Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 19 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i19.58317

PAPER

Streaming Load and Rendering Optimization for Mobile VR or AR Art Exhibitions

Zhijuan Chen[1](✉), Xijin Li[2]

[1]Industrial Arts Department, Kaifeng Vocational College of Culture and Arts, Kaifeng, China

[2]Fakulti Ekonomi & Pengurusan, Universiti Kebangsaan Malaysia, Selangor, Malaysia

chenzhijuan@kfwyxy.edu.cn

ABSTRACT

With the growing application of mobile VR or AR technologies in art exhibitions, immersive experiences demand higher efficiency in large-scale data streaming and stereoscopic rendering quality. However, mobile devices are constrained by limited computing power, storage capacity, and network bandwidth, making it challenging for traditional loading and rendering approaches to balance data transmission efficiency with user experience. Current research often relies on fixed-pattern viewpoint prediction based on historical data, which fails to adapt to the dynamic and personalized nature of user viewpoints in mobile environments. This mismatch leads to inefficient preloading, where loaded content does not align well with users' actual focus areas. Furthermore, conventional stereoscopic rendering optimization strategies frequently overlook device performance variability and localized user attention, resulting in low rendering efficiency and excessive power consumption. To address these challenges, this study focuses on two core aspects: (1) viewpoint prediction for dynamic streaming preloading, which integrates user gaze patterns, attention distribution, and exhibition content structure to build a dynamic prediction model that enhances preloading relevance; and (2) viewpoint-driven stereoscopic rendering optimization, where high-detail rendering is applied to predicted focus regions while non-focus areas are simplified. This approach ensures visual quality in key scenes while reducing computational load. The outcomes of this study provide efficient technical solutions for mobile VR or AR art exhibitions, significantly improving streaming speed and rendering smoothness and promoting deeper integration between artistic presentation and digital technology.

KEYWORDS

mobile VR or AR art exhibitions, streaming load, rendering optimization, viewpoint prediction, stereoscopic rendering

1 INTRODUCTION

With the rapid development of digital technology, mobile VR or AR technology, due to its immersive and interactive characteristics [1–4], has shown great application

Chen, Z., Li, X. (2025). Streaming Load and Rendering Optimization for Mobile VR or AR Art Exhibitions. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(19), pp. 151–165. https://doi.org/10.3991/ijim.v19i19.58317

Article submitted 2025-06-22. Revision uploaded 2025-08-08. Final acceptance 2025-08-09.

potential in the field of art exhibitions. Mobile VR or AR art exhibitions break through the spatial limitations of traditional exhibitions, allowing users to experience rich and diverse artistic content anytime and anywhere, and have become a new trend in art dissemination and presentation. However, mobile devices are limited in terms of computing power, storage capacity, and network bandwidth [5–7], resulting in challenges in loading and rendering large-scale art exhibition data. How to achieve efficient streaming load and high-quality rendering in mobile environments has become a key issue to improve user experience.

At present, there have been some studies on loading and rendering optimization for mobile VR or AR. In terms of viewpoint prediction, some studies adopt fixed-pattern prediction methods (such as the multi-viewpoint system attribute fusion prediction method proposed in reference [8]), but such methods do not fully consider the dynamic changes and personalized differences of user viewpoints in mobile scenarios, resulting in a low matching degree between preloaded content and users' actual focus areas, causing data waste or untimely loading. In terms of stereoscopic rendering optimization, traditional methods based on uniform resolution and rendering parameters (such as the global rendering optimization strategy used in reference [9]) do not take into account the performance characteristics of mobile devices and the local focus characteristics of user viewpoints, resulting in low rendering efficiency, screen stuttering, high power consumption, and other problems, which make it difficult to meet the requirements of real-time and smoothness in mobile VR or AR art exhibitions.

This paper mainly focuses on two core aspects. The first is viewpoint prediction for streaming dynamic preloading in mobile VR or AR art exhibitions. By analyzing the movement patterns of user viewpoints during mobile processes, the characteristics of attention distribution, and the structural characteristics of art exhibition content, a dynamic viewpoint prediction model is constructed to accurately predict users' future viewpoint areas, thereby guiding streaming preloading strategies and improving the targeting and effectiveness of data loading. The second is stereoscopic rendering optimization based on mobile viewpoints for mobile VR or AR art exhibitions. Based on viewpoint prediction results, high-resolution and high-detail rendering is applied to the user's current focus areas, while non-focus areas are rendered in a simplified manner, ensuring the rendering quality of key areas while reducing overall rendering computation, improving rendering efficiency, and adapting to the performance limitations of mobile devices. The value of this study lies in proposing a streaming loading and rendering optimization scheme that is more suitable for mobile VR or AR art exhibition scenarios, addressing the shortcomings of existing methods. Through accurate viewpoint prediction and intelligent rendering optimization, the loading speed and rendering quality of art exhibitions on mobile devices can be effectively improved, bringing users a smoother and more immersive viewing experience.

2 VIEWPOINT PREDICTION FOR STREAMING DYNAMIC PRELOADING IN MOBILE VR OR AR ART EXHIBITIONS

From the perspective of the resource optimization objective of streaming dynamic preloading, the bandwidth and computing power limitations of mobile devices require that preloaded data must have a very high "target hit rate," and integrating head movement trajectories and scene saliency features can construct a more comprehensive user attention model. To this end, this paper proposes a prediction

model that combines user head movement trajectories and the saliency features of mobile VR or AR art exhibition scenes. The head movement trajectory reflects the user's immediate movement trend, providing time-series information of viewpoint changes, which is the basis for predicting short-term viewpoints; the scene saliency features analyze the visual hierarchy of the exhibition scene to quantify the "attractiveness weight" of different areas, revealing the core content that the user may focus on during long-term browsing. This multidimensional fusion deep learning model can not only capture the inertia rules of user behavior but also analyze the semantic guiding role of exhibition content, so that the preloading strategy maintains dual coverage of the "user active attention area" and the "content key display area" in the dynamically changing browsing process, fundamentally improving the pertinence and efficiency of preloading and providing a more accurate input basis for subsequent rendering optimization.

