Comparison of Visual Saliency for Dynamic Point Clouds: Task-free vs. Task-dependent

Xuemei Zhou, Irene Viola, Silvia Rossi, Pablo Cesar

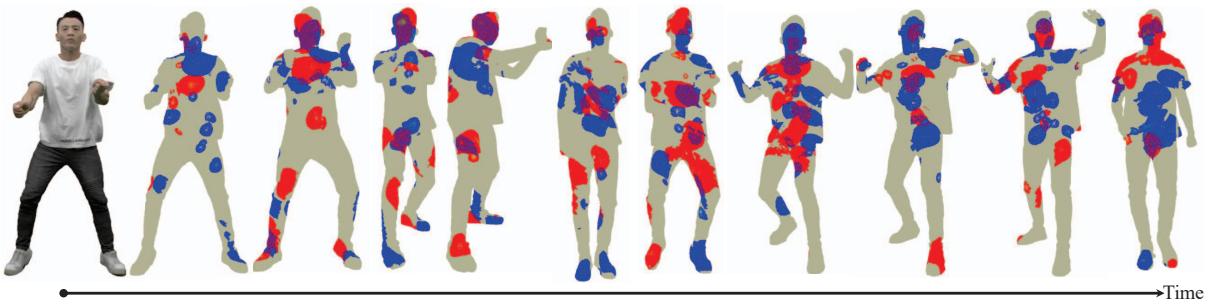


Fig. 1: Fixation maps of *dancer* sequences with uniform temporal sampling every 30 frames. The blue regions represent task-free conditions, while the red regions indicate task-dependent conditions. Gray areas denote nonsalient regions in both conditions, and overlapping areas are shown as a blend of the two colormaps.

Abstract— This paper presents a Task-Free eye-tracking dataset for Dynamic Point Clouds (TF-DPC) aimed at investigating visual attention. The dataset is composed of eye gaze and head movements collected from 24 participants observing 19 scanned dynamic point clouds in a Virtual Reality (VR) environment with 6 degrees of freedom. We compare the visual saliency maps generated from this dataset with those from a prior task-dependent experiment (focused on quality assessment) to explore how high-level tasks influence human visual attention. To measure the similarity between these visual saliency maps, we apply the well-known Pearson correlation coefficient and an adapted version of the Earth Mover's Distance metric, which takes into account both spatial information and the degrees of saliency. Our experimental results provide both qualitative and quantitative insights, revealing significant differences in visual attention due to task influence. This work enhances our understanding of the visual attention for dynamic point cloud (specifically human figures) in VR from gaze and human movement trajectories, and highlights the impact of task-dependent factors, offering valuable guidance for advancing visual saliency models and improving VR perception.

Index Terms—dynamic point cloud, eye-tracking, task-free, visual saliency metric, similarity measurement

1 Introduction

The Human Vision System (HVS) processes vast amounts of visual information by selectively focusing on relevant parts of the surrounding environment. This mechanism, known as *visual saliency* or *visual attention*, allows for efficient interpretation of complex scenes. Visual saliency has become a key focus in image and video processing due to its ability to efficiently identify regions of interest, improving both processing and transmission [20, 33], with extensive studies already conducted in this area [10, 11, 28, 43]. In particular, researchers have investigated how the oculomotor behavior and attention is affected by high-level visual tasks [32], such as Image Quality Assessment (IQA) or Video Quality Assessment (VQA), compared with free viewing, where users observe the media content as they normally would, which results in so-called natural scene saliency. For example, Liu [36] and Le Meur [32] have collected eye-tracking data under both free viewing and quality assessment scenarios. Their findings suggest that the main region of interest for image/video remains highly similar, with certain deviations observed during quality assessment tasks.

- *Xuemei Zhou and Pablo Cesar are with Centrum Wiskunde en Informatica, Amsterdam, The Netherlands, and with TU Delft, Delft, The Netherlands. E-mail: name.surname@cwi.nl.*
- *Irene Viola, and Silvia Rossi are with Centrum Wiskunde en Informatica, Amsterdam, The Netherlands. E-mail: name.surname@cwi.nl*

Received 18 September 2024; revised 13 January 2025; accepted 13 January 2025. Date of publication 11 March 2025; date of current version 31 March 2025. Digital Object Identifier no. 10.1109/TVCG.2025.3549863

Recent advancements in immersive media have shifted the focus on 3D content. Specifically, volumetric video, such as dynamic point cloud, has become one of the most popular formats [6]. Unlike 2D images and videos, where visual saliency has been extensively studied, dynamic point clouds present unique challenges that have not been fully addressed in the literature. For example, dynamic point clouds differ from traditional video in terms of data volume, and the use of Head-Mounted Displays (HMDs) for their consumption introduces additional complexities. Thus, established findings for visual saliency in image and video, such as the spatial bias [45] and central bias [62] in fixation data, may not hold for dynamic point clouds.

One of the main challenges hindering the advancement of saliency-guided applications for dynamic point clouds is the lack of ground-truth saliency data. To address this gap, several studies have attempted to collect eye-tracking data to generate ground-truth saliency maps for point clouds. For instance, Alexiou *et al.* [7] conducted an eye-tracking experiment in VR under task-dependent scenario. Nguyen *et al.* [40] released an open source, task-free eye-tracking dataset for 4 dynamic point clouds in mixed reality using Hololens 2. Zhou *et al.* [69] presented a task-dependent eye-tracking dataset for 50 dynamic point clouds. A summary of existing visual attention datasets for point clouds is shown in Table 1. These datasets, in which gaze patterns are recorded under free viewing or different task-dependent conditions, have been instrumental in creating ground-truth visual saliency maps used for model design and validation. Despite these efforts, the impact of task-free and task-dependent conditions on human visual attention deployment in point clouds is still unexplored, unlike in its 2D counterpart. A dataset that captures saliency maps for the same content across

Table 1: Publicly available visual attention datasets for point clouds.

Dataset	Type	Stimuli	Display	Interaction *	Visual Attention	Task-free
ViAtPCVR [7]	Static	8	VR	✓	✓	✗
QAVA-DPC [69]	Dynamic	50	VR	✓	✓	✗
ComPEQ-MR [40]	Dynamic	4	AR	✓	✓	✓
TF-DPC (Ours)	Dynamic	19	VR	✓	✓	✓

* Interaction here refers to being able to move around and observe the point cloud from different angles.

different perceptual tasks in VR/AR is still needed to assess the impact of tasks.

In this study, we aim to address these challenges by creating a novel Task-Free dataset for Dynamic Point Clouds (TF-DPC), which will benefit both the research community and provide extensive training data. The dataset is composed of eye gaze and head movements collected from 24 participants observing 19 scanned dynamic point cloud in a Virtual Reality (VR) environment with 6 Degrees of Freedom (DoF). Based on the collected data, we investigate how human visual attention is affected by high-level visual tasks, by comparing our task-free saliency maps with those obtained in a subjective quality assessment scenario presented in [69]. To better quantify the difference between saliency maps in task-free and task-dependent scenarios, we use Pearson's Correlation Coefficient (PCC) and a modified version of the Earth Mover's Distance (EMD) metric for image retrieval [55]. Our experimental results provide both qualitative and quantitative insights, revealing significant differences in visual attention due to task influence. For example, Figure 1 shows the fixation maps for the *dancer* sequence in both task-free and task-dependent conditions (represented by blue and red areas, respectively). Users tend to focus on different regions of the content based on the experiment condition. Specifically, in the task-dependent scenario, participants show a more consistent focus on facial expressions or fine details, reflecting the specific task of evaluating the quality of the content. To conclude, our contributions are threefold and can be summarised as follows:

- We create a visual attention dataset for 19 original dynamic point clouds in a task-free VR experiment with 6-DoF. We release all raw data, containing the gaze samples and movement trajectory collected in our study, along with the code to compute and compare the dynamic point cloud visual saliency maps. https://github.com/cwi-dis/TVCG2025-TaskFree_PointCloudEyeTracking
- We provide an in-depth analysis of the collected dataset, using quantitative measures to explore the dataset in terms of gaze and trajectory; furthermore, we use qualitative methods to draw further insights from interviews.
- We compare the visual saliency maps under task-free and task-dependent conditions, to explore the impact of the high-level quality assessment task on human visual attention.

This novel dataset offers valuable opportunities for developing reliable saliency models for 3D representations, which are essential for augmented and mixed reality applications [23, 24]. For instance, they can enable advancements in several areas, including saliency-guided compression [44, 66] and live reconstruction [51] for point cloud streaming, saliency-aware point cloud mixup for data augmentation [67], volume visualization [27], foveated rendering [54], point cloud transmission [51] and visual quality assessment [12, 61, 68].

