Temporo-Spatial Inaccuracy of Video See-Through Augmented Reality Head-Mounted Displays

Chumin Zhao and Ryan Beams
Center for Devices and Radiological Health, U.S. Food and Drug Administration

Abstract
Latency and motion lead to temporo-spatial inaccuracy on video see-through (VST) head-mounted displays (HMDs). We develop a bench setup that enables translation and rotation of a physical target and HMD. VST latency and temporo-spatial inaccuracy are evaluated using various motion schemes involving target translation, HMD translation, and HMD rotation.

Author Keywords
Temporo-spatial accuracy, video see-through, augmented reality, head-mounted display, latency, temporal warping.

1. Introduction
Video see-through (VST) augmented reality (AR) mixes a digital scene into the virtual physical passthrough using outward-facing cameras, which differs from optical see-through (OST) AR with a real physical visualization. Unfortunately, VST content on a head-mounted display (HMD) is subject to image quality degradation in both spatial and temporal domains due to camera and display limitations, geometric distortion, and latency. Among the image artifacts, spatial accuracy of the physical view is particularly important for potential uses of VST AR HMDs in medical applications [1]. Two major sources contributing to the VST spatial inaccuracy are geometric distortion of the physical passthrough and latency-induced temporo-spatial error with motion. It has been reported that mismatched camera and eye positions is the primary cause for geometric distortion on VST AR HMDs in static [2]. When the target or HMD (user's head) is in motion, the motion-to-photon (i.e., physical-to-VST) latency creates a spatial misalignment between the physical and VST visualization that we define as the temporo-spatial inaccuracy. Evaluation methods to quantify temporo-spatial inaccuracy on VST AR HMDs have not been established in standards.

In this work, we develop experimental setups and methodologies to evaluate latency and temporo-spatial inaccuracy on two VST AR HMDs. The temporo-spatial measurements are repeated using various motion configurations involving target translation, HMD translation, and HMD rotation. In addition, we develop a temporal warping model to simulate the effect of motion correction on temporo-spatial accuracy.

2. Methods
VST AR HMDs: Two VST AR HMDs were evaluated in this study: the Meta Quest 3 and HTC VIVE XR Elite. Both HMDs implement a fast-switching LCD technology featuring about a 2k × 2k pixel resolution per eyepiece with a 90 Hz display refresh rate. The Quest 3 captures a binocular view of the physical world using two VST cameras integrating at 60 Hz, while the VIVE XR Elite use a single VST camera with a refresh rate of 90 Hz.

VST Latency Measurement: As illustrated in Fig. 1(a), prior to evaluate the temporo-spatial inaccuracy, we first measured the VST latency without motion. The bench setup is similar to a video latency measurement on a flat-panel display [3]. Specifically, a white target LED was pulsed at a low frequency of 5 Hz as the reference input signal from the physical world. The optical output from the VST AR HMD was captured using a photodiode placed approximately at the eye position. As shown in Fig. 1(b), both the input (5 Hz LED pulses) and output waveforms (90 Hz HMD illumination) were shown on an oscilloscope. The VST latency ($t_{latency}$) is determined by the delay between the rising edges of the input pulse and VST output. The measured $t_{latency}$ is about 50.4 ± 5.3 ms and 52.4 ± 4.3 ms on the Meta Quest 3 and HTC VIVE XR Elite HMDs respectively, by repeating the measurements five times on each evaluated HMD.

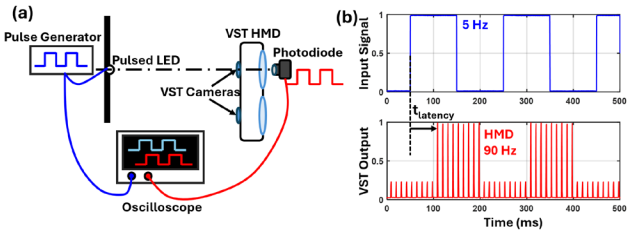

Figure 1 (a) Schematic illustration and (b) example measurement of VST latency using a photodiode setup with pulsed LED input signal.

