CleAR: Robust Context-Guided Generative Lighting Estimation for Mobile Augmented Reality

YIQIN ZHAO, Worcester Polytechnic Institute, USA
MALLESHAM DASARI, Northeastern University, USA
TIAN GUO*, Worcester Polytechnic Institute, USA

High-quality environment lighting is essential for creating immersive mobile augmented reality (AR) experiences. However, achieving visually coherent estimation for mobile AR is challenging due to several key limitations in AR device sensing capabilities, including low camera FoV and limited pixel dynamic ranges. Recent advancements in generative AI, which can generate high-quality images from different types of prompts, including texts and images, present a potential solution for high-quality lighting estimation. Still, to effectively use generative image diffusion models, we must address two key limitations of *content quality* and *slow inference*. In this work, we design and implement a generative lighting estimation system called CleAR that can produce high-quality, diverse environment maps in the format of 360° HDR images. Specifically, we design a two-step generation pipeline guided by AR environment context data to ensure the output aligns with the physical environment's visual context and color appearance. To improve the estimation robustness under different lighting conditions, we design a real-time refinement component to adjust lighting estimation results on AR devices. To train and test our generative models, we curate a large-scale environment lighting estimation dataset with diverse lighting conditions. Through a combination of quantitative and qualitative evaluations, we show that CleAR outperforms state-of-the-art lighting estimation methods on both estimation accuracy, latency, and robustness, and is rated by 31 participants as producing better renderings for most virtual objects. For example, CleAR achieves 51% to 56% accuracy improvement on virtual object renderings across objects of three distinctive types of materials and reflective properties. CleAR produces lighting estimates of comparable or better quality in just 3.2 seconds—over 110X faster than state-of-the-art methods. Moreover, CleAR supports real-time refinement of lighting estimation results, ensuring robust and timely updates for AR applications.

CCS Concepts: • **Computing methodologies** → **Mixed / augmented reality**; • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: mobile augmented reality; lighting estimation; generative model

ACM Reference Format:
Yiqin Zhao, Mallesham Dasari, and Tian Guo. 2025. CleAR: Robust Context-Guided Generative Lighting Estimation for Mobile Augmented Reality. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 154 (September 2025), 26 pages. https://doi.org/10.1145/3749535

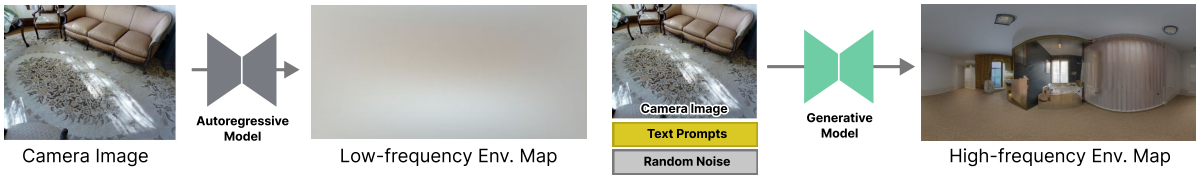
1 INTRODUCTION

As new augmented reality (AR) hardware and software enter consumer markets, mobile AR technologies have positively impacted various industries, including e-commerce, education, and engineering [12, 37]. The growing public adoption of AR technologies demands new standards for content quality and application user experiences,

*Corresponding author.

Authors' Contact Information: Yiqin Zhao, Worcester Polytechnic Institute, Worcester, MA, USA, yzhao11@wpi.edu; Mallesham Dasari, Northeastern University, Boston, MA, USA, m.dasari@northeastern.edu; Tian Guo, Worcester Polytechnic Institute, Worcester, MA, USA, tian@wpi.edu.

ACM 2474-9567/2025/9-ART154
https://doi.org/10.1145/3749535



(a) Autoregressive Lighting Estimation (b) Generative Lighting Estimation

Fig. 1. Comparison between autoregressive and generative lighting estimation methods. (1a) An autoregressive lighting estimation system, Xihe [58], estimates omnidirectional low-frequency lighting information from camera images with autoregressive models. The low-frequency lighting information misses important visual details, as visualized in the example environment map. (1b) Generative lighting estimation models can create high-frequency environment lighting estimation results from limited environment observations. The estimation process can be conditioned in several ways, such as partial environment observations and text prompts.

particularly emphasizing the need for *visual coherency* between virtual and physical content to ensure high-quality user experiences. To create visual coherency, AR applications require an accurate and robust environment *lighting estimation*, which ensures that virtual objects blend naturally with the physical environment.

To address this challenging task, traditional systems often employ autoregressive models [58] to extract parametric lighting information from AR device camera images (shown in Fig. 1a). While these methods effectively capture coarse environmental lighting conditions, they struggle to reliably capture high-frequency details, such as geometric features of the environment. To address the critical needs of high-frequency information in environment lighting, recent works [42, 48, 52] have sought new solutions by leveraging the advancements in *controllable generative models*. The primary reason is that these models have the potential to generate fine-grained environmental details (shown in Fig. 1a), thereby empowering AR systems with a more realistic rendering effect. However, it is challenging to achieve robust mobile AR lighting estimation with generative models in real-world environments. We identify two key challenges below.

First, robust lighting estimation demands accurate and timely estimations. Therefore, it is crucial to ensure that the system can provide reliable and accurate estimation outputs under diverse and complex lighting conditions. Additionally, generative models often exhibit long inference latencies, which slow down the estimation process. Directly integrating generative model-based lighting estimation methods may not satisfy the needs of real-time mobile AR experiences. Second, training and evaluating robust lighting estimation models require careful attention to data bias. Specifically, the datasets used must be curated to reflect diverse and balanced lighting conditions to ensure fair and generalizable performance. However, our measurement study on existing datasets reveals distribution biases in key lighting properties, such as intensity and color temperature, which in turn affect the generalization and robustness of several recent lighting estimation models [3, 48].

In this paper, we address the above issues with CleAR, a novel generative lighting estimation system for mobile AR. To support robust lighting estimation, CleAR builds on two key ideas: *(i)* AR context-guided generative estimation and *(ii)* edge-device collaborative estimation. It leverages *multimodal AR context* to guide 360° HDR environment map estimation using generative diffusion models. To ensure temporal consistency, CleAR uses a multi-output strategy to generate environment maps that best match real-time lighting conditions. To improve estimation quality and responsiveness, CleAR applies color appearance matching for fine-grained adjustments to environment maps.

At the core of CleAR, the lighting estimation process in CleAR revolves around a *two-step generative pipeline* that estimates 360° HDR environment maps from *partial LDR environment observations*. Our key design insight is to separate the generative model training objective into two domains: *LDR environment map completion* and *high-intensity pixel value estimation*. This novel learning objective design addresses the practical challenge of the

scarcity of high-quality lighting estimation training data, allowing us to leverage pre-trained large models to tackle each generation step effectively. Beyond its two-step generative pipeline, CleAR leverages AR context—*semantic maps* to guide visual details and *ambient light* sensor data to inform lighting intensity and color temperature. This context-guided generative estimation design effectively helps CleAR to tackle the challenge of estimating omnidirectional environment maps from limited environment observations on mobile AR devices.

To evaluate CleAR, we conduct a comprehensive evaluation that includes in-lab deployment tests, data-driven evaluations, and a user study. In the deployment test, CleAR excels in supporting virtual object rendering quality compared to three representative baselines: unwrapping a mirror ball (physical reference) [14], ARKit (commercial) [4], and LitAR (academic) [60]. Quantitatively, we compare CleAR with state-of-the-art lighting estimation models [3, 19, 35, 48] and show that CleAR outperforms the best performing baseline, DiffusionLight [35], by up to 56%. Our user study also confirms the effectiveness and robustness of CleAR, showing at least a 12% improvement in quality ratings compared to StyleLight, the second-best method. Furthermore, we introduce a robustness testing protocol, which leverages our augmented Laval dataset to test the estimation accuracy under diverse lighting intensity and temperature conditions. We observe that CleAR consistently achieves low estimation errors across diverse lighting conditions, whereas the estimation quality of baseline methods varies significantly with changes in lighting.

Related works on mobile AR lighting estimation systems seek to extract environment information from physical light probes [36], user dynamics, and learning-based solutions [20, 48, 52, 58]. While physical light probes provide the most comprehensive environment observations, their use is limited by the need for physical setup. Consequently, AR applications often use learned models to estimate lighting from camera images. Over the past couple of decades, learning-based methods have evolved from discovering scene lighting cues from image details, such as highlights and shadows [53], to regressing omnidirectional environment lighting representations [20, 57, 58]. However, autoregressive models cannot effectively tackle the environment information generation in lighting estimation. For example, Xihe [58] provides real-time low-frequency lighting estimation for AR applications but lacks support for detailed environment reflections in object rendering. In contrast, recent advances in generative models enable the estimation of highly detailed environment maps [48, 52], opening new possibilities for visually coherent rendering. Our work explores the integration of image-generative models into AR systems to enable high-quality environment lighting estimation, with novel AR context-aware design to improve generation speed, accuracy, and robustness.

We summarize our main contribution as follows:

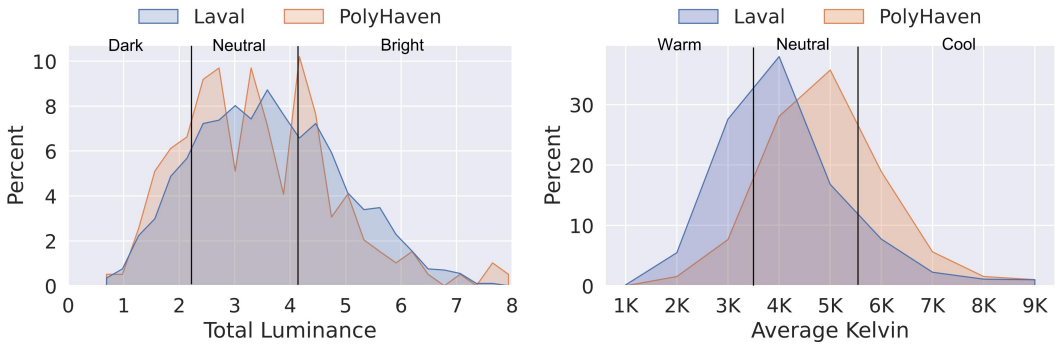
- We design and implement CleAR, a generative lighting estimation system that allows mobile AR applications to acquire environment lighting to support visually coherent AR experiences. The relevant research artifacts are at https://github.com/cake-lab/CleAR.
- We introduce a two-step generative pipeline to produce more accurate and visually coherent environment maps aligned with the user's physical surroundings. This design effectively overcomes the limited environmental sensing of AR devices and smartly leverages multi-modal AR context to achieve accurate estimation.
- To further improve estimation accuracy and real-time robustness, we develop a set of refinement techniques that work in tadam with the generative estimation pipeline. These refinement components use a multi-output estimation strategy to address ambiguity and adapt the estimated environment color based on real-time AR device camera observations.
- We evaluate CleAR through both quantitative and qualitative experiments, comparing it against SoTA lighting estimation systems and models on a commonly used dataset [19] and a robustness-focused variant generated by us. A user study based on perceptual quality ratings shows that CleAR outperforms all baselines, scoring 12% higher than the second-best method, with 7% lower rating variance.

