

Perceptually Guided 3DGS Streaming and Rendering for Mixed Reality

Yunxiang Zhang^{1*} Sai Harsha Mupparaju^{1*} Kenneth Chen¹ Jenna Kang¹ Xinyu Zhang²
Maito Omori² Kazuyuki Arimatsu² Qi Sun¹

¹New York University, USA ²Sony Interactive Entertainment, Japan
{yunxiang.zhang, qisun}@nyu.edu kazuyuki.arimatsu@sony.com

Abstract

Recent breakthroughs in radiance fields, particularly 3D Gaussian Splatting (3DGS), have unlocked real-time, high-quality rendering of complex environments, enabling a wide range of applications. However, the stringent requirements of mixed reality (MR) rendering, such as rapid refresh rates, high-resolution stereo viewing, and constrained computing budgets, remain out of reach for current 3DGS techniques. Nevertheless, the wide field-of-view design of MR displays, which mimics human vision, presents a unique opportunity to exploit human visual system’s own perceptual limitations to reduce computational overhead while not compromising user-perceived rendering quality.

To this end, we propose a perception-guided, continuous level-of-detail (LOD) framework for 3DGS that maximizes perceived quality under given compute resources. We distill a visual quality metric, which encodes the spatial, temporal, and peripheral characteristics of human visual perception, into a lightweight, gaze-contingent model that predicts and adaptively modulates rendering LOD across the visual field based on each region’s contributions to perceptual quality. This budget-driven LOD modulation, guided by both scene content and gaze behavior, enables significant computation reduction with minimal loss in perceived quality. To support low-power, untethered MR setups, we design an edge-cloud collaborative rendering framework to partially offload computation to the cloud, further reducing overhead on the edge MR devices. Objective metrics and MR user study evidence that, compared to vanilla and foveated LOD baselines, our method achieves superior trade-offs between computational efficiency and user-perceived visual quality.

1. Introduction

Recent advances in neural rendering, most notably 3DGS, have transformed traditional graphics pipelines and enabled breakthroughs across various domains, including scientific

visualization [6], interactive gaming [21], autonomous driving [61], robotic controls [54], and virtual productions [45]. Building upon the 3DGS framework, numerous follow-up efforts have been made to further enhance rendering quality [55, 58], support dynamic content [29, 56], accelerate rendering speed [15, 18, 38], reduce memory footprint [15, 16], and introduce LOD mechanisms for scalability [24, 36, 44]. Despite these advances, the current state-of-the-art remains insufficient to satisfy the stringent demands of MR systems, such as rapid refresh rates (≥ 60 FPS) [52], high-resolution stereo rendering ($\geq 2K \times 2K$ per eye) [51, 59], and compute and power constraints, particularly for untethered MR [8].

At the same time, the wide field-of-view design of MR displays, intended to mirror natural human vision, offers an opportunity to leverage the perceptual limitations of human visual system (HVS) [11] to reduce computation while not compromising perceived visual quality. In particular, the HVS exhibits strong eccentricity effects: peripheral vision (high eccentricity) has significantly lower acuity and sensitivity than foveal vision (low eccentricity) [10, 39, 48]. This perceptual non-uniformity has motivated foveated rendering, which prioritizes compute resources in the foveal region over the periphery [19, 22, 40, 49, 50]. Recent efforts have extended these principles to the 3DGS framework to enable real-time stereo rendering tailored for MR settings [14, 17, 28]. However, existing methods typically partition the visual field into a few concentric zones (often 3-4), each tied to a discrete LOD. This coarse, fixed stratification limits the ability to smoothly modulate LOD or dynamically adjust rendering cost in response to changes in scene complexity or power budgets. Furthermore, foveation alone does not account for the implications of local scene features, such as luminance, contrast, and spatial frequency, on visual sensitivity [1]. For instance, artifacts on a high-contrast object in the periphery may be more perceptible than those on a low-contrast object near the fovea, underscoring the need for more content-aware perceptual guidance.

To address these limitations, we propose a perceptually guided 3DGS framework that maximizes perceived visual quality for MR stereo rendering under constrained compute

*Both authors contributed equally to this research.

budgets. Specifically, we partition the visual field of each eye into a dense, uniform grid of $N_h \times N_w$ viewing frustums, and independently control rendering quality for each frustum via LOD modulation. Building upon the hierarchical 3D Gaussian representation from [24], we define LOD as a projected size threshold (in screen pixels) for 3D Gaussians and only retrieve those of appropriate granularity for rendering. To enable efficient execution, we implement custom CUDA kernels that fuse LOD fetching with frustum culling, allowing all frustums to be processed in a single pass with linear complexity in the number of Gaussians. To ensure perceptual consistency across the visual field, we distill FovVideoVDP [31], a visual quality metric that models the spatial, temporal, and eccentricity-dependent aspects of the HVS, into a lightweight, gaze-contingent prediction model. This model takes as input a viewpoint (position, direction, and eccentricity) and a target perceptual quality level (measured in Just-Objectable Differences, or JODs [42]), and outputs the minimal LOD needed to meet the desired quality under the given viewpoint. This approach enables continuous, adaptive LOD modulation across the visual field based on each region’s contribution to perceptual quality, while also supporting flexible, resource-aware control by adjusting the target quality level.

To better support low-power, untethered MR setups, we develop an edge-cloud collaborative rendering framework that further offloads computation to the cloud. Specifically, both the cloud and the edge store a copy of the scene represented as hierarchical 3D Gaussians. At runtime, the edge transmits the user’s current viewpoint and desired quality level (in JODs, determined by local compute budget) to the cloud, while the cloud runs the LOD prediction model to determine the optimal LOD values across the visual field, fetches the corresponding Gaussians, and returns a binary mask indicating the subset of Gaussians that should be rendered on the edge side, all in real time.

