Article Type: Regular

Probing the Augmented Reality Scene Analysis Capabilities of Large Multimodal Models: Toward Reliable Real-Time Assessment Solutions

Lin Duan, Elias Rotondo, Yanming Xiu, Sangjun Eom, Ryan Chen, Conrad Li, Yuhe Hu, and Maria Gorlatova,
Duke University, Durham, NC, 27708, USA

Abstract—Augmented Reality (AR) is transforming everyday experiences across domains like education, entertainment, and healthcare. As AR technologies become increasingly widespread, human-aligned and scalable AR quality evaluation is critical for optimizing immersive user experiences. This paper investigates the potential of Large Multimodal Models (LMMs) for automating AR quality assessment. We curate DiverseAR+, a new dataset of 1,405 scenes collected from diverse sources and environments, and use it to evaluate four commercial LMMs. Our results demonstrate that LMMs can perceive, describe, and judge AR content with promising accuracy. To deliver real-time, robust, and scalable AR quality evaluation under diverse network conditions, we propose a hybrid cloud-edge architecture that combines LMMs with traditional machine learning models. We argue that task-tailored AR-LMM systems can make AR experience assessment more efficient, adaptive, and user-centered.
Keywords: *Scene understanding, real-time system architecture, mixed / augmented reality*

How might a technology enthusiast determine whether a new sofa will match the design and layout of their living room? Augmented Reality (AR) offers a promising solution: by overlaying a virtual sofa onto the physical space, users can preview size, style, and fit without any heavy lifting. However, if the sofa casts unrealistic shadows, intersects with nearby furniture, or floats above the floor, the illusion falls apart. In such cases, unsuccessful AR content integration not only breaks immersion but also degrades the overall user experience.

As AR technologies become increasingly widespread, improving user experience and accelerating the development cycle remain critical. A key enabler for both objectives is effective human-centered AR quality evaluation, which ensures that the rendered experiences align with human perception and expectations. However, current assessment methods face significant limitations in scalability and alignment with subjective human experiences.

User studies have served as the gold standard for assessing AR scene quality due to their ability to reflect human-perceived experiences. While effective, these studies are labor-intensive and time-consuming, posing a bottleneck for the iterative design and deployment of AR applications. Although numerous automatic quality evaluation methods exist for conventional image processing tasks [1], they seldom capture the human-centric criteria essential for AR. Recent advances in Large Multimodal Models (LMMs) offer a promising alternative. Trained on massive textual and visual corpora, LMMs exhibit strong capabilities in visual understanding, natural language grounding, and complex scene interpretation. Their ability to describe and reason about visual content suggests potential for scalable, human-aligned AR quality assessment. Prior works [2], [3], [4] have demonstrated the promise of LMMs for image quality assessment, including foreground-background coherence, CT image

Digital Object Identifier 10.1109/XXX.0000.0000000

clarity, and general image distortions. Yet, automatic and human-centric quality assessment in AR remains largely unexplored, particularly for AR-specific perceptual factors such as object-placement coherence and object-size appropriateness, which are crucial for maintaining user immersion.

In this work, we explore the AR scene analysis capabilities of leading commercial LMMs, including GPT, Gemini, Claude, and DeepSeek. We empirically evaluate their performance using tailored prompts on DiverseAR+, our curated benchmark comprising 1,405 AR images, each exhibiting variations in one or more quality factors such as incorrect shadows, misaligned objects, or implausible intersections (e.g., instances where virtual objects penetrate real-world objects). Additionally, we analyze LMMs' performance in evaluating key perceptual factors and compare their results to traditional machine learning (ML) baselines we develop for this study. Based on our findings, we propose a hybrid cloud-edge architecture that integrates LMMs with traditional models to enable accurate, low-latency, and cost-aware AR quality assessment under diverse network conditions. Our contributions are summarized as follows:

- We create and release the DiverseAR+ dataset publicly on GitHub[1]. It contains 1,405 AR scenes captured from various platforms and scenarios. To the best of our knowledge, DiverseAR+ is the first dataset to cover a diverse range of user-perceived AR visual factor states.
- We evaluate four commercial LMMs on DiverseAR+, quantifying their strengths and limitations in AR scene analysis. The best-performing models achieve a true positive rate of 94.4% for AR perception and 86.0% for AR description, demonstrating the potential of LMMs for human-aligned AR quality evaluation.
- We propose a hybrid cloud-edge architecture that combines LMMs with traditional ML models to achieve accurate, low-latency, and cost-aware AR quality evaluation under unstable real-world network conditions.

