Understanding User Behavior in Volumetric Video Watching: Dataset, Analysis and Prediction

Kaiyuan Hu†
FNii and SSE, CUHK-Shenzhen
kaiyuanhu@link.cuhk.edu.cn

Haowen Yang†
FNii, CUHK-Shenzhen
Versee Inc.
119010377@link.cuhk.edu.cn

Yili Jin
McGill University
yili.jin@mail.mcgill.ca

Junhua Liu
FNii, CUHK-Shenzhen
junhualiu@link.cuhk.edu.cn

Yongting Chen
FNii and SSE, CUHK-Shenzhen
yongtingchen@link.cuhk.edu.cn

Miao Zhang
Simon Fraser University
mza94@sfu.ca

Fangxin Wang*
SSE and FNii, CUHK-Shenzhen
The Guangdong Provincial Key Laboratory of Future Networks of Intelligence
wangfangxin@cuhk.edu.cn

ABSTRACT

Volumetric video emerges as a new attractive video paradigm in recent years since it provides an immersive and interactive 3D viewing experience with six degree-of-freedom (DoF). Unlike traditional 2D or panoramic videos, volumetric videos require dense point clouds, voxels, meshes, or huge neural models to depict volumetric scenes, which results in a prohibitively high bandwidth burden for video delivery. Users' behavior analysis, especially the viewport and gaze analysis, then plays a significant role in prioritizing the content streaming within users' viewport and degrading the remaining content to maximize user QoE with limited bandwidth. Although understanding user behavior is crucial, to the best of our best knowledge, there are no available 3D volumetric video viewing datasets containing fine-grained user interactivity features, not to mention further analysis and behavior prediction.

In this paper, we for the first time release a volumetric video viewing behavior dataset, with a large scale, multiple dimensions, and diverse conditions. We conduct an in-depth analysis to understand user behaviors when viewing volumetric videos. Interesting findings on user viewport, gaze, and motion preference related to different videos and users are revealed. We finally design a transformer-based viewport prediction model that fuses the features of both gaze and motion, which is able to achieve high accuracy at various conditions. Our prediction model is expected to further benefit volumetric video streaming optimization.

Our dataset, along with the corresponding visualization tools is accessible at https://cuhksz-inml.github.io/user-behavior-in-vv-watching/

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; • **Information systems** → **Multimedia databases**.

KEYWORDS

Volumetric videos, Dataset, User Behavior Analysis

ACM Reference Format:
Kaiyuan Hu, Haowen Yang, Yili Jin, Junhua Liu, Yongting Chen, Miao Zhang, and Fangxin Wang. 2023. Understanding User Behavior in Volumetric Video Watching: Dataset, Analysis and Prediction. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3613810

1 INTRODUCTION

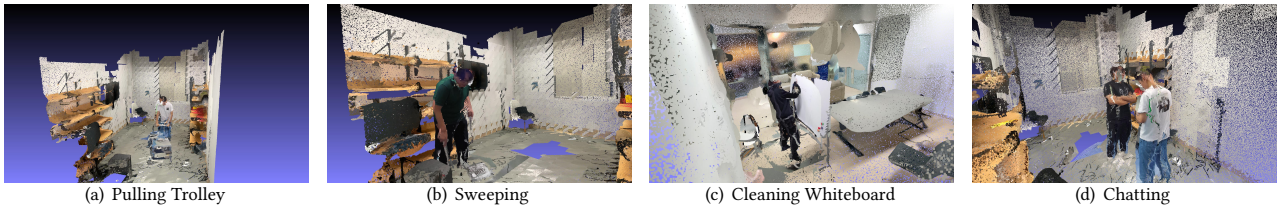
The confluence of video and the recently booming 3D representation technology embraces a new video paradigm, i.e., the *volumetric video* (VV). Different from traditional 2D video that has mature codecs based on frames and pixels, volumetric video is still in its infant stage with various representation formats, such as point cloud [11, 34], voxel [33], mesh [34], and even neural representations [23]. Volumetric video is envisioned as a fundamental service that is able to facilitate various new applications such as extended reality (XR) and Metaverse, empowering entertainment [20], healthcare [7], and education [2], etc. The global industry VV market is expected to reach 22.5 billion USD by 2024 [21].

Unlike traditional or 360-degree videos that only provide flat or curved 2D experience, volumetric video captures the scene and objects in 3D format, providing 6 degree-of-freedom (DoF) viewing

*Fangxin Wang is the corresponding author.
†Both authors contributed equally to this research.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3613810



(a) Pulling Trolley (b) Sweeping (c) Cleaning Whiteboard (d) Chatting

Figure 1: Example of used VV: 'Pulling Trolley', 'Sweeping', 'Cleaning Whiteboard', and 'Chatting'

experience, including three dimensions of position (X, Y, Z) and three dimensions of orientation (yaw, pitch, roll). This new viewing paradigm revolutionizes the way we consume video content, offering an unprecedented full immersive and interactive experience. Such interactivity between the user and the 3D video already demonstrates great value in various fields, e.g., revealing mental activity, inferring user preference, and even identifying different users.

Due to the extreme complexity in volumetric video representation, e.g., extensive points or meshes using point cloud or 3D mesh formats, or huge neural models using implicit neural representation, the size of a volumetric video is usually much larger (up to 100x) than the 2D representation in the same condition. Thus, streaming volumetric video through the current network infrastructure tends to become a key challenge. Users' behavior analysis, especially the field of view (FoV) and gaze analysis, then plays a significant role because we can prioritize the content streaming within FoV and reduce or even ignore the content out of FoV to maximize user's QoE with limited network transmission capacity [8].

Although understanding user behavior is crucial, to our best knowledge, there is no available 3D volumetric video viewing datasets containing fine-grained user interactivity features. Pioneer researchers in the community of multimedia have contributed some 3D datasets on objects or scenes [9, 26], but they never focus on the analysis and understanding of user behavior in volumetric video. Thus, an open dataset in this context is in urgent need to reveal the viewing characteristics, optimize the video streaming, and further facilitate the research in the related community.

