# Vision-Language Model-Based Evaluation of Shadows & Lighting in AR Scenes

**Yuhe (Alice) Hu, Lin Duan, Maria Gorlatova**
Intelligent Interactive Internet of Things (I^3T) Lab

## Background

### Motivation

- Realistic visual presentation is essential for delivering immersive augmented reality (AR) experiences
- Assessing realism traditionally depends on human-subject studies, which are time-consuming, costly and not scalable [1,2]

### Vision Language Models (VLMs)

- VLMs offer semantic reasoning that more closely reflects human perception than pixel-based methods [3]
- This study examines how VLMs assess shadow and lighting realism in AR scenes compared to traditional evaluation methods

***Figure 1:*** Examples of Visual Realism Artifacts in AR Scenes
Left: shadow realism example. Right: light direction coherence example

## Data Collection

Two datasets were curated for evaluating shadow realism and light direction coherence respectively

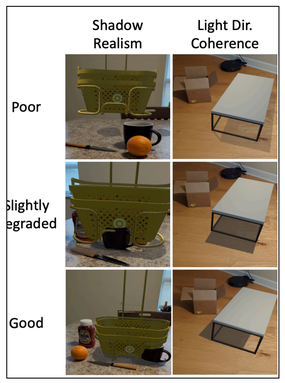


***Figure 2:*** Sample Images from Collected Dataset

**Sources:**

- AR scenes: Apple Vision Pro, Android smartphones
- VR scenes: VIDTI dataset [3]

**Image Design**:

- One virtual object per image
- Real or simulated background environments chosen
- Scenarios include both plausible and implausible shadow and lighting

## Experimental Design

### Baseline Evaluation

| | Model Type | Training Context | Output |
|---|---|---|---|
| Shadow Realism | CNN-based binary classifier [4] | Synthetic shadow datasets | Realism **confidence score per object** |
| Light Direction Coherence | CNN-based multi-class classifier [5] | VR datasets with known lighting labels | **Categorical label**: front, back, left, right |

### Vision Language Model Setup (GPT - 4o)

- Model Input:
  - AR Image: sirtual object placed into the scene
  - Background Image: same scene, without the virtual object
  - Natural Language Prompt
- Prompts were structured to reduce bias and ensure consistent context for evaluation across samples

| | Example Prompt |
|---|---|
| Shadow Realism | "In this image, all objects are virtual and placed in a real-world environment. Could you please rate each virtual object's shadow quality on a scale from 1 to 5?" |
| Light Direction Coherence | "Can you identify the direction of the main light source in the image?" |

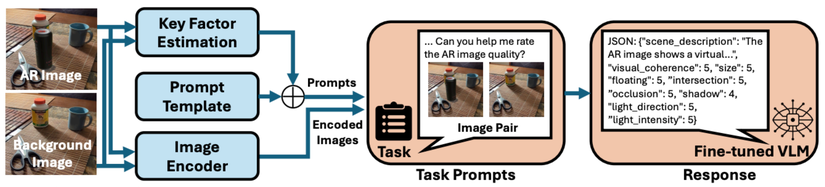


***Figure 3:*** Experimental Design Workflow

## Evaluation

### Classification Accuracy

- For each image sample, the model's output (e.g., a score, label, or judgment) is evaluated against known ground truth annotations or human agreement labels
- Direct performance comparison between baseline CNNs and VLMs

## Results

### Factor Accuracy Criteria:

- Shadow Realism: accuracy based on correctly ranking the virtual shadow as less realistic
- Light Direction: accuracy based on correct classification of the scene's primary light direction

### Key Findings

- GPT-4o outperforms baseline models on both factors
- VLM excels in light direction (80% accuracy compared to 47% for the CNN-based classifier_
- Improvement in shadow realism detection was modest (55% vs. 50%)

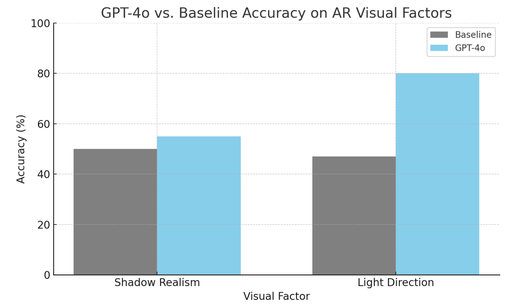


***Figure 4:*** Accuracy Comparison for VLM vs. Baseline

## Future Work

- Expand evaluation to additional perceptual factors in AR realism, such as floating plausibility and occlusion
- Explore integration of depth maps, object segmentation, or lighting estimation tools to improve shadow realism detection
- Investigate temporal consistency by applying VLM evaluation to AR video sequences

## References

[1] Hu, Y., Duan, L., & Gorlatova, M. Bridging Human Perception and Automated Evaluation: Vision-Language Model-Based Visual Quality Assessment of AR-Generated Scenes. ISMAR 2025 (submitted).
[2] Barkowsky, M., Le Callet, P., Tourancheau, B., et al. Subjective Quality Assessment of Heterogeneous AR Content: Limitations and Challenges. IEEE Trans. on Image Processing, 2020.
[3] Chen, T., Mishra, P., & Girdhar, R. Shadows Don't Lie: Shadow-Based Image Forgery Detection. CVPR 2021.
[4] Yan, T., Li, J., & Wang, Z. Light Direction Estimation for Photorealistic Rendering. ICCV 2019.
[5] Lee, S., et al. VIDTI: A Virtual Reality Dataset with Lighting and Depth for Visual Inference. NeurIPS Datasets Track, 2022.

Duke UNIVERSITY