Annorama: Enabling Immersive At-Desk Annotation Experiences in Virtual Reality with 3D Point Cloud Dioramas

Subramanian Chidambaram
subbuc@amazon.com
AWS AI, Amazon
Santa Clara, CA, USA

Alex C. Williams
acwio@amazon.com
AWS AI, Amazon
Santa Clara, CA, USA

Min Bai
baimin@amazon.com
AWS AI, Amazon
New York City, NY, USA

Satyugjit Virk
svirk@amazon.com
AWS AI, Amazon
Santa Clara, CA, USA

Patrick Haffner
haffnerp@amazon.com
AWS AI, Amazon
New York City, NY, USA

Matthew Lease
ml@utexas.edu
The University of Texas at Austin
Austin, TX, USA
Last Mile Driver Science, Amazon
Austin, TX, USA

Li Erran Li
lilimam@amazon.com
AWS AI, Amazon
Santa Clara, CA, USA

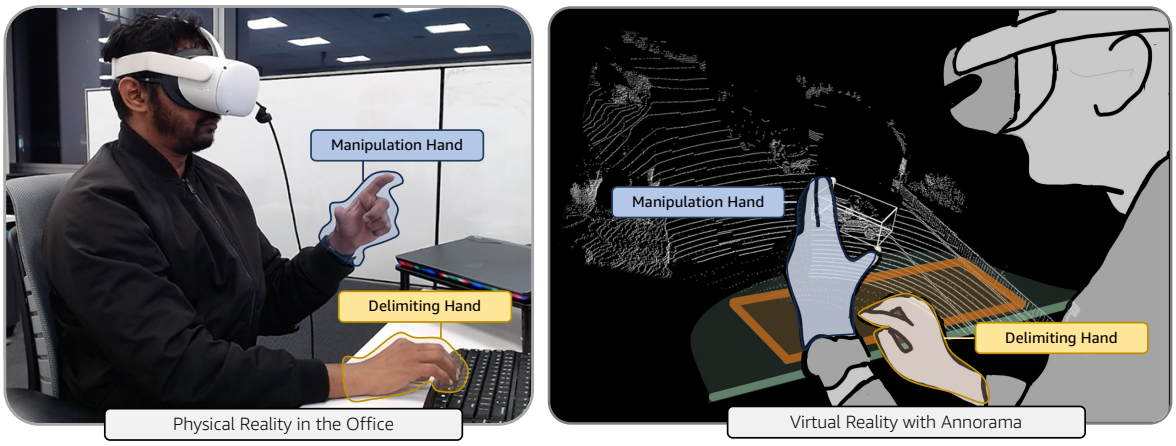


Figure 1: Annorama enables immersive point cloud data annotation experiences in virtual reality for space-constrained desk environments (A). Annorama facilitates these experiences by rendering 3D point cloud dioramas (B), a miniaturized representation of point cloud scenes that scales the scene to the environment's desk space while simultaneously preserving its point cloud density. Annorama provides a set of delimited mid-air gestures that allow users to create and manage 3D cuboid annotations in this 3D space using principles of direct manipulation.

ABSTRACT

Point cloud annotation plays a pivotal role in computer vision and machine learning by facilitating the creation of volumetric annotations in 3D space. While prior research has explored point cloud annotation in VR environments, its practical implementation in space-constrained office settings, where data annotation is typically conducted, remains an open question. In this paper, we introduce Annorama, an interactive system that translates 3D point cloud scenes into miniature desk-scale dioramas, enabling annotation using a unique family of keyboard-assisted mid-air gestures inspired by direct manipulation. Through a within-subjects study with 16 participants, we demonstrate the feasibility of our system by assessing the efficacy of four types of mid-air gestures for drawing cuboid annotations. Our findings suggest that Annorama allows for rapid and accurate annotation of point cloud data, particularly with the Sizing and Two Point Gestures.

SUI '24, October 07–08, 2024, Trier, Germany

ACM ISBN 979-8-4007-1088-9/24/10
https://doi.org/10.1145/3677386.3682081

CCS CONCEPTS

• **Human-centered computing** → *Virtual reality*.

KEYWORDS

Virtual reality, point cloud data annotation, annotator productivity.

ACM Reference Format:
Subramanian Chidambaram, Alex C. Williams, Min Bai, Satyugjit Virk, Patrick Haffner, Matthew Lease, and Li Erran Li. 2024. Annorama: Enabling Immersive At-Desk Annotation Experiences in Virtual Reality with 3D Point Cloud Dioramas. In *ACM Symposium on Spatial User Interaction (SUI '24), October 07–08, 2024, Trier, Germany*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3677386.3682081

1 INTRODUCTION

Point cloud annotation is a computer-based task that involves the detection and annotation of relevant objects in 3D point cloud scenes. Point cloud annotations are commonly represented as 3D cuboids and are often collected to support the training of machine learning algorithms to automatically identify the annotated objects. As an annotation context, point cloud annotation has been employed across a range of domains, including autonomous driving [60], digital twinning [33, 48], drone management [13], AR/VR authoring [69], and robotics [43]. Reports suggest that data annotation will scale to an $8-billion industry by 2028 [53], with the growth of point cloud annotations being driven by continued investments in the autonomous driving sector. As the data annotation industry continues to grow, it will become increasingly important for service providers, such as Amazon SageMaker GroundTruth [1] and Scale.AI [2], to understand how point cloud annotation can be optimized to reduce costs both for themselves and for their customers.

Point cloud annotation is often recognized as a time-consuming, labor-intensive, and expensive task. Prior research has studied how pre-trained machine learning models and clustering algorithms, for example, can ease the burden of manually annotating or classifying relevant 3D objects [67, 76]. However, these techniques require pre-processing steps that can also be time-consuming and require additional data collection, model training, and the application of density-reduction techniques to the point cloud scene. In contrast to automated techniques, research has explored how virtual reality (VR) environments can expedite the task by allowing users to annotate point cloud scenes in an immersive fashion [71]. Despite substantially accelerating point cloud annotation, these VR-based approaches are ill-suited to space-constrained environments where physical space is limited. Prior research suggests that such constraints are particularly pertinent to professional data annotation contexts, which mirror the constraints of conventional information workplaces [68]. To this end, a recent trend involves utilizing head-mounted displays (HMDs) to enhance 2D screen applications by extending traditional displays into 3D, which has been found to be easier to use and more intuitive [47]. This is exemplified by products like Apple Vision Pro's infinite canvas [3], Lenovo ThinkReality Workspace [4], and Oculus Infinite Office [5], offering additional benefits and overcoming limitations imposed by screen real estate. In the context of 3D user interface research, point cloud annotation is essentially a '3D selection' task [52]. For Computer Vision (CV) applications, this involves selecting a group of points and labeling them, which constitutes an annotation task.

In this paper, we explore the efficacy of adapting immersive approaches in point cloud annotation to the physical constraints of professional data annotation. To that end, we introduce Annorama, an interactive system that facilitates immersive, at-desk annotation experiences in virtual reality. Drawing inspiration from Stoakley et al.'s World in Miniature (WiM) metaphor [59] and Magnoramas [73], Annorama creates, manages, and renders point cloud scenes as point cloud dioramas, a representation of point cloud scenes that are miniaturized and, by default, anchored to the user's desk. Using principles of direct manipulation, Annorama allows users to scale, rotate, transform, and annotate its rendered dioramas using a set of delimited mid-air gestures. In a "First-use" [27] study with 16 participants, we found that some of Annorama's gestures provided more assistance to participants in annotating 3D bounding boxes faster and more precisely than others. Our investigation revealed that the design of mid-air gestures played a crucial role in the annotation process using dioramas. A well-designed gesture eliminated the need for additional editing operations, resulting in a reduction in overall annotation time. Additionally, we observed variations in user preferences regarding ease, comfort, and usability among different gestures. Users generally favored gestures that provided additional fidelity over control points; however, an excessive level of fidelity was deemed unnecessary and time-consuming. In this paper, we present the following contributions:

(1) We introduce Annorama, a system enabling immersive point cloud annotation at the desk using 3D dioramas and one-handed mid-air parametric gestures.
(2) We provide a set of design recommendations based on a within-subject study with 16 participants, focusing on mid-air gestures for future hybrid spatial user interaction in at-desk immersion.

2 RELATED WORK

2.1 Point Cloud Annotation

High-quality 3D point cloud annotation is critical for enabling the training of novel machine learning algorithms, with applications across various industries [63] such as autonomous driving [60], robotics [43], UAV/drone navigation [13], AR/VR authoring [14, 49], and geospatial data analysis. However, current 3D point cloud annotation practices often rely on 2D visualization tools, such as monitors, keyboards, and mouse interfaces. While this approach ensures accessibility and broad compatibility, it falls short in fully exploiting the 3D nature of point cloud data and lacks the advantages of immersive data analysis and depth perception.

Several research and commercial products have employed variations of this 2D interface approach in their design strategies. For instance, ScanNet [18] developed a web-based tool for semantic labeling using mouse clicks on models generated from RGBD data. Wong et al. [72] introduced SmartAnnotator, which suggests labels based on prior annotations, streamlining the annotation process.

[1] Amazon Web Services, "Amazon SageMaker Data Labeling," Sep 2023.
[2] Scale.ai, "3D Sensor Fusion," Sep 2023.
[3] Apple, "Apple Vision Pro infinite canvas," Sep 2023.
[4] Lenovo, "ThinkReality A3 Smart Glasses Workspace," Sep 2021.
[5] Oculus, "Facebook Connect: Oculus Quest 2 Infinite Office," Sep 2021.