Experimental Setup for Temporo-Spatial Inaccuracy Measurements: Although the photodiode method is very effective to capture the VST latency, it cannot measure the temporo-spatial error due to lack of a spatial component in the temporal data (i.e., waveforms). To evaluate the temporo-spatial inaccuracy of VST AR HMDs, as shown in Fig. 2, we developed a bench setup using two synchronized monochromatic cameras (Teledyne FLIR Blackfly S, BFS-U3-04S2M-C, Teledyne FLIR, USA) that can capture video recordings up to 522 frame per second (fps). A 5 mm focal length wide field of view (FoV) lens (f/2.8) with a low distortion (<0.5%) was attached to each camera. Initially, without placing the HMD, the synchronized cameras were mounted vertically such that their optical axes were aligned with the physical target (a point white LED light source) at zero position. As illustrated in Fig. 2(b) and (c), the top camera is used to capture the physical target positions (x_{phys}) without occlusion of the HMD. Then we mounted the VST HMD such that the bottom camera is aligned to the VST signal approximately at the right eye position. The bottom camera was used to measure position of the digitized target passing through the HMD (x_{VST}). During the measurement, the synchronized cameras were operated at a frame rate of 90 Hz matching that of the HMD to capture two synchronized video streams for 30 seconds.

The VST AR HMD and synchronized cameras were mounted on a motorized rotary stage (RSW60A-E03T3A, Zaber Technologies, Canada) with a maximum rotation speed of 115°/s that emulates the head rotation (φ_{HMD}) in the axial plane. The target was placed at a 90 cm distance from the rotational axis, while the HMD was mounted 7 cm in front of the rotational axis, which is approximately the distance from the front of the head to the rotation axis. The rotary stage was mounted on a set of linear translational stages (X-LC40B1000, Zaber Technologies, Canada) with a 1000 mm travel range in lateral (x_{HMD}) and depth

ISSN 0097-996X/25/5602-2317-$1.00 © 2025

direction (z). The translational stages simulate linear head movement of the user. The physical target was mounted on another linear translation stage (X-LSQ450B-E01C, Zaber Technologies, Canada) with a maximum speed of 240 mm/s.

Note that the synchronized-camera setup features a wide FoV measurements that emulates a pupil rotation scheme [3,4]. To enable eye rotation, the cameras need to track the target instantaneously, which will substantially add complexity to the bench setup. Nevertheless, as the measurement goal is to quantify latency-induced spatial error, we believe the wide FoV measurement is capable to measure the spatial difference between the target positions captured by the cameras based on the results presented in this study. For static spatial measurement such as geometric distortion, eye rotation geometry is recommended [2].

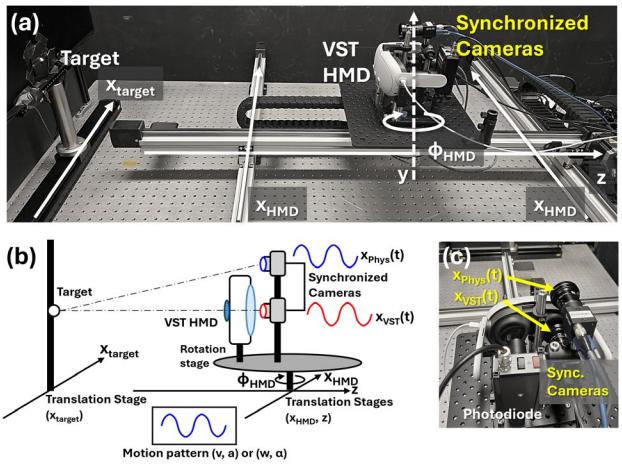

Figure 2 (a) Experimental setup and (b) schematic illustration of temporo-spatial inaccuracy measurement on VST AR HMDs using synchronized cameras as shown in (c) to capture video streams of the target position in the physical ($x_{phys}(t)$) and digital ($x_{VST}(t)$) domains.

Motion Schemes: Three motion configurations were investigated: **1) Target translation**: the physical target was in continuous motion on the translation stage that varies the x-position of the target (x_{target}), while the HMD was static; **2) HMD translation**: the physical target was fixed at the zero position, while the HMD was in linear motion along the x-direction (x_{HMD}) without HMD rotation; and **3) HMD rotation**: x_{target} and x_{HMD} were fixed at the zero position, while the HMD was in rotational motion by an angle φ_{HMD}.

The motion schemes were determined by three parameters: the range of motion, the maximum velocity, and acceleration. The active stage accelerates at a constant acceleration until it reaches its maximum velocity and continues motion with the constant velocity. Then the stage will deaccelerate with the same acceleration such that the velocity is zero at the maximum travel distance determined by the motion range.

