INCLUVERSO 5G

Tecnologías de realIdad exteNdida y Comunicaciones para la incLUsión de personas en situación de VulnERabilidad psicoSOcial mediante redes 5G avanzadas

Entregable E3.3	Evaluación de las tecnologías de XR

Índice

1 Introducción

Este documento detalla la evaluación de las tecnologías de XR accesibles desarrolladas en Incluverso 5G. Incluye la validación de la calidad de experiencia y de interacción (QoE/QoI) de las distintas funcionalidades de XR y el estudio del compromiso entre coste y calidad de los algoritmos de computación delegada. Se detallan las pruebas realizadas y se relacionan con los requisitos detallados en el entregable E1.2.

La estructura del documento se corresponde con el análisis de escenarios base definidos en el entregable E1.2. Para facilitar la investigación en las tecnologías del proyecto y la implementación de los casos de uso, se definieron tres escenarios base de despliegue. Estos escenarios ofrecen soluciones potenciales para la implementación de los distintos casos de uso, y describen para ello una arquitectura de alto nivel que se puede analizar para obtener los requisitos técnicos. Como se detalla en E1.2, este análisis está alineado con el descrito en el informe técnico 3GPP TR 26.928 ("XR in 5G") y en el entregable "D1: Network Requirements and Capabilities" de Metaverse Standard Forum. La asociación entre los casos de uso y los escenarios base se muestra en la siguiente tabla:

Tabla 1: Correspondencia entre los casos de uso y los escenarios base

	Terapia	Telepresencia	Teleformación
Teleportación Virtual		X	X
Computación Delegada en MR	X		X
Regulación Emocional	X		X

Como se comprueba en la Tabla 1, le ejecución de algunos casos de uso extremo a extremo puede implicar combinar varios de los escenarios base, bien simultáneamente, bien de forma secuencial. La definición de los escenarios base facilita también la investigación en tecnología dentro del proyecto. La Tabla 2 muestra la relación entre las tareas del paquete de trabajo P3 y los escenarios base.

Tabla 2: relación entre las tareas del paquete de trabajo P3 y los escenarios base

		Teleportación Virtual	Computación Delegada en MR	Regulación Emocional
A3.1	Presencia autónoma		X	
A3.2	Interacción natural		X	

A3.3	Telepresencia inmersiva	X		
A3.4	Biomarcadores digitales			X
A3.5	Sistema XR	X	X	X
A3.6	Co-diseño	X	X	X

Los capítulos 2, 3 y 4 analizan cada uno de lose scenarios base desde el punto de vista de las tecnologías implementadas (tareas A3.1 – A3.4). El capítulo 2 describe los resultados del escenario de teleportación virtual (tarea A3.3). El capítulo 3 describe los resultados del escenario de computación delegada (tareas A3.1 y A3.2). El capítulo 4 describe los resultados del escenario de biomarcadores digitales (tarea A3.4). El capítulo 5, finalmente, analiza de forma global los tres casos de uso del proyecto (terapia, telepresencia y teleformación), mostrando cómo se ha llevado el desarrollo tecnológico a los casos de uso a los usuarios finales, incluyendo el proceso de co-diseño tanto con personas beneficiarias como profesionales (tareas A3.5 y A3.6).

2 Escenario de Teleportación Virtual: tecnologías de telepresencia

2.1 Tecnología de telepresencia utilizada en el caso de uso

Las tecnologías evaluadas en esta sección son las presentadas en el entregable E3.2: el búho y el búho con la herramienta adicional de la lupa. Sin embargo, para tener una línea base con la que comparar esta tecnología y por ser la tecnología más utilizada actualmente, se añade como condición una videollamada 2D tradicional. De esta forma, las tecnologías consideradas en el estudio son:

Condición #2: videollamada 2D ordenador-móvil
Condición #3: videollamada basada en vídeo 360º con el búho
Condición #4: videollamada basada en vídeo 360º con el búho y la lupa

De esta forma, evaluamos la videollamada 2D realizada entre personas conectadas desde el portátil y una persona conectada desde el móvil. La Condición #2 puede extenderse pasando de una videollamada estándar a una transmisión de vídeo 360 grados (Condición #3). Esto se puede lograr utilizando una cámara de 360º en el entorno real, integrada en el prototipo búho, que permite capturar completamente el entorno y transmitirlo a la persona experta remota. El vídeo puede ser visualizado con unas gafas de Realidad Virtual (VR) en el lado del instructor/a, aumentando así su sensación de inmersión y participación en la tarea. Este escenario puede extenderse aún más proporcionando a la persona que se encuentra con el búho una herramienta de zoom (Condición #4). En este caso, aunque el marco general es similar al de la Condición #3, el/la usuario/a puede transmitir un vídeo 2D capturado por un móvil inteligente, enfocando áreas específicas de la escena para que el instructor/a pueda verlas con más detalle y así ayudarle más fácilmente. Para permitir que la persona experta remota visualice las dos fuentes de vídeo, este vídeo 2D se presenta en una ventana flotante en el entorno virtual.