2 BACKGROUND

Environment Lighting Representations. Lighting estimation has been a long-standing research question in the vision and graphics community [30]. Pioneering works have established several methods for capturing high-fidelity lighting of physical environments. For instance, environmental lighting can be captured with mirror balls, panoramic cameras, and photometric stereo techniques [14, 47, 50]. The captured environment lighting can also be represented in various formats, including parametric format [22], neural representations [34], and image-based format [14]. In modern computer graphics, mirror ball-based capturing and image-based lighting are widely used for accurate lighting rendering. However, mirror ball setups can be cumbersome in practice [36], making image-based lighting a more practical approach for mobile AR, which this work adopts. Image-based lighting relies on acquiring high-dynamic range (HDR) environment maps that capture omnidirectional lighting at a specific location. HDR images are critical because they preserve the full range of luminance—including bright highlights and subtle shadows—often lost in standard low-dynamic range (LDR) images [16]. To enable realistic lighting effects and photorealistic virtual content, this work focuses on generating HDR environment maps.

Lighting Estimation in Mobile AR. In mobile AR, accurately estimating environment lighting is crucial for creating immersive visual experiences when overlaying virtual content on physical environments. Such application includes virtual try-ons [56] and interactive games [40]. Obtaining accurate environmental lighting in mobile AR faces several additional challenges compared to traditional image— or video-based lighting estimation tasks. For example, mobile AR devices typically have limited environmental sensing capabilities, such as the field of view (FoV) of their cameras. Because visually coherent AR experiences require omnidirectional environment lighting, the limitations in environment sensing render lighting estimation in mobile AR a highly ambiguous process and, thus, an ill-posed problem. Traditional methods [19, 20, 43] often aim to solve this problem using deep models, which result in lengthy computation times. However, mobile AR applications require timely environment lighting updates to maintain visual coherence in temporally changing lighting environments. Recently, mobile system researchers have begun leveraging enhanced device capabilities [58] and user-in-the-loop interactions [60] to acquire more reliable environmental information, thereby improving the accuracy and robustness of lighting estimation in mobile AR. Despite these advances, current state-of-the-art methods still fall short of providing the accuracy and robustness required for reliable lighting estimation in mobile AR applications.

Conditional Generative Models. Recent years have witnessed significant advancements in the data synthesis capabilities of generative models [13, 52]. Various model architectures, such as Generative Adversarial Networks (GANs) [21], Variational Autoencoders (VAEs) [27], have achieved notable success in generating realistic images, text, and other forms of data. However, the content generation process in these models typically cannot be conditioned. In other words, users of these models cannot control the outputs through external guidance or specific input signals. Conditional generative models extend generative models by incorporating additional input information, referred to as conditions, to guide the generation process. These conditions can range from class labels in image generation [31] to textual descriptions in text-to-image models [38] or even semantic features for structured data generation [25]. In this work, we base CleAR on the ControlNet [54] architecture, a type of state-of-the-art conditional diffusion model. ControlNet works by extending a pre-trained diffusion model, such as Stable Diffusion [39], with additional neural network layers that allow external conditions to influence the diffusion process. Specifically, ControlNet uses a trainable copy of the diffusion model's encoder to encode the input conditioning latent code and merge it with the diffusion model's latent code. By doing so, it maintains the strong generative priors of the base diffusion model while allowing the generation process to be guided by external conditioning inputs.



(a) Light Intensity (b) Color Temperature

Fig. 2. Distributions of lighting condition in two datasets. We analyze the distributions of light intensity and color temperature on environment maps collected from the Laval indoor dataset and PolyHaven. Both datasets exhibit data biases. For example, a significant part of the PolyHaven data items lies in the cool color range. Also, both datasets have an imbalanced distribution between neutral and other lighting conditions.

3 MOTIVATION AND CHALLENGES

Our analysis uses two standard lighting estimation datasets: the *Laval dataset* [19], an academic open research dataset, and the *PolyHaven* website [1], a royalty-free HDR environment map data website. On these data, we calculate the *light intensity* as the total luminance of a given HDR environment map image, along with the environment map image's color temperature values in Kelvin. Additional details of analysis setup, measurement calculations, and technical details are included in Appendix §A. Figure 2 shows the lighting condition distributions with annotated human-perceivable categories for light intensity and color temperature.

Our analysis reveals two critical challenges. First, *real-world lighting conditions are complex and diverse*. Both datasets include a wide range of light intensity and color temperature values. In particular, the measured light intensity spans from as low as 0.6 to over 8 in luminance values, capturing both dim and brightly lit scenes. Similarly, the color temperature varies widely, ranging from below 1000K, typical of incandescent lighting, to over 9000K, which is seen in artificial lights. This suggests that real-world lighting conditions have complex and diverse properties. In many real-world scenarios, environmental lighting conditions often vary across scenes and over time. For example, indoor rooms lit by sunlight often exhibit neutral to warm color temperatures, while those illuminated by LED sources tend to appear cooler. Therefore, lighting estimation systems must have rich built-in knowledge of diverse environmental lighting conditions, as well as adaptive application-time policies to react to real-time environment lighting changes. To address this challenge, in §4, we introduce an end-to-end system that takes multimodal *AR context* to guide the generation of a visually coherent HDR environment map with color refinement and color palette matching.

Another key challenge revealed by our study is the *inherent data biases*, which pose challenges to training and evaluating generative models in lighting estimation systems. These biases in training data can lead to overfitting on dominant lighting conditions and poor generalization to underrepresented lighting features. For instance, models trained predominantly on neutral or cool lighting conditions may fail to produce plausible estimations under warm or high-intensity illumination. Furthermore, biased data distributions in testing data can skew evaluation results, making it difficult to evaluate the robustness of lighting estimation systems. To address the data bias issue, We introduce a *data balancing technique* in Appendix §B, and present a robustness testing protocol in §6, through which we evaluate several recent lighting estimation models.

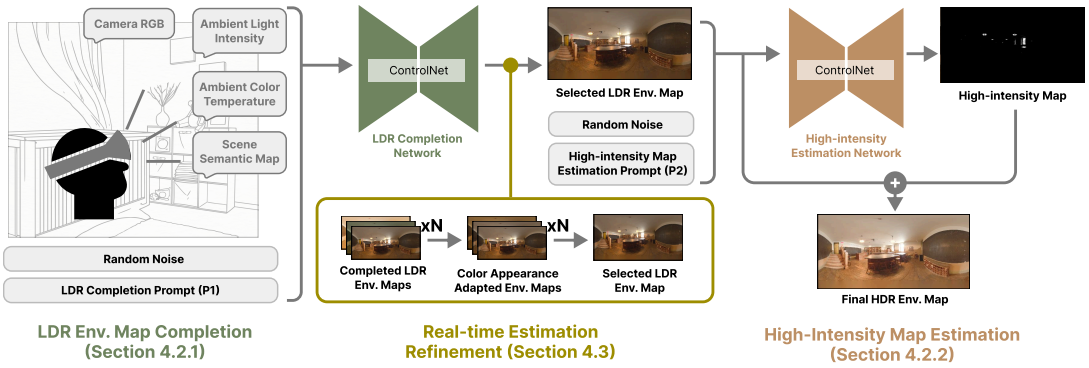


Fig. 3. Overview of CLEAR. CLEAR uses four types of AR context data to guide a two-step high-quality lighting generation pipeline. The first step (§4.2.1) completes partial environment observations in the LDR pixel range domain. The completed environment maps will be post-processed and selected using our real-time estimation refinement components (§4.3) and then passed on to the subsequent high-intensity map estimation step (§4.2.2). Finally, CLEAR outputs an HDR environment map by combining the completed LDR environment map and the high-intensity map.

4 CLEAR DESIGN

4.1 Overview

At a high level, estimating HDR omnidirectional lighting from low-FoV LDR environment observations is a process of significant information increase. To tackle this challenging problem, CLEAR leverages the recent advancements in conditional image generative models [54] to create missing environment lighting information, guided by AR context data. The key design of CLEAR is a novel integration of image generative models to mobile AR lighting estimation systems to ensure high-quality and robust estimation results.

As illustrated in Figure 3, the design of CLEAR revolves around an AR context-guided generative lighting estimation pipeline (§4.2). CLEAR estimates omnidirectional environment lighting, a 360° HDR environment map image, from multimodal AR context data. In this work, we consider four types of AR data, including low-FoV LDR images, ambient light intensity and color temperature, and scene semantic map. Our context-guided generation design addresses the key limitation of image generative models. While generative models excel at image generation in general, producing accurate HDR environment maps requires precise conditioning to ensure visual consistency with real-world scenes, especially under complex lighting conditions and limited environmental observations. We address this challenge by *using AR context data to guide generation*, offering both coarse- and fine-grained conditioning.

Another critical aspect of robust mobile AR lighting estimation is the timely update of estimation results to address changes in environmental lighting. However, the high inference latency of generative models limits the frequency of lighting updates. To address this challenge, CLEAR employs an estimation refinement workflow (§4.3) that adapts the color profile of estimated environment maps to real-time AR camera images, ensuring temporally consistent lighting estimation. Specifically, CLEAR first employs a multi-output generation strategy (§4.3.2) to create accurate yet diverse estimation results for unobserved environments. Then, CLEAR uses several techniques (§4.3.1 and §4.3.2) to refine the previously estimated environment maps and select the best one as the estimation output.

4.2 Context-guided Generative Lighting Estimation

4.2.1 LDR Environment Map Completion. In the first step, CLEAR completes the environment map from limited camera observations, treating the standard 8-bit pixel range as the LDR range. Our LDR completion model builds

on the ControlNet [54] architecture with pre-trained Stable Diffusion [39] model weights. We choose ControlNet for its flexible controllability using text and image conditioning, which integrates well with the AR context. The pre-trained StableDiffusion model provides strong prior knowledge learned from large-scale image generation datasets[1]. These priors enable us to train environment map generation models without starting from scratch entirely, also allowing for better generalization to diverse real-world scenarios. While Stable Diffusion excels at general image generation, it must be fine-tuned on environment map data to adapt it to our task of generating panoramas in the correct format and style.