Related Work

AR Content Quality Evaluation

AR quality evaluation traditionally relies on user studies, which are accurate but impractical for the timely deployment of AR systems. Recent ML approaches, including deep classifiers [5] and regressors, often target specific AR qualities, such as shadow realism or object size. While these methods capture high-level visual patterns, their task-specific designs require extensive labeled data and lack generalizability across platforms, scenes, and AR styles. This highlights the need for a scalable and adaptive alternative that requires less data, handles diverse scenarios, and supports real-time deployment.

LMMs for Perceptual Tasks

Trained on large-scale datasets, LMMs exhibit strong human-aligned scene interpretation capabilities [6]. Prior studies have applied these models to perceptual tasks such as scene understanding [7] and image quality assessment [2], [3]. However, most existing benchmarks assess LMMs on general-purpose datasets that do not account for the inherent complexity in AR scenes. As AR deployments demand robust semantic understanding for tasks such as content moderation and user safety validation, there is a growing need to evaluate how well LMMs perform under the contextual and perceptual constraints of AR environments.

Applications of LMMs in AR Contexts

LMMs are increasingly utilized in AR applications to guide virtual content generation and placement. Text-to-3D pipelines leverage LMMs to synthesize virtual assets from natural language prompts [8]. Similarly, scene-aware LMMs have been applied to identify semantically appropriate and physically plausible locations for virtual object placement [9]. However, these applications largely focus on what LMMs can create or decide during the content generation phase, while largely overlooking their potential in closing the user experience loop through perceptual feedback. Our work highlights this missing link and explores how LMMs can recognize and evaluate virtual objects within AR scenes.

DiverseAR+ Dataset

We introduce **DiverseAR+**, a benchmark meticulously curated to capture a broad range of AR scenarios and artifacts. The DiverseAR+ dataset is publicly available on GitHub[1].

Dataset composition. DiverseAR+ comprises *1,405* AR samples, extending an earlier limited collection of *321* AR images released in [10]. The dataset includes AR scenarios spanning both common and specialized environments, such as homes, labs, and medical offices. Each image shows virtual objects typical in AR applications, including stationary objects

[1] https://github.com/ARResearcher/DiverseARPlus

(e.g., chair, table), dynamic, user-manipulable objects (e.g., basketball, glove), and medical anatomical structures (e.g., eye, brain).

AR platforms. We collect AR samples from diverse sources and environments using custom-developed applications: *320* from Apple Vision Pro (AVP); *826* from Android phones (Galaxy S25 and Pixel 7); *135* from Microsoft HoloLens 2; and *124* from external sources (*29* public websites, *53* commercial apps, *42* prior studies).

Visual quality factors. The dataset captures a diverse range of visual quality states across three factors, recognized as key dimensions of photorealistic coherence [11], [12]: shadow realism, floating plausibility, and intersection coherence. All three factors are applicable for each visual sample collected.

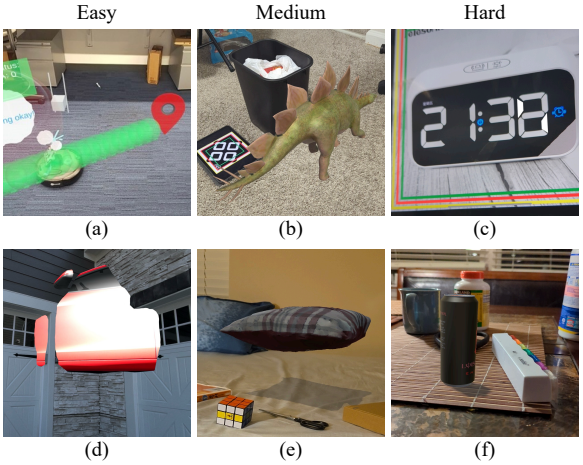

FIGURE 1. Cropped DiverseAR+ images showcasing the three complexity levels: (a) Robot's virtual traversal path; (b) Virtual dinosaur; (c) Virtual white bars; (d) Virtual red vehicle; (e) Virtual pillow; (f) Virtual beverage can.

AR Perception and Description Capability of LMMs

Evaluation Setup

We evaluate four widely used and readily accessible commercial LMMs to analyze their perceptual and descriptive capabilities: 1) **GPT-4o-2024-08-06**; 2) **Gemini-1.5-Pro-002**; 3) **Claude-3.5-Sonnet-20241022**; 4) **DeepSeek-VL2**. All models are tested on *DiverseAR+*.