2.2 Estudio del compromiso entre coste y calidad en el caso de uso de telepresencia

La colaboración en entornos virtuales pronto se convertirá en una realidad. Esto es posible gracias al rápido desarrollo de dispositivos de consumo que soportan experiencias de Realidad Extendida (XR) fáciles de usar, y a las capacidades ofrecidas por las tecnologías de comunicación más recientes, como 5G y WiFi7. Los entornos remotos colaborativos son elementos clave en muchos sectores que van desde el entretenimiento hasta la formación y la educación. En los últimos años, los sistemas tradicionales consistían típicamente en comunicaciones de audio uno a uno y rara vez involucraban a más participantes. Hoy en día,

este escenario ha cambiado y las teleconferencias, ampliamente utilizadas durante la pandemia de COVID-19, están ahora ampliamente adoptadas. Las teleconferencias permiten la participación de múltiples usuarios con modalidades diferentes y heterogéneas (por ejemplo, solo audio, audio y vídeo). Además, la posibilidad de compartir la pantalla o documentos de los usuarios, y de interactuar fácilmente con personas de todo el mundo, ha aumentado significativamente la efectividad de la colaboración. El efecto inmediato de esta mejora es que los usuarios y usuarias tienden a estar más dispuestos a utilizar estas herramientas. Esto también tiene un impacto significativo desde un punto de vista ambiental: diferentes estudios muestran que las reuniones remotas permiten reducir las emisiones globales de gases. Sin embargo, además de los problemas técnicos (es decir, caída en la calidad de audio/vídeo, conexión a internet), todavía existen algunos problemas abiertos que reducen la calidad percibida por los usuarios y usuarias finales, como la dificultad para mantenerse concentrado/a o la distracción causada por el fondo ambiental de las otras personas que participen. En este contexto, los entornos virtuales colaborativos podrían representar una solución efectiva, ya que proporcionan una sensación de copresencia, permitiendo así superar algunos de los problemas mencionados anteriormente.

Los entornos inmersivos colaborativos se pueden dividir en dos categorías: sistemas simétricos y asimétricos. En escenarios simétricos, los/as usuarios/as suelen emplear el mismo tipo de dispositivo e interactúan como pares, mientras que en escenarios asimétricos los dispositivos adoptados son heterogéneos y los roles de las personas que participan en la comunicación también difieren (por ejemplo, instructor/alumno). Debido a la amplia variedad de configuraciones, los sistemas asimétricos pueden ser significativamente diferentes entre sí, lo que hace que la evaluación de la Calidad de Experiencia (QoE) sea extremadamente desafiante. Hasta donde sabemos, a pesar de la cantidad de estudios que se han realizado en el estado del arte, no existen estándares o recomendaciones que definan metodologías acordadas para este propósito. Para contribuir a la investigación de este caso de uso concreto, presentamos una metodología para la evaluación del compromiso entre coste y calidad en aplicaciones asimétricas interactivas y colaborativas. Además, introducimos una tarea versátil que se valida en diferentes tecnologías, como videollamadas 2D (condición #2) y videollamadas 360º (condición #3 y #4). Es importante subrayar que nuestro objetivo es evaluar la telepresencia considerando tanto la percepción de las personas que utilizan la tecnología (profesionales y personas usuarias) como su efectividad en la realización de tareas específicas. De hecho, para aplicaciones no relacionadas con el entretenimiento, una buena experiencia no solo abarca el disfrute de la persona, sino también el rendimiento exitoso de una tarea que involucra tecnologías inmersivas, fundamental en este caso de uso.

Uno de los elementos clave en el diseño de una metodología para la evaluación de la QoE es la selección de tareas adecuadas. El estudio presentado en [1] presenta que las tareas de comunicación inmersiva se pueden clasificar principalmente en tres categorías: deliberación, exploración y manipulación. La primera implica solo conversación, la segunda utiliza la

comunicación para explorar un entorno e identificar sus elementos, mientras que la manipulación abarca la interacción con objetos. La tarea base del caso de uso de telepresencia se caracteriza principalmente por la exploración, ya que las personas pueden navegar y explorar colaborativamente un entorno, en este caso real, utilizando el búho. Sin embargo, también incluye elementos de deliberación, ya que los usuarios y profesional se comunican para realizar la tarea, y manipulación, ya que interactúan con objetos dentro del entorno. Para incluir estas características y crear un escenario asimétrico, implementamos un juego de búsqueda del tesoro, distribuyendo la información necesaria para completar la tarea entre los participantes, creando así roles de experto (instructores/as) y no experto (explorador/a). Esto permite la simulación de escenarios colaborativos realistas, como el de apoyo remoto en la realización de tareas domésticas.

En resumen, se presenta una metodología diseñada para la evaluación de la QoE en entornos interactivos y colaborativos, que permite:

Considerar tanto los aspectos técnicos como los elementos socioemocionales, como los factores contextuales y humanos.

Proporcionar un protocolo experimental común para comparar tecnologías que van desde configuraciones tradicionales (es decir, videollamadas 2D) hasta entornos virtuales (vídeo 360º).

Simular entornos realistas en un entorno de experimentación controlado

Por lo tanto, este capítulo aborda las siguientes preguntas de investigación:

Q1: ¿Es posible diseñar una metodología integral y general para la evaluación de la QoE en comunicaciones inmersivas?

Q2: ¿Cómo impactan las condiciones experimentales (en términos de información transmitida y tecnología empleada) en la efectividad de la realización de una tarea asignada?

Q3: ¿Cómo impactan las condiciones experimentales (en términos de información transmitida y tecnología empleada) en la experiencia percibida por la persona participante?

Los resultados de este estudio pueden apoyar también la creación de un marco común para la evaluación de comunicaciones inmersivas, gracias a su versatilidad y aplicabilidad en múltiples configuraciones y tecnologías.

Para abordar la primera pregunta de investigación (Q1), se necesitan tres componentes clave: un protocolo experimental con cuestionarios y métricas utilizadas para la evaluación, la definición de una tarea a realizar y la definición de las condiciones de prueba, incluyendo las soluciones tecnológicas empleadas.