During the completion process, we encode four types of AR context data into three input modalities for ControlNet: *(i)* a camera RGB image as direct environmental observation, *(ii)* a scene semantic map for structural conditioning, and *(iii)* a text prompt encoding ambient light intensity and color temperature. As part of input preparation, we stitch multi-view camera RGBs into 360 ° panoramic environment map images in advance. This process can be extended with more sophisticated reconstruction methods like [60] or combined with multi-user observation sharing to increase the observed environments. Similarly, environment semantic maps are also processed into 360 ° panoramic images, following the ADE20K [63] definition with 150 instance labels.

To encode ambient lighting conditions, we convert numerical sensor data into natural language descriptions of the environment's lighting characteristics. We use the following text prompt template to describe ambient lighting conditions. The lighting condition words are marked in bold font.

P1: *A panoramic photo of an indoor room. The room is in a **[dark/neutral/bright]** lighting condition. The room has a **[warm/neutral/cool]** ambient color.*

The ambient light property labels are created based on partial LDR image pixels with empirically derived threshold values. The light intensity label is defined by the mean pixel intensity [7] with values between 0.25 and 0.40 as *neutral*, and the rest of the values as *dark* and *bright*. We define the color temperature label based on the mean pixel color temperature: values between 3500K and 5500K are considered *neutral*, while values outside this range are categorized as *warm* or *cool* [32].

4.2.2 High-intensity Map Estimation. The second step of our pipeline is to estimate high-intensity components in the completed LDR environment maps. The high-intensity components describe environment light intensities and directionalities, which are critical for rendering photorealistic highlight and shadow effects on AR objects. However, due to limited camera capabilities, mobile AR devices often fail to capture high-intensity lighting information. As shown in Figure 4, we decompose each HDR environment map into two parts using Equation (1)—an LDR representation and a high-intensity component (in the form of a high-intensity map)—for visualization. We define a *high-intensity map* as a 360° LDR image that describes the high-intensity light positions, directions, and pixel values from the original HDR environment map.

$$I_m = 2.0/(1 + e^{-I_i}) - 1.0 \tag{1}$$

To generate high-intensity maps, we again condition a generation model on the completed LDR environment map produced in §4.2.1. One of the key challenges in building reliable high-intensity map estimation models is the scarcity of high-quality HDR environment maps. For example, the Laval dataset includes only about 2K data items. Training generative models on small datasets can be highly unreliable. Additionally, pre-trained backbone models, like Stable Diffusion [39], cannot be directly used for high-intensity map estimation because they are trained on LDR images.

To address these challenges, we propose a novel learning formulation to allow high-intensity estimation in the LDR pixel range domain. Specifically, our design includes two key steps. First, we use a scaling transformation

[1]Note that CLEAR's design is not tied to ControlNet nor Stable Diffusion; other conditional image generative models can serve as drop-in replacements to improve the performance of CLEAR in the future.

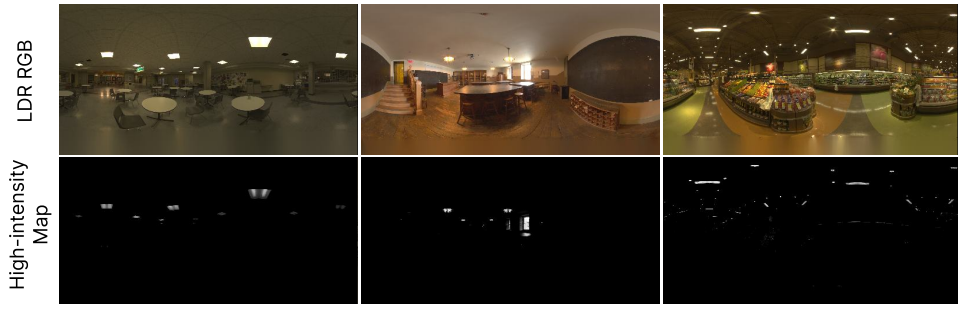


Fig. 4. Visualization of high-intensity pixels. We decompose three examples of HDR environment maps into an LDR RGB component (row 1) and a high-intensity component (row 2). The high-intensity pixels, i.e., bright light spots on the environment map, are extracted using Equation (1).

(Equation(1)) to convert HDR environment maps into LDR and high-intensity maps. Then, we fine-tune a pre-trained generative model, using the transformed Laval dataset, to learn the mapping from LDR environment maps to high-intensity maps. This is feasible because *both input and output are represented in the LDR pixel range*. More importantly, the high-intensity values often align with simple LDR image features, such as bright spots and light strips, making high-intensity map estimation a simpler task compared to the LDR completion task. During training and inference, we use the following text prompt to guide the high-intensity estimation process in producing consistent image style and reducing visual artifacts.

P2: *A grayscale panoramic image describing the bright spots of an indoor room. Brighter spots get more bright color. Regions without light sources should stay pure black.*

This novel design allows us to overcome the HDR data availability challenge and use the small amount of ground-truth HDR data from the Laval dataset to train the high-intensity estimation model. Combined with the completed LDR environment map from the first step (§4.2.1), our generative lighting estimation pipeline outputs a 360° HDR environment map.

4.3 Real-time Estimation Refinement

Completing 360° environment maps from low-FoV environment observations is an inherently ambiguous process. Although AR context data helps condition the generation process, each generation can take several seconds—leading to potential mismatches between the current AR camera image and the one used during generation. To address this, we introduce a multi-output estimation process co-designed with our two-step generation pipeline. Our key insight is to generate multiple completion variants and select the one that best matches the real-time AR view. At a high level, the process consists of three steps as illustrated in Figure 5.

Our multi-output generation design not only enables CleAR to output accurate lighting estimation results, but also provides practical support for responsive lighting estimation in the real-time AR application workflow. The high memory footprint of generative models, often exceeding 2–4 GB during inference [39], makes them unable to run on resource-constrained mobile platforms that must concurrently handle rendering and sensing tasks. In contrast, our on-device refinement method enables the estimated environment maps to adapt to minor changes in environmental lighting, thereby reducing the need for frequent inference of the generative model. To facilitate real-time deployments, CleAR adopts an edge-device collaborative estimation architecture, similar to recent works [6, 58]. Specifically, LDR environment map completion and high-intensity map estimation can be offloaded to an edge server, while lightweight color refinement and result selection algorithms are executed on-device.

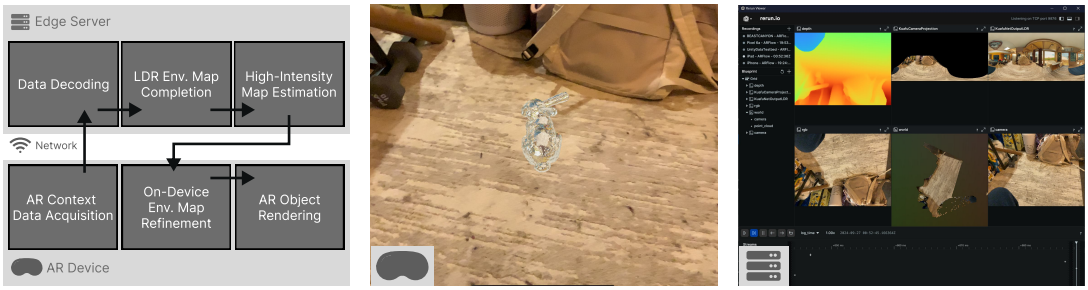
Fig. 5. Estimation refinement workflow. CLEAR uses a hybrid architecture to offload heavy generative model inference to the edge server while adaptively refining estimation results on AR devices in real time. We adopt a multi-output generation strategy to tackle the estimation ambiguity. On the client, our system first matches the color appearances between the completed environment maps and the real-time camera observations (§4.3.1). This operation helps improve estimation accuracy in challenging lighting conditions and enables real-time adaptation of the generative estimation result. Then, our system client chooses the best estimation result based on real-time environment observation. The chosen environment map is combined with its corresponding high-intensity map estimation to create the final HDR environment map.

4.3.1 Color Appearance Adaptation. Addressing unpredictable changes in environmental lighting is crucial for supporting temporally consistent rendering of AR objects. Therefore, CLEAR employs a lightweight on-device refinement technique to refine the multi-output LDR completion results to adapt their color appearances to real-time AR camera images. While other additional metrics like semantic consistency or structural similarity are also important for environment map quality, we consider recovering these properties as orthogonal problems that can be addressed separately in real-time environment reconstruction systems [46].

Our design selects the AR camera image as the adaptation target because the ultimate goal of mobile AR lighting estimation is to support the visually coherent insertion of AR objects into camera images. Additionally, our on-device real-time color appearance adaptation design is well-positioned to address the real-time estimation needs in mobile AR. By applying color adaptation to the estimated environment maps, CLEAR achieves consistently high estimation accuracy with few generative model inference requests.

A recent work, DiffusionLight [35], uses a similar strategy that has to invoke large quantities of inferences to achieve desirable quality results. However, generative models, particularly diffusion-based models, typically suffer from long inference latency. Under the default setting, DiffusionLight requires 60 diffusion model inference calls to generate one estimation result. As we will show in §6, DiffusionLight incurs a very high end-to-end inference time (359.9s), making its multi-output estimation strategy infeasible for real-time AR applications. In comparison, using our color adaptation technique, CLEAR can achieve better estimation accuracy than DiffusionLight while using approximately 110X less time. Additional implementation details and testing examples of color refinement are included in Appendix §D.

4.3.2 Estimation Output Selection. As the final step of the real-time estimation refinement process, CLEAR selects the best environment map image from the previously refined LDR environment map images. In this process, CLEAR searches for the best estimation candidate by identifying the highest cosine similarity between the color palettes of the observation and estimation environment map images. The color palettes are selected using the K-means algorithm, which identifies the five most common colors for each environment map image. The total cosine similarity is calculated as the sum of per-color cosine similarity between color palettes. The selected LDR environment map will later be combined with its corresponding high-intensity map to form a complete HDR environment map for mobile AR applications.



(a) System Architecture Overview (b) Example AR Application (c) Edge Server Data Visualization

Fig. 6. System implementation and AR application integration. Figure 6a shows an overview of the software components of CleAR. Figure 6b depicts an AR application screenshot of a rendered virtual bunny with CleAR estimated environment lighting. Figure 6c illustrates the visualization component at the edge server for the collected AR data.

5 IMPLEMENTATIONS

5.1 System Implementation

We implement CleAR based on a recent AR data streaming framework, ARFlow [59], which enables low-latency bidirectional data streaming between the AR device and the edge server. We also implement an example object placement AR application with CleAR to demonstrate the end-to-end rendering results. Figure 6 shows an overview of the CleAR architecture. We implement the client as a Unity3D[2] package. The client is responsible for streaming and receiving data from mobile AR devices and providing AR applications with estimated environment maps. We use ARFlow's built-in data collection feature to capture camera RGB, depth, and device tracking data provided by ARFoundation[3], Unity's low-level AR framework. The captured data is streamed to our server via the ARFlow gRPC service. The server component is a Python-based service for the generative lighting estimation pipeline. During generative model inference, we combine the RGB and other context information using the `MultiControlNet` pipeline from the HuggingFace `Diffusers`[4] library to run different ControlNet models. We configure the ControlNet model inference to use the UniPC [55] sampler with 20 steps. Our ControlNet models output environment maps at a resolution of `512x256`. `xformers`[5] is used to accelerate model inference.