In this paper, we propose the first large-scale user behavior dataset on volumetric video viewing with rich dimensions across various scenes, including the six DoF viewport, gaze, and motion features. We next conduct a comprehensive data analysis to deeply understand the user behavior, fully capture the potential correlations among viewport, gaze, and motion trajectory, and further reveal the future viewing activity. We find that VV users exhibit distinct regions of interest and display varying movement patterns based on different scenarios and personalities. Based on our observations and findings, we conduct a pilot study on viewport adaptive 3D volumetric video streaming. We design a transformer-based model to well capture the inherent relationship between the motion and gaze, and further achieve an accurate and robust viewport prediction for video streaming optimization.

The contributions of our work are summarized as follows:

- ▷ We for the first time release a volumetric video viewing behavior dataset, with large scale (50 users), multiple dimensions (8 attributes), and diverse conditions (including both static and dynamic scenes, both single and multi-user activities).
- ▷ We conduct an in-depth analysis to understand user behaviors when viewing volumetric videos. Interesting findings on user viewport, gaze, and motion preference related to different videos and users are revealed.
- ▷ We design a transformer-based viewport prediction model that fuses the features of both gaze and motion, which is able to achieve high accuracy and strong robustness.

Table 1: User Information

Gender	Female		Male	
	27		23	
Age	16-20	20-24	24-30	30+
	25	17	5	3
VR Exp (Times)	Never	1-5	6-10	10+
	32	9	6	3
VV Exp (Times)	Never	1-5	6-10	10+
	41	3	3	3

The rest of this paper is organized as follows. Section 2 gives an overall description of the dataset, including how data is collected as well as the video and dataset attribute description. Section 3 gives an initial visualization of the dataset, plotting headset movement and gaze direction. Section 4 introduces our analysis of user behavior in detail, and also reveals some interesting findings based on our observation. Motivated by these, Section 5 proposes a transformer-based viewport prediction for six DoF volumetric video viewing. We further give some potential applications in Section 6 and conclude this work in Section 7.

2 DATASET

In this section, we introduce the details of our dataset regarding the collection procedure, dataset description, and user information.

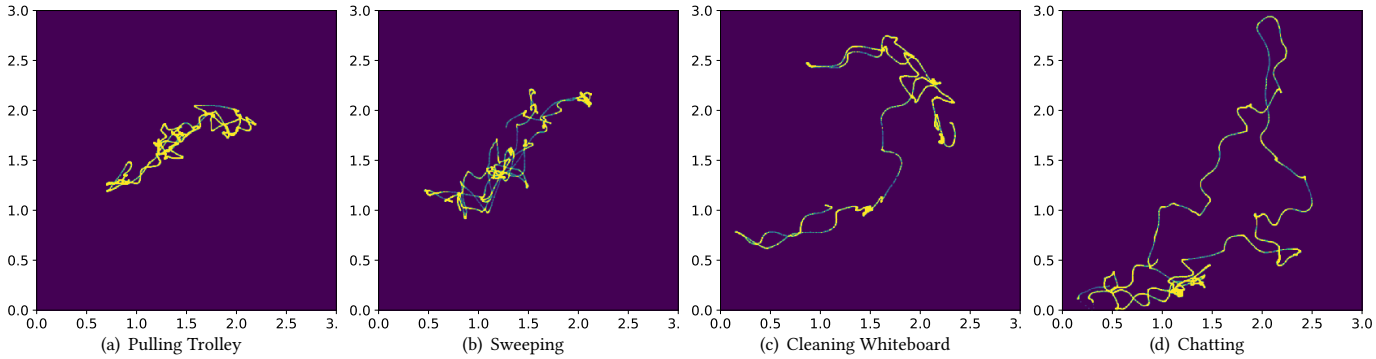
2.1 Data Collection Procedure

For convenience, we select volumetric videos from the current most appropriate public volumetric dataset FSVVD [9] related to our context, which contains 26 volumetric videos represented by point cloud covering multiple common scenarios such as education, exercise, daily life, and entertainment. We seek 50 volunteers to participate in this dataset collection. These volunteers are given enough time and guidance to get familiar with the 3D volumetric environment. Videos are preloaded and played through Unity[1] when a volunteer is wearing a Meta Quest Pro[2] headset. People are able to freely navigate the 3D scenes and watch the activities from

[1] https://unity.com/
[2] https://www.meta.com/quest

Table 2: Description of selected volumetric videos:

Name	#Actors	Spatial Movements	Body Movements	Environment Interaction	#Frame
Chatting	2	Small	Small	-	300
Cleaning Whiteboard	1	Static	Large	✓	300
News Interviewing	2	Small	Small	-	300
Pulling Trolley	1	Large	Small	✓	300
Presenting	2	Static	Small	-	300
Sweeping	1	Middle	Middle	✓	300



(a) Pulling Trolley (b) Sweeping (c) Cleaning Whiteboard (d) Chatting

Figure 2: The aerial view for movement trajectory heatmap of different volumetric scenes. The lighter yellow color indicates a longer dwelling time and vice versa for the darker blue color.

any viewing angle and any position within a 5x5 square meters space, as required by the FSVVD video dataset.

The VR headset has a built-in accelerometer and we are able to easily calculate the current headset position (X,Y,Z) and the rotation of the headset (yaw, pitch, and roll). Besides, gaze information is also important as it provides more fine-grained features [12]. For the gaze data collection, we rely on the built-in eye tracker in the headset with a sample rate of 144 Hz. The collected data consisted of 8 dimensions, including 3 rotational angles corresponding to the position of each eye, plus the confidence level. Since there are subtle differences (usually less than 3°) in the gaze data between the two eyes, we use the weighted average of the two eyes as the gaze in our later analysis.

2.2 Viewer Selection

Different viewers can also have quite personalized preferences on the same video content and conduct diverse behaviors. Therefore, we try our best to choose volunteers with different backgrounds, majors, hobbies, ages, genders, and familiarity levels with VR. Detailed information is listed in Table 1. Once the recording ends, the volunteers are asked to fill out a questionnaire about these information, and the overall experience of watching volumetric videos.

2.3 Video Selection

We argue that the video content should have a significant impact on the viewer's behavior feature. A viewer's attention can largely change if provided with different video content. To analyze the impact of video content on users, we selected 6 different scenes aiming to cover more representative scenarios. Specifically, we mainly evaluate the impact of actor numbers and the movement level of the actors. We divide the movement of target actors as spatial movement (e.g., moving from one position to another) and self-movement (e.g., body movement without obvious position change). Table. 2 indicates the detailed taxonomy of our selected video.