A continuación, presentamos el protocolo experimental definido. Dada la naturaleza colaborativa de la tarea, sugerimos reclutar participantes en grupos de personas que se

conozcan entre sí. Este enfoque minimiza el riesgo de problemas de comunicación que podrían surgir si los participantes se conocieran por primera vez durante la prueba. Para aumentar aún más la naturalidad de la comunicación, los participantes deben usar el idioma que normalmente utilizan para interactuar. Además, todos los participantes deben leer una hoja informativa que detalle el experimento antes de la prueba. Asimismo, los participantes deben firmar el formulario de consentimiento, de acuerdo con el GDPR, y realizar una evaluación inicial. Esta última tiene como objetivo identificar posibles personas que no sean adecuadas para realizar la tarea. Después de esta fase, los participantes completan un cuestionario previo a la prueba, que se utiliza para recopilar información demográfica y general que pueda influir en su rendimiento. La prueba experimental comienza con la sesión de entrenamiento, que proporciona instrucciones prácticas a las personas participantes (por ejemplo, cómo utilizar los mandos para modificar las ventanas flotantes del entorno virtual) y les ayuda a familiarizarse con la tarea y las tecnologías empleadas. Además, la fase de entrenamiento se aprovecha para asegurar que todos los componentes del sistema de comunicación funcionen correctamente. Después de la sesión de entrenamiento, se lleva a cabo la condición experimental, los cuestionarios post-condición y un descanso. Estos tres pasos pueden repetirse para evaluar la QoE de los usuarios bajo diferentes condiciones experimentales. Las características de las condiciones experimentales dependen en gran medida de los objetivos del estudio y un elemento clave está representado por la selección de la tecnología inmersiva a emplear.

El protocolo experimental debe incluir un conjunto de métricas y cuestionarios para evaluar la experiencia. En este análisis, para responder a las preguntas de investigación Q2 y Q3, proponemos evaluar el impacto de las condiciones experimentales en la efectividad de las personas participantes al realizar una tarea, así como en la QoE percibida. Para alcanzar el primer objetivo, consideramos una evaluación objetiva del rendimiento de los/as participantes mediante la medición del tiempo de finalización de las tareas. En cuanto al segundo objetivo, seleccionamos los cuestionarios post-condición para evaluar los siguientes componentes de la QoE.

- El cibermareo se refiere a las sensaciones desagradables que pueden surgir durante y después de la exposición a la XR. Los síntomas suelen aumentar con el tiempo de exposición. Para evaluarlo existen dos posibilidades principales: el Cuestionario de Simulator Sickness (SSQ) [2] o la Escala de Valoración de Vértigo (VSR) [3]. El primero incluye 16 síntomas de malestar (por ejemplo, náuseas, malestar y dolor de cabeza). En contraste, el VSR consiste en una única pregunta sobre el nivel actual de malestar. Como uno de los aspectos clave de la tarea considerada es la exploración de un entorno virtual o real, es necesario monitorizar y controlar cualquier síntoma durante la sesión.
- La presencia social representa una condición psicológica en la que las personas usuarias se perciben a sí mismas como parte de un entorno interpersonal [4]. Para evaluar la

presencia social, empleamos el cuestionario propuesto en [5] adaptado de [6]. Las preguntas sobre presencia social tienen como objetivo evaluar el nivel de participación mutua en la experiencia y, por lo tanto, su evaluación es extremadamente relevante para aplicaciones colaborativas asimétricas.

- La presencia espacial representa una condición psicológica en la que las personas se sienten presentes dentro de un entorno físico diferente al que están [4]. Para evaluar la presencia espacial, empleamos una versión submuestreada del Cuestionario de Presencia [7], eliminando las preguntas que no aplicaban a nuestra tarea.
- La carga cognitiva se refiere al esfuerzo requerido para realizar una tarea. Para evaluarla, empleamos el cuestionario NASA-TLX [8]. Para la tarea considerada, la evaluación de la carga cognitiva es muy recomendable debido a la asimetría en términos de tecnología, así como de roles (experto vs. no experto), para comprobar si estas diferencias están asociadas a demandas desequilibradas.
- La Calidad de Interacción (QoI) se refiere a la eficacia con la que las personas usuarias interactúan y se relacionan entre sí, abarcando la calidad de la conversación y el rendimiento de la comunicación [9]. Empleamos el cuestionario propuesto en [10]. Como indicador de la calidad de las interacciones, su evaluación es clave para la valoración de aplicaciones colaborativas.
- La QoE general indica el nivel de satisfacción de la persona que utiliza la tecnología con una aplicación o servicio. Resulta del cumplimiento de sus expectativas con respecto a la utilidad o el disfrute de la aplicación o servicio y se ve afectada por la personalidad y el estado actual del usuario. Para su evaluación, utilizamos una Calificación de Categoría Absoluta (ACR) para la calidad de audio y video, y la QoE percibida en general. Además, evaluamos la percepción de la latencia de transmisión con la única pregunta "¿Percibió alguna reducción en su capacidad para interactuar durante la comunicación debido al retardo?". Aunque no se introdujeron deterioros de red durante las sesiones, evaluamos la calidad del vídeo, audio y la latencia para evaluar su impacto potencial en otros componentes de la QoE.

En cuanto a la definición de la tarea, se presenta una tarea gamificada que puede aplicarse a diversas situaciones de la vida real. Como ya se ha comentado, se plantea un juego de búsqueda del tesoro en el que los/as participantes, es decir, exploradores no expertos, tienen que encontrar pistas ocultas en un entorno basándose en las instrucciones proporcionadas por el(los) otro(s) usuario(s), denominado(s) instructor(es). Las pistas se codifican mediante un código de sustitución simple, y las pistas decodificadas se utilizan para avanzar en el juego. Para la codificación de los símbolos, optamos por el alfabeto Matoran, en el que los símbolos se asocian a letras (ver Figura 3b para conocer el alfabeto). El rol de los participantes difiere según sus capacidades y la información disponible para resolver la tarea. Basándose en esto, se pueden identificar dos posibles subtareas:

Guía la navegación (instructor) y explora el entorno (explorador): instructores guían al explorador a través del entorno para llegar a la sala donde se ocultan las pistas. Una vez allí, el explorador busca las pistas en la sala. La información disponible para el instructor consiste en las pistas que deben ser descubiertas y/o un mapa del entorno.