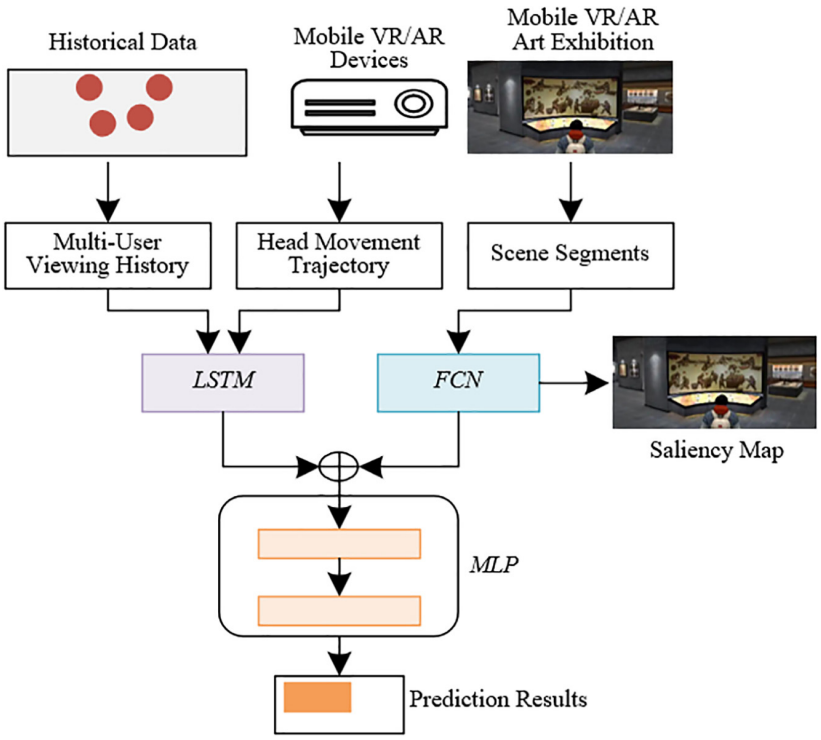

Fig. 1. Structure of the viewpoint prediction model for streaming dynamic preloading in mobile VR or AR art exhibitions

The viewpoint prediction model for streaming dynamic preloading in mobile VR or AR art exhibitions proposed in this paper mainly includes three parts: saliency map, trajectory prediction, and fully connected layer. The model architecture is shown in Figure 1. Among them, the core principle of the saliency map extraction module is to provide the key feature of "content guidance" for viewpoint prediction by quantifying the visual attraction distribution of the scene content, addressing the contradiction between the large size of panoramic images and the focus of user attention. Since art exhibition scenes are often presented as panoramic videos or high-resolution panoramic images, traditional convolutional neural networks (CNN) are limited by the requirement of fixed input size in the fully connected layers [10, 11], making it difficult to directly process such large-size content, while fully convolutional networks (FCN), by replacing the fully connected layers with 1×1 convolution layers [12], eliminate the input size limitation and can adapt to arbitrary resolutions

of panoramic images. The essence of the saliency map is a visual representation of the "ability to attract user attention" of each region in the image. The more salient the region, the more likely it is to become the target of the user's viewpoint. By performing pixel-level classification of panoramic videos using FCN, heatmaps reflecting the saliency of each region can be generated. These heatmaps not only retain the spatial structure information of the scene but also quantify the visual priority of different content elements, providing "content-driven" prior knowledge for the subsequent viewpoint prediction model, enabling the system to anticipate the user's visual bias caused by scene saliency differences, and thus prioritize the scheduling of data in highly salient regions in the preloading strategy. Specifically, let the scene frame content at time s be denoted by n_s, and the FCN network be denoted by FCN, then the saliency feature h_s at this moment can be expressed as:

$$h_s = FCN(n_s) \tag{1}$$

The core principle of the trajectory prediction module is to explore the temporal dependency [13] of viewpoint trajectories based on the temporal continuity of user viewpoint movement, using the long short-term memory (LSTM) network to provide a "behavioral inertia-driven" dynamic prediction basis for the preloading strategy. When users browse art exhibitions, the viewpoint trajectory is essentially time series data generated by head movements, whose variation patterns include not only short-term dynamics of immediate turning but also imply long-term movement trends. Traditional recurrent neural networks (RNN) are difficult to capture long-term dependencies due to the gradient vanishing problem [14], while LSTM effectively solves this defect through a cell state mechanism and can memorize movement patterns over longer periods. The module takes the user's current viewpoint coordinates and historical viewpoint sequence as input and uses LSTM's gate structure to filter key temporal features to predict the movement trajectory of the viewpoint in the next period. Specifically, suppose the input viewpoint coordinate sequence of the past v frames at time s is represented by a_{s-v+m}, a_{s-v+2}, ..., a_s, and the LSTM network is denoted by $LSTM$, then the output viewpoint feature, i.e., the hidden state of the LSTM at time s, can be expressed as:

$$d_{s+1} = LSTM(a_{s-v+1}, a_{s-v+2}, \cdots, a_s) \tag{2}$$

The core principle of the fully connected module is to convert the saliency features and viewpoint trajectory features into quantitative indicators that can directly guide the preloading strategy through feature fusion and probability mapping, solving the cross-modal information integration problem between "content attractiveness" and "behavioral inertia." This module first concatenates the scene visual priority features output by the saliency map module with the temporal movement features extracted by the trajectory prediction module to form a high-dimensional feature vector that contains both spatial saliency and temporal continuity. This fusion mechanism breaks the limitation of single-modal features. The saliency features describe "which areas in the scene are more likely to attract users," and the trajectory features describe "how the user's viewpoint moves over time." The combination of the two enables the model to capture both the guiding effect of the exhibition content on attention and model the inertia of user head movements. Through multi-layer nonlinear transformations in the fully connected network, high-dimensional features are abstracted layer by layer into more discriminative representations and finally mapped by the *softmax* activation function into the viewing probability distribution of each image *tile*, providing a quantitative basis for "regional importance"