To support our investigation of perceptual sensitivity in immersive MR setups, we construct four high-quality image datasets featuring four photorealistic scenes for 3DGS training and evaluation. As evidenced by a perceptual quality metric tailored to human vision [31], our method achieves superior trade-offs between computational efficiency and perceptual quality compared to both the vanilla hierarchical 3DGS [24] and foveated 3DGS baselines. A psychophysical user study conducted in MR further confirms that, under similar compute budgets, participants consistently perceive our method as delivering higher visual quality.

We hope this work inspires future research toward more perceptually efficient MR systems. Our source code, trained models, and evaluation dataset are released at <https://github.com/NYU-ICL/perceptual-3dgs>.

In summary, our main contributions include:

- a lightweight, gaze-contingent prediction model that en-

ables continuous and adaptive LOD modulation across the visual field to ensure perceptual consistency;

- a perceptually guided 3DGS framework that maximizes perceived visual quality under given compute resources;
- an edge-cloud collaborative 3DGS rendering framework tailored to low-power, untethered MR setups;
- four high-quality image datasets to support MR research.

2. Related Work

2.1. Neural Scene Representation and Rendering

Recent advances in neural rendering have transformed traditional graphics pipelines, achieving photorealistic visual synthesis while bypassing the need for explicit scene modeling, complex lighting setups, and handcrafted assets [43]. A seminal milestone in this area is Neural Radiance Fields (NeRFs), which employ neural networks to implicitly encode scene geometry and appearance [35]. Building on this framework, numerous extensions have been proposed to improve visual fidelity [3–5] and accelerate inference [7, 9, 37, 57]. Despite improved performance, NeRF-based methods remain constrained by the computational overhead of volumetric ray marching, which hinders their ability to deliver high-resolution rendering at real-time frame rates.

To address these limitations, Kerbl et al. proposed 3DGS [23], an explicit scene representation that enables real-time, high-resolution rendering. 3DGS models a scene as a set of anisotropic 3D Gaussians, with scene geometry defined by the Gaussian distributions and view-dependent color encoded via spherical harmonics. By replacing ray marching with GPU-efficient rasterization, 3DGS achieves substantial gains in rendering speed while maintaining competitive visual quality. Several recent follow-ups have enhanced vanilla 3DGS along multiple directions: improving visual quality [55, 58], supporting dynamic content [29, 56], enabling semantic feature learning [60], reducing computational cost [15, 16, 18, 25, 38], and introducing efficient LOD mechanisms [24, 36, 44]. While these advances have expanded 3DGS’s applicability across diverse domains, current state-of-the-art still falls short of meeting the stringent demands of untethered mixed reality (MR) systems, such as high refresh rates (≥ 60 FPS) [52], high-resolution stereo rendering ($\geq 2K \times 2K$ per eye) [51, 59], and strict constraints on compute and power budgets [8, 26].

2.2. Perceptually Guided Rendering

The HVS exhibits a highly non-uniform ability to receive and resolve light across the visual field [11]. A key characteristic of human vision is eccentricity dependence: peripheral vision (high eccentricity) exhibits significantly lower visual acuity, reduced contrast sensitivity, and diminished color perception compared to foveal vision (low eccentricity) [10, 39, 48]. Several computational models have

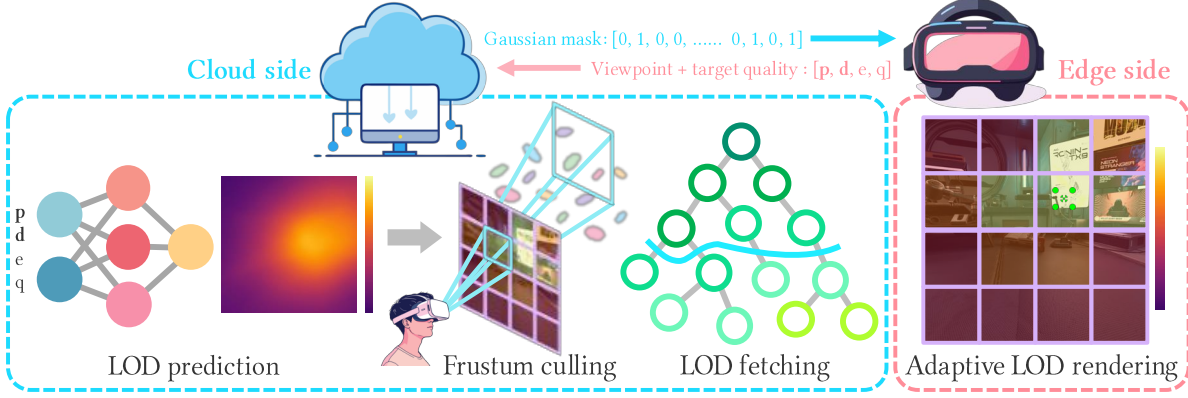


Figure 1. The proposed perceptually guided, adaptive level-of-detail 3DGS streaming and rendering framework for mixed reality.

been developed to describe these perceptual variations by analyzing stimulus visibility across different eccentricities [1, 31, 32]. These models have informed a broad class of perceptually guided rendering techniques that exploit the anisotropic sensitivity of the HVS. Among them, foveated rendering reduces computational load by focusing resources to the foveal region, where visual acuity is highest, while allocating fewer resources to the periphery, where detail perception drops, thereby improving performance without compromising perceived quality [19, 22, 40, 49, 50].

Recently, these principles have been adapted to neural rendering, resulting in foveated variants of NeRF [12, 46, 53] and 3DGS [14, 17, 28]. While these methods improve rendering speed without compromising perceptual fidelity, they typically divide the visual field into a small number of concentric eccentricity-based zones (usually 3–4), each assigned a discrete LOD [12, 14, 28]. This coarse, fixed stratification limits the system’s ability to smoothly modulate LOD or dynamically adjust rendering resources based on scene complexity or device compute & power constraints. Moreover, foveation alone overlooks important content-aware factors, such as local luminance, contrast, and spatial frequency, that also strongly influence visual sensitivity [1]. For example, artifacts on a high-contrast object in the periphery may be more perceptible than those on a low-contrast object near the fovea.