Decodifica las pistas (instructor y explorador): esta subtarea consiste en decodificar las pistas en un orden predeterminado para permitir la progresión en la tarea. Para ello, el explorador podría describir verbalmente o mostrar visualmente las pistas al instructor. Luego, el instructor utiliza un diccionario para descifrarlas y uno de los participantes introduce la solución decodificada en una Interfaz de Usuario (UI) proporcionada, que podría ser basada en web o instalada en un dispositivo local. Este paso permite la evaluación de la secuencia decodificada, permitiendo así a los participantes avanzar en la tarea solo si la decodificación es exitosa. En esta subtarea, las herramientas necesarias son el diccionario para la decodificación por parte del instructor y la UI por parte del explorador o del instructor.

La implementación específica de la tarea puede incluir una o más subtareas. Además, cada subtarea puede implementarse utilizando diferentes niveles de dificultad (por ejemplo, aumentando el tamaño del entorno, el número de pistas o la complejidad de la decodificación). Las pistas pueden ser objetos 2D (es decir, carteles) u objetos 3D (es decir, cajas).

La tarea de búsqueda del tesoro es adecuada para probar entornos colaborativos donde hay un participante experto y uno no experto. En este trabajo, implementamos tres casos de uso: teleasistencia, visita a un museo y soporte técnico (ver Figura 1). En el caso de la teleasistencia, la tarea se desarrolla en un apartamento supervisado donde los residentes (exploradores) requieren el apoyo de un profesional remoto (instructor) para encontrar diferentes ingredientes en un estante para cocinar u objetos para realizar tareas diarias. En el escenario del museo, el explorador se hace pasar por un visitante de un museo y el instructor desempeña el papel de guía turístico. En el caso de uso de soporte técnico, el explorador aprende a reconocer diferentes elementos en un laboratorio bajo la guía de un instructor experto.

Figura 1: 3 Ejemplos de las cajas y pósteres utilizados en cada caso de uso.

Como siguiente paso para llevar a cabo este estudio, se diseña un entorno real para los requisitos de la dinámica del juego de la búsqueda del tesoro. Configuramos un pasillo y dos habitaciones, presentados en la Figura 2. Cada habitación o pasillo fue diseñado para representar un caso de uso específico: teleasistencia (Habitación 1), museo/turismo (Pasillo) y soporte técnico (Habitación 2). En cada habitación colocamos dos objetos que contenían tres pistas cada uno. Para los casos de uso de telepresencia y soporte técnico, había un objeto 2D (póster) y un objeto 3D (caja), mientras que en el caso de uso del museo, había dos objetos 2D (pósteres). Tal y como se presenta en la Figura 1, explicamos con los ejemplos los detalles de estos objetos.

Teleasistencia:

- **Estante de cocina (póster).** La idea es probar la tecnología simulando una situación en la que una persona remota ayuda a los locales a encontrar ingredientes en el estante para hacer una receta o colocar artículos de compra.
- **Productos de limpieza (caja).** La tarea es similar a la del estante de especias, pero basada en productos de limpieza alrededor/dentro de una caja. Los usuarios locales pueden mover la caja, acercándola al búho o girándola para ver otras partes.

Visita guiada en un museo (dos pósteres). El objetivo es imitar una ruta guiada por un museo por una persona experta en arte, donde el/la participante local debe identificar partes seleccionadas de las pinturas. Esta sala está compuesta por dos cuadros. En el ejemplo, "Autorretrato con Collar de Espinas y Colibrí" de Frida Kahlo y "La Escuela de Atenas" de Rafael.

Soporte Técnico. De nuevo, como en las otras dos salas, hay dos tareas que resolver:

Circuito Eléctrico (póster). La idea es ayudar a las personas participantes proporcionando soporte técnico basado en un circuito que no se puede mover y sobre el cual se debe ejecutar alguna operación.

Elementos del Circuito (caja). Esta tarea es similar a la anterior, pero esta vez los elementos del circuito están alrededor o dentro de la caja y, de manera similar a la tarea de teleasistencia de productos de limpieza, la persona local puede mover la caja.

En cada objeto (póster o caja), las personas locales con ayuda de las personas remotas deben descifrar tres símbolos del alfabeto Matoran en un orden específico (ver Figura 1).

De esta forma, la persona no experta es guiada por dos participantes remotos expertos (instructores/as) que colaboraron para ayudarle a encontrar las pistas en el entorno físico y decodificarlas lo más rápido posible. Los/as instructores/as se ubicaron en habitaciones separadas (pequeñas habitaciones A y B en la Figura 2). Cabe destacar que el orden en que se exploró el espacio se mantuvo igual en todos los experimentos: Habitación 1, Pasillo y Habitación 2.

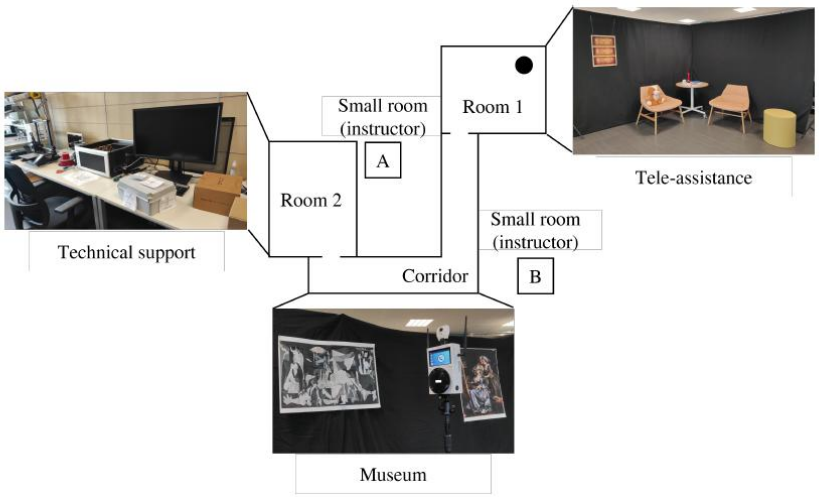
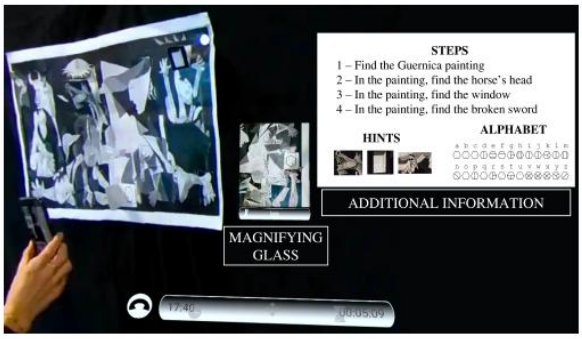