2.4 Dataset Description

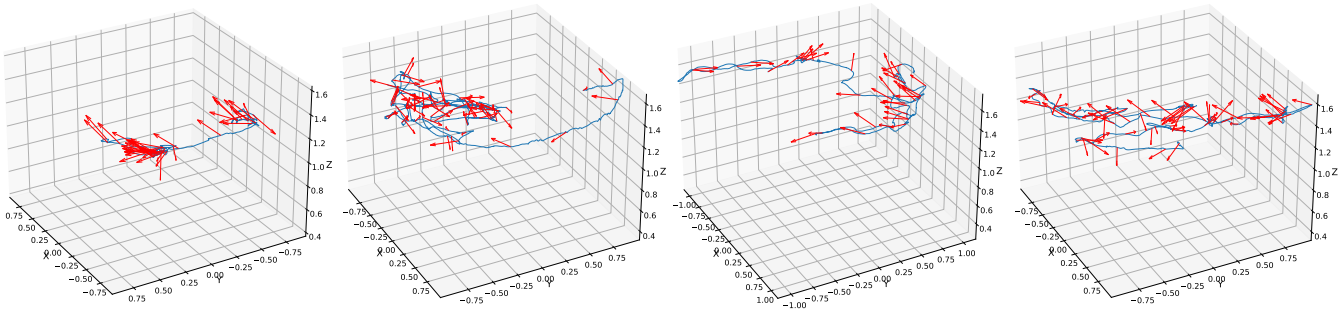
Our collected dataset consists of 28 dimensions, including the frame number and time stamp of each sample, the spatial movement (the spatial coordinates of X, Y, and Z axes), the rotational orientation (rotation angles of Yaw, Pitch, and Roll) information of the headset and two wireless controllers, and the gaze information of both eyes with two confidence indexes.

3 VISUALIZATION

To help better understand our dataset and promote further study, we first give a visualization of the dataset and provide preliminary analysis on headset movement and gaze information. We select four representative scenes for observation and subsequent analysis, i.e., pulling trolley, sweeping, cleaning whiteboard, and chatting.

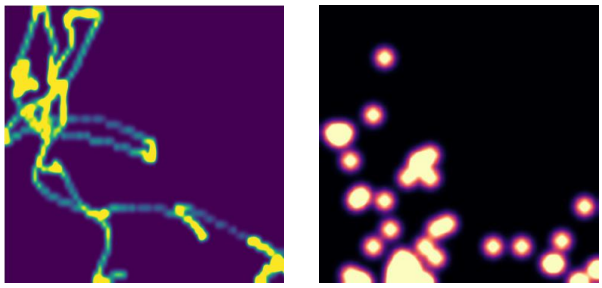
3.1 Headset Movement Trajectory

We first observe the user movement, represented by the headset movement trajectory in our dataset. Among all the participating volunteers, we randomly select one and compare his/her movement trajectory. According to our observation, the values at the Z axis almost keep stable. This is because people rarely crouch down and stand up, which follows our intuition about people's behaviors. Thus, we select to use an aerial view to better depict the trajectory. Fig. 2(a) shows the heatmap of movement trajectory across different scenes from a randomly selected user. Some interesting findings can be obtained. For 'Pulling Trolley' in Fig. 2(a) and 'Sweeping' in Fig. 2(b), the movement trajectories are relatively uniform and



(a) Pulling Trolley (b) Sweeping (c) Cleaning Whiteboard (d) Chatting

Figure 3: Gaze Direction with Movement Trajectory. The blue line represents the movement trajectory and the red arrows indicate the gaze direction.



(a) Original Trajectory (b) Gaze Intersection Points

Figure 4: Illustration of the intersection between movement trajectory and gaze ray. The left figure shows the original user movement trajectory, and the right figure indicates the point where the movement trajectory coincides with the gaze ray.

concentrated, indicating a slow movement within a small region. This matches our findings that **for volumetric videos with large movement, viewers tend to follow the moving object and are prone to pay more attention therein**. While for 'Cleaning Whiteboard' in Fig. 2(c) and 'Chatting' in Fig. 2(d), the trajectory is more dispersive. This indicates that **for small-movement or even static scenes, viewers may go around and observe the object more from different angles**.

3.2 Gaze Direction

Users' gaze information is also a significant indicator of user VV interactivity. We then try to visualize the gaze direction in our dataset. However, different from the traditional 2D video, the gaze can be simply projected onto the video surface, in 3D volumetric scene, the starting point of the gaze is changing along with the movement. Therefore, we need to combine these two together.

Since the rotational angles returned from the headset are represented using degrees in Euler angles, for the convenience of subsequent calculation and visualization, we transform the data into a rotation matrix. We convert the angle into radians to compute the viewport area, where α, β, and γ stand for yaw, pitch, and roll, respectively.

As denoted in Eq. 1, matrix R comprises the product of the rotation matrices about the yaw, pitch, and roll axes to represent the rotation matrix of users' headset movement and gaze movement.

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\gamma & -\sin\gamma \\ 0 & \sin\gamma & \cos\gamma \end{bmatrix} \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \begin{bmatrix} \cos\alpha & \sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

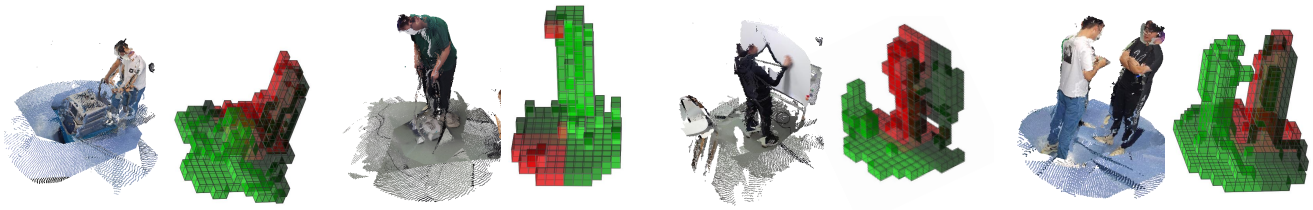
Using the above transformation formula, we get the rotation matrix of both the gaze and the headset. To transform the gaze direction from the local coordinate system of the headset to the global coordinate system, we apply a rotation matrix using the orientation data provided by the headset, representing the transformation from the local coordinate system of the headset to the global coordinate system. The above transformation could be represented as:

$$R_g = R_h * R_e \tag{2}$$

where R_g represents the global gaze orientation matrix, R_h and R_e represent the headset and local gaze (eye) orientation matrix respectively.

