Audio-visual aware Foveated Rendering

Xuehuai Shi, Yucheng Li, Jiaheng Li, Jian Wu, Jieming Yin, Xiaobai Chen, and Lili Wang

(a) audio-visual aware foveated rendering result (middle) based on extracted auditory and visual perception features (left) vs. the conventional foveated rendering result (right)

(b) time cost plots

Fig. 1: *Illustration of the proposed audio-visual aware foveated rendering method (AvFR).* In a VR scene containing auditory content, AvFR extracts auditory and visual features (a, left) and generates an optimized foveated rendering result (a, middle) based on the extracted features to accelerate rendering performance while preserving visual fidelity compared to the state-of-the-art foveated rendering method (FR) shown in (a, right) for VR head-mounted displays [26]. The time cost comparison between AvFR (b, bottom) and FR (b, top) shows that AvFR achieves a 1.3× speedup compared to FR, which indicates that AvFR significantly improves foveated rendering performance in VR scenes containing auditory content.

Abstract—With the increasing complexity of geometry and rendering effects in virtual reality (VR) scenes, existing foveated rendering methods for VR head-mounted displays (HMDs) struggle to meet users' demands for VR scene rendering with high frame rates ($\geq 60fps$ for rendering binocular foveated images in VR scenes containing over $50M$ triangles). Current research validates that auditory content affects the perception of the human visual system (HVS). However, existing foveated rendering methods primarily model the HVS's eccentricity-dependent visual perception ability on the visual content in VR while ignoring the impact of auditory content on the HVS's visual perception. In this paper, we introduce an auditory-content-based perceived rendering quality analysis to quantify the impact of visual perception under different auditory conditions in foveated rendering. Based on the analysis results, we propose an audio-visual aware foveated rendering method (AvFR). AvFR first constructs an audio-visual feature-driven perception model that predicts the eccentricity-based visual perception in real time by combining the scene's audio-visual content, and then proposes a foveated rendering cost optimization algorithm to adaptively control the shading rate of different regions with the guidance of the perception model. In complex scenes with visual and auditory content containing over $1.17M$ triangles, AvFR renders high-quality binocular foveated images at an average frame rate of $116fps$. The results of the main user study and performance evaluation validate that AvFR achieves significant performance improvement (up to 1.4× speedup) without lowering the perceived visual quality compared with the state-of-the-art VR-HMD foveated rendering method.

Index Terms—Virtual Reality, Foveated Rendering, Perception-driven Rendering

1 Introduction

Foveated rendering, a rendering acceleration technique, takes advantage of the capabilities and limitations of the human visual system (HVS) to improve rendering performance in a way that is unnoticeable to users. As the geometric structures, lighting, and animation effects in virtual reality (VR) scenes become increasingly complex, current foveated rendering can hardly meet users' demands for rendering high-quality binocular images at high frame rates in head-mounted displays (HMDs). In VR, current foveated rendering methods often rely on visual-perception models. These methods are based on the understanding that the HVS has a limited capacity to perceive visual content, but they have not yet taken into account how auditory factors can affect the HVS's perception.

Research shows that audio exerts effects on the HVS's visual perception. From the physiological perspective, the human auditory system (HAS) and the HVS are among the main sensory systems involved in postural control and balance [7]. When vestibular function is compromised, it weakens HVS's perception through visual-vestibular integration [1, 39]. When the perceived loudness of audio reaches a certain threshold, rod cells in the retina exhibit reduced sensitivity to brightness discrimination [14]. The HAS is also sensitive to audio frequencies, and even if the actual sound pressure level remains constant, the perceived loudness of audio in the HAS varies with frequency [29, 43]. Moreover, the semantic coherence between auditory and visual stimuli affects HVS's perception [21]. Thus, in VR scenes containing auditory content, the HVS's visual perception shifts under varying auditory conditions. Furthermore, when HVS's visual perception ability degrades

- *Xuehuai Shi is with School of Computer Science, Nanjing University of Posts and Telecommunications, Jiangsu, China, 210023; State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, 100191. E-mail: xuehuai@njupt.edu.cn*
- *Jiaheng Li and Yucheng Li are with School of Computer Science and Engineering, Beihang University, Beijing, 100191.*
- *Jian Wu, and Lili Wang are with State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, 100191.*
- *Jieming Yin and Xiaobai Chen are with School of Computer Science, Nanjing University of Posts and Telecommunications, Jiangsu, China, 210023.*
- *Jiaheng Li and Jieming Yin are corresponding authors.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

under specific auditory conditions, more aggressive foveated rendering can be implemented without sacrificing perceived fidelity.

Existing foveated rendering techniques model the HVS's eccentricity-dependent perceptual capability based on the visual content of VR scenes, disregarding the effects of auditory content on visual perception. To optimize foveated rendering by considering both auditory and visual content, two main challenges must be addressed. First, to quantify the eccentricity-dependent visual perceptual ability of the HVS, a perception model must be developed to account for heterogeneous features from auditory and visual content in VR scenes. Second, guided by the perception model, the foveated rendering cost needs to be optimized to achieve significant performance improvement without sacrificing visual fidelity.

In this paper, we first introduce an auditory-content-based perceived rendering quality analysis to provide theoretical foundation and experimental validation for audio-visual aware foveated rendering. To address the above two challenges, we propose the Audio-Visual aware Foveated Rendering method (AvFR). For the first challenge, we present an audio-visual feature-driven perception model to predict the HVS's visual perceptual capability in foveated rendering under varying audio-visual conditions. For the second challenge, we propose a foveated rendering cost optimization algorithm that achieves significant shading rate optimization guided by the predicted eccentricity-dependent visual perceptual ability, without sacrificing perceived visual quality.

Fig. 1 gives the illustration of AvFR in VR scene *library*. AvFR first extracts auditory and visual features (Fig. 1(a) left) to construct the audio-visual feature-driven perception model. Using the cost optimization algorithm, AvFR generates optimized foveated rendering results (Fig. 1(a) middle). Compared with the state-of-the-art VR-HMD foveated rendering method (FR) [26] (Fig. 1(a) right), AvFR achieves comparable perceived visual quality. Fig. 1(b) compares rendering performance: AvFR (bottom) achieves an average 83 fps in *library* (8.85M triangles) – 1.3× speedup over FR (up).

In summary, the contributions of this paper are as follows:

- An auditory-content-based perceived rendering quality analysis to measure the eccentricity-dependent perceived visual quality under varying auditory conditions;
- An audio-visual feature-driven perception model to quantify the eccentricity-dependent visual perceptual ability based on the auditory and visual content of VR scenes;
- A foveated rendering cost optimization algorithm to achieve significant foveated rendering cost savings compared with the state-of-the-art VR-HMD foveated method according to the eccentricity-dependent visual perceptual ability.

Source code can be found on the web page [1].

2 Related Work

In this section, we review the recent advances of foveated rendering and perception-driven rendering in VR that are related to our work.

2.1 Foveated Rendering in VR

Foveated rendering technology takes advantage of the non-uniform feature of the HVS sensitivity of the retina by dynamically adjusting the image quality in different regions of the visual field. This technology enhances rendering performance without losing visual fidelity [11, 49]. The recent research of foveated rendering in VR that is related to our work focuses on leveraging perception models to accelerate rendering in the graphics pipeline.

Patney et al. [26] first design a VR foveated rasterization system based on the visual acuity fall-off model. It proposes a novel foveated anti-aliasing algorithm that is implemented in the coarse pixel shading technique (CPS), aiming to recover peripheral details that are resolvable by the HVS but with lower rendering quality, thus achieving a significant reduction in the number of shading operations in VR HMDs. To further improve foveated rasterization performance in VR, Meng et al. [24] introduce a kernel-mapping-based foveated rendering framework (KMF), which uses a kernel transformation function to map foveated rendering computations into a low-resolution and non-uniform shading space, thereby achieving variable-resolution rendering effect within a single shading process. Additionally, Ye et al. [52] propose a rectangular mapping function to preserve peripheral sharp details in the KMF. Later, Fan et al. [9] further improve the rendering quality in the KMF by employing a novel convolutional kernel mapping function to increase the shading density in peripheral salient regions. Zhang et al. [53] reduce unnecessary rendering costs in the upper and lower visual fields for foveated rendering based on the horizontal-vertical and vertical-meridian asymmetries of the HVS.

Besides the decline in visual acuity characteristics of the HVS, properties such as ocular dominance, contrast sensitivity, and attention mechanisms further enhance the quality and performance of foveated rendering in VR. Meng et al. [23] utilize the ocular dominance characteristics of the HVS to reduce the rendering quality of the non-dominant eye in VR HMD rendering, thereby further improving rendering performance without sacrificing perceived visual quality. Shi et al. [32] leverage changes in HVS visual acuity under different motion conditions to parametrically adjust the shading of peripheral regions based on various motion patterns, further accelerating foveated rendering. Based on the spatiotemporal contrast sensitivity function (CSF), Stengel et al. [38] propose a foveated sampling method to optimize performance without affecting perceived visual quality by shading only regions with important image features and interpolating the remaining areas. Some studies further refine foveated rendering by using luminance CSFs, which improve computational efficiency by reducing the number of sample rays in regions with low luminance contrast sensitivity [31, 46]. Krajancich et al. [18] introduce an attention-aware contrast sensitivity model into the foveated rendering framework to dynamically adjust rendering quality based on user attention.

Although existing methods utilize various physiological characteristics of the HVS to optimize foveated rendering, they overlook the impact of auditory content on visual perception in VR. To further enhance performance without compromising perceived visual quality, we specifically optimize the cost of foveated rendering based on both auditory and visual content in VR scenes.

2.2 Perception-driven Rendering in VR

Another related research field is perception-driven rendering in VR, which focuses on optimizing user experience by leveraging cross-modal perception to guide rendering in VR.

Many researchers leverage the spatiotemporal features of visual cues to optimize VR rendering. Jindal et al. [16] dynamically control local shading and refresh rates based on motion perception to optimize the trade-off between rendering quality and performance. Krajancich et al. [17] introduce a new model to jointly measure eccentricity-dependent critical flicker fusion thresholds for both space and time to guide perception-driven rendering. Additionally, Tursun et al. [45] measure the temporal aspect of visual perception in the periphery, demonstrating how foveated rendering methods can be evaluated and optimized to limit the visibility of temporal aliasing based on the proposed model.

In addition to visual cues, the impact of sound on visual perception is studied in VR. To improve the sound fidelity of virtual environments, Tang et al. [44] learn the acoustic characteristics of different environments, ensuring that the sounds emitted by virtual objects match the current environmental semantics. Malpica et al. [19] first demonstrate that sound effects influence the perception of material rendering in VR through a series of user experiments. They validate that sound effects have a more significant impact on low-quality rendering than on high-quality ones during material perception. Subsequently, they explore the effect of auditory stimuli on visual performance in VR and conclude that the detrimental effect of auditory stimuli on visual performance is significant regardless of cognitive load levels [20]. Jimenez et al. [15] further study the audiovisual suppression effect (ASE) and demonstrate that the ASE is influenced by sound volume, frequency, and cogni-

[1] https://drive.google.com/drive/folders/1rkVElG156xzAwDWLOAOOEhEgXp0hR5cH

tive load, suggesting that these effects can be applied in practical VR applications such as redirected walking.

Existing research shows that visual perception is significantly affected by audio effects. However, no studies have established a perception model yet to evaluate and quantify the impact of audio-visual content in VR scenes on eccentricity-dependent visual perception. To fill this gap, based on the auditory and visual content of VR scenes, we propose an audio-visual feature-driven perception model to quantify the perceptual capabilities of the HVS and provide guidance for optimizing the foveated rendering cost.

3 Auditory-Content-based Perceived Rendering Quality Analysis

In this section, we first discuss the impact of audio on the HVS and how this effect benefits foveated rendering in Section 3.1. Based on the theoretical basis, we construct the supplemental pilot user studies to quantify the differences in the perceived visual quality and the visual-perception attention between auditory and non-auditory conditions. The supplemental pilot user study 1 evaluates the effect of auditory content on visual perception in VR. The experimental results show that viewing scenes with visual-semantic-consistent auditory content enhances perceived visual quality compared to silent scenes in VR (see more details in Supplement Section 1.1). Subsequently, the supplemental pilot user study 2 explores the factors contributing to the perceived visual quality gap between auditory and non-auditory conditions in VR (see more details in Supplement Section 1.2). Based on the theoretical basis and supplementary pilot user studies, in Section 3.2, we conduct the pilot user study to further quantify the impact of audio with different auditory factors on the eccentricity-dependent perceived visual quality.

3.1 Theoretical Basis

In this section, we discuss the impact of audio on the HVS and how this effect benefits foveated rendering. Related researches show that audio affects the visual perception [4, 14, 50]. From an anatomical perspective, the vestibular organs in the HAS regulate position and balance. Audio can affect these vestibular functions, subsequently affecting the visual perception through visual-vestibular integration [1]. From the perspective of scene semantics, the semantic consistency between auditory and visual cues also impacts the HVS's perception [21].

In VR, audio spectra are diverse and often lack distinct primary frequencies and waveform characteristics. It is difficult to directly extract auditory perception features from the audio spectrum to quantify its impact on visual perception. Audio loudness and frequency, as the most direct auditory attributes affecting human perception, are important for understanding how audio affects visual perception [25], which makes them ideal for studying the impact of audio on visual perception in VR. Therefore, to investigate the relationship between audio and visual perception, this paper primarily extracts perceptual features from two fundamental audio properties: loudness and frequency.

3.1.1 The Effect of Audio Loudness on Visual Perception

Audio loudness refers to the subjective perception of audio intensity by the HAS [42], and it can be quantified and standardized through objective measurement methods [30, 54]. Audio loudness can be measured by sound pressure, which represents the changes in volumetric pressure caused by sound disturbances. Direct use of sound pressure results in a wide range of values, which may not accurately reflect the HAS's perception of sound. Current acoustic and medical research often uses the sound pressure level (SPL) to indicate changes in loudness, although it does not perfectly correspond to the loudness perceived by the HAS. SPL is calculated by taking the logarithm of the ratio between the actual and reference sound pressure, typically expressed in decibels (dB). Loudness Units relative to Full Scale ($lufs$) [40] is a unit for measuring audio loudness that better aligns with human perception of loudness compared with dB, which provides improved consistency and standardization.

The loudness of the audio affects the visual perception capability of the HVS. Audio at higher loudness levels can shift visual attention, reducing the HVS's perception of certain regions within the current visual field [34]. The loudness level of audio affects the integration of visual and auditory information, where high-decibel audio can enhance the coherence and consistency of perception, improving the overall perception experience [36]. Further, Ayres and colleagues [1] validate that the sensitivity of retinal rod cells to brightness discrimination decreases at noise levels of $90dB$, leading to prolonged response times to dim light stimuli; at $95dB$, the pupils dilate; at $115dB$, the eyes' adaptability to brightness changes decreases by 20%. Additionally, the sudden onset of audio can affect vestibular function [13], thereby impacting visual sensitivity. Therefore, in foveated rendering, dynamically adjusting the shading rate of the peripheral region based on the loudness level of the audio can achieve higher rendering performance without reducing the perceived visual quality.