Data Analysis: Two synchronized video streams of the target were recorded at 90 Hz matching the HMD refresh rate for each motion configuration. In each frame at time t, the center of the target was determined by a simple segmentation algorithm using MATLAB. As the HMD and target only move in the horizontal plane, we only recorded the x (or φ) position of the target in physical ($x_{phys}(t)$ or $\varphi_{phys}(t)$) and VST ($x_{VST}(t)$ or $\varphi_{VST}(t)$) spaces. To eliminate the spatial error from geometric distortion of VST AR HMDs, the $x_{VST}(t)$ and $\varphi_{VST}(t)$ were normalized such that the travel range matches the physical recordings. This normalization allows for separation of spatial geometric distortion error and temporo-spatial error caused by latency. In theory, as illustrated in Fig. 3, latency ($t_{latency}$) adds a temporal delay on $x_{VST}(t)$ resulting in a temporo-spatial error $\Delta x_{ts}(t) = x_{VST}(t) - x_{phys}(t)$. For each measurement, the root mean square error (RMSE) of $\Delta x_t(t)$ (i.e., $\Delta x_{ts,rmse}$ in mm or degree) was used as the figure of merit to quantify the temporo-spatial inaccuracy of VST AR HMDs.

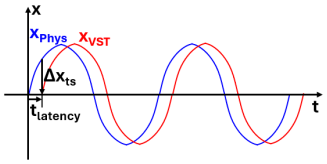

Figure 3 Schematic illustration of target position profiles in the physical ($x_{phys}(t)$) and VST ($x_{VST}(t)$) domains. $\Delta x_t(t)$ represents the temporo-spatial error induced by latency.

3. Results

Target Translation: Fig. 4(a) and (b) show the extracted $x_{phys}(t)$ and $x_{VST}(t)$ obtained using the synchronized camera acquisition method on the Meta Quest 3 HMD under the target translation motion scheme. For both target and HMD translation, the translation range was fixed at ±150 mm (x_{range} of 300 mm) with an acceleration of a = 50 mm/s^2. The maximum translational velocity (v_{max}) was varied from 30 to 122 mm/s. The velocity profile ($v(t)$ calculated as the derivative of dx_{phys}/dt) is also shown for comparison with the temporo-spatial error $\Delta x_{ts}(t)$. It is illustrated in Fig. 4(a) and (b) that for target translation, $\Delta x_{ts}(t)$ can be estimated as the product of target motion velocity and latency, i.e., $\Delta x_{ts}(t) \approx -v(t) \cdot t_{latency}$. The result on the HTC VIVE XR Elite HMD demonstrates the same trend. As the latency is about the same, there is no major difference in $\Delta x_{ts,rmse}$ between the two evaluated HMDs under target translation, which increases from about 1.7 to 3.6 mm by increasing the maximum velocity from 30 to 122 mm/s as shown in Fig. 5(a).

HMD Translation: Fig. 4 (c) and (d) show the temporo-spatial measurement results for HMD translation using the same motion configurations as (a) and (b) (i.e., same translation range, maximum velocity, and acceleration) on the Meta Quest 3 HMD. As shown in Fig. 5(a) and (b), the temporo-spatial error for HMD translation is smaller than that with target translation under the same motion characteristics on the Meta Quest 3 HMD. On the other hand, there is no obvious difference in temporo-spatial error between target and HMD translation on the HTC VIVE XR Elite HMD (see Fig. 5(a, b) and Fig. 6) with various maximum velocity. The assumption of $\Delta x_{ts}(t) \approx -v(t) \cdot t_{latency}$ is not valid for the Meta Quest 3 under HMD translation, indicating potential implementation of motion compensation (e.g., temporal warping [5]) on the HMD.

HMD Rotation: Fig. 7 show the temporo-spatial measurement results for HMD rotation on the Meta Quest 3 (a, b) and HTC VIVE XR Elite HMDs (c, d). The setup emulates human head rotation with a rotation range of ±10° (φ_{range} of 20°), a constant angular acceleration of α = 50°/s^2, and a maximum angular velocity (ω_{max}) ranging from 10 to 31.6°/s. It is indicated that temporal warping performs an instantaneous rotational correction on the angular pose to minimize the impact of latency on temporo-spatial inaccuracy. As shown in Fig. 5(c), the RMSE of temporo-spatial error ($\Delta\varphi_{ts,rmse}$) is less than 0.13° at the maximum angular velocity of 31.6°/s. We suspect that the difference in temporo-

spatial error with HMD rotation between the two HMDs is due to different timing scheme for motion compensation.