Russell et al. [54] introduced an image labeling tool using a lasso technique, later expanding it to generate 3D data.

In the realm of commercial 3D point cloud data annotation services, platforms such as Amazon SageMaker Ground Truth, Scale.ai, MathWorks' Ground Truth Labeler [6], and Pointly.ai [7], along with research-based tools like those developed by Zimmer et al. [75] and 3D BAT [76], allow point cloud labeling while viewing corresponding images captured with a surround camera. They employ label interpolation for sequence annotation, reducing effort. Veit et al. [65] explored the use of a mobile phone for 3D interaction with a scene viewed on a monitor. While not significantly boosting annotation performance, it simplified the process. LATTE [67] explored one-click annotation through clustering algorithms like DBSCAN [21], requiring preprocessing like ground point removal. Bacim et al. [2] employed hand-based gestures for data selection. Despite these innovative concepts, these approaches fall short of fully leveraging the 3D nature of point cloud data, confining annotators to 2D monitors and missing out on the benefits of immersive data analysis and depth perception [61].

2.2 Point Cloud Annotation in Virtual Reality

Many researchers have explored how to leverage this higher level of immersion and enable 3D point cloud interaction in immersive environments. Work such as VRFromX [32, 33], Garrido et al. [23], and Stets et al. [58] have introduced a paintbrush metaphor for highlighting and selecting points. Approaches like Immersive-Labeler [19] and PointCloudLab [20] have explored 3D bounding box annotation with VR controllers, where one controller provides constraint functions, enhancing the efficiency and accuracy of labeling point cloud data compared to traditional keyboard and mouse interfaces. Through a hybrid immersive desktop-based virtual reality (VR) annotation system, Franzluebbers et al. [22] explored the differences in point cloud annotation between 3D interfaces with tracked controllers and 2D interfaces with a mouse and keyboard. These approaches are often constrained by physical space limitations and require complex VR navigation algorithms, which are incompatible with most information workplaces.

Recent research has demonstrated the efficacy of free-hand, natural gestures as alternatives to controller-based interactions through a number of system evaluations. SemanticPaint [64] presents a workflow involving a Virtual Reality headset and depth camera for capturing design point clouds, converting them to 3D meshes, and using hands for automated labeling. However, it is limited by special hardware requirements and is unsuitable for sparse point clouds, which are common in LiDAR scans. Other works like Slice-n-Swipe [3], Burgess et al. [11], and Krug et al. [34] offer guidance on interaction techniques for aligning 3D models to point clouds in space. Most relevant to our own work, Lubos et al. [37] introduced Touching the Cloud, an interface that allows users to annotate point cloud data with bimanual, hands-free gestures in VR while seated at a desk. An important limitation of Lubos et al.'s work is that it did not evaluate the introduced system in any capacity.

[6]MatWorks, "Get Started with Ground Truth Labeling," Sep 2023.
[7]Pointly.ai, "Point Cloud Custom Classifier," Sep 2023

2.3 Virtual Reality at the Desk

The concept of using stereo displays or 3D viewing alongside a desktop environment was introduced with Fish Tank VR [70], where a stereo display mounted to a control arm was used instead of a 2D screen to work alongside a desktop environment. The advancements in hardware for HMDs and hand-tracking technology since Fish Tank VR have led to the exploration of mid-air interaction alongside desktop environments in Hybrid Spaces [7] via a formal study. Hybrid Spaces provides compelling evidence that hybrid interaction techniques, including transitions between 2D and 3D and mid-air interactions, outperform both 2D-only and 3D-only interactions. Numerous research efforts [25, 39] and commercial products since then have sought to harness the advantages of both 2D and 3D interactions. For instance, In-Depth Mouse [74] has developed a Depth-Adaptive Cursor capable of enabling a 2D mouse pointer for 3D selection by continuously interpolating the cursor's depth based on user intentions, cursor position, viewpoint, and selectable objects. Vremiere [44] offers an At-Desk VR immersive 360 video editing interface that uses a traditional keyboard and mouse for video editing while providing real-time previews through a VR headset. Biener et al. [6] have demonstrated that users can work for prolonged periods while using VR in a seated desk environment. These hybrid interactions, combining the strengths of immersiveness alongside an at-desk environment, have given rise to what we term 'Immersive At-Desk' experiences.

2.4 Research Scope & Contribution

Past research suggests that hybrid interaction, i.e., combining traditional 2D environments like keyboards with mid-air 3D interactions, mitigates the disadvantages of relying solely on either 3D or 2D interfaces [7]. Consequently, we did not perform a comparative analysis between hybrid spatial user interfaces and traditional methods, relying instead on past evidence [7]. Our system addresses the limitations of previous VR annotation tools by providing an annotation environment in a compact desk setup that enables users to annotate with free-hand interactions. Due to the lack of sufficient research on such hybrid interaction design, there is limited knowledge of design guidance for these systems. Therefore, in this paper, we explored this gap through a comparative study of interactive gestures and provided design recommendations for future 3D mid-air interface designs in at-desk environments.

3 DESIGN MOTIVATION

Dan Olsen [46] outlines various ways to contribute to system development. Inspired by Olsen et al.'s philosophy, we sought to pinpoint potential areas of friction and oriented our system design toward minimizing "solution viscosity." Solution viscosity refers to the resistance encountered in the design process when a tool is cumbersome, making it more difficult to explore a wide range of diverse alternatives efficiently. Our research team distilled essential insights from past literature into a set of design goals (DG) to minimize solution viscosity.

(1) ***Support Natural Interaction in 3D Space with Mid-Air Gestures***. In contemporary office settings, organizations operate within constrained spaces [68]. Additionally, VR has been shown to enhance productivity in conventional

knowledge work settings. Drawing inspiration from established works like Hutchins et al. [31] and the HybridSpace framework [7], we recognize the strength of mid-air gestures alongside traditional interaction devices such as keyboards and mice.

(2) ***Resolve Intent Uncertainty with Physical Delimiters***. While advocating for mid-air interaction, it's important to note that relying solely on gestures might not be sufficient, as emphasized by Handy Potter [66]. Depending solely on gestures can result in unintentional gesture recognition. Therefore, established interaction techniques often incorporate multiple input modalities to distribute task-related functions, as seen in methods like Pen+Touch [30] and Gunslinger [36]. In the context of mode switching, a delimiter is defined as "something different" that a system can use to determine the structure of input phrases [30].

(3) ***Enable Environmental Navigation with Direct Manipulation***. Facilitating navigation in an immersive environment is essential. Ashtari et al. identified that designing VR for user navigation is a current challenge faced by developers, for which at-desk immersive hybrid interaction could be a potential solution, as proposed by Ruminova et al. [55]. Constrained environments like cubicles often contain fragile, stationary objects susceptible to damage from physical collisions (e.g., computer monitors, picture frames [57]). Additionally, interaction designs for supporting immersive, at-desk annotation experiences may involve a range of input devices, including those native to virtual reality (e.g., VR controllers) and non-native ones (e.g., mice and keyboards). Immersive, at-desk annotation experiences should, therefore, prioritize system designs that incorporate a family of mid-air gestures. This approach enables users to interact with the virtual world without concerns about harming nearby physical objects or hindering their ability to use other present input devices.

3.1 Design Approach

We have shaped the design of our system, "Annorama," based on these key design motivations. To facilitate *interactivity with mid-air gestures*, Annorama is designed to accommodate users seated at a desk with a keyboard, enabling mid-air interaction. The preference for free-hand interaction, as opposed to controller-based methods, stems from its advantages, allowing users unrestricted movement of both hands, tracking of individual fingers (enabling sophisticated gestures), and ergonomic access to objects within the workspace (e.g., the keyboard). *Expressing intention with physical delimiters* is supported by our design, where one hand serves as the primary manipulating hand (i.e., for direct manipulation), while the user's other hand serves as the delimiting hand (i.e., which triggers delimiting). We frame the choice of a delimiter as a design decision that can be triggered by an arbitrary event (e.g., a specific keypress on a keyboard or a specific button on a VR controller). As the delimiter for our study, we chose the keyboard's Space Bar because its position is spatially familiar and conducive to a natural resting hand posture. The Space Bar's large size and central location on most standard keyboards make it an ideal choice. Previous research has investigated how visual representations of the keyboard and users' hands assist in 'homing' after performing tasks such as mid-air gestures [24]. Combining this with techniques like Keystroke-Level Modeling [12] can help enhance the selection and design of future delimiters.

3.2 The 3D Point Cloud Diorama

Existing systems for supporting VR-based point cloud annotation have primarily employed first-person perspectives that allow users to immerse themselves in the annotation activity by "walking through" the point cloud data [71]. However, there is compelling evidence that suggests users are more capable of identifying relevant objects in point cloud data when viewed from a third-person perspective, such as an isometric view or a bird's-eye view [42, 56], as a by-product of perceiving the point cloud's sparsity.