2 Related Work

2.1 Visual Attention for Point Clouds

In the early stages of visual attention computation, due to the limitations of eye-tracking technologies, different collection procedures for salient points were pursued. For example, Chen *et al.* [14] investigate "Schelling points" on 3D meshes, feature points selected by people in a pure coordination game due to their salience. They designed an online experiment that asked people to select points via mouse-tracking technology on 3D surfaces that they expected would be selected by other people. This dataset is widely used as a benchmark for objective saliency detection algorithms for colorless point cloud/mesh [16, 59]. Later methods employ handcrafted descriptors [16, 35] from more low-level geometric properties to detect the point cloud/mesh saliency, but these approaches lack expressiveness and overlook real human viewing behaviors [39]. More recently, to explore the visual attention of 3D point clouds, eye-tracking experiments remain the main way to understand human visual behaviors. Abid *et al.* [2] compute the visual saliency of the point cloud considering the viewpoint from which the 3D content was seen/rendered, using an offline-computed view-based saliency map. One eye-tracking experiment on 2D screen is conducted to verify the proposed saliency map. Alexiou *et al.* [7] conduct an eye-tracking experiment in an immersive 3D scene. A method to exploit the high-quality recorded gaze measurements is introduced based on per-session profiling, and a scheme to determine areas of fixations in a static point cloud is proposed. Zhou *et al.* [69] collect a dataset containing the subjective opinion scores and visual saliency maps in a VR environment using eye-tracking technology, which first establishes a link between quality assessment and visual attention within the context of the dynamic point clouds. Nguyen *et al.* [40] propose a dataset with compressed dynamic point clouds, rating scores, and eye-tracking data with Augmented Reality (AR) HMD. However, only 4 reference dynamic point clouds have an associated visual saliency map. In our dataset, we collected a dynamic point cloud dataset in VR with free viewing. By using the same content as [69] and [40] and extending it with other dynamic sequences, our dataset provides the possibility to investigate the task impact or device impact for visual attention deployment in VR or between VR and AR, as well as using the collected data for other applications (i.e., saliency-guided compression).

2.2 Task Impact on Visual Attention

Understanding how the allocation of human visual attention changes depending on perceptual tasks offers clear benefits in developing techniques and improving the quality of experience in VR/AR. This is a complex behavior that holds great importance for the field of IQA/VQA. Specifically, task-free means that the user observes the content as naturally as possible, with fixation data from such free viewing commonly used to evaluate visual saliency. In contrast, task-dependent means that the user observes the media content to fulfill a specific task; in the case of IQA/VQA, to evaluate the visual quality. In these experiments, the mean opinion score (typically ranging from 1 to 5) across users serves as the ground truth for quality evaluation.

Meur *et al.* [32] carry out two eye-tracking experiments on 10 original video sequences in a free viewing and a quality assessment task, separately. The comparison between eye movements indicates that the degree of similarity between human priority maps is rather high. They observe that saliency-based distortion pooling does not significantly improve the performances of the VQA metric. Liu *et al.* [36] and Hani *et al.* [5] perform a similar experiment procedure for IQA, Liu evaluates whether and to what extent the addition of natural scene saliency is beneficial to objective quality prediction in general terms, and Hani conclude that it is not fair to compare the effect of adding saliency in objective metrics without specifying how the saliency was measured.

In larger contexts, task effects more broadly influence visual attention in immersive environments. Hadnett-Hunter *et al.* [21] investigated free-viewing, search, and navigation tasks in interactive virtual environments and found task-specific differences in several human visual attention measures, particularly during navigation. Their findings demonstrated the potential for using attention data to dynamically adapt virtual simulations and games. Hu *et al.* [25] analyzed eye and head movements of participants performing free-viewing, visual search, saliency, and tracking tasks in 360-degree VR videos. They revealed significant task-driven differences in fixation durations, saccade amplitudes, head rotation velocities, and eye-head coordination. EHTask–a learning-based method that employs eye and head movements to recognize user tasks in VR is proposed. Their work provides meaningful insights into human visual attention under different VR tasks and guides future work on recognizing user tasks in VR. Malpica *et al.* [38] systematically examined the impact of free exploration, memory, and visual search tasks on visual behavior in immersive scenes. They reported consistent task-specific differences in eye and head movement patterns, offering practical insights for designing task-oriented immersive applications. To the best of our knowledge, we are the first to

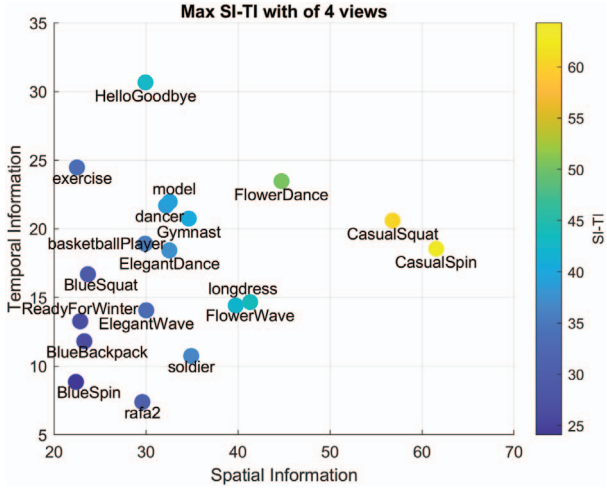


Fig. 2: Distribution of SI and TI of 19 source dynamic point clouds from 3 datasets, the color value is computed by $\sqrt{(SI^2+TI^2)}$.

investigate the impact of tasks on human attention deployment in the context of dynamic point clouds, building on insights from video, VR, and immersive media studies.

2.3 Evaluation Metrics of Saliency Maps

The deviation between two saliency maps is often quantified depending on how the visual saliency is represented [48]. Following the evaluation metrics on 2D image saliency maps, we can divide the evaluation metric into location-based metrics and distribution-based metrics. Location-based metrics, such as AUC [22], NSS [42], IG [30], are designed specifically for saliency evaluation, and they operate on the ground truth represented as discrete fixation locations. On the other hand, distribution-based metrics, such as SIM [47], PCC, KL [34], and EMD [50], have been adapted from information theory (IG, KL-divergence), statistics (PCC) and image matching and retrieval (SIM, EMD), and operate on the ground truth represented as a continuous fixation map. Interested readers can refer to [13] for more information about the recommendation for metric selections under specific assumptions and for specific applications. However, the aforementioned metrics are designed for grid-based 2D saliency maps, which makes them difficult to apply to point cloud saliency maps due to the intrinsic characteristics of dynamic point clouds. Based on the recommendations for metric selection, we chose PCC and adapted EMD as they are well-suited for evaluating distribution-based saliency maps and can be easily extended to 3D scenarios, aligning with the nature of our point cloud saliency maps. We further clarify our choice in Section 5.2.

3 Dataset Construction

To investigate how visual attention is deployed on dynamic point clouds and compare it with task-dependent saliency maps [69], we conducted a task-free eye-tracking experiment in a VR environment. During the experiment, we recorded the position (x, y, z coordinates) and rotation (three Euler angles around the x, y, and z axes) of the camera associated with each participant's HMD, along with timestamped data. This information was used to analyze participants' navigation movements within the physical space (i.e., the floor). Gaze-related data (gaze origin in x, y, z, and normalized gaze direction vector, the position of the point cloud frames) was collected following the same method as in [69] to generate saliency maps.

3.1 Materials

We select all 12 point cloud sequences from UVG-VPC dynamic point cloud dataset [19], 5 reference sequences from the QAVQ-DPC dataset [69], and 2 sequences from the Owlii dataset [64] for the task-free eye-tracking experiment. We selected all the reference contents from the QAVA-DPC dataset as it contains task-dependent visual attention maps, thus aiding us in our purpose of comparing task-dependent and task-free viewing, and we complemented it with additional high-quality contents to provide additional saliency data. We compute the Spatial Information (SI) and Temporal Information (TI) for each content [1], by projecting the point cloud into 4 views, namely, left, right, front, and back view, of its bounding box to apply SI and TI separately, then obtain the maximum value among the 4 views over all the frames as the final SI/TI for one sequence. The distribution of all dynamic point clouds can be seen in Figure 2. The dispersed state in SI (horizontal axes)/TI (vertical axes) shows the diversity of our contents in the spatial/temporal domain. All the stimuli are reference quality (without any compression distortion).

3.2 Apparatus

To ensure that the high-level task is the only variable, we used the same apparatus as [69], to enable a fair comparison with the other task-dependent experiment. Our experiment software is developed in Unity (version 2021.3.10.f1). The CWI point cloud unity package (version 0.10.0) is used to import and playback the dynamic point clouds [46]. For the UVG-VPC dataset, each sequence contains 250 frames, while other sequences contain 300 frames. The frame rate is 30 frames per second, with each video being displayed 3 times. We use HTC Vive Pro Eye devices with eye-tracking capabilities and Vive hand controllers for participant interaction. The eye-tracking applications are developed using the native HTC Vive SRanipal SDK.

We ensured a watertight appearance of all the stimuli by adjusting the point size to the average distance among its 10 nearest neighbors all over all points in the point cloud [57]. They are rescaled to a similar size, around 1.8m in height, to mimic realistic tele-immersive scenarios. The VR scene is illuminated by a virtual lamp on the ceiling centered above the models. The lamp is set as an area light with intensity values of 2 in Unity to simulate ordinary lighting in a room.