for streaming preloading. Specifically, let the viewpoint prediction be performed at time s, and the predicted probability of each tile being viewed at time $s + 1$ be denoted by o_{s+1}. The saliency feature of the scene at time $s + 1$ is denoted by h_{s+1}, and the temporal viewpoint feature at time $s + 1$ is denoted by d_{s+1}. The two-layer fully connected layers that fuse the saliency feature and the viewpoint feature are denoted by function e. The calculation formula of the model is given as follows:

$$O_{s+1} = e([h_{s+1}; d_{s+1}]) \tag{3}$$

Assume the predicted tiles within the field of view (FOV) are denoted by S'_u, the predicted tiles outside the FOV are denoted by S'_p, the actual tiles within the FOV are denoted by S_u, and the actual tiles outside the FOV are denoted by S_p, then the tile overlapping ratio can be expressed as:

$$ACC = \frac{S'_u \cap S_u + S'_p \cap S_p}{S_u + S_p} \tag{4}$$

3 OPTIMIZATION OF STEREOSCOPIC RENDERING IN MOBILE VR OR AR ART EXHIBITIONS BASED ON MOVING VIEWPOINTS

3.1 Algorithm idea

Traditional stereoscopic projection relies on rendering from a single virtual camera viewpoint [15], resulting in the fixed left and right eye images remaining unchanged during user movement. This leads to deformation or displacement of the fused stereoscopic virtual image as the viewing position shifts, especially prominent in VR or AR exhibitions where users move freely. For example, when a viewer walks around a 3D sculpture exhibit, traditional systems cause distortion in the stereoscopic image of the sculpture, disrupting the spatial integrity of the artwork. The technique based on moving viewpoints captures the user's position and perspective in real time and dynamically adjusts the virtual camera's position and projection matrix, ensuring that the stereoscopic virtual image viewed by users from different positions always maintains the correct spatial position and shape. This "user-centered" rendering strategy allows viewers to continuously observe a stable and physically consistent stereoscopic scene during movement. It is particularly suitable for sculptures and installation art in exhibitions that require multi-angle appreciation, preventing visual deviation from interrupting the viewing experience and enhancing the authenticity and immersion of artistic expression.

In the implementation of stereoscopic rendering and projection in mobile VR or AR art exhibitions, real-time adaptation to the moving viewpoint can be achieved by dynamically constructing the camera frustum. This method first converts the physical characteristics of the projection screen into fixed convergence plane parameters in the virtual scene. This plane, as the intersection reference of the left and right camera frustums, has four vertices, *MI*, *MF*, *EI*, *EF* that remain unchanged in the virtual scene coordinate system and directly correspond to the physical boundaries of the screen. The position of the stereoscopic cameras is dynamically determined based on the real-time position of the user's eyes, forming two camera nodes corresponding to the left and right eyes. The system calculates the orientation matrix of the camera through real-time acquisition of the user's head position and line of sight direction, then transforms the convergence plane vertices from world coordinates to camera coordinates. Combined with the pre-set near clipping plane v and

far clipping plane d, it finally determines the edge parameters m, e, s, y of the left and right camera frustums. The perspective transformation matrix of the camera is as follows:

$$L = \begin{bmatrix} \frac{2v}{e-m} & 0 & -\frac{e+m}{e-m} & 0 \\ 0 & \frac{2v}{s-y} & -\frac{s+y}{s-y} & 0 \\ 0 & 0 & \frac{d}{d-v} & -\frac{dv}{d-v} \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{5}$$

Curved screens, with a wider FOV and curvature characteristics more in line with human vision [16–18], can significantly enhance user immersion and are particularly suitable for artistic exhibitions requiring panoramic experiences. However, their non-planar geometric form disrupts the simple rectangular mapping relationship between traditional flat screens and virtual convergence planes. When stereoscopic cameras directly project images of T_m and T_e rendered according to flat-screen logic onto a curved screen, differences in screen curvature cause changes in light reflection paths. As a result, the image disparity received by the user's eyes does not match the actual observation position, ultimately causing displacement and distortion of the stereoscopic virtual image. In addition, curved screens often require multiple projectors to jointly stitch high-resolution images. The geometric calibration and disparity consistency control among projection areas are more complex. If a fixed-disparity rendering strategy is still used, users at different positions may experience visual fatigue or stereoscopic discontinuity due to disparity deviation caused by screen curvature, severely damaging the immersive experience of the art exhibition. Therefore, stereoscopic rendering for curved screens must break through the traditional flat mapping framework and establish a disparity adjustment mechanism dynamically matching the screen geometry. To address this, this paper proposes a disparity dynamic adjustment method to resolve the combined interference of moving viewpoint and screen geometry dynamics on stereoscopic imaging.

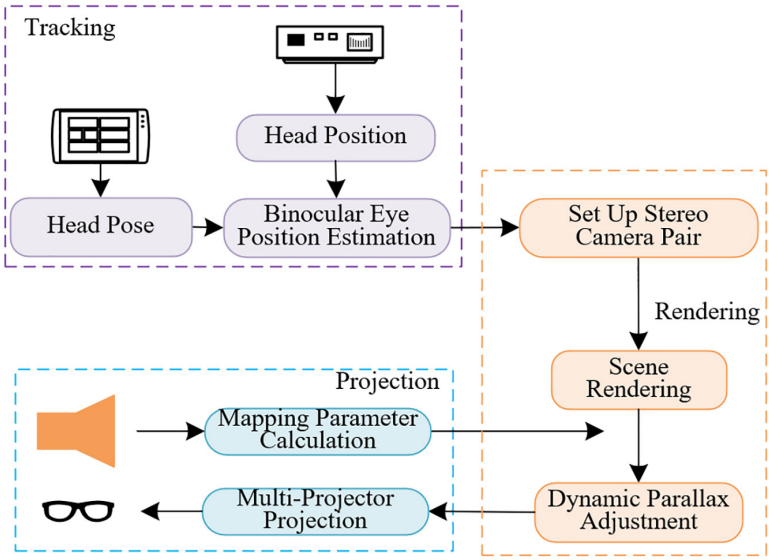

Fig. 2. Algorithm flowchart of stereoscopic rendering optimization based on moving viewpoints in mobile VR or AR art exhibitions