To support each of Annorama's design goals described in Section 3, we conceptualized *3D Point Cloud Dioramas*, a representation of point cloud data that is miniaturized in scale and can be directly manipulated. We base our concept on Stoakley et al.'s *World-in-Miniature* (WiM) metaphor [59], which introduced a secondary viewport of a virtual world within the virtual world. Through their demonstration of the metaphor, Stoakley et al. illustrated how miniaturizing the environment can offer advantages for accelerated navigation and natural interaction (e.g., with the fingers) while introducing other challenges related to fatigue and two-handed interaction [59]. Our concept of point cloud dioramas departs from Stoakley et al.'s WiM metaphor by relying on a sole representation of the virtual world that can be directly manipulated rather than two viewports that replicate the same information at varying scales. The *3D Point Cloud Dioramas* also help us address the issue of *navigation in an immersive environment* within constrained spaces. The miniature point cloud diorama rendered within a reachable volume of the user eliminates the requirement for an open area and controller-based navigational interaction techniques.

4 SYSTEM ARCHITECTURE

As a system, Annorama builds on prior VR-based point cloud annotation systems [20, 71] with an architecture that can be deployed to commercial VR devices.

As illustrated in 2, Annorama has four components with their responsibilities being described as follows:

- **Input Manager.** Handles the management of reading and storing annotation and scene data from industry-standard data file formats (i.e., PLY and TFRecord).
- **State Manager.** Handles the management of environmental state in response to user activity, including annotation data and interface telemetry data.
- **Canvas Manager.** Handles parallelization of activities related to rendering with the virtual canvas.
  - **Diorama Layer.** Renders the 3D Point Cloud Diorama representation of a provided point cloud scene
  - **Annotation Layer.** Renders information drawn on the Diorama Layer, such as the user's hands (e.g., as inferred through handtracking SDKs) and cuboid annotations.
- **Interaction Manager.** Handles the management of event handlers for interacting with the virtual canvas.

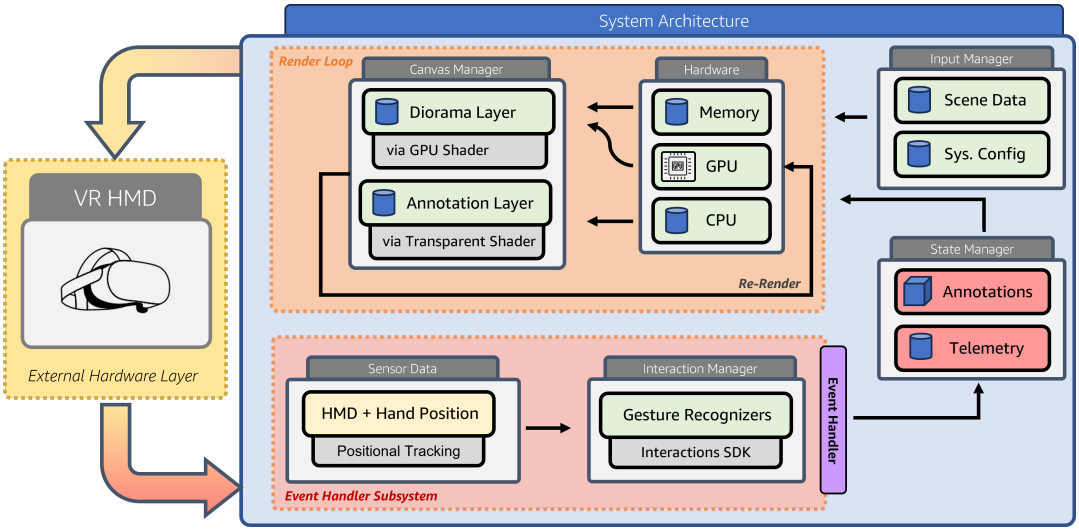


Figure 2: Annorama's architecture operates on the basis of four subcomponents: (1) the *Input Manager*, which facilitates system initialization and configuration; (2) the *State Manager*, which facilitates system-wide state management; (3) the *Canvas Manager*, which manages Annorama's Diorama and Annotation Layers; and (4) the *Interaction Manager*, which manages system-wide event handling with hand-tracking and recognizers.

Unlike prior systems, Annorama's architecture is uniquely inspired by the technical limitations and constraints of commercially available VR headsets and point cloud rendering within VR headsets. By considering these constraints, we not only improve Annorama's ease of use in practical settings but also implement Annorama in a way that takes advantage of the robust tooling provided by device manufacturers (e.g., gesture recognizers). We chose to integrate Annorama with Meta's Oculus Quest 2[8] as it is generally recognized as the most popular consumer device [62]. Furthermore, the Oculus Quest has a family of developer tools and APIs, such as the Insight and Interaction SDK, that ease the burden of crafting complex yet reliable gesture-based and head-pose interactions.

4.1 Creating Annotations with Mid-Air Gestures

Annorama allows users to create 3D cuboid annotations based on four types of mid-air gestures. The design of these gestures was driven by two specific design parameters: inspiration from Deep Extreme Cut [38], where they evaluated 2D bounding box annotation by clicking the extremities of the object of interest.

(1) **Sizing Gesture:** This gesture utilizes the positions of the index fingertip and thumb tip of the manipulative hand. A vectorial distance between these two points is calculated in real-time, serving as the volumetric diagonal for rendering a 3D cuboid.
(2) **One-Point Gesture:** Inspired by common 3D point cloud web interfaces, where users create a box of arbitrary size at the cursor location, this gesture follows a similar principle. In our gesture, we also create a 3D bounding box of arbitrary size, similar to the 2D approach. We leverage the full three-dimensionality. The initiated bounding box's location is set at the local position of the index fingertip, serving as the centroid of the bounding box. Users can later edit the annotation for size, position, and orientation in Edit mode.
(3) **Two-Point Gesture:** This gesture, akin to the sizing gesture, involves a two-step trigger process. First, a trigger sequence is used to obtain the initial position information, and then a second sequence captures the end position. The vectorial distance between these two positions is computed, resulting in the rendering of a cuboid annotation.
(4) **Three-Point Gesture:** Drawing inspiration from 3D cuboid creation features in computer-aided design tools like Autodesk Fusion 360 or SolidWorks, this gesture requires three triggers. Each trigger sets one of the dimensions of the cuboid, totaling three triggers for completion.

All generated bounding boxes are axis-aligned bounding boxes (AABBs). This choice was made because most real-world objects requiring annotation lie on the same plane (the ground plane due to gravity). Thus, an AABB with the option to reorient around that plane is sufficient for the task. However, for objects in unique orientations, like flying objects (e.g., birds or banking planes) or objects oriented in 3D space (e.g., cranes), the current interface may pose challenges. Nonetheless, the described edit mode can be simply extended to support non-AABB objects to address such cases effectively.

4.1.1 Modifying Cuboid Annotations with Mid-Air Gestures. The edit mode in Annorama is designed to facilitate adjustments to the position, scale, and orientation of initially created box annotations. To access the edit mode, users can simply touch their index finger to the annotation box of interest and then press the Alt key; this action triggers the appearance of edit widgets at the annotation's location. The edit widget offers three main functions:

[8]Oculus, "Oculus Quest 2," 2020, Retrieved April 4, 2021.

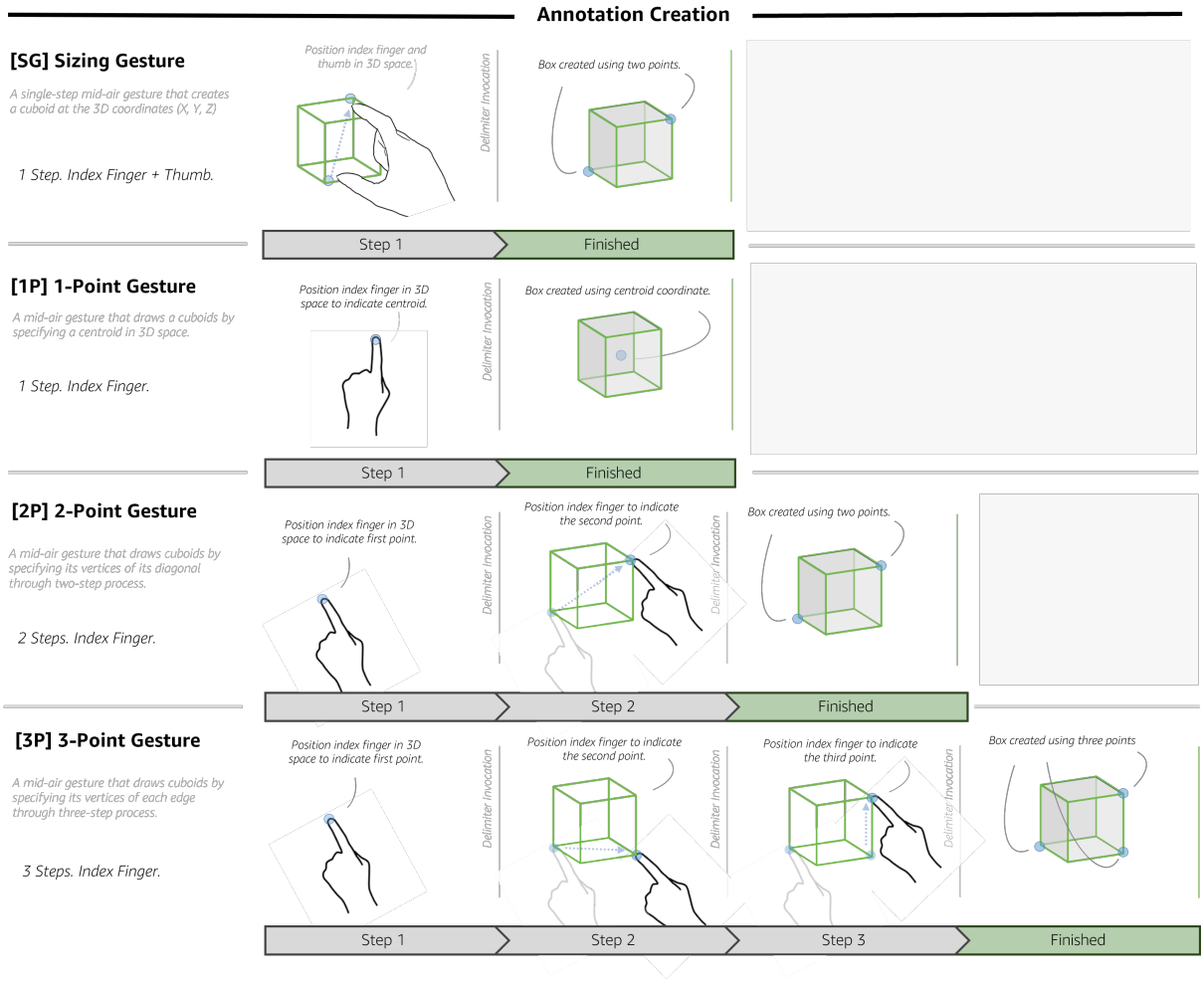


Figure 3: Annorama's four gestures for facilitating annotation creation.