3.3 Procedure

In this study, we use a within-subject design. To avoid the effects of contextual or memory comparisons, we randomly generated a playlist for each subject. Before the experiment, the visual acuity and color vision of every subject was tested using Snellen [18] and Ishihara [15] charts. Participants were briefed and signed a consent form prior to taking part in the study. At the beginning of the session, the inter-pupillary distance was measured and the headset was adjusted by the subject accordingly. Then, a training session was conducted to help familiarize the subjects with the system, including the controllers and the naming of each button to interact more easily. Two training sequence, namely *loot* and *redandblack*, were used, which were not included in the dataset. The subjects always started at the same location, which is 1.5 meters away from the center of the virtual room, but could move freely from there onward and ended anywhere they preferred. A stimulus was located in the center of the virtual room, and each stimulus was randomly rotated between $[0°, 360°]$ to avoid bias. During the experiment, the subjects were instructed to view each model freely in the VR environment during the playback of each sequence. The subjects were also required to stand still while doing the calibration and error profiling.

For each subject, the test was split into two rounds, lasting for around 17 minutes each, with a mandatory 5-minute break in between. Before and after each round, participants were requested to fill in a Simulator Sickness Questionnaire (SSQ) on a 1-4 discrete scale (1=none to 4=severe) [26]. For every model and subject, a round was split into three consecutive steps:

1. *Calibration* was done at the beginning of the experiment, and only when calibration was successful users could enter into the dynamic point cloud playback stage.
2. *Inspection of models* is the step where the participants are observing the dynamic point cloud naturally, while their movement trajectory and gaze-related information are recorded.
3. *Error profiling* is issued as the last step in order to estimate the accuracy of the gaze measurements due to calibration inaccuracies, or HMD displacements.

After participants finished the two rounds, they were requested to fill out the Immersive Presence Questionnaire (IPQ)[1]. Last, the researchers conducted a semi-structured interview. The interview was conducted individually in a non-VR setting, and the entire conversation was recorded for analysis purposes.

3.4 Participants

A total of 24 participants took part in the subjective tests of this study, with a diverse composition that includes 1 non-binary individual, 12 males, and 11 females. The participants' ages ranged from 23 to 35, with an average age of 28.33 and a standard deviation of 3.10. Each participant observed all the dynamic point cloud stimuli. In terms of occupation, the majority (66.67%) of the participants were students, ranging from master to PhD levels. The remaining 33.34% were researchers (scientist and lecturer), one landscape designer, and one accountant. Regarding familiarity with VR devices, 5 participants had never experienced VR before the experiment, 13 participants had intermediate experience (using VR 1 to 3 times), and 6 of them were considered experts, having backgrounds as VR designers or researchers. Additionally, 17 out of 24 participants wore glasses during the experiment. No ethical approval was sought for this study, due to the absence of an established ethical review board at the institution where the research was conducted. The experimental protocol, including participant consent and data collection, was reviewed through an internal board to be compliant with current GDPR legislation. Participants consented to the collection and usage of their data at the start of the experiment, after being informed about the study.

4 Experiment Results

4.1 Analysis of movements on the physical space

The analysis of the movements on the physical space is based on the recorded data associated with the position and rotation of HMD collected during experiments. For the following analysis, the data was resampled at 30Hz. A general overview of the navigation behaviour of participants on the floor (plane XY) is given in Figure 3 for three selected contents, *rafa2*, *HelloGoodbye* and *CasualSpin*. We chose these volumetric point clouds based on their SI and TI values to investigate how the users movements change in relation with content characteristics. As shown in Figure 2, *rafa2* has low TI and SI, *CasualSpin* has high value of SI while *HelloGoodbye* is characterised by high TI. The volumetric content is initially placed approximately at the center of the floor plane and since the sequences are dynamic, we also represent their position over time with a trajectory of pink dots. It can be noted that the first sequence is the less dynamic since *rafa2* stays in its initial position (Figure 3 (a)). This brings to a more static behaviour also from the participants who mainly stay in one location without exploring the area around the content: there are indeed some strong red spots which represent the position where users spent most of their time and the shadow of the user position is quite compacted around the content. The point cloud *CasualSpin* is instead spinning around itself. In this case, participants are more spread around the content to display it from different perspective as shown in Figure 3 (b) but they are still quite compact. On the contrary, Figure 3 (c) shows a more dynamic exploratory behaviour from the users while displaying *HelloGoodbye*. To be noted that this sequence is also the most dynamic one since it walks back and forward. Thus, users tend to explore more while watching dynamic sequences, as already observed in [49].

4.2 Analysis of gaze data

To understand deeper visual exploration, we now analyze the relationship between gaze and contents. Following the same gaze data processing in [69], we ignored the initial 400 ms gaze data of each user to avoid unintentional gaze because of the unexpected appearance of the dynamic point cloud. Then, only the valid gaze samples provided by the native HTC Vive SRanipal SDK were selected. Each valid gaze sample was processed as follows: 1) Verify the data validity of gaze data by calculating the weighted average angular error to each gaze sample with the help of GazeMetrics [9]; 2) Identify fixation points of gaze data by dispersion threshold identification algorithm; 3) Map gaze data to dynamic point cloud frame with truncated-cone-sector algorithm [7]; 4) Fuse multiple users' gaze data to dynamic point cloud frames. After the four steps, we obtained the saliency map per frame. Each point cloud frame has a normalized heat value range in [0,1] for each point, 0 meaning non-salient and 1 meaning the most salient. For the processing details, please refer to [69]. Figure 4 represents the number of fixations of each subject on each content. Specifically, each row denotes the number of fixation points per content across the different users. Blue colors indicate a low value of fixations while yellow ones indicate high values. Vertically, we can notice consistent behavior per participant across the different content. For example, User 14 always has a low value of fixation, independent of the visualized volumetric content, indicating a more erratic behavior. On the contrary, User 1 appears to have more consistent fixations across the content. Thus, participants tend to preserve similar gaze behavior (highly erratic or quite static) independently of the volumetric content. Similar outcomes were observed also in [49]. Looking at Figure 4 per row (i.e., a single content across different users), we can notice that contents with higher TI got more attention: *FlowerDance* and *model*, which are characterized by higher TI, present more fixations than *rafa2*. To further our analysis, in Figure 5, we show the saliency map (randomly selected frame 150_{th}) for these three sequences. We can see that all three sequences show fixations on semantically relevant areas, such as the face. However, in *FlowerDance*, who is in the middle of a spinning motion, and *model*, who is simply adjusting her dress, the fixation areas are smaller and more dispersed across the content, as the users' attention is drawn by the motion of the dresses or any patterns on them. We further analyse and discuss gaze data in Section 5.1.

4.3 Analysis of SSQ and IPQ data

SSQ comprises 16 symptoms which are further grouped into three different categories: Oculomotor, Nausea, and Disorientation; we also computed the total score according to [26]. The simulator scores increased after the experiment. Specifically, the total scores rose from 6.37 to 10.33 before and after Session 1, and from 5.91 to 10.08 before and after Session 2. However, it can be seen that breaks help in reducing simulator sickness. The current version of the IPQ has three subscales (Spatial Presence, Involvement, Experienced Realism) and one additional general item not belonging to a subscale. We calculate the mean across the users for each factor. The participants experience high levels of Spatial Presence ($M_{SP} = 4.5$) and Involvement ($M_{INV} = 3.8$), whereas lower levels of Realisms ($M_{REAL} = 3.3$). The possible reason is that there is no interaction between the user and the content, as mentioned in Section 4.4.4, and there is no eye contact. The virtual room is empty for better capturing the visual attention, which normally get a lower score for the question: "the virtual world seemed more realistic than the real world."

4.4 Qualitative results

22 valid interview audio recordings were transcribed into texts and coded using Dovetail[2]. Following Maguire's guideline on thematic analysis [37], we initially reviewed and labeled the text, organized these labels into themes, and subsequently convened to establish the connection between content and visual attention during the subjective test. Each participant is denoted as P1-P24, with the number of participants concurring with each statement indicated in parentheses.

4.4.1 Factors that Capture Visual Attention Allocation

Temporal information Participants (18) pointed out that movement is the most attractive factor in our dynamic point cloud playback scene(P21: "when you are watching a video, it's easy to follow the direction of the movements"). 11 of them interpreted the information conveyed by the content as interesting to attract their attention. However, participants (16) also noted that high-motion sequences do not necessarily attract more attention than low-motion sequences.