3. Method

3.1. Continuous Level-of-detail 3DGS via Hierarchical 3D Gaussian Representation

3D Gaussian Splatting (3DGS) [23] represents a scene as a point cloud composed of semi-transparent 3D Gaussians, each carrying spatial, color, and opacity information. These Gaussians can be efficiently sampled using simple ray casting operations, which accumulate contributions along each ray in a view-dependent manner. This process ultimately

enables the rendering of high-quality 2D images from arbitrary camera viewpoints.

$$I : (\mathbf{p}, \mathbf{d}; \{G_j\}) \rightarrow \mathbb{R}^{H \times W}, \text{ where } j \in \{1..n\}, \quad (1)$$

where $\mathbf{p} \in \mathbb{R}^3$ and $\mathbf{d} \in \mathbb{R}^3$ represent the camera pose and G denotes the parameters of each of the n Gaussians.

Building on hierarchical 3DGS [24], we represent each scene as a binary tree of 3D Gaussians, where progressively smaller Gaussians are placed deeper in the hierarchy to capture increasingly fine-grained scene details. We partition the visual field of each monocular view into a dense, uniform grid of $N_h \times N_w$ viewing frusta, and modulate each viewing frustum independently using an appropriately chosen LOD value. The LOD value l is defined as a projected size threshold (in screen pixels) for the bounding boxes of 3D Gaussians. Within each viewing frustum, a cut is performed over the binary tree using the selected LOD value (as illustrated in the LOD fetching stage of Fig. 1), ensuring that Gaussians on the two sides of the cut edge have their projected sizes above and below the LOD value, respectively. 3D Gaussians on both sides of all cut edges are fetched for rendering.

3.2. Measuring Perceptual Quality

Rendering images using only coarse LOD Gaussians significantly reduces computational cost but often results in noticeable degradation in visual quality. To make principled use of this trade-off, we require an objective metric that quantifies the perceptual impact of such quality loss and determines whether it is likely to be noticed by a human observer. To this end, we adopt the FovVideoVDP (fvvdvp) metric [31], which estimates the perceptual difference between a test image rendered at a lower LOD l and a reference image rendered at the highest fidelity level l_{ref} . In our case, $l_{ref} = 0$ as it indicates the minimum possible projected size, thus the best quality rendering. This metric accounts for human visual sensitivity by incorporating

the viewer’s gaze direction, thereby aligning the evaluation with perceptual importance across the field of view.

Formally, this metric is defined as a function, $\text{fvdp} : (I_{\text{test}}, I_{\text{ref}}, e) \rightarrow q$ where I_{test} and I_{ref} denote the test and reference renderings, respectively, and e indicates the visual eccentricity. Output quality is quantified in Just-Objectable Difference (JOD) [42], where differences can be interpreted as preference percentages. For example, a 1-JOD advantage of condition A over B corresponds to a 75% preference for A. The resulting score q denotes the proportion of observers who would perceive the test image as objectionably different from the reference, given the gaze direction. For a given LOD value l , we evaluate:

$$q_l = \text{fvdp} \left[I(\mathbf{p}, \mathbf{d}; \{G_j^l\}), I(\mathbf{p}, \mathbf{d}; \{G_j^{l_{\text{ref}}}\}), e \right], \quad (2)$$

This formulation allows us to adaptively modulate the LOD per view and gaze direction while ensuring that perceptual degradation remains within acceptable bounds, effectively balancing quality and efficiency.

3.3. Transforming Image-space Perceptual Quality into LOD-modulated 3D Gaussians

Due to the varying geometric and perceptual complexity present in 3D scenes, fine-grained control over which Gaussians are included during rendering is essential for balancing computational cost and image quality. Fortunately, the hierarchical organization of Gaussians enables such control by allowing selective masking of high LOD Gaussians during ray casting, based on perceptual quality requirements along different viewing directions.

To achieve this, we define a spatially varying LOD map, denoted as $L(q) \in [0, 100]^{N_h \times N_w}$, which specifies the minimum LOD value required to satisfy a target perceptual quality threshold q in each frustum. Using this map, we can cull unnecessary Gaussians from the hierarchy during rendering, effectively enabling adaptive LOD selection that achieves an optimal trade-off between visual fidelity and computational efficiency (see Figure 1). This operation can be performed efficiently thanks to the recursive structure of the hierarchical 3D Gaussian representation [24], which supports fast cut based on LOD. However, computing the full LOD map at test time by exhaustively rendering the image at all LOD levels would negate the computational savings our approach aims to achieve. To overcome this, we propose a learning-based strategy that predicts the optimal LOD map without explicit per-pixel comparisons.

Specifically, we train a lightweight MLP that predicts L from the camera parameters $(\mathbf{p}, \mathbf{d})$, eccentricity e , and desired perceptual quality q . Since the FovVideoVDP metric operates on spatially local neighborhoods of pixels [31], we simplify the prediction task by learning at the patch level rather than the full image level.

To construct our training data, we generate a dataset of image patches per scene. Each patch $P_k^l \in \mathbb{R}^{w_p \times h_p}$ is rendered at multiple LOD levels $l_k \in [0, 100]$ for $k = 1, \dots, n_p$. By choosing patches that subtend small visual angles, we further simplify the perceptual evaluation by fixing the eccentricity e , allowing us to model the quality prediction as: $\text{fvdp} : (P_k^l, P_k^{l_{\text{ref}}}, e) \rightarrow q_l$.

Using the perceptual scores q and applying Equation (2), we determine the JOD-LOD relationship for each patch given the associated camera and gaze parameters. This yields a training set: $(\mathbf{p}, \mathbf{d}, e, q) \mapsto l$. The LOD predictor is then trained to regress this mapping using standard stochastic gradient descent with a mean squared error (MSE) loss. Once trained, it can efficiently infer L for novel viewpoints and gaze directions at runtime as depicted in Fig. 2, enabling perception-aware and compute-efficient rendering.