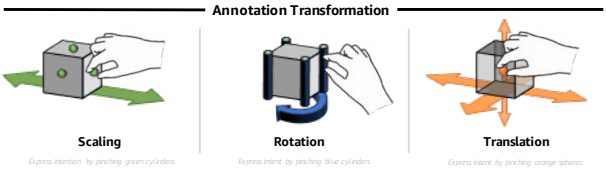


Figure 4: Annorama's 3D Edit Widgest that allows users to scale, rotate, and translate annotations annotations with pinch gestures.

(1) **Scaling:** Users can adjust the edge length (scale) of the annotation by pinching and pulling or pushing one of the green spheres attached to each face of the annotation.
(2) **Reorientation:** To change the orientation of the box, users can utilize any of the blue cylinders attached to the edges of the annotations.
(3) **Repositioning:** For relocating the annotation, users can use the orange sphere positioned at the centroid of the cuboid.

All annotation widgets automatically readjust themselves after initial interactions. To exit the edit mode, users can simply press the Alt key again. The constraints for each of the widgets were programmed with inspiration from past relevant work [40], applying constraints to two virtual models presented in planes, rays, and points [28]. Control over the rotation and scale of the diorama is achieved through two distinct widgets: the zoom sphere and the rotate sphere, both positioned on the same layer as the 3D miniature point cloud dioramas. To delete a created annotation, users can simply interact (collide) their index fingertip with the desired annotation and press both the Alt and Space keys simultaneously.

5 USER STUDY

We conducted a "first-use" study to evaluate Annorama [27]. The goal of this study was twofold: 1) to demonstrate Annorama's ability to enable immersive, at-desk point cloud annotation, and 2) to understand how each of Annorama's four hands-free gestures for creating cuboid annotations affects annotation quality and efficiency.

To support these goals, we employed a within-subjects study design in which each participant was exposed to each creation gesture in a non-uniform order. Conditions were counterbalanced to prevent ordering effects that might otherwise influence participant experiences.

5.1 Task

To facilitate our study, we designed a 3D object detection To facilitate our study, we designed a 3D object detection task using the Waymo Open Dataset [60], an open-source dataset of point cloud scenes with ground-truth cuboid annotations commonly used to support machine learning and computer vision research. To evaluate Annorama's ability to facilitate point cloud annotation for objects that vary in volumetric shape and size, we designed our task so that three objects of Level-1 difficulty in the Waymo dataset—Pedestrian, Vehicle, and Street Signs—would be annotated and classified.

We sampled a total of 20 scenes (i.e., video frames) from a total of 2 segments (i.e., videos). One segment was dedicated to training activities, while the other was used exclusively for the "real" task. Segments were sampled on the basis that the total number of relevant objects be equal to 8 to provide enough observational data while simultaneously limiting the length of each condition in our study.

5.2 Procedure

The study began with participants seated at the workstation. They received a study information letter, read the consent letter, and completed the Pre-Study Survey. Next, they equipped the VR headset and spent 10 minutes in a training environment with a pre-annotated point cloud scene to learn how to use Annorama. Participants then used versions of Annorama, each with a single enabled annotation creation gesture, for up to 10 minutes each to annotate and classify objects in a point cloud scene. After using each system version, they completed the Post-Condition Survey. The study concluded with participants completing the Post-Study Survey.

5.3 Participants

Sixteen participants (4 female, 12 male) were recruited for study participation via a large technology corporation's internal instant messaging application. Participants' job roles included software engineer, software engineer intern, applied scientist, and project manager. Four participants were moderately familiar with point cloud annotation, while only one had similar familiarity with VR headset devices. The majority (9) had only slight familiarity with VR, and none claimed high expertise. Regarding point cloud annotation, five participants were somewhat familiar, two had slight familiarity, and five had no prior experience.

5.4 Apparatus

All participants used an Oculus Quest 2 VR headset to complete the study. The headset was connected to a laptop computer equipped with a 2.7 GHz Intel Core i7 12th Generation processor, 16 GB RAM, and an Nvidia GeForce RTX 3070 Ti graphics card via a USB-C 'link' cable. The system was developed in Unity 2021.3.15f1 [9] game engine. Participants were compensated with a $25 Amazon gift card after completing the study.

5.5 Data Collection

In support of our study, we collect the following types of data:

(1) **Survey Data.** We administered three unique surveys to each user at different stages of our study, with one survey repeated four times:
   (a) ***Pre-Study Survey:*** Collected information about users' familiarity with point cloud data annotation and VR.
   (b) ***Post-Condition Survey:*** Collected feedback on liked features, disliked features, and overall system usability.
   (c) ***Post-Study Survey:*** Collected preference and ranking information among the studied techniques. This survey also asked users to write short sentences describing aspects of interaction techniques that were liked or disliked.

(2) **Cuboid Annotations Data.** By design, Annorama collects cuboid annotation data when used to annotate point cloud scene data. We collected and stored centroid positions and box sizes following the guidelines outlined by Ravi et al. [50]. We measured the similarity of ground-truth cuboid annotation coordinates and participants' cuboid annotation coordinates by calculating Intersection over Union (IoU) both per class and in aggregate across all classes [35, 60].

(3) **Interface Telemetry Data.** Inspired by Rechkemmer et al. [51], we analyzed timestamped events logged by the Annorama system (e.g., mode switches, annotation creation start, annotation creation stop, etc.). Total Annotation Time is defined as the sum of Annotation Creation Time and Annotation Edit Time.

The complete list of metrics is shown in Table 1.

5.6 Methods of Analysis

We employ a combination of quantitative and qualitative methods to analyze the data shown in Table 1. First, we evaluated the normality of data distributions with Kolmogorov-Smirnov tests and performed one-way analysis of variance (ANOVA) tests to examine whether statistically significant differences exist between groups (e.g., between conditions). We determined which groups had statistically significant differences using Tukey-Kramer tests. Lastly, we analyzed all qualitative data collected in each of the administered surveys using Brau et al.'s method of Thematic Analysis [8]. In our case, we analyzed our qualitative data collectively to identify themes that illustrate the system's strengths and shortcomings across participants' usage.

6 RESULTS

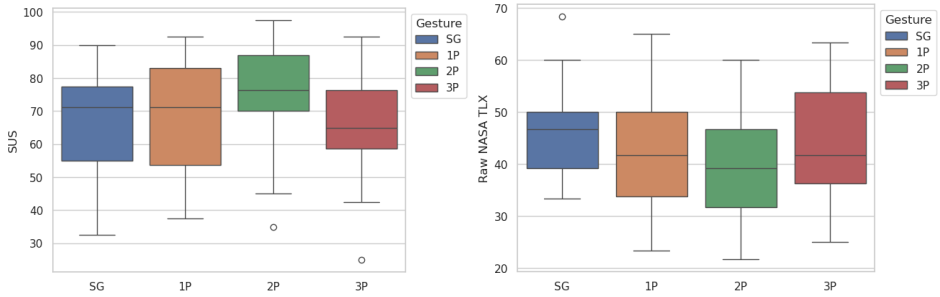
6.1 Supporting VR-Based Labeling at the Desk

Annorama's approach to enabling VR-based point cloud annotation at the desk was generally well-received by all 16 participants. For example, P1 described Annorama's approach to annotation as being *"way better than drawing them using a mouse and faster as well"*. As shown in Figure 5a, SUS scores across each gesture were

[9] Unity, "Unity 2021.3.15," Dec 1, 2022.

Table 1: An overview of all the different metrics computed for evaluating Annorama from collected data.