[1] https://www.igroup.org/pq/ipq/index.php

[2] https://dovetail.com/

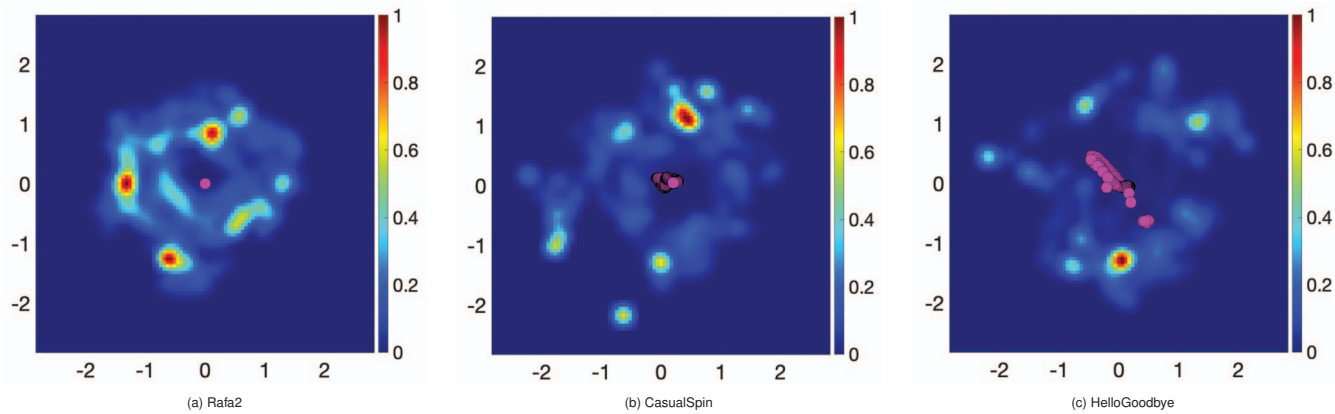


Fig. 3: Spatial distribution over time of the main location visited by users while displaying three different content: (a) *rafa2*, (b) *HelloGoodbye* and (c) *CasualSpin*. The centroid position of each volumetric content is represented by a sequence of pink points on the floor.

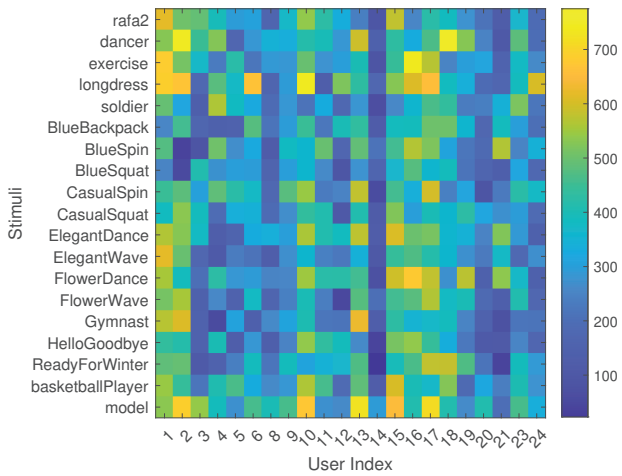


Fig. 4: The fixations per subject content in the proposed TF-DPC dataset. Each row denotes the fixations on a specific content and each column denotes the fixations for each subject, respectively.

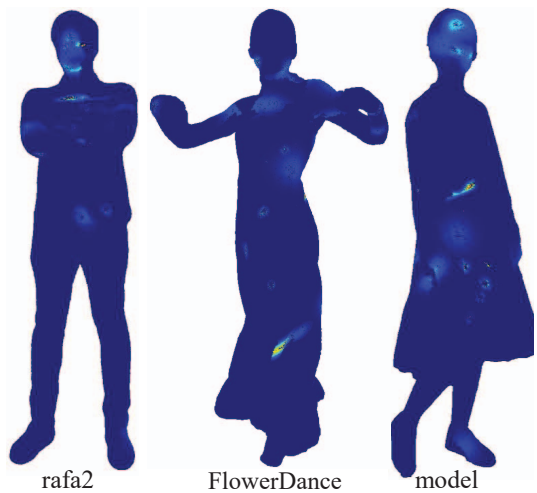


Fig. 5: The visual saliency map of the 150th frame of the dynamic point cloud with the front view.

Artifacts and Details Artifacts (9) and details (9) are identified as the co-second factors attracting people's attention. (P8: "what I focused on also negative things are, on the edges of the point calls often there was like rippling, sort of flickering, attracts a lot of attention, distracts me, other than that, I think eyes like faces in general, people like the expression.")

Geometry and Texture Geometry (2) and texture (7) are identified as the second and third factors influencing the subjective rating of point clouds under scrutiny. (P3: " I was observing precisely two things, the edges of the body and how distorted they are and also some distortions inside the costume.")

In terms of visual attention allocation, temporal information proves to be more crucial than either geometry or texture, with both geometry and texture showing relatively low importance. The details of the dynamic point cloud fall somewhere in between, while negative artifacts in the point cloud attract significant attention, aligning with findings from a previous study [60].

4.4.2 Factors Affecting Visual Attention

Participants (12) reported the realism of the content and naturalness of the action would change their attention. (P1: "I have to say there's an effect, if I see the quality is good, I usually will look closer. I will check the details. But if the quality is so poor that I can see distortion everywhere, then I will consciously, I will realize this is not real. So I will be less interested.") Abrupt distortions of the sequence will shift attention, (P5: "The point cloud's intended focal point might end up being overlooked because the flaws draw my attention away from it, instead I focus on the imperfections."). It is worth noting that all the point clouds under test were of reference quality; that is, any impairment was derived from the acquisition itself, and was not due to any additional processing such as compression. Thus, the acquisition methods themselves can have a significant impact on visual attention. This observation aligns with Zhang's conclusion [65] that distortions always change the attended regions.

4.4.3 Factors Influencing User Interaction

Participants (14) attributed most of their movement to the need to observe the front face to have more understanding of the human figure. They noted that sequences showing the same human figure with only slight variations in movement and clothing, as in the UVG-VPC dataset, led to decreased movement and reduced interest. This repetition (5) and the monotonous actions (5) made the task feel not engaging and dull. Limited space (8) and cable (1) result in less movement of the participants.

4.4.4 Designing the Construction of a Visual Attention Dataset

Content Participants favored the "longdress" (7), "soldier" (6), and "Gymnast" (5) point cloud sequences among all the contents, describing them as both realistic and engaging. However, some participants (3) noted that there are only human figures. Additionally, they

expressed a desire for more varied objects and increased interactivity, such as eye contact between themselves and the content, to enhance the immersive experience.

Display equipment for dynamic point cloud Participants (16) stated that using an HMD in VR is a better alternative to a 2D screen, as it provides greater immersion and freedom. (P7: "I think it's more intuitive if you feel more real when you see it, by 1 to 1 ratio is like your size, it's like next to you while on the screen it's like really small, you can zoom in but then the screen is not as big or you only see maybe one part of it even though it's a big screen, it's not 3D.") However, the HMD is heavy (2) and uncomfortable for prolonged use (2), while 5 participants noted that its effectiveness depends on the specific application.

5 Comparison between task-free and task-dependent

To explore how visual tasks impact the visual attention, we quantitatively analyze gaze statistics and saliency map similarity between task-free and task-dependent scenarios. To be noted these analyses are limited to the five shared sequences across the proposed dataset and the one presented in [69]: *rafa2* (low SI, TI), *dancer* (medium SI, high TI), *exercise* (low SI, high TI), *longdress* (high SI, medium TI), and *soldier* (medium SI, TI).

5.1 Comparison Consistency of Gaze

To analyze the allocation of visual attention depending on the task, we propose three measurements. We choose the total fixation number instead of other statistics of the gaze [4] (the mean duration or scan-path magnitude), because since the fixation is obtained through the dispersion threshold identification algorithm, the duration of consecutive gaze samples is implicitly considered. Apart from gaze behavior, our focus is on where the gaze is allocated within 3D point cloud frames. We select the Volumes of Interest (VoI) [56], which can show how many volumes have been observed by humans, and the distribution of the VoI, which can tell us how their attention is dispersed across the point cloud. VoI is computed as the total number of points whose heat value is larger than zero, the spread of VoI is the average pairwise distance of the VoI within the point cloud. Figure 6, from left to right, shows the fixation, VoI, and the spread of VoI across participants in both a task-free and task-depended experiment. We can observe the following: 1) Fixations for all 5 sequences with variant SI and TI perform consistently. The fixation number under task-free is lower than under task-dependent conditions since people need to focus relatively more to evaluate the quality of the sequences. 2) Generally, more fixations mean larger VoI and sparser distribution of the VoI. However, this is not true for *dancer* and *rafa2* sequences.

To analyze the difference between tasks with respect to these measures of visual attention, we ran a set of analysis of variance (ANOVA) tests. We grouped all fixations by task and aggregated measures by participant for each content per frame. One-way ANOVAs indicate the overall effect of the task on these measures. The p-value is below the threshold (0.05) of significance for all the contents per measure except for the spread of distributions for *rafa2* and the RoI for *dancer*, which are 0.1641 and 0.8008, separately. In conclusion:

- Across all 5 sequences, the number of fixations is significantly different between task-free and task-dependent scenarios. Task-dependent viewers, who were evaluating the quality of the content, consistently had more fixations compared to task-free viewers, who likely scanned the content more freely. This supports the idea that task-related goals require more focused attention, leading to a higher fixation count. Sequences with higher SI and TI, such as *longdress* and *dancer*, tend to capture more attention, evidenced by the higher number of fixations. In contrast, lower SI and TI sequences like *rafa2* generally had fewer fixations, as they may not have been as visually engaging.
- There is a significant difference for most contents, with task-dependent conditions leading to larger VoIs. This suggests that when participants are given specific tasks, they distribute their attention more widely across the point cloud (multiple specific areas), perhaps because the tasks prompt them to explore more regions for relevant information. While in free-viewing, they explored generally, driven by personal curiosity or passive observation rather than the active search for specific details. *dancer* stands out as the only content where both conditions cover the same. This could mean that the nature of the *dancer* does not lead to a noticeable change in the areas participants attend to, regardless of whether they are given a task or not.
- There is a significant difference for most contents, with task-dependent conditions leading to a broader spread of attention. However, for *rafa2*, there is no significant difference between the two conditions since it lacks of a main attention area, likely due to its low SI and TIand no particularly engaging features to attract viewers' attention. As a result, people tend to look around more. The possible reason for the higher spread of VoI for *dancer* while remaining the same VoI is due to its continuous movements over time, with the dynamic dance gestures evenly capturing attention across the point cloud.