3.1.2 The Effect of Audio Frequency on Visual Perception

Audio frequency refers to the number of vibrations per second of a sound wave at a given point, typically measured in Hertz (Hz) [2]. Audio frequency is the primary physical attribute determining the perceived pitch in the HAS; higher frequencies generally correspond to higher perceived pitches, while lower frequencies correspond to lower ones [22]. Existing methods typically employ Fast Fourier Transform (FFT) to measure audio frequencies [28, 51], which involves converting the audio signal from the time domain to the frequency domain. This process generates a spectrum that displays the intensity or amplitude of various frequency components, identifying the peak amplitude as the fundamental frequency of the audio.

Audio frequency affects the visual perception. Van der Burg et al. [47] demonstrate that audio frequency has a cross-modal effect on the visual modality, where audio frequency changes can modulate visual attention distribution, with high frequencies enhancing focus on visual stimuli. Experiments by Hagtvedt et al. [12] indicate a cross-modal link between sound frequency and visual color brightness, where high-frequency sounds shift visual attention to lighter colors, while low-frequency sounds have the opposite effect, thereby affecting visual attention. Furthermore, Fletcher et al. [10] demonstrate that the HAS's sensitivity to different frequencies is uneven; even at the same sound pressure level, the perceived loudness varies across frequencies. The equal-loudness contours quantify that the HAS is more sensitive to mid to high-frequency sounds, particularly between 2 to $5kHz$. Thus, leveraging high-frequency audio to enhance attention concentration and reduce the foveal region in foveated rendering, and adjusting the shading rate of the peripheral region based on frequency-induced loudness perception, merits further investigation for more efficient foveated rendering.

3.2 Pilot User Study: Quantifying Audio Effects on Foveated Rendering

The results of supplemental pilot user studies demonstrate that incorporating semantically relevant auditory content in VR scenes enhances the perceived visual quality of rendering results (Supplement Sections 1.1 and 1.2). In this section, based on the theoretical basis in Section 3.1 and the results of supplemental pilot user studies, we conduct the pilot user study to further quantify the impact of auditory content on foveated rendering in VR. We formulate the hypothesis for the pilot user study:

H1 Viewing scenes with visual-semantic-consistent auditory content within a specific range of loudness and frequency significantly enhances perceived visual quality through foveated rendering in VR.

3.2.1 User Study Design

Setup We use a PICO 4 Pro HMD powered by a workstation with a 3.9Hz Intel® Core™ i9-12900K CPU, 32GB RAM, and an NVIDIA GeForce GTX 3080 Ti graphic card. The resolution of the HMD is 2160×2160 pixels for each eye, and the field-of-view is 105°. The program is developed with C# and is run in Unity 2021.3.13f1. We implement the variable-rate shading pipeline (VRS) [5] in Unity. All scenes are presented to the participants through VRS to achieve foveated rendering effects in HMDs [26].

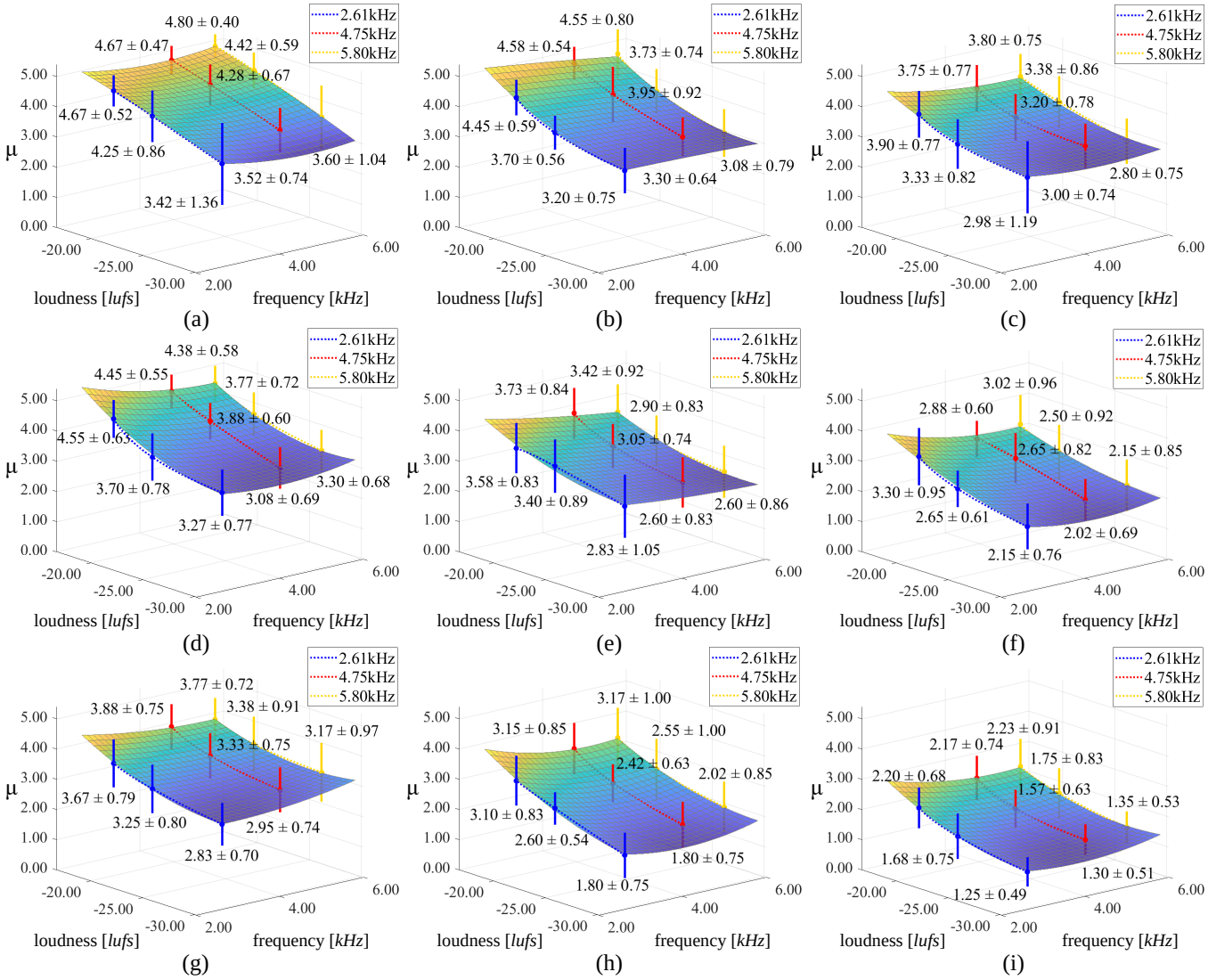


Fig. 2: Aaverage values and standard deviations of μ under different foveated rendering conditions as a function of audio loudness and frequency in the pilot user study.

Dataset The supplementary pilot user study 1 constructs a 360° panoramic video dataset D with auditory content (see more details in Supplement Section 1.1.1). To facilitate the comparison of visual perception under different auditory conditions, we expand D to obtain the expanded dataset D' with different frequencies and loudness, which is detailed in Supplement Section 1.3.

Participants We recruit 20 participants, the same as the supplementary pilot user study 1. Participants include 10 males and 10 females, aged 18 to 50, with an average age of 35. All participants have normal hearing and vision or have corrected vision through glasses. Ten of them have experience using HMD VR applications before the study.

Condition We divide the view of foveated rendering into the foveal, transitional, and peripheral regions, which is the same as [11]. The foveal layer uses the highest shading rate in VRS, the peripheral layer uses the lowest, and the transitional layer uses a linearly declining shading rate from the foveal to the peripheral region. We define the foveated rendering coefficients in the graphics pipeline as $FR(E_f, E_t)$, where E_f is the eccentricity angle at the edge of the foveal region, and E_t is the eccentricity angle at the edge of the transitional region. In the pilot user study, we divide the foveated rendering into three levels FR_{L1}, FR_{L2}, and FR_{L3}, each containing three conditions $C1$, $C2$, and $C3$. E_f is constant within the same level but decreases as the level increases. Within the same level, E_t decreases as the condition sequence progresses. Specifically, for FR_{L1-C1}, FR_{L1-C2}, and FR_{L1-C3}, (E_f, E_t) are set to (14.34°, 27.08°), (14.34°, 23.49°), and (14.34°, 18.39°); for FR_{L2-C1}, FR_{L2-C2}, and FR_{L2-C3}, (E_f, E_t) are set to (12.26°, 27.08°), (12.26°, 23.49°), and (12.26°, 18.39°); for FR_{L3-C1}, FR_{L3-C2}, and FR_{L3-C3}, (E_f, E_t) are set to (9.44°, 27.08°), (9.44°, 23.49°), and (9.44°, 18.39°). (E_f, E_t) of FR_{L1-C1} are set the same as the state-of-the-art VR-HMD foveated rendering method [26].

Procedure Participants are instructed to view one randomly selected scene for each scene type in dataset D'. First, they view the selected scenes using the full-resolution rendering method without auditory content and are informed that these are high-quality rendering versions. Each scene contains auditory content with low, mid, and high frequencies, presented with varying levels of loudness and frequency. The scenes are rendered using foveated rendering coefficients ranging from FR_{L1-C1} to FR_{L3-C3} in random order, generating 81 rendering sequences for each scene. Before viewing the scenes, we visualize the artifacts generated by the 360° panoramic video stitching and instruct participants to ignore any artifacts resulting from the stitching process in the 360° panoramic videos. For each rendering sequence, participants are instructed to view it for 20s, and then provide a corresponding visual perceptual quality score, μ. After completing one sequence, the next one follows. Each participant completes the entire task in two sessions, averaging 30min each, for a total of 60min. A total of 20 (participants) × 2 (scenes) × 9 (different audio loudness and frequencies for each scene) × 9 (foveated rendering conditions) = 3240 trials are collected.

Table 1: Compared with FR_{L1-C1}, the items of (audio loudness, audio frequency, $\bar{\mu}$) under other foveated rendering conditions with no significant differences evaluated in p-value.

Condition	Auditory Factor with no Significant Difference
FR_{L1-C2} vs. FR_{L1-C1}	(-18.54$lufs$, 5.42kHz, 4.65), (-18.32$lufs$, 5.32kHz, 4.63), (-15.20$lufs$, 2.62kHz, 4.58), (-12.56$lufs$, 1.76kHz, 4.52), (-19.46$lufs$, 3.45kHz, 4.55), (-11.63$lufs$, 1.54kHz, 4.55), (-19.98$lufs$, 4.07kHz, 4.45), (-19.35$lufs$, 1.80kHz, 4.48), (-21.14$lufs$, 1.84kHz, 4.45), (-13.56$lufs$, 5.05kHz, 4.25) (-31.46$lufs$, 5.05kHz, 3.98), (-28.28$lufs$, 5.05kHz, 3.95), (-12.13$lufs$, 5.05kHz, 3.89), (-12.98$lufs$, 2.13kHz, 3.30), (-31.55$lufs$, 2.01kHz, 3.29), (-11.07$lufs$, 3.21kHz, 3.20)
FR_{L1-C3} vs. FR_{L1-C1}	(-13.56$lufs$, 5.05kHz, 4.14), (-14.28$lufs$, 5.05kHz, 3.93), (-12.13$lufs$, 5.05kHz, 3.84), (-12.89$lufs$, 2.13kHz, 3.35), (-11.07$lufs$, 3.21kHz, 3.18), (-11.94$lufs$, 4.09kHz, 3.02), (-12.99$lufs$, 3.73kHz, 2.98), (-11.84$lufs$, 4.01kHz, 2.96)
FR_{L2-C1} vs. FR_{L1-C1}	(-19.13$lufs$, 4.76kHz, 4.52), (-18.32$lufs$, 5.32kHz, 4.52), (-19.46$lufs$, 3.45kHz, 4.51), (-15.20$lufs$, 2.62kHz, 4.50), (-19.35$lufs$, 1.80kHz, 4.46), (-21.42$lufs$, 5.21kHz, 4.42), (-28.28$lufs$, 5.05kHz, 4.04), (-31.46$lufs$, 5.05kHz, 3.95), (-12.13$lufs$, 5.05kHz, 3.88), (-22.44$lufs$, 5.80kHz, 3.35), (-29.56$lufs$, 3.51kHz, 3.29)
FR_{L2-C2} vs. FR_{L1-C1}	(-12.56$lufs$, 3.45kHz, 3.24) (-13.86$lufs$, 3.21kHz, 3.16) (-13.31$lufs$, 3.67kHz, 2.90) (-12.89$lufs$, 3.21kHz, 2.88) (-11.84$lufs$, 4.01kHz, 2.85) (-11.94$lufs$, 1.98kHz, 2.83)
FR_{L3-C1} vs. FR_{L1-C1}	(-12.56$lufs$, 3.45kHz, 3.20) (-13.86$lufs$, 3.21kHz, 3.18) (-15.20$lufs$, 2.62kHz, 3.14) (-12.89$lufs$, 3.21kHz, 2.94) (-11.94$lufs$, 1.98kHz, 2.88)

3.2.2 Results and Discussion

Fig. 2 visualizes the average values and standard deviations of the visual perceptual quality score μ across different foveated rendering conditions based on audio loudness and frequency, where (a)-(c) represent FR_{L1-C1} to FR_{L1-C3}, (d)-(f) represent FR_{L2-C1} to FR_{L2-C3}, and (g)-(i) represent FR_{L3-C1} to FR_{L3-C3}, respectively. To ensure the consistency of visual content across different audio loudness and frequency levels, in Fig. 2, we categorize the loudness and frequency into low, medium, and high levels for each foveated rendering condition. We then merge the data of μ based on the factor levels to create nine points for visual perception trend fitting. Since each scene in D' contains low, mid, and high loudness and frequency, this ensures that the visual content for each point in Fig. 2 under each foveated rendering condition is consistent.

Fig. 2 shows that under the same audio loudness and frequency levels, as the eccentricity angle of the foveal region decreases, μ shows a downward trend. Similarly, μ shows a downward trend as the eccentricity angle of the transitional region decreases. Higher audio loudness positively affects μ. In all foveated rendering conditions, as loudness increases from -30$lufs$ to 0$lufs$, μ shows an upward trend. Regarding audio frequency, under more aggressive foveated rendering conditions (e to i), increasing the frequency to higher ranges improves μ when the loudness is low ($\leq$-25$lufs$). Thus, we conclude that audio loudness and frequency within specific ranges enhance perceived visual quality.