Metric Theme	Metric Name	Metric Description
Time of Task	Annotation Creation Time	Time taken to create a new annotation
	Annotation Edit Time	Time taken to edit a previously created annotation
Quality of annotation	mean Intersection Over Union (mIoU)	3D mIoU [50] of user annotation over ground truth from [60]
Usability	System Usability Scale: For Gestures	Usability Metric of System for gestures [9]
	System Usability Scale: VR Environment	Usability Metric of System for VR environment [9]
	NASA-TLX	Method to assesses cognitive load [26]
	Preference Ranking: Speed	Rank order of gesture preference based on Speed
	Preference Ranking: Accuracy	Rank order of gesture preference based on Accuracy
	Post-Condition Survey	Comments from user collected after each condition
	Post-Study Survey	Comments from user collected at the end of the study



(a) System Usability Scale (SUS) scores. **(b) NASA TLX scores.**

Figure 5: Boxplots for System Usability Scale (SUS) and NASA TLX scores across all 16 participants.

positive (M=76.56, SD=17.19), indicating that participants perceive the system to be highly usable. Similar observations were made for participants' NASA TLX scores in Figure 5b, which were also consistent among Annorama's four gestures (M=43.18, SD=10.55).

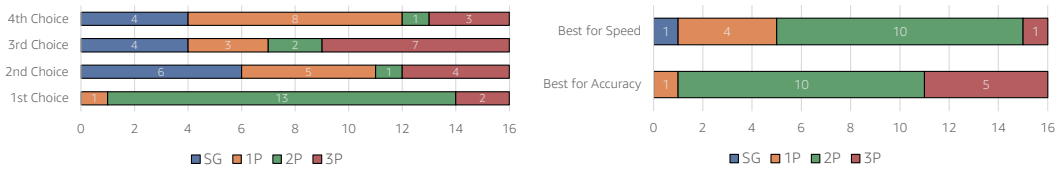
6.2 Perceptions and Preferences of Mid-Air Gestures

Participants' attitudes varied significantly across each of Annorama's family of point-based gestures. As shown in Figure 6a, the 2P gesture was widely appreciated by most users, with 13 users rating the 2P gesture as their most preferred option. In contrast, the 1P gesture was the least preferred, with 11 out of 16 users rating it as either their third or last preferred option. The 2P gesture also rated highly for questions regarding speed and accuracy, with 10 out of 16 users perceiving it to be the fastest and most accurate. The Sizing Gesture and 3P gestures occupied intermediate rankings in preferences, as shown in Figure 6a.

Through our analysis of user comments across all surveys, we found that the 1P gesture was among the most negatively received due to its inherent connection to unavoidable editing. Despite finding it substantially fast for creating annotations, 12 participants reported significant challenges during the editing process. As shown in Figure 6a, the 1P gesture was reported by 11 participants (69%) as the least preferred, finding it less favorable. In the preference ranking, 8 users rated the 1P gesture as their least preferred choice. Users reported that while creating a 3D annotation box using a single point was quick, the subsequent editing process consumed more time. For example, P3 mentioned, "*The gesture itself was simple, but the editing afterward was a hassle*," and P2 highlighted, "*Compared to other gestures, I would say this gesture requires more post-adjustments.*"

In contrast to the incremental, point-based gestures, the Sizing Gesture was generally perceived as intuitive and easy to *learn*, as reported by 8 users. P6 mentioned, "*It's easy to learn for beginners and easy to use.*" However, some participants noted room for improvement, particularly in controlling the orientation of the annotation box during its creation. P5 pointed out, "*The sizing was pretty cool to control but could use more refinement to control the orientation while creating the box instead of afterward with editing.*" Mirroring the post-creation experience with the 1P gesture, several participants remarked that the Sizing Gesture's usability was partially limited due to the cognitive overhead of simultaneously managing two points in three-dimensional space: "*I like this gesture. Though with two fingers, we can't control the orientation of the box (or make it rotate with my hand), so we need to rotate the point cloud constantly to avoid post-adjustment.*" (P8). Participants' remarks were often accompanied by a disclaimer noting that the need to edit was nullified if the initial cuboid was created in a satisfactory fashion.

Taken collectively, our data suggest that participants prefer mid-air gestures that provide a larger degree of control in a point-by-point fashion. For example, in describing their support for the 2P gesture, P13 emphasized, "*This gives us more control over drawing the boxes. Using 2 points is very useful for drawing boxes.*" Similarly, P15 stated, "*The gesture is useful for creating the right box because I can only focus on one point at a time.*" However, this affinity for greater control did not extend to the 3P gesture, which, while allowing for a higher number of individual point placements, was appreciated by some users. However, questions arose regarding the necessity of the additional click, given that the 2P gesture already delivered



(a) Ranked choices across the four gestures. **(b) Gesture preferences for speed and accuracy.**

Figure 6: Rankings and preferences for Annorama's four mid-air gestures.

highly accurate results. Some users found controlling the Z-axis with the 3P gesture to be challenging and largely unnecessary. They mentioned that the placement of an additional control point for 3P slowed them down compared to the 2P gesture. P4 stated, "*The third point slowed it down*," and P7 noted, "*The third point is unnecessary to ensure accuracy.*"

6.3 Supporting Rapid Annotation

Through our analysis, we observed significant variations in the duration logged for performing specific gestures. For example, when using the Sizing gesture, participants' duration for creating annotations averaged 9.23 seconds, while the duration was marginally longer when using the 2P gesture at 15.36 seconds. In contrast, the 3P and 1P gestures exhibited relatively longer annotation times, with averages of 21.23 and 22.77 seconds, respectively. This range underscores the influence of each gesture on the temporal aspects of annotation creation.

Moving beyond the duration of annotation creation events, our investigation extends to the time users spent editing annotations. A one-way ANOVA test yielded a statistically significant difference among the four gestures ($p<0.05$, $F=244.16$). Post-hoc Tukey-Kramer tests further elucidated that while there was no significant difference in annotation time between the 1P and 3P gestures, there were significant differences between the Sizing and 2P gestures. This suggests that the efficiency observed in annotation creation is not mirrored in the editing phase, emphasizing the need to carefully design gestures for user interaction. A gesture that provides users with a more precise annotation, without the need for additional gestural interaction, is preferred over a gesture that enables users to create annotations quickly but necessitates editing afterward.

Based on our analysis, we find that annotation time is shortest when participants use the Sizing gesture to create cuboid annotations, followed closely by the 2P gesture.

Table 2: mIoU scores across the three object classes with standard deviation reported in parentheses.

Object Type	SG	1P	2P	3P
Vehicle	0.93 (0.04)	0.83 (0.05)	0.92 (0.05)	0.87 (0.05)
Pedestrian	0.87 (0.1)	0.75 (0.1)	0.84 (0.1)	0.79 (0.1)
Sign	0.69 (0.2)	0.62 (0.2)	0.64 (0.2)	0.63 (0.2)
All	0.83 (0.1)	0.73 (0.1)	0.80 (0.1)	0.76 (0.1)

6.4 Supporting Precise Annotation

We find that Annorama enabled users to create high-quality cuboid annotations with high precision across all three types of objects used in our study. Using ground truth cuboid annotations from the Waymo dataset, we find that average 3D IoU (mIoU) scores across the three objects types to be generally satisfactory. As shown in Table 2, vehicles were most frequently annotated at the highest level of accuracy (M=0.9, SD=0.04), followed by Pedestrian (M=0.87, SD=0.1) and Sign (M=0.69, SD=0.2) respectively.

As with annotation time, we observe that Annorama's gestures impacted the quality of annotations that were created by participants. Using a one-way ANOVA, we observe significant differences in 3D mIoU scores among Annorama's four gestures ($p< 0.0001$, $F = 15.9$, $\eta^2=0.086$). Mirroring our findings related to annotation time, a post-hoc Tukey-Kramer test revealed significant differences in 3D mIoU scores exist between the Sizing Gesture (SG) and both the 1P and 3P gestures as well as between the 2P gesture and both the 1P and 3P gestures. However, no significant difference was observed between SG and 2P, suggesting that SG and 2P gestures perform comparably.

To deepen our analysis of differences across Annorama's gestures, we examined the possibility of a correlation existing between annotation time for each object category (Total Annotation Time) and the accuracy of user annotations (mIoU). As shown in Figure 7, we observe significant visual distinctions in the time ranges that exist across each of Annorama's gestures. Each gesture is distinctly grouped, underscoring significant differences in annotation choices. Across all three object classes, the Sizing Gesture exhibits a notable advantage in terms of both time efficiency and accuracy, followed closely by the 2P gesture. Conversely, there is no discernible difference between the 1P and 3P gestures, affirming the findings of the earlier performed ANOVA analysis.

A positive trend is evident across all gestures for objects of varying sizes, suggesting that as annotation accuracy increases, more time is required for annotation. This phenomenon can be rationalized by the user's endeavor to enhance the precision and accuracy of the annotation bounding box, necessitating additional adjustments and focused attention for greater precision during annotation. The trend is particularly pronounced in the case of the Sizing Gesture, as indicated by the steeper slope of its trend line in the scatter plot. Consequently, a conclusion can be drawn regarding the relationship between different gestures and accuracy across various object sizes. This analysis provides compelling evidence that the choice of gesture design significantly impacts the development of such interactive systems. Among the designed gestures, the Sizing

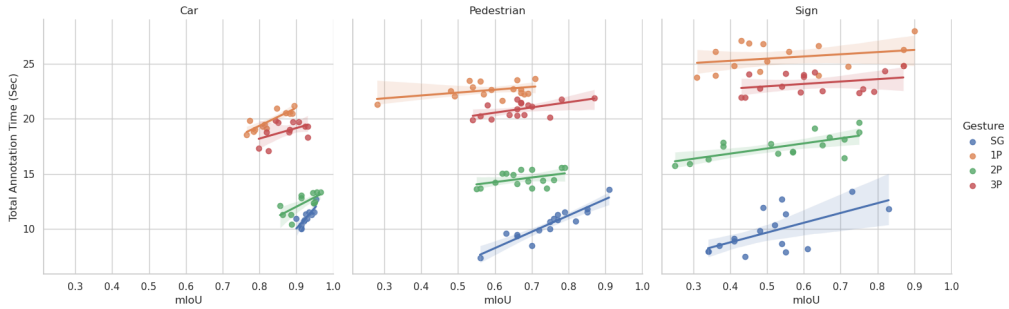


Figure 7: Relationships between annotation time (Total Annotation Time) and annotation quality (mIoU) across Annorama's four mid-air gestures. Best-fit lines with 95% confidence shadings are shown for each gesture.