Table 2: Property of Evaluation Metrics for Image Saliency Map

	AUC	NSS	IG	SIM	KL	PCC	EMD
Location-based	✓	✓	✓				
Distribution-based				✓	✓	✓	✓
Similarity	✓	✓	✓	✓		✓	
Dissimilarity					✓		✓
Sensitive to 0 values			✓	✓	✓		
With spatial distance							✓

5.2 Comparison Consistency of Visual Saliency Map

We aim to compare the point cloud saliency map in task-free and task-dependent scenarios. Commonly used metrics for such a comparison are listed in Table 2. The key properties include location or distribution-based, similarity or dissimilarity measurement, sensitivity to 0 values, and consideration of spatial distance. Since the generated saliency map for dynamic point clouds uses exactly the same method in [69], which does not include an explicit fixation point on the point cloud, the location-based metrics are not applicable to our continuous point cloud saliency maps. Among the distribution-based metrics, SIM, as a similarity metric, penalizes misalignment and is sensitive to missing values and 0 values, while KL, as a dissimilarity metric, is also sensitive to 0 values. Thus, based on the recommendation for metric selection [13, 48] and the characteristics of our dynamic point cloud saliency map, i.e., the majority of the points are non-salient (i.e., heat values equal to 0), we opt not to use them. EMD, as a dissimilarity, is the only metric that considers spatial distance. Herein we choose PCC to measure the similarity and adapt EMD, which is used to measure the 2D saliency map, to measure the dissimilarity.

The PCC is a statistical method to measure how correlated or dependent two variables are. In our scenario, given the visual saliency maps obtained from a task-free $\mathbf{F}$ and task-dependent $\mathbf{D}$ experiment, PCC can be defined as follows: [31]:

$$PCC(\mathbf{F},\mathbf{D}) = \frac{cov(\mathbf{F},\mathbf{D})}{\sigma_{\mathbf{F}}\sigma_{\mathbf{D}}}. \quad (1)$$

where $cov(\cdot)$ is the covariance and σ is the standard deviation. PCC ranges from -1 to 1, with higher absolute values indicating stronger correlation between visual saliency maps. However, PCC is sensitive to outliers and only compares the magnitudes of corresponding points. This makes it unable to account for shifts in point locations or partial matches in attended areas, which are common in eye-tracking experiments due to device limitations or participant preferences. This issue is especially noticeable in large point clouds. To address this, we propose to adapt EMD for dissimilarity measurement [50], as it better captures the distribution of attention by incorporating spatial information. EMD helps to alleviate the issues of point shifts and partial matches in large volumetric content cases. Specifically, we generate the "signature" (a feature that can represent the saliency map) by calculating a histogram of the heat value at each point in 3D space. We denote a discrete,

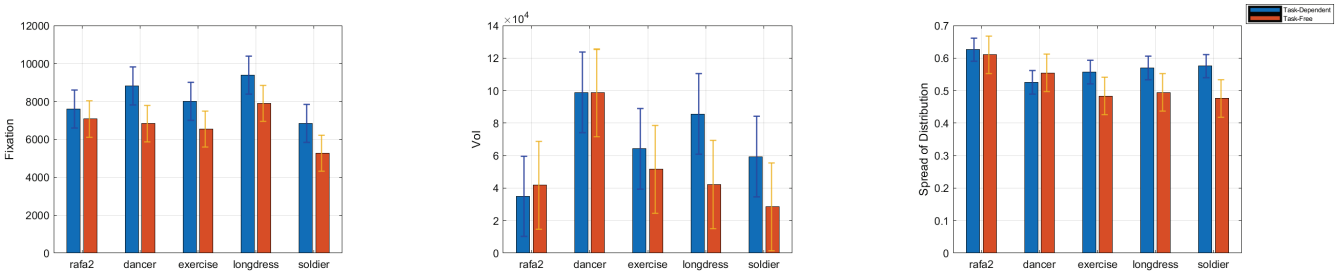
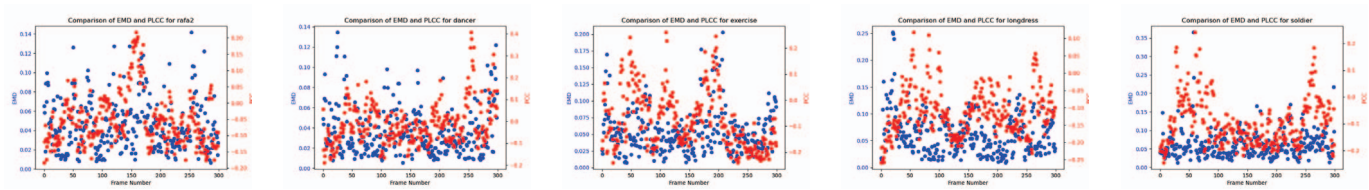


Fig. 6: Aggregation of fixations, Vol, and the spread of the distribution across participants of task-free and task-dependent experimental scenarios for the 5 shared dynamic point clouds from both QAVQ-DPC and proposed TF-DPC datasets, separately.



(a) rafa2 (b) dancer (c) exercise (d) longdress (e) soldier

Fig. 7: Similarity of point cloud saliency maps between task-free and task-dependent scenarios through EMD (•) and PCC (•) for the shared 5 sequences per frame, separately.

finite distribution $\mathbf{p}$ from the saliency map obtained in the task-free experiment as

$$\mathbf{p} = \{(p_1, w_1), \ldots, (p_m, w_m)\} \equiv (\mathbf{P}, \mathbf{w}) \in \mathbb{D}^{K \times m} \tag{2}$$

where $\mathbf{P} = [p_1, \ldots, p_m] \in \mathbb{R}^{K \times m}$ represents the signature with m points (or clusters), $w_i \geq 0$ represents the weight or fraction associated with the i-th point (or cluster) for all $i = 1, \cdots, m$. Here K is the dimension of ambient space (Euclidean space for 3D point cloud) of the points $p_i \in \mathbb{R}^K$, and m is the number of points (or clusters). The total weight of the distribution $\mathbf{p}$ is $w_\Sigma = \Sigma_{i=1}^m w_i$. Given two distributions in task-free and task-dependent scenarios as $\mathbf{p} = (\mathbf{P}, \mathbf{w}) \in \mathbb{D}^{K,m}$ and $\mathbf{q} = (\mathbf{Q}, \mathbf{u}) \in \mathbb{D}^{K,n}$. We used the following EMD [50]:

$$\mathrm{EMD}(\mathbf{p}, \mathbf{q}) = \frac{\min_{F=(f_{ij}) \in \mathscr{F}(\mathbf{p},\mathbf{q})} \mathrm{WORK}(F, \mathbf{p}, \mathbf{q})}{\min(w_\Sigma, u_\Sigma)}. \tag{3}$$

The EMD distance $\mathrm{EMD}(\mathbf{p}, \mathbf{q})$ between $\mathbf{p}$ and $\mathbf{q}$ is the minimum amount of work to match between distribution $\mathbf{p}$ and $\mathbf{q}$, normalized by the weight of the lighter distribution. Thus, to obtain the EMD value, we need to find the optimal flow by solving the transportation problem. The work done by a feasible flow $F \in \mathscr{F}(\mathbf{p}, \mathbf{q})$ in matching $\mathbf{p}$ and $\mathbf{q}$ is given by

$$\mathrm{WORK}(F, \mathbf{p}, \mathbf{q}) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}, \tag{4}$$

where $d_{ij} = d(p_i, q_j)$ is the "ground distance" between p_i and q_j. We consider the degree of salience and the spatial information of the point cloud jointly, the ground distance is now defined as

$$d_{ij} = \lambda |h_i - h_j| + (1 - \lambda)[(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{\frac{1}{2}}, \tag{5}$$

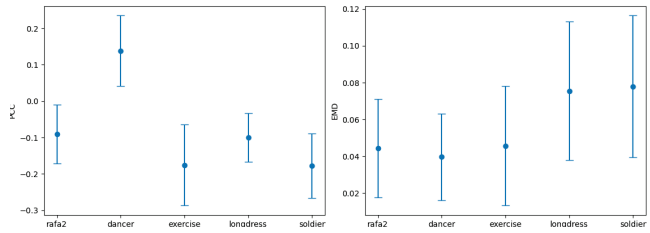
where h_i is the middle value of the i_{th} bin of the histogram in $\mathbf{p}$, and (x_i, y_i, z_i) is the location of the centroid point located in i_{th} bin of $\mathbf{p}$. λ is a weight used to balance the importance between spatial information and the magnitude of the heat value. The flow F is a feasible flow between $\mathbf{p}$ and $\mathbf{q}$ iff

$$f_{ij} \geq 0 \quad i = 1, \ldots, m, \; j = 1, \ldots, n, \tag{4.1}$$

$$\sum_{j=1}^{n} f_{ij} \leq w_i \quad i = 1, \ldots, m, \tag{4.2}$$

$$\sum_{i=1}^{m} f_{ij} \leq u_j \quad j = 1, \ldots, n, \quad \text{and} \tag{4.3}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min(w_\Sigma, u_\Sigma). \tag{4.4}$$



(a) Average PCC (b) Average EMD

Fig. 8: Similarity of point cloud visual saliency maps between task-free and task-dependent for the shared 5 sequences averaged over 300 frames, separately.

