ASKLIGHT: GROUNDING ACTIVE-LIGHT REASONING FOR MULTI-ILLUMINANT WHITE BALANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-illuminant white balance requires a spatially varying illumination estimate, but visible light sources poorly indicate the actual illumination: they may be switched off, miss sources outside the view, or share a region with other sources. We present **AskLight**, which connects semantic source reasoning to continuous illumination decomposition through an explicit source-to-region representation. A shared vision–language model (VLM) classifies detected proposals, then revises the active inventory by removing false actives, adding missed-visible or hidden hypotheses, and associating every retained hypothesis with a dominant illumination region. These relations initialize a variable-cardinality set of illuminant colors and spatial priors, which a recurrent decomposition backend refines into mixture weights and candidate chromaticities through differentiable physical composition. A central obstacle is that no real dataset provides per-source supervision of which sources are active, which are visible, and where each one casts light. We therefore introduce **Infinigen-Light**, a procedurally rendered multi-illuminant dataset with per-source annotations of activity, visibility, and region influence, which makes active-light reasoning both learnable and separately evaluable from final white balance accuracy. By grounding each illuminant candidate in an explicit source hypothesis rather than an unstructured slot or a global estimate, our method achieves the lowest mean illumination error on Infinigen-Light, LSMI, and MIIW, reducing the LSMI illumination error by 16.8% over the strongest baseline. We will release our source code.

1 INTRODUCTION

Consider a café at dusk. Daylight enters through a window, warm pendant lamps illuminate the counter, and an unseen source contributes light from an adjacent room. A camera records the combined illumination, so the same nominally neutral surface can exhibit different color casts across the image (Finlayson et al., 1995). Correcting this scene requires more than a global white point: the camera must estimate how the illuminant chromaticity varies spatially. Multi-illuminant datasets and decomposition models explicitly address this setting (Kim et al., 2021; 2024).

The relevant question is not *which objects look like lights?* but *which sources explain the illumination reaching the scene?* A visible source may be off, a source may illuminate the scene from outside the view, and an illumination region may be shared by several sources or split across disconnected surfaces. Treating detections, regions, and illuminant components as interchangeable therefore imposes an inappropriate correspondence.

Photometric methods and vision–language models (VLMs) offer complementary evidence. Gray-Pixel statistics provide image-derived illuminant cues without requiring source visibility (Qian et al., 2019), whereas VLMs use scene context and object appearance to make structured judgments. However, VLM perception of fine image attributes is not uniformly reliable (Wu et al., 2024a), so asking a VLM to directly output camera-space illuminant RGB would place the numerical burden on the least constrained part of the system. Recent VLM-guided color constancy instead uses qualitative feedback to refine a global white point (Li et al., 2026); our focus is different: a variable set of simultaneous source hypotheses and their spatial influence.

Learning this form of reasoning requires supervision connecting source semantics to spatial illumination, which a target illumination field alone does not provide (Kim et al., 2021). We therefore

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.

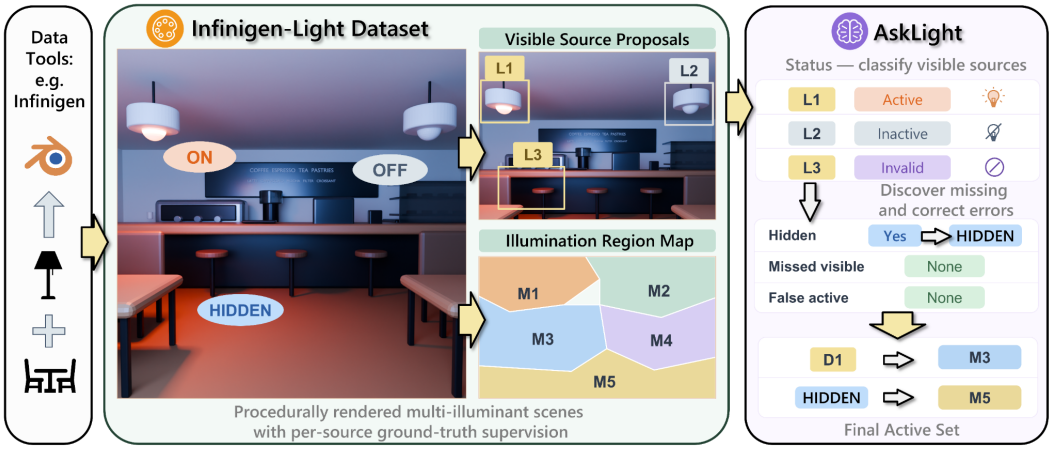

Figure 1: Overview of **Infinigen-Light** dataset and **AskLight**. ***Left***: Infinigen-Light provides procedurally rendered multi-illuminant scenes with per-source annotations of activity, visibility, and illumination region. ***Right***: AskLight classifies visible sources, revises the active-source inventory, and associates each source with a dominant region.

introduce **Infinigen-Light** (Fig. 1 left), a synthetic multi-illuminant dataset procedurally rendered from Infinigen-Indoor scenes (Raistrick et al., 2023; 2024), pairing mixed-illumination images and white-balanced targets with per-source annotations of activity, visibility, and spatial contribution. Crucially, its contribution maps tie each source to its spatial influence, allowing source to region supervision to be derived from illumination regions rather than source locations (see Fig. 2). This makes Infinigen-Light a resource for both learning active-light reasoning and evaluating source understanding separately from final white balance accuracy.

Building on this supervision, we introduce **AskLight**, which uses a source-to-region representation as the interface between semantic reasoning and photometric estimation. A shared VLM first classifies frozen detector proposals as `ACTIVE`, `INACTIVE`, or `INVALID`, then revises the inventory: removing false actives, supplementing missed-visible and hidden-source hypotheses, and assigning each retained hypothesis to a dominant region. Region assignment is many-to-one and refers to source influence rather than lamp location. A deterministic module then turns these discrete relations into candidate colors and soft priors, which a shared recurrent backend refines into spatial weights and chromaticities through differentiable physical composition, yielding a spatially varying illumination field for multi-illuminant white balance.

Our contributions are threefold.

- Formulating active-light reasoning as inventory revision and source-to-region association, we separate source validity, visibility, and spatial illumination support.
- We introduce **Infinigen-Light**, a procedurally rendered multi-illuminant dataset with per-source activity, visibility, and region-influence annotations that real datasets cannot provide, making active-light reasoning both learnable and separately evaluable from final white balance accuracy.
- We propose **AskLight**, a VLM that uses the source-to-region representation to initialize a variable-cardinality recurrent decomposition that jointly refines spatial mixtures and illuminant chromaticities, achieving state-of-the-art (SOTA) multi-illuminant white balance on both synthetic and real-world benchmarks.

2 Related Work

Statistical and learned color constancy. Color constancy estimates the scene illuminant so that captured colors can be corrected toward a canonical appearance (Gijsenij et al., 2011). Classical statistical methods rely on low-level assumptions such as Grey-World, White-Patch, and their higher-order Grey-Edge extensions (Buchsbaum, 1980; van de Weijer et al., 2007). Learning-based estimators improve accuracy: like FFCC (Barron & Tsai, 2017) performs frequency-domain inference,

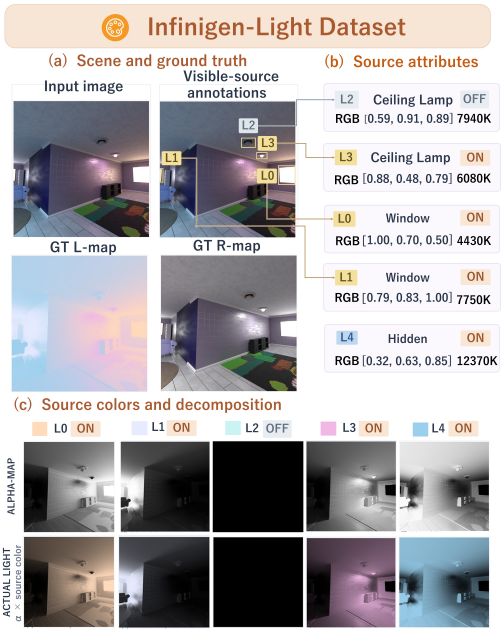

Figure 2: **Infinigen-Light dataset.** (a) Annotated input and targets $L^\star$, $R^\star$. (b) Source attributes. (c) Source contributions.

FC4 (Hu et al., 2017) learns confidence-weighted feature pooling, and CNN-based regressors directly predict the illuminant (Lou et al., 2015). Cross-camera approaches further address sensor dependence, with C5 (Afifi et al., 2021) adapting convolutional filters from image content and CCMNet (Kim et al., 2025) using calibrated color correction matrices. A few methods further address color constancy specifically in nighttime scenes (Lee et al., 2026; Li & Tan, 2024). These methods provide useful baselines, but a global estimator is not inherently a dense multi-illuminant model, and its pixel-wise adaptation must be specified rather than assumed.