5.2 Training and Testing Data Generation

We curate a large-scale dataset for training and testing CleAR. For LDR completion model training, we assemble 30K LDR environment map images from two prominent LDR indoor environment map sources: the Matterport3D dataset [9] and Structured3D datasets [62]. Matterport3D dataset provides real-world captured environment maps, and Structured3D dataset consists of photorealistic synthetic ones. For high-intensity map estimation model training, we use the training split of the Laval dataset [19] to create 1,489 pairs of LDR environment map images and corresponding high-intensity map images. To evaluate CleAR under diverse lighting conditions, we create variants of the Laval test set that preserve the original environment map visual context. We generate these variants by uniformly scaling the environment map images. For light-intensity editing, all color channels are scaled equally; for color temperature editing, only the red and blue channels are adjusted [2]. See Appendix §B for additional details on the data generation process, and Appendix §C for details of model training.

[2]Unity3D: https://unity.com
[3]ARFoundation: https://docs.unity3d.com/Packages/com.unity.xr.arfoundation@6.0/
[4]Diffusers: https://huggingface.co/docs/diffusers
[5]xFormers: https://facebookresearch.github.io/xformers/

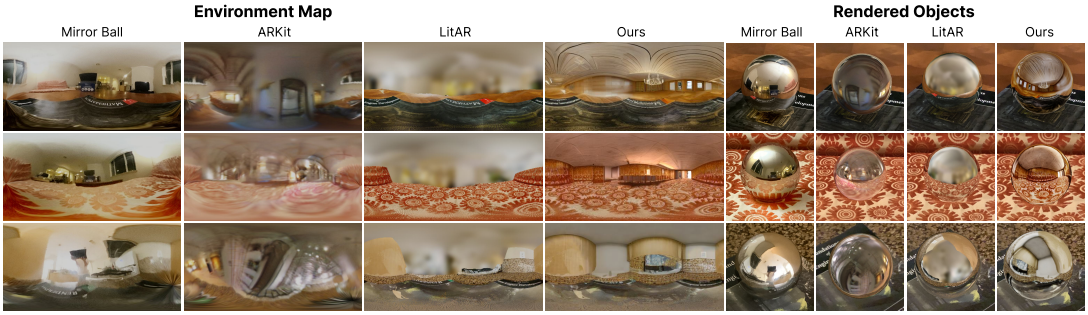


Fig. 7. AR object rendering results. We show AR virtual object rendering qualitative comparison to other environment lighting acquired by unwrapping a physical mirror ball [14] (reference), ARKit [4] (commercial), and LitAR [60] (academic). Compared to ARKit, CLEAR estimates the environment map with significantly more realistic visual details. Compared to LitAR, CLEAR can generate more visually coherent far-field environment map details, while LitAR can only generate blurry ones.

6 EVALUATION

To comprehensively evaluate CLEAR, we use a combination of quantitative and qualitative methods to assess key aspects of system performance, including generation latency, output quality, and user perception. We conduct a lab-based evaluation using an example AR application built on top of CLEAR to assess rendering quality and latency. Our data-driven evaluation compares CLEAR to several state-of-the-art lighting estimation methods [3, 35, 48] using both the standard three-sphere protocol and our proposed robustness testing protocol under varied lighting conditions. In addition, we use an online survey to understand the impact of lighting estimations on human perceptual preferences. Our results demonstrate that CLEAR achieves high-quality lighting estimation with low latency. For example, our end-to-end estimation results show that CLEAR achieves the best performance in two image-based metrics while taking significantly less time. Participants also rated CLEAR-generated lighting as producing better renderings for most virtual objects.

6.1 Testbed-based Evaluation

6.1.1 Testbed Setup. We conduct our evaluation using an example AR application running on a 4th-generation 11-inch iPad Pro with a built-in LiDAR sensor. The edge server is a high-end workstation with an Intel Core i9-13900K CPU, 128GB of RAM, and an NVIDIA RTX 4090 GPU. The edge server runs an Ubuntu 22.04 operating system. The mobile device and edge server are connected via an 802.11ac Wi-Fi network. To ensure consistent and reliable results, all system runtime measurements are performed 10 times, and the reported values are the average across these runs.

6.1.2 Qualitative Visual Comparison. We showcase CLEAR's lighting estimation quality with our testbed AR application in real-world environments. We followed the experiment setup of LitAR [60], a recent lighting estimation mobile system, to compare the quality of lighting estimation on rendered virtual sphere images. In Figure 7, we show the visual comparison between two baselines: LitAR [60] and ARKit [4]. We also include a physical mirror ball as a reference for physical environment lighting. For fair comparison, we use the same near-field reconstruction data from LitAR as the RGB input to CLEAR. Overall, we observe that CLEAR outputs the best environment map image quality among estimated environment maps (ARKit and LitAR). CLEAR generates environment maps with more visual details than LitAR and fewer artifacts than ARKit. Particularly, CLEAR can generate detailed far-field environments while LitAR can only generate blurry ones. When comparing rendered virtual mirror balls to the physical reference mirror ball, CLEAR produces the closest visual match.

Table 1. System performance breakdown. (Left) We show the breakdown of system latency under the default generative estimation settings, which use five LDR completion outputs and do not use semantic maps in the input. In the left table, on-device operations are marked with `D`, edge operations are marked with `E`, and network operations are marked with `N`. (Right) We show the end-to-end system execution time (excluding network time) under various system configurations. Our system can generate an environment map as fast as 2,383 ms. Under the default system configuration, which is marked in **bold** font, the end-to-end execution time is 3,269 ms.

	System Component	Avg. Time (ms)
D	Device data preparation	0.5 (± 0.1)
N	AR Data Offloading	31 (± 10.1)
E	LDR completion x5	1957 (± 9.3)
E	LDR env. map retrieval	72 (± 8.1)
D	Estimation refinement	3 (± 0.2)
E	Refinement result synchronization	10 (± 0.6)
N	High-intensity map estimation	1166 (± 10.7)
N	High-intensity map retrieval	30 (± 8.9)

System Configuration	Avg. Time (ms)
LDR completion x1 (RGB)	2383 (± 11.4)
LDR completion x1 (RGB + Semantics)	3021 (± 51.3)
LDR completion x3 (RGB)	2825 (± 17.6)
LDR completion x3 (RGB + Semantics)	4526 (± 31.5)
LDR completion x5 (RGB)	**3269 (± 19.5)**
LDR completion x5 (RGB + Semantics)	5855 (± 38.5)
LDR completion x7 (RGB)	3741 (± 14.2)
LDR completion x7 (RGB + Semantics)	8082 (± 56.9)

6.1.3 System Performance Breakdown. Table 1 shows CleAR's component-wise latency and its end-to-end execution latency under different configurations. In the default configuration, CleAR takes RGB images and ambient light data from the AR device as input to generate five LDR completion variants and one corresponding high-intensity map. The end-to-end execution takes 3,269 ms, primarily consisting of the generative LDR completion time. When combined with semantics in LDR completion, the end-to-end execution takes 5,855 ms. We observe that the main impact factor for the LDR completion process is the number of environment map outputs. Increasing generation output introduces a near-linear yet slow increase in the estimation latency. Although our default system configuration uses five outputs, the estimation latency is only about double that of using a single estimation output. Note that these long execution time could be masked by pipelining per-frame requests to a powerful edge server, if needed. However, with our on-device refinement, we have the feasibility to skip expensive generative inferences yet still meet the real-time and quality goals. Specifically, once the high-intensity map is generated, CleAR can generate visually coherent HDR environment maps in real-time by matching the LDR environment map to the current camera view's color appearance. If the environmental lighting conditions do not change much, our design allows for the reuse of expensive generative results across frames without sacrificing the visual quality. To put CleAR's latency in context, we note that a recent AR-specific lighting estimation system LitAR [60] needs 46.6/134.38ms to generate low/high quality near-field portion of the environment map. Another AR-specific system Xihe can achieve real-time lighting estimation (20.1ms) but can only produce low-fidelity spherical harmonics coefficients. In contrast, CleAR can deliver real-time HDR environment maps, which are much higher quality lighting information.

6.2 Data-driven Evaluation

6.2.1 Experiment Setups. We quantitatively evaluate the lighting estimation quality of CleAR against three state-of-the-art lighting estimation models: DiffusionLight [35], StyleLight [48], and OmniDreamer [3]. Among these baselines, DiffusionLight is most similar to CleAR because it uses both diffusion models and a multi-output generation pipeline. We use the Laval test set and its augmented variants (§5.2) to evaluate the accuracy and robustness of lighting estimation methods using the following two protocols:

- **The three-sphere evaluation protocol** [35, 48] uses a proxy method to evaluate the HDR lighting estimation accuracy through pixel-wise error measurements on three rendered spheres with different material properties. Specifically, the used sphere materials are: *matte*, *silver matte*, and *silver*. Figure 8 shows a set of rendered

Fig. 8. Visualizations of two experiment protocols for data-driven evaluation. For the three-sphere protocol (left), we use environment maps (middle) from the Laval dataset to render virtual spheres of different material properties. For the robustness testing protocol (right), we use the robustness testing dataset, consisting of environment maps from Laval with edited ambient light intensity and color temperature. As an example, we show an environment map (middle) with four variants with different light intensities and color temperatures. The results (right) of light intensity augmentation are marked as *dark* and *bright*, and the results of color temperature augmentation are marked as *warm* and *cold*.

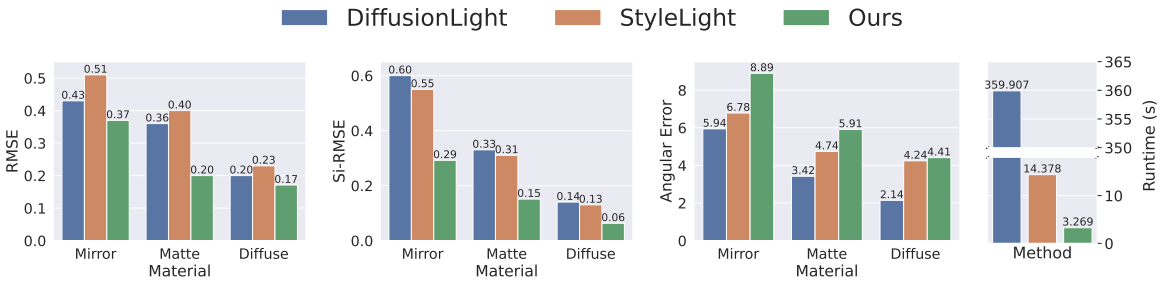
virtual spheres used in this evaluation protocol. Each material represents a mixture of specific reflectance and roughness properties. Compared to direct environment map-wise comparison, this evaluation protocol helps us understand the *impact of lighting estimation on object rendering quality*.