The algorithm flow is shown in Figure 2. The method centers around the core logic of "dynamic tracking—intelligent rendering—precise projection," and achieves dual adaptation to curved screen geometry and user moving viewpoint through a three-stage technical architecture. The first stage, "tracking," uses *Kinect* sensors and *IMU* components to capture the user's head 3D position and posture data in real time. Based on an ergonomic model, it estimates the accurate coordinates of the left and right eyes, providing dynamic viewpoint input for subsequent rendering. This stage addresses the problem of traditional fixed-viewpoint rendering being unable to perceive user position changes, enabling the system to respond in real time to viewer movement in front of the curved screen. The second stage, "rendering," dynamically configures the frustum parameters of the stereoscopic camera pair based on the positions of both eyes and the geometric parameters of the curved screen: first, a virtual convergence plane is constructed according to the screen shape, so that its geometry completely matches the physical screen. Then, through coordinate transformation, the positions and orientations of the user's eyes are mapped to the positions and orientations of the left and right cameras, generating asymmetric frustums to simulate binocular disparity. To address the disparity distortion caused by the curved screen, the system further introduces a disparity dynamic adjustment algorithm, which corrects the edge parameters of the camera frustum in real time according to the curvature of each screen region. This ensures that the rendered left and right eye images T_m and T_e are projected onto the curved screen such that, after light reflection, they form correctly aligned stereoscopic virtual images in the user's eyes. The third stage, "projection," uses multi-projector calibration and blending fusion technology to accurately project the dynamically adjusted left and right eye images onto the curved screen, forming a complete high-resolution stereoscopic image. Considering the multi-device coordination required by curved screens, the system pre-calibrates the projection areas of each projector, eliminating brightness and geometric deviations between devices, and ensuring that the stitched images maintain disparity consistency in regions with curvature changes. When users wear 3D glasses to view, the disparity-adjusted images for the left and right eyes are reflected into the eyes through the screen. After being fused by the brain, they form an undistorted stereoscopic virtual image, whose position and shape always strictly align with the target object in the virtual scene, unaffected by user movement or screen curvature. This three-stage technical architecture achieves high-precision stereoscopic rendering for moving viewpoints on curved screens through a closed-loop control of "real-time user position perception—disparity parameter dynamic computation—multi-device collaborative projection."

3.2 Parallax dynamic adjustment

The curvature variation of the curved screen results in inconsistencies in geometric parameters such as the normal direction and projection distance of different points on the screen. Direct calculation of parallax relationships in curved space requires handling high-dimensional nonlinear transformations, which involves extremely high computational complexity and is difficult to guarantee real-time performance. Therefore, in the parallax dynamic adjustment of curved screens for mobile VR or AR art exhibitions, this paper defines a virtual planar screen. The virtual planar screen serves as an intermediate medium, with a geometric form consistent with traditional planar screens, allowing the reuse of mature stereoscopic rendering theory under planar screens. Ideal parallax-correct images are first generated

on the virtual planar screen and then transformed through spatial mapping to be "fitted" into the physical space of the curved screen. The image space rendered by a stereoscopic camera is denoted as (i, n), the parameterized planar screen space is denoted as (t_1, s_1), the 3D coordinates corresponding to (t_1, s_1) on the planar screen are also denoted as (t_1, s_1). The mapping relationship between the planar screen space (t_1, s_1) and image space (i, n) is denoted as $L_{o \Leftrightarrow U}$, and the mapping relationship between the planar screen space (t_1, s_1) and the 3D coordinate point (a_1, b_1, c_1) on the planar screen is denoted as L_1. The parameterized curved screen space is denoted as (t_2, s_2), the 3D coordinates corresponding to (t_2, s_2) on the curved screen are denoted as (a_2, b_2, c_2), the mapping relationship between the curved screen space (t_2, s_2) and image space (i, n) is denoted as $L_{Z \Leftrightarrow U}$, and the mapping relationship between the curved screen space (t_2, s_2) and the 3D coordinate point (a_2, b_2, c_2) on the curved screen is denoted as L. The mapping relationship among curved screen space, planar screen space, and projected image space is expressed by the following equations:

$$(a_1, b_1, c_1) \overset{L_1}{\leftrightarrow} (t_1, s_1) \overset{L_{0 \leftrightarrow 1}}{\leftrightarrow} (i, n) \tag{6}$$

$$(a_2, b_2, c_2) \overset{L_2}{\leftrightarrow} (t_2, s_2) \overset{L_{Z \leftrightarrow 1}}{\leftrightarrow} (i, n) \tag{7}$$