Multi-illuminant white balance correction. Unlike global color constancy, multi-illuminant settings require a spatially varying estimate. Early methods extend low-level statistics locally: Gray-Pixel, retinex-style cues, and radiance clustering methods estimate spatially varying illumination from local image statistics (Gu et al., 2014; Yang et al., 2015; Qian et al., 2019). Learning-based methods later added dedicated data and models: LSMI (Kim et al., 2021) provides RAW observations with dense mixed-illumination supervision, and AID (Kim et al., 2024) represents illuminants as slots combined with spatial weights. IFFCC (Wei et al., 2025) extends FFCC to spatial illumination via integral histograms, and nonlinear preset fusion operates in rendered sRGB rather than linear RGB (Serrano-Lozano et al., 2025). Prior methods treat illuminants as anonymous components; AskLight instead grounds each in a source hypothesis and the region it lights.

VLMs for color and illumination understanding. Recent work probes how well VLMs perceive low-level color and lighting attributes. Benchmarks such as Q-Bench (Wu et al., 2024a), Q-Instruct (Wu et al., 2024b), and ColorBench (Liang et al., 2025) reveal persistent gaps in fine-grained color and lightness perception, and Sun et al. (Sun et al., 2026) report similar instability for color constancy under varying illumination. Other work uses VLMs to reason about photometric attributes (Wang et al., 2026), or drive color and tone editing (Dutt et al., 2025; Lin et al., 2025; Yao et al., 2026). Most related to us, VLM-CC (Li et al., 2026) recasts single-illuminant color constancy as iterative VLM-guided feedback, yet remains limited to a global uniform correction. We extend this paradigm to multi-illuminant scenes, where the VLM grounds source hypotheses in illumination regions to guide a pixel-wise illumination field.

3 Infinigen-Light Dataset

Controlled rendering. To complement existing synthetic indoor datasets (Roberts et al., 2021; Li et al., 2021) with dedicated supervision for multi-illuminant white balance, Infinigen-Light provides source-level supervision for multi-illuminant reasoning (see Fig. 2). We construct it from procedurally generated Infinigen Indoors scenes (Raistrick et al., 2024), each containing 3–5 registered light sources. We render multiple lighting states by changing which sources are switched on, while keeping the scene geometry, camera, source placement, source colors, and source powers fixed, varying only the active-source configuration. Unregistered illumination and environment emission are disabled. The mixed image I is rendered directly under each active-source configuration rather than synthesized from source-isolated renders.

Source-level supervision. Each source m is annotated with its type, linear-RGB color, activity state, visibility, visible-source bounding box, and dense contribution map α. The contribution map describes where the source casts light rather than where its source is located. Each lighting state

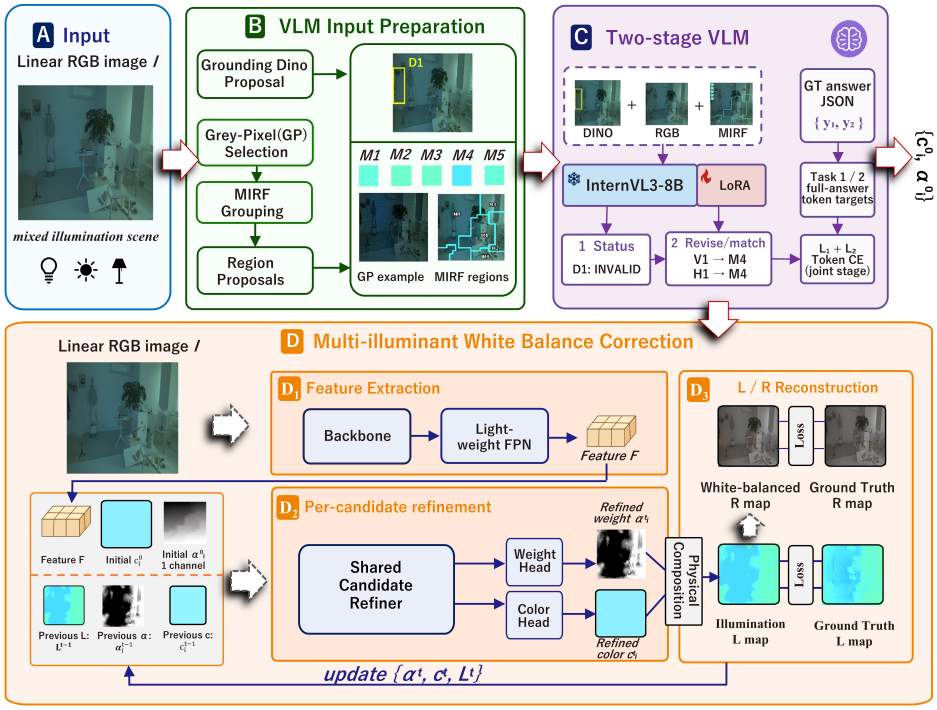

Figure 3: **Overview of AskLight.** (A) Linear RGB input. (B) Visible-source proposals and coarse illumination regions. (C) A shared VLM classifies proposals (Task 1) and revises the source inventory with region associations (Task 2). (D) Each retained source initializes one candidate: (D_1) feature extraction, (D_2) three-round refinement of mixture weights and chromaticities, and (D_3) composition of illumination L and white-balanced output R.

(i.e., a specific configuration of active and inactive light sources in the scene) additionally provides the dense illumination target L and white-balanced reference R. These annotations provide the supervision used for visible-source status reasoning and source-to-region grounding in Sec. 4.1. All lighting states from the same scene are assigned to the same split. Detailed rendering, target construction, and quality-control procedures are provided in Appendix A.

4 AskLight

Given a linear RGB image I, AskLight models the illumination L at each pixel p as a mixture of N illuminant candidates and recovers the white-balanced image R:

$$R(p) = \frac{I(p)}{L(p) + \epsilon}, \qquad L(p) = \sum_{i=1}^{N} \alpha_i(p)\, c_i, \tag{1}$$

where $c_i = [r_i, 1, b_i]^\top$ is the green-normalized chromaticity of candidate i, and the mixture weights satisfy $\alpha_i(p) \geq 0$ and $\sum_i \alpha_i(p) = 1$.

AskLight decides how many candidates there are and where each one starts by reasoning about light sources: each *source* retained by the VLM yields one *candidate* (c_i, α_i). In Fig. 3, the VLM judges the detected proposal D_1 invalid, adds a missed visible source V_1 and a hidden source H_1, and links both to the same illumination region; the two resulting candidates are initialized from that region and refined by a recurrent network into L. Sec. 4.1 covers source reasoning (Fig. 3(B, C)), Sec. 4.2 the decomposition (Fig. 3(D)), and Sec. 4.3 the losses.

4.1 Active-Light Reasoning

Input preparation. Two annotated views are prepared for the VLM (Fig. 3(B)). A frozen Grounding DINO (Liu et al., 2024) detects visible-source proposals $\{D_j\}$ and numbers them. At the same

time, Gray-Pixel (GP) selection (Qian et al., 2019) extracts local illuminant cues from the linear RGB image, and a procedure inspired by the Multi-Illuminant Random Field (MIRF) (Beigpour et al., 2014) groups these cues with spatial regularization into coarse illumination regions $\{M_k\}$. Overlaying the numbered proposals, or the region boundaries and identifiers, on the preview gives the *proposal view* and the *region view*.

Task 1: visible-source status. A single InternVL3-8B (Zhu et al., 2025) performs both tasks with shared parameters and task-specific inputs (Fig. 3(C)). We freeze its pretrained weights and train Low-Rank Adaptation (LoRA) (Hu et al., 2021) adapters in the vision and language branches, together with the vision–language projector. Given the preview and the proposal view, Task 1 labels each proposal as `ACTIVE` (illuminating), `INACTIVE` (switched off), or `INVALID` (false or redundant). For the scene in Fig. 3, it outputs

```
{"sources": [{"id": "D1", "status": "INVALID"}]}
```

The predicted statuses are drawn on the preview as a *status view* and passed to Task 2.

Task 2: inventory revision and region association. Given the preview, the Task 1 outputs, and the region view, Task 2 removes false actives, adds missed visible (V_i) and hidden or out-of-view (H_i) sources, and links each retained source to the region it lights most:

```
{"active": [], "extra": [["V1", "M4"], ["H1", "M4"]]}
```

Several sources may share a region, as V_1 and H_1 do here, and the number of entries in `active` and `extra` gives the candidate count N.

Source-grounded candidate initialization. Each retained source then initializes one candidate (left of Fig. 3(D)). We collect the Gray Pixels inside its assigned region; when several sources share a region, these pixels are split by position into one group per source. Each group gives an initial color $c_i^{(0)}$, normalized to a green channel of one, and a center μ_i at its mean pixel position; if too few Gray Pixels are available, we use the region color and center instead. The initial weight $\alpha_i^{(0)}$ combines a Gaussian centered at μ_i with the blurred region mask and is normalized across candidates at each pixel, so every candidate keeps a nonzero weight outside its region.