We use the ANOVA method to evaluate the impacts of audio loudness and frequency on μ. Significant effects on μ are found for audio loudness ($F_{59,23207} = 38.35$, $p = 1.03 \times 10^{-296}$, $\eta_p^2 = 0.39$) and frequency ($F_{59,23207} = 64.58$, $p = 1.61 \times 10^{-98}$, $\eta_p^2 = 0.14$), with both effect sizes classified as *large*. The ANOVA results indicate that audio loudness and frequency significantly influence perceived visual quality. In conjunction with the conclusions of Fig. 2, **H1** is proven to be valid.

Table 1 shows the items (audio loudness, audio frequency, and average visual perceptual quality score μ) under different foveated rendering conditions where no significant differences in perceived visual quality were found compared to the state-of-the-art foveated rendering method (i.e., FR_{L1-C1}) based on p-values. Results indicate that within specific loudness and frequency ranges, reducing the eccentricity angle of the foveal region from $L1$ to $L2$, or reducing the eccentricity angle of the transitional region from $C1$ to $C2$, yields no significant difference in μ. Further decreasing the eccentricity angles in either the foveal or transitional region still yields no significant difference in p-values. Therefore, in foveated rendering, under specific auditory factors, reducing shading in the foveal or transitional region does not significantly affect perceived visual quality. We believe that by further combining foveated sampling based on visual content, more aggressive foveated rendering optimization can be achieved.

4 Audio-visual aware Foveated Rendering

In this section, we propose the audio-visual aware foveated rendering method (AvFR) to afford foveated rendering cost savings without sacrificing the perceived visual quality in VR scenes containing auditory content. We first introduce the audio-visual feature-driven perception model (AvPM) in AvFR, which predicts the HVS's visual perceptual ability under varying auditory and visual content in real time to guide the foveated rendering cost optimization. Then, we present the foveated rendering cost optimization, which imperceptibly optimizes the shading rate in foveated rendering with the guidance of the AvPM.

Fig. 3 visualizes the pipeline of AvFR. There are three steps in AvFR. Step 1 is the audio-visual feature-driven perception model construction, which constructs a perception model based on the audio-visual content to guide the foveated rendering cost optimization. Step 2 is the foveated rendering cost optimization, which generates a required shading rate map to control the shading rate in the output framebuffer based on the guidance of the audio-visual feature-driven perception model. Step 3 is the VRS-based foveated rendering, which renders the foveated framebuffer in the VRS pipeline with the shading rate defined by the required shading rate map. To demonstrate the details of the steps, we give Algorithm 1.

Given the inputs experimental results *data* in the pilot user study, 3D content S, VR scene auditory content *aud*, current viewpoint V, gaze position *gaze*, length of time for clamping the audio T_a, sampling interval of audio features T_s, full resolution of the output framebuffer (w,h), the fixed size of pixel block (w_b,h_b), and the predefined shading rate array SR, Algorithm 1 outputs the foveated rendering framebuffer fb. AvFR is implemented on VRS [5] due to the excellent compatibility and performance of VRS in the universal rendering pipeline. VRS divides pixels in the output framebuffer into fixed-size pixel blocks, (w_b,h_b) defines the size of each pixel block, and SR is the shading rate array that contains multiple shading rate options for rendering each pixel block in VRS.

In Algorithm 1, we first build the training data set $\mathscr{D}$ based on *data* for the construction of AvPM (lines 1-8). In line 1, we initialize $\mathscr{D}$ as an empty set. Each *item* in *data* includes the scene audio clip *item.audio*, the scene video clip *item.video* and the visual perceptual quality score *item.*μ [32] when the participants view the scene. For each *item* in *data*, we calculate the average auditory perception-loudness feature value $\overline{\gamma_{ld}}$ (line 3) and the average auditory perception-frequency feature value $\overline{\gamma_{fq}}$ (line 4) based on the audio clip in *item*. The calculation details of *audLoudFeature* and *audFreqFeature* are described in Sections 4.1.1.1 and 4.1.1.2. We calculate the average CSF-based visual loss feature $\overline{\gamma_{vis}}$ based on the video clip in *item* in line 5. The details of *csfLossFeature* are described in Section 4.1.1.3. Then we merge {[$\overline{\gamma_{ld}}$, $\overline{\gamma_{fq}}$, $\overline{\gamma_{vis}}$], *item.*μ} as an entry and add it into $\mathscr{D}$ (line 6). After iterating through all items in *data* (line 7), we merge the entries in $\mathscr{D}$ with identical auditory and visual features, with this entry's $\overline{\mu}$ in $\mathscr{D}$ being the average μ of these entries with identical auditory and visual features (line 8). Then, we construct AvPM to fit $\overline{\mu}$ with [$\overline{\gamma_{ld}}$, $\overline{\gamma_{fq}}$, $\overline{\gamma_{vis}}$] in $\mathscr{D}$ (line 9), the details are demonstrated in Section 4.1.2.

We perform foveated rendering with the guidance of AvPM. We initialize the auditory perception-loudness feature value γ_{ld}, the auditory perception-frequency feature value γ_{fq}, and the CSF-based visual loss feature γ_{vis} as 0 (line 10). Then, we perform the rendering loop (lines 11-19). In the rendering loop, we first capture the audio clip $aud^t_{t-T_a}$ within the scene auditory content *aud* at the time period T_a (line 12). Due to the temporal continuity of $aud^t_{t-T_a}$, we update audio perceptual features γ_{ld} and γ_{fq} with a period of T_s (lines 13-15), which can enhance algorithm performance while maintaining the accuracy of audio perceptual features. Then, we calculate the CSF-based visual loss feature γ_{vis} based on the 3D content S, current viewpoint V, and gaze position *gaze* (line 16). We perform the foveated rendering cost optimization algorithm $FRCO$ to generate the required shading rate

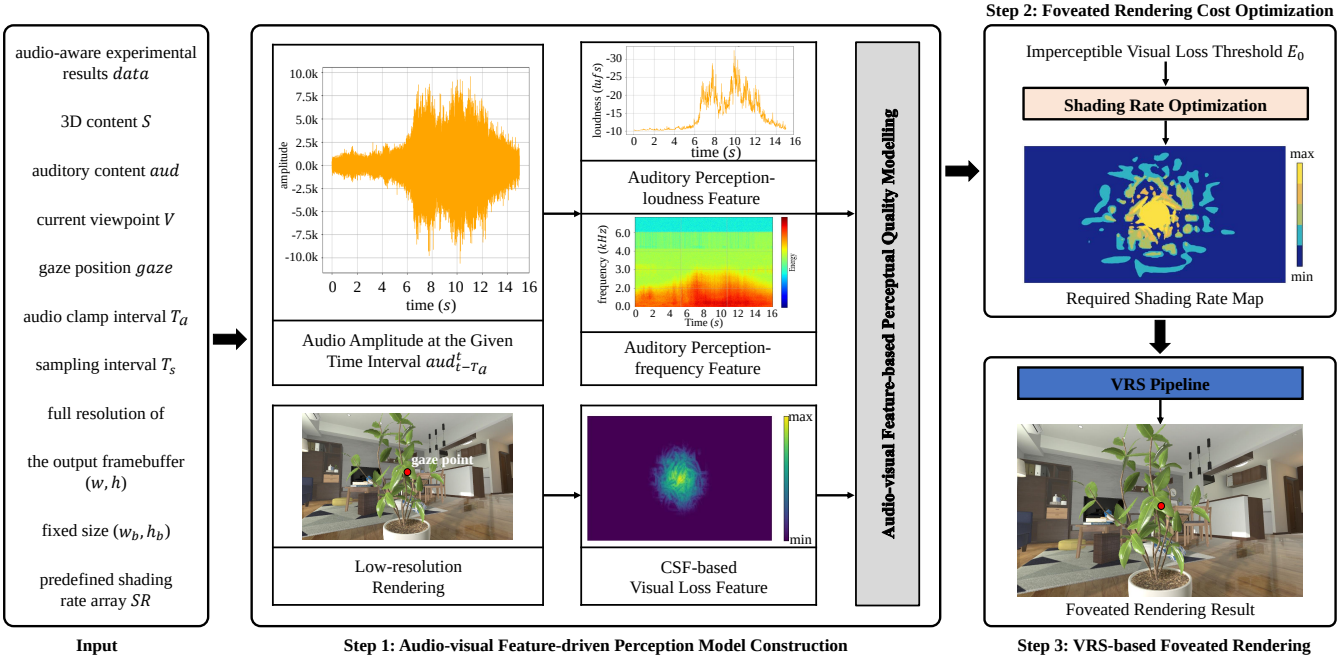


Fig. 3: Pipeline of the Audio-visual aware Foveated Rendering Method

map map_{sr} with the guidance of *AvPM* (line 17). The details of *FRCO* are described in Section 4.2. Finally, the output foveated rendering framebuffer fb for the current viewpoint V is rendered by VRS, and shading rates of all pixel blocks in fb are determined by SR (line 18).

4.1 Audio-visual Feature-driven Perception Model

The theoretical basis in Section 3 and pilot user studies validate that the loudness and frequency of auditory content have effects on the HVS's perception in VR. Many studies show that the contrast is the main factor that significantly affects the perceptual ability of the HVS due to the photosensitivity of ganglion cells [18, 46, 48]. Therefore, foveated rendering can be further optimized based on the above characteristics of the HVS according to the audio-visual content in VR. However, existing methods cannot leverage the characteristics of the HVS in VR scenes containing auditory content, leading to limited improvement in foveated rendering performance. The proposed AvPM provides guidance for minimizing foveated rendering cost by modeling the HVS's visual perceptual ability based on the audio-visual content. AvPM has two steps: the first step is the auditory and visual features extraction, which are related to the HVS's visual perceptual ability; the second step is the audio-visual feature-based perceptual quality modeling, which predicts the HVS's visual perceptual ability based on the extracted auditory and visual features.

4.1.1 Auditory and Visual Features Extraction

Section 3.1 demonstrates that the loudness and frequency of auditory content affect the HVS's perception. Many foveated rendering studies show that the effect of visual content on the HVS's perception mainly depends on contrast sensitivity [18, 31]. Therefore, in the auditory feature extraction, we extract the **auditory perception-loudness feature** and the **auditory perception-frequency feature** based on the auditory content. In visual feature extraction, we extract the **CSF-based visual loss feature** in foveated rendering. The auditory perception-loudness feature measures the users' perceptual ability to the audio loudness in the current VR scene, where a higher feature value indicates a stronger perception of the audio loudness by the HAS. The auditory perception-frequency feature measures the users' perceptual ability to the audio frequency in the current VR scene, where a higher feature value indicates a stronger perception of the audio frequency. The CSF-based visual loss feature measures the contrast loss perceivable by the HVS in foveated rendering compared with the full-resolution rendering in the current VR scene, where a higher feature value indicates a stronger visual loss perceivable by the HVS.

4.1.1.1 Auditory Perception-loudness Feature Extraction

In the auditory perception-loudness feature extraction, since the HAS's perception of audio loudness is not linear, the common sound pressure level units (dB) cannot accurately model the human perception of audio loudness well [40]. The ITU-R BS.1770-4 recommendation [37] has been shown to correlate well with the perceived loudness of the HAS, which is an efficient algorithm consisting of frequency weighting filters and gated energy measurements. Due to the effectiveness of the ITU-R BS.1770-4 recommendation in modeling the perceived loudness of the HAS and its high computational efficiency, we use it to extract the auditory perception-loudness feature γ_{ld}. Given the auditory content $aud^t_{t-T_a}$ at the current time t from the past period T_a, the calculation of γ_{ld} is shown in Equation 1.

$$\gamma_{ld} = clamp(itu(aud^t_{t-T_a}), ITU_{min}, ITU_{max}) - ITU_{min} \quad (1)$$

where $clamp(x, val_{min}, val_{max})$ clamps x within a range of values between val_{min} and val_{max}; $itu(aud)$ calculates the loudness units relative to full scale ($lufs$) [33] with the given audio aud by the ITU-R BS.1770-4 recommendation; ITU_{min} and ITU_{max} are the minimum and maximum thresholds in the ITU-R BS.1770-4 recommendation. Equation 1 avoids the negative infinite result when the sound pressure level of $aud^t_{t-T_a}$ is $0dB$, and converts the perception-loudness features in the range of $[0, ITU_{max}\text{-}ITU_{min}]$.

4.1.1.2 Auditory Perception-frequency Feature Extraction

In the auditory perception-frequency feature extraction, since the spectrum of auditory content in VR at any moment contains numerous frequencies with various intensities, and the relationship between the HAS's perception and the audio frequency is non-linear, it is difficult to effectively express the changes in the HVS's perception at different frequencies by directly representing frequency in Hertz (Hz) [54]. Moreover, the HAS's perception of frequencies varies across different frequency bands [54], making it challenging to effectively represent the perceived frequency by simply summing all frequencies within the spectrum based on intensity weighting. Thus, to effectively extract the auditory perception-frequency feature, two issues need to be solved:

Algorithm 1: Audio-visual aware Foveated Rendering

input : audio-aware experimental results $data$, 3D content S, auditory content aud, current viewpoint V, gaze position $gaze$, audio clamp interval T_a, sampling interval T_s, full resolution of the output framebuffer (w,h), fixed size of pixel block (w_b,h_b), predefined shading rate array SR

output : output framebuffer fb

$\mathscr{D} \leftarrow \emptyset$
for $item \in data$ **do**
  $\overline{\gamma_{ld}} \leftarrow$ audLoudFeature($item.audio$)
  $\overline{\gamma_{fq}} \leftarrow$ audFreqFeature($item.audio$)
  $\overline{\gamma_{vis}} \leftarrow$ csfLossFeature($item.video$)
  $\mathscr{D} \leftarrow \mathscr{D} \cup (\{[\overline{\gamma_{ld}},\overline{\gamma_{fq}},\overline{\gamma_{vis}}], item.\mu\})$
end
$\mathscr{D} \leftarrow$ avgScore($\mathscr{D}$)
$AvPM \leftarrow$ constructModel($\mathscr{D}$)
$\gamma_{ld}, \gamma_{fq}, \gamma_{vis} \leftarrow 0,0,0$
for *time t in rendering loop* **do**
  $aud^t_{t-T_a} \leftarrow$ capAudio(aud)
  if $t\%T_s$ *is 0* **then**
    $\gamma_{ld} \leftarrow$ audLoudFeature($aud^t_{t-T_a}$)
    $\gamma_{fq} \leftarrow$ audFreqFeature($aud^t_{t-T_a}$)
  end
  $\overline{\gamma_{vis}} \leftarrow$ csfLossFeature(S, V, $gaze$)
  $map_{sr} \leftarrow$ FRCO(w, h, w_b, h_b, SR, γ_{ld},γ_{fq},γ_{vis}, $AvPM$)
  $fb \leftarrow$ VRS(S, V, map_{sr})
end

Issue 1: What scale unit is used to express the audio frequency?