3.4. Perceptually Guided 3DGS in an Edge-Cloud Collaborative Rendering Framework

To better support low-power, untethered MR setups, we develop an edge-cloud collaborative rendering framework that further offloads computation to the cloud. Specifically, both the cloud and the edge store a copy of the scene represented as hierarchical 3D Gaussians. At runtime, the edge transmits the user’s current viewpoint and desired quality level in JODs (determined by local compute budget) to the cloud, while the cloud runs the LOD prediction model to determine the optimal LOD values across the visual field, fetches the corresponding Gaussians, and returns a binary mask indicating the subset of Gaussians that should be rendered on the edge side, all in real time. Figure 1 illustrates our perceptually guided 3DGS streaming and rendering pipeline, showcasing the division of responsibilities between edge and cloud to enable scalable, perceptually optimized rendering in resource-constrained environments.

4. Evaluation

4.1. Implementation Details

Image dataset To support our investigation of perceptual sensitivity in immersive MR setups, we constructed four high-fidelity image datasets featuring four photorealistic scenes for evaluation. All scenes were rendered using high-resolution textures and the Cycles path tracer in Blender to ensure realistic lighting and material appearance. Cameras were uniformly placed at varying positions and angles throughout each scene to ensure good coverage, with 100 randomly selected views held out for evaluation. All images were rendered at a resolution of 1920×1920 pixels using an 18 mm full-frame equivalent focal length, corresponding to a $90^\circ \times 90^\circ$ monocular field of view, aligned with that of most commercial MR displays.

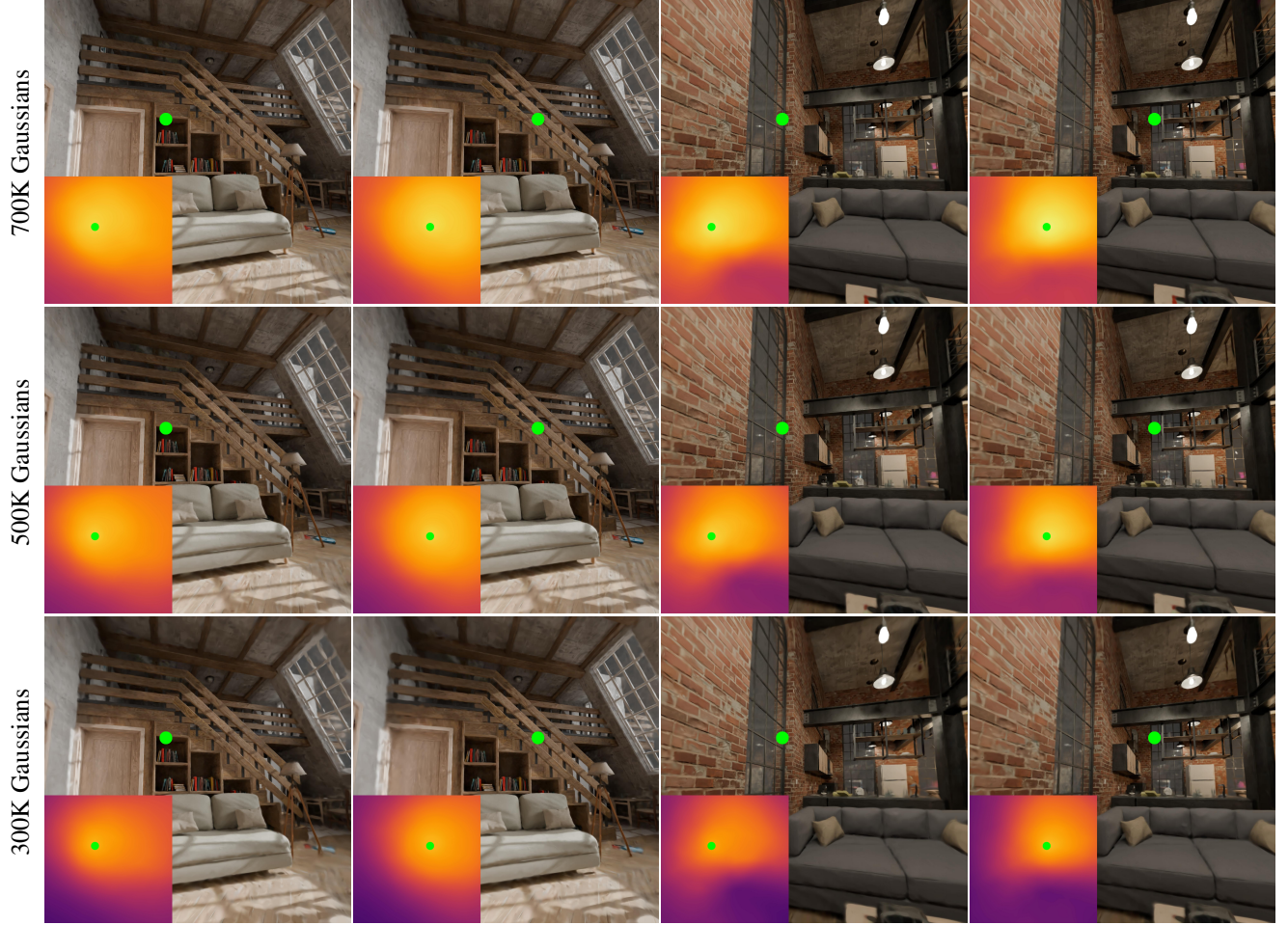


Figure 2. Visual comparison of our perception-guided, continuous LOD modulation across varying eye gaze positions and compute budget levels. Each row corresponds to a different compute budget level, determined by the average number of Gaussians used for rendering. Each column corresponds to a different eye gaze position, marked by a green dot. The inset in the bottom-left corner shows the LOD map, where brighter regions indicate finer LODs.