The non-planar curvature of the curved screen renders the fixed parallax rendering strategy of traditional planar screens ineffective. That is, when images rendered according to planar screen logic by the left and right cameras are directly projected onto a curved screen, differences in normal directions and projection distances of different points on the screen alter the light reflection paths. This causes the parallax received by the user's eyes to be inconsistent with the actual observation position, ultimately resulting in the stereoscopic virtual image point o' deviating from the ideal position o on the planar screen. Especially during user movement, the parallax deviation caused by the screen curvature changes dynamically with the observation angle, making it impossible for viewers to obtain a stable stereoscopic visual experience, which severely affects the spatial presentation of artworks. Therefore, it is necessary to dynamically adjust the parallax of stereoscopic images to compensate for the interference of the curved screen with the light propagation path, ensuring that the stereoscopic projection effect on the curved screen is consistent with that of a planar screen and that the virtual image point o' always maintains the correct spatial position. This satisfies the immersive, distortion-free visual experience required by mobile VR or AR art exhibitions. Specifically, the system first uses the virtual planar screen as the baseline, defining the perspective projection relationship between the image points T_{lm}, T_{le} on the planar screen and the virtual scene point O. When rendering the curved screen image, for the original left camera image $U_1(i_1, n_1)$, a new image $U_2(i_2, n_2)$ is generated through parallax dynamic adjustment so that the image points T_{2m}, T_{2e} on the curved screen satisfy the collinearity condition with the eye position and the virtual point O. This constraint is achieved by real-time calculation of curvature parameters in each region of the curved screen, dynamically adjusting the parallax offset of each pixel in the left and right camera images, i.e., horizontal displacements Δi, Δn, so that the pixel coordinates projected onto the curved screen, after light reflection, visually match the ideal projection of the planar screen. Assuming that the images $U_1(i_1, n_1)$, $U_2(i_2, n_2)$, and the image points corresponding to the projection point P on the planar screen are represented by T_1 and T_2, the spatial coordinate of image point T_1 is (a_1, b_1, c_1), and the spatial coordinate of image point T_2 is (a_2, b_2, c_2). The perspective projection matrix of the left camera

corresponding to the left eye is denoted as L, and the coordinate transformation from spatial coordinate system to the left camera coordinate system is denoted as S. T_1 and T_2 satisfy the perspective projection relationship:

$$\begin{bmatrix} a_1 \\ b_1 \\ c_1 \end{bmatrix} = S^{-1}LS \begin{bmatrix} a_2 \\ b_2 \\ c_2 \end{bmatrix} \tag{8}$$

The first step of parallax dynamic adjustment is to establish an accurate mapping between output image pixels and the physical space of the curved screen. Each image point in the processed image $U_2(i_2, n_2)$ is back-projected to the 3D coordinate point $T_2(a_2, b_2, c_2)$ on the curved screen through Equation (7).

The second step is to convert the curved coordinate T_2 to the 3D coordinate $T_1(a_1, b_1, c_1)$ of the virtual planar screen through Equation (8). This virtual plane is aligned with the physical size and center position of the curved screen and serves as the reference space for parallax computation.

The third step is to map the planar screen coordinate T_1 to the image point position (i_1, n_1) of the original input image U_1 through Equation (6), completing the pixel filling of the processed image U_2.

Assuming the mapping from (i_1, n_1) to (t_1, s_1) is denoted as $L_{o\leftarrow U}$, the mapping from (t_2, s_2) to (i_2, n_2) is denoted as $L_{U\leftarrow Z}$, and the mapping from (t_1, s_1) to (t_2, s_2) is denoted as $L_{Z\leftarrow P}$. The entire process can be described by the following equations:

$$(i_2, n_2) = L_{U\leftarrow Z}(L_{Z\leftarrow O}(L_{O\leftarrow U}(i_1, n_1))) \tag{9}$$

3.3 User perception optimization method

In stereoscopic projection of mobile VR or AR art exhibitions, the drifting sensation caused by head pose changes is essentially a dynamic mismatch problem between disparity calculation and the actual gaze direction. Traditional stereoscopic rendering systems can capture head position but often ignore the influence of pose parameters on the binocular visual axis angle. When the user's head slightly tilts or pitches, the actual intersection point of the binocular visual axes deviates from the convergence plane of the virtual scene. If the system does not update the orientation matrix of the stereoscopic camera in real time, the disparity of the left and right eye images is still calculated based on the initial pose, resulting in a mismatch between the virtual object's image position on the retina and the real gaze direction, forming a "drifting" illusion. This deviation is particularly sensitive in art exhibitions, where users require precise observation of exhibit details, meaning the stereoscopic imaging must be strictly aligned with the visual focal point. Real-time tracking of head pose and disparity correction ensures that regardless of the viewing angle, the spatial position of the virtual object always coincides with the intersection point of the binocular visual axes, eliminating visual conflicts caused by pose changes and enhancing the realism and immersion of the artistic experience.

Traditional binocular position estimation methods rely only on the head center position O and the average interpupillary distance r to estimate the left and right eye positions. This simplified model ignores the actual influence of head pose on the spatial positions of the eyes. When the user turns sideways to the screen or tilts the head, the actual positions of the eyes deviate from the estimated ones in space, causing a mismatch between the view frustum parameters of the left and right cameras and

the actual disparity. The projected virtual point P deviates from the target position P', resulting in drifting and distortion. This deviation can significantly disrupt immersion in art exhibitions. For example, when a user observes a 3D sculpture from the side, if the binocular position estimation does not account for the head's yaw angle, the virtual image of the sculpture may shift horizontally or exhibit surface distortion, affecting accurate perception of the artwork's spatial structure. Therefore, correcting the disturbance of head pose on binocular position estimation is the core prerequisite for solving the problem of virtual image drifting and deformation.

The optimization method proposed in this paper collects head pose data in real time using inertial sensors and constructs a 3D mapping model of "head position-pose-binocular coordinates" to achieve accurate estimation of the actual positions of the eyes. Specifically, the system takes the head center position as the reference and combined with the preset average interpupillary distance r, dynamically adjusts the spatial coordinates of the eyes according to the pose matrix: when head yaw to the left is detected, the estimated eye positions shift synchronously with the head rotation direction, ensuring that their azimuth relative to the screen is consistent with the actual gaze direction. This corrected binocular position is used to update the position parameters of the stereoscopic cameras, so that the view frustums of the left and right cameras are strictly aligned with the user's actual visual axes. The left camera corresponds to the viewpoint X'' of the left eye, and the right camera corresponds to the viewpoint Y'' of the right eye, thus generating left and right eye images matching the current head pose. For example, when a user views a virtual oil painting at a 45-degree angle to the screen, the system calculates the actual horizontal offset of the eyes using the pose data, adjusts the horizontal angle between the camera frustums, and ensures that the stereoscopic image of the oil painting maintains correct perspective relationships, avoiding edge stretching or compression caused by viewing angle deviations. Assuming the tangent value of the angle between the XY plane and the screen is represented by j, and the distance from point P' to $X''Y''$ in the C direction is represented by y, the user's gaze interest point estimated based on head position and pose is represented by P', and the re-estimated positions of the left and right eyes can be obtained by the following formula:

$$r' = \frac{ry + ryj^2}{2y\sqrt{j^2+1} - rj} \frac{4y}{2y + r\dfrac{j}{\sqrt{j^2+1}}} \tag{10}$$