- **The robustness testing protocol** is designed by us to test the *generalizability and robustness of lighting estimation systems* under different lighting conditions, specifically changes in light intensity and color temperature. In this protocol, we test lighting estimation systems on the Laval variants and compare estimated environment maps with the ground truth to calculate pixel-wise errors. The comparison is performed in the LDR image domain to capture lighting condition-related color appearance differences between environment map images.

Evaluation Metrics. Following prior works [3, 20, 35, 48], we use the FID score [41] to measure the diversity of generated environment maps. For evaluating the accuracy of environment maps, we follow [60] to use RMSE. We use three image-based metrics—scale-invariant Root Mean Square Error (si-RMSE) [17], Angular Error [29], and RMSE—to evaluate lighting estimation accuracy on rendered virtual spheres. The si-RMSE and RMSE quantify pixel-wise differences, with si-RMSE being insensitive to global intensity scaling. Angular Error focuses on differences in pixel chromaticity, emphasizing the assessment of environment lighting color properties. We follow [35] to map the 0.1st and 99.9th value percentiles to 0 and 1 when calculating RMSE.

6.2.2 End-to-End Visual Quality and Performance. In Figure 9, we show qualitative comparisons between CLEAR and baseline methods in three scenes, representing the neutral, warm, and cool lighting conditions. Compared to other methods, CLEAR allows more accurate overall color tones rendering and creates more diverse reflection details on the rendered virtual spheres. Figure 10 presents a quantitative comparison between CLEAR and two SoTA methods [35, 48] that can output HDR environment maps in both rendering quality and estimation latency. Under the same environmental observations, CLEAR achieves the lowest scale-invariant and normalized RMSE values across all three virtual sphere types. Noticeably, for the mirror sphere, the most challenging material of the three, CLEAR achieves approximately 50% reduction in the scale-invariant RMSE values. On the other hand, we observe that CLEAR generates environment maps with slightly higher angular errors. Upon further inspection, we suspect that the diverse environment map image details potentially caused the higher angular error because the angular error metric is sensitive to pixel color differences. Furthermore, CLEAR achieves the lowest lighting estimation latency by large margins. In particular, compared to DiffusionLight [35], a diffusion model-based method, CLEAR takes significantly less time (110X).

Fig. 9. Qualitative comparisons. We show examples of lighting estimation results on three scenes and the corresponding rendered virtual sphere images. For each method, we show rendered virtual spheres (row top) and estimated environment maps (row bottom). Note, OmniDreamer can only output LDR environment maps.



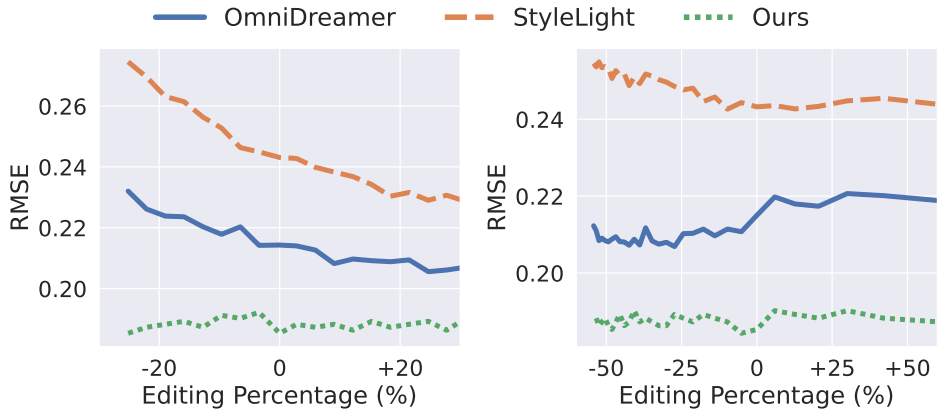
(a) RMSE (b) Si-RMSE (c) Angular Error (d) Latency

Fig. 10. End-to-end qualitative comparison. We compare the end-to-end estimation accuracy and latency between CleAR and two state-of-the-art works [35, 48]. We test estimation accuracy through the three-sphere evaluation protocol, which directly measures the visual quality impact of environment lighting on AR virtual object rendering. CleAR achieves the lowest si-RMSE and RMSE values, on all three testing object material types in the three-sphere evaluation protocol. However, CleAR has slightly higher angular errors on the mirror and matte spheres. Moreover, CleAR achieves this competitive accuracy with significantly lower end-to-end latency compared to the other generative model-based methods.

Impact of Context Data Availability. We follow the setups in [35, 48] to use a single centered view of a 75° horizontal FoV camera as the default configuration for CleAR. This configuration represents a baseline performance of CleAR because our system can take in more AR context data during runtime. Table 2 shows

Table 2. Impact of context data. We evaluate CLEAR's estimation accuracy under different configurations based on the three-sphere protocol. The first row shows the default configuration of CLEAR, marked in gray, which already outperforms DiffusionLight and StyleLight in both si-RMSE and RMSE. As more context data is used—larger FoV, environmental semantics, and more camera observations, CLEAR can further improve the rendering quality.

Configuration	Scale-invariant RMSE ↓			Angular Error ↓			RMSE ↓		
	Diffuse	Matte	Mirror	Diffuse	Matte	Mirror	Diffuse	Matte	Mirror
RGB 75° FoV 1 view	0.06	0.15	0.29	4.42	5.92	8.89	0.17	0.21	0.37
RGB+semantics 75° FoV 1 view	0.06	0.14	0.28	4.41	5.91	8.89	0.17	0.20	0.37
RGB+semantics 110° FoV 1 view	0.05	0.13	0.27	4.19	5.77	8.66	0.16	0.19	0.35
RGB+semantics 110° FoV 3 views	0.03	0.10	0.24	3.89	5.21	8.23	0.13	0.17	0.31
RGB+semantics 110° FoV 5 views	0.02	0.09	0.22	3.71	5.13	8.17	0.11	0.16	0.29
RGB+full semantics 110° FoV 3 views	**0.02**	**0.07**	**0.20**	**3.59**	**4.96**	**8.09**	**0.10**	**0.15**	**0.27**



(a) Light Intensity (b) Color Temperature

Fig. 11. Robustness testing result. We test the robustness of our lighting estimation results against two recent lighting estimation models [3, 48] using our robustness testing protocol. CLEAR shows consistently lower estimation error rates on different lighting conditions. This observation indicates CLEAR can generalize better and provide more consistent result quality in different lighting conditions.

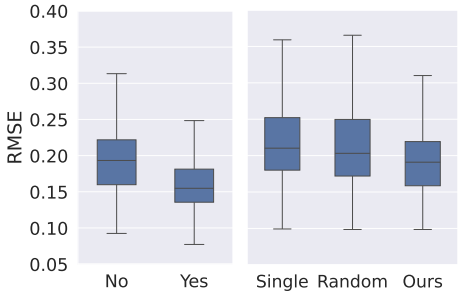
CLEAR's performance under different configurations of camera FoVs, the number of input views, and the usage of semantic maps. Note that the ambient light data has been included in all tests in Table 2 because it is inseparable from our estimation pipeline design. The added environment observations through new views and increased FoVs can both improve the accuracy of lighting estimation. Additionally, full-scene semantic maps are important guiding information for high-accuracy lighting estimation. Full-scene semantic maps can often be obtained as floor maps or scene design layouts in real-world applications. This ablation study suggests that sharing scene semantics can greatly benefit lighting estimation accuracy with our system.

Estimation Robustness Evaluation. We test the lighting estimation robustness of CLEAR under several environmental lighting conditions with the robustness testing protocol. We compare CLEAR against two recent lighting estimation methods that are built on top of generative models: OmniDreamer [3] and StyleLight [48] [6]. Figure 11 shows the RMSE values calculated on the estimated LDR environment maps. We choose to compare the LDR environment maps, instead of the HDR environment maps, because lighting conditions mainly affect

[6]We were unable to complete the robustness testing for DiffusionLight [35] within a reasonable time given its high inference cost.

Method	FID ↓ (full env. map)	FID ↓ (selected regions)
Gardner et al. [19]	307.5	197.4
OmniDreamer [3]	<u>106.3</u>	<u>46.2</u>
StyleLight [48]	137.7	97.2
DiffusionLight [35]	207.2	193.5
Ours	**86.3**	**44.31**

(a) LDR Completion Output Diversity



(b) Color Refinement (c) Output Selection

Fig. 12. Evaluation of generative estimation refinement. (12a) Our method achieves the lowest FID scores, showing strong generation diversity. (12b) Our color appearance matching technique improves estimation accuracy. (12c) Our output selection method reduces RMSE and variance compared to baselines.

the accuracy of LDR environment map estimation. We observe that the accuracy of both OmniDreamer and StyleLight is affected by the changing lighting intensities and color temperatures. While CleAR, on the other hand, shows consistently lower estimation error rates under different lighting conditions. Particularly, these two models exhibit much higher lighting estimation errors on lower environmental lighting intensity and warmer color temperature. The observed error increases potentially are caused by the unbalanced training datasets used by these models.

6.2.3 System Ablation Study. In this section, we evaluate the estimation accuracy of CleAR by comparing its performance across different system configurations and design choices.

Analysis of LDR Environment Map Completion. For the completion of the LDR environment map, we specifically examine the pixel-wise errors and the overall diversity of the environment map content. The former assesses CleAR's LDR lighting estimation accuracy, and the latter evaluates the generation richness of CleAR generative estimation pipeline. For CleAR, we use the LDR environment maps after applying the color appearance adaptation. Following [3, 48], we adopt two methods for measuring the generation content diversity: *(i)* calculate the FID score on the full estimated LDR environment map, and *(ii)* converting the environment map into a cube map and calculate FID scores on each face *without the top and bottom faces* as these two faces containing little information. In Table 12a, we show that our LDR environment map completion model can output environment map images with greater diversity than state-of-the-art models. Our LDR completion model outperforms other models by 27% to 370% in the first calculation method and 4% to 344% in the second calculation method. This observation confirms the effectiveness of our large LDR environment map completion dataset. Next, we compare the accuracy of the LDR environment map completion. Under the same environment observations as [35, 48], our method achieves comparable results to DiffusionLight even though our LDR completion model is based on a smaller-sized pre-trained diffusion model.