To examine how LMMs handle AR scenes of varying difficulty, we classify DiverseAR+ images into three complexity tiers, as shown in Figure 1. The **Easy** class contains clearly identifiable virtual content (e.g., transparent or low-quality rendered objects). The **Medium** class features high-quality virtual objects that still break immersion due to conflicting expectations or interactions (e.g., objects that appear to defy gravity). Lastly, the **Hard** class blends digital and real content seamlessly, making virtual objects difficult to detect.

An AR researcher with four years of experience manually assigned complexity levels to all 1,405 images following the above complexity guidelines, resulting in 866 easy, 449 medium, and 90 hard samples. The scarcity of hard cases reflects challenges in achieving immersive AR.

To evaluate virtual content identification and context understanding in AR scenes, we measure the LMMs' performance using true positive rates, TPR_P and TPR_D:

$$TPR_P = \frac{N_P}{N_{VC}} \quad (1) \qquad TPR_D = \frac{N_D}{N_{VC}} \quad (2)$$

where N_P and N_D denote the number of samples correctly perceived and described by the LMM, respectively. N_{VC} is the total number of samples containing virtual contents. TPR_P expresses the proportion of images successfully identified as containing AR objects, regardless of the accompanying description accuracy. Alternatively, TPR_D quantifies the percentage of AR scenes in which the model not only correctly detects the presence of virtual content but also provides a human-like description appropriate to the scene.

Prompt Design

To evaluate the perceptual and comprehension abilities of LMMs, we design two prompts and use them as-is:

General image captioning prompt. *"Can you explain what is happening in this image?"* This prompt evaluates the model's ability to distinguish and recount AR content without task-related guidance.

Task-aware image captioning prompt. *"Does this image include virtual content superimposed on the real world? If yes, could you list all the virtual elements present and explain why each one is considered virtual?"* This prompt assesses the model's ability to caption images when it is explicitly primed about the presence of digital elements. By comparing responses across the two prompts, we investigate whether this task-aware prompt improves AR perception and description.

Performance Analysis

We assess the perception and description capabilities of four LMMs with both the general prompt (G) and the

task-aware prompt (T) on the DiverseAR+ dataset and its three scene-complexity levels.

Perception performance. Figure 2 shows the Perception True Positive Rate (TPR_P). When queried with prompt T, Gemini achieves the highest overall TPR_P at 94.4%, closely followed by GPT-4o at 94.3%. Performance then drops to 85.8% for DeepSeek, and 49.6% for Claude. Thus, while Gemini and GPT-4o exhibit comparable, near-ceiling sensitivity to AR objects, DeepSeek is merely adequate and Claude lags substantially, making it unsuitable for accuracy-critical applications without additional tuning. This ordering indicates that LMMs' sensitivity to AR content is uneven and strongly model-dependent, likely reflecting differences in pre-training data and model architecture.

Explicitly mentioning AR cues in the prompt markedly enhances perception: across models, TPR_P rises by 37.3%–78.3% when moving from prompt G to prompt T, indicating the importance of crafting clear, targeted prompts. Notably, DeepSeek shows the most dramatic improvement, with TPR_P jumping from 7.5% under prompt G to 85.8% under prompt T. Based on the analysis of its raw responses, DeepSeek shows a tendency to emphasize spatial relationships and basic attributes (e.g., color and shape), making it less sensitive to contextual cues such as unnatural shadows or implausible placement. By introducing those AR-specific cues, prompt T redirects the model's attention to the missing context and thus unlocks its detection capability.

Furthermore, perception performance declines consistently as scenes become more complex. From the easy to the hard level, TPR_P drops by 3.6%–44.8% under prompt G and 6.1%–25.7% under prompt T. This finding suggests that LMMs are proficient at detecting overtly digital content, but struggle with improved virtual object integration (e.g., enhanced rendering, appropriate lighting, and realistic shadows). Meanwhile, for DeepSeek under prompt T, TPR_P of medium level exceeds the hard level score by only a negligible margin of 0.5%. The medium-to-hard gap in the other models is similarly small (G: 0.2%–17.9%, T: 0.5%–16%). These limited differences likely arise due to several scenes being borderline between the medium and hard levels, as the annotator noted.

Description performance. Figure 2 also demonstrates the Description True Positive Rate (TPR_D), which reflects how well a caption aligns with human scene understanding. We observe three trends that closely parallel the perception results: 1) the prompt T consistently boosts performance, increasing TPR_D by 15.0%–41.0% across the full dataset, showing that explicit AR cues improve caption quality; 2) description accuracy declines with scene complexity: from easy to hard scenes, TPR_D drops by 5.3%–47.5% with prompt G and 9.4%–56.1% with prompt T; 3) GPT and Gemini outperform Claude and DeepSeek by 22.9%–65.6% under both prompts, reflecting their superior contextual-reasoning abilities.