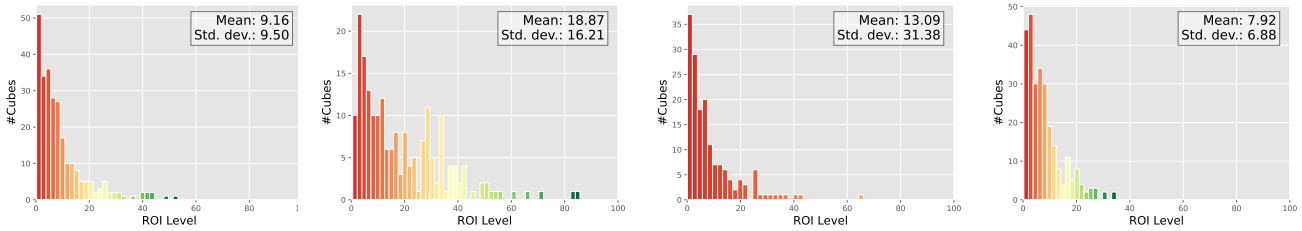
The combination of gaze direction and the headsets' movement trajectory is visualized in Fig. 3. The blue line indicates the user's motion trajectory and the red arrows attached to the blue line indicate the gaze direction at the corresponding position. Not surprisingly, we can find that **users' gaze often follows the activity of the object inside the video**. Specifically, they can be divided into two categories. On the one hand, **for volumetric scenes with relatively large movement, users' gaze tends to precede users' movement by a short period of time**. This phenomenon can be observed from Fig. 3 (a) and Fig. 3 (b), where the trolley and dustpan follow a regular movement. Then users' gaze can be focused on these objects and appear to have a similar movement feature. On the other hand, **for volumetric scenes with small movements, the gaze may move back and forth with irregular movement, but it generally still focuses on the target object**. Fig. 3 (c) and Fig. 3 (d) verify this observation that the endpoints of the gaze arrows mostly locate at the target objects.

Fig. 4 reaffirms this observation. The left figure shows the original aerial view heatmap of the movement trajectory, and the light part of the right figure indicates the point where the movement trajectory coincides with the gaze ray. It demonstrates a strong



(a) Pulling Trolley (b) Sweeping (c) Cleaning Whiteboard (d) Chatting

Figure 5: The Volumetric ROI level together with 4 representative scenes. Here the light green color indicates a higher ROI level and the dark red color indicates a lower ROI level.



(a) Pulling Trolley (b) Sweeping (c) Cleaning Whiteboard (d) Chatting

Figure 6: The distribution of the volumetric ROI level of four scenes. X-axis indicates the ROI levels, and Y-axis indicates the number of cubes with the corresponding ROI.

correlation that a large portion of the movement trajectory and the gaze ray indeed have interaction.

4 ANALYSIS ON USER BEHAVIOR

In this section, we conduct a comprehensive analysis of user behaviors based on the dataset, aiming to reveal the implicit correlations between various observed features, and further provide insight for future user behavior prediction. We mainly focus on user attention and movement features.

4.1 Volumetric ROI Calculation

Users' region of interest (ROI) is the most important feature when viewing volumetric videos. However, different from 2D or 360-degree videos where ROI can be directly obtained, ROI calculation in volumetric video is not so intuitive given its 3D nature. On one hand, there can be multiple objects alongside a user's eyesight and it is hard to uniquely determine the interested object. On the other hand, users are moving most of the time and the viewing angles are constantly changing. Thus, we define the *volumetric ROI level*, as a quantitative indicator, to represent how much attention a user pays to a region.

Calculating the volumetric ROI level includes the following steps:

▷ **Scene segmentation.** We first divide the whole volumetric scene into small blocks, where each block is a cube after slicing the space from x, y, and z dimensions. Since most of the cubes do not contain any points or only contain very few points, we set a threshold to filter out those near-empty cubes and only preserve those representing practical objects. Note that users' sensitivity to the point cloud density decreases with the increase of observing distance [8], we also vary such threshold accordingly.

▷ **Gaze frustum calculation.** By exploiting the pre-processed headset trajectory and gaze data, we are able to calculate the viewing directions of the user at every position. Normally, people's effective viewing angle is about 30° [27, 29, 32], we therefore define a virtual viewing frustum with an angle of 30°. And objects within this frustum will be viewed by the user.

▷ **Intersection calculation.** The ROI level of one cube can be calculated as how frequently this cube is covered by the gaze frustum of the user. In practice, we calculate the direction vector formed by the coordinates of the headset and the center of the cubes and then compare the angle between the direction vector and the gaze direction vector obtained from previous processing. The cube is counted once every time the angle is less or equal to 30°. By going through all of the effective cubes, we obtain the total counts for the whole volumetric video.

▷ **Volumetric ROI level calculation.** Inspired by the ROI mechanism used in 360 videos [5], we propose to calculate the volumetric ROI level F_a of a cube according to the density weight, the appearance frequency, and the distance between the user's eyes and the cube. The calculation formula is given as:

$$F_a = \frac{\rho_c * f_g}{D_c} \tag{3}$$

$$f_g = \frac{\sum_{i=1}^{N} N_g(i)}{N_{sample}} \tag{4}$$

where ρ_c is the point cloud density of the cube, f_g is the frequency of each cube falling into the viewing frustum, D_c is the distance between the headset and the cube center, $N_g(i)$ is the total counts of the current cube, and N_{sample} is the total number of user behavior samples.