4.2 Recurrent Illumination Decomposition

RepViT-M1.1 (Wang et al., 2024) with a lightweight FPN (Lin et al., 2017) extracts a shared feature map F once per image (Fig. 3(D_1)). Starting from $L^{(0)} = \sum_i \alpha_i^{(0)} c_i^{(0)}$, the backend refines all candidates over three rounds (Fig. 3(D_2)):

$$\left(\alpha_i^{(t)}, c_i^{(t)}\right) = f_\theta\left(F, \mu_i, c_i^{(0)}, \alpha_i^{(0)}, \alpha_i^{(t-1)}, c_i^{(t-1)}, L^{(t-1)}\right), \qquad t \in \{1,2,3\}, \tag{2}$$

where f_θ consists of a shared candidate refiner, a *Weight Head* that predicts mixture weights, and a *Color Head* that predicts bounded residual chromaticity updates. All parameters are shared across candidates and rounds. Each round recomposes the illumination and the white-balanced image (Fig. 3(D_3)):

$$L^{(t)}(p) = \sum_i \alpha_i^{(t)}(p)\, c_i^{(t)}, \qquad R^{(t)}(p) = \frac{I(p)}{L^{(t)}(p) + \epsilon}. \tag{3}$$

Because this composition is differentiable, the losses on $L^{(t)}$ and $R^{(t)}$ train both heads without matching candidates to annotated sources.

4.3 Supervision and Losses

AskLight is trained in two stages: the VLM is first optimized with $\mathcal{L}_{\text{VLM}}$ and then frozen, after which the recurrent backend is optimized with $\mathcal{L}_{\text{WB}}$. No gradient flows from $\mathcal{L}_{\text{WB}}$ to the VLM.

Region labels. Task 2 labels are built from the ground-truth contribution maps and are used only for training. For each active source m, we threshold $\alpha_m^\star$ to obtain a region G_m and, over valid pixels,



Figure 4: **Qualitative comparison on Infinigen-Light and LSMI.** The first two groups are from Infinigen-Light dataset, and the last two are from LSMI dataset (Kim et al., 2021).

select the MIRF region with the largest Dice overlap (Dice, 1945):

$$k_m^\star = \arg\max_k \frac{2|G_m \cap M_k|}{|G_m| + |M_k| + \epsilon}, \tag{4}$$

VLM loss. Each task is trained with token-level cross-entropy over its target JSON answer $y_q^\star$:

$$\mathcal{L}_q = -\frac{1}{T_q} \sum_{u=1}^{T_q} \log p_\phi\left(y_{q,u}^\star \mid \mathbf{Z}_q,\, y_{q,<u}^\star\right), \quad q \in \{\text{status}, \text{region}\} \text{ for Task 1 and Task 2}, \tag{5}$$

where $\mathbf{Z}_q$ denotes the images and text prompt given to task q (Sec. 4.1), T_q is the number of answer tokens, and p_ϕ is the predicted token probability; image, prompt, and padding tokens are excluded.

We first train Task 1 with $\mathcal{L}_{\text{status}}$, then Task 2 from its checkpoint with $\mathcal{L}_{\text{region}}$, and finally jointly minimize $\mathcal{L}_{\text{VLM}} = \mathcal{L}_{\text{status}} + \mathcal{L}_{\text{region}}$.

White-balance loss. Each round t is supervised on both the illumination field and the white-balanced image:

$$\mathcal{L}^{(t)} = \sum_{Y \in \{L,R\}} \Big[D_{\text{mse}}\big(Y^{(t)}, Y^{\star}\big) + \lambda_{\text{ang}}\, D_{\text{ang}}\big(Y^{(t)}, Y^{\star}\big) \Big], \tag{6}$$

where D_{mse} averages over RGB channels and valid pixels, D_{ang} is the angular error in radians averaged over valid pixels with nonzero RGB vectors, and $\lambda_{\text{ang}} = 0.05$. Following RAFT (Teed & Deng, 2020), later rounds are emphasized with geometric weights:

$$\mathcal{L}_{\text{WB}} = \frac{\sum_{t=1}^{3} \gamma^{3-t}\, \mathcal{L}^{(t)}}{\sum_{t=1}^{3} \gamma^{3-t}}, \qquad \gamma = 0.8. \tag{7}$$

5 Experiments

5.1 Experimental Setup

Datasets and metrics. We evaluate on Infinigen-Light, the Galaxy, Nikon, and Sony subsets of LSMI (Kim et al., 2021) and their combination (All), and a mixed-illumination variant of MIIW (Murmann et al., 2019). For Infinigen-Light, it contains 2,992 images from 672 scenes, split by scene into 2,096/450/446 images (471/101/100 scenes) for training, validation, and testing. For LSMI, we use the official camera-wise splits. For MIIW (Murmann et al., 2019), since it provides single-illumination captures, we combine multiple illumination conditions of each photographed scene to construct mixed-illumination samples with corresponding targets, yielding 4,440 samples from 1,110 scenes with scene-disjoint splits (Appendix C). We report angular errors in degrees for the illumination field L and the corrected image R, averaged over valid pixels per image and then over images; P25/P50/P75 summarize the per-image error distribution.

Baselines. We compare with single-illuminant methods (C5 (Afifi et al., 2021), CCMNet (Kim et al., 2025), FC4 (Hu et al., 2017), GCC (Chang et al., 2025), VLM-CC (Li et al., 2026)) and multi-illuminant methods (LSMI-U (Kim et al., 2021), AID (Kim et al., 2024)). All baselines except GCC are retrained on the same splits; GCC uses its public pretrained models. Calibration targets are excluded from the inputs and supervision of all methods (Appendix C.5).

Implementation. We first train InternVL3-8B on Infinigen-Light with supervised Task 1 and Task 2 learning. From this checkpoint, we separately adapt the VLM to LSMI and MIIW for three epochs. Since these datasets lack complete source-level annotations, LSMI adaptation supervises only Task 2, whereas MIIW adaptation additionally supervises Task 1 under an all-hidden formulation. Ours (Inf.) uses the Infinigen-Light checkpoint, and Ours (LSMI) and Ours (MIIW) use the corresponding adapted checkpoints. Ours (LSMI) models share the LSMI-adapted VLM but use independently trained backends. The VLM is frozen during all downstream training, and only the recurrent illumination decomposition network is optimized. Each backend takes 512×512 linear RGB inputs, performs three refinement rounds, and is trained with AdamW for 100 epochs with batch size 4. Further details are provided in Appendix C.

5.2 White Balance Results

Quantitative comparison. Table 1 reports results on Infinigen-Light and LSMI (Kim et al., 2021). On Infinigen-Light, AskLight reduces the mean illumination error from 4.365° (LSMI-U) to 3.935°, a 9.9% reduction, and lowers the P75 error from 5.597° to 4.930°, indicating fewer large failures. Single-illuminant methods predict a single global illuminant and cannot represent spatially varying light, which is most evident for VLM-CC (8.101°), whose VLM-guided correction remains global. On LSMI, we evaluate two training settings: joint training on the complete **LSMI-ALL** dataset, which combines images from all three cameras, and device-specific training where models are trained separately on each camera subset:**Galaxy,Nikon and Sony**. AskLight achieves the lowest mean illumination error on every camera subset, with reductions of 10.5%, 12.7%, and 7.5% over the strongest baseline on Galaxy, Nikon, and Sony, respectively, and 16.8% on the combined set.

Table 1: Angular errors on Infinigen-Light and LSMI (Kim et al., 2021). "All" denotes the combination of the Galaxy, Nikon, and Sony subsets. P25/P50/P75 denote per-image error percentiles. **Best** and second-best results are highlighted.