Gesture stands out by statistically improving both Total Annotation time and label accuracy, as measured quantitatively. However, considering user preferences from earlier qualitative metrics, which favored the 2P gesture for ease of use, further research in this area would be valuable for making conclusive design decisions.

7 DISCUSSION

Recognizing the intuitiveness of interacting in 3D space with natural gestures, we address associated limitations by introducing the use of a miniature diorama. Our approach mitigates challenges inherent to the 3D point cloud annotation task, such as navigation within constrained spaces and the need for a 1:1 environment for FPOV control. All participants completed the annotation of all objects with all the different gestures. The results demonstrate that a hybrid interface system like Annorama enables immersive 3D point cloud annotation with mid-air gestures at one's desk.

Notably, the time taken to complete the annotation tasks indicates that users were able to annotate more efficiently compared to values reported for controller-based VR annotation [71], with 55 sec/annotation, and desktop-based annotation [41], with 92 sec/annotation. This reported improvement in the throughput of mid-air interaction contrasts with a mid-air pointing task study conducted by Brown et al. [10] and Cockburn et al. [17]. These past works explore mid-air pointing tasks on a 2D screen, exploring various conditions with the space bar as a delimiter triggered via the non-dominant hand. Brown's findings revealed that using a mouse for pointing and selecting virtual assets on a 2D screen resulted in significantly higher throughput, but movement speed was more pronounced in in-air interaction. However, as our focus is on interacting with 3D virtual assets directly, Brown's study outcomes do not precisely align with our context. Notably, Brown et al. themselves advocate for in-air interaction in scenarios involving "*bi-manual interaction, gesture-based systems, and applications where the user cannot touch or hold a device*," all of which align with the characteristics of our system.

We attribute this reduction in average annotation time to the unique way of annotating point cloud data in VR at the desk utilizing a miniature 3D point cloud diorama environment. The ease with which users can navigate through these scaled-down 3D spaces, as opposed to a 1:1 scale, facilitated swift transitions between annotation points, resulting in rapid and efficient annotation processes.

To gain a better understanding of the process behind annotating 3D point cloud labels, we carefully studied the annotation creation and annotation edits separately. Gestures demanding precise input for adjusting annotation size and orientation, specifically the one-point gestures that necessitate edit mode usage, were less favored by our users, as shown in Figure 6b. Qualitative questionnaire responses further confirm the substantial cognitive load associated with annotation editing and adjustments. This issue can be attributed to the inherent jitter experienced by human hands due to physiological factors. Existing literature [1, 4, 5] offers substantial documentation on the effects of jitter in natural hand interactions. In contrast, gestures offering higher fidelity for crafting accurately sized and oriented annotations, such as the two-point gesture, were highly preferred due to their efficiency.

Remarkably, the Edit mode garnered limited favor among users, primarily due to the constraints posed by the popular VR pinch gesture, which couldn't provide the requisite level of precision when interacting with 3D widgets. This limitation is tied to how the pinch gesture is recognized and supported by the system's hardware. In the Oculus VR environment, there is a noticeable lag between the recognition of the *"pinch state"* and its transition to the *"unpinch state"*. The pinch doesn't release immediately, causing the widget to interact with the fingertips a few milliseconds after the release action. This inherent lag led to user frustration when using the edit mode and resulted in longer edit times.

7.1 Design Recommendations

Based on our study results, we present the following recommendations for creating a hybrid mid-air 3D UI design for an at-desk immersive annotation experience:

(1) **Use the Sizing Gesture:** For actions requiring volumetric adjustment, such as precision in position and orientation, use the Sizing Gesture.
(2) **Use the Two-Point Gesture:** For tasks requiring a higher degree of control and user comfort during interaction, the Two-Point Gesture is recommended.
(3) **Utilize Dioramas:** Dioramas enable faster 3D interactions due to the miniature nature of the environmental representation.

(4) **Avoid/Reduce Small Object Representations:** Minimize the number of small objects within the diorama, as smaller objects result in longer and more imprecise interactions.
(5) **Avoid Pinch-Based Gestures:** Avoid using pinch-based gestures for delimiting actions. Instead, rely on external actions such as button presses to overcome current hardware limitations with distinct pinch recognition.

8 LIMITATION AND FUTURE WORK

We only tested Annorama on static point cloud data, which may not fully represent the dynamic nature of real-world scenes. We did not evaluate the system's performance on sequential point cloud data, such as streams of LiDAR scans, which could introduce additional challenges. However, our current rendering pipeline is capable of loading sequential point cloud data, making this a promising direction for future research.

To address issues of imprecision due to jitter, two potential approaches can be considered. First, implementing a clustering algorithm for point cloud data could improve precision [67]. Exploring the integration of a snapping function within Annorama's workflow could enhance the efficiency of 3D cuboid bounding box annotations, as suggested by participants who expressed a desire for an auto-snapping feature [29, 45]. Additionally, VR headset disorientation and physical fatigue were identified as challenges, with participants noting the physical strain of prolonged arm elevation. Second, research suggests that mouse interactions are better suited for tasks requiring precision and finer adjustments, making them ideal for refining 3D annotations. Future research could explore a system combining VR-based annotation creation with mouse-based edits and adjustments.

Annorama does not anchor the miniature diorama to any physical object in space, such as a desktop or keyboard. Instead, the dioramas are positioned within the 3D space for user convenience. Conducting an empirical study to evaluate various reach volumes and anchoring strategies could provide valuable insights for developing design guidelines. Additionally, offering multiple perspectives for viewing annotations [15, 16] and enabling automatic movement to the next annotation during editing interactions represent promising avenues for future research.

9 CONCLUSION

Manual annotation of 3D point clouds with traditional 2D GUI interfaces is time-consuming, labor-intensive, and expensive. To address these challenges, we introduced Annorama, a 3D point cloud annotation system that combines the strengths of VR with the convenience of a desk. By using a VR headset and natural gestural interactions, users can efficiently annotate 3D point clouds while seated. Annorama utilizes miniature 3D point cloud dioramas and one-handed mid-air gestures, supplemented by an editing widget for precision adjustments. Our within-subjects study with 16 participants assessed the effectiveness of these gestures, focusing on annotation quality, efficiency, and user comfort. For tasks such as 3D point cloud annotation, gestures like Sizing and Two-Point gestures are particularly useful.

ACKNOWLEDGMENTS

We express our gratitude to the Amazon Science Postdoctoral Science Program for its invaluable support. Our sincere appreciation goes to the study participants for generously contributing their time and providing valuable feedback during the evaluation of the system. Special acknowledgment is extended to Koushik Kalyanaraman for his insightful feedback and suggestions. We would also like to extend our thanks to Ammar Chinoy and Kumar Chellapilla for their support of this project. Additionally, we acknowledge the members and staff within the AWS Human-in-the-Loop (HIL) organization for their constructive feedback and suggestions.