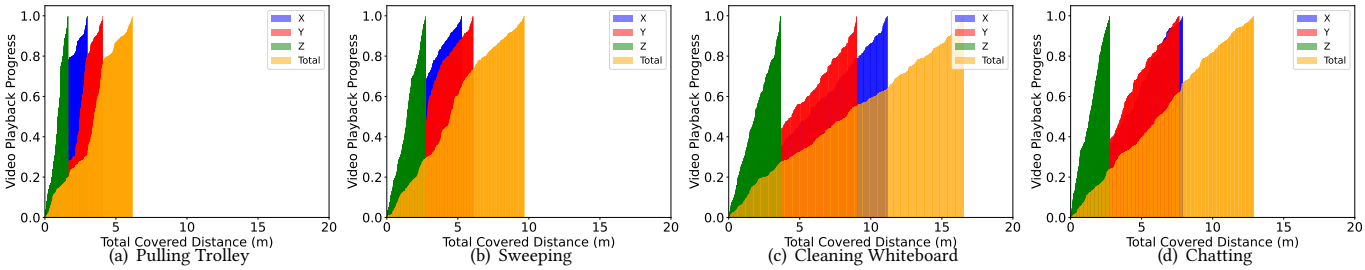


Figure 7: Movement Distance of X, Y, Z axes

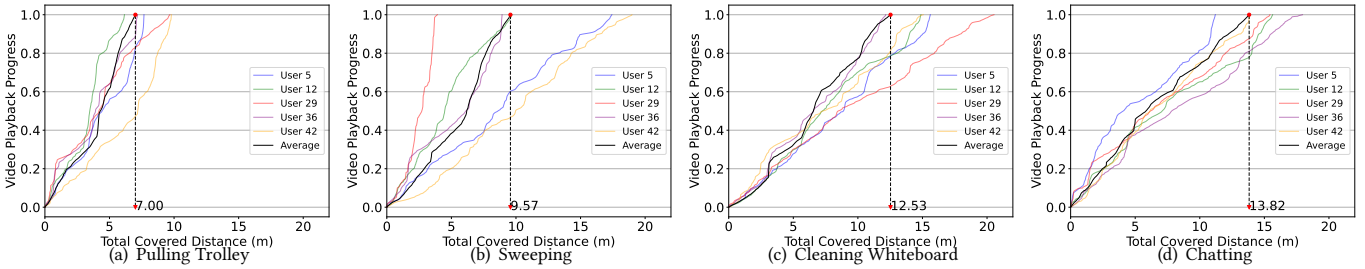


Figure 8: The total movement distance of different scenes. Here the colored line and the black line indicate each user's total movement distance and the average respectively along the video playback progress, the dotted line with a number represents the average value.

4.2 Analysis on User Attention

We next analyze users' attention (which can be directly reflected by the ROI level) when they are viewing different volumetric videos. Fig. 5 visually shows the different ROI levels for different volumetric video scenes. Here we randomly select 5 viewers and illustrate their average ROI level. We can find that users' attention is highly correlated with the volumetric content, and is **particularly on the actors and the objects they are manipulating**. For example, in the 'Chatting' scene, most attention is focused on the right person. In the 'Sweeping' scene, the dustpan instead attracts even more attention than the person.

Another interesting finding lies in the personalized preference, i.e., **users may pay higher attention to their preferred object or person**. Like in Fig. 5(d), the right person obviously has a higher ROI level than the left person, which is largely due to the user's personalized preference.

We next consider the distribution of user attention in different volumetric scenes. Fig. 6 plots the distribution of cubes with different volumetric ROI levels together with the mean value (Mean) and the standard deviation (Std. dev.). Note that we already remove those rarely-watched cubes. Comparing the different volumetric scenes, we can obtain several interesting findings: 1) **The ROI dispersion level of different volumetric videos is quite diverse, depending on the scene content**. For example, the ROIs of the 'Sweeping' scene concentrate with the range from 0 to 60, while the ROIs of the 'Chatting' scene mainly spread between 0 to 15. This means that users are more focused when watching the former more 'dynamic' video while they are more distracted when watching the latter more 'static' one. And it further reaffirms that people's attention is more easily captured by moving objects. 2) **Only a small portion of cubes have relatively high ROI levels.** This is because a volumetric scene can have a lot of effective cubes, while only a small portion of them, especially those representing the target actors or objects, will gain enough attention.

4.3 Analysis on User Movement

We conduct a more in-depth analysis of user movement to examine the correlations between movement behavior and video content. We first define the movement mode. Taking the user's headset as the origin, moving along the lateral direction of the body is indicated as the x-axis, along the vertical direction of the body is indicated as the y-axis, and the z-axis represents the up-down movement.

Fig. 7 shows the average moving distance along video playback progress in the three directions as well as the total distance of the 4 volumetric scenes. Naturally, moving laterally means that the user prefers to observe from different angles while moving vertically means that users would like to follow the moving objects. From this figure, we can find that the vertical distance is clearly larger than the lateral distance in the 'Pulling Trolley' and 'Sweeping' scenes, and vice versa for the rest two scenes. This observation matches exactly with our previous finding that **people tend to follow the moving object while observing the static object from various angles**.

We also investigate the movement features from the perspective of different users. Fig. 8 shows the cumulated moving distances of five randomly selected users. We find that the first two dynamic-scene videos have an average moving distance of 7.0m and 9.57m,

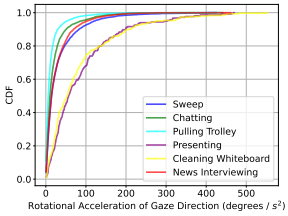


Figure 9: Rotational Acceleration in Different Scenes

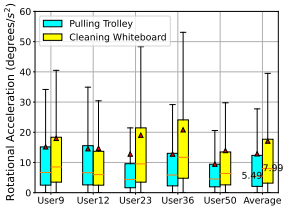


Figure 10: Rotational Acceleration of 'Pulling Trolley' and 'Cleaning Whiteboard'

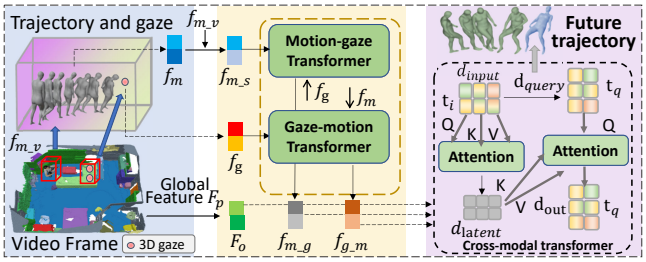


Figure 11: Transformer-based viewport prediction model.

respectively, while the rest two static-scene videos reach an average moving distance of 12.53m and 13.82m. Thus we can verify that **users tend to perform more spatial movements in static scenes compared to dynamic scenes to explore more areas in volumetric scenes.**

In Fig. 9, we show the average data to observe the difference in rotational acceleration for different scenes. From the figure, we find that in more static scenes, users change their orientation more frequently at a faster rate. According to the CDF plot, for even more than 30% sampling points the head moving speeds exceed 100 degree/s^2 in the 'cleaning whiteboard' scene. In contrast, for the relatively dynamic scene 'Pulling Trolley', there are about 80% sampling points with moving speed less than 20 degree/s^2. Diversity across different users is shown in Fig. 10, which depicts the specific rotational acceleration speed of 5 randomly selected users and their average for the scenes of 'Pulling Trolley' and 'Cleaning Whiteboard'. Observation from these figures reaffirms the finding that user movement in static scenes is usually faster than in dynamic scenes.