Issue 2: How to integrate the frequency components in various frequency bands to effectively extract the auditory perception-frequency feature?

For **Issue 1**, we use the Mel scale to express the perceptual audio frequency, which has shown excellent ability in modeling frequency perception [41]. Given the current frequency fq in Hz of the auditory content, the Mel scale mel is calculated by Equation 2:

$$mel = 2595lg(1+\frac{fq}{700}) \tag{2}$$

Due to the fact that the basilar membrane of the cochlea has 24 points where the maximum resonance occurs at 24 different frequencies, the audible frequency range for the HAS is divided into 24 critical bands [55]. Audios that arrive at the same critical band of the cochlea within a short period of time are difficult to distinguish due to auditory masking. This results in the perception of frequency by the HAS having the characteristic of frequency band division. We divide the Mel scale into multiple overlapping frequency bands, and use the granularity of bands to filter and synthesize the intensity of various frequencies of audio, which is more in line with the auditory characteristics of the HAS. Therefore, for **Issue 2**, we extract the auditory perception-frequency feature by performing a weighted sum based on frequency intensity on the frequencies within 24 specific critical bands. Specifically, we utilize a filter set FB to compute the audio's Filter Banks (FBanks) perceived by the HAS. FBanks represents the intensity level of the audio in various frequency critical bands. Then, we perform a weighted sum of the center frequencies in critical bands using the computed FBanks as weights to obtain the auditory perception-frequency feature γ_{fq}, as shown in Equation 3:

$$\gamma_{fq} = \sum_{i=1}^{24} \frac{FB[i]}{\sum_{m=1}^{24} FB[m]} \cdot \overline{mel_i} \tag{3}$$

where $\overline{mel_i}$ is the center frequency in the i-th critical frequency band, as shown in Equation 4:

$$\overline{mel_i} = \begin{cases} minMel & i = 1 \\ \frac{i-1}{24}(maxMel - minMel) + minMel & 1 < i < 24 \\ maxMel & i = 24 \end{cases} \tag{4}$$

where $minMel$, $maxMel$ are the minimum and maximum frequencies of the audio. $FB[i]$ is the calculated FBanks of FB for the i-th critical frequency band, the calculation is shown in Equation 5:

$$FB[i] = \sum_{mel \in [minMel, maxMel]} P(mel) \cdot H_i(mel) \tag{5}$$

where $P(mel)$ is the intensity spectrum calculation based on Hamming window and Fourier transformer, similar to Mel Cepstral Coefficients (MFCC) [6]; $H_i(mel)$ is the triangular filtering function for the i-th filter in FB, as shown in Equation 6:

$$H_i(mel) = \begin{cases} 0 & mel < \overline{mel_{i-1}} \\ \frac{mel - mel(i-1)}{mel(i) - mel(i-1)} & \overline{mel_{i-1}} \le mel < \overline{mel_i} \\ 1 & mel = \overline{mel_i} \\ \frac{mel(i+1) - mel}{mel(i+1) - mel(i)} & \overline{mel_i} < mel \le \overline{mel_{i+1}} \\ 0 & mel > \overline{mel_{i+1}} \end{cases} \tag{6}$$

4.1.1.3 CSF-based Visual Loss Feature Extraction

In the CSF-based visual loss feature extraction, since luminance contrast sensitivity can efficiently model the visual perceptual capability of the HVS [8, 49], we use the luminance contrast sensitivity function to quantify the visual perceptual loss of the foveated rendering. The CSF-based visual loss feature extraction includes three processes. In process 1, we calculate the eccentricity-based luminance contrast sensitivity of each pixel. In process 2, we quantify the overall perceived visual quality of the current rendering result by a weighted sum of the contrast sensitivity values of all pixels. In process 3, we calculate the difference in overall contrast sensitivity between the full resolution and foveated rendering results to obtain the CSF-based visual loss feature.

In process 1, for each pixel p, the frequency-based luminance contrast sensitivity is calculated by Equation 7 [46]:

$$C_n(f,p) = C(f,p) S_{csf}(f, \theta(p), L_a(f,p)) \tag{7}$$

where f is the spatial frequency in cpd units (cycles-per-visual-degree); $C(f,p)$ is the luminance contrast pyramid decomposition result of the given frequency f at p; $\theta(p)$ is the retinal eccentricity of p; $L_a(f,p)$ models the adaptation luminance based on the Laplacian pyramid decomposition [3]; S_{csf} is the contrast sensitivity function proposed in [27]. Then, the eccentricity-based luminance contrast sensitivity $C_n(p)$ of p is obtained by accumulating the frequency-based luminance contrast sensitivity of all peak frequencies in the frequency band set $\mathscr{B}$ proposed in [46], as shown in Equation 8:

$$C_n(p) = \sum_{f \in \mathscr{B}} C_n(f,p) \tag{8}$$

According to [46], for a pixel p with a shading rate of sr, its peak frequency f will decrease to $\sqrt{sr} \cdot f$ in frequency bands contained in $\mathscr{B}$. Thus, the eccentricity-based luminance contrast sensitivity $C_n(p,sr)$ of p with the shading rate $p.sr$ is calculated by Equation 9:

$$C_n(p,sr) = \sum_{f \in \mathscr{B}} C_n(\sqrt{p.sr} \cdot f, p) \tag{9}$$

In process 2, the perceived visual quality $C_v(fb)$ of the foveated rendering result fb is qualified by performing a weighted sum of the

luminance contrast sensitivity of all pixels p in fb, as shown in Equation 10:

$$C_v(fb) = \sum_{p \in fb} w(p) C_n(p) \tag{10}$$

where $w(p)$ is the pixel weight; $w(p)$ is set to $\frac{1}{w \cdot h}$; (w,h) are the width and height of the output framebuffer; $\sum_{p \in fb} w(p) = 1$.

In process 3, the CSF-based visual loss feature γ_{vis} of the foveated rendering result fb is obtained by the difference between the perceptual quality in the full resolution $C_v(gt)$ and that in the foveated rendering $C_v(fb)$, as shown in Equation 11:

$$\gamma_{vis} = C_v(gt) - C_v(fb) \tag{11}$$

4.1.2 Audio-visual Feature-based Perceptual Quality Modelling

We jointly model the perceptual quality based on auditory and visual features. Specifically, given the auditory perception-loudness feature γ_{ld}, the auditory perception-frequency feature γ_{fq}, and the CSF-based visual loss feature γ_{vis}, we construct the model $AvPM(\gamma_{ld}, \gamma_{fq}, \gamma_{vis})$ to predict the visual perceptual quality score μ of the foveated rendering result.

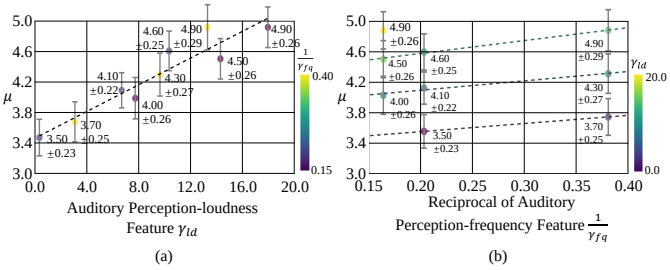


Fig. 4: Plots of μ as a function of the auditory perception loudness feature (a) and the reciprocal of the auditory perception frequency feature (b).

The experimental results of the pilot user study demonstrate that both loudness and frequency significantly affect perceived visual quality in foveated rendering. To describe the relationship between the visual perceptual quality score μ and audio loudness and frequency, we conduct fitting analyses based on audio loudness and frequency for the experimental results. Specifically, we combine the data in the pilot user study across all foveated rendering conditions to form a merged dataset. Each entry in this dataset includes audio loudness, frequency, and the corresponding μ. We visualize μ as a function of auditory features in Fig. 4 based on the merged dataset. In Fig. 4 (a), we convert audio loudness $lufs$ into the auditory perception-loudness feature γ_{ld} on the x-axis. We find an approximately positive linear relationship between γ_{ld} and μ. This indicates that auditory perception-loudness affects visual perception; higher perception-loudness audio reduces visual sensitivity, allowing users to accept lower rendering quality without a decrease in perceived quality, which is consistent with previous research findings [34, 36]. In Fig. 4 (b), we convert audio frequency kHz to the reciprocal of the auditory perception-frequency feature $\frac{1}{\gamma_{fq}}$ on the x-axis. Based on different audio loudness conditions, we fit μ according to $\frac{1}{\gamma_{fq}}$. As shown in Fig. 4 (b), when the auditory perception-loudness feature value is essentially consistent, $\frac{1}{\gamma_{fq}}$ and μ exhibit an approximately positive correlation. Experimental data demonstrate that as perception frequency increases, visual perception declines. Previous research indicates that the loudness perceived by the HAS varies with frequency, showing greater sensitivity to high frequencies in the 2-5kHz range [10]. Therefore, we can utilize high-perception-frequency audio to dynamically adjust the shading rate, achieving more aggressive foveated rendering without compromising visual fidelity.

In terms of the visual feature, the higher the value of the CSF-based visual loss feature, the more likely the user perceives the visual loss, resulting in lower μ. According to [11, 18, 26], the finding is that there is a negative relationship between the CSF-based visual loss feature and μ. Based on the above findings, we derive four rules to instruct the model construction:

Rule 1: γ_{ld} is linearly positively correlated with μ;

Rule 2: $\frac{1}{\gamma_{fq}}$ is linearly positively correlated with μ;

Rule 3: γ_{vis} is negatively correlated with μ;

Rule 4: all features should be normalized to the same scale to facilitate the fitting of $AvPM$.

According to the above rules, we adopt the most straightforward polynomial to fit the visual perceptual quality score μ based on the auditory and visual features, as shown in Equation 12:

$$AvPM(\gamma_{ld}, \gamma_{fq}, \gamma_{vis}) = a \cdot \gamma_{ld} + \frac{b}{\gamma_{fq}} - c \cdot \frac{\gamma_{vis}}{10000} + d \tag{12}$$

where γ_{ld}, $\frac{1}{\gamma_{fq}}$, and $\frac{\gamma_{vis}}{10000}$ are in the same scale; a, b, c, d are coefficients to be fitted.

Given that the training data set $\mathscr{D}$ contains n entries, Equation 13 describes the optimization of $AvPM$ according to Rules 1-4:

$$\begin{aligned} minimize \quad & \frac{1}{2}||X \cdot A - H|| \\ s.t. \quad & a > 0, b > 0, c > 0, d > 0 \\ & A = [a, b, c, d]^T \\ & X = \begin{bmatrix} \overline{\gamma_{ld}}^1 & \overline{\frac{1}{\gamma_{fq}}}^1 & \overline{\frac{\gamma_{vis}}{10000}}^1 & 1 \\ \dots & \dots & \dots & 1 \\ \overline{\gamma_{ld}}^n & \overline{\frac{1}{\gamma_{fq}}}^n & \overline{\frac{\gamma_{vis}}{10000}}^n & 1 \end{bmatrix} \\ & H = \begin{bmatrix} \overline{\mu}^1 \\ \dots \\ \overline{\mu}^n \end{bmatrix} \end{aligned} \tag{13}$$

where A is the coefficient matrix, X is the feature matrix, and H is the score matrix. Each row in X represents the average feature values $[\overline{\gamma_{fq}}, \overline{\gamma_{ld}}, \overline{\frac{\gamma_{vis}}{10000}}]$ contained in the corresponding entry in $\mathscr{D}$. We add a constant 1 to the last column of each row in X to facilitate the dot product with A. Each row in H represents the average visual perceptual quality score $\overline{\mu}$ of the corresponding entry in $\mathscr{D}$. At last, we employ the constrained least square method [35] to optimize Equation 13, yielding the optimized model $AvPM$, as shown in Equation 14:

$$AvPM(\gamma_{ld}, \gamma_{fq}, \gamma_{vis}) = 0.05\gamma_{ld} + \frac{0.45}{\gamma_{fq}} - 0.19\frac{\gamma_{vis}}{10000} + 3.63 \tag{14}$$

where the coefficient of determination R^2 is 0.80 for the fitted results of $AvPM$ on the training set $\mathscr{D}$.

4.2 Foveated Rendering Cost Optimization

According to the guidance of AvPM, we propose the foveated rendering cost optimization algorithm (FRCO) to further enhance the foveated rendering performance without perceived visual loss. FRCO aims to minimize the shading rate for each pixel block in the output framebuffer while ensuring no significant perceived visual loss compared with the state-of-the-art VR-HMD foveated rendering method. Specifically, given the full resolution of the output framebuffer (w,h), the fixed size of pixel block (w_b, h_b), the predefined shading rate array SR, and the audio-visual feature-driven perception model $AvPM$, FRCO outputs the required shading rate map map_{sr} to define the shading rate of each pixel block in the output framebuffer, as shown in Algorithm 2. Since VRS's minimum fixed-size block is 16×16. the predefined shading rate provided by VRS contains limited options, including $\frac{1}{4} \times \frac{1}{4}$, $\frac{1}{2} \times \frac{1}{4}$, $\frac{1}{2} \times \frac{1}{2}$, and 1×1. Thus, (w_b, h_b) is set to 16×16, SR is an array formed by sorting predefined shading rate values in VRS according to ascending order.

Algorithm 2: Foveated Rendering Cost Optimization

input : full resolution of the output framebuffer (w,h), fixed size of pixel block (w_b,h_b), predefined shading rate array SR, auditory perception-loudness feature γ_{ld}, auditory perception-frequency feature γ_{fq}, the CSF-based visual loss feature γ_{vis}, audio-visual feature-driven perception model $AvPM$

output : required shading rate map map_{sr}

1 $B \leftarrow$ genPxBlocks($[\frac{w}{w_b},\frac{h}{h_b}]$)
2 $map_{sr} \leftarrow$ initSRmap($[\frac{w}{w_b},\frac{h}{h_b}]$, max(SR))
3 $E_0 \leftarrow$ impVisualLossThr($SvPM$)
4 **for** $b \in B$ **do**
5 **for** $sr \in SR$ **do**
6 **if** *isMinSR(b, sr, E_0) is True* **then**
7 $map_{sr}[b] \leftarrow sr$
8 break
9 **else**
10 continue
11 **end**
12 **end**
13 **end**
14 **return** map_{sr}

In Algorithm 2, we divide all pixels in the output framebuffer into pixel block map B based on the full resolution of the output framebuffer (w,h) and fixed size of pixel block (w_b,h_b), and each pixel block b in B contains $w_b \times h_b$ pixels (line 1). We initialize the required shading rate map map_{sr} in line 2, with the dimensions identical to those of B. Each value in map_{sr} defines the corresponding shading rate for pixel blocks within B, and each value is initialized to the maximum shading rate in SR, i.e., 1×1.