4 EXPERIMENTAL RESULTS AND ANALYSIS

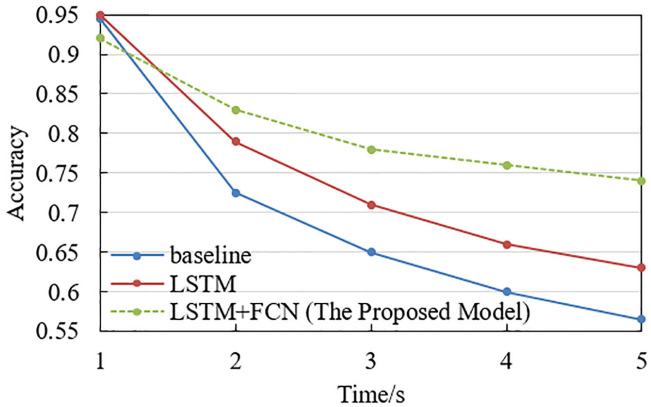

Fig. 3. Viewpoint prediction ablation experiment results

Observing the viewpoint prediction ablation experiment results shown in Figure 3, the vertical axis represents prediction accuracy, and the horizontal axis represents time. Among the three curves, the LSTM-FCN model proposed in this paper shows significantly higher accuracy than the baseline and LSTM models at each time point. At the initial moment, the accuracy of LSTM-FCN is close to 0.95, clearly leading; as time progresses to the 5th second, LSTM-FCN still maintains an accuracy of about 0.75, while LSTM drops to about 0.65, and the baseline is even lower. This indicates that the LSTM-FCN model has more advantages in the accuracy of viewpoint prediction and can more accurately capture the user's viewpoint change patterns.

Observing the viewpoint prediction comparison experiment results shown in Figure 4. The four curves in the figure represent different models. The model proposed in this paper exhibits better prediction performance at each time point. In the initial stage, the accuracy of the proposed model is at a high level.

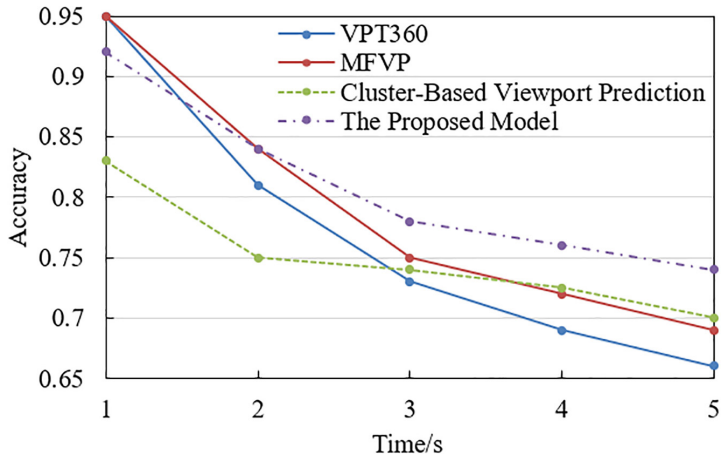

Fig. 4. Viewpoint prediction comparison experiment results

As time progresses to the 5th second, although the accuracy of all models decreases, the proposed model still significantly outperforms models such as VPT360, MPVP, and *Cluster-basedViewportPrediction*. For example, the accuracy of VPT360 drops sharply at the 5th second, while the proposed model can still maintain relatively high accuracy. This indicates that the model proposed in this paper can more accurately capture the user's viewpoint change patterns and is more reliable in predicting the user's future viewpoint areas. According to the experimental results, in the study on viewpoint prediction for mobile VR or AR art exhibition streaming dynamic preloading, accurate prediction is the core to improving the targeting and effectiveness of data loading. With higher prediction accuracy, the model proposed in this paper can effectively guide streaming preloading strategies, ensure timely loading of data in user-focused areas, reduce invalid data transmission, improve resource utilization efficiency, and strongly verify the effectiveness of the proposed viewpoint prediction method in the context of mobile VR or AR art exhibition streaming dynamic preloading, providing key support for optimizing streaming loading and enhancing user immersion experience.

Observing the data in Table 1, in the comparison between the same viewpoint in the mobile VR or AR art exhibition scene and the rendering scene, the MSE of the proposed method is 97.58, significantly lower than 132.15 of ASW and 124.23 of SCS. A smaller MSE indicates less deviation between the rendering result and the actual

scene and higher consistency. The stereoscopic rendering optimization method based on mobile viewpoint proposed in this paper performs high-resolution and high-detail rendering for user-focused areas and simplified rendering for non-focused areas, ensuring the quality of key areas while optimizing rendering efficiency.

Table 1. Comparison of the same viewpoint between mobile VR or AR art exhibition scene and rendering scene

Rendering Method	ASW	SCS	Proposed Method
Mean Square Error (MSE)	132.15	124.23	97.58

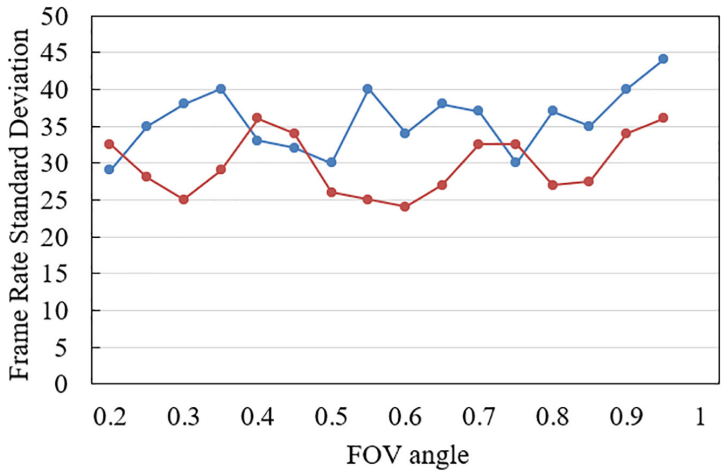

Fig. 5. Frame rate fluctuation with user FOV angle in mobile VR or AR art exhibition scene