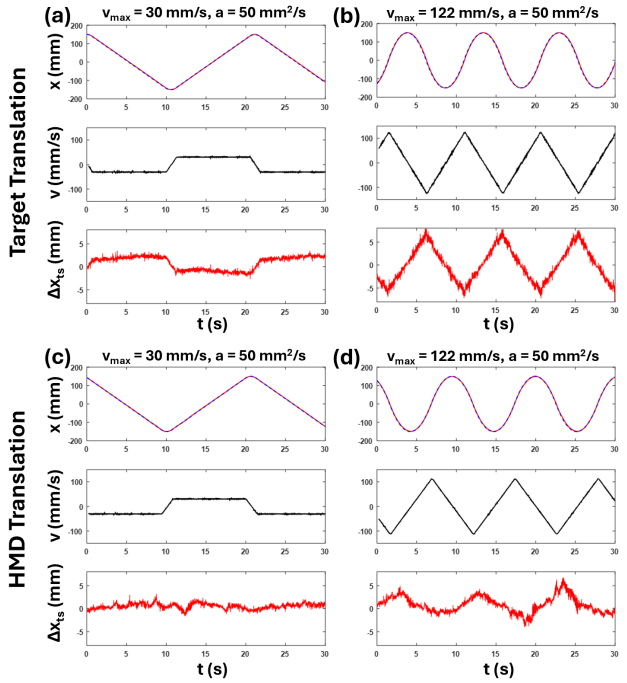

Figure 4 Target physical (blue dash line) and VST positions (red line) on the Meta Quest 3: x(t) (top row), velocity of target: v(t) (second row), and temporo-spatial error: $\Delta x_{ts}(t)$ (third row) for (a) target translation with maximum velocity of 30 and (b) 122 mm/s, and (c) HMD translation with maximum velocity 30 and (d) 122 mm/s. The acceleration is fixed at 50 mm^2/s.

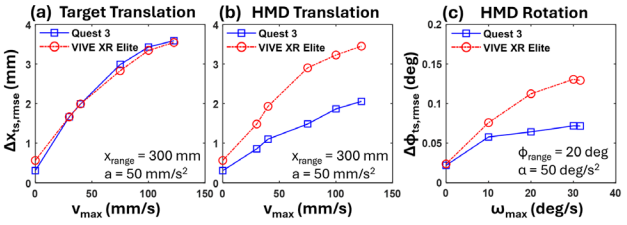

Figure 5 Measured temporo-spatial error under various maximum velocity configurations on the Meta Quest 3 and HTC VIVE XR Elite HMDs for (a) target translation, (b) HMD translation, and (c) HMD rotation.

Temporal Warping: As indicated in the temporo-spatial characteristics, temporal warping can be applied to partially compensate the latency induced spatial inaccuracy, when the HMD is in motion. It should be clarified that temporal warping cannot correct the temporo-spatial inaccuracy for target motion. In general, 3D positional warping and 2D rotational warping can be implemented for motion prediction and compensation of HMD translational and rotational errors, respectively [5]. Specifically, the inertial measurement unit (IMU) of the HMD constantly tracks the rotation speed and linear acceleration using gyro and accelerometer providing a 2D rotational pose of the HMD. On the other hand, 3D position of the HMD is determined by the tracking cameras that operates at a much lower refresh rate compared to the IMU.

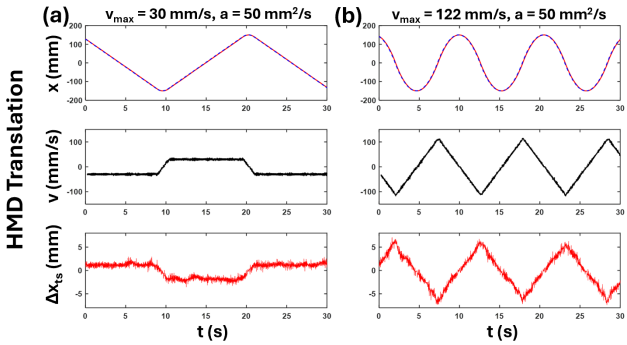

Figure 6 Target physical and VST positions on the HTC VIVE XR Elite: x(t) (top row), velocity of target: v(t) (second row), and temporo-spatial error: $\Delta x_{ts}(t)$ (third row) for (a) HMD translation with maximum velocity of 30 and (b) 122 mm/s. The acceleration is fixed at 50 mm^2/s.