Dataset	*Subset*	*Method*	L error (°)				R error (°)			
			Mean↓	P25↓	P50↓	P75↓	Mean↓	P25↓	P50↓	P75↓
Infinigen-Light	–	C5 (Afifi et al., 2021)	5.154	3.042	4.494	6.605	5.034	2.990	4.376	6.565
		CCMNet (Kim et al., 2025)	5.386	3.229	4.886	7.181	5.256	3.173	4.751	7.082
		FC4 (Hu et al., 2017)	4.638	3.010	4.206	5.899	4.560	2.914	4.132	5.809
		GCC (Gehler) (Chang et al., 2025)	5.111	3.357	4.803	6.369	4.983	3.190	4.683	6.258
		AID (Kim et al., 2024)	5.234	3.128	4.624	6.889	5.100	3.029	4.404	6.749
		VLM-CC (Li et al., 2026)	8.101	4.381	7.524	11.125	7.570	4.227	7.147	10.454
		LSMI-U (Kim et al., 2021)	4.365	2.525	3.979	5.597	4.217	2.592	3.863	5.258
		Ours (Inf.)	**3.935**	**2.399**	**3.484**	**4.930**	**3.888**	**2.311**	**3.390**	**4.890**
LSMI	Galaxy	C5 (Afifi et al., 2021)	2.361	0.783	1.746	3.199	3.179	**1.304**	2.707	4.329
		CCMNet (Kim et al., 2025)	2.421	**0.756**	**1.684**	3.304	3.245	1.307	2.536	4.444
		FC4 (Hu et al., 2017)	2.303	1.013	1.733	2.932	3.059	1.662	2.502	4.081
		GCC (Chang et al., 2025)	2.915	1.386	2.343	3.847	3.578	1.958	3.035	4.710
		AID (Kim et al., 2024)	2.835	1.044	1.984	3.313	3.670	1.700	2.825	4.508
		VLM-CC (Li et al., 2026)	3.508	1.667	2.994	4.557	4.339	1.974	3.755	5.897
		LSMI-U (Kim et al., 2021)	2.571	1.480	2.241	3.331	2.589	1.373	**2.210**	3.343
		Ours (LSMI)	**2.061**	1.092	1.759	**2.675**	**2.542**	1.371	2.225	**3.319**
	Nikon	C5 (Afifi et al., 2021)	2.748	1.302	2.335	3.674	3.281	1.512	2.862	4.437
		CCMNet (Kim et al., 2025)	2.326	0.852	1.737	3.392	2.806	1.066	2.194	4.200
		FC4 (Hu et al., 2017)	2.233	0.947	1.742	3.147	2.657	1.203	2.231	3.695
		GCC (Chang et al., 2025)	3.143	1.756	2.859	4.073	3.718	2.075	3.367	5.093
		AID (Kim et al., 2024)	2.536	**0.702**	1.820	3.214	3.068	**0.910**	2.446	4.075
		VLM-CC (Li et al., 2026)	2.802	1.024	2.342	3.705	3.299	1.211	2.942	4.666
		LSMI-U (Kim et al., 2021)	2.237	1.123	1.920	2.890	2.545	1.254	2.188	3.411
		Ours (LSMI)	**1.954**	0.853	**1.658**	**2.552**	**2.368**	1.160	**2.032**	**3.183**
	Sony	C5 (Afifi et al., 2021)	2.426	0.837	1.948	3.419	3.191	1.266	2.711	4.590
		CCMNet (Kim et al., 2025)	2.286	**0.703**	1.808	3.335	2.972	**1.005**	2.576	4.454
		FC4 (Hu et al., 2017)	2.341	1.007	1.923	3.114	3.047	1.353	2.683	4.232
		GCC (Chang et al., 2025)	3.214	1.891	2.855	4.185	3.927	2.210	3.646	5.247
		AID (Kim et al., 2024)	2.331	0.903	1.882	3.128	3.095	1.154	2.697	4.200
		VLM-CC (Li et al., 2026)	2.922	1.469	2.380	3.804	3.613	1.844	3.184	4.817
		LSMI-U (Kim et al., 2021)	2.146	1.348	1.999	**2.606**	**2.425**	1.477	**2.212**	**3.066**
		Ours (LSMI)	**1.986**	0.830	**1.780**	2.799	2.525	1.062	2.282	3.578
	All	C5 (Afifi et al., 2021)	2.171	**0.658**	1.682	3.089	2.849	1.015	2.359	4.187
		CCMNet (Kim et al., 2025)	2.305	0.768	1.773	3.292	2.991	1.149	2.560	4.290
		FC4 (Hu et al., 2017)	2.202	0.808	1.723	2.946	2.869	1.204	2.393	3.929
		GCC (Chang et al., 2025)	3.037	1.616	2.570	4.009	3.647	1.891	3.154	5.025
		AID (Kim et al., 2024)	2.693	0.955	1.911	3.258	3.425	1.397	2.624	4.307
		VLM-CC (Li et al., 2026)	3.385	1.831	2.840	4.210	3.846	2.048	3.286	5.018
		LSMI-U (Kim et al., 2021)	2.340	1.410	2.039	3.021	2.524	1.428	2.171	3.285
		Ours (LSMI)	**1.806**	0.785	**1.494**	**2.444**	**2.252**	**1.010**	**1.883**	**3.030**

Table 2: Comparison of VLM backbones under identical training settings. Task1 performs active source identification, while Task2 adds source counting and photometric region grounding. Img Acc./Src Rec. denote image-level accuracy and source-level recall.

VLM Backbone	**Task1**				**Task2**			
	Status	**F1**	**Img Acc.**	**Src Acc.**	**Region Acc.**	**GT Rec.**	**Img Acc.**	**Src Acc.**
Qwen2.5-VL-3B	93.40	87.80	24.66	68.28	51.41	44.21	48.21	84.98
Qwen2.5-VL-7B	92.85	85.33	25.11	68.54	53.70	44.34	41.93	81.36
InternVL3-2B	93.22	85.36	**26.46**	68.80	58.30	50.23	54.48	86.15
InternVL3-8B (Ours)	**94.50**	**89.01**	24.89	**69.26**	**63.38**	**58.71**	**69.73**	**92.69**

Table 3: White balance performance on the **MIIW** (Murmann et al., 2019) dataset.

Method	L (°)↓	R (°)↓
C5 (Afifi et al., 2021)	2.83	2.95
CCMNet (Kim et al., 2025)	2.77	2.92
FC4 (Hu et al., 2017)	2.82	3.03
GCC (Chang et al., 2025)	3.85	4.27
AID (Kim et al., 2024)	2.64	2.85
VLM-CC (Li et al., 2026)	2.73	2.89
LSMI-U (Kim et al., 2021)	2.93	2.99
Ours(MIIW)	**2.63**	**2.82**

Table 4: **Iteration ablation.** Effects of training and inference rounds.

Setting	Iters.	L (°)↓	R (°)↓
Inference	1	4.14	4.09
Inference	2	3.97	3.92
Training	1	4.36	4.36
Training	2	3.98	3.94
Training	3	3.94	3.89

Table 5: **Initialization ablation.** Gray-Pixel (GP) clustering and region priors; N: VLM-predicted source count.

Initialization	L (°)↓	R (°)↓
N + GP clustering	4.07	4.00
N + region priors	4.01	3.94
Ours (full)	**3.94**	**3.89**

The corrected image R follows the same trend, except on Sony, where LSMI-U obtains a slightly lower mean R error ($2.425°$ vs. $2.525°$) while AskLight achieves lower P25 error. On our mixed-illumination MIIW (Table 3), AskLight obtains the lowest errors of $2.63°$ and $2.82°$ for L and R, although the margin over AID is small.

Qualitative comparison. Fig. 4 shows that single-illuminant baselines produce uniform illumination maps and leave residual casts in regions lit by a different source, whereas AskLight recovers the boundaries between differently lit regions and yields corrected images closer to the ground truth. More qualitative results on LSMI, Infinigen-Light, and MIIW are provided in Figs. 7 and 8 in the supplementary material.

5.3 Ablation Study

VLM backbone and task design. Table 2 compares VLM backbones under identical training.Image accuracy measures the ratio of images with completely recovered active-light inventories, source accuracy measures the ratio of correctly recovered sources, we evaluate them both on task1 and 2. We evaluate Task 1 with status accuracy (correct source states among all predicted states) and F1 (correctly classified source states among all sources). For Task 2, GT recall measures the correctly recovered and region-grounded GT sources, and region accuracy measures the correctly assigned regions.

All backbones achieve similar Task 1 status accuracy (92.9–94.5%), indicating that visible source classification mainly relies on general visual-semantic knowledge. Task 2 further requires recovering missed and hidden sources with region grounding. InternVL3-8B achieves the best results, with 69.73% image accuracy compared with at most 54.48% for other backbones. Its Task 2 improves image accuracy from 24.89% to 69.73% and Source accuracy from 69.26% to 92.69%, demonstrating the effectiveness of our two-task VLM beyond visible proposal classification.

Refinement rounds. Table 4 varies the number of refinement rounds with the VLM frozen. Training with more rounds steadily reduces the final illumination error ($4.36° \rightarrow 3.98° \rightarrow 3.94°$). Moreover, the first-round output of the three-round model ($4.14°$) is already better than a model trained with a single round ($4.36°$), showing that multi-round supervision also improves intermediate states rather than merely adding inference steps.

Candidate initialization. Table 5 fixes the VLM-predicted source count and varies the initialization. Global Gray-Pixel clustering without region guidance yields $4.07°$, and region priors alone yield $4.01°$. Combining source-grounded colors with region-guided priors performs best ($3.94°$), showing that source-to-region association provides a better starting point.

6 Conclusion

We presented AskLight for multi-illuminant white balance. Instead of treating illuminants as anonymous components, it asks a VLM which sources are active and where each one casts light, and uses these answers to initialize illumination decomposition. To train this reasoning, we built Infinigen-Light, which labels each source with its activity, visibility, and spatial influence. AskLight gives more accurate illumination estimates on both synthetic and real data. Future work will add source-level labels on real data, allow association with multiple regions, and speed up the reasoning stage.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript preparation, including writing refinement, language editing, translation, literature search, reference verification, and code debugging. They were not used to generate experimental data or replace the authors' decisions in developing the proposed framework, conducting experiments, and interpreting results. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the manuscript.

REFERENCES

Mahmoud Afifi, Jonathan T. Barron, Chloe LeGendre, Yun-Ta Tsai, and Francois Bleibel. Cross-camera convolutional color constancy. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1961–1970, 2021. doi: 10.1109/ICCV48922.2021.00199.

Jonathan T. Barron and Yun-Ta Tsai. Fast fourier color constancy. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6950–6958, 2017. doi: 10.1109/CVPR.2017.735.

Shida Beigpour, Christian Riess, Joost van de Weijer, and Elli Angelopoulou. Multi-illuminant estimation with conditional random fields. *IEEE Transactions on Image Processing*, 23(1):83–96, 2014. doi: 10.1109/TIP.2013.2286327.

G. Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin Institute*, 310(1):1–26, 1980. ISSN 0016-0032. doi: https://doi.org/10.1016/0016-0032(80)90058-7. URL https://www.sciencedirect.com/science/article/pii/0016003280900587.

Chen-Wei Chang, Cheng-De Fan, Chia-Che Chang, Yi-Chen Lo, Yu-Chee Tseng, Jiun-Long Huang, and Yu-Lun Liu. Gcc: Generative color constancy via diffusing a color checker. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10868–10878, 2025. doi: 10.1109/CVPR52734.2025.01015.

Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. ISSN 00129658, 19399170. URL http://www.jstor.org/stable/1932409.

Niladri Shekhar Dutt, Duygu Ceylan, and Niloy J. Mitra. Monetgpt: Solving puzzles enhances mllms' image retouching skills. *ACM Trans. Graph.*, 44(4), July 2025. ISSN 0730-0301. doi: 10.1145/3730926. URL https://doi.org/10.1145/3730926.

G.D. Finlayson, B.V. Funt, and K. Barnard. Color constancy under varying illumination. In *Proceedings of IEEE International Conference on Computer Vision*, pp. 720–725, 1995. doi: 10.1109/ICCV.1995.466867.

Arjan Gijsenij, Theo Gevers, and Joost van de Weijer. Computational color constancy: Survey and experiments. *IEEE Transactions on Image Processing*, 20(9):2475–2489, 2011. doi: 10.1109/TIP.2011.2118224.

Lin Gu, Cong Phuoc Huynh, and Antonio Robles-Kelly. Segmentation and estimation of spatially varying illumination. *IEEE Transactions on Image Processing*, 23(8):3478–3489, 2014. doi: 10.1109/TIP.2014.2330768.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Yuanming Hu, Baoyuan Wang, and Stephen Lin. Fc^4: Fully convolutional color constancy with confidence-weighted pooling. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 330–339, 2017. doi: 10.1109/CVPR.2017.43.

Dongyoung Kim, Jinwoo Kim, Seonghyeon Nam, Dongwoo Lee, Yeonkyung Lee, Nahyup Kang, Hyong-Euk Lee, ByungIn Yoo, Jae-Joon Han, and Seon Joo Kim. Large scale multi-illuminant (lsmi) dataset for developing white balance algorithm under mixed illumination. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2390–2399, 2021. doi: 10.1109/ICCV48922.2021.00241.

Dongyoung Kim, Jinwoo Kim, Junsang Yu, and Seon Joo Kim. Attentive illumination decomposition model for multi-illuminant white balancing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25512–25521, 2024. doi: 10.1109/CVPR52733.2024.02410.

Dongyoung Kim, Mahmoud Afifi, Dongyun Kim, Michael S. Brown, and Seon Joo Kim. Ccmnet: Leveraging calibrated color correction matrices for cross-camera color constancy. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6198–6208, 2025. doi: 10.1109/ICCV51701.2025.00585.

Yuan-Kang Lee, Kuan-Lin Chen, Chia-Che Chang, and Yu-Lun Liu. Rl-awb: Deep reinforcement learning for auto white balance correction in low-light night-time scenes, 2026. URL https://arxiv.org/abs/2601.05249.

Shuwei Li and Robby T. Tan. Nighttcc: Nighttime color constancy via adaptive channel masking. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25522–25531, 2024. doi: 10.1109/CVPR52733.2024.02411.

Shuwei Li, Lei Tan, and Robby T. Tan. White-balance first, adjust later: Cross-camera color constancy via vision-language evaluation, 2026. URL https://arxiv.org/abs/2605.19613.

Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, Sai Bi, Hong-Xing Yu, Zexiang Xu, Kalyan Sunkavalli, Milos Hasan, Ravi Ramamoorthi, and Manmohan Chandraker. Openrooms: An open framework for photorealistic indoor scene datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7190–7199, June 2021.

Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-4803. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d34497330b1fd6530f7afd86d0df9f76-Paper-Datasets_and_Benchmarks_Track.pdf.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017. doi: 10.1109/CVPR.2017.106.

Yunlong Lin, Zixu Lin, Kunjie Lin, Jinbin Bai, Panwang Pan, Chenxin Li, Haoyu Chen, Zhongdao Wang, Xinghao Ding, Wenbo Li, and Shuicheng Yan. Jarvisart: Liberating human artistic creativity via an intelligent photo retouching agent. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 52088–52130. Curran Associates, Inc., 2025. doi: 10.52202/085713-1736. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/4ac4365b98bc242acd5ab974a05c68a8-Paper-Conference.pdf.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. URL https://arxiv.org/abs/2303.05499.

Zhongyu Lou, Theo Gevers, Ninghang Hu, Marcel P Lucassen, et al. Color constancy by deep learning. In *BMVC*, volume 2, pp. 3, 2015.

Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A dataset of multi-illumination images in the wild. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4079–4088, 2019. doi: 10.1109/ICCV.2019.00418.

Yanlin Qian, Joni-Kristian Kamarainen, Jarno Nikkanen, and Jiri Matas. On finding gray pixels, June 2019.

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12630–12641, 2023. doi: 10.1109/CVPR52729.2023.01215.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21783–21794, 2024. doi: 10.1109/CVPR52733.2024.02058.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10892–10902, 2021. doi: 10.1109/ICCV48922.2021.01073.

David Serrano-Lozano, Aditya Arora, Luis Herranz, Konstantinos G. Derpanis, Michael S. Brown, and Javier Vazquez-Corral. Revisiting image fusion for multi-illuminant white-balance correction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8275–8284, 2025. doi: 10.1109/ICCV51701.2025.00775.

Haoran Sun, Bingyang Wang, Suyang Yu, Yijiang Li, Qingying Gao, Haiyun Lyu, Lianyu Huang, Zelong Hong, Jiahui Ge, Qianli Ma, Hang He, Yifan Zhou, Lingzi Guo, Lantao Mei, Maijunxian Wang, Dezhi Luo, and Hokin Deng. Probing perceptual constancy in large vision-language models, 2026. URL https://arxiv.org/abs/2502.10273.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 402–419, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58536-5.

Joost van de Weijer, Theo Gevers, and Arjan Gijsenij. Edge-based color constancy. *IEEE Transactions on Image Processing*, 16(9):2207–2214, 2007. doi: 10.1109/TIP.2007.901808.

Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Rep vit: Revisiting mobile cnn from vit perspective. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15909–15920, 2024. doi: 10.1109/CVPR52733.2024.01506.

Zhihua Wang, Yu Long, Qiuping Jiang, Chao Huang, and Xiaochun Cao. Harnessing multi-modal large language models for measuring and interpreting color differences. *IEEE Transactions on Image Processing*, 35:2292–2304, 2026. doi: 10.1109/TIP.2024.3522802.

Wenjun Wei, Yanlin Qian, Huaian Chen, Junkang Dai, and Yi Jin. Integral fast fourier color constancy. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26420–26429, 2025. doi: 10.1109/CVPR52734.2025.02460.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 12547–12573, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/363d4c97bb411e4b07612915b76c06ae-Paper-Conference.pdf.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25490–25500, 2024b. doi: 10.1109/CVPR52733.2024.02408.

Kai-Fu Yang, Shao-Bing Gao, and Yong-Jie Li. Efficient illuminant estimation for color constancy using grey pixels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Mingde Yao, Zhiyuan You, King-Man Tam, Menglu Wang, and Tianfan Xue. PhotoAgent: Exploratory Visual Aesthetic Planning with Large Vision Models. *arXiv e-prints*, art. arXiv:2602.22809, February 2026. doi: 10.48550/arXiv.2602.22809.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