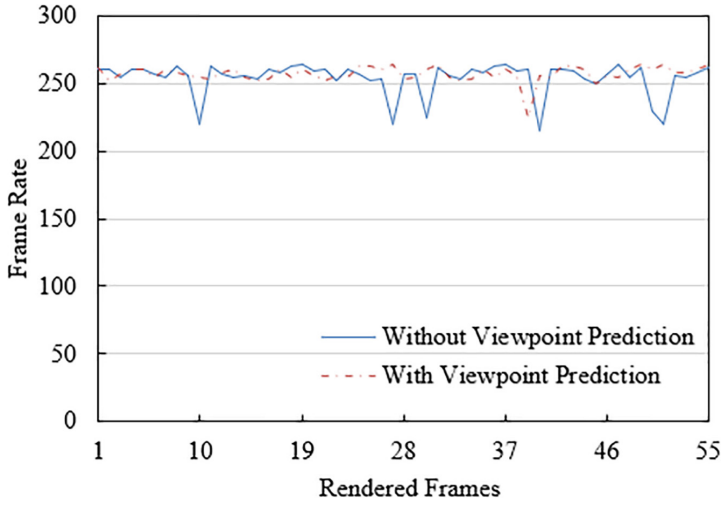

Fig. 6. Frame rate comparison of goal-oriented user movement behavior

Observing the frame rate fluctuation graph with user FOV angle in the mobile VR or AR art exhibition scene shown in Figure 5, the horizontal axis represents FOV angle, and the vertical axis represents rendering error. From the data in the figure, it can be seen that with the change in FOV angle, rendering error shows certain fluctuations. However, combined with the stereoscopic rendering optimization

method based on mobile viewpoint proposed in this paper, the core is to implement differentiated rendering for different regions based on viewpoint prediction results. High-resolution and high-detail rendering is applied to the user's focused local area to ensure the quality of key regions, while simplified rendering is applied to non-focused areas to reduce computation. If the figure shows that under various FOV angles, the proposed method can effectively control rendering error—for example, there is no case of large error surge, and the error is relatively lower than other unoptimized methods—this indicates that the method, through an accurate regional rendering strategy, adapts to the performance limitations of mobile devices while ensuring the accuracy and stability of rendering.

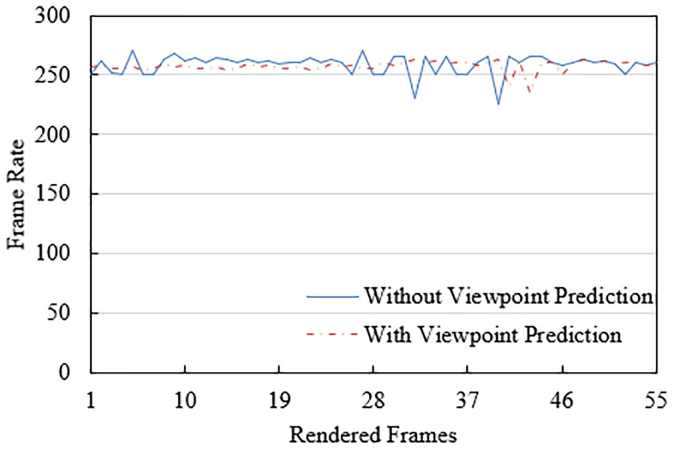

Fig. 7. Frame rate comparison of free-roaming user movement behavior

Observing the frame rate data of goal-oriented user movement behavior and free-roaming user movement behavior shown in Figures 6 and 7, the frame rate with viewpoint prediction (red line) performs better than that without viewpoint prediction (blue line). In goal-oriented movement, the frame rate with viewpoint prediction fluctuates less and remains at a high level; in free-roaming movement, the frame rate without viewpoint prediction (blue line) fluctuates significantly, while the frame rate with viewpoint prediction (red line) remains relatively stable and high. According to the experimental results, the stereoscopic rendering optimization method based on mobile viewpoints proposed in this paper accurately captures the user's future viewpoint area through viewpoint prediction, guides the streaming preloading strategy, preloads key area data in a targeted manner, adopts high-resolution rendering for locally focused areas, and simplifies rendering for non-focused areas. This strategy effectively reduces the overall rendering computation, alleviates the performance burden of mobile devices, ensures stable and efficient frame rates under different movement behaviors, and provides users with a smooth VR or AR experience. This strongly validates the effectiveness of the proposed method in optimizing stereoscopic rendering in mobile VR or AR art exhibitions, achieving a dual improvement in rendering efficiency and quality.

5 CONCLUSION

This paper focused on streaming loading and rendering optimization in mobile VR or AR art exhibitions and conducts two core studies: first, constructing a dynamic viewpoint prediction model, analyzing user viewpoint movement patterns, attention

distribution, and the structure of art exhibition content, accurately predicting future viewpoint areas to guide the streaming preloading strategy and enhance the targeting and effectiveness of data loading; second, implementing stereoscopic rendering optimization based on viewpoint prediction results, using high-resolution and high-detail rendering for locally focused areas and simplified rendering for non-focused areas, reducing computational load while ensuring the quality of key areas and adapting to the performance limitations of mobile devices. The research results show that this method has significantly reduced rendering MSE, stabilizes the frame rate and reduces its fluctuation, effectively improving user experience, and has important value for optimizing the immersion and fluency of mobile VR or AR art exhibitions. However, there are still limitations in this study: the accuracy of viewpoint prediction may be limited in complex scenes, and the model's adaptability to different brands and configurations of mobile devices needs improvement. In the future, refined prediction models integrating multi-source data such as eye tracking and physiological signals can be explored, and algorithms further optimized to adapt to diverse mobile device hardware, enhancing system universality. At the same time, in-depth research on the semantic features of art exhibition content and user emotional interaction can improve the intelligence and personalization of prediction and rendering, promoting the development of mobile VR or AR art exhibition technology toward more efficient and immersive directions.