Description accuracy is uniformly lower than perception accuracy, declining by 1.6%–65.4% on the full dataset. This is expected, as captioning demands both accurate AR object perception and human-aligned narration. Furthermore, even with task-aware guidance, DeepSeek's TPR_D improves by only 16.0%, far less than its perception gain, demonstrating the added difficulty of aligning textual descriptions with human perspective.

Summary. Overall, these results indicate that modern LMMs can perceive and describe AR content in a human-aligned fashion; however, their effectiveness depends critically on both model selection and prompt design. By matching a model's strengths (e.g., DeepSeek for spatial relations, GPT or Gemini for contextual reasoning) to task requirements and supplying clear, explicit instructions, researchers can substantially improve the accuracy of AR quality assessments.

Pilot Study: AR Visual Quality Assessment Capability of LMMs

Recent studies [6] demonstrate that LMMs can achieve holistic, human-aligned understanding of complex scenes. Our empirical results show that LMMs can perceive and describe AR content, suggesting their potential for human-centric AR quality assessment. However, a follow-up question arises: *Can LMMs detect AR visual artifacts, such as unnatural intersections, implausible floating, and incoherent shadows, and do they surpass traditional evaluation methods?*

Intersection Coherence

Improper object placement often causes partial "penetration" into real surfaces, which we define as *intersection*. We evaluate this using a dataset of 20 background-AR image pairs collected by AVP. Conventional intersection detection methods used in virtual-physical collision detection compare the real scene's depth map with the rendered object's depth buffer. However, this requires privileged depth data that AR developers usually cannot access. As a baseline without such data, we apply traditional CV techniques based on the observation that severe intersections visually split the virtual object, as demonstrated in Figure 1(d). Specifically, we subtract each background

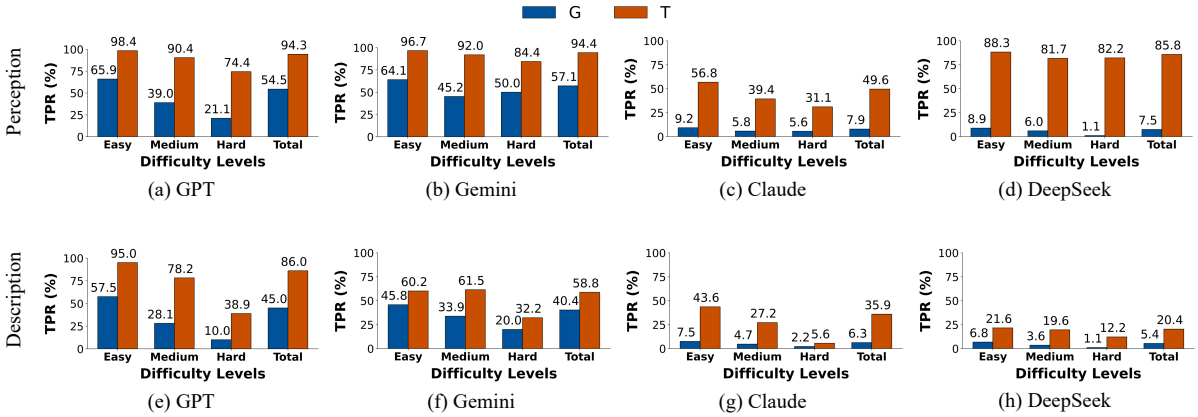

FIGURE 2. The TPR_P perception results (top, [a]-[d]) and the TPR_D description results (bottom, [e]-[h]) of the four LMMs across varying complexity levels using G (general) and T (task-aware) image captioning prompts. Although LMMs' perception performance tends to decrease as scene complexity increases, models such as GPT and Gemini demonstrate their abilities to perceive AR content across diverse settings. Regarding LMMs' description capabilities, GPT overall outperforms all competitors (see [e] "Total"), particularly when prompted with task-aware language.

image from its corresponding AR image to obtain the virtual object's binary mask, defined as:

$$M(p) = \begin{cases} 1, & \text{if the object contains } p = (x, y), \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where p denotes a pixel coordinate in the AR image. Then, we use connected component analysis with OpenCV to count distinct regions in M. If the object is segmented into multiple regions, the sample is flagged as a non-coherent intersection. Alternatively, we prompt the GPT-4o with: *"When comparing the raw image to the AR image, does the virtual object appear to sink into the vertical plane, the horizontal plane, or is it being penetrated by the real object?"*

Our results show that the traditional method achieves 50% accuracy, while GPT-4o attains *85%*. This indicates that pre-training on the large-scale training corpora enables the LMM to infer complete object shapes from partial fragments and contextual cues to detect intersections with real-world surfaces even when artifacts are subtle, whereas the baseline approach often overlooks such cases.