Analysis of Estimation Refinement. We evaluate the performance of our on-device estimation refinement components on the impacts of lighting estimation accuracy. In Figure 12b, we show that our color appearance matching technique improves the overall estimation accuracy by 31%. The estimation refinement technique allows our system to achieve high-quality estimation with limited estimation outputs. By default, our system only requires five generation outputs while DiffusionLight [52] requires more than 90 generation outputs. Reducing the required generation output will also reduce the estimation latency and computation resources requirements during deployment. Next, we evaluate the effectiveness of our generation output selection policy. In Figure 12c, we show that our color-matching technique can reduce the estimation error rate compared to other selection

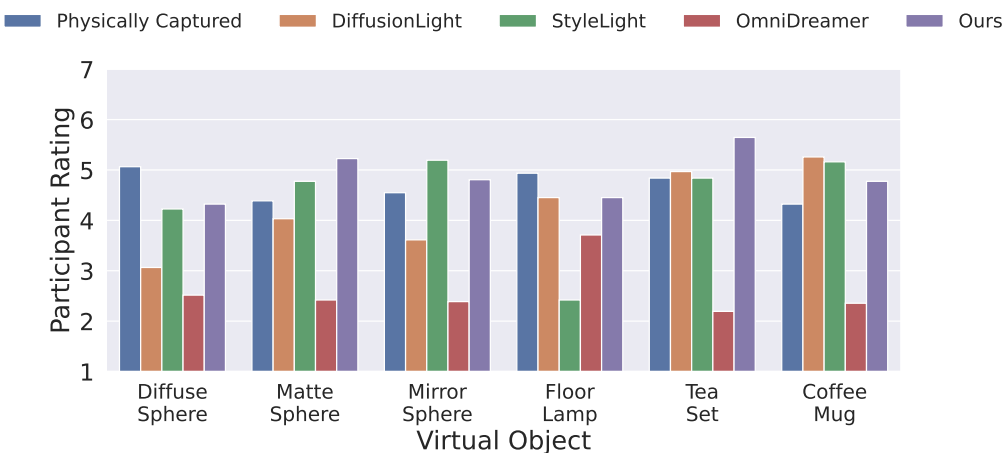


Fig. 13. Human perceptual preference comparisons. We compare virtual object rendering quality between CLEAR and four other methods. Based on responses from 31 participants, CLEAR received the highest average rating of 4.87, outperforming the second-best method, StyleLight, by 12%. CLEAR also shows more consistent quality, with a lower standard deviation.

methods. We observe that, with our output selection policy and a total of five generation outputs, CLEAR can reduce 15% of the estimation error compared to using a single generation output. Furthermore, when using five-generation outputs, our generation output selection component can reduce the average estimation error rate by 11% compared to random selection.

6.3 Perceptual User Study

We conduct an online user study, via Qualtrics[7], to assess the impact of lighting estimation on perceived rendering quality. Our study was approved by our organization's Institutional Review Board (IRB) and then distributed through personal and professional networks.

6.3.1 Study Protocol. Our survey consists of three chapters. The first chapter surveys the participants' past experiences with mobile AR and their impressions of virtual object rendering qualities with existing mobile AR applications. The second chapter is a training section that guides participants in completing the quality assessment study. This chapter first shows a quality rating question where the participant will be given five images of the same virtual sphere rendered under lighting conditions from five different sources. Participants are instructed to rate the visual qualities using a Likert scale of 1 to 7, where 1 represents the lowest and 7 represents the highest quality. To avoid bias, participants are not informed which lighting source was used for each image. An example rating for this question is shown to the user for training purposes. Next, we will show a follow-up question to ask participants for feedback, specifically on the quality issues of images generated by our system. The last chapter is the formal quality assessment, consisting of six question groups. Each group uses different environments and virtual object setups, representing environments with varying lighting conditions and objects with various geometries and materials. Six virtual objects are used in total, including a diffuse sphere, a matte sphere, a mirror sphere, a floor lamp, a tea set, and a coffee mug. Specifically, lighting conditions used in these questions were generated using CLEAR, DiffusionLight [35], StyleLight [48], OmniDreamer [3], and physically captured HDR environment maps. Each group follows the same format used in the training chapter. The lighting conditions used to generate the images are randomized in order and not revealed to participants.

[7]Qualtrics: https://www.qualtrics.com/

6.3.2 Results and Analysis. We received responses from 31 participants (48% Respondents are familiar with multiple areas of graphics technology, most commonly image editing (51.6%), followed by video editing (41.9%) and 3D modeling and rendering (35.5%). 25.8% reported no familiarity with specific graphics technologies. 58% had used AR devices, most commonly mobile phones or tablets (48%), followed by Meta Quest (29%).

Figure 13 shows the visual quality assessment results from the last chapter. Excluding ratings for objects rendered with physically captured lighting, participants rated CleAR-generated lighting as producing better renderings for 4 out of the six virtual objects. The visual quality rating of CleAR ranked the top with an average score of 4.87 across all responses. This score is about 12% better than that of the second-best method, StyleLight, which received an average rating of 4.35. Additionally, CleAR has a lower overall standard deviation in ratings (1.56) compared to StyleLight (1.67), indicating more robust estimation quality.

Interestingly, in two question groups—the tea set and coffee mug—participants rated CleAR 's renderings as higher quality than those produced using physically captured environment maps. Upon further inspection, we found that, while physically captured environment lighting more accurately reflects the radiance conditions in the scene, human perception tends to favor the slightly brighter outputs generated by CleAR. We hope this finding encourages future research to design and evaluate lighting estimation systems with considerations of human perceptual preferences.

7 RELATED WORK

Environment Lighting and Rendering. Pioneering works have established several ways of capturing high-fidelity physical environment lighting and representing it in digital formats. Omnidirectional HDR environment map is a commonly adopted solution [14] for representing the environment lighting because it stores the incoming radiance information that can be easily integrated with modern computer graphics rendering. Environment maps can be commonly captured using mirror balls, 360° cameras, or bracketed image stitching [8, 15]. A recent work, GLEAM [36], incorporates the mirror ball-assisted environment lighting capturing process into AR applications. However, the requirement for the presence of the mirror ball limits the practicality of AR device usage. Instead, our system uses the generative estimation approach to provide flexible, high-quality lighting estimation. Additionally, HDR environment maps can be challenging to obtain on AR devices due to many sensor limitations. Recent lighting estimation systems [58, 60] can only output LDR ones. But our novel two-step generative estimation design allows CleAR to support HDR lighting estimation.

Image Generative Models. Recently, generative models have demonstrated impressive capabilities in image synthesis, as well as generating audio, 3D models, and other data modalities [13, 52]. The generative diffusion model [24] has attracted significant attention from the research community because of its high-quality image generation capabilities. The generative diffusion model uses a physical diffusion process inspired by the Gaussian signal denoising generation process. Combining this novel generation process with the large model parameter sizes, generative diffusion models outperform generative models with prior architectures, such as GAN [21] or VAE [18]. Several recent works [3, 11, 35, 48] also propose to adopt generative models of different architectures to solve the generic lighting estimation problem. Most notably, DiffusionLight [35] achieves state-of-the-art estimation accuracy with diffusion models. In this work, we present a generative model-based approach to lighting estimation tailored for mobile AR applications, addressing key challenges in achieving visually coherent and fast generation.

Context-Aware Mobile System. Context-aware computing is a classical computing paradigm in which applications sense and adapt to contextual information [10]. Early research in context-aware AR systems [44] demonstrated that important environment information can be extracted from camera frames for task planning and decision-making. In recent years, new developments of AR systems have also sought to leverage broad types of environment context information to assist several AR tasks and achieve better user experiences [28, 45, 49]. Our

work leverages four types of AR context data—camera RGB, scene semantic map, and ambient light intensity and color temperature—to guide environment map estimation and align virtual rendering with real-world lighting.

8 DISCUSSION

Application Use Cases. CleAR's design enables a wide range of mobile AR applications that demand visually coherent environment lighting. For instance, CleAR can significantly enhance virtual object rendering in entertainment applications by ensuring that rendered objects consistently blend with real-world scenes under diverse lighting conditions. Additionally, CleAR is also well suited for commercial applications. For example, in interior retail applications, CleAR enables the creation of visually appealing virtual furniture that accurately reflects realistic ambient lighting. Finally, CleAR's tight integration with mobile AR can open new opportunities for versatile solutions in digital production, cinematic content creation, and immersive storytelling.

Implications for Privacy and Security. Our work demonstrates the feasibility and effectiveness of generating high-quality environment lighting information from limited camera observations. While this capability can significantly enhance the realism of virtual object rendering, it also introduces potential risks to user spatial privacy, as the reconstructed environment maps may inadvertently expose details about the user's physical surroundings. For example, recent work demonstrates that sensitive user information can inadvertently be captured by advanced lighting estimation systems and leaked in AR streaming applications [61]. Additional types of information, such as users' location, may be inferred from the completed LDR environment map [23]. Future work should consider defense mechanisms to prohibit the leakage of sensitive user spatial privacy information.

Limitations and Future Directions. While CleAR demonstrates strong performance in generating high-quality environment lighting, it can be further improved in two main ways. First, a useful extension of CleAR would be to support spatially variant lighting estimation, where lighting conditions differ across physical locations. Supporting this capability requires two key changes to the lighting estimation workflow: *(i)* enabling 3D reconstruction of the surrounding environment to allow spatial transformation and alignment of environment observations, and *(ii)* training generative lighting models on spatially variant environment map datasets, which often contain diverse distortion patterns. Existing solutions, such as the near-field reconstruction method in LitAR [60], can be leveraged to develop 3D environment reconstruction methods. However, the primary challenge for training such generative models lies in the lack of high-quality datasets that capture spatially varying lighting. Future research could explore new data curation pipelines or integrate 3D transformation operations directly into generative models. Second, CleAR's temporal consistency mechanism can be extended beyond the color appearance refinement to better support long AR sessions and scenes with dynamic lighting changes. Future work may investigate methods for progressive environment map updates to enhance consistency and realism across time.

9 CONCLUSION

In this work, we introduced a lighting estimation framework CleAR that can be easily integrated into many existing mobile AR applications. Our novel two-step generative lighting estimation pipeline ensures high-quality and robust results under multiple environmental lighting conditions, including various light intensities and color temperatures. Our design uses pre-trained large generative models and AR context data to generate HDR environment maps more accurately from limited environment observations. Our real-time refinement steps enhance the quality of lighting and improve responsiveness to estimation. Comprehensive quantitative evaluation and user study confirm the accuracy and robustness of CleAR compared to recent generative models.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This work was supported in part by NSF Grants #2105564, #2236987, #2346133, #2350189, #2402383, and a VMware grant.

References

[1] 2024. Poly Haven. https://polyhaven.com/. Accessed: 2024-05-01.

[2] Mahmoud Afifi and Michael S Brown. 2020. Deep white-balance editing. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 1397–1406.