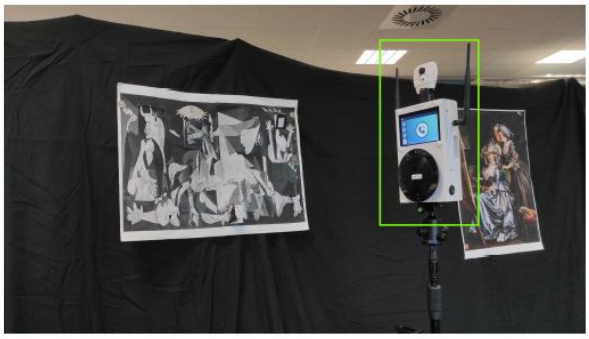
Figura 2: La imagen muestra un mapa del espacio físico utilizado para la evaluación. Se consideran 3 salas asociadas a 3 casos de uso de telepresencia en los que hay un apoyo remoto: visita a museos, soporte técnico y teleasistencia para ayuda en la realización de tareas cotidianas.

Las Condiciones #2, #3 y #4 (vídeo 2D, vídeo 360 y vídeo 360 con lupa, respectivamente) fueron experimentadas por cada grupo de tres participantes en el entorno físico descrito. Todos los objetos, cajas y pósteres, los objetos a encontrar en los mismos y los símbolos, se cambiaron en cada iteración para evitar efectos de memoria.

La Figura 3 proporciona un ejemplo de la vista del instructor extraída de las gafas de RV y un ejemplo del entorno físico en el lado del explorador en la condición #4.



(a) Instructor viewport



(b) Physical environment (explorer environment)

Figura 3: (a): Vista del instructor desde el HMD con información adicional. B) Entorno real en el que se encuentran los pósteres y el prototipo búho.

Como se empleó el búho, que permite la comunicación de vídeo de 360 grados, se utilizó sobre un trípode con ruedas, lo que permitía al explorador moverlo y navegar por el entorno físico. El explorador visualizaba en la pantalla los avatares de los instructores, cuyos movimientos de labios y cabeza estaban sincronizados con las acciones de los instructores mediante los sensores de las gafas de VR. Además, cada instructor estaba representado por su avatar en la vista del otro instructor, facilitando la conversación entre los tres participantes.

Además del avatar del otro instructor, como se ilustra en la Figura 3a, la vista de cada instructor se componía de tres ventanas flotantes: la ventana del búho, la ventana de la lupa y la ventana de información adicional. Los/as instructores/as tenían la capacidad de mover las ventanas y reubicarlas dentro de su vista utilizando los mandos. La primera ventana mostraba el vídeo capturado por el búho (véase el lado izquierdo de la Figura 3a), mientras que la ventana de la lupa mostraba un vídeo 2D adicional transmitido por el explorador a través de un teléfono inteligente. Esto se introdujo como una herramienta adicional para proporcionar la funcionalidad de aumento/lupa. La razón de esta elección fue que, si la calidad del vídeo de 360 grados no era suficiente para discernir detalles específicos (por ejemplo, el símbolo a decodificar), el explorador podía usar la cámara de este móvil para enfocar el área relevante y transmitir esa porción de la escena a los instructores. La ventana de información adicional proporcionaba al instructor el conocimiento necesario para guiar al explorador, es decir, el póster o la caja objetivo, las pistas a encontrar dentro de esos objetos y el diccionario para decodificar el símbolo encontrado por el explorador. Los dos instructores tenían acceso a la misma información.

En la condición #3, el explorador estaba en las mismas condiciones que en la condición #4 (el búho, aunque en la #4 manejaba el móvil también), y la vista del instructor tenía los mismos componentes excepto la ventana de la lupa. Por el contrario, en la condición #2, el explorador estaba equipado con un móvil que transmitía el video al instructor. Cada instructor disponía

de un portátil para ver el vídeo que transmitía un explorador y un conjunto de diapositivas que contenían los mismos datos mostrados en el panel de información adicional en las otras dos condiciones.

Las personas participantes en la evaluación fueron reclutados a través de un formulario de registro, donde se les pedía que se inscribieran en grupos de tres. El procedimiento de selección incluyó la prueba de Snellen y la prueba de daltonismo de Ishihara. La primera se realizó para evaluar la agudeza visual de las personas participantes. Además, se comprobó si las personas participantes padecían daltonismo y solo se permitió participar a las personas con una visión cromática correcta. Concretamente, 24 participantes formaron parte de la prueba, con una edad promedio de 44.00 ± 11.11 años, y edades que oscilaban entre los 23 y los 57 años. De estas personas, el 45.8 % se identificó como mujer y el 54.2 % como hombre. Además, el 58.33% de los participantes usaba gafas o lentes de contacto. En cuanto al nivel educativo, el 54.17% tenía un máster, el 33.33% una licenciatura, el 4.17% un doctorado, el 4.17% reportó "otro" y el 4.17% había completado la escuela secundaria. En términos de experiencia previa con tecnología XR, el 87.5% se describió como principiante, mientras que el resto de los participantes reportó una experiencia intermedia.

La prueba consistió en ocho sesiones, cada una con grupos de tres participantes, quienes experimentaron las condiciones #2, #3 y #4.