LOD prediction To train the LOD prediction model described in Section 3.3, we built a dedicated dataset for each scene by computing FovVideoVDP [31] on images rendered at varying LOD values. Specifically, for each training camera view with position $\mathbf{p} \in \mathbb{R}^3$ and viewing direction $\mathbf{d} \in \mathbb{R}^3$ (a unit vector in 3D), we randomly sampled 40 LOD values from $[0, 50]$ and 20 from $[50, 100]$. The view was then rendered at each sampled LOD value l . Each rendered image was compared against the reference image (rendered at $l = 0.0$) using FovVideoVDP to produce a pixel-wise perceptual quality map in JODs. From each obtained JOD map, we randomly sampled 100 peripheral directions $\theta \in [0^\circ, 360^\circ]$ and eccentricity values $e \in [0^\circ, 45^\circ]$, cropping 64×64 pixel patches centered at the sampled coordinates. Each patch was then pooled into a single scalar JOD value q , representing the perceptual quality of the corresponding image region when rendered at LOD

l and viewed from viewpoint $(\mathbf{p}, \mathbf{d})$ at eccentricity e . As a result, each camera view yielded 6K data tuples of the form $(\mathbf{p}, \mathbf{d}, e, q, l)$. All such tuples collected from the training views of a given scene were used to optimize a scene-specific LOD prediction model. This model predicts the minimal LOD l needed to achieve a target perceptual quality q under a given viewpoint defined by $(\mathbf{p}, \mathbf{d}, e)$.

4.2. Objective Evaluation

Baselines To quantitatively evaluate our perceptually guided 3DGS method (**PERCEPTUAL**), we compare it against two baselines: the vanilla hierarchical 3DGS **UNIFORM** and foveated 3DGS **FOVEATED**. **UNIFORM** applies a constant LOD value uniformly across the visual field, regardless of eye gaze position. **FOVEATED** adaptively modulates LOD based on eye gaze position, assigning finer LOD near the fovea and progressively coarser LOD



Figure 3. Visual comparison of different LOD modulation methods. All images, except the reference, were rendered using 300K Gaussians.

toward the periphery. We evaluate perceptual quality using FovVideoVDP and assess computational efficiency by measuring the number of Gaussians used during rendering, which determines the edge-side rendering speed.

Experimental setup For each method (**UNIFORM**, **FOVEATED**, and **PERCEPTUAL**), we rendered the 100 test camera views at three target compute budget levels, corresponding to average Gaussian count of 300K, 500K, and 700K used during rendering. These levels were obtained differently for each method: **UNIFORM** by adjusting a constant LOD value; **FOVEATED** by varying a strength parameter that governs the slope of its foveated LOD falloff; and **PERCEPTUAL** by modulating the perceptual quality input q to the LOD prediction model. For both **FOVEATED** and **PERCEPTUAL**, we further evaluated their gaze-contingent performance by rendering with

three different eye gaze positions, each producing a distinct LOD map. In total, this resulted in 3 rendering setups for **UNIFORM**, and 9 (3 gaze positions $\times$ 3 compute budget levels) for each of **FOVEATED** and **PERCEPTUAL**. The highest-quality rendering ($l = 0.0$) was used as the reference for computing FovVideoVDP.

Results and discussion Figure 3 presents a randomly sampled test view rendered using different LOD modulation approaches. Figure 4 summarizes the JOD scores achieved by the three LOD modulation methods, aggregated across all four scenes and three eye gaze positions. As shown, **PERCEPTUAL** consistently outperforms both **FOVEATED** and **UNIFORM** across all computational efficiency levels. Notably, **PERCEPTUAL** achieves a JOD of -0.84 ± 0.14 using an average of 700K Gaussians, just one quarter of the Gaussian count used in the reference ren-

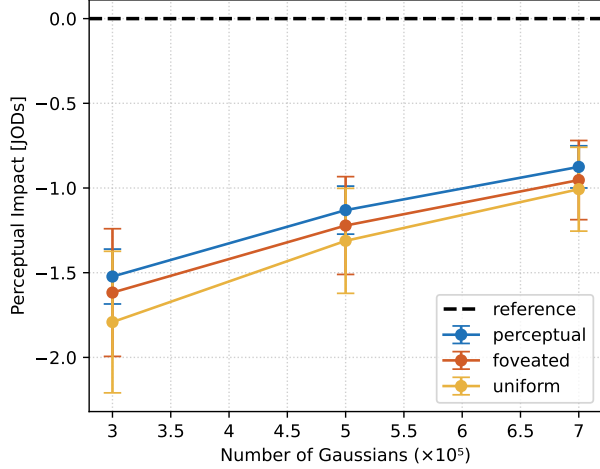


Figure 4. Objective evaluation results. Perceptual visual quality is measured in JODs, where higher scores indicate better quality.

dering, corresponding to a 75% reduction in computation. Even at an extremely low Gaussian count of 300K (an 86% computation reduction), **PERCEPTUAL** still maintains a JOD of -1.51 ± 0.18 . We attribute this superior performance to **PERCEPTUAL**’s ability to adaptively allocate Gaussians based on the perceptual importance of scene content. The resulting reduction in computation results in substantial rendering speedup. While the reference rendering uses an average of 2.56M Gaussians and runs at 63.51 FPS across the four scenes, our LOD-modulated 300K variant provides $5\times$ acceleration and runs at 331.53 FPS. In addition, LOD prediction takes 2.64 ms on average, and our efficient implementation of LOD fetching and frustum culling via custom CUDA kernels enables an average overhead of 17.76 ms. Note that both operations can be offloaded to the cloud under our proposed edge-cloud collaborative rendering framework. All the time measurements were obtained locally on an Nvidia RTX A6000 GPU.

4.3. User Study

Participants We recruited 16 participants aged 21-29 (11M, 5F). All participants signed informed consent and reported normal or corrected-to-normal vision. The study was approved by an Institutional Review Board (IRB).