5 GAZE-ASSISTED VIEWPORT PREDICTION FOR VOLUMETRIC VIDEO STREAMING

In this section, we give a case of the dataset application in volumetric video streaming. By fusing the correlated features between video content and gaze information, we are able to improve the accuracy of viewport prediction, further benefiting VV streaming.

5.1 Background and Motivation

Viewport adaptive video streaming together with tile-based partition strategy [8, 15] has been widely explored in traditional 2D and recently 360-degree videos. By reducing the bitrate of the video content outside users' viewport, the whole transmitted video size can be saved and thus relieving the network bandwidth pressure. This idea is intuitive to move to the 3D scenario if the scene is partitioned into small cubes for cube-based streaming. However, though it applies well in 2D videos, a critical challenge arises when it comes to 3D volumetric videos. The major difficulty lies in the flexible six DoF spatial feature, where the significant uncertainty in spatial position and viewing angle makes the viewport prediction error easy to accumulate.

Several pioneer works have made attempts for six DoF viewport prediction [6, 15, 25]. For example. `ViVo` [8] and `Vues` [19] employ linear regression (LR) and multilayer perceptron (MLP) to predict the viewport, and have also explored the use of advanced deep learning models such as LSTM for prediction. Extending from `Parima` [1], `VolParima` [18] utilizes 3D object detection and tracking techniques to achieve improved accuracy in viewport prediction. However, these works either consider each DoF separately or mainly focus on the video content, which cannot fully capture the implicit features in volumetric videos to yield accurate viewport prediction towards various volumetric scenes. Motivated by our previous observations and findings, **we realize that the features in user movement, gaze direction, and video content are tightly correlated so that the multi-modal information, as well as their mutual impacts, should be combined together for consideration.**

5.2 Design

We extract the multimodal features and present an architecture with a bidirectional fusion model that facilitates the communication of different features in Fig. 11. This is a paradigm for accurate viewport predictions based on video content, interaction, and intention. Followed by a variety of cross-modal transformers to transcend information from multi-modality.

Cross-modal transformer. The cross-modal transformer [10] is used to capture the interplay of several elements and to establish communications among the multi-modal information.

Instead of extracting the multi-modal features independently[14], we propose a pipeline to overall integrate the history viewport feature, 3D gaze feature, and video features, which enhances the in-between feature communication to mutually decrease their future uncertainties on interaction and intention.

Video feature extraction. To learn the constraints (e.g. Surface and topology of furniture) from the 3D video and retrain the network for attention on locally interacted structures, we apply PointNet ++ [24], to extract both global (the video content) and local video features (interacted region). We derive the per-point feature and global descriptor of video as F_P, F_o.

Gaze feature extraction. The gaze point feature f_g is retrieved from the per-point video feature map F_P into $F_{P|g}$. Consequently, the interacted gaze feature with corresponding video information provides indications to infer the intention.

Viewport feature extraction. We use a linear layer to extract the viewport feature embedding f_m from multidimensional viewport trajectories input. The viewport is well-aligned with the video content. To endow the feature awareness of the 3D video content, we further query the video features with the viewport features.

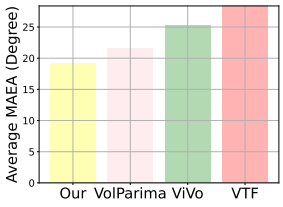


Figure 12: Average MAEA for viewport prediction

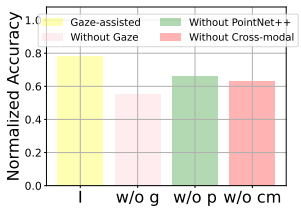


Figure 13: Average Accuracy for each prediction part

These interacted video features are then supplied to PointNet++ to get the contextual video feature f_{m_v} of the current viewport.

In lieu of directly concatenating the features, which would bring modalities features redundancy and impair the prediction accuracy [16], we propose a model by deploying a cross-modal transformer [22] to fuse the gaze, viewport, and video features.

Feature fusion. As an intermediary element, the viewport features strive to be cognizant of the 3D video features and the subject's intention inferred from the gaze features. First, we utilize the video feature f_{m_v} acquired from the 3D environment as the query to update the viewport feature f_m in the viewport-video transformer. Then, the output viewport embedding f_{m_s} is expected to be aware of the 3D video, which results in the final viewport embedding f_{m_g}. Inspired by [35], we handle the gaze embedding in a bidirectional manner, i.e., the viewport embedding f_m is also utilized as the query to update the gaze features into f_{g_m}. The bidirectionally fused multi-modal features are then assembled into holistic temporal input representations to perform human viewport prediction. As shown in Fig. 11, the updated gaze feature f_{g_m}, viewport feature f_{m_g} and the global video feature F_O are used to predict the future viewport trajectories from t to T by:

$$V_{T:T+t} = \mathfrak{R}\left(h_{\text{pos}}, \text{concat}\left(f_{g\_m}, f_{m\_g}, F_O\right)_{T-n:T-1}\right) \quad (5)$$

where concat denotes operator of concatenation, and h_{pos} is the latent vector containing temporal positional encodings for the output[24]. We evaluate our gaze-assisted viewport prediction against representative VV system and methods `ViVo`, `VolParima` and transformer-based `Vanilla-TF (VTF)` [31] using the Average Mean Absolute Error Angle (MAEA) as a metric. We also do an ablation study to compare the effect of each part.

As depicted in Figure 12, our proposed model is capable of reducing MAED by 13.3%, 19.8%, and 34.5% in comparison with `VolParima`, `ViVo`, and `VTF`, respectively. Furthermore, we conducted experiments to evaluate the accuracy of our gaze-assisted model and performed an ablation study comprising three variations: without gaze (w/o g), without PointNet ++ (w/o p), and without a cross-modal transformer (w/o cm). The results indicate that each component has a positive contribution to the overall performance. Our model, which effectively integrates and utilizes video content and gaze information, is demonstrated to produce more accurate predictions than the previous methods.