Floating Plausibility

Some virtual objects appear to hover above real surfaces. To detect floating artifacts, we used a dataset of 22 background and AR image pairs collected by AVP. As with intersection detection, most existing methods for floating artifact detection in plane anchoring rely on privileged depth buffers that developers are frequently not privy to. Consequently, we establish a baseline that relies on classical ML and CV techniques: we employ a monocular depth estimator to predict a depth map for the virtual object and its surrounding background, then compare the relative depth differences at the object's base to flag implausible floating. Specifically, we first extract the virtual object's binary mask M. Using this mask, we focus on the region around the virtual object and define $H = 30$ (an empirically determined value) as the vertical extent, in pixels, measured from the object's lower boundary. The virtual object's bottom region is obtained by:

$$P_{\text{obj}} = \{(x, y) \mid M(x, y) = 1\}, \tag{4}$$

$$y_{\max} = \max_{(x,y)\in P_{\text{obj}}} y, \tag{5}$$

$$P^H_{\text{obj}} = \{(x, y) \in P_{\text{obj}} \mid y \geq y_{\max} - H\}, \tag{6}$$

where P_{obj} is the set of all object pixels, $y_{\max}$ denotes the the bottom-most row index of the object, and P^H_{obj} is the set of pixels in the bottom H rows of the virtual object. We then apply ZoeDepth [13] on the AR image to generate a pixel-wise depth map and compute the average depth D_{obj} in the bottom region:

$$D_{\text{obj}} = \frac{1}{|P^H_{\text{obj}}|} \sum_{p\in P^H_{\text{obj}}} d(p), \tag{7}$$

where $d(p)$ is the depth at pixel p. We then extract P^H_{bg}, which is the set of pixels in the H rows immediately below the virtual object's lower boundary in the AR image, and compute its average depth D_{bg}:

$$D_{\text{bg}} = \frac{1}{|P^H_{\text{bg}}|} \sum_{p\in P^H_{\text{bg}}} d(p). \tag{8}$$

We define the depth difference as:

$$\Delta_F = |D_{\text{obj}} - D_{\text{bg}}|. \tag{9}$$

If Δ_F exceeds an empirically determined threshold T_F of 10 cm, the sample is flagged as floating. Alternatively, we prompt the GPT-4o:*"In this image, every object is a virtual overlay integrated into a real-world scene. Could you evaluate each virtual object's floating artifact level on a scale from 1 to 5?"*

Our results show that the traditional method achieves 77% accuracy, whereas GPT-4o attains *95%* accuracy. This disparity indicates that LMMs, by leveraging large-scale training data, more effectively infer object height and identify implausible floating. In contrast, conventional ML techniques are constrained by the accuracy of their depth estimates.

Shadow Realism

Non-realistic shadows often exhibit mismatched strength, geometry, or direction relative to real-world lighting. To detect such artifacts, we conducted experiments on 20 AR images collected by Android phones, each containing real objects and one virtual object. As a baseline, we use a widely adopted pre-trained shadow-realism classifier [5], which produces confidence scores indicating each object's shadow realism, allowing objects to be ranked within each image. With the LMM, we observe a bias: when the model perceives an object as virtual, it tends to rate that object's shadow realism poorly. To ensure fair comparisons, we standardize the rating space by treating all objects as "virtual". Subsequently, we prompt the GPT-4o with: *"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?"*

A prediction is considered "correct" if the truly virtual object receives a lower shadow quality rating than all real objects. GPT-4o achieves a detection accuracy of 55%, slightly outperforming the pre-trained shadow-realism model, which achieves 50%. Given that multiple objects appear in each image, this accuracy is not a trivial random baseline, indicating that the LMM can perceive lighting conditions, such as intensity and direction, to assess shadow realism.

Summary. Our pilot study reveals that GPT-4o significantly outperforms traditional approaches in the challenging task of intersection detection, while both methods still struggle with shadow analysis. Additionally, although GPT-4o demonstrates high accuracy in floating detection, the traditional method still achieves fair performance. These findings highlight the potential of LMMs for AR visual quality analysis and motivate further exploration into their use for human-aligned quality assessment.

Hybrid Cloud-Edge AR Quality Evaluation

Our preliminary results demonstrate that commercial LMMs can analyze nuanced AR scenes from a human-centric perspective, outperforming lightweight CV/ML methods. However, LMMs suffer from high latency and monetary cost, whereas traditional approaches support more rapid estimation. To address these trade-offs, we propose a hybrid AR quality evaluation framework that integrates both approaches to achieve accurate, responsive, and scalable assessments.