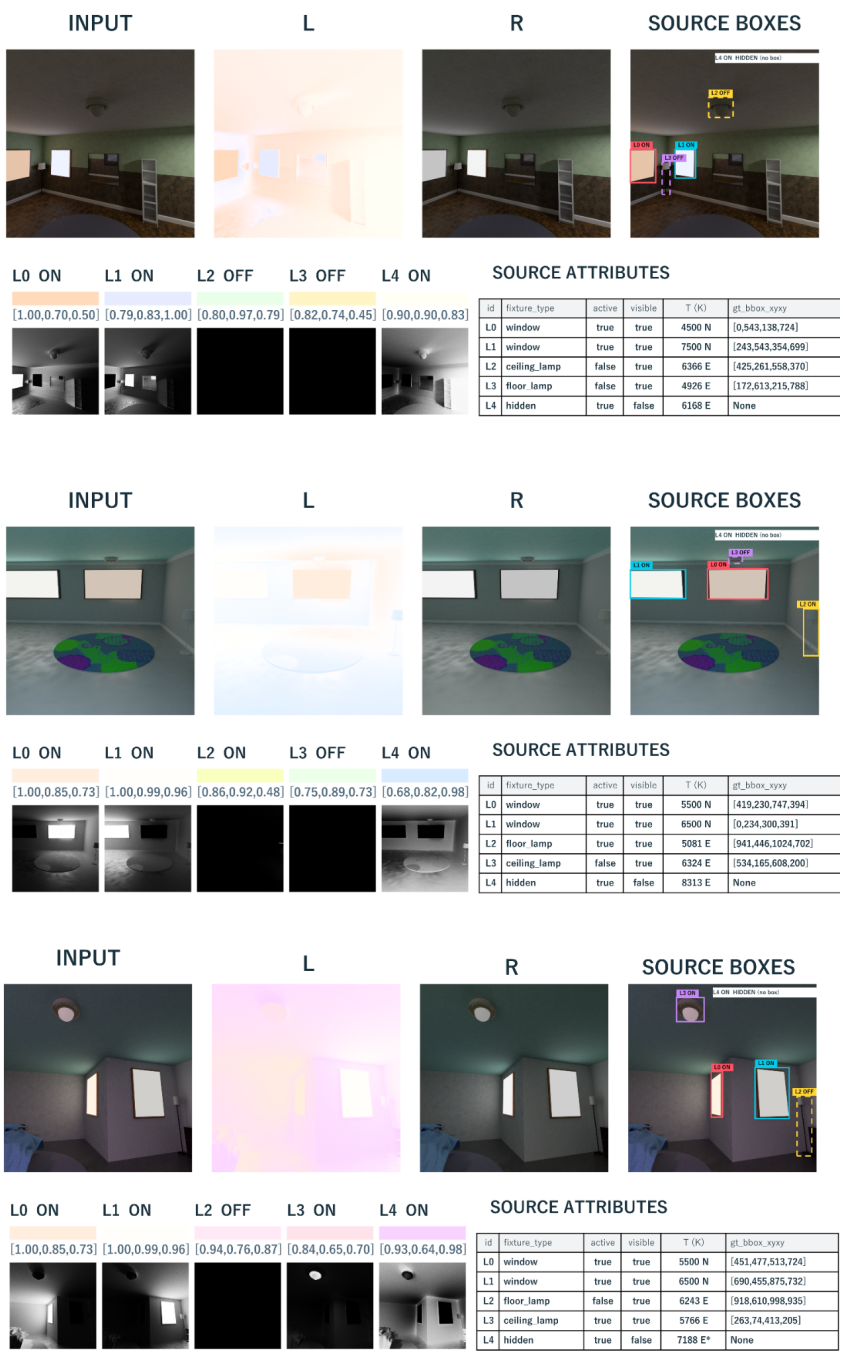

Figure 5: **Infinigen-Light dataset.** Infinigen-Light provides physically grounded multi-illuminant scenes with source-level supervision, including illumination maps, white-balanced images, and source attributes such as chromaticity, activity, and visibility. It explicitly models the gap between visible fixtures and active illuminants, where hidden sources can still contribute to scene illumination.

A Additional Details of Infinigen-Light Construction

As illustrated in Fig. 5, Infinigen-Light provides source-level supervision beyond conventional illumination maps, enabling models to reason about the physical causes of spatially varying illumination. We next introduce the dataset construction and annotation details in depth.

A.1 Scene and Source Configurations

We generate bedrooms, living rooms, and dining rooms using Infinigen Indoors (Raistrick et al., 2024), then apply our controlled source registration, color assignment, and rendering protocol. Each accepted camera view contains 3–5 registered sources, including windows, ceiling lamps, floor lamps, and, in some scenes, one hidden source that illuminates the observed scene without being visible from the camera.

Each registered source groups a Blender light object and its associated emissive surface, such as a bulb surface or window panel, under one identity. The two components share their color and activation state. Switching a source off sets its light energy and surface emission strength to zero while preserving its geometry and reflective material. An inactive source can therefore remain visible under illumination from other sources. Scene geometry, materials, camera parameters, and source positions remain fixed across lighting states; each source uses the same color and power whenever it is active. In retained lighting states, windows and hidden sources remain active, while other sources can be switched on or off. We select five states per scene without a hidden source and four with one.

A.2 Source Colors and Annotations

Let q_m^{raw} denote the scene-linear RGB emission color of source m. For non-window sources, including hidden sources, we independently sample

$$H \sim \mathcal{U}(0,1), \qquad S \sim \mathcal{U}(0.02, 0.38), \qquad V \sim \mathcal{U}(0.90, 1.00). \tag{8}$$

We convert HSV to display-encoded sRGB and then decode it to scene-linear RGB. Window colors are obtained from nominal 4500, 5500, 6500, or 7500K̇These templates provide controlled color variation without spectral daylight simulation. The renderer uses q_m^{raw}.For illumination-target construction, we divide each source's linear-RGB color by its green-channel value:

$$q_m = \left[\frac{q_m^{\mathrm{raw},R}}{q_m^{\mathrm{raw},G}}, 1, \frac{q_m^{\mathrm{raw},B}}{q_m^{\mathrm{raw},G}}\right]^\top. \tag{9}$$

Each source has a persistent identity, type, position, emission parameters, linear-RGB color q_m^{raw}, activity state, visibility annotation, and visible-source bounding box. Activity specifies whether the source emits light in the current state. Visibility records whether it is semantically visible from the selected camera. Bounding boxes are computed separately from object-index renders using the saved camera and geometry. Each box encloses the pixels where the complete registered lamp or window geometry is directly visible from the camera, including its emissive surface; sources with no such pixels receive a null box. The box localizes visible source geometry, whereas α describes its spatial illumination contribution.

A.3 Rendering and Target Construction

We use Blender 4.2 and Cycles with unregistered lights, unregistered material emission, and environment emission disabled. For each registered source m, we render a scene-linear RGB response C_m with only that source active, using its assigned color q_m^{raw} and power. These source-isolated responses share the same scene geometry, materials, and camera, and are reused across lighting states. For each retained state, the observation I is rendered directly with the selected sources emitting together.

Following the green-channel mixture estimation in LSMI (Kim et al., 2021), for each lighting state, we construct the contribution maps α from the source-isolated responses of its active sources. At each pixel, an active source's weight is its green-channel response divided by the summed green-channel responses of all active sources. Inactive sources receive zero weight.

We construct the illumination target as

$$L = \sum_m \alpha_m q_m, \tag{10}$$

where $\alpha \cdot q_m$ is the source-specific illumination component shown in the dataset figure.

The white-balanced reference is obtained by channel-wise division:

$$R = \frac{I}{L}. \tag{11}$$

Dataset statistics and rendering details. Infinigen-Light contains 672 indoor scenes with 2,992 rendered samples, including 471 training scenes (2,096 images), 101 validation scenes (450 images), and 100 test scenes (446 images). All images are rendered using Blender 4.2 with the Cycles CUDA renderer at 1024×1024 resolution and 128 SPP without denoising or adaptive sampling. The rendering is performed on four NVIDIA Tesla V100-SXM2 GPUs (16GB memory per GPU), with each task assigned to a single GPU and 16 CPU threads. The average rendering time is 13.3 seconds per input image, and approximately 50.7 seconds per sample including reference rendering, ground-truth generation, and validation.

B Additional Details of AskLight

We use m for annotated physical sources, j for Grounding DINO proposals, P_j for Gray-Pixel color candidates, k for MIRF illumination regions, and i for predicted illuminant candidates.

B.1 Proposal and Illumination-Region Construction

Visible-source proposals. We use a frozen Grounding DINO (Liu et al., 2024) with the Swin-T OGC checkpoint to locate potential visible sources in the RGB preview.

These proposals provide candidate locations for Task 1 to determine source validity and activity. We retain proposals with scores above 0.35, apply non-maximum suppression with an IoU threshold of 0.5, and assign identifiers $D_1, D_2, \ldots$ for VLM reference.

Gray-Pixel selection. We select Gray Pixels (GP) (Qian et al., 2019) from the 512×512 linear RGB image independently of detection. The selected GP pixels are clustered by their two-dimensional positions using K-means. We obtain five groups, each providing a color seed P_j together with its spatial center. The seed color is computed by summing the linear RGB values of the group and applying L_2 normalization for angular comparisons.

Coarse illumination regions. We construct coarse illumination regions using a procedure inspired by the Multi-Illuminant Random Field (MIRF) (Beigpour et al., 2014).

The five GP seeds provide initial color cues and region seeds, which are combined with spatial regularization to generate coarse illumination regions:

$$\{M_k\}_{k=1}^{5}.$$

Specifically, we optimize illumination labels on a patch grid by jointly considering color agreement, spatial proximity, and pairwise smoothness. The resulting labels are further refined using SLIC superpixels on linear RGB. Regions with identical labels are merged and indexed as $M_1, \ldots, M_K$, whose boundaries and identifiers form the MIRF region view.

B.2 Recurrent Refinement Details

Shared image features. RepViT-M1.1 (Wang et al., 2024) and a lightweight FPN (Lin et al., 2017) extract a shared feature representation F once per image. The multi-scale features are fused through the FPN to provide spatially detailed representations for candidate refinement. We fine-tune the backbone while keeping BatchNorm running statistics fixed. The input image remains in linear RGB space for reconstruction.

Candidate refiner. For each predicted illuminant candidate i, we construct a Gaussian center map from its spatial center μ_i. At each refinement round, the candidate refiner receives the shared feature map, spatial center prior, initial weight prior α_i^0, previous weight estimate α_i^{t-1}, previous illumination estimate L^{t-1}, and previous candidate color c_i^{t-1}.

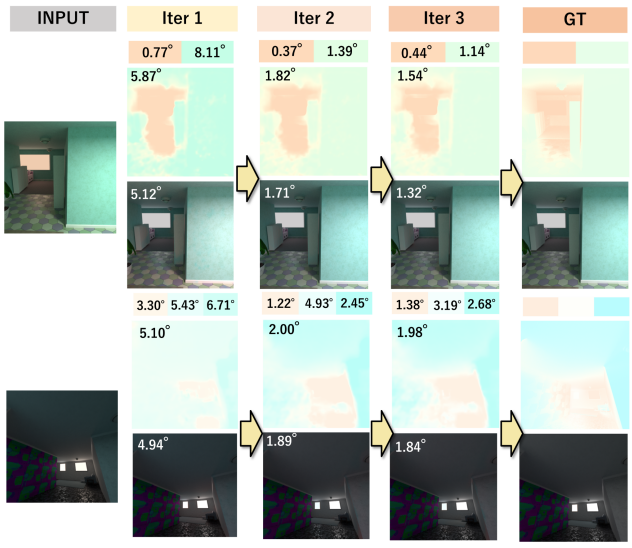

Figure 6: **Iterative refinement of the illumination estimation.** The downstream weight network progressively improves the estimated illumination through recurrent refinement. The numbers indicate the mean angular error (°) at each iteration.