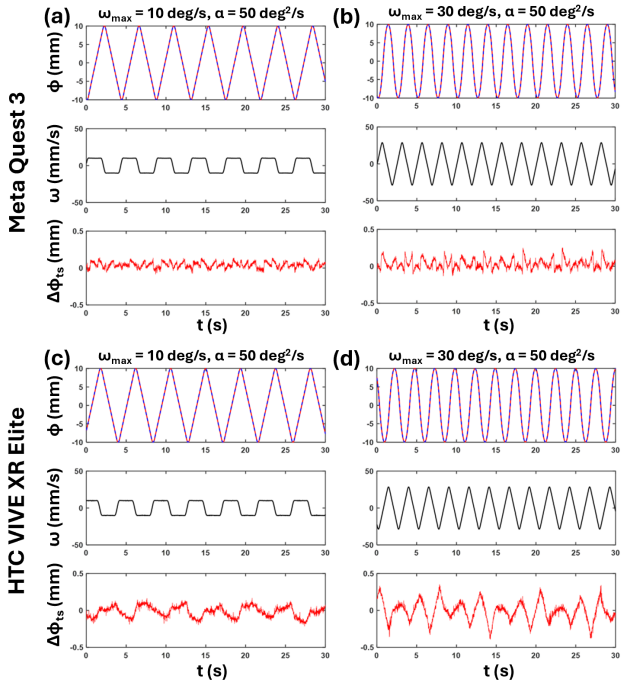

Figure 7 Target physical and VST angular positions: φ(t) (top row), velocity of target: ω(t) (second row), and temporo-spatial error: $\Delta\varphi_{ts}(t)$ (third row) on the Meta Quest 3 (a,b) and HTC VIVE XR Elite (c,d) for HMD rotation with maximum angular velocity of 10 (a,c) and 30°/s (b,d). The angular acceleration is fixed at 50 deg^2/s.

We simulate the 3D positional temporal warping and 2D rotational temporal warping using the timing shown in Fig. 8, assuming a 90 Hz refresh rate on display rendering and warping. Without temporal warping, e.g., for target translation, the VST content is delayed by the time between camera integration (t_c) and display emission (t_d). For HMD translation, starting from time t_c, the 3D position of the HMD is determined by tracking cameras (e.g., x_c) with an integration frame rate of f_{cam}. 3D positional temporal warping can be performed to predict the position of the HMD at display cycle t_d using velocity and acceleration after the integration phase at time t_r. Effectively the predicted position of 3D positional temporal warping is estimated by $x_{VST,warp} = x_c +$

$v_r(t_d - t_r) + a_r(t_d - t_r)^2/2$ in our model. In Fig. 8, it is shown that the motion during camera integration time from t_c to t_r cannot be corrected. Therefore, it is desirable to shorten the camera integration time for reduced temporo-spatial inaccuracy. As illustrated in Fig. 9, increasing f_{cam} from 30 to 90 Hz can substantially reduce $\Delta x_{ts,rmse}$ especially for high-speed HMD translation. However, a shorter VST and tracking camera integration time may affect the VST image quality and signal-to-noise ratio. In addition, translation of the 3D scene can lead to occlusion and misalignment between the eye and virtual camera positions resulting in additional visualization artifact without proper mitigation.

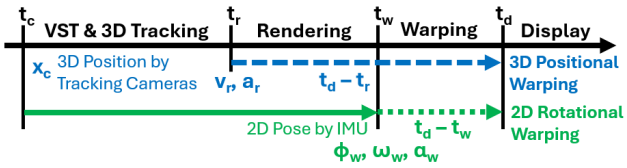

Figure 8 Timeline for 3D positional and 2D rotational temporal warping, where t_c, t_r, t_w, and t_d represents the starting time for camera integration, rendering, warping, and display emission phases in a frame cycle.