REFERENCES

[1] Wei Tech Ang. 2004. *Active tremor compensation in handheld instrument for microsurgery.* Ph. D. Dissertation. Carnegie Mellon University, the Robotics Institute.
[2] Felipe Bacim, Regis Kopper, and Doug A. Bowman. 2013. Design and evaluation of 3D selection techniques based on progressive refinement. *International Journal of Human-Computer Studies* 71, 7 (2013), 785–802. https://doi.org/10.1016/j.ijhcs.2013.03.003
[3] Felipe Bacim, Mahdi Nabiyouni, and Doug A. Bowman. 2014. Slice-n-Swipe: A free-hand gesture user interface for 3D point cloud annotation. In *2014 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, Minneapolis, Minnesota, USA, 185–186. https://doi.org/10.1109/3DUI.2014.6798882
[4] Anil Ufuk Batmaz, Mohammad Rajabi Seraji, Johanna Kneifel, and Wolfgang Stuerzlinger. 2020. No jitter please: Effects of rotational and positional jitter on 3D mid-air interaction. In *Proceedings of the Future Technologies Conference*. Springer International Publishing, Cham, 792–808.
[5] Anil Ufuk Batmaz and Wolfgang Stuerzlinger. 2019. The Effect of Rotational Jitter on 3D Pointing Tasks. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312752
[6] Verena Biener, Snehanjali Kalamkar, Negar Nouri, Eyal Ofek, Michel Pahud, John J Dudley, Jinghui Hu, Per Ola Kristensson, Maheshya Weerasinghe, Klen Čopič Pucihar, et al. 2022. Quantifying the effects of working in VR for one week. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (2022), 3810–3820.
[7] Natalia Bogdan, Tovi Grossman, and George Fitzmaurice. 2014. HybridSpace: Integrating 3D freehand input and stereo viewing into traditional desktop applications. In *2014 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, Singapore, 51–58. https://doi.org/10.1109/3DUI.2014.6798842
[8] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.
[9] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
[10] Michelle A Brown, Wolfgang Stuerzlinger, and EJ Mendonça Filho. 2014. The performance of un-instrumented in-air pointing. In *Graphics Interface 2014*. AK Peters/CRC Press, 59–66.
[11] Robin Burgess, António J. Falcão, Tiago Fernandes, Rita A. Ribeiro, Miguel Gomes, Alberto Krone-Martins, and André Moitinho de Almeida. 2015. Selection of Large-Scale 3D Point Cloud Data Using Gesture Recognition. In *Technological Innovation for Cloud-Based Engineering Systems*, Luis M. Camarinha-Matos, Thais A. Baldissera, Giovanni Di Orio, and Francisco Marques (Eds.). Springer International Publishing, Cham, 188–195.
[12] Stuart K Card, Thomas P Moran, and Allen Newell. 1980. The keystroke-level model for user performance time with interactive systems. *Commun. ACM* 23, 7 (1980), 396–410.
[13] Han Chen and Peng Lu. 2022. Real-time identification and avoidance of simultaneous static and dynamic obstacles on point cloud for UAVs navigation. *Robotics and Autonomous Systems* 154 (2022), 104124. https://doi.org/10.1016/j.robot.2022.104124
[14] Subramanian Chidambaram, Hank Huang, Fengming He, Xun Qian, Ana M Villanueva, Thomas S Redick, Wolfgang Stuerzlinger, and Karthik Ramani. 2021. ProcessAR: An Augmented Reality-Based Tool to Create in-Situ Procedural 2D/3D AR Instructions. In *Designing Interactive Systems Conference 2021* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 234–249. https://doi.org/10.1145/3461778.3462126
[15] Subramanian Chidambaram, Rahul Jain, Sai Reddy, Asim Unmesh, and Karthik Ramani. 2024. AnnotateXR: An Extended Reality Workflow for Automating Data Annotation to Support Computer Vision Applications. *Journal of Computing and Information Science in Engineering* (08 2024), 1–13. https://doi.org/10.1115/1.4066180

[16] Subramanian Chidambaram, Sai Swarup Reddy, Matthew Rumple, Ananya Ipsita, Ana Villanueva, Thomas Redick, Wolfgang Stuerzlinger, and Karthik Ramani. 2022. EditAR: A Digital Twin Authoring Environment for Creation of AR/VR and Video Instructions from a Single Demonstration. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, Singapore, 326–335. https://doi.org/10.1109/ISMAR55827.2022.00048

[17] A. Cockburn, P. Quinn, C. Gutwin, G. Ramos, and J. Looser. 2011. Air pointing: Design and evaluation of spatial target acquisition with and without visual feedback. *International Journal of Human-Computer Studies* 69, 6 (2011), 401–414. https://doi.org/10.1016/j.ijhcs.2011.02.005

[18] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New York, NY, USA, 5828–5839.

[19] Achref Doula, Tobias Güdelhöfer, Andrii Matviienko, Max Mühlhäuser, and Alejandro Sanchez Guinea. 2022. Immersive-Labeler: Immersive Annotation of Large-Scale 3D Point Clouds in Virtual Reality. In *ACM SIGGRAPH 2022 Posters* (Vancouver, BC, Canada) *(SIGGRAPH '22)*. Association for Computing Machinery, New York, NY, USA, Article 27, 2 pages. https://doi.org/10.1145/3532719.3543249

[20] Achref Doula, Tobias Güdelhöfer, Andrii Matviienko, Max Mühlhäuser, and Alejandro Sanchez Guinea. 2023. PointCloudLab: An Environment for 3D Point Cloud Annotation with Adapted Visual Aids and Levels of Immersion. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 11640–11646. https://doi.org/10.1109/ICRA48891.2023.10160225

[21] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. AAAI Press, Portland, Oregon, 226–231.

[22] Anton Franzluebbers, Changying Li, Andrew Paterson, and Kyle Johnsen. 2022. Virtual Reality Point Cloud Annotation. In *Proceedings of the 2022 ACM Symposium on Spatial User Interaction* (Online, CA, USA) *(SUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 14, 11 pages. https://doi.org/10.1145/3565970.3567696

[23] Daniel Garrido, Rui Rodrigues, A. Augusto Sousa, Joao Jacob, and Daniel Castro Silva. 2021. Point Cloud Interaction and Manipulation in Virtual Reality. In *2021 5th International Conference on Artificial Intelligence and Virtual Reality (AIVR)* (Kumamoto, Japan) *(AIVR 2021)*. Association for Computing Machinery, New York, NY, USA, 15–20. https://doi.org/10.1145/3480433.3480437

[24] Alexander Giovannelli, Lee Lisle, and Doug A. Bowman. 2022. Exploring the Impact of Visual Information on Intermittent Typing in Virtual Reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 8–17. https://doi.org/10.1109/ISMAR55827.2022.00014

[25] Jens Grubert, Eyal Ofek, Michel Pahud, and Per Ola Kristensson. 2018. The Office of the Future: Virtual, Portable, and Global. *IEEE Computer Graphics and Applications* 38, 6 (2018), 125–133. https://doi.org/10.1109/MCG.2018.2875609

[26] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, USA, 139–183.

[27] Björn Hartmann, Leith Abdulla, Manas Mittal, and Scott R. Klemmer. 2007. Authoring sensor-based interactions by demonstration with direct manipulation and pattern recognition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 145–154. https://doi.org/10.1145/1240624.1240646

[28] Devamardeep Hayatpur, Seongkook Heo, Haijun Xia, Wolfgang Stuerzlinger, and Daniel Wigdor. 2019. Plane, ray, and point: Enabling precise spatial manipulations with shape constraints. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 1185–1195.

[29] Devamardeep Hayatpur, Seongkook Heo, Haijun Xia, Wolfgang Stuerzlinger, and Daniel Wigdor. 2019. Plane, Ray, and Point: Enabling Precise Spatial Manipulations with Shape Constraints. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 1185–1195. https://doi.org/10.1145/3332165.3347916

[30] Ken Hinckley, Koji Yatani, Michel Pahud, Nicole Coddington, Jenny Rodenhouse, Andy Wilson, Hrvoje Benko, and Bill Buxton. 2010. Pen + Touch = New Tools. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology* (New York, New York, USA) *(UIST '10)*. Association for Computing Machinery, New York, NY, USA, 27–36. https://doi.org/10.1145/1866029.1866036

[31] Edwin L Hutchins, James D Hollan, and Donald A Norman. 1985. Direct manipulation interfaces. *Human–computer interaction* 1, 4 (1985), 311–338.

[32] Ananya Ipsita, Runlin Duan, Hao Li, Subramanian C, Yuanzhi Cao, Min Liu, Alexander J Quinn, and Karthik Ramani. 2023. The Design of a Virtual Prototyping System for Authoring Interactive VR Environments from Real World Scans. *Journal of Computing and Information Science in Engineering* (07 2023), 1–18. https://doi.org/10.1115/1.4062970

[33] Ananya Ipsita, Hao Li, Runlin Duan, Yuanzhi Cao, Subramanian Chidambaram, Min Liu, and Karthik Ramani. 2021. VRFromX: From Scanned Reality to Interactive Virtual Experience with Human-in-the-Loop. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 289, 7 pages. https://doi.org/10.1145/3411763.3451747

[34] Katja Krug, Marc Satkowski, Reuben Docea, Tzu-Yu Ku, and Raimund Dachselt. 2023. Point Cloud Alignment through Mid-Air Gestures on a Stereoscopic Display. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 230, 7 pages. https://doi.org/10.1145/3544549.3585862

[35] Loic Landrieu and Mohamed Boussaha. 2019. Point cloud oversegmentation with graph-structured deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7440–7449.

[36] Mingyu Liu, Mathieu Nancel, and Daniel Vogel. 2015. Gunslinger: Subtle Arms-down Mid-Air Interaction. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 63–71. https://doi.org/10.1145/2807442.2807489

[37] Paul Lubos, Rüdiger Beimler, Markus Lammers, and Frank Steinicke. 2014. Touching the Cloud: Bimanual annotation of immersive point clouds. In *2014 IEEE Symposium on 3D User Interfaces (3DUI)*. 191–192. https://doi.org/10.1109/3DUI.2014.6798885

[38] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. 2018. Deep Extreme Cut: From Extreme Points to Object Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New York, NY, USA.

[39] Mark Mcgill, Aidan Kehoe, Euan Freeman, and Stephen Brewster. 2020. Expanding the Bounds of Seated Virtual Workspaces. *ACM Trans. Comput.-Hum. Interact.* 27, 3, Article 13 (may 2020), 40 pages. https://doi.org/10.1145/3380959

[40] Daniel Mendes, Fabio Marco Caputo, Andrea Giachetti, Alfredo Ferreira, and Joaquim Jorge. 2019. A survey on 3d virtual object manipulation: From the desktop to immersive virtual environments. In *Computer graphics forum*, Vol. 38. Wiley Online Library, 21–45.

[41] Riccardo Monica, Jacopo Aleotti, Michael Zillich, and Markus Vincze. 2017. Multi-label point cloud annotation by selection of sparse control points. In *2017 International Conference on 3D Vision (3DV)*. IEEE, 301–308.

[42] Mong H Ng, Kaahan Radia, Jianfei Chen, Dequan Wang, Ionel Gog, and Joseph E Gonzalez. 2020. Bev-seg: Bird's eye view semantic segmentation using geometry and semantic point cloud. *arXiv preprint arXiv:2006.11436* (2020).