The refiner predicts candidate-specific updates through the Weight Head and Color Head. All parameters are shared across candidates and refinement rounds.

Weight and color updates. The Weight Head predicts a spatial score for each candidate. After normalization across candidates, the updated mixture weights are computed as:

$$\alpha_i^t(p) = \mathrm{softmax}_i \left[u_i^t(p) + 0.1 \log \left(\frac{\alpha_i^0(p) + \alpha_i^{t-1}(p)}{2} + \epsilon \right) \right]. \tag{12}$$

The softmax is applied across candidates at each pixel. Candidates padded for variable cardinality are masked and assigned zero weight.

The Color Head predicts bounded residual updates for candidate chromaticities:

$$c_{i,k}^t = c_{i,k}^{t-1} \exp\left(0.4 \tanh \delta_{i,k}^t\right), \quad k \in \{R, B\}, \qquad c_{i,G}^t = 1. \tag{13}$$

This update is performed in log-color space and preserves positive red and blue components.

Reconstruction and supervision. After each refinement round, we reconstruct L^t and R^t using Eq. equation 3. The recurrent refiner and prediction heads share parameters across all candidates and three refinement rounds, and gradients are propagated through the recurrent states.

Before training this network, we freeze the VLM and use the candidate priors generated from its predictions. The losses in equation 6–equation 7 supervise only the composed illumination field and white-balanced output, without requiring explicit matching between predicted candidates and annotated sources. Losses are averaged over valid pixels within each image and then across images. Angular errors are computed in radians.

Figure 6 illustrates the whole iterative refinement process. With the VLM prediction fixed, the downstream weight network progressively reduces the illumination estimation error through recurrent optimization, demonstrating the effectiveness of the proposed refinement strategy.

C Additional Experimental Details

C.1 Dataset Splits and MIIW Reconstruction

For the camera-specific LSMI experiments, we use the official camera-wise splits. The training/validation/test sample counts are $1,740/498/250$ for Galaxy, $1,386/390/202$ for Nikon, and $2,094/598/304$ for Sony. Each backend is trained and evaluated using its corresponding camera split.

For MIIW evaluation, we construct a mixed-illumination reconstruction protocol based on the MIIW dataset (Murmann et al., 2019). Following previous multi-illuminant white balance studies (Kim et al., 2021; 2024), we combine multiple illumination conditions within each scene to generate single-, two-, and three-illuminant variants.

Specifically, each MIIW scene provides multiple illumination settings. We select different combinations of these settings to synthesize mixed- illumination observations while preserving the original scene content. Each scene generates two single-illuminant variants, one two-illuminant variant, and one three-illuminant variant. All variants from the same scene are assigned to the same split to avoid scene-level leakage.

Our MIIW reconstruction contains 801/90/219 scenes for training, validation, and test, resulting in 3,204/360/876 samples. The reconstructed samples provide corresponding illumination targets and white-balanced references for evaluating multi-illuminant white balance. Gray and chrome sphere regions contained in MIIW scenes are masked during training and evaluation to prevent models from exploiting calibration objects as image cues.

C.2 Adaptation on LSMI and MIIW

After supervised training on Infinigen-Light, we further adapt the VLM to LSMI and MIIW. Unlike Infinigen-Light, these real datasets do not provide complete physical source-level annotations, including source identity, visibility, and ACTIVE/INACTIVE states. However, both datasets provide illumination-related supervision, which can be used to construct Task 2 targets.

For both datasets, InternVL3-8B is initialized from the Blender-trained checkpoint and adapted for three epochs before being frozen for downstream white balance training. The VLM adaptation follows the same two-stage reasoning framework: Task 1 predicts the status of visible proposals, and Task 2 generates source hypotheses with illumination-region associations.

LSMI adaptation. LSMI provides dense contribution maps without physical source identities. We represent each annotated contribution as an anonymous source for Task 2 supervision and associate it with a MIRF region. Here, an anonymous source is a supervision label and does not identify a physical lamp.

Specifically, we derive Task 2 target JSONs from the ground-truth contribution maps. Each illumination contributor is matched to its dominant MIRF region, and the resulting region associations supervise the generated source inventory:

$$Y_2^{\mathrm{GT}} = \{(S_i, M_i)\}_{i=1}^{N},$$

where S_i denotes an illumination-source hypothesis and M_i denotes its corresponding illumination region. These hypotheses do not represent known physical lamp identities.

Since LSMI does not provide reliable ACTIVE/INACTIVE/INVALID labels for detected proposals, Task 1 is not directly supervised. During adaptation, Task 1 predictions are generated by the current VLM and used only as the context input for Task 2. The training loss is therefore:

$$\mathcal{L}_{\mathrm{LSMI}} = \mathcal{L}_2^{\mathrm{CE}}, \tag{14}$$

where the cross-entropy loss supervises only the Task 2 JSON tokens.

MIIW adaptation. MIIW also lacks explicit physical source annotations. In MIIW, illumination is provided by the reflected light captured by light spheres rather than directly visible light sources. To prevent the VLM from exploiting sphere appearance cues, all light spheres are masked

and inpainted before being fed into the VLM. Therefore, each illumination contributor is treated as a hidden-source hypothesis, and its dominant illumination region is obtained from the ground-truth contribution weights.

The Task 2 target is therefore represented as:

$$Y_2^{\mathrm{GT}} = \{(H_i, M_i)\}_{i=1}^{N},$$

where H_i denotes a hidden-source hypothesis and M_i is its associated illumination region. The model does not receive the identity of the physical light producing this contribution.

For Task 1, we follow the all-hidden formulation. Since all illumination contributors are represented as hidden sources, the GT status labels for detected proposals are set to INACTIVE:

$$Y_1^{\mathrm{GT}} = \{D_j : \mathrm{INACTIVE}\}_{j=1}^{K}.$$

The MIIW adaptation jointly optimizes Task 1 and Task 2 answer-token cross-entropy:

$$\mathcal{L}_{\mathrm{MIIW}} = \lambda_1 \mathcal{L}_1^{\mathrm{CE}} + \lambda_2 \mathcal{L}_2^{\mathrm{CE}}. \tag{15}$$

C.3 Calibration-target handling

To prevent models from exploiting calibration targets as shortcuts, we remove calibration regions from the inputs of both VLM and downstream backends. For LSMI, color-chart regions are removed using their provided masks. For MIIW, annotated gray-sphere regions are masked out before model processing. The removed regions are inpainted for VLM inputs, while their pixels remain excluded from downstream losses and evaluation.

For LSMI, the masked chart regions are also excluded from GP selection and backend feature aggregation. For MIIW, the gray-sphere regions are handled in the same manner. Calibration information is only used when constructing the supervision targets, and the reference objects themselves are never provided as model inputs.

C.4 Downstream Backend Optimization

All downstream illumination decomposition backends use the same training configuration. Each backend is initialized from an ImageNet-pretrained RepViT-M1.1, while the FPN and prediction modules are randomly initialized. AdamW is used with a learning rate of 10^{-5} for the backbone and 10^{-4} for the remaining modules, with weight decay 0.01 and gradient-norm clipping at 1. All backends are trained for 100 epochs with batch size 4, using a 5% linear warm-up followed by cosine decay to 1% of the initial learning rates.

We use FP16 mixed precision training without random geometric or color augmentation. All results are obtained with a single random seed. During backend optimization, VLM predictions are cached and fixed, and the predicted candidate count is retained without replacing it with ground-truth counts. The three-round recurrent loss is defined in the main paper, and no additional supervision is applied to individual candidate colors or mixture weights.

C.5 Baseline Implementation Details

To ensure a consistent evaluation protocol, we do not use calibration targets (e.g., color charts or reference spheres) as training inputs or supervision signals for baseline comparison.

For GCC Chang et al. (2025) evaluation, the original training protocol relies on color-chart information. Since calibration targets are excluded in our setting, we use the publicly available pretrained GCC models without additional fine-tuning. Among the provided pretrained models, we select the model with the lower illumination-field (L) angular error on the validation set and report its performance as the GCC baseline.

For AID (Kim et al., 2024), the official implementation is not publicly available. Therefore, we reproduce the method following the experimental settings and training configurations described in the original paper. The reproduced model is used as the comparison baseline under our evaluation protocol.

Table 6: **Image-level active-source recovery accuracy.** Img Acc. is the percentage of test images for which all annotated active sources are recovered under our dataset-specific evaluation protocol. Ours uses task-adapted InternVL3-8B (Zhu et al., 2025). Higher is better; bold indicates the best result for each test set.