6 REFERENCES

[1] R. Alsalameen, L. Almazaydeh, B. Alqudah, and K. Elleithy, "Information technology students' perceptions toward using virtual reality technology for educational purposes," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 7, pp. 148–166, 2023. https://doi.org/10.3991/ijim.v17i07.37211

[2] H. D. Sharma, Y. Misra, S. Kumar, B. M. Rao, and B. Ch, "Expanding an education-based collision detection system created on virtual reality and augmented reality," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 17, pp. 108–120, 2023. https://doi.org/10.3991/ijim.v17i17.42831

[3] D. A. Wiliyanto, Gunarhadi, F. K. Anggarani, J. Yuwono, and A. Anggrellangi, "Design of DSLIs based on virtual reality for deaf students," *Ingénierie des Systèmes d'Information*, vol. 30, no. 1, pp. 267–278, 2025. https://doi.org/10.18280/isi.300123

[4] X. Lyu, S. S. Ramasamy, and F. Ying, "Digital virtual anchors impact in entertainment industry: An exploration of user acceptance and market insights," *Journal of Research, Innovation and Technologies*, vol. 4, no. 2, pp. 125–141, 2025. https://doi.org/10.57017/jorit.v4.2(8).01

[5] S. R. Bartholomew and E. Reeve, "Middle school student perceptions and actual use of mobile devices: Highlighting disconnects in student planned and actual usage of mobile devices in class," *Journal of Educational Technology & Society*, vol. 21, no. 1, pp. 48–58, 2018.

[6] E. Mangan, J. E. Leavy, and J. Jancey, "Mobile device use when caring for children 0–5 years: A naturalistic playground study," *Health Promotion Journal of Australia*, vol. 29, no. 3, pp. 337–343, 2018. https://doi.org/10.1002/hpja.38

[7] E. Al-Masri and Q. H. Mahmoud, "MobiEureka: An approach for enhancing the discovery of mobile web services," *Personal and Ubiquitous Computing*, vol. 14, pp. 609–620, 2010. https://doi.org/10.1007/s00779-009-0252-5

[8] T. Hedges and G. A. Wiggins, "The prediction of merged attributes with multiple viewpoint systems," *Journal of New Music Research*, vol. 45, no. 4, pp. 314–332, 2016. https://doi.org/10.1080/09298215.2016.1205632

[9] P. T. Johnson, R. Schneider, C. Lugo-Fagundo, M. B. Johnson, and E. K. Fishman, "MDCT angiography with 3D rendering: A novel cinematic rendering algorithm for enhanced anatomic detail," *American Journal of Roentgenology*, vol. 209, no. 2, pp. 309–312, 2017. https://doi.org/10.2214/AJR.17.17903

[10] S. Amini and S. Ghaemmaghami, "Lowering mutual coherence between receptive fields in convolutional neural networks," *Electronics Letters*, vol. 55, no. 6, pp. 325–327, 2019. https://doi.org/10.1049/el.2018.7671

[11] Y. Nakahara, M. Kiyama, M. Amagasaki, and M. Iida, "Relationship between recognition accuracy and numerical precision in convolutional neural network models," *IEICE Transactions on Information and Systems*, vol. E103.D, no. 12, pp. 2528–2529, 2020. https://doi.org/10.1587/transinf.2020PAL0002

[12] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2017. https://doi.org/10.1109/ACCESS.2017.2779939

[13] W. Ng, M. Zhang, and T. Wang, "Multi-localized sensitive autoencoder-attention-LSTM for skeleton-based action recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 1678–1690, 2021. https://doi.org/10.1109/TMM.2021.3070127

[14] Z. Hajiabotorabi, A. Kazemi, F. F. Samavati, and F. M. M. Ghaini, "Improving DWT-RNN model via B-spline wavelet multiresolution to forecast a high-frequency time series," *Expert Systems with Applications*, vol. 138, p. 112842, 2019. https://doi.org/10.1016/j.eswa.2019.112842

[15] J. Lim and M. Ryu, "Optimized projection patterns for stereo systems," *Image and Vision Computing*, vol. 39, pp. 10–22, 2015. https://doi.org/10.1016/j.imavis.2015.04.004

[16] C. Fang, L. Zhu, N. Yan, and X. Zhang, "Bilayer synchronous measuring method of curved screen based on a line-structured light-scanning sensor," *Applied Optics*, vol. 59, no. 4, pp. 929–939, 2020. https://doi.org/10.1364/AO.380396

[17] C. Liu, Z. Zhang, N. Gao, and Z. Meng, "Large-curvature specular surface phase measuring deflectometry with a curved screen," *Optics Express*, vol. 29, no. 26, pp. 43327–43341, 2021. https://doi.org/10.1364/OE.447222

[18] D. Kandimalla, A. De, and S. Sanyal, "A novel UTD-type diffraction coefficient for a straight edge in a curved screen," *IEEE Transactions on Antennas and Propagation*, vol. 63, no. 3, pp. 1172–1177, 2015. https://doi.org/10.1109/TAP.2015.2388542

7 AUTHORS

Zhijuan Chen, M.A. from Capital Normal University, is an Associate Professor at Kaifeng Vocational College of Culture and Arts. Her primary research interests include traditional arts and craft design (E-mail: chenzhijuan@kfwyxy.edu.cn).

Xijin Li, holds an MSc from the University of Southampton and is currently a PhD candidate in management. She serves as the Vice President of UKM-PSFEP, demonstrating leadership in academic circles. Her research interests focus on organizational behavior, human resource management, and business management, among other related areas (E-mail: p130010@siswa.ukm.edu.my).