Experimental Setup We used one scene for the study. We sampled four test camera views from the scene and interpolated their 6-degree-of-freedom poses to obtain a smooth trajectory of $72 \times 18 = 1296$ camera views (displayed as a 72-FPS stereo video for 18 seconds) for evaluation, using linear interpolation for translation and spherical linear interpolation for rotation. For each of **UNIFORM**, **FOVEATED**, and **PERCEPTUAL**, we rendered this tra-

jectory at three compute budget levels and eye gaze positions similar to Section 4.2, resulting in 3 rendering setups for **UNIFORM**, and 9 (3 gaze positions $\times$ 3 compute levels) for each of **FOVEATED** and **PERCEPTUAL**. A Meta Quest Pro was used to collect data during the study.

Experiment Procedure & Stimuli The experimental procedure followed the two-interval forced choice (2IFC) task using the method of paired comparisons. We chose this protocol because it has been used in previous studies evaluating the subjective quality of view synthesis techniques [27] and has been shown to produce less noisy results by presenting a simpler task to users [30].

At the start of each trial, the user is shown the reference video and can freely switch between this and two test videos. Participants used a regular keyboard’s arrow keys to switch between the reference video and the two test videos and were not able to continue to the next trial unless all three videos were viewed at least once. Selections were made using the space key. A 500ms gray screen was added while switching between conditions to provide a clear distinction between conditions.

Participants were seated during the study. During each trial, a fixation is displayed using a square crosshair, the user’s gaze is tracked, and the screen is grayed out if the user’s gaze deviates from the point of fixation for more than 500ms. This is done to ensure the user is focused on the point of fixation throughout the study. Participants were instructed to keep their gaze on the crosshair and select the video closer to the reference, in terms of which had higher quality and fewer distortions. The users were also instructed to spend 15-30 seconds on each trial, but time constraints were not enforced. Before the start of each study, a 5-trial warmup session is conducted to familiarize users with the study setup and stimuli.

All stimuli were stereo rendered in Unity and displayed in the head-mounted display. The study design included a single scene, three levels of Gaussian densities, three fixations, and three Gaussian LOD modulation methods (described in Section 4.2) including a reference condition. Within trials, we show users the same fixation location. For our study, a naïve full pairwise comparison design would lead to 432 comparisons, which is quite large. Instead, we used the active sampling ASAP algorithm [34] to determine the optimal conditions to probe in each trial, resulting in 105 trials per participant. On average, participants spent 45 minutes completing the study, including training and inter-pupillary distance (IPD) adjustment.

Results The pairwise comparison data from our study were converted into JOD units using the *pwcmp* algorithm Perez-Ortiz and Mantiuk [41].

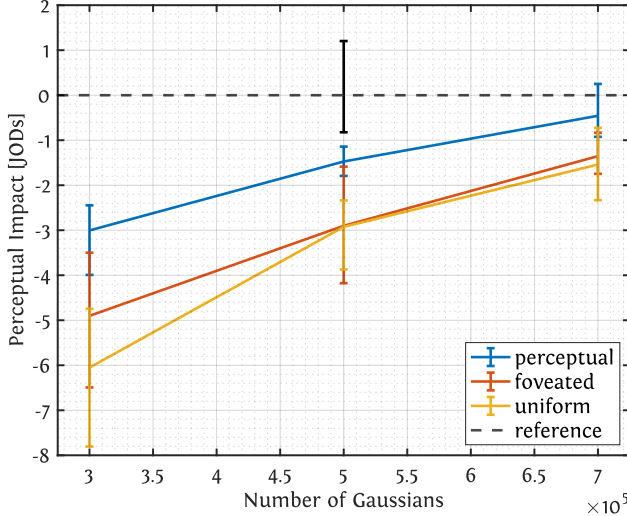


Figure 5. User study results. The x -axis represents the average number of Gaussians used for rendering, while the y -axis indicates perceived visual quality in JOD scores. Error bars are 95% confidence intervals.

We plot the results of the study averaged across fixation positions in Figure 5. We can see that, across all Gaussian densities, our method consistently outperforms the other two techniques, with the largest margin observed at the lower densities (300 K and 500 K). As the JOD scale is relative, the reference condition is pegged at 0, and the scores for each method represent the distance from the reference, which itself is a condition in the study and therefore has a confidence interval associated with it. The 95% confidence intervals are computed with 1,000 bootstrap samples using *pwcmp*. Per-fixation results are shown in the supplement.

An ANOVA analysis was run on the scaled results from the study to determine the effect of relevant independent variables (average Gaussian count, LOD modulation method, and fixation) on the JOD scores. Significant effects were shown by the average Gaussian count ($p = 0.0032$) and LOD modulation method ($p = 0.039$). The main effect of fixation position was not found to be significant.

5. Limitations and Future Work

Computation of Run-time Gaussian Subset Decision.

In our cloud-edge game streaming scenario, computation is split between LoD-based Gaussian subset selection on the cloud and splatting-based rendering on the edge. By selecting a compact subset of Gaussians guided by a perceptual metric (fvvdp), we obtain a representation that is both efficient to render and suitable for real-time inference on edge devices with latency as low as 3.02 ms, well below the 11 ms frame duration of a 90 FPS VR display. Computing this subset, however, through prediction and fetching in a tree-

based structure, remains computationally demanding. Although cloud servers provide ample resources, further optimization could yield significant savings in energy consumption and improve scalability across concurrent clients. We anticipate that high-performance parallelization techniques, such as GPU-based tree traversal [33], could substantially reduce cloud-side costs.

Eye Tracker for Gaze Contingency. Our method is gaze-contingent, which means the subset decision, streaming, and rendering require runtime information on the gaze position on the display. Therefore, it requires eye trackers on the display. While a large body of today’s MR displays, such as Apple Vision Pro, supports eye-tracking, we plan to extend the research toward solutions that do not rely on eye-tracking. For example, we will extend the foveal region in the display center, as human gaze does not commonly extend beyond 15 degrees [2], similar to the foveated rendering approach in low-cost headsets such as Meta Quest 3. Additionally, saliency-based prediction models have been extensively studied [20, 47]. We envision that predicting gaze saliency instead of direct tracking would provide an effective computation performance gain without eye-tracking.