Para abordar las preguntas de investigación Q2 (es decir, el impacto de la condición en la efectividad de los/as participantes al realizar la tarea) y Q3 (es decir, el impacto de la condición en la QoE percibida por las personas usuarias), complementamos el análisis basado en cuestionarios con la evaluación objetiva del tiempo de finalización de la tarea. En cuanto a la QoE percibida, realizamos pruebas estadísticas para evaluar el impacto de las diferentes condiciones en los componentes de QoE identificados. Con este fin, promediamos las respuestas de cada cuestionario post-condición para cada persona y realizamos análisis separados para instructores y exploradores. Respecto al rendimiento objetivo, el tiempo de finalización se midió como el intervalo necesario para encontrar las tres pistas por objeto y decodificar los símbolos Matoran.

Para abordar la pregunta de investigación sobre el impacto de la condición en la efectividad de los usuarios al realizar la tarea (Q2), presentamos el tiempo promedio de finalización de la tarea por condición experimental e iteración en la Figura 4. Cada grupo realizó el experimento tres veces, experimentando una condición diferente en cada iteración, y los grupos realizaron las condiciones en órdenes distintos. Debido a problemas técnicos durante la ejecución de la prueba, el orden de las condiciones no estuvo completamente equilibrado entre los grupos de participantes, lo que representa una limitación del estudio. Para tener esto en cuenta, analizamos el efecto de la condición controlando la iteración del experimento utilizando un modelo lineal de efectos mixtos, con la condición y la iteración como parámetros fijos y el grupo como variable aleatoria. Para obtener estimaciones robustas de

las medias de las condiciones, calculamos las medias de mínimos cuadrados (medias marginales ajustadas por iteración) y derivamos intervalos de confianza del 95% mediante muestreo Bootstrap a nivel de grupo (Figura 4a). Se realizaron comparaciones por pares entre las condiciones con valores p ajustados por Bonferroni. El mismo procedimiento se llevó a cabo para analizar el efecto de la iteración del experimento controlando la condición (Figura 4b).

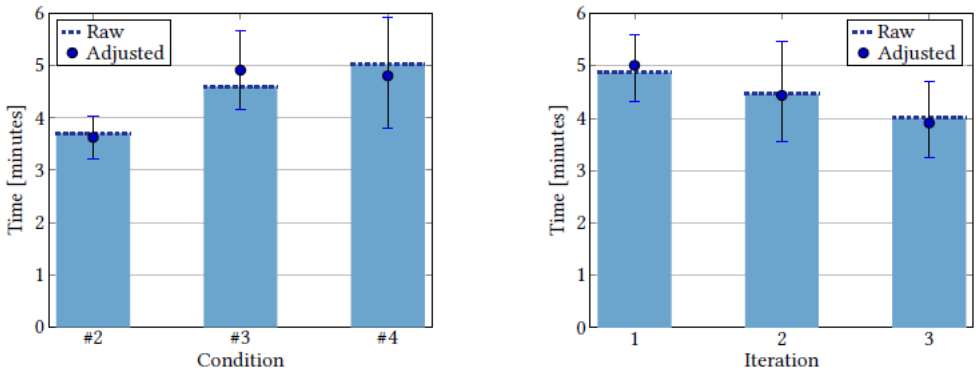


Figura 4: Tiempo promedio invertido para completar las tareas por experimento y condición y por iteración. Las barras de error indican el Intervalo de Confianza (IC) del 95% obtenido mediante bootstrapping.

La comparación de las condiciones sugiere que la familiaridad con la tecnología influyó en el tiempo de finalización: los participantes en la condición #2 (sin dispositivos inmersivos) completaron la tarea más rápido que en las condiciones #3 y #4 (con dispositivos inmersivos) ($p < 0.05$). Los instructores en el entorno de 360 grados requirieron tiempo adicional para posicionar los paneles del alfabeto y la lupa, lo que aumentó el tiempo necesario para enfocar su atención. Por el contrario, la condición #2 proporcionó una interfaz 2D familiar (diapositivas y video transmitido en un portátil) más cercana a la práctica diaria de los/as instructores/as.

Además, se analizó el rendimiento de las personas participantes a lo largo de las iteraciones. Aunque los objetos y símbolos variaron entre iteraciones, aprendieron la dinámica de la tarea y los tipos de objetos/símbolos con el tiempo, lo que permitió una finalización más rápida del juego. Nos referiremos a este fenómeno como efecto de aprendizaje. Este efecto fue estadísticamente significativo entre la primera y la tercera iteración ($p < 0.05$), como se muestra en la Figura 4b.

En cuanto al impacto en la QoE percibida (Q3), la Figura 5 presenta la Puntuación Media de Opinión (MOS) con IC del 95% para la calidad de audio percibida, la calidad de vídeo (solo instructores/as), el impacto de la latencia y la QoE general después de las condiciones #2, #3 y #4.

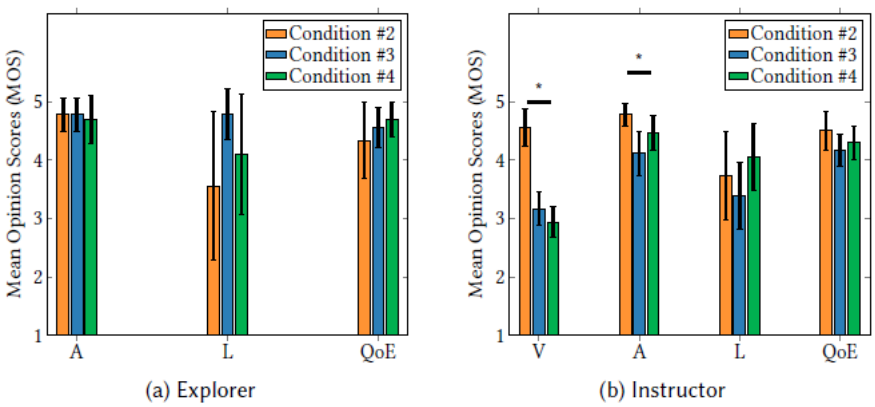


Figura 5: MOS y el IC del 95% asociado para los cuestionarios sobre la calidad de video (V) y audio (A), el impacto de la latencia (L) y la QoE después de las condiciones #2, #3 y #4. Los asteriscos indican diferencias estadísticamente significativas.

