterpreting sensor noise as high-frequency textures. In ultra-low-light acquisitions (≤ 0.01 lux), high-frequency color shot noise and read noise are sometimes misconstrued as physical surface details. Because our context pathway injects a uniform global shift Δr, it lacks localized spatial adaptability to filter out severe noise at the sub-pixel level, justifying future exploration into spatially resolved RAW adapters.