[3] Naofumi Akimoto, Yuhi Matsuo, and Yoshimitsu Aoki. 2022. Diverse Plausible 360-Degree Image Outpainting for Efficient 3DCG Background Creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] Inc Apple. 2022. *Augmented Reality - Apple Developer*. Retrieved Oct 14, 2022 from https://developer.apple.com/augmented-reality/

[5] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. 2021. ARKitScenes - A Diverse Real-World Dataset for 3D Indoor Scene Understanding Using Mobile RGB-D Data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. https://openreview.net/forum?id=tjZjv_qh_CE

[6] Ali J Ben Ali, Marziye Kouroshli, Sofiya Semenova, Zakieh Sadat Hashemifar, Steven Y Ko, and Karthik Dantu. 2022. Edge-SLAM: Edge-assisted visual simultaneous localization and mapping. *ACM Transactions on Embedded Computing Systems* 22, 1 (2022), 1–31.

[7] Christophe Bolduc, Justine Giroux, Marc Hébert, Claude Demers, and Jean-François Lalonde. 2023. Beyond the pixel: a photometrically calibrated HDR dataset for luminance and color prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8071–8081.

[8] Matthew Brown and David G Lowe. 2007. Automatic panoramic image stitching using invariant features. *International journal of computer vision* 74 (2007), 59–73.

[9] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision (3DV)* (2017).

[10] Guanling Chen and David Kotz. 2000. A survey of context-aware mobile computing research. (2000).

[11] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. 2022. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)* (2022).

[12] Mathew Chylinski, Jonas Heller, Tim Hilken, Debbie Isobel Keeling, Dominik Mahr, and Ko de Ruyter. 2020. Augmented reality marketing: A technology-enabled approach to situated customer experience. *Australasian Marketing Journal* 28, 4 (2020), 374–384.

[13] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (2023), 10850–10869.

[14] Paul Debevec. 2006. Image-based lighting. In *ACM SIGGRAPH 2006 Courses*. 4–es.

[15] Paul E Debevec and Jitendra Malik. 2023. Recovering high dynamic range radiance maps from photographs. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 643–652.

[16] Frederic Dufaux, Patrick Le Callet, Rafal Mantiuk, and Marta Mrak. 2016. *High dynamic range video: from acquisition, to display and applications*. Academic Press.

[17] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems* 27 (2014).

[18] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12873–12883.

[19] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to Predict Indoor Illumination from a Single Image. *ACM Transactions on Graphics* (2017).

[20] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. 2019. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7175–7183.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[22] Robin Green. 2003. Spherical harmonic lighting: The gritty details. In *Archives of the game developers conference*, Vol. 56. 4.

[23] Jaybie Agullo de Guzman, Aruna Seneviratne, and Kanchana Thilakarathna. 2021. Unravelling Spatial Privacy Risks of Mobile Mixed Reality Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1 (March 2021), 1–26.

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[25] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. 2018. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1335–1344.

[26] Noor A Ibraheem, Mokhtar M Hasan, Rafiqul Z Khan, and Pramod K Mishra. 2012. Understanding color models: a review. *ARPN Journal of science and technology* 2, 3 (2012), 265–275.

[27] Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[28] Kit Yung Lam, Lik Hang Lee, and Pan Hui. 2021. A2w: Context-aware recommendation system for mobile augmented reality web browser. In *Proceedings of the 29th ACM international conference on multimedia*. 2447–2455.

[29] Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. 2019. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5918–5928.
[30] Zhengqin Li, Li Yu, Mikhail Okunev, Manmohan Chandraker, and Zhao Dong. 2023. Spatiotemporally consistent hdr indoor lighting estimation. *ACM Transactions on Graphics* 42, 3 (2023), 1–15.
[31] Mehdi Mirza. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).
[32] Hiroki Noguchi and Toshihiko Sakaguchi. 1999. Effect of illuminance and color temperature on lowering of physiological activity. *Applied human science* 18, 4 (1999), 117–123.
[33] Yoshi Ohno. 2014. Practical use and calculation of CCT and Duv. *Leukos* 10, 1 (2014), 47–55.
[34] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. 2021. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–21.
[35] Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Amit Raj, Varun Jampani, Pramook Khungurn, and Supasorn Suwajanakorn. 2023. DiffusionLight: Light Probes for Free by Painting a Chrome Ball. In *ArXiv*.
[36] Siddhant Prakash, Alireza Bahremand, Linda D Nguyen, and Robert LiKamWa. 2019. Gleam: An illumination estimation framework for real-time photorealistic augmented reality on mobile devices. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 142–154.
[37] Philipp A Rauschnabel, Reto Felix, and Chris Hinsch. 2019. Augmented reality marketing: How mobile AR-apps can improve brands through inspiration. *Journal of Retailing and Consumer Services* 49 (2019), 43–53.
[38] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*. PMLR, 1060–1069.
[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]
[40] Michelangelo Scorpio, Roberta Laffi, Ainoor Teimoorzadeh, Giovanni Ciampi, Massimiliano Masullo, and Sergio Sibilio. 2022. A calibration methodology for light sources aimed at using immersive virtual reality game engine as a tool for lighting design in buildings. *Journal of Building Engineering* 48 (2022), 103998.
[41] Maximilian Seitzer. 2020. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid. Version 0.3.0.
[42] Gowri Somanath and Daniel Kurz. 2021. HDR Environment Map Estimation for Real-Time Augmented Reality. https://arxiv.org/pdf/2011.10687.pdf
[43] Shuran Song and Thomas Funkhouser. 2019. Neural Illumination: Lighting Prediction for Indoor Environments. *CVPR* (2019).
[44] Thad Starner, Bernt Schiele, and Alex Pentland. [n. d.]. Visual contextual awareness in wearable computing. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No. 98EX215)*.
[45] Tomu Tahara, Takashi Seno, Gaku Narita, and Tomoya Ishikawa. 2020. Retargetable AR: Context-aware augmented reality in indoor scenes based on 3D scene graph. In *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 249–255.
[46] Charalambos Theodorou, Vladan Velisavljevic, Vladimir Dyo, and Fredi Nonyelu. 2022. Visual SLAM algorithms and their application for AR, mapping, localization and wayfinding. *Array* 15 (2022), 100222.
[47] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. 2009. Dynamic shape capture using multi-view photometric stereo. In *ACM SIGGRAPH Asia 2009 papers*. 1–11.
[48] Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. 2022. StyleLight: HDR Panorama Generation for Lighting Estimation and Editing. In *European Conference on Computer Vision (ECCV)*.
[49] Tianyi Wang, Xun Qian, Fengming He, Xiyun Hu, Ke Huo, Yuanzhi Cao, and Karthik Ramani. 2020. CAPturAR: An augmented reality tool for authoring human-involved context-aware applications. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 328–341.
[50] Greg Ward, Erik Reinhard, and Paul Debevec. 2008. High dynamic range imaging & image-based lighting. In *ACM SIGGRAPH 2008 classes*. 1–137.
[51] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*. 418–434.
[52] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion models: A comprehensive survey of methods and applications. *Comput. Surveys* 56, 4 (2023), 1–39.
[53] Yizhou Yu, Paul Debevec, Jitendra Malik, and Tim Hawkins. 1999. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 215–224.
[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

[55] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. 2024. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems* 36 (2024).

[56] Yiqin Zhao, Sean Fanello, and Tian Guo. 2023. Multi-camera lighting estimation for photorealistic front-facing mobile augmented reality. In *Proceedings of the 24th International Workshop on Mobile Computing Systems and Applications*. 68–73.

[57] Yiqin Zhao and Tian Guo. 2020. Pointar: Efficient lighting estimation for mobile augmented reality. In *European Conference on Computer Vision*. Springer, 678–693.

[58] Yiqin Zhao and Tian Guo. 2021. Xihe: A 3D Vision-Based Lighting Estimation Framework for Mobile Augmented Reality. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys'21)*. 13 pages.

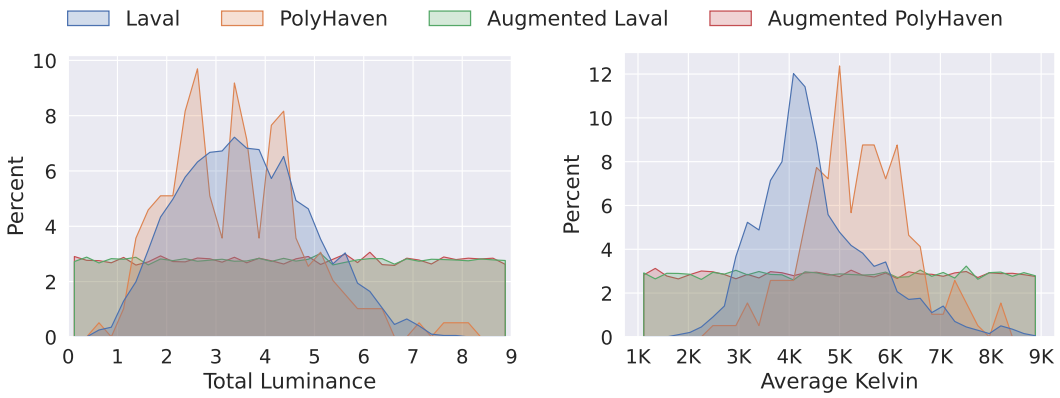
[59] Yiqin Zhao and Tian Guo. 2024. Demo: ARFlow: A Framework for Simplifying AR Experimentation Workflow. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications* (, San Diego, CA, USA,) *(HOTMOBILE '24)*. Association for Computing Machinery, New York, NY, USA, 154. https://doi.org/10.1145/3638550.3643617

[60] Yiqin Zhao, Chongyang Ma, Haibin Huang, and Tian Guo. 2022. LITAR: Visually Coherent Lighting for Mobile Augmented Reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–29.

[61] Yiqin Zhao, Sheng Wei, and Tian Guo. 2022. Privacy-preserving Reflection Rendering for Augmented Reality. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 2909–2918. https://doi.org/10.1145/3503161.3548386

[62] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. 2020. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 519–535.

[63] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 633–641.



(a) Light Intensity (b) Color Temperature

Fig. 14. Lighting condition measurement. We measure the distributions of light intensity (14a) and color temperature (14b) on environment maps collected from the Laval indoor dataset and PolyHaven and their augmented variants. Compared to the original data, our augmentation technique effectively creates greater diversity in the lighting features of the dataset.