The detailed explanation for the constraints can be found in [50]. The coordinates of the distribution points are not used directly in the EMD formulation, only the ground distances d_{ij} between points are needed. A larger EMD indicates a larger difference between two distributions while an EMD of zero indicates that two distributions are the same. In this paper, we remove the points that are non-salient in both experiments before we compute the PCC and EMD to obtain an accurate measurement. The bin number of the histogram is set to 30, λ is set to 0.5.

To fairly compare similarity and dissimilarity metrics, we normalize the EMD values to [0,1] range and convert dissimilarity into similarity. This is achieved by dividing the computed EMD by the maximum possible EMD for a given histogram, assuming all the mass (i.e., salient points) starting in the leftmost bin need to be moved to the rightmost bin. The similarity score for EMD is then calculated as 1 minus the normalized EMD. Figure 7 compares PCC and EMD values for the shared 5 contents per frame, separately. We observe that PCC exhibits greater variance for *exercise*, *longdress*, and *soldier*, as evidenced by fluctuations in the PCC values across frames. This variability suggests

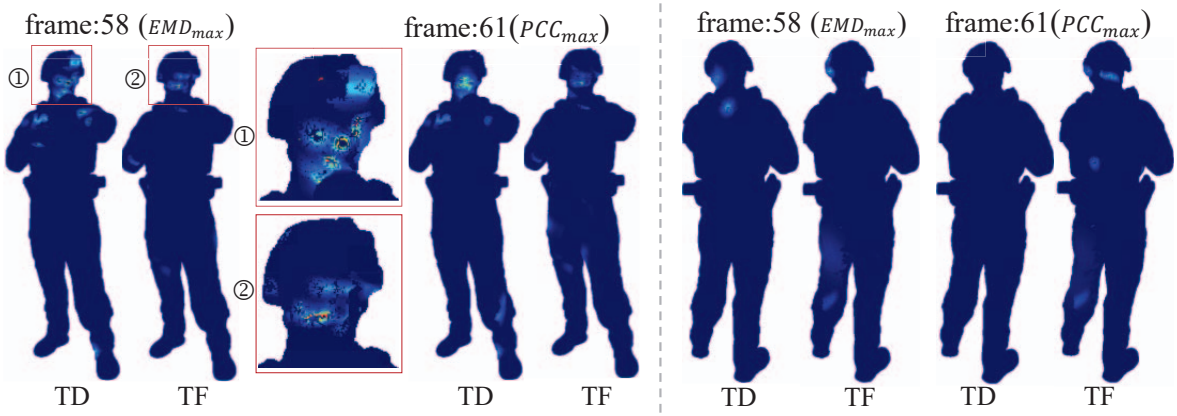


Fig. 9: Saliency map visualization of *soldier* in frame 58 and frame 61, identified as the most similar maps using the adapted EMD and PCC metrics. The left side of the dotted line shows the front view of the *soldier*, while the right side shows the back view. TD refers to the saliency map collected under task-dependent conditions, and TF refers to task-free.

that PCC is sensitive to outliers in the saliency map, leading to greater variation in visual similarity over time for these contents. In contrast, EMD demonstrates more stable and consistent behavior, with values that remain within a narrower range, indicating reduced fluctuations. This stability arises from EMD's consideration of spatial information and its partial match property. Figure 8b and 8a show the average similarity across frames in task-free and task-dependent scenarios. Notably, *dancer* is identified as the most similar sequence by PCC, while *soldier* is the most similar according to EMD. PCC's emphasis on matching magnitudes at the same points leads to high similarity scores for *dancer*, where obvious salient regions identified by humans remain consistent over time, independently of the task.

Combining Figure 7 and Figure 8, it becomes clear that both EMD ([0, 0.35]) and PCC ([-0.25, 0.4]) exhibit low similarity values, suggesting substantial differences between task-free and task-dependent scenarios. This highlights that task-dependent scenarios in dynamic point clouds significantly alter human visual attention. EMD identifies overlapping regions of attention in both scenarios, providing a more spatially-aware similarity measure, while PCC captures sharp variations for specific content. Figure 9 shows saliency maps for *soldier* at the frames with maximum similarity under EMD and PCC metrics. Visually, the saliency in the 58_{th} frame appears more similar than in the 61_{th} frame, with the inset of the head showing greater overlap, particularly from the back view. This comparison further demonstrates that while both PCC and EMD have their strengths, EMD's consideration of spatial information makes it more suitable for evaluating saliency in point cloud data.

5.3 Summary

Quality assessment, as a high-level perceptual task, significantly influences how visual attention is deployed when evaluating dynamic point clouds in VR. As discussed in Section 5.1, one key observation is that participants exhibit fewer fixations in task-free conditions compared to task-dependent ones. This is evident in Figure 9, where task-dependent viewers focus more on specific details, such as the spotlight on the soldier's hat. In contrast, task-free viewers typically form a general impression, primarily attending to broader features like facial expressions, rather than thoroughly exploring "less critical" details once they have grasped the overall scene.

In task-dependent conditions, the demand for precise quality evaluation prompts participants to observe the sequence more carefully. Their goal is to gather visual cues to assess the content's quality, which explains why saliency maps under the quality assessment task tend to have a larger VoI. Additionally, due to content repetition (same content with different quality level), participants in task-dependent conditions are less inclined to explore the back of the point cloud, preferring the primary areas in the front view that they deem relevant for the quality assessment task. In task-free conditions, participants generally scan the content broadly, focusing on prominent movements or artifacts. Since they are not bound by a specific objective, they tend to observe both the front and back views of the point clouds without particular focus.

The spread of the VoI, however, varies between different conditions for different reasons. in task-dependent, participants' attention is drawn to specific features from head to toe, like the spotlight on the hat, the watch on the hand, and the shoes, as shown in Figure 5 the frame 58 under task-dependent condition. Participants' attention is more targeted, with individual differences in strategies for assessing quality. This variability contributes to the spread of the VoI but with greater focus on elements that are crucial to quality judgment. In contrast, the task-free condition reflects a more passive viewing approach. Participants form a holistic view of the scene, only directing their gaze toward areas of movement or obvious artifacts. Without the demand to assess quality, their focus is less concentrated on specific details, and their viewing patterns reflect a broader exploration of the scene.

Movement and semantic information in the dynamic point clouds, such as facial expressions or body movements, consistently attract visual attention in both scenarios. For example, in Figure 1, participants frequently fixate on faces across multiple frames. Interestingly, visual attention appears to be more consistent in task-dependent conditions, especially when it comes to fine details, regardless of whether the scene has high or low TI. Participants are more likely to scrutinize these details to detect subtle distortions, which are critical for assigning quality scores. This difference in attention deployment highlights how task-driven objectives shape visual behavior, with task-dependent viewers engaging in top-down mechanisms and task-free viewers adopting a more relaxed, impressionistic approach.

6 Discussion

6.1 Visual attention collection limitations

In this study, we collect a task-free saliency dataset for dynamic point clouds and investigate the task impact on human attention allocation. We observed that a central bias persists to some extent when viewing human faces, regardless of whether the conditions are task-free or task-dependent. However, our study is limited by the fact that TF-DPC focuses solely on human figures, excluding other immersive content types like landscapes or interactive objects. This limitation stems directly from the lack of high-quality, realistic datasets of dynamic point cloud objects, as to date, only synthetic datasets including dynamic objects are present in the literature [58, 63]. Thus, the outcomes of this study are valid only for the dynamic human category, and future work should explore broader content types. We chose a wired HMD to maintain consistency with the conditions of the previous study; however, this choice restricted physical movement due to the HMD's cable, and the device's weight and discomfort may have increased cognitive load, potentially resulting in fewer and less stable fixation points. To explore this, future studies should consider assessing pupil dilation and blink rate, reliable indicators of cognitive load, alongside gaze amplitude and fixation patterns. These constraints may limit the

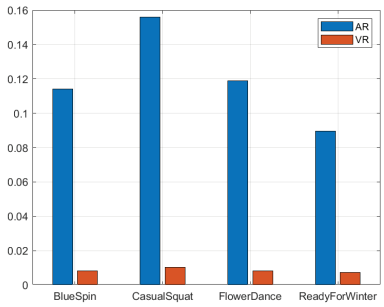


Fig. 10: The average ratio of the VoI for the shared 4 dynamic point cloud sequences in AR and VR.

ability to collect naturally viewing saliency maps and could introduce systematic biases. Using wireless HMDs, such as the HTC Vive Focus Vision, could improve ecological validity. Additionally, dynamic point clouds in high-quality XR scenarios are inherently dense, but the visual saliency regions occupy only a small portion of the content. Increasing the participant sample size in future studies would enhance statistical power and improve the generalizability of the findings.