Then, we calculate the imperceptible visual loss threshold E_0 (line 3). When the perceived visual quality of the current foveated rendering result qualified by $AvPM$ is greater than or equal to this value, the perceived visual quality of the current foveated rendering is comparable to that of the state-of-the-art VR-HMD foveated rendering method. According to the optimized result of the model $AvPM$ in Equation 14, when γ_{ld}, $\frac{1}{\gamma_{fq}}$ and γ_{vis} are approaching 0, the predicted visual perceptual quality score μ is 3.63. It means that when there is no effect of auditory features in VR scenes, the CSF-based visual loss is 0 compared with the state-of-the-art VR-HMD foveated rendering when the visual perceptual quality score is 3.63. Thus, E_0 is set to 3.63.

In lines 4-13, we aim to minimize the shading rate sr of every pixel block b in B while ensuring no significant perceived visual loss compared with the state-of-the-art VR-HMD foveated rendering method, which can be formulated by Equation 15:

$$\begin{aligned} &minimize\ \ \Lambda = \sum_{sr \in map_{sr}} sr \\ &s.t.\ \ AvPM(\gamma_{ld},\gamma_{fq},\gamma_{vis}) \geq E_0 \end{aligned} \tag{15}$$

where Λ is the cumulative sum of the shading rate in map_{sr}.

According to Equations 11 and 14, $AvPM$ can be formulated as Equation 16:

$$AvPM = 0.05\gamma_{ld} + \frac{0.45}{\gamma_{fq}} - 0.19\frac{C_v(gt) - C_v(fb)}{10000} + 3.63 \tag{16}$$

Therefore, we derive the constraint inequality in Equation 15 and obtain Equation 17:

$$C_v(gt) - C_v(fb) \leq \frac{10000}{0.19}(0.05\gamma_{ld} + \frac{0.45}{\gamma_{fq}} + 3.63 - E_0) \tag{17}$$

For any moment in VR scenes, auditory features γ_{ld} and γ_{fq} are independent of shading rates and can be computed before shading. Therefore, the right-hand side of the above inequality is a constant at any moment, which we simplify to a constant α, i.e., $\alpha = \frac{10000}{0.19}(0.05\gamma_{ld} + \frac{0.45}{\gamma_{fq}} + 3.63 - E_0)$. Equation 17 can be formulated as Equation 18:

$$\sum_{b \in B} w(b)(C_v(gt[b]) - C_v([b])) \leq \sum_{b \in B} w(b)\alpha \tag{18}$$

where B is the pixel block map. For optimizing the shading rate of each pixel block b in B, we adopt a sufficient solution of the above inequality, wherein the difference in luminance contrast sensitivity for b is less than the constant α. This aims to conservatively select the shading rate for b without reducing visual loss, as shown in Equation 19:

$$\sum_{f \in \mathscr{B}} C_n(\sqrt{b.sr} \cdot f, b) - C_n(f,b) \leq \alpha \tag{19}$$

where f is the peak frequency of the corresponding frequency band in $\mathscr{B}$, and $b.sr$ is the shading rate of b. Since $\sum_{f \in \mathscr{B}} C_n(f,b)$ is computed in the CSF-based visual loss feature extraction, which can be considered as a constant in Equation 19. We set $\beta = \alpha + \sum_{f \in \mathscr{B}} C_n(f,b)$, the function $isMinSR$ in line 6 can be formulated by Equation 20:

$$\begin{aligned} &minimize\ \ sr \\ &s.t. \sum_{f \in \mathscr{B}} C_n(\sqrt{sr} \cdot f, b) \leq \beta \end{aligned} \tag{20}$$

We traverse the shading rate options in SR in ascending order to find the minimum shading rate sr that satisfies Equation 20 and set map_{sr}[b]=sr, which means sr is the shading rate of b (lines 5-12). After the shading rates of all pixel blocks in B are determined (line 13), we return the required shading rate map map_{sr} (line 14). Finally, VRS renders the output foveated rendering framebuffer based on the shading rate of all pixels determined by map_{sr}.

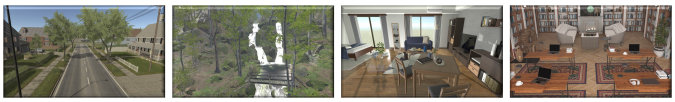


Fig. 5: Visualization of all tested VR scenes for the main user study.

5 Evaluation

In this section, we evaluate the perceived visual rendering quality and performance of our method (AvFR). In Section 5.1, we give the implementation details of this evaluation. In Section 5.2, we conduct a user study to evaluate the perceptual fidelity between AvFR and the state-of-the-art VR-HMD foveated rendering method [26]. Then, in Section 5.3, we measure the performance improvement by further analyzing time savings in each step of the rendering pipeline. To validate the effectiveness of AvFR, we also ablate different audio features used to optimize foveated rendering in Supplement Section 2.2.

We formulate two hypotheses for the evaluation:

H2 AvFR achieves a perceived visual quality similar to the state-of-the-art VR-HMD foveated rendering method with significant performance improvement in VR scenes with auditory content.

H3 AvFR significantly improves the perceived visual quality compared with state-of-the-art VR-HMD foveated rendering method with a similar rendering performance in VR scenes with auditory content.

5.1 Implementation

To evaluate AvFR in VR, we construct four VR scenes containing auditory content: *street*, *forest*, *library*, and *room*, as shown in Fig. 5. And the consistency of the auditory content with the semantics of the visual content is proven valid. We apply a reuse strategy to accelerate AvFR. The details of test scenes and reuse strategy are demonstrated in Supplement Section 2.1.

5.2 Main User Study

Motivated by the experiments of Patney et al. [26], we conduct a psychophysical user study to measure the perceived visual rendering quality degradation experienced by the participants during the free exploration of VR scenes. "Free exploration" is notably a condition where the participants can freely roam and rotate their heads/eyes to naturally investigate immersive scenes.

5.2.1 User Study Design

Setup The hardware setup and the program operating environment of this study are the same as those of the pilot user study.

Table 2: Post-hoc analysis of between $AvFR$ and other conditions for the percentage of the observed artifacts.

measure	Comparisons		mean dif.	std. dif.	p
Bonferroni	$AvFR$	FR	-0.01	0.02	1.00
		FR'	-0.39	0.02	3.25×10^{-54}
Tukey	$AvFR$	FR	-0.01	0.02	0.78
		FR'	-0.39	0.02	5.10×10^{-9}

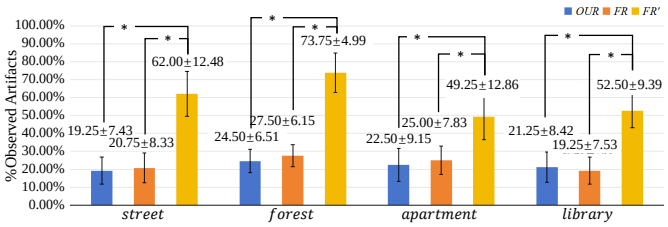


Fig. 6: The percentage of trials where the participants identify artifacts in all tested VR scenes. Asterisks indicate significant differences.

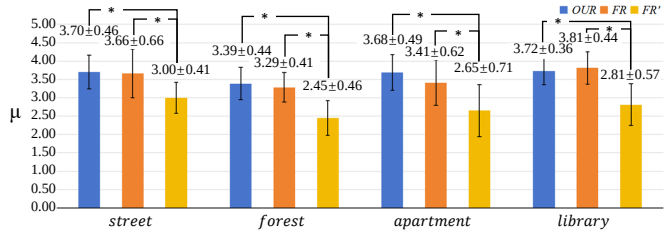


Fig. 7: Average values and standard deviations of μ in user study under $AvFR$, FR, and FR'. Asterisks indicate significant differences.

Participants We recruit 25 participants, consisting of 13 males and 12 females, aged between 18 and 50, with an average age of 31. None of the participants have taken part in pilot user studies, and 11 of them have prior experience using VR HMDs. All participants have normal hearing and vision or have their vision corrected to normal levels through glasses.

Conditions The methods used to render VR scenes include the audio-visual foveated rendering method ($AvFR$), the state-of-the-art VR-HMD foveated rendering method (FR) [26], and the state-of-the-art VR-HMD foveated rendering method with a similar average rendering cost as $AvFR$ (FR').

Procedure All participants are instructed to freely explore the four scenes shown in Fig. 5 using three different rendering methods. Each exploration lasts for 90s. In the scene exploration, participants' initial positions and segments of auditory content are fixed, with the audio starting to play simultaneously as they begin their exploration. Since the audio length far exceeds the exploration time, this design helps prevent participants from feeling constrained by the exploration duration, allowing them to remain immersed in the experience even as the audio nears its end. After each exploration trial, participants are required to answer a two-alternative forced-choice question: "Do you notice any visual artifacts?" They then provide a visual perceptual quality score μ for their exploration. The order of trials is randomized. Since the auditory content duration for all scenes exceeds 90s, there are no unnatural audio interruptions or switches during the exploration process. Each experiment consists of 12 trials, and each participant takes an average of 21min to complete all trials. A total of 25 (participants) $\times$ 4 (scenes) $\times$ 3 (conditions) = 300 trials are collected.

5.2.2 Results and Discussion

We conduct an ANOVA analysis to evaluate the perceived visual quality differences among $AvFR$, FR, and FR'. Additionally, we perform Tukey and Bonferroni post-hoc analyses to examine individual differences between $AvFR$ and both FR and FR'. The following are the results of these analyses.

For the two-alternative forced-choice question, we use the proportion of trials in which participants notice artifacts as one of the perceptual quality metrics. Lower values indicate better perceived visual quality, meaning less noticeable visual modulation. Fig. 6 plots the user-reported values for each scene under each condition. As shown in the figure, the average percentage of observed artifacts is 21.88±7.66 under $AvFR$, 23.13±7.46 under FR, and 59.38±9.93 under FR'. The effect test for the three conditions regarding observed artifacts yields ($F_{2,237} = 273.79$, $p = 2.47\times10^{-62}$, $\eta_p^2 = 0.70$), indicating a significant difference among the three conditions. Table 2 presents the post-hoc statistical results comparing $AvFR$ with the other two conditions using both the Bonferroni and Tukey methods. The p-values from both methods show no significant difference in observed artifacts when comparing $AvFR$ with FR. However, the percentage of artifacts observed under $AvFR$ is significantly lower than that of FR'.

Table 3: Post-hoc analysis of between $AvFR$ and other conditions for μ.

measure	Comparisons		mean dif.	std. dif.	p
Bonferroni	$AvFR$	FR	0.03	0.09	1.00
		FR'	0.90	0.09	1.04×10^{-17}
Tukey	$AvFR$	FR	0.03	0.09	0.93
		FR'	0.90	0.09	5.10×10^{-9}

Fig. 7 gives the average values and standard deviations of the visual perceptual quality score μ when exploring four tested VR scenes under three conditions: $AvFR$, FR, and FR'. Fig. 7 presents the average values and standard deviations of the visual perceptual quality score μ for the exploration of four tested VR scenes under three conditions: $AvFR$, FR, and FR'. The effect test for the three conditions in μ yields ($F_{2,237} = 57.31$, $p = 4.97\times10^{-21}$, $\eta_p^2 = 0.33$), indicating a significant difference among the three conditions in μ. Consistent with the trends observed in the two alternative forced-choice statistics, the average μ for $AvFR$ in *street*, *forest*, and *apartment* is higher than that of FR. However, in the *library*, the average μ for $AvFR$ is slightly lower than that for FR. Participants reported that the lighting in *library* is dim, and slight artifacts appear at the edges of some books during exploration. This is attributed to the reduced accuracy of the CSF-based visual loss feature extraction, which relies on luminance-CSF in low-light scenes, thereby weakening the guiding effect of the audio-visual feature-driven perception model. Consequently, certain artifacts emerge during the rendering process, affecting perceived visual quality. Table 3 presents the post-hoc statistical results comparing $AvFR$ with the other two conditions for μ, using both the Bonferroni and Tukey methods. Compared with FR, the p-values from both methods indicate no significant difference in μ. Therefore, we get **Conclusion 1**: $AvFR$ achieves a perceived visual quality similar to that of the state-of-the-art VR-HMD foveated rendering method.

The average μ of $AvFR$ is significantly higher than that of FR' across all tested scenes. According to the participants' feedback, the reduction in the shading rate of FR' in periphery leads to a noticeable decrease in the rendering quality. According to the statistical results of Bonferroni

(a) Full-resolution Rendering (b) $AvFR$ (c) FR (d) FR'

Fig. 8: Rendering results comparison among the full-resolution rendering, $AvFR$, FR, and FR' in *apartment*.

and Tukey in Table 3, $AvFR$ achieves significant improvement in μ when compared with FR'. Thus, we arrive at **Conclusion 2**: AvFR significantly improves the perceived visual quality compared with the state-of-the-art VR-HMD foveated rendering method with a similar average rendering cost in VR scenes with auditory content. Therefore, the results support **H3**.

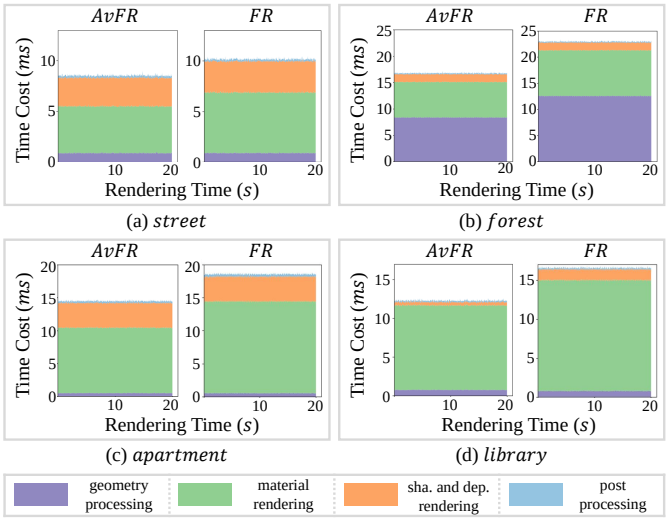


Fig. 9: Time cost plots of $AvFR$ and FR in four tested VR scenes.

To further evaluate the rendering quality of $AvFR$, Fig. 8 compares $AvFR$, FR, and FR' with the full-resolution rendering result in *apartment*. In Fig. 8, the shading rate maps for $AvFR$, FR, and FR' are visualized in the bottom right corner of the rendering results. Green and red boxes highlight the salient regions in the periphery, which are magnified on the right side of each rendering result. Compared to FR, $AvFR$ effectively preserves the geometric contours of the books and enhances the salient glossy rendering effects in regions near the fovea, demonstrating improved performance. Additionally, when compared to FR', $AvFR$ shows significant improvements in the rendering results.