Design Objectives

Our proposed hybrid framework is guided by the following design objectives:

Human-perspective alignment. Since AR quality estimation aims to enhance immersion, the resulting quality estimations should align with users' perceptions of key visual factors, such as shadow realism, intersection coherence, and floating plausibility, which are widely regarded as essential to visual quality.

Low latency. To support downstream tasks such as real-time rendering quality adjustments, the system must deliver evaluations with low latency.

Reliability reporting. Downstream quality-improvement tasks need to understand the trustworthiness of each quality result. Each evaluation should therefore attach an explicit confidence metric, enabling subsequent components to prioritize high-reliability results and treat uncertain ones with appropriate caution.

Network variability robustness. Variations in network conditions can delay responses from cloud-based LMMs and disrupt real-time operations. Remote-rendering systems tackle this problem through client-side reprojection or speculative rendering when cloud frames arrive late or are lost [14], [15]. Inspired by these strategies, our quality evaluation system should degrade gracefully under poor connectivity and recover as network conditions improve.

Cost efficiency. Given the usage-based pricing of commercial LMMs and the performance-latency trade-offs identified in inference offloading and edge-first inference studies [16], [17], we aim to minimize cloud queries by relying on edge models when their predictions are sufficiently accurate.

System Framework

Building on these objectives, we propose an AR quality evaluation system composed of four components, as illustrated in Figure 3.

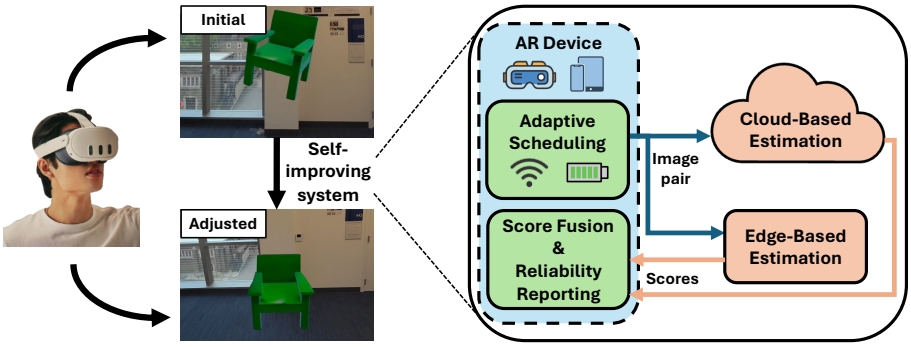

FIGURE 3. Self-improving AR furniture placement via our hybrid cloud-edge quality evaluation framework. The user first sees a hovering virtual sofa (top). Our proposed system evaluates the scene and generates corrective instructions to adjust the virtual sofa's position, lowering it to rest on the floor (bottom).

Cloud-based estimation. We will employ a cloud-hosted LMM that takes an ⟨image, structured prompt⟩ pair as input and returns quality scores for each factor. We will also explore prompt-engineering strategies (e.g., chain-of-thought prompting and few-shot exemplars) to improve accuracy.

Edge-based estimation. To enable real-time responsiveness and reduce cloud dependence, we will deploy lightweight quality estimators on the edge. We envision three approaches: 1) Customized classifiers that predict quality scores end-to-end; 2) Interpretable rule-based proxies that approximate visual quality through domain heuristics; and 3) Open-sourced on-device LMMs fine-tuned on large-scale AR quality datasets to deliver quality scores with acceptable latency. Implementing this module involves: constructing human-aligned quality datasets, applying AR-specific data augmentation, defining suitable rules, and tuning parameters to achieve high accuracy.

Adaptive scheduling. To balance cost and latency while remaining resilient to network variability, we propose a runtime scheduler that monitors latency and bandwidth and decides which frames should be evaluated on the cloud or on the edge. The scheduler supports two modes:

- *Cloud-preferred mode:* Used when connectivity is stable or high-reliability estimates are required. The scheduler will upload frames to the cloud while using the edge estimator as a fallback.
- *Edge-preferred mode:* Used under constrained bandwidth, strict latency budgets, or cost limits. The system will score most frames locally and send a key frame every k seconds to the cloud for recalibration, following key-frame offloading in SLAM [18].

Mode selection can be threshold-based or driven by a reinforcement-learning (RL) policy that optimizes cost, latency, and quality for long-term performance.