6.2 Visual saliency collection under various perceptual tasks

The findings of our study on the impact of high-level tasks for human visual attention deployment differ from previous research on images [17] but align with conclusions drawn from static 3D models [53]. Specifically, similarity metrics indicate lower saliency collection for static 3D models (PCC: 0.35) [53] compared to images (PCC: 0.84) [17]. While task-dependent, top-down mechanism effects on overt visual attention have been well-studied for 2D media [29], how these findings translate to dynamic point clouds remains largely unexplored. Additionally, there is evidence that traditional attention paradigms may not fully apply to newer media formats, such as panoramic videos [52]. Our findings have shown that quality assessment has a significant impact on human visual attention deployment, with both saliency maps under task-free and quality assessment tasks focusing on semantic area and movement. However, their focus differs, as mentioned in the above Section 5.3. A critical question that emerges from our study is whether saliency collected under task-free conditions or task-dependent conditions provides greater value for specific applications, such as point cloud quality assessment. Exploring the temporal dynamics of saliency in dynamic point clouds–how it evolves over time under varying task demands–critical for optimizing visual representations. Future research should focus on exploring the temporal dynamics of saliency across various perceptual tasks, clarifying the benefits of different saliency detection methods, and incorporating these insights into prediction models tailored to dynamic point clouds for specific applications.

6.3 Visual saliency collection in AR

3D visual saliency has been measured using various devices, including eye-tracking glasses [41], AR HMD [40], and VR HMD [69]. Understanding the differences between these devices is essential for accurately predicting saliency while accounting for factors such as spatial bias [28], center bias [45], and systematic error [3]. Nguyen [40] released saliency maps for four dynamic point clouds (namely *BlueSpin*, *CasualSquat*, *FlowerDance*, and *ReadyForWinter*) in AR, overlapping with our proposed TF-DPC dataset. Thus, using these four sequences, we were able to conduct an initial analysis of saliency maps across different devices. Notably, not every frame in the AR sequences contains fixation data, so we retained only the frames with salient areas present in both AR and VR. We computed the average VoI ratio (salient area relative to the entire point cloud across the sequence), as shown in Figure 10. Our findings indicate that the VoI in the AR condition is significantly smaller than in the VR laboratory setting, with participants primarily focusing on limited regions of the point cloud. This reduction may be attributed to the HoloLens' limited field of view (about 52°) compared to VR headset (about 110°). Furthermore, since AR blends virtual context with the real environment, users must frequently switch contexts and refocus their gaze [8], which can further reduce fixations on dynamic point clouds. Additionally, participants cannot view the entire life-sized point cloud unless they step back. Thus, the experimental protocol for saliency collection in AR requires careful consideration.

6.4 Evaluation metrics for the similarity of point cloud saliency maps

Several metrics exist for quantitatively measuring the similarity of 2D saliency maps, some of which can be adapted to static point cloud saliency maps with minimal adjustments. However, location-based metrics like NSS, which depend on precise fixation points, may not be directly applicable to point clouds. Human gaze fixation corresponds to a specific pixel in 2D images, but in 3D point clouds, the gaze ray may not intersect with any point in space, requiring approximation methods that introduce inaccuracies. Thus, metrics relying on fixation locations may not be suitable for point clouds unless these approximations are properly addressed. For distribution-based metrics, which compare the overall spread of attention, present a different challenge: how should we balance coverage similarity (whether the same areas are salient, regardless of magnitude) against magnitude similarity (whether the saliency levels are comparable)? Some scenes may show full spatial matches but differ in magnitude, or vice versa, making it unclear which aspect should be prioritized. This decision depends on the specific application.

Riche *et al.* [48] argues that no single metric is sufficient for evaluating saliency map similarity. The 3D nature of point clouds and the relatively small salient regions further complicate this task. For dynamic point clouds, the added dimension of time introduces variability due to motion, requiring spatial-temporal saliency distributions to be more effective in measuring similarity. Especially for human dynamic point clouds, for example, in Figure 1, the 151_{st} frame of *dancer* sequence, should the saliency of symmetric semantic areas (the left and right feet) be treated equivalently when we measure the similarity? Incorporating metrics that consider temporal consistency and semantic relationships could help capture nuances in saliency similarity, particularly in dynamic scenarios where motion and semantic equivalency, such as symmetrical regions, play a significant role.

7 Conclusion

In this work, we constructed a task-free visual saliency dataset in virtual reality with 6-DoF, comprising 19 dynamic point clouds. We analyze gaze and movement trajectories to explore how visual attention is allocated in dynamic point clouds. To compare the generated saliency maps in task-free and task-dependent conditions, we evaluate gaze statistics and the similarity of the saliency maps. Additionally, we introduced a novel metric based on the earth mover's distance, which incorporates both spatial information and salience levels, enabling us to quantify the dissimilarity of saliency maps in dynamic point clouds. Our experimental results show that high-level tasks, such as quality assessment, significantly affect human visual attention, and this effect varies based on content characteristics, particularly the temporal information.

Acknowledgments

This work was supported through the NWO WISE grant and the European Commission Horizon Europe program, under the grant agreement 101070109, *TRANSMIXR* https://transmixr.eu/. Funded by the European Union.

References

[1] ITU-T Rec. p.910 (04/2008) subjective video quality assessment methods for multimedia applications. 2009. 3

[2] M. Abid, M. P. Da Silva, and P. Le Callet. Towards visual saliency computation on 3d graphical contents for interactive visualization. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 3448–3452. IEEE, 2020. 2

[3] M. Adamove, D. Padúch, and P. Kapec. Evaluation of visual saliency models in immersive analytics. In K. Arai, ed., *Advances in Information*

and Communication, pp. 375–392. Springer Nature Switzerland, Cham, 2024. 9

[4] H. Alers, L. Bos, and I. Heynderickx. How the task of evaluating image quality influences viewing behavior. In *2011 Third International Workshop on Quality of Multimedia Experience*, pp. 167–172. IEEE, 2011. 6

[5] H. Alers, J. Redi, H. Liu, and I. Heynderickx. Effects of task and image properties on visual-attention deployment in image-quality assessment. *Journal of Electronic Imaging*, 24(2):023030–023030, 2015. 2

[6] E. Alexiou, Y. Nehmé, E. Zerman, I. Viola, G. Lavoué, A. Ak, A. Smolic, P. Le Callet, and P. Cesar. Subjective and objective quality assessment for volumetric video. In *Immersive Video Technologies*, pp. 501–552. Elsevier, 2023. 1

[7] E. Alexiou, P. Xu, and T. Ebrahimi. Towards modelling of visual saliency in point clouds for immersive applications. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4325–4329. IEEE, 2019. 1, 2, 4

[8] M. S. Arefin, N. Phillips, A. Plopski, J. L. Gabbard, and J. E. Swan. Impact of ar display context switching and focal distance switching on human performance: Replication on an ar haploscope. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 571–572, 2020. 9

[9] I. B. Adhanom, S. C. Lee, E. Folmer, and P. MacNeilage. Gazemetrics: An open-source tool for measuring the data quality of hmd-based eye trackers. In *ACM symposium on eye tracking research and applications*, pp. 1–5, 2020. 4

[10] T. Betz, T. C. Kietzmann, N. Wilming, and P. Koenig. Investigating task-dependent top-down effects on overt visual attention. *Journal of vision*, 10(3):15–15, 2010. 1

[11] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2012. 1

[12] S. Bourbia, A. Karine, A. Chetouani, M. El Hassouni, and M. Jridi. No-reference point clouds quality assessment using transformer and visual saliency. In *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, pp. 57–62, 2022. 2

[13] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2019. 3, 6

[14] X. Chen, A. Saparov, B. Pang, and T. Funkhouser. Schelling points on 3D surface meshes. *ACM Transactions on Graphics (TOG)*, 31(4):1–12, 2012. 2

[15] J. Clark. The ishihara test for color blindness. *American Journal of Physiological Optics*, 1924. 3

[16] X. Ding, W. Lin, Z. Chen, and X. Zhang. Point cloud saliency detection by local and global feature fusion. *IEEE Transactions on Image Processing*, 28(11):5379–5393, 2019. 2

[17] U. Engelke, H. Liu, H.-J. Zepernick, I. Heynderickx, and A. Maeder. Comparing two eye-tracking databases: The effect of experimental setup and image presentation time on the creation of saliency maps. In *28th Picture Coding Symposium*, pp. 282–285, 2010. 9

[18] F. L. Ferris III, A. Kassoff, G. H. Bresnick, and I. Bailey. New visual acuity charts for clinical research. *American journal of ophthalmology*, 94(1):91–96, 1982. 3

[19] G. Gautier, A. Mercat, L. Fréneau, M. Pitkänen, and J. Vanne. Uvg-vpc: Voxelized point cloud dataset for visual volumetric video-based coding. In *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 244–247. IEEE, 2023. 3