Per-Scene LoD Prediction. We trained the LoD predictor separately for each scene, which yielded consistently high validation accuracy. While effective, this per-scene strategy constrains deployment efficiency, as a new predictor must be retrained whenever a novel scene is introduced. Such retraining not only adds computational overhead but also hinders scalability in large-scale or dynamic content pipelines. To address this limitation, recent advances in deep image characterization [13] suggest promising directions for learning compact, scene-agnostic feature spaces. These representations could enable the design of a universal LoD predictor capable of generalizing across diverse scenes, thereby eliminating the need for scene-specific training and significantly improving practicality for real-world deployment.

6. Conclusion

In this paper, we present an integrated approach combining comprehensive human visual perception models with 3D Gaussian splatting for next-generation MR streaming. By adaptively controlling the level of detail in Gaussian space, our method achieves a superior quality-bandwidth trade-off, validated by both objective metrics and user studies. We envision this work as a step toward advancing human-centered 3D representations for future MR teleportation and gaming experiences.

Acknowledgement

This research is supported by a SONY Focused Research Award.

References

- [1] Maliha Ashraf, Rafal K Mantiuk, Alexandre Chapiro, and Sophie Wuerger. castlcsf—a contrast sensitivity function of color, area, spatiotemporal frequency, luminance and eccentricity. *Journal of vision*, 24(4):5–5, 2024. 1, 3
- [2] A Terry Bahill, Deborah Adler, and Lawrence Stark. Most naturally occurring human saccades have magnitudes of 15 degrees or less. *Investigative Ophthalmology & Visual Science*, 14(6):468–469, 1975. 8
- [3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021. 2
- [4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022.
- [5] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023. 2
- [6] Yuanhao Cai, Yixun Liang, Jiahao Wang, Angtian Wang, Yulun Zhang, Xiaokang Yang, Zongwei Zhou, and Alan Yuille. Radiative gaussian splatting for efficient x-ray novel view synthesis. In *European Conference on Computer Vision*, pages 283–299. Springer, 2024. 1
- [7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pages 333–350. Springer, 2022. 2
- [8] Kenneth Chen, Thomas Wan, Nathan Matsuda, Ajit Ninan, Alexandre Chapiro, and Qi Sun. Pea-pods: Perceptual evaluation of algorithms for power optimization in xr displays. *ACM Transactions on Graphics (TOG)*, 43(4):1–17, 2024. 1, 2
- [9] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16569–16578, 2023. 2
- [10] Michael A Cohen, Thomas L Botch, and Caroline E Robertson. The limits of color awareness during active, real-world vision. *Proceedings of the National Academy of Sciences*, 117(24):13821–13827, 2020. 1, 2
- [11] Michael F Deering. The limits of human vision. In *2nd international immersive projection technology workshop*, 1998. 1, 2
- [12] Nianchen Deng, Zhenyi He, Jiannan Ye, Budmonde Duinkharjav, Praneeth Chakravarthula, Xubo Yang, and Qi Sun. Fov-nerf: Foveated neural radiance fields for virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3854–3864, 2022. 3
- [13] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 8
- [14] Runze Fan, Jian Wu, Xuehuai Shi, Lizhi Zhao, Qixiang Ma, and Lili Wang. Fov-gs: Foveated 3d gaussian splatting for dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 1, 3
- [15] Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, Zhangyang Wang, et al. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. *Advances in neural information processing systems*, 37: 140138–140158, 2024. 1, 2
- [16] Guangchi Fang and Bing Wang. Mini-splatting: Representing scenes with a constrained number of gaussians. In *European Conference on Computer Vision*, pages 165–181. Springer, 2024. 1, 2
- [17] Linus Franke, Laura Fink, and Marc Stamminger. Vr-splatting: Foveated radiance field rendering via 3d gaussian splatting and neural points. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 8(1):1–21, 2025. 1, 3
- [18] Sharath Girish, Kamal Gupta, and Abhinav Shrivastava. Eagles: Efficient accelerated 3d gaussians with lightweight encodings. In *European Conference on Computer Vision*, pages 54–71. Springer, 2024. 1, 2
- [19] Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. Foveated 3d graphics. *ACM transactions on Graphics (TOG)*, 31(6):1–10, 2012. 1, 3
- [20] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2007. 8
- [21] Ying Jiang, Chang Yu, Tianyi Xie, Xuan Li, Yutao Feng, Huamin Wang, Minchen Li, Henry Lau, Feng Gao, Yin Yang, et al. Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–1, 2024. 1
- [22] Anton S Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. Deep-fovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019. 1, 3
- [23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 3
- [24] Bernhard Kerbl, Andreas Meuleman, Georgios Kopanas, Michael Wimmer, Alexandre Lanvin, and George Drettakis. A hierarchical 3d gaussian representation for real-time rendering of very large datasets. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024. 1, 2, 3, 4