6 OTHER APPLICATIONS

In addition to our proposed viewport prediction systems, we provide several potential application cases that could be derived from our dataset.

User Identification for VV. User identification is a crucial task in 360-degree video, yet it poses a new challenge for volumetric video. Such a technique has the potential to improve user experience or enhance privacy.

For headset-movement-based identification, Li et al. [17] achieved an identification accuracy of 95.57% while participants nodded when listening to music. Gaze data could also be used for identification, Sluganovic et al. [28] proposed gaze-based authentication using a gaze-tracking device, which achieves an error rate of 6.3% at an authentication time of five seconds.

Given that our dataset on VV user behavior encompasses a wider range of attributes, an identification method utilizing both headset and gaze data could be developed to enhance accuracy.

Personalized Content Delivery. Many works have been conducted for content recommendation traditional in 2D video [3, 4] and 360-degree video [30], but for volumetric video, such field is still undefined. By analyzing the behavior of the users, developers can gain insights into users' preferences and adapt personalized content to better suit their needs.

7 CONCLUSION

In this paper, we focused on understanding user behavior patterns when watching volumetric videos. We released the first large-scale volumetric video user behavior dataset, including movement information, headset direction, user gesture, and user gaze information. This dataset involved data from 50 users with strong diversity and covered multiple representative volumetric scenes. We then conducted a comprehensive analysis aiming to reveal the behavior features. We defined the volumetric ROI level calculation mechanism in this context and focused on the feature analysis on user attention and user movement. Some interesting findings were therefore derived. Further, based on our analysis and observation, we designed a transformer-based volumetric video viewport prediction model, which fused all the correlated features and outperformed the state-of-the-art baseline solutions.

ACKNOWLEDGMENTS

The work was supported in part by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, by NSFC (Grant No. 62293482 and No. 62102342), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515012668), the Shenzhen Science and Technology Program (Grant No. RCBS20221008093120047), the Shenzhen Outstanding Talents Training Fund 202002, the Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104, the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055).

REFERENCES

[1] Lovish Chopra, Sarthak Chakraborty, Abhijit Mondal, and Sandip Chakraborty. 2021. Parima: Viewport adaptive 360-degree video streaming. In *Proceedings of the Web Conference 2021*. 2379–2391.

[2] Alexander Clemm, Maria Torres Vega, Hemanth Kumar Ravuri, Tim Wauters, and Filip De Turck. 2020. Toward Truly Immersive Holographic-Type Communication: Challenges and Solutions. *IEEE Communications Magazine* 58, 1 (2020), 93–99. https://doi.org/10.1109/MCOM.001.1900272

[3] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. 2010. The YouTube Video Recommendation System. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (Barcelona, Spain) *(RecSys '10)*. Association for Computing Machinery, New York, NY, USA, 293–296. https://doi.org/10.1145/1864708.1864770

[4] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, Pietro Piazzolla, and Massimo Quadrana. 2016. Content-Based Video Recommendation System Based on Stylistic Visual Features. *Journal on Data Semantics* 5 (2016), 99–113.

[5] Yu Guan, Chengyuan Zheng, Xinggong Zhang, Zongming Guo, and Junchen Jiang. 2019. Pano: Optimizing 360° Video Streaming with a Better Understanding of Quality Perception. In *Proceedings of the ACM Special Interest Group on Data Communication* (Beijing, China) *(SIGCOMM '19)*. Association for Computing Machinery, New York, NY, USA, 394–407. https://doi.org/10.1145/3341302.3342063

[6] Serhan Gül, Dimitri Podborski, Thomas Buchholz, Thomas Schierl, and Cornelius Hellge. 2020. Low-latency cloud-based volumetric video streaming using head motion prediction. In *Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*. ACM. https://doi.org/10.1145/3386290.3396933

[7] Matthew Hackett, Basiel Makled, Elliot Mizroch, Simon Venshtain, and Matthias Mccoy-Thompson. 2022. Volumetric Video and Mixed Reality for Healthcare Training. In *Interservice/Industry Training, Simulation, and Education Conference. IMX*, Vol. 22. 22–24.

[8] Bo Han, Yu Liu, and Feng Qian. 2020. ViVo: Visibility-Aware Mobile Volumetric Video Streaming. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (London, United Kingdom) *(MobiCom '20)*. Association for Computing Machinery, New York, NY, USA, Article 11, 13 pages. https://doi.org/10.1145/3372224.3380888

[9] Kaiyuan Hu, Yili Jin, Haowen Yang, Junhua Liu, and Fangxin Wang. 2023. FSVVD: A Dataset of Full Scene Volumetric Video. In *Proceedings of the 14th Conference on ACM Multimedia Systems* (Vancouver, BC, Canada) *(MMSys '23)*. Association for Computing Machinery, New York, NY, USA, 410–415. https://doi.org/10.1145/3587819.3592551

[10] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2021. Perceiver IO: A General Architecture for Structured Inputs & Outputs. *arXiv: Learning* (2021).

[11] Jack Jansen, Shishir Subramanyam, Romain Bouqueau, Gianluca Cernigliaro, Marc Martos Cabré, Fernando Pérez, and Pablo Cesar. 2020. A Pipeline for Multiparty Volumetric Video Conferencing: Transmission of Point Clouds over Low Latency DASH. In *Proceedings of the 11th ACM Multimedia Systems Conference* (Istanbul, Turkey) *(MMSys '20)*. Association for Computing Machinery, New York, NY, USA, 341–344. https://doi.org/10.1145/3339825.3393578

[12] Yili Jin, Junhua Liu, Fangxin Wang, and Shuguang Cui. 2022. Where Are You Looking?: A Large-Scale Dataset of Head and Gaze Behavior for 360-Degree Videos and a Pilot Study. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*. ACM, 1025–1034.

[13] Yili Jin, Junhua Liu, Fangxin Wang, and Shuguang Cui. 2023. Ebublio: Edge Assisted Multi-User 360-Degree Video Streaming. *IEEE Internet of Things Journal* (2023).

[14] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. 2020. Learning to Generate Diverse Dance Motions with Transformer. *arXiv: Computer Vision and Pattern Recognition* (2020).