[20] H. Hadizadeh and I. V. Bajić. Saliency-aware video compression. *IEEE Transactions on Image Processing*, 23(1):19–33, 2013. 1

[21] J. Hadnett-Hunter, G. Nicolaou, E. O'Neill, and M. Proulx. The effect of task on visual attention in interactive virtual environments. *ACM Transactions on Applied Perception (TAP)*, 16(3):1–17, 2019. 2

[22] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982. 3

[23] Z. Hu. Gaze analysis and prediction in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 543–544, 2020. 2

[24] Z. Hu. [dc] eye fixation forecasting in task-oriented virtual reality. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 707–708, 2021. 2

[25] Z. Hu, A. Bulling, S. Li, and G. Wang. Ehtask: Recognizing user tasks from eye and head movements in immersive virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 29(4):1992–2004, 2021. 2

[26] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993. 3, 4

[27] Y. Kim and A. Varshney. Saliency-guided enhancement for volume visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):925–932, 2006. 2

[28] S. Kollmorgen, N. Nortmann, S. Schröder, and P. König. Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLoS computational biology*, 6(5):e1000791, 2010. 1, 9

[29] S. Kollmorgen, N. Nortmann, S. Schröder, and P. König. Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLOS Computational Biology*, 6(5):1–20, 05 2010. 9

[30] M. Kümmerer, T. S. Wallis, and M. Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52):16054–16059, 2015. 3

[31] O. Le Meur, P. Le Callet, and D. Barba. Predicting visual fixations on video based on low-level visual features. *Vision research*, 47(19):2483–2498, 2007. 6

[32] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba. Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric. *Signal Processing: Image Communication*, 25(7):547–558, 2010. 1, 2

[33] H. Li, G. Chen, G. Li, and Y. Yu. Motion guided attention for video salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7274–7283, 2019. 1

[34] J. Li, C. Xia, Y. Song, S. Fang, and X. Chen. A data-driven metric for comprehensive evaluation of saliency models. In *Proceedings of the IEEE international conference on computer vision*, pp. 190–198, 2015. 3

[35] M. Limper, A. Kuijper, and D. W. Fellner. Mesh saliency analysis via local curvature entropy. In *Eurographics (Short Papers)*, pp. 13–16, 2016. 2

[36] H. Liu and I. Heynderickx. Visual attention in objective image quality assessment: Based on eye-tracking data. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(7):971–982, 2011. 1, 2

[37] M. Maguire and B. Delahunt. Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *All Ireland journal of higher education*, 9(3), 2017. 4

[38] S. Malpica, D. Martin, A. Serrano, D. Gutierrez, and B. Masia. Task-dependent visual behavior in immersive environments: A comparative study of free exploration, memory and visual search. *IEEE transactions on visualization and computer graphics*, 2023. 2

[39] D. Martin, A. Fandos, B. Masia, and A. Serrano. Sal3d: a model for saliency prediction in 3d meshes. *The Visual Computer*, pp. 1–11, 2024. 2

[40] M. Nguyen, S. Vats, X. Zhou, I. Viola, P. Cesar, C. Timmerer, and H. Hellwagner. Compeq-mr: Compressed point cloud dataset with eye tracking and quality assessment in mixed reality. In *Proceedings of the 15th ACM Multimedia Systems Conference*, pp. 367–373, 2024. 1, 2, 9

[41] L. Paletta, K. Santner, G. Fritz, H. Mayer, and J. Schrammel. 3d attention: measurement of visual saliency using eye tracking glasses. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pp. 199–204. 2013. 9

[42] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18):2397–2416, 2005. 3

[43] P. Polatsek, M. Waldner, I. Viola, P. Kapec, and W. Benesova. Exploring visual attention and saliency modeling for task-based visual analysis. *Computers & Graphics*, 72:26–38, 2018. 1

[44] E. Psatha, D. Laskos, G. Arvanitis, and K. Moustakas. Aggressive saliency-aware point cloud compression. *arXiv preprint arXiv:2307.10741*, 2023. 2

[45] S. Rahman and N. Bruce. Visual saliency prediction and evaluation across different perceptual tasks. *PloS one*, 10(9):e0138053, 2015. 1, 9

[46] I. Reimat, E. Alexiou, J. Jansen, I. Viola, S. Subramanyam, and P. Cesar. Cwipc-sxr: Point cloud dynamic human dataset for social xr. In *Proceedings of the 12th ACM Multimedia Systems Conference*, pp. 300–306, 2021. 3

[47] L. Renninger, J. Coughlan, P. Verghese, and J. Malik. An information maximization model of eye movements. *Advances in neural information processing systems*, 17, 2004. 3

[48] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *Proceedings of the IEEE international conference on computer vision*, pp. 1153–1160, 2013. 3, 6, 9

[49] S. Rossi, I. Viola, and P. Cesar. Behavioural analysis in a 6-dof vr system: Influence of content, quality and user disposition. In *Proceedings of the 1st Workshop on Interactive eXtended Reality*, pp. 3–10, 2022. 4

[50] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40:99–121, 2000. 3, 6, 7

[51] P. Ruiu, L. Mascia, and E. Grosso. Saliency-guided point cloud compression for 3d live reconstruction. *Multimodal Technologies and Interaction*, 8(5):36, 2024. 2

[52] A. Schmitz, A. MacQuarrie, S. Julier, N. Binetti, and A. Steed. Directing versus attracting attention: Exploring the effectiveness of central and peripheral cues in panoramic videos. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 63–72, 2020. 9

[53] O. Sidorov, J. S. Harvey, H. E. Smithson, and J. Y. Hardeberg. Overt visual attention on rendered 3d objects. *arXiv preprint arXiv:1905.10444*, 2019. 9

[54] R. Singh, M. Huzaifa, J. Liu, A. Patney, H. Sharif, Y. Zhao, and S. Adve. Power, performance, and image quality tradeoffs in foveated rendering. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 205–214, 2023. 2

[55] A. K. Sinha and K. Shukla. A study of distance metrics in histogram based image retrieval. *Int. J. Comput. Technol*, 4(3):821–830, 2013. 2

[56] I. Stein, H. Jossberger, and H. Gruber. Map3d: An explorative approach for automatic mapping of real-world eye-tracking data on a virtual 3d model. *Journal of Eye Movement Research*, 15(3), 2022. 6

[57] S. Subramanyam, J. Li, I. Viola, and P. Cesar. Comparing the quality of highly realistic digital humans in 3dof and 6dof: A volumetric video case study. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 127–136. IEEE, 2020. 3

[58] Y.-C. Sun, I.-C. Huang, Y. Shi, W. T. Ooi, C.-Y. Huang, and C.-H. Hsu. A Dynamic 3D Point Cloud Dataset for Immersive Applications. In *Proceedings of the 14th Conference on ACM Multimedia Systems*, pp. 376–383, 2023. 8

[59] F. P. Tasse, J. Kosinka, and N. Dodgson. Cluster-based point set saliency. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 163–171, 2015. 2

[60] I. Viola, S. Subramanyam, J. Li, and P. Cesar. On the impact of vr assessment on the quality of experience of highly realistic digital humans: A volumetric video case study. *Quality and User Experience*, 7(1):3, 2022. 5

[61] Z. Wang, Y. Zhang, Q. Yang, Y. Xu, J. Sun, and S. Liu. Point cloud quality assessment using 3d saliency maps. In *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–5. IEEE, 2023. 2

[62] C. Wloka and J. K. Tsotsos. Overt fixations reflect a natural central bias. *Journal of Vision*, 13(9):239–239, 2013. 1

[63] L. Xie, X. Mu, G. Li, W. Gao, et al. PKU-DPCC: A New Dataset for Dynamic Point Cloud Compression. *APSIPA Transactions on Signal and Information Processing*, 13(6), 2024. 8

[64] Y. Xu, Y. Lu, and Z. Wen. Owlii dynamic human mesh sequence dataset. ISO/IEC JTC1/SC29/WG11 MPEG Document m41658, ISO/IEC JTC1/SC29/WG11, 120th MPEG Meeting, Macau, October 2017. 3

[65] W. Zhang and H. Liu. Toward a reliable collection of eye-tracking data for image quality research: Challenges, solutions, and applications. *IEEE Transactions on Image Processing*, 26(5):2424–2437, 2017. 5

[66] Y. Zhang, K. Ding, N. Li, H. Wang, X. Huang, and C.-C. J. Kuo. Perceptually weighted rate distortion optimization for video-based point cloud compression. *IEEE Transactions on Image Processing*, 32:5933–5947, 2023. 2

[67] T. Zheng, C. Chen, J. Yuan, B. Li, and K. Ren. Pointcloud saliency maps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1598–1606, 2019. 2

[68] W. Zhou, G. Yue, R. Zhang, Y. Qin, and H. Liu. Reduced-reference quality assessment of point clouds via content-oriented saliency projection. *IEEE Signal Processing Letters*, 30:354–358, 2023. 2

[69] X. Zhou, I. Viola, E. Alexiou, J. Jansen, and P. Cesar. QAVA-DPC: eye-tracking based quality assessment and visual attention dataset for dynamic point cloud in 6 DoF. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 69–78. IEEE, 2023. 1, 2, 3, 4, 6, 9