La Figura 6 muestra la MOS y los IC del 95% para la carga cognitiva, la QoI y la presencia social y espacial (solo instructores/as). Dado que cada grupo incluía una persona exploradora local y dos instructoras, el número de evaluaciones de las personas instructoras fue el doble que el de las exploradoras. Por lo tanto, los análisis se realizaron por separado para ambos roles en las diferentes condiciones.

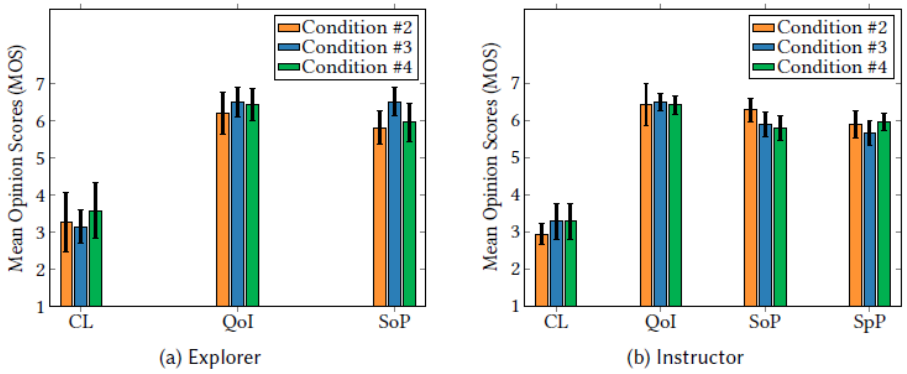


Figura 6: MOS y el IC del 95% asociado para los cuestionarios de carga cognitiva (CL), Calidad de Interacción (QoI), y presencia social y espacial en las condiciones #2, #3 y #4.

Para analizar los datos del post-cuestionario, se realizó la prueba de normalidad de Pearson-D'Agostino para validar la distribución normal de los datos recopilados. Para los datos distribuidos normalmente, se utilizó el Análisis de Varianza (ANOVA) para examinar las diferencias entre las condiciones #2, #3 y #4. En caso contrario, se aplicó la prueba de Kruskal-Wallis. Además, se empleó la prueba de Dunn con corrección de Bonferroni para comparaciones múltiples. El umbral de significancia se estableció en 0.05.

Dado que la evaluación de la calidad de vídeo en una escala de 5 niveles puede modelarse como un proceso aleatorio gaussiano, realizamos un análisis paramétrico de las puntuaciones, siguiendo las prácticas estándar para evaluar datos de calidad de vídeo. Los resultados revelaron diferencias significativas entre las condiciones ($F2,50 = 34.1116$, $p < 0.01$). La prueba de Diferencia Honestamente Significativa (HSD) de Tukey mostró además que la calidad de vídeo en la condición no inmersiva (#2) fue calificada significativamente más alta que en las condiciones inmersivas (#3 y #4).

Para la calidad de audio, la prueba de Kruskal-Wallis presentó una diferencia significativa entre las evaluaciones de instructores ($p = 0.01$). Posteriormente, la prueba de Dunn con corrección de Bonferroni reveló una diferencia significativa entre las condiciones #2 (vídeo 2D) y #3 (configuración basada en el búho) con $p = 0.01$. Este resultado probablemente se debe a la pérdida de paquetes en las sesiones inmersivas que dificultó la comunicación.

Las evaluaciones de latencia (donde 5 corresponde a ningún impacto en el rendimiento de la tarea) no mostraron diferencias significativas entre las condiciones ni para los exploradores ni para los instructores ($p > 0.05$). Sin embargo, la condición #3 reveló una diferencia de MOS de 1.2 entre exploradores e instructores. En esta condición, el explorador se basó principalmente en la comunicación de audio, y los instructores tanto en audio como en el vídeo capturado por el búho. En este sentido, fue posible para los instructores percibir si había algún problema de desincronización entre audio y vídeo. Esto puede justificar el aumento de la latencia percibida por los instructores. También percibieron latencia del otro instructor remoto, representado solo por un avatar con información contextual limitada. Esta falta de realimentación visual hizo que las interrupciones fueran más difíciles de manejar y causó mayor incomodidad en el lado del instructor de la comunicación. Otro factor que podría influir en los resultados es que los instructores tuvieron que colaborar entre sí para dar instrucciones al explorador, mientras que el papel principal del explorador era seguir las instrucciones.

Finalmente, las puntuaciones de QoE fueron altas en todas las condiciones para exploradores e instructores, sin diferencias significativas ($p > 0.05$).

De manera similar, la carga cognitiva fue muy baja en todas las condiciones, tanto para exploradores como para instructores (ver Figura 6), sin diferencias significativas entre las condiciones ($p > 0.05$), lo que confirma la efectividad del diseño gamificado para escenarios del mundo real como la teleasistencia, visitas a museos y soporte técnico. La QoI (Calidad de Interacción) también recibió altas evaluaciones, con MOS (Mean Opinion Scores) casi idénticos en ambos roles, lo que destaca el potencial de estas tecnologías para casos de uso en el mundo real. A pesar de las asimetrías de comunicación en las configuraciones basadas en el búho (condiciones #3 y #4), no se observaron diferencias significativas ($p > 0.05$).