5.3 Performance Evaluation

In the performance evaluation, we divide the VRS-based foveated rendering pipeline into four steps: geometry processing, material rendering, shadow and depth rendering, and post processing. In FR, the geometry processing stage is responsible for various tasks, including vertex processing, clipping, perspective division, and viewport transformation. In $AvFR$, this stage includes an additional process for computing the required shading rate map. Since $AvFR$ needs to compute this shading rate map to guide frame rendering, and the time cost for this computation is less than 0.5ms across all four scenes, it is challenging to list it separately. Therefore, we combine it into the geometry processing stage. The material rendering stage is primarily responsible for texture mapping and material-based lighting shading. Shadow and depth rendering handles shadow mapping, depth testing, depth of field effect, etc. The post processing stage focuses on tone mapping, anti-aliasing, color correction, etc. Fig. 9 visualizes the plots of the stacked time cost of $AvFR$ and FR in *street* (a), *forest* (b), *library* (c), and *apartment* (d). As shown in Fig. 9, the time costs of $AvFR$ are lower than those of FR in all four scenes.

Table 4 shows the time-cost speedup of each step in the rendering pipeline and the overall performance improvement of $AvFR$ over FR. Since post processing operates on uniform pixels in screen space, the performance improvement of $AvFR$ is mainly reflected in the first three steps, and there is no performance improvement in the post processing step. Material rendering is the key step for $AvFR$ to achieve rendering performance improvement, accounting for 63% of the total time cost on average. AvFR improves performance by 1.3-1.4× compared with FR in this step. The large number of geometric meshes in *forest* leads to significantly higher time consumption in the geometry processing stage than in the other three scenes. However, due to the reduction in the number of geometric processing objects required in $AvFR$, performance in this stage improves by 1.5× compared to FR. In *apartment*, although the time cost in the post processing stage is around 0.3ms, the performance of color correction and tone mapping sees further improvement due to the reduced shading quantity achieved by the constructed shading rate map in $AvFR$. This leads to a 1.4× speedup in the post processing stage compared to FR. In *library*, the large number of light sources increases the number of shadow maps. $AvFR$ more aggressively reduces the overall number of sampled shadow maps, greatly enhancing the performance of the shadow and depth rendering stage by 3.6×. In all scenes, $AvFR$ achieves an overall performance improvement of 1.2-1.4× compared with FR. We compare the p-value and Cohen's d of the total time cost per frame by $AvFR$ and FR in the four scenes. p-values are all 0.00, and the values of Cohen's d are 17.11, 55.54, 33.14, and 37.16 in *street*, *forest*, *apartment*, and *library*, with the effect sizes all being *huge*. Therefore, we have **Conclusion 3**: The rendering performance improvement of $AvFR$ compared with FR is significant. Thus, based on **Conclusion 1** in the main user study and **Conclusion 3** in the performance evaluation, the results support **H2**.

6 Conclusion, Limitation, and Future Work

The analysis of auditory-content-based perceived rendering quality has shown that auditory content significantly influences the perceived visual quality in foveated rendering for VR scenes. Based on these findings, we propose the audio-visual aware foveated rendering method (AvFR) to significantly accelerate foveated rendering performance without sacrificing visual fidelity in VR scenes with auditory content. Compared to the state-of-the-art foveated rendering methods, AvFR achieves up to a 1.4× speedup while maintaining similar perceived visual rendering quality.

Since the human auditory system perceives auditory content in VR scenes as background audio through headphones in HMDs, the proposed AvFR in this paper does not consider the effects of auditory content coming from different directions on visual perception in an immersive VR environment. This limitation arises from the lack of high-quality 360° panoramic video datasets that include spatial auditory content, which need to uniformly cover low, medium, and high levels of loudness and frequency. As a result, there is insufficient data support for building perceptual models based on spatial characteristics. Additionally, the spatial features of auditory content often need to be integrated with specific scene object content, necessitating more complex perceptual modeling, which can challenge rendering performance. Conversely, experimental results indicate that loudness and frequency of audio are sufficient to significantly enhance rendering performance, suggesting that the investment in spatial perception may not provide

Table 4: Time cost comparison between *AvFR* and *FR* in each step of the rendering pipeline.

scene	geometry processing		material rendering		shadow and depth rendering		post processing		total time cost	speedup
	time cost	speedup	time cost	speedup	time cost	speedup	time cost	speedup		
str.	0.84	1.0	4.56	1.3	2.76	1.1	0.15	0.9	8.31	1.2
for.	8.34	1.5	6.67	1.3	1.43	1.0	0.13	1.0	16.57	1.4
apa.	0.46	1.1	9.88	1.4	3.70	1.0	0.22	1.4	14.26	1.3
lib.	0.69	1.1	10.89	1.3	0.37	3.6	0.13	1.0	12.08	1.3

adequate returns in terms of performance improvement. The first possible future work is to build a VR scene dataset that includes spatial auditory content, associating auditory elements with scene objects; then advances the audio-visual perception model to enhance foveated rendering quality and performance in VR. On the other hand, AvFR focuses on optimizing the shading rate within the graphics pipeline and does not address potential aliasing caused by rendering effects such as motion blur and complex materials. Therefore, the other potential future work is to quantify the visual perception of different rendering effects in VR scenes with auditory content, and use this as a basis to optimize the quality of complex rendering effects in foveated rendering.

Acknowledgments

This work is supported by the National Natural Science Foundation of China through Project 62402231, 92473205, 62102193; the Natural Science Foundation of the Jiangsu Higher Education Institutions of China 24KJB520027; the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2024C03), and the National Key R&D Program of China Grant No. 2023YFB4404400.

References

[1] T. J. Ayres and P. Hughes. Visual acuity with noise and music at 107 dba. *Journal of Auditory Research*, 1986. 1, 3

[2] L. Bartel and A. Mosabbir. Possible mechanisms for the effects of sound vibration on human health. In *Healthcare*, vol. 9, p. 597. MDPI, 2021. 3

[3] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, p. 532–540, Apr 1983. doi: 10.1109/tcom.1983.1095851 7

[4] A. Carlini and E. Bigand. Does sound influence perceived duration of visual motion? *Frontiers in Psychology*, 12:751248, 2021. 3

[5] N. corporation. Vrworks-variable rate shading(vrs), 2020. 3, 5

[6] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980. 7

[7] B. L. Day and R. C. Fitzpatrick. The vestibular system. *Current biology*, 15(15):R583–R586, 2005. 1

[8] B. Duinkharjav, K. Chen, A. Tyagi, J. He, Y. Zhu, and Q. Sun. Color-perception-guided display power reduction for virtual reality. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. 7

[9] R. Fan, X. Shi, K. Wang, Q. Ma, and L. Wang. Scene-aware foveated rendering. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[10] H. Fletcher and W. A. Munson. Loudness, its definition, measurement and calculation. *Bell System Technical Journal*, 12(4):377–430, 1933. 3, 8

[11] B. Guenter, M. Finch, S. Drucker, D. Tan, and J. Snyder. Foveated 3d graphics. *ACM transactions on Graphics (tOG)*, 31(6):1–10, 2012. 2, 4, 8

[12] H. Hagtvedt and S. A. Brasel. Cross-modal communication: sound frequency influences consumer responses to color lightness. *Journal of Marketing Research*, 53(4):551–562, 2016. 3

[13] G. Halmagyi, I. Curthoys, J. Colebatch, and S. Aw. Vestibular responses to sound. *Annals of the New York Academy of Sciences*, 1039(1):54–67, 2005. 3

[14] S. Hidaka and M. Ide. Sound can suppress visual perception. *Scientific reports*, 5(1):10483, 2015. 1, 3

[15] D. Jiménez-Navarro, A. Serrano, and S. Malpica. Minimally disruptive auditory cues: their impact on visual performance in virtual reality. *The Visual Computer*, pp. 1–15, 2024. 2

[16] A. Jindal, K. Wolski, K. Myszkowski, and R. K. Mantiuk. Perceptual model for adaptive local shading and refresh rate. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021. 2

[17] B. Krajancich, P. Kellnhofer, and G. Wetzstein. A perceptual model for eccentricity-dependent spatio-temporal flicker fusion and its applications to foveated graphics. *ACM Transactions on Graphics (TOG)*, 40(4):1–11, 2021. 2

[18] B. Krajancich, P. Kellnhofer, and G. Wetzstein. Towards attention–aware foveated rendering. *ACM Transactions on Graphics (TOG)*, 42:1 – 10, 2023. 2, 6, 8

[19] S. Malpica, A. Serrano, M. Allue, M. G. Bedia, and B. Masia. Crossmodal perception in virtual reality. *Multimedia Tools and Applications*, 79:3311–3331, 2020. 2

[20] S. Malpica, A. Serrano, J. Guerrero-Viu, D. Martin, E. Bernal, D. Gutierrez, and B. Masia. Auditory stimuli degrade visual performance in virtual reality. In *ACM SIGGRAPH 2022 Posters*, pp. 1–2. 2022. 2

[21] S. Malpica, A. Serrano, D. Gutierrez, and B. Masia. Auditory stimuli degrade visual performance in virtual reality. *Scientific Reports*, 10(1), Jul 2020. doi: 10.1038/s41598-020-69135-3 1, 3

[22] J. H. McDermott and A. J. Oxenham. Music perception, pitch, and the auditory system. *Current opinion in neurobiology*, 18(4):452–463, 2008. 3

[23] X. Meng, R. Du, and A. Varshney. Eye-dominance-guided foveated rendering. *IEEE transactions on visualization and computer graphics*, 26(5):1972–1980, 2020. 2

[24] X. Meng, R. Du, M. Zwicker, and A. Varshney. Kernel foveated rendering. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 1(1):1–20, 2018. 2

[25] A. J. Oxenham. How we hear: The perception and neural coding of sound. *Annual review of psychology*, 69:27–50, 2018. 3

[26] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics*, p. 1–12, Nov 2016. doi: 10.1145/2980179.2980246 1, 2, 3, 4, 8, 9, 10

[27] E. Peli. Contrast in complex images. *Journal of the Optical Society of America A*, p. 2032, Oct 1990. doi: 10.1364/josaa.7.002032 7

[28] J. Púčik, P. Kubinec, and O. Ondřáček. Fft with modified frequency scale for audio signal analysis. In *2014 24th International Conference Radioelektronika*, pp. 1–4. IEEE, 2014. 3

[29] D. M. Rasetshwane, A. C. Trevino, J. N. Gombert, L. Liebig-Trehearn, J. G. Kopun, W. Jesteadt, S. T. Neely, and M. P. Gorga. Categorical loudness scaling and equal-loudness contours in listeners with normal hearing and hearing loss. *The Journal of the Acoustical Society of America*, 137(4):1899–1913, 04 2015. doi: 10.1121/1.4916605 1

[30] B. Series. Algorithms to measure audio programme loudness and true-peak audio level. In *International Telecommunication Union Radiocommunication Assembly*, 2011. 3

[31] X. Shi, L. Wang, J. Wu, R. Fan, and A. Hao. Foveated stochastic lightcuts. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3684–3693, 2022. 2, 6

[32] X. Shi, L. Wang, J. Wu, W. Ke, and C.-T. Lam. Locomotion-aware foveated rendering. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 471–481. IEEE, 2023. 2, 5

[33] E. Skovenborg. Loudness range (lra)-design and evaluation. In *Audio Engineering Society Convention 132*. Audio Engineering Society, 2012. 6

[34] C. Spence and J. Driver. Audiovisual links in exogenous covert spatial

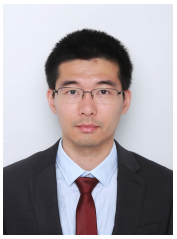
orienting. *Perception & psychophysics*, 59(1):1–22, 1997. 3, 8
[35] P. Stark and R. Parker. Bounded-variable least-squares: an algorithm and applications. *Computational Statistics*, 10, 01 1995. 8
[36] B. E. Stein and M. A. Meredith. *The merging of the senses*. MIT press, 1993. 3, 8
[37] C. J. Steinmetz and J. Reiss. pyloudnorm: A simple yet flexible loudness meter in python. In *Audio Engineering Society Convention 150*. Audio Engineering Society, 2021. 6
[38] M. Stengel, S. Grogorick, M. Eisemann, and M. Magnor. Adaptive image-space sampling for gaze-contingent real-time rendering. In *Computer Graphics Forum*, vol. 35, pp. 129–139. Wiley Online Library, 2016. 2
[39] M. N. Stevens, D. L. Barbour, M. P. Gronski, and T. E. Hullar. Auditory contributions to maintaining balance. *Journal of Vestibular Research*, 26(5-6):433–438, 2016. 1
[40] S. S. Stevens. The measurement of loudness. *The Journal of the Acoustical Society of America*, 27(5):815–829, 1955. 3, 6
[41] S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190, 1937. 7
[42] P. Susini, G. Lemaitre, and S. McAdams. Psychological measurement for sound description and evaluation. *Measurements with persons: Theory, methods, and implementation areas*, 227, 2012. 3
[43] Y. Suzuki and H. Takeshima. Equal-loudness-level contours for pure tones. *The Journal of the Acoustical Society of America*, 116(2):918–933, 08 2004. doi: 10.1121/1.1763601 1
[44] Z. Tang, N. J. Bryan, D. Li, T. R. Langlois, and D. Manocha. Scene-aware audio rendering via deep acoustic analysis. *IEEE transactions on visualization and computer graphics*, 26(5):1991–2001, 2020. 2
[45] C. Tursun and P. Didyk. Perceptual visibility model for temporal contrast changes in periphery. *ACM Transactions on Graphics*, 42(2):1–16, 2022. 2
[46] O. T. Tursun, E. Arabadzhiyska-Koleva, M. Wernikowski, R. Mantiuk, H.-P. Seidel, K. Myszkowski, and P. Didyk. Luminance-contrast-aware foveated rendering. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2, 6, 7
[47] E. Van der Burg, C. N. Olivers, A. W. Bronkhorst, and J. Theeuwes. Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5):1053, 2008. 3
[48] D. R. Walton, R. K. Dos Anjos, S. Friston, D. Swapp, K. Aksit, A. Steed, and T. Ritschel. Beyond blur: Real-time ventral metamers for foveated rendering. *ACM Transactions on Graphics*, 40(4):1–14, 2021. 6
[49] L. Wang, X. Shi, and Y. Liu. Foveated rendering: A state-of-the-art survey. *Computational Visual Media*, 9(2):195–228, 2023. 2, 7
[50] J. R. Williams, Y. A. Markov, N. A. Tiurina, and V. S. Störmer. What you see is what you hear: sounds alter the contents of visual perception. *Psychological science*, 33(12):2109–2122, 2022. 3
[51] C. Wu and M. Low. Fft-based simultaneous calculations of very long signal multi-resolution spectra for ultra-wideband digital radio frequency receiver and other digital sensor applications. *Sensors*, 24(4):1207, 2024. 3
[52] J. Ye, A. Xie, S. Jabbireddy, Y. Li, X. Yang, and X. Meng. Rectangular mapping-based foveated rendering. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 756–764. IEEE, 2022. 2
[53] Y. Zhang, K. You, X. Hu, H. Zhou, K. Kiyokawa, and X. Yang. Retinotopic foveated rendering. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 903–912. IEEE, 2024. 2
[54] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and models*, vol. 22. Springer Science & Business Media, 2013. 3, 6
[55] E. Zwicker, G. Flottorp, and S. S. Stevens. Critical band width in loudness summation. *The Journal of the Acoustical Society of America*, 29(5):548–557, 1957. 7

Xuehuai Shi got his Ph.D. degree from the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China. He is now a lecturer at the School of Computer Science, Nanjing University of Posts and Telecommunications. His research interests include foveated rendering, real-time rendering, human-computer interaction, and augmented reality.