Score fusion and reliability reporting. To combine sparse, human-aligned scores from the cloud with rapid but coarser scores from the edge, we propose to fuse the two predictions into a single rating. As the cloud estimate of a previous frame becomes outdated, its influence is gradually reduced to prevent stale information from dominating the final score. This adaptive weighting can be RL-driven, rule-based, or a learned fusion model integrating scores with metadata such as network round-trip time and past errors to produce the final rating.

Since scores without trust estimates may mislead downstream modules, reliability reporting is essential. Commercial LMMs and our rule-based edge estimators typically do not expose explicit or standardized confidence metrics for their perceptual outputs. We therefore propose a three-tier scheme—*high*, *medium*, and *low*—based on the disagreement between the most recent cloud and edge scores: zero difference yields high reliability, a difference of one yields medium, and greater than one yields low. Future refinements could employ regression models trained on past errors and contextual metadata.

Summary. Our proposed hybrid system integrates the high-level perceptual understanding capability of LMMs and the immediacy of edge models. To demonstrate our design's utility for downstream tasks, we have built a working prototype of a self-improving AR placement app: a Meta Quest 3 captures and streams AR frames and prompts to a cloud server, receives quality assessments and actionable adjustments from GPT-o3, and applies the resulting content-position updates on-device. In preliminary (unoptimized) mea-

surements, the end-to-end latency was around 50 seconds, demonstrating feasibility while highlighting substantial optimization headroom. By adaptively selecting computation paths and integrating results through reliability-aware fusion, the proposed system is positioned to deliver efficient, robust, and perceptually human-aligned AR quality assessments across diverse deployment environments.

Future Work

Although LMMs demonstrate competitive performance in describing virtual content and assessing certain visual artifacts, their assessments are sensitive to pretraining data, architectural choices, and prompt design. As these factors shape each model's task-specific strengths and influence response accuracy and efficiency, some misalignment with human perception remains unavoidable. In this study, we employed off-the-shelf commercial models without any adaptation. While this approach is training-data-independent, it may also limit the model's ability to capture subtle perceptual cues that matter to humans in AR contexts.

To address this gap, we are actively exploring strategies to better align model outputs with human judgments. In our ongoing work, we incorporate human-annotated scores on visual factors to fine-tune the models. We are also developing real-time feedback loops that calibrate model predictions with user responses. These efforts aim to bridge the semantic gap between LMMs and human perception, ultimately enabling LMMs to serve as more trustworthy evaluators in AR systems.

Conclusion

In this work, we evaluate the AR scene understanding capabilities of LMMs to perceive, describe, and assess virtual content using our DiverseAR+ dataset, a comprehensive benchmark of 1,405 AR images encompassing an array of scene complexities and content of various quality levels. Our results demonstrate that LMMs can identify and characterize AR objects and motivate further studies into using such models for user-centered quality assessment. Building on our analysis, we propose a hybrid cloud-edge framework for accurate, low-latency, and cost-aware AR quality evaluation. In addition, we outline ongoing efforts toward user-centric AR quality assessment to advance AR development.

ACKNOWLEDGMENTS

This work was supported in part by NSF grants CSR-2312760, CNS-2112562, and IIS-2231975, NSF CAREER Award IIS-2046072, NSF NAIAD Award 2332744, a CISCO Research Award, a Meta Research Award, Defense Advanced Research Projects Agency Young Faculty Award HR0011-24-1-0001, and the Army Research Laboratory under Cooperative Agreement Number W911NF-23-2-0224. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency, the Army Research Laboratory, or the U.S. Government. This paper has been approved for public release; distribution is unlimited. No official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein.

REFERENCES

1. A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proceedings of IEEE ICPR*, 2010.
2. Z. Xu, H. Duan, G. Ma, L. Yang, J. Wang, Q. Wu, X. Min, G. Zhai, and P. L. Callet, "Harmony-IQA: Pioneering benchmark and model for image harmonization quality assessment," *arXiv preprint arXiv:2501.01116*, 2025.
3. Z. Chen, B. Hu, C. Niu, T. Chen, Y. Li, H. Shan, and G. Wang, "IQAGPT: Image quality assessment with vision-language and ChatGPT models," *Visual Computing for Industry, Biomedicine, and Art*, vol. 7, no. 20, 2024.
4. Y. Xiu, T. Scargill, and M. Gorlatova, "ViDDAR: Vision language model-based task-detrimental content detection for augmented reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 05, 2025.
5. A. Sarkar, H. Mai, A. Mahapatra, S. Lazebnik, D. A. Forsyth, and A. Bhattad, "Shadows don't lie and lines can't bend! Generative models don't know projective geometry... for now," in *Proceedings of IEEE/CVF CVPR*, 2024.
6. C. R. Jones, B. Bergen, and S. Trott, "Do multimodal large language models and humans ground language similarly?" *Computational Linguistics*, vol. 50, no. 4, pp. 1415–1440, 2024.
7. D. Zheng, S. Huang, and L. Wang, "Video-3D LLM: Learning position-aware video representation for 3D scene understanding," in *Proceedings of the IEEE/CVF CVPR*, 2025.