Test set	**AID** (Kim et al., 2024)	**Ours (ASKLIGHT)**
MIIW (Murmann et al., 2019)	55.59%	**67.12%**
LSMI (Kim et al., 2021)	**77.91%**	72.49%
Infinigen-Light	23.54%	**69.73%**

D ADDITIONAL RESULTS

Image-level active-source recovery. Table 6 compares our task-adapted VLM with our reproduction of AID (Kim et al., 2024), evaluating active-source recovery separately from final white balance accuracy. For LSMI and reconstructed MIIW, the evaluation targets refer to illumination contributors rather than identifiable physical sources, consistent with Appendix C.2. ASKLIGHT achieves 67.12% on MIIW and 69.73% on Infinigen-Light, exceeding AID by 11.53 and 46.19 percentage points, respectively. On LSMI (All), which combines the Galaxy, Nikon, and Sony subsets, AID achieves 77.91%, compared with 72.49% for our method.

D.1 MORE QUALITATIVE RESULTS

We provide additional qualitative comparisons on LSMI (Kim et al., 2021), Infinigen-Light, and MIIW (Murmann et al., 2019) in Figures 7 and 8. The compared methods often produce inaccurate illumination estimation under complex multi-illuminant conditions, resulting in noticeable residual color casts. In contrast, our method consistently recovers spatially varying illumination and generates white-balanced results closer to the ground truth. These examples further demonstrate the effectiveness of VLM-guided source reasoning and illumination refinement in challenging real and synthetic scenarios.

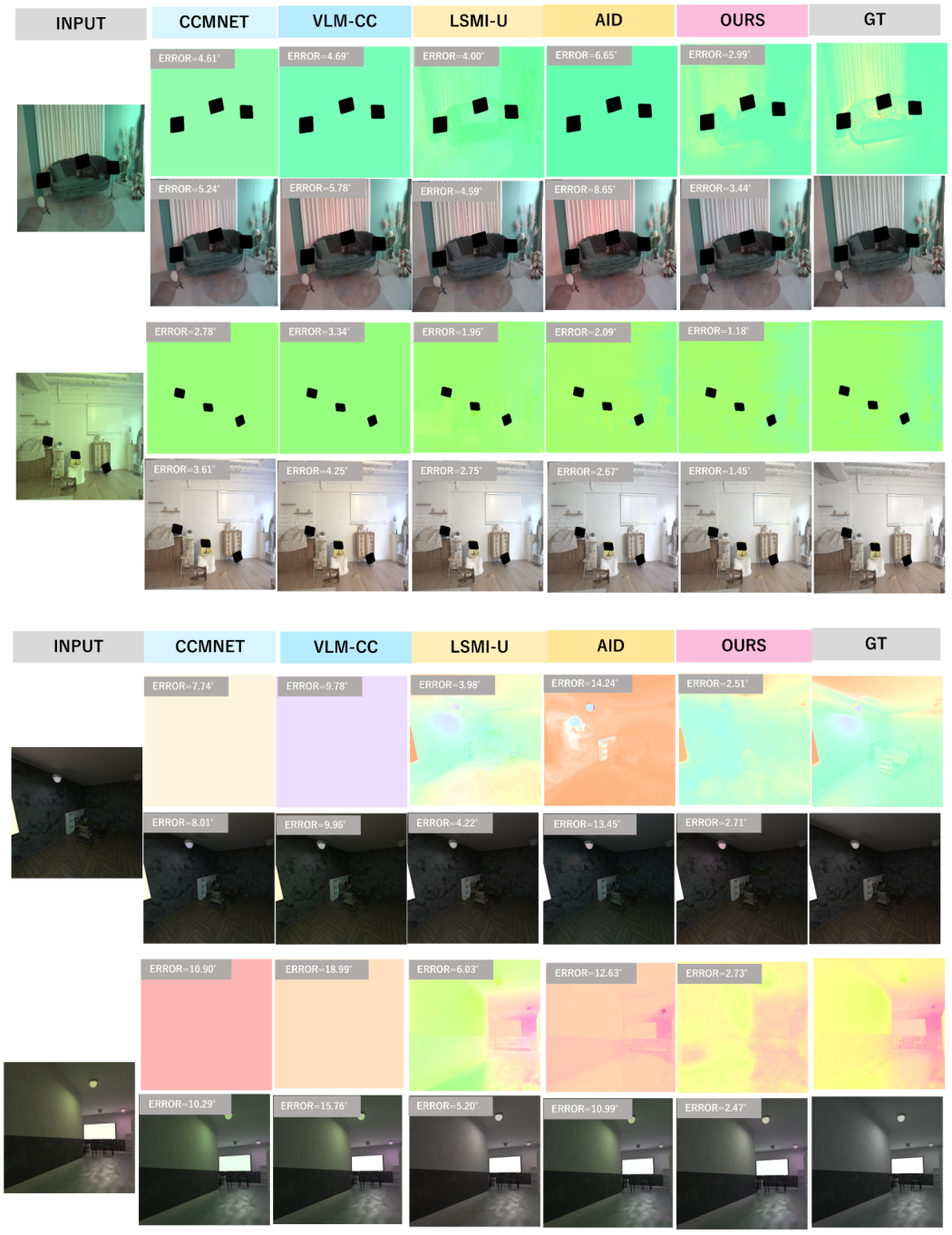

Figure 7: More visualization results on LSMI and Infinigen-Light. We visualize the estimated illumination maps and white-balanced results from different methods, showing that our approach better captures spatially varying illumination under complex multi-illuminant conditions. The first two groups correspond to LSMI, while the last two groups are from Infinigen-Light.

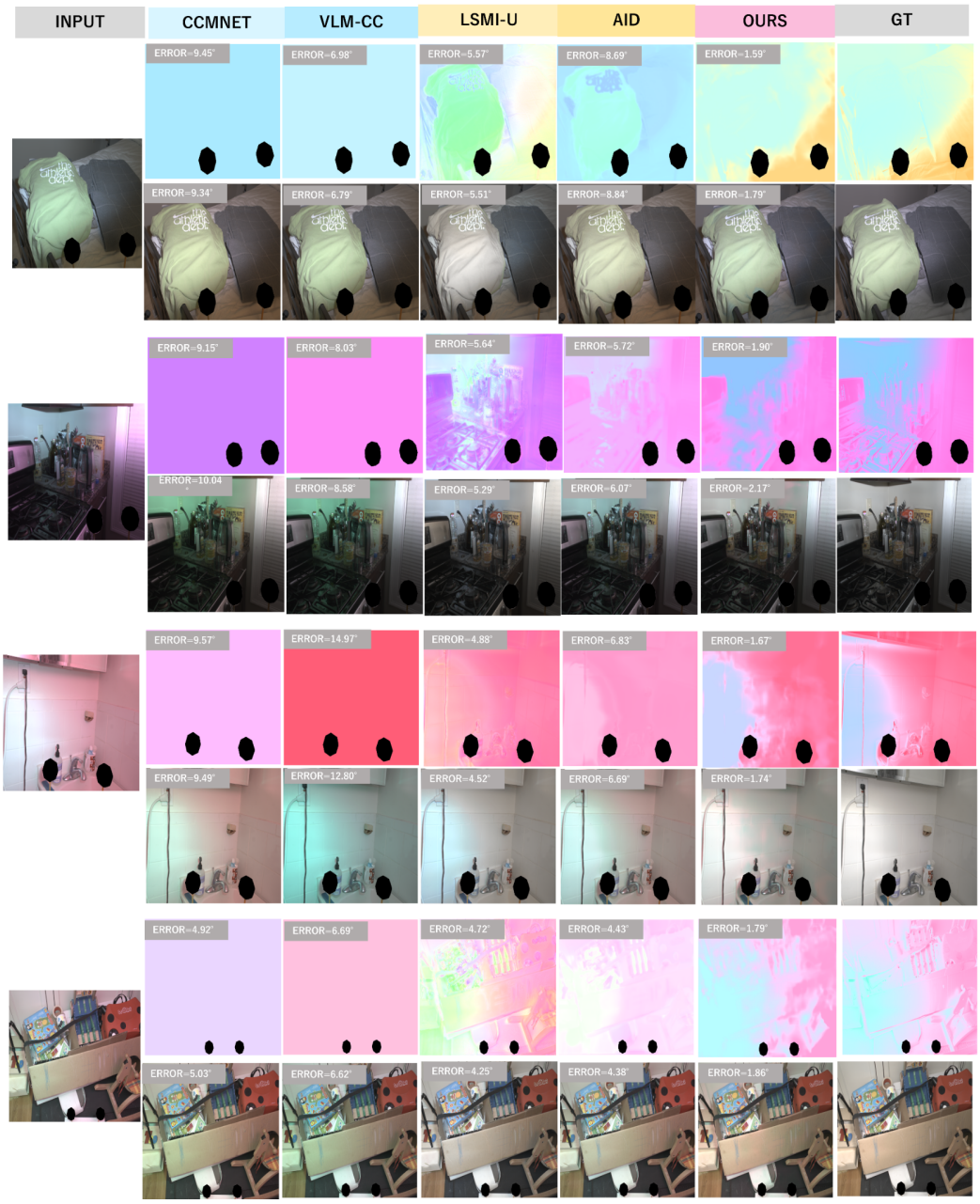

Figure 8: More Visualization Results on MIIW. We visualize the estimated illumination maps and white-balanced results from different methods, demonstrating that our approach produces more accurate spatially varying illumination estimation under complex multi-illuminant conditions.