Yucheng Li is an undergraduate student at the State Key Laboratory of Virtual Reality Technology and systems, School of Computer Science and Engineering, Beihang University, Beijing, China. His current research focuses on virtual reality, computer graphics, and visualization.

Jiaheng Li is currently an undergraduate student at the State Key Laboratory of Virtual Reality Technology and systems, School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include computer graphics and virtual reality.

Jian Wu received his Ph.D. degree from the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China. He is currently an assistant professor at the School of Computer Science and Engineering, Beihang University, Beijing, China. His current research focuses on virtual reality, augmented reality, human-computer interaction and visualization.



Jieming Yin is a professor with the School of Computer Science, Nanjing University of Posts and Telecommunications, China. He obtained his PhD degree from University of Minnesota, Twin Cities in 2015. He used to be a faculty member at Lehigh University, and a researcher at AMD. His research interests lie in computer architecture, machine learning aided computer system design, and virtual reality.

Xiaobai Chen is a professor with the School of Computer Science, Nanjing University of Posts and Telecommunications, China. He obtained his PhD degree from Sun Yat-sen University-Carnegie Mellon University Joint Institute of Engineering at Sun Yat-sen University. His research interests lie in computer architecture, AI computing system, and virtual reality.

Lili Wang received her Ph.D. degree from the Beihang University, Beijing, China. She is a professor with the School of Computer Science and Engineering of Beihang University, and a researcher with the State Key Laboratory of Virtual Reality Technology and Systems. Her interests include virtual reality, augmented reality, mixed reality, real-time rendering and realistic rendering.

Audio-visual aware Foveated Rendering
Supplementary Material

Xuehuai Shi, Yucheng Li, Jiaheng Li, Jian Wu, Jieming Yin, Xiaobai Chen, and Lili Wang

In this document, we provide supplemental pilot user studies and evaluation details in support of the main text. The experimental results in supplemental pilot user studies validate that, compared with the non-auditory condition, the enabling visual-semantic-consistent auditory content creates significant differences in visual-perception attention, and enhances the perceived visual quality in VR. Additionally, the supplementary details of the evaluation illustrate the specific implementation of the audio-visual aware foveated rendering method (AvFR) and the test scenes, and ablate the different features of AvFR in terms of the rendering performance.

1 Pilot User Studies

In this section, we conduct the supplementary pilot user study 1 to evaluate the effect of auditory content on visual perception in Section 1.1, explore the factors contributing to the perceived visual quality gap between auditory and non-auditory conditions in Section 1.2, and demonstrate the dataset auditory content expansion details for the pilot user study in the main text in Section 1.3.

1.1 Supplementary Pilot User Study 1: Evaluating Audio Effects on Visual Perception

Based on the theoretical basis in Section 3 in the main text, we conduct the supplementary pilot user study 1 to evaluate the impact of having auditory content versus not having it on visual perception in VR. We formulate the hypothesis for the supplementary pilot user study 1:

H3 Viewing scenes with visual-semantic-consistent auditory content enhances perceived visual quality compared to silent scenes in VR.

1.1.1 User Study Design

Setup We use a PICO 4 Pro HMD powered by a workstation with a 3.9Hz Intel® Core™ i9-12900K CPU, 32GB RAM, and an NVIDIA GeForce GTX 3080 Ti graphic card. The resolution of the HMD is 2160×2160 pixels for each eye, and the field-of-view is 105°. The program is developed with C# and is run in Unity 2021.3.13f1.

Participants We recruit 20 participants, including 10 males and 10 females, aged 18 to 50, with an average age of 35. All participants have normal hearing and vision or have corrected vision through glasses. Ten of them have experience using HMD VR applications before the study.

Fig. 1: Visualization of all scenes in the 360° panoramic video dataset D.

Dataset To compare the impact of different auditory conditions on the visual perception in VR, we construct a 360° panoramic video dataset D with auditory content, as visualized in Fig. 1. D includes two types of scenes, namely, forest stream and urban street, with 10 entries for each scene type. All scenes in D utilize original audio to eliminate the impact of semantic inconsistencies between auditory and visual content on visual perception. The selection of forest stream and urban street as scene types is based on their auditory attributes, which cover the perceivable loudness and frequency range of the HAS. This ensures a diversity of auditory content under visually similar conditions, facilitating an understanding of how visual perception is affected by different auditory conditions in VR. Due to the ability of $lufs$ and Hz to accurately describe the loudness and frequency of the auditory content perceived by the HAS, we use $lufs$ and Hz as the units to represent the loudness and frequency of the auditory content in VR scenes. The value range of $lufs$ is from negative infinity to 0, with values closer to 0 indicating higher loudness. The value range of Hz is from 0 to positive infinity, with larger values representing higher frequencies and higher sound pitches.

To compare the impact of different auditory conditions on visual perception in VR, we construct a 360° panoramic video dataset D with auditory content, as visualized in Fig. 1. D includes two types of scenes, namely, forest stream and urban street, with 10 entries for each scene type. All scenes in D utilize original audio to eliminate the impact of semantic inconsistencies between auditory and visual content on visual perception. The selection of forest stream and urban street as scene types is based on their auditory attributes, which cover the perceivable loudness and frequency range of the hearing aid system (HAS). This ensures a diversity of auditory content under visually similar conditions, facilitating an understanding of how visual perception is affected by different auditory conditions in VR. Because $lufs$ (Loudness Units Full Scale) and Hz (hertz) accurately describe the loudness and frequency of auditory content perceived by the HAS, we use $lufs$ and Hz as the units to represent the loudness and frequency of the auditory content in VR scenes. The value range of $lufs$ is from negative infinity to 0, with values closer to 0 indicating higher loudness. The value range of Hz is from 0 to positive infinity, with larger values representing higher frequencies and higher sound pitches.

In the real world, ranges of loudness and frequency are [-35, -10]$lufs$ and [1, 6]kHz, which can be perceived by the HAS and are safe for prolonged exposure without risking hearing damage [2, 6]. The audio loudness and frequency range in the scenes included in D uniformly cover these real-world audio loudness and frequency ranges. To avoid the effect of audio spatial location on user perception, audio is integrated as background sound in VR scenes and is not linked to any specific object. It is played through the HMD headset during the viewing of audio-enabled panoramic videos.

Condition In the supplementary pilot user study 1, the panoramic video with its original audio enabled in D presented in full resolution is regarded as the experimental condition (EC), while the full-resolution video with audio disabled is regarded as the control condition (CC).

Procedure We ask each participant to view all 20 scenes in D. These scenes are presented using two conditions, EC and CC, in a randomized order for participants to view for 20s. Before viewing the scenes, we visualize the artifacts generated by the 360° panoramic video stitching and instruct participants to ignore the artifacts from the stitching process in the 360° panoramic videos. After each scene presentation, participants are required to give the visual perceptual quality score, μ,

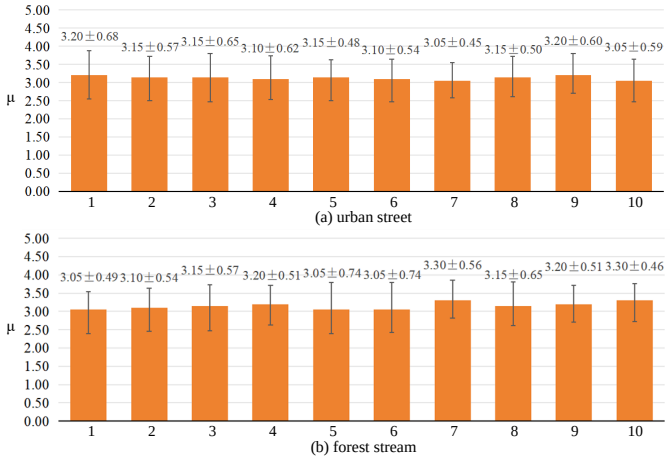


Fig. 2: Average values and standard deviations of μ under *CC* in all scenes included in two scene types in the supplementary pilot user study 1.

of the presented scene, and then the next scene is presented. The visual perceptual quality score μ [7] includes five confidence levels: 5 represents that they cannot perceive artifacts at all, 4 represents that they can perceive acceptable artifacts for only a few very short moments, 3 represents that they can perceive acceptable artifacts, 2 represents that they can perceive noticeable artifacts, and 1 represents that they can perceive obvious artifacts. On average, each participant spends 15*min*. A total of 20 (participants) $\times$ 2 (scene types) $\times$ 10 (scenes in one scene type) $\times$ 2 (conditions) = 800 trials are collected.

1.1.2 Results and Discussion

We use the ANOVA method for pair-wise comparisons. The following are the results of the analysis.

Firstly, we compared the differences in perceived quality among scenes within the two scene types in *CC*. Fig. 2 presents the average values and standard deviations of μ for all urban street and forest stream scenes. In all scenes, the average values of μ range from 3.05 to 3.30. Then, we run an ANOVA analysis to measure the impacts of scene types and individual scenes on μ. The effect of scene types on μ is ($F_{1,780} = 1.92$, $p = 0.17$, $\eta_p^2 = 0.00$), and the effect of individual scenes on μ is ($F_{9,780} = 0.21$, $p = 0.99$, $\eta_p^2 = 0.00$). We use the partial eta squared η_p^2 to measure the effect size of the pair-wise difference. The effect sizes for scene types and individual scenes are *small*, indicating no significant perceived visual differences between the two scene types among all scenes.

Fig. 3 gives the average values and standard deviations of μ under *EC* and *CC* in both forest stream and urban street scene types. The average μ for *EC* is higher than that for *CC* in both scene types. The effect test between *EC* and *CC* is ($F_{1,798} = 188.77$, $p = 1.03 \times 10^{-38}$, $\eta_p^2 = 0.19$), which shows that there is a significant difference between *EC* and *CC* in μ. In conclusion, enabling visual-semantic-consistent auditory content in VR significantly enhances perceived visual quality for participants. Thus, the results support **H3**.

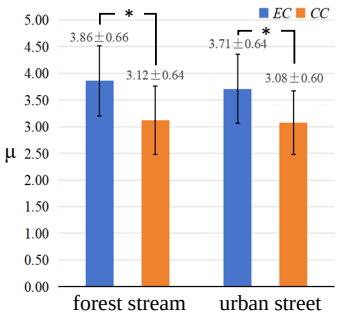


Fig. 3: Average values and standard deviations of μ under different scene types in the supplementary pilot user study 1. Asterisks indicate significant differences.

Fig. 4 visualizes the gaze motion heatmaps under two auditory conditions for specific scenes in both scene types. The results indicate that the salient regions in gaze motion are similar between the two auditory

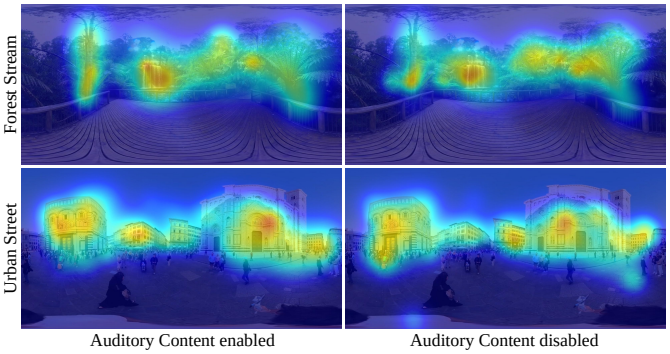


Fig. 4: Visualization of gaze motion when the auditory content is enabled/disabled in two scene types in the supplementary pilot user study 1.

conditions. However, according to participant feedback, three participants report that viewing audio-enabled scenes enhances immersion and directs their attention more toward objects likely to emit sounds. This inconsistency arises from the discrepancy between explicit gaze trajectories and implicit cognitive resource-based attention allocation. Specifically, gaze motion heatmaps reflect only explicit gaze trajectories, while participants' report involves implicit attention allocation influenced by cognitive resources. Implicit attention is not solely determined by gaze motion; it is also regulated by neural interactions and attention shifts, which can lead to visual perception suppression [3, 5]. As a result, even when gaze trajectories are similar, differences in attention allocation can still occur, ultimately affecting participants' visual perception. To investigate this further, we conduct the supplementary pilot user study 2 in Section 1.2 to evaluate visual-perception attention between *EC* and *CC*.

1.2 Supplementary Pilot User Study 2: Evaluating Visual-perception Attention under Different Auditory Conditions

In this section, we conduct the supplementary pilot user study 2 to evaluate the visual-perception attention with auditory content enabled/disabled in VR. We formulate the hypothesis for the supplementary pilot user study 2:

H4 Visual-perception attention is significantly different when the visual-semantic-consistent auditory content is enabled versus disabled in VR.

1.2.1 User Study Design

Setup and Dataset The hardware setup and the program operating environment of this study are the same as those of the supplementary pilot user study 1. We utilize the 360° panoramic video dataset *D* constructed in the supplementary pilot user study 1 to evaluate the visual-perception attention.

Participants and Condition We recruit 20 participants, including 10 males and 10 females, aged 18 to 50, with an average age of 27. All participants have normal hearing and vision or have corrected vision through glasses. Twelve of them have experience using HMD VR applications before the study. The conditions in the supplementary pilot user study 2 are the same as the supplementary pilot user study 1, meaning that the panoramic video with its original audio enabled in *D* presented in full resolution is regarded as the experimental condition (*EC*), while the full-resolution video with audio disabled is regarded as the control condition (*CC*).