A LIGHTING PROPERTY MEASUREMENT DETAILS

We describe additional details for the environmental lighting conditions measurement study described in §3. Our measurement focuses on two properties of environmental lighting conditions: *light intensity* and *color temperature*, representing an environment's overall brightness and color appearances. We choose these two lighting condition properties because of their significant impact on virtual object rendering of all material kinds [60] We select two standard lighting estimation data sources: the *Laval dataset* [19], an academic open research dataset, and the *PolyHaven* website [1], a royalty-free HDR environment map data website. In total, we collected 2235 and 196 data items from the Laval dataset and PolyHaven, respectively. Following [35, 48], we perform a standard data preprocessing procedure of a color correction process with gamma correction with $\gamma = 2.4$ and setting the 99th percentile of pixel intensity to 0.9.

Next, we calculate the properties of the lighting condition. We calculate the *light intensity* as the total luminance of a given HDR environment map image. To do so, we first calculate the individual pixel luminance l, which converts the original environment map image from RGB color space to the CIE XYZ color space [26] and then derives the pixel luminance component. We derive the total image luminance by summing the pixel luminance l weighted by their differential solid angles $\mathrm{d}\omega_i$ throughout the HDR environment map image:

$$L = \sum_{i=1}^{N} (0.212671R_i + 0.71516G_i + 0.072169B_i)\mathrm{d}\omega_i \tag{2}$$

where i represents the i-th pixel in the environment map with N pixels. For color temperature value extraction, we first calculate the average pixel RGB value from each environment map and then calculate its correlated color temperature using a recent method [33]. The calculated environment map image color temperature values are given in Kelvin. In Figure 14, we visualize the measured lighting condition distributions. While both datasets contain a wide selection of lighting conditions, the data distribution shows several biases. Predominantly, neutral lighting conditions of light intensity and color temperatures are seen in two datasets. For light intensity, more low-light environments than bright-lighting conditions data items are seen in the Laval dataset. As for the color temperature distribution, a significant part of the PolyHaven data items lies in the cool color range. These data distribution biases have also led to biased training and evaluation of lighting estimation systems.

We also evaluated the augmented versions of the Laval and PolyHaven datasets generated using our data augmentation technique (§B.3). The results demonstrate that our augmentation method successfully produces new dataset variants with more diverse and evenly distributed lighting characteristics. These augmented datasets are well-suited for assessing the robustness of lighting estimation systems under a wide range of lighting conditions.

B DATASET GENERATION DETAILS

Below, we describe how we construct the training datasets for the LDR environment map completion and high-intensity map estimation tasks, from three existing open-source datasets.

B.1 LDR Environment Map Completion Data

We first collect a large set of LDR environment map images from two large LDR indoor environment map sources: the Matterport3D dataset [9] and Structured3D datasets [62]. The Matterport3D dataset provides a large set of real-world captured environment maps, and the Structured3D dataset provides a large set of synthetic environment map images with photorealistic visuals. The combined training dataset consists of 29461 data items, *10X* larger than the Laval dataset [19].

Next, we mask the LDR environment map images at random angles to generate AR camera observation images. The masks are generated using the pin-hole camera model to simulate the real-world partial environment observations. We also combine multiple image masks to simulate multi-view environment observations. Specifically, the number of views is randomly chosen from 1 to 5, and the camera's horizontal FoV is randomly chosen between 60 and 120 degrees. For environment semantics, we use a pre-trained semantic map estimation model [51] to estimate semantic maps directly from the collected LDR environment maps. We also use the LDR image mask to generate masked semantic maps representing the environment's semantic information received by AR devices. For the ambient lighting condition prompt generation, we create the ambient light property labels using the masked partial environment map image pixel values and our defined system threshold values in §4.2.1.

B.2 High-intensity Map Estimation Data

Our novel learning objective enables the high-intensity map estimation model to leverage its pre-trained knowledge on LDR images, allowing it to be effectively fine-tuned using a limited amount of HDR environment map data. This capability is crucial for training with the small yet high-quality HDR dataset provided by the Laval dataset. Since the original Laval dataset contains HDR environment map images, we applied Equation 1 to convert them into 1,489 paired samples of LDR environment maps and their corresponding high-intensity maps. Specifically, we first clamp the raw HDR environment map to the range $[1, +\infty]$ to isolate pixel intensities beyond the LDR range, and then rescale the values to $[0, +\infty]$ to obtain the high-intensity map I_i. Next, we convert the range of the raw high-intensity $[0, +\infty]$ to the standard LDR pixel range in $[0, 1]$. Note that although this transformation is not lossless, our evaluation suggests minimal impact on the overall lighting estimation. In addition to the generated image pair data, we add the text prompt **P2** during training.

B.3 Robustness Evaluation Data Generation

Evaluating the robustness of lighting estimation systems is particularly challenging because it requires controlled lighting condition changes. To avoid costly real-time data capturing, we propose an image editing-based method that creates variants of standard lighting estimation testing datasets to represent diverse lighting conditions while maintaining the original environment map visual context. Specifically, our method includes the following three steps. First, we generate a set of edited variants of the Laval dataset by applying a uniform scaling term s to all environment map images. Then, we measure the total light intensity and average color temperatures using the measurement method introduced in §A. Finally, we uniformly sample from the generated data to

Fig. 15. Examples of color appearance refinement. We use examples of extreme lighting conditions, extremely cool temperature (row 1) and extremely warm temperature (row 2), to show how our color refinement algorithm can improve the estimation result color accuracy. Compared to the original LDR completion result, our refined environment map color is closer to the original full environment map (marked as Reference). Columns 5 and 6 show additional results to visualize the effects of median blur filtering and patch splitting on the refinement results. Particularly, visual artifacts can be observed on the image regions between the transitioning edges of the observed and unobserved environments.

ensure equal representation of each edited environment map variant across different light intensity and color temperature ranges. The scaling term s are selected from $[0.25, 4]$ using a step of 0.125. For the light-intensity editing, the scaling term is uniformly applied to all three color channels, while only red and blue channels are scaled with s and $1/s$ for color temperature editing [2]. Figure 8 shows examples of the edited environment maps. We created a set of augmented Laval indoor datasets using this method, with edited average light intensities and color temperatures in the ranges of $[-20\%, 20\%]$ and $[-50\%, 50\%]$. In total, we generated approximately 60,000 data items from the 290 original environment maps in the test split of the Laval dataset.

C MODEL TRAINING DETAILS

Our generative lighting estimation pipeline consists of three ControlNet [54] models, two for LDR environment map completion with RGB and semantics context, and one for high-intensity map estimation. We train all the models by fine-tuning the pre-trained `StableDiffusion 1.5 inpaint` [39] checkpoint. We apply several data augmentation techniques to the previously generated datasets to train the models and improve generalization in different lighting conditions and environmental contexts. We augment the lighting condition by applying a random scaling s similar to the scaling term used in the robustness test data generation. This augmentation is only applied to LDR environment map completion training. For environment context augmentation, we apply horizontal rotations to environment map images. This augmentation is used for both LDR completion and high-intensity map estimation training. Our generative lighting estimation models are trained on affordable commercial PC hardware, specifically, a high-end workstation PC with an I9-13900K CPU and an RTX4090 GPU. The LDR completion model requires an average of 12 hours of training, while the high-intensity map estimation model, due to its smaller training dataset, only requires an average of 4 hours.

D ON-DEVICE REAL-TIME REFINEMENT DETAILS

In this section, we show additional results and technical details of the color appearance refinement technique in CLEAR. In the refinement process, our technique seeks to create a color refinement matrix of pixel color multipliers for the completed LDR environment maps. The color refinement matrix consists of two types of multiplier values: *(i)* the global color multiplier and *(ii)* the local color multiplier. The global color multiplier adjusts the overall image colors of completed environment map images, while the local color multiplier adjusts fine-grained colors on the observed environment map regions. The two color refinement terms are combined into a color refinement matrix and applied to the completed LDR environment maps.

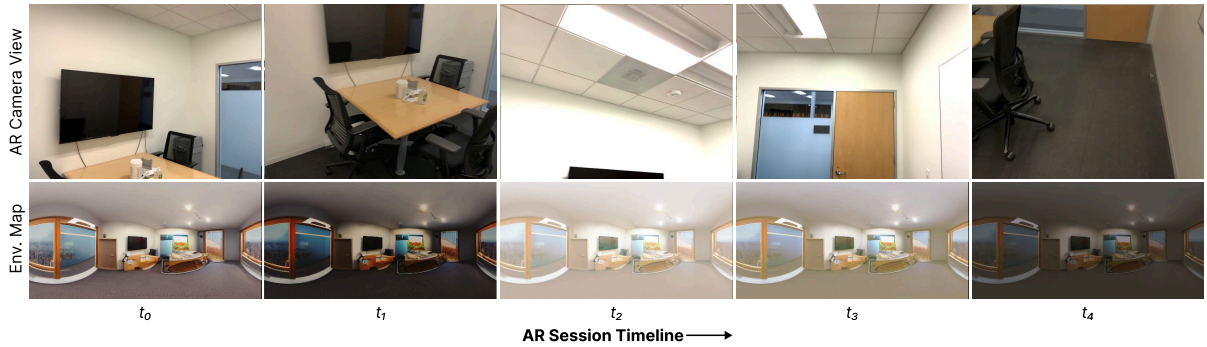


Fig. 16. Examples of environment map refinement over an AR session. We demonstrate the effectiveness of our refinement step using selected frames from an AR session in the ARKitScene [5] dataset. In this example, the initial environment map is estimated by CLEAR at time t_0, and subsequently refined at t_1 through t_4 using the corresponding AR camera images. As the AR session progresses, the ambient lighting intensity varies. The color refinement step in CLEAR effectively adapts the estimated environment maps to reflect these changing lighting conditions. Note that the original AR camera video has been lightly edited to enhance intensity variations for improved visual clarity.

Our technique calculates the global color multiplier by deriving per-channel multipliers as the ratio between average colors in the estimated and observed environment map images. The local multipliers are calculated by first splitting the estimated environment map images and the partial environment observation image into NxM patches. Then, our technique calculates local multipliers as the average color ratios for each patch. A 3x3 median blur filter is used to address color smoothness between the observation edges. Empirically, we found 8x8 is the best image patch size.

We show the effectiveness of our refinement technique via two visual quality evaluation. As shown in Figure 15, our refinement technique can adjust environment map colors even in very challenging lighting conditions. Furthermore, our design choices of median blurring and patch splitting are useful to smooth out the artifacts on the edges. Figure 16 illustrates the effectiveness of this refinement process over time during an AR session. The example video is sourced from the ARKitScene [5] dataset, with minor edits applied to enhance lighting intensity variations. CLEAR first estimates the environment map at frame t_0, and subsequently refines it at frames t_1 through t_4 using the corresponding AR camera images. The refined environment maps clearly reflect the observed changes in ambient lighting intensity. This refinement step is essential for maintaining coherence between the physical environment and the rendered virtual content as lighting conditions change over time. In practice, it is especially effective for adapting to minor lighting changes that do not involve significant alterations to scene geometry or object placement.