[43] Anh Nguyen and Bac Le. 2013. 3D point cloud segmentation: A survey. In *2013 6th IEEE Conference on Robotics, Automation and Mechatronics (RAM)*. 225–230. https://doi.org/10.1109/RAM.2013.6758588

[44] Cuong Nguyen, Stephen DiVerdi, Aaron Hertzmann, and Feng Liu. 2017. Vremiere: In-Headset Virtual Reality Video Editing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 5428–5438. https://doi.org/10.1145/3025453.3025675

[45] Benjamin Nuernberger, Eyal Ofek, Hrvoje Benko, and Andrew D. Wilson. 2016. SnapToReality: Aligning Augmented Reality to the Real World. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1233–1244. https://doi.org/10.1145/2858036.2858250

[46] Dan R. Olsen. 2007. Evaluating User Interface Systems Research. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology* (Newport, Rhode Island, USA) *(UIST '07)*. Association for Computing Machinery, New York, NY, USA, 251–258. https://doi.org/10.1145/1294211.1294256

[47] Leonardo Pavanatto, Shakiba Davari, Carmen Badea, Richard Stoakley, and Doug A Bowman. 2023. Virtual monitors vs. physical monitors: an empirical comparison for productivity work. *Frontiers in Virtual Reality* 4 (2023), 1215820.

[48] Karthik Ramani, Subramaniam Chidambaram, Hank Huang, and Fengming He. 2022. System and method for generating asynchronous augmented reality instructions. US Patent 11,380,069.

[49] Karthik Ramani, Subramaniam Chidambaram, and Sai Swarup Reddy. 2024. Digital twin authoring and editing environment for creation of ar/vr and video instructions from a single demonstration. US Patent App. 18/480,173.

[50] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. 2020. Accelerating 3D Deep Learning with PyTorch3D. *arXiv:2007.08501* (2020).

[51] Amy Rechkemmer, Alex C Williams, Matthew Lease, and Li Erran Li. 2023. Characterizing Time Spent in Video Object Tracking Annotation Tasks: A Study of Task Complexity in Vehicle Tracking. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 11. 140–151.

[52] Gang Ren and Eamonn O'Neill. 2013. 3D selection with freehand gesture. *Computers & Graphics* 37, 3 (2013), 101–120. https://doi.org/10.1016/j.cag.2012.12.006

[53] Grand View Research. 2021. Data Collection And Labeling Market Size, Share & Trends Analysis Report By Data Type (Audio, Image/ Video, Text), By Vertical (IT,

Automotive, Government, Healthcare, BFSI), By Region, And Segment Forecasts, 2023 - 2030. Retrieved April 3, from https://www.grandviewresearch.com/industry-analysis/data-collection-labeling-market.

[54] Bryan C. Russell and Antonio Torralba. 2009. Building a database of 3D scenes from user annotations. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2711–2718. https://doi.org/10.1109/CVPR.2009.5206643

[55] Anastasia Ruvimova, Junhyeok Kim, Thomas Fritz, Mark Hancock, and David C. Shepherd. 2020. "Transport Me Away": Fostering Flow in Open Offices through Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376724

[56] Khaled Saleh, Ahmed Abobakr, Mohammed Attia, Julie Iskander, Darius Nahavandi, Mohammed Hossny, and Saeid Nahvandi. 2019. Domain Adaptation for Vehicle Detection from Bird's Eye View LiDAR Point Cloud Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

[57] Nikil Saval. 2015. *Cubed: The Secret History of the Workplace*. Anchor.

[58] Jonathan Dyssel Stets, Yongbin Sun, Wiley Corning, and Scott W. Greenwald. 2017. Visualization and Labeling of Point Clouds in Virtual Reality. In *SIGGRAPH Asia 2017 Posters* (Bangkok, Thailand) *(SA '17)*. Association for Computing Machinery, New York, NY, USA, Article 31, 2 pages. https://doi.org/10.1145/3145690.3145729

[59] Richard Stoakley, Matthew J. Conway, and Randy Pausch. 1995. Virtual Reality on a WIM: Interactive Worlds in Miniature. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. ACM Press/Addison-Wesley Publishing Co., USA, 265–272. https://doi.org/10.1145/223904.223938

[60] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. 2020. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[61] Robert J. Teather and Wolfgang Stuerzlinger. 2007. Guidelines for 3D Positioning Techniques. In *Proceedings of the 2007 Conference on Future Play* (Toronto, Canada) *(Future Play '07)*. Association for Computing Machinery, New York, NY, USA, 61–68. https://doi.org/10.1145/1328202.1328214

[62] Tech Times. 2022. Meta's Oculus Quest 2: Top-Selling VR Headset in 2021. https://www.techtimes.com/articles/273303/20220321/meta-s-oculus-quest-2-top-selling-vr-headset-2021.htm Accessed: 2023-09-14.

[63] Asim Unmesh, Rahul Jain, Jingyu Shi, V. K. Chaithanya Manam, Hyung-Gun Chi, Subramanian Chidambaram, Alexander Quinn, and Karthik Ramani. 2024. Interacting Objects: A Dataset of Object-Object Interactions for Richer Dynamic Scene Representations. *IEEE Robotics and Automation Letters* 9, 1 (2024), 451–458. https://doi.org/10.1109/LRA.2023.3332554

[64] Julien Valentin, Vibhav Vineet, Ming-Ming Cheng, David Kim, Jamie Shotton, Pushmeet Kohli, Matthias Nießner, Antonio Criminisi, Shahram Izadi, and Philip Torr. 2015. SemanticPaint: Interactive 3D Labeling and Learning at Your Fingertips. *ACM Trans. Graph.* 34, 5, Article 154 (nov 2015), 17 pages. https://doi.org/10.1145/2751556

[65] Manuel Veit, Antonio Capobianco, and Dominique Bechmann. 2009. Influence of Degrees of Freedom's Manipulation on Performances during Orientation Tasks in Virtual Reality Environments. In *Proceedings of the 16th ACM Symposium on Virtual Reality Software and Technology* (Kyoto, Japan) *(VRST '09)*. Association for Computing Machinery, New York, NY, USA, 51–58. https://doi.org/10.1145/1643928.1643942

[66] Vinayak, Sundar Murugappan, Cecil Piya, and Karthik Ramani. 2013. Handy-Potter: Rapid Exploration of Rotationally Symmetric Shapes Through Natural Hand Motions. *Journal of Computing and Information Science in Engineering* 13, 2 (04 2013), 021008. https://doi.org/10.1115/1.4023588

[67] Bernie Wang, Virginia Wu, Bichen Wu, and Kurt Keutzer. 2019. LATTE: Accelerating LiDAR Point Cloud Annotation via Sensor Fusion, One-Click Annotation, and Tracking. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. 265–272. https://doi.org/10.1109/ITSC.2019.8916980

[68] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. Whose AI Dream? In Search of the Aspiration in Data Annotation.. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 582, 16 pages. https://doi.org/10.1145/3491102.3502121

[69] Zeyu Wang, Cuong Nguyen, Paul Asente, and Julie Dorsey. 2023. PointShopAR: Supporting Environmental Design Prototyping Using Point Cloud in Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 34, 15 pages. https://doi.org/10.1145/3544548.3580776

[70] Colin Ware, Kevin Arthur, and Kellogg S. Booth. 1993. Fish Tank Virtual Reality. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 37–42. https://doi.org/10.1145/169059.169066

[71] Florian Wirth, Jannik Quehl, Jeffrey Ota, and Christoph Stiller. 2019. PointAtMe: Efficient 3D Point Cloud Labeling in Virtual Reality. In *2019 IEEE Intelligent Vehicles Symposium (IV)*. 1693–1698. https://doi.org/10.1109/IVS.2019.8814115

[72] Yu-Shiang Wong, Hung-Kuo Chu, and Niloy J Mitra. 2015. Smartannotator an interactive tool for annotating indoor rgbd images. In *Computer Graphics Forum*, Vol. 34. Wiley Online Library, 447–457.

[73] Kevin Yu, Alexander Winkler, Frieder Pankratz, Marc Lazarovici, Dirk Wilhelm, Ulrich Eck, Daniel Roth, and Nassir Navab. 2021. Magnoramas: Magnifying Dioramas for Precise Annotations in Asymmetric 3D Teleconsultation. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. 392–401. https://doi.org/10.1109/VR50410.2021.00062

[74] Qian Zhou, George Fitzmaurice, and Fraser Anderson. 2022. In-Depth Mouse: Integrating Desktop Mouse into Virtual Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 354, 17 pages. https://doi.org/10.1145/3491102.3501884

[75] Brian Zimmer, Cynthia Liutkus-Pierce, Scott T Marshall, Kevin G Hatala, Adam Metallo, and Vincent Rossi. 2018. Using differential structure-from-motion photogrammetry to quantify erosion at the Engare Sero footprint site, Tanzania. *Quaternary Science Reviews* 198 (2018), 226–241.

[76] Walter Zimmer, Akshay Rangesh, and Mohan Trivedi. 2019. 3D BAT: A Semi-Automatic, Web-based 3D Annotation Toolbox for Full-Surround, Multi-Modal Data Streams. In *2019 IEEE Intelligent Vehicles Symposium (IV)*. 1816–1821. https://doi.org/10.1109/IVS.2019.8814071