- [25] Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3d gaussian representation for radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21719–21728, 2024. 2
- [26] Chen Li, Tongyu Zong, Yueyu Hu, Yao Wang, and Yong Liu. Spatial visibility and temporal dynamics: Rethinking field of view prediction in adaptive point cloud video streaming. In *Proceedings of the 16th ACM Multimedia Systems Conference*, pages 24–34, 2025. 2
- [27] H. Liang, T. Wu, P. Hanji, F. Banterle, H. Gao, R. Mantiuk, and C. Öztireli. Perceptual quality assessment of nerf and neural view synthesis methods for front-facing views. *Computer Graphics Forum*, 43(2):e15036, 2024. 7
- [28] Weikai Lin, Yu Feng, and Yuhao Zhu. Metasapiens: Real-time neural rendering with efficiency-aware pruning and accelerated foveated rendering. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, pages 669–682, 2025. 1, 3
- [29] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 800–809. IEEE, 2024. 1, 2
- [30] Rafał K. Mantiuk, Anna Tomaszewska, and Radosław Mantiuk. Comparison of four subjective methods for image quality assessment. *Computer Graphics Forum*, 31(8):2478–2491, 2012. 7
- [31] Rafał K Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)*, 40(4):1–19, 2021. 2, 3, 4, 5
- [32] Rafał K Mantiuk, Maliha Ashraf, and Alexandre Chapiro. stelacsf: a unified model of contrast sensitivity as the function of spatio-temporal frequency, eccentricity, luminance and area. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 3
- [33] Duane Merrill, Michael Garland, and Andrew Grimshaw. High-performance and scalable gpu graph traversal. *ACM Transactions on Parallel Computing (TOPC)*, 1(2):1–30, 2015. 8
- [34] Aliaksei Mikhailiuk, Clifford Wilmot, Maria Perez-Ortiz, Dingcheng Yue, and Rafał Mantiuk. Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *2020 IEEE International Conference on Pattern Recognition (ICPR)*, 2021. 7
- [35] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 2
- [36] Nicholas Milef, Dario Seyb, Todd Keeler, Thu Nguyen-Phuoc, A Božič, Sushant Kondguli, and Carl Marshall. Learning fast 3d gaussian splatting rendering using continuous level of detail. In *Computer Graphics Forum*, page e70069. Wiley Online Library, 2025. 1, 2
- [37] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 2
- [38] Michael Niemeyer, Fabian Manhardt, Marie-Julie Rakotsaona, Michael Oechsle, Daniel Duckworth, Rama Gosula, Keisuke Tateno, John Bates, Dominik Kaeser, and Federico Tombari. Radsplat: Radiance field-informed gaussian splatting for robust real-time rendering with 900+ fps. In *International Conference on 3D Vision 2025*. 1, 2
- [39] Cynthia Owsley. Contrast sensitivity. *Ophthalmology Clinics of North America*, 16(2):171–177, 2003. 1, 2
- [40] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions On Graphics (TOG)*, 35(6):1–12, 2016. 1, 3
- [41] Maria Perez-Ortiz and Rafał K Mantiuk. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint arXiv:1712.03686*, 2017. 7
- [42] Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafał K Mantiuk. From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing*, 29:1139–1151, 2019. 2, 4
- [43] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation*. MIT Press, 2023. 2
- [44] Kerui Ren, Lihan Jiang, Tao Lu, Mulin Yu, Linning Xu, Zhangkai Ni, and Bo Dai. Octree-gs: Towards consistent real-time rendering with lod-structured 3d gaussians. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1, 2
- [45] Boxiang Rong, Artur Grigorev, Wenbo Wang, Michael J Black, Bernhard Thomaszewski, Christina Tsalicoglou, and Otmar Hilliges. Gaussian garments: Reconstructing simulation-ready clothing with photorealistic appearance from multi-view video. In *International Conference on 3D Vision 2025*. 1
- [46] Xuehuai Shi, Lili Wang, Xinda Liu, Jian Wu, and Zhiwen Shao. Scene-aware foveated neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 3
- [47] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetstein. Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4):1633–1642, 2018. 8
- [48] LN Thibos, FE Cheney, and DJ Walsh. Retinal limits to the detection and resolution of gratings. *Journal of the Optical Society of America A*, 4(8):1524–1529, 1987. 1, 2
- [49] Okan Tarhan Tursun, Elena Arabadzhyska-Koleva, Marek Wernikowski, Radosław Mantiuk, Hans-Peter Seidel, Karol Myszkowski, and Piotr Didyk. Luminance-contrast-aware foveated rendering. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 1, 3

- [50] David R Walton, Rafael Kuffner Dos Anjos, Sebastian Fris-ton, David Swapp, Kaan Akşit, Anthony Steed, and To-bias Ritschel. Beyond blur: Real-time ventral metamers for foveated rendering. *ACM Transactions on Graphics*, 40(4): 1–14, 2021. [1](#), [3](#)
- [51] Jialin Wang, Rongkai Shi, Zehui Xiao, Xueying Qin, and Hai-Ning Liang. Effect of render resolution on gameplay ex-perience, performance, and simulator sickness in virtual re-ality games. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(1):1–15, 2022. [1](#), [2](#)
- [52] Jialin Wang, Rongkai Shi, Wenxuan Zheng, Weijie Xie, Do-minic Kao, and Hai-Ning Liang. Effect of frame rate on user experience, performance, and simulator sickness in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2478–2488, 2023. [1](#), [2](#)
- [53] Zijun Wang, Jian Wu, Runze Fan, Wei Ke, and Lili Wang. Vprf: Visual perceptual radiance fields for foveated image synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 2024. [3](#)
- [54] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Con-ference on Computer Vision and Pattern Recognition*, pages 19595–19604, 2024. [1](#)
- [55] Zhiwen Yan, Weng Fei Low, Yu Chen, and Gim Hee Lee. Multi-scale 3d gaussian splatting for anti-aliased rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20923–20931, 2024. [1](#), [2](#)
- [56] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Pro-ceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20331–20341, 2024. [1](#), [2](#)
- [57] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenotrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5752–5761, 2021. [2](#)
- [58] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splat-ting. In *Proceedings of the IEEE/CVF conference on com-puter vision and pattern recognition*, pages 19447–19456, 2024. [1](#), [2](#)
- [59] Yang Zhao, Dave Lindberg, Bruce Cleary, Olivier Mercier, Ryan McClelland, Eric Penner, Yu-Jen Lin, Julia Majors, and Douglas Lanman. Retinal-resolution varifocal vr. In *ACM SIGGRAPH 2023 Emerging Technologies*, pages 1–3. 2023. [1](#), [2](#)
- [60] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Ze-hao Zhu, Dejie Xu, Pradyumna Chari, Suyu You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Pro-ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024. [2](#)
- [61] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian:

Composite gaussian splatting for surrounding dynamic au-tonomous driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21634–21643, 2024. [1](#)