Attention Evaluation Task We present a rapid serial visual presentation (RSVP) [1] to modulate the visual-perception attention when viewing scenes in VR. Inspired by [4], the RSVP stimulus consists of N $1^\circ \times 1^\circ$ letters, each lasting 300/*Nms* with 0*ms* blank in between. In the RSVP, the color of the letters alternates between red and yellow, and the sequence of letter colors for each scene is randomized and fixed. The participant is instructed to view the VR scene freely and to navigate through all areas of the scene as thoroughly as possible during each trial. The RSVP stimulus is triggered at a random time point

Fig. 5: Photograph of the supplementary pilot user study 2 setup. The inset shows an enlarged illustration of the stimulus presented in the HMD; the participant is viewing a forest stream-type scene, and the RSVP letter task is randomly presented within an eccentricity range of 0–49°.

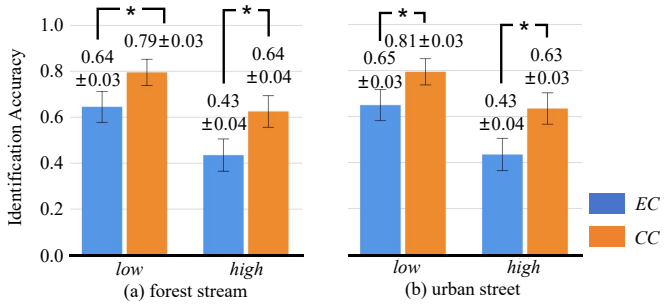


Fig. 6: Average values and standard error of the identification accuracy of the attention evaluation task under different scene types in the supplementary pilot user study 2. Asterisks indicate significant differences.

in the range of 5-8s in the scene view, with its presentation position following a uniform spatial probability distribution across the 0-49° range of the eccentric angle. To control for the interference of irrelevant variables on visual-perception attention, the trigger timing, position, and color parameters of the RSVP stimuli are kept consistent under *EC* and *CC*. Fig. 5 shows that RSVP is presented when the participant is in the scene view. The view time for each trial is 15s. After the scene view, the participant is asked to identify the color of the "target letter" (predefined before each scene view). Increasing N increases the difficulty of the visual-perception attention evaluation task. Two task levels are chosen, with N values set to 2 (easy) and 4 (hard), to force the implementation of two difficulty levels—*low* and *high*—for the visual-perception attention evaluation task. In the same task difficulty, the identification accuracy of the target letter color is used to evaluate the level of visual perceptual attention in VR.

Procedure We ask each participant to view all 20 scenes in D. These scenes are presented using two conditions, *EC* and *CC*, and two attention evaluation task levels, *low* and *high*, in a randomized order for participants to view for 15s. The participant is instructed to complete the attention evaluation task after each scene view. Specifically, after each scene view, the participant has 10s to use the controller's joystick to choose the color of the target letter, with the left side representing red and the right side representing green. At the end of each trial, the screen will turn black, and the participant will have a 5s rest period to clear the visual afterimage of the pattern. To control attention bias in consecutive testing, the experiment adopts a phased execution plan: each participant completes 8 trials per day over 10 consecutive days to finish the entire experimental procedure. Before each day's experiment begins, participants are required to explore the entire scene to eliminate the interference from goal-directed behavior in the visual-perception attention evaluation. On average, each participant spends 45min. A total of 20 (participants) × 2 (scene types) × 10 (scenes in one scene type) × 2 (conditions) × 2 (levels) = 1600 trials are collected.

1.2.2 Results and Discussion

We conduct an ANOVA analysis to evaluate the differences in visual-perception attention based on the identification accuracy in visual-perception attention evaluation tasks.

Fig. 6 presents the average values and standard error of the identification accuracy in both *low*-level and *high*-level visual-perception attention evaluation tasks. The effect test for scene types on the identification accuracy is ($F_{1,1520} = 0.01$, $p = 0.91$, $\eta_p^2 = 0.00$), and is ($F_{19,1520} = 0.66$, $p = 0.74$, $\eta_p^2 = 0.00$) for individual scenes, which indicate that there are no significant differences in the visual-perception attention between two scene types among all scenes.

The effect test for *EC* and *CC* regarding the identification accuracy is ($F_{1,1520} = 132.17$, $p = 0.00$, $\eta_p^2 = 0.08$) in *low* level, and is ($F_{1,1520} = 268.24$, $p = 0.00$, $\eta_p^2 = 0.15$) in *high* level. From forest stream to urban street, under *low* and *high* difficulty conditions, the p-values are 7.96×10^{-4}, 1.28×10^{-4}, 1.16×10^{-3}, and 5.38×10^{-5}, respectively. All p-values are less than 0.01, indicating that the identification accuracy under *EC* is significantly lower than that under *CC* under all scene types and task difficulty levels. The identification accuracy in *high*-level task is lower than in *low*-level task across all conditions and scene types, and the gap between *EC* and *CC* further widens as task difficulty increases. Thus, the results support **H4**. Experimental results indicate that, compared with the non-auditory condition, the auditory condition significantly diminishes the visual-perception attention, and this diminishment effect becomes increasingly pronounced as the visual task difficulty increases. This effect contributes to the significant perceived visual quality gap between *EC* and *CC* in the supplementary pilot user study 1.

1.3 Dataset Auditory Content Expansion in Pilot User Study

To facilitate the comparison of how auditory content with different frequencies and loudness affects visual perception in the pilot user study, we expand dataset D by adding two different frequency auditory contents for each scene, resulting in each scene containing three types of auditory content covering low, mid, and high frequencies, thus obtaining the expanded dataset D'. To ensure the auditory content is semantically consistent with the visual content in D', we recruit 10 participants to evaluate the semantic consistency between the auditory and visual content of all scenes. Each participant views each scene in D' with three different auditory contents in random order, each viewing session lasting 20s. At the end of each scene view, the participant answers a two-alternative forced-choice question: "Is the audio semantics consistent with the visual semantics of the scene?" The positive consistency rates for all auditory contents in each scene are $\geq 80\%$ in D'. Therefore, we conclude that the visual and auditory contents of all scenes in D' are semantically consistent. Fig. 7 visualizes the coverage range of audio loudness and frequency of all scenes in D'. The loudness coverage range for each scene in Fig. 7 is obtained by adjusting the audio volume on the workstation to low, medium, and high levels, then merging the three auditory contents for calculation. The frequency coverage range corresponds to the frequency ranges of the three types of auditory content. As shown in Fig. 7, the audio loudness and frequency of each scene in D' uniformly cover the perceivable and safe-exposed audio loudness range (-35 to -10$lufs$) and the frequency range (1 to 6kHz).

2 Supplemental Details of Evaluation

2.1 Implementation Details

In the implementation of step 1 in AvFR, we reuse the full-resolution rendering result from the previous frame, reducing both its width and height to $\frac{1}{4}$ of the original size to serve as the low-resolution rendering result for the current frame. This approach aims to extract the CSF-based visual loss feature. Next, we combine real-time audio from the scene to extract auditory features, constructing the audio-visual feature-driven perception model. In step 2 of AvFR, E_0 is set to 3.63. This implementation enables the required shading rate map computation in AvFR to be $\leq$0.5ms, resulting in a negligible impact on performance.

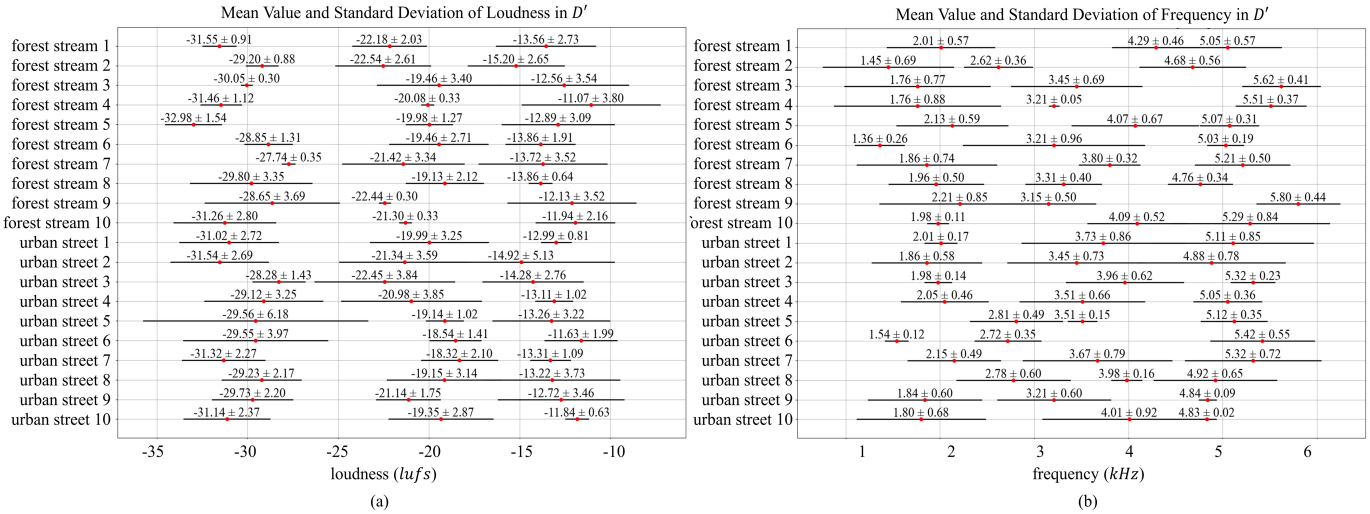
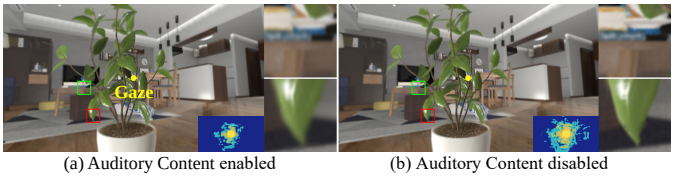


Fig. 7: Coverage range of audio loudness (a) and frequency (b) in the 360° panoramic video dataset D'.

To evaluate AvFR in VR, we construct four VR scenes containing auditory content: *street*, *forest*, *library*, and *room*. In *street*, the number of triangles contained in *street* is 11781.32k, and the auditory content consists of birdcalls and fountain sounds, lasting 150s, with an audio loudness range of [-34.70, -10.50]$lufs$ and an audio frequency range of [1.57, 5.20]kHz. The number of triangles contained in *forest* is 54789.28k, and the auditory content in *forest* is the sound of a waterfall, lasting 150s, with an audio loudness range of [-33.80, -10.4l]$lufs$ and an audio frequency range of [1.47, 5.50]kHz. The number of triangles contained in *apartment* is 3036.64k, and the auditory content is the sound of a music video playing on the television, lasting 253s, with an audio loudness range of [-34.91, -13.30]$lufs$ and an audio frequency range of [1.32, 5.75]kHz. The number of triangles contained in *library* is 8847.41k, and the auditory content is the sound of commonly used library closing music, lasting 205s, with an audio loudness range of [-34.70, -12.71]$lufs$ and an audio frequency range of [1.41, 5.58]kHz.

To ensure the consistency of the auditory content with the semantics of the visual content, we recruit 10 participants to evaluate the semantic consistency between the auditory and visual content of four VR scenes. Each participant is allowed to freely explore the four scenes using the full-resolution rendering method, with each scene lasting 90s. Before each exploration, we introduce the auditory content for that scene to the participant. At the end of each scene exploration, participants answer the question: "Is the audio semantics consistent with the visual semantics of the scene?" The positive consistency rates for the audio selected for all four scenes are all $\geq 80\%$.



(a) Auditory Content enabled (b) Auditory Content disabled

Fig. 8: Rendering results of *AvFR* in *apartment* with auditory content enabled (a) and disabled (b).

2.2 Ablation Evaluation

To validate the rendering advantages of *AvFR* in VR scenes with auditory content, we visualize the shading rate map of *AvFR* along with its corresponding rendering results in *apartment* in Fig. 8, both with auditory content enabled (a) and disabled (b). When the auditory content is disabled, a high shading rate for *AvFR* is observed not only in the foveal region but also in salient peripheral areas due to its eccentricity-CSF-based characteristics. This results in a performance decrease compared to *FR*, with an average frame rate of 45fps in *apartment*. When auditory content is enabled, *AvFR* further optimizes the shading rate, achieving an average frame rate of 70fps, which represents a 1.6× speedup compared to the auditory-content-disabled *apartment*. Additionally, it preserves rendering details in salient peripheral regions, such as the contours of book objects and glossy rendering effects.

Several features of the proposed AvPM influence the rendering performance of *AvFR*. In Table 1, we visualize the performance acceleration of *AvFR* compared to *FR* when enabling and disabling different features of AvPM in the apartment scene. When the CSF-based visual loss feature γ_{vis}, the auditory perception-loudness feature γ_{ld}, and the auditory perception-frequency feature γ_{fq} are disabled, *AvFR* defaults to full-resolution rendering, resulting in a 2.2× decrease in rendering performance. With only γ_{vis} enabled, *AvFR* reverts to eccentricity-CSF-based foveated rendering, which focuses on high-CSF regions in the periphery, leading to a decrease in performance compared to *FR*. Furthermore, when AvPM accelerates *AvFR* using bothγ_{vis} and γ_{ld}, the performance improves by 1.2× compared to *FR*. Disabling γ_{ld} while enabling γ_{vis} and γ_{fq} also results in a 1.1× speedup. The ablation experiment demonstrates that in *apartment*, both auditory loudness and frequency contribute to the performance enhancement of *AvFR*, with loudness having a greater impact.

Table 1: Performance Ablation on AvPM's features of *AvFR* in *apartment*.

Components			speedup
γ_{vis}	γ_{ld}	$\frac{1}{\gamma_{fq}}$	
✗	✗	✗	0.4×
✓	✗	✗	0.8×
✓	✓	✗	1.2×
✓	✗	✓	1.1×
✓	✓	✓	1.3×

REFERENCES

[1] K. R. Dobkins and L. Huang. Attentional effects on contrast discrimination in humans: Evidence for both contrast gain and response gain. *Journal of Vision*, 4(8):456–456, Aug 2004. doi: 10.1167/4.8.456 2

[2] ITU-R. Bs.1770 : Algorithms to measure audio programme loudness and true-peak audio level, 2000. 1

[3] D. Jiménez-Navarro, A. Serrano, and S. Malpica. Minimally disruptive auditory cues: their impact on visual performance in virtual reality. *The Visual Computer*, pp. 1–15, 2024. 2

[4] B. Krajancich, P. Kellnhofer, and G. Wetzstein. Towards attention–aware foveated rendering. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2

[5] S. Malpica, A. Serrano, J. Guerrero-Viu, D. Martin, E. Bernal, D. Gutierrez, and B. Masia. Auditory stimuli degrade visual performance in virtual reality. In *ACM SIGGRAPH 2022 Posters*, pp. 1–2. 2022. 2

[6] A. R. Moller. *Hearing: its physiology and pathophysiology*. Academic Press, 2000. 1

[7] X. Shi, L. Wang, J. Wu, W. Ke, and C.-T. Lam. Locomotion-aware foveated rendering. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 471–481. IEEE, 2023. 2