8. D. Tochilkin, D. Pankratz, Z. Liu, Z. Huang, A. Letts, Y. Li, D. Liang, C. Laforte, V. Jampani, and Y.-P. Cao, "TripoSR: Fast 3D object reconstruction from a single image," *arXiv preprint arXiv:2403.02151*, 2024.
9. S. Srinidhi, E. Lu, and A. Rowe, "XaiR: An XR platform that integrates large language models with the physical world," in *Proceedings of IEEE ISMAR*, 2024.
10. L. Duan, Y. Xiu, and M. Gorlatova, "Advancing the understanding and evaluation of AR-generated scenes: When vision-language models shine and stumble," in *Proceedings of IEEE VR GenAI-XR*, 2025.
11. A. Alhakamy and M. Tuceryan, "Real-time illumination and visual coherence for photorealistic augmented/mixed reality," *ACM Computing Surveys*, vol. 53, no. 3, pp. 1–34, 2020.
12. H. Adams, J. Stefanucci, S. Creem-Regehr, and B. Bodenheimer, "Depth perception in augmented reality: The effects of display, shadow, and position," in *Proceedings of IEEE VR*, 2022.
13. S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "ZoeDepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.
14. K. Lee, D. Chu, E. Cuervo, J. Kopf, Y. Degtyarev, S. Grizan, A. Wolman, and J. Flinn, "Outatime: Using speculation to enable low-latency continuous interaction for mobile cloud gaming," in *Proceedings of MobiSys*, 2015.
15. S. Gao, J. Liu, Q. Jiang, F. Sinclair, W. Sentosa, B. Godfrey, and S. Adve, "XRgo: Design and evaluation of rendering offload for low-power extended reality devices," in *Proceedings of the ACM MMSys*, 2025.
16. L. Yuan, D. Han, S. Wang, and C. G. Brinton, "Local-cloud inference offloading for LLMs in multi-modal, multi-task, multi-dialogue settings," *arXiv preprint arXiv:2502.11007*, 2025.
17. S. Jang and R. Morabito, "Edge-first language model inference: Models, metrics, and tradeoffs," *arXiv preprint arXiv:2505.16508*, 2025.
18. A. Dhakal, X. Ran, Y. Wang, J. Chen, and K. Ramakrishnan, "SLAM-Share: Visual simultaneous localization and mapping for real-time multi-user augmented reality," in *Proceedings of CoNEXT*, 2022.

Lin Duan is a Ph.D. candidate at Duke University, Durham, NC, 27708, USA. Her research interests include augmented reality, scene understanding, and distributed learning. Contact her at lin.duan@duke.edu.

Elias Rotondo is a Ph.D. student at Duke University, Durham, NC, 27708, USA. His research interests include computer vision, augmented reality, and AI perception. Contact him at eli.rotondo@duke.edu.

Yanming Xiu is a Ph.D. candidate at Duke University, Durham, NC, 27708, USA. His research interests include augmented reality, computer vision, and AI. Contact him at yanming.xiu@duke.edu.

Sangjun Eom is a Ph.D. candidate at Duke University, Durham, NC, 27708, USA. Her research interests include extended reality for medical applications and the Internet of Things. Contact her at sangjun.eom@duke.edu.

Ryan Chen received a B.S.E. in ECE with distinction from Duke University, Durham, NC, 27708, USA. His research interests include augmented reality and computer graphics. Contact him at ryan.j.chen@duke.edu.

Conrad Li is a M.S. student at Duke University, Durham, NC, 27708, USA. His research interests include multi-modal learning and reinforcement learning. Contact him at conrad.li@duke.edu.

Yuhe Hu is a B.S.E. student in ECE at Duke University, Durham, NC, 27708, USA. Her research interests include scene understanding and computer graphics. Contact her at alice.hu@duke.edu.

Maria Gorlatova is an associate professor at Duke University, Durham, NC, 27708, USA. Her research interests include augmented reality, the Internet of Things, and mobile pervasive systems. Gorlatova received her Ph.D. degree in electrical engineering from Columbia University. Contact her at maria.gorlatova@duke.edu.