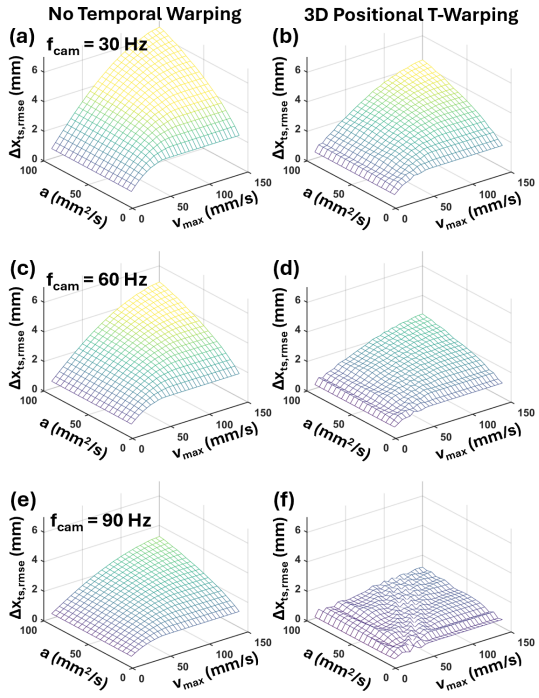

Figure 9 Simulated temporo-spatial error under various HMD translation configurations without (a,c,e) and with 3D positional temporal warping (b,d,f) with a camera frame rate from 30 to 90 Hz.

For HMD rotation, as illustrated in the timing budget in Fig. 8, current rotational pose φ_w can be obtained before temporal warping phase at time t_w from the fast-sampling IMU. 2D rotational temporal warping predicts the angular position for the warping phase $t_d - t_w$ by rotating the scene, i.e., $\varphi_{VST,warp} = \varphi_w + \omega_w(t_d - t_w) + \alpha_w(t_d - t_w)^2/2$. As shown in Fig. 10, the simulation result shows that 2D rotational temporal warping is very effective in correction of the rotational temporo-spatial error at high rotation speed up to 100°/s. It is also easier to implement compared to 3D positional temporal warping.

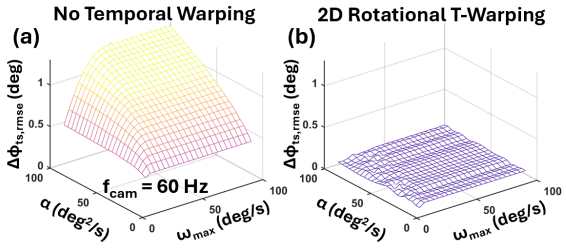

Figure 10 Simulated temporo-spatial error under various HMD rotation configurations without (a) and with 2D rotational temporal warping (b) with a camera frame rate of 60 Hz.

4. Conclusions

We present experimental and data analysis methodologies to evaluate the temporo-spatial inaccuracy on VST AR HMDs. The results show that for target translation, the temporo-spatial error is generally proportional to the motion velocity and latency. However, different temporo-spatial characteristics are observed for HMD translation and rotation using temporal warping technique. We perform analytical simulations to investigate the impact of 3D positional and 2D rotational temporal warping on temporo-spatial accuracy. It is critical to optimize the camera integration time and timing of temporal warping to minimize the temporo-spatial inaccuracy. We believe the findings in this study can be used to evaluate the HMD latency and temporal-spatial inaccuracy and guide the design of future VST AR HMDs.

5. Acknowledgements

The authors would like to acknowledge Dr. Ashraf Bader (CDRH/FDA) for technical review of this manuscript.

6. Disclosures

The mention of commercial products, their resources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.

7. References

1. Arensmeyer J, *et al*. A System for Mixed-Reality Holographic Overlays of Real-Time Rendered 3D-Reconstructed Imaging Using a Video Pass-through Head-Mounted Display—A Pathway to Future Navigation in Chest Wall Surgery. *Journal of Clinical Medicine*. 2024;13(7):2080.
2. Zhao C, Beams R. Geometric distortion on video see-through head-mounted displays. *J Soc Inf Display*. 2024; 32(5): 184–193.
3. Information Display Measurement Standards, version 1.2, 2023, *SID*.
4. IEC 63145-20-10:2019 *Eyewear display - Part 20-10: Fundamental measurement methods – Optical properties*.
5. Ishihara A, *et al*. Integrating both parallax and latency compensation into video see-through head-mounted display. IEEE Transactions on Visualization and Computer Graphics. 2023;29(5):2826-36.