[15] Jie Li, Cong Zhang, Zhi Liu, Richang Hong, and Han Hu. 2022. Optimal Volumetric Video Streaming with Hybrid Saliency based Tiling. *IEEE Transactions on Multimedia* (2022), 1–1. https://doi.org/10.1109/TMM.2022.3153208

[16] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. 2021. Learn to Dance with AIST++: Music Conditioned 3D Dance Generation. *CoRR* abs/2101.08779 (2021). arXiv:2101.08779 https://arxiv.org/abs/2101.08779

[17] Sugang Li, Ashwin Ashok, Yanyong Zhang, Chenren Xu, Janne Lindqvist, and Marco Gruteser. 2016. Whose move is it anyway? Authenticating smart wearable devices using unique head movement patterns. In *2016 IEEE International Conference on Pervasive Computing and Communications, PerCom 2016, Sydney, Australia, March 14-19, 2016*. IEEE Computer Society, 1–9. https://doi.org/10.1109/PERCOM.2016.7456514

[18] Junhua Liu, Boxiang Zhu, Fangxin Wang, Yili Jin, Wenyi Zhang, Zihan Xu, and Shuguang Cui. 2023. CaV3: Cache-assisted Viewport Adaptive Volumetric Video Streaming. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 173–183.

[19] Yu Liu, Bo Han, Feng Qian, Arvind Narayanan, and Zhi-Li Zhang. 2022. Vues: Practical Mobile Volumetric Video Streaming through Multiview Transcoding. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking* (Sydney, NSW, Australia) *(MobiCom '22)*. Association for Computing Machinery, New York, NY, USA, 514–527. https://doi.org/10.1145/3495243.3517027

[20] Zhi Liu, Qiyue Li, Xianfu Chen, Celimuge Wu, Susumu Ishihara, Jie Li, and Yusheng Ji. 2021. Point Cloud Video Streaming: Challenges and Solutions. *IEEE Network* 35, 5 (2021), 202–209. https://doi.org/10.1109/MNET.101.2000364

[21] Zhi Liu, Qiyue Li, Xianfu Chen, Celimuge Wu, Susumu Ishihara, Jie Li, and Yusheng Ji. 2021. Point Cloud Video Streaming: Challenges and Solutions. *IEEE Network* 35, 5 (sep 2021), 202–209. https://doi.org/10.1109/mnet.101.2000364

[22] Shujie Luo, Hang Dai, Ling Shao, and Yong Ding. 2020. C4AV: learning cross-modal representations from transformers. In *European Conference on Computer Vision*. Springer, 33–38.

[23] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Commun. ACM* 65, 1 (dec 2021), 99–106. https://doi.org/10.1145/3503250

[24] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5099–5108. https://proceedings.neurips.cc/paper/2017/hash/d8bf84be3800d12f74d8b05e9b89836f-Abstract.html

[25] Feng Qian, Bo Han, Jarrell Pair, and Vijay Gopalakrishnan. 2019. Toward Practical Volumetric Video Streaming on Commodity Smartphones. In *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications*. ACM. https://doi.org/10.1145/3301293.3302358

[26] Ignacio Reimat, Evangelos Alexiou, Jack Jansen, Irene Viola, Shishir Subramanyam, and Pablo Cesar. 2021. CWIPC-SXR: Point Cloud Dynamic Human Dataset for Social XR. In *Proceedings of the 12th ACM Multimedia Systems Conference* (Istanbul, Turkey) *(MMSys '21)*. Association for Computing Machinery, New York, NY, USA, 300–306. https://doi.org/10.1145/3458305.3478452

[27] P.J. Savino and H.V. Danesh-Meyer. 2012. *Color Atlas and Synopsis of Clinical Ophthalmology – Wills Eye Institute – Neuro-Ophthalmology*. Wolters Kluwer/Lippincott Williams & Wilkins Health. https://books.google.com.tw/books?id=6RgSZGWQZGIC

[28] Ivo Sluganovic, Marc Roeschlin, Kasper B. Rasmussen, and Ivan Martinovic. 2016. Using Reflexive Eye Movements for Fast Challenge-Response Authentication. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Vienna, Austria) *(CCS '16)*. Association for Computing Machinery, New York, NY, USA, 1056–1067. https://doi.org/10.1145/2976749.2978311

[29] W.B. Trattler, P.K. Kaiser, and N.J. Friedman. 2012. *Review of Ophthalmology E-Book: Expert Consult - Online and Print*. Elsevier Health Sciences. https://books.google.com.tw/books?id=AazA_9TQnHYC

[30] Ioannis Tsingalis, Ioannis Pipilis, and Ioannis Pitas. 2014. A statistical and clustering study on Youtube 2D and 3D video recommendation graph. In *2014 6th International Symposium on Communications, Control and Signal Processing (ISCCSP 2014)*. Institute of Electrical and Electronics Engineers (IEEE), United States, 294–297. https://doi.org/10.1109/ISCCSP.2014.6877872 6th International Symposium on Communcations, Control and Signal Processing (ISCCSP 2014) ; Conference date: 21-05-2014 Through 23-05-2014.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[32] C.P. Wilkinson, D.R. Hinton, S.V.R. Sadda, P. Wiedemann, S.J. Ryan, and A.P. Schachat. 2012. *Retina E-Book: 3 Volume Set*. Elsevier Health Sciences. https://books.google.com.tw/books?id=PdAsuzFRv5oC

[33] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Xiaojun Tong. 2018. Weighted Voxel: A Novel Voxel Representation for 3D Reconstruction. In *Proceedings of the 10th International Conference on Internet Multimedia Computing and Service* (Nanjing, China) *(ICIMCS '18)*. Association for Computing Machinery, New York, NY, USA, Article 33, 4 pages. https://doi.org/10.1145/3240876.3240888

[34] Emin Zerman, Cagri Ozcinar, Pan Gao, and Aljosa Smolic. 2020. Textured Mesh vs Coloured Point Cloud: A Subjective Study for Volumetric Video Compression. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6. https://doi.org/10.1109/QoMEX48832.2020.9123137

[35] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C. Karen Liu, and Leonidas J. Guibas. 2022. GIMO: Gaze-Informed Human Motion Prediction In Context. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII* (Tel Aviv, Israel). Springer-Verlag, Berlin, Heidelberg, 676–694. https://doi.org/10.1007/978-3-031-19778-9_39