Las puntuaciones de presencia social fueron altas en todas las condiciones, sin diferencias significativas ($p > 0.05$). Es importante señalar que las evaluaciones de los instructores

reflejaron tanto su percepción del explorador como del otro instructor remoto representado por un avatar ya que no había preguntas específicas. Los valores obtenidos en las condiciones #3 y #4 son consistentes con estudios previos [5], a pesar de la complejidad adicional de dos participantes remotos (instructores) y el movimiento del búho.

El impacto de estos factores en la presencia social y la calidad es un aspecto novedoso explorado por primera vez en esta prueba. La presencia espacial se analizó desde la perspectiva de los instructores y no mostró diferencias significativas ($p > 0.05$). Contrariamente a este resultado, se preveía que el sentido de presencia aumentaría en las condiciones #3 y #4. De hecho, se esperaba que el vídeo omnidireccional mejorara la sensación de presencia del instructor, quien podría "teletransportarse" al entorno físico.

Finalmente, el cibermareo se evaluó utilizando el SSQ y el VSR. Dado que se incluyeron dos condiciones consecutivas basadas en HMD, el SSQ se midió al principio y al final, mientras que el VSR se preguntó entre ellas para detectar molestias que pudieran impedir el uso del HMD para una condición experimental adicional. El VSR proporcionó la información necesaria de manera oportuna y efectiva. La Figura 7 muestra las respuestas del SSQ de una persona que abandonó el experimento debido al inicio de cibermareo mientras usaba el HMD. Como se puede observar, el único síntoma que recibió la puntuación máxima es "Malestar general", mientras que los demás no lograron capturar la gravedad de las sensaciones que esa persona estaba experimentando. Además, dado que el experimento no se centró en investigar los síntomas de malestar que una experiencia con el HMD podría producir, sino solo la presencia o ausencia de malestar, recomendamos la pregunta del VSR para calificar el cibermareo en estudios similares. Es útil señalar que los participantes que completaron todas las condiciones no experimentaron molestias significativas según las puntuaciones de SSQ y VSR.

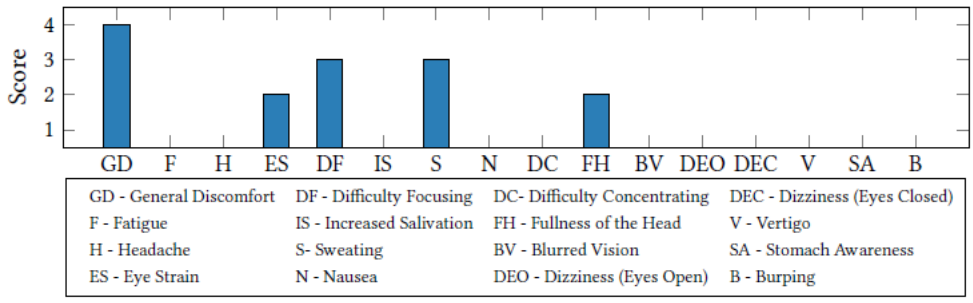


Figura 7: Los valores de SSQ recopilados de una persona participante que sufrió cibermareo durante las pruebas preliminares.

Conclusiones

Los resultados obtenidos en esta evaluación que simula entornos de colaboración de persona remota experta versus persona local no remota son muy positivos y esperanzadores. Por supuesto que existe una curva de aprendizaje hasta que una persona con el HMD se acostumbra al entorno inmersivo pero, a pesar de esta limitación y desventaja de las tecnologías inmersivas respecto a las no inmersivas, la experiencia es muy positiva y no se encuentran diferencias estadísticamente significativas.

Se observaron puntuaciones altas de Calidad de Experiencia (QoE) y una carga cognitiva muy baja en todas las condiciones, tanto para exploradores como para instructores, lo que valida la efectividad del diseño gamificado para diversos escenarios del mundo real. La Calidad de Interacción (QoI) recibió altas puntuaciones, y la presencia social fue consistentemente alta en todas las condiciones, sin diferencias significativas, subrayando el potencial de estas tecnologías inmersivas para casos de uso prácticos. Aunque no hubo diferencias significativas generales en la latencia, los instructores percibieron una mayor latencia en las condiciones con el búho, posiblemente debido a la desincronización de audio/vídeo y la limitada información contextual de los avatares. La condición no inmersiva (video 2D) fue calificada significativamente mejor en calidad de audio que la configuración inmersiva basada en el búho, probablemente debido a la pérdida de paquetes en las sesiones inmersivas. Esto es algo con lo que podíamos contar ya que la condición #2 se trata de una tecnología madura y enfrentarla a las condiciones #3 y #4, basadas en un prototipo, no es del todo justo. Contrariamente a las expectativas, la presencia espacial desde la perspectiva de los instructores no mostró diferencias significativas entre las condiciones, a pesar del vídeo omnidireccional en los entornos inmersivos. Finalmente, aunque se identificó el cibermareo como un factor potencial (con un participante abandonando el experimento), la mayoría de los participantes que completaron todas las condiciones no experimentaron molestias significativas, y se recomienda el uso del VSR para su detección en estudios similares. En resumen, a pesar de desafíos, la evaluación demuestra que las tecnologías inmersivas ofrecen una experiencia colaborativa robusta y efectiva, con altos niveles de QoE, baja carga cognitiva y una fuerte presencia social, lo que las posiciona como herramientas prometedoras para aplicaciones experto-no experto en el mundo real.

2.3 Requisitos e indicadores de rendimiento

La siguiente tabla muestra un resumen de los distintos escenarios en los que se ha validado la teleportación virtual: las pruebas sistemáticas presentadas en el apartado anterior, el caso de uso de telepresencia en pisos tutelados (CU2) descrito en el entregable E4.2 y el caso de use de teleformación (CU3) descrito en el entregable E4.3. Se han probado distintas combinaciones de usuarios locales y remoto, y se ha evaluado en diferentes contextos, pero empleando herramientas y cuestionarios similares.

Tabla 3: Análisis de KPIs para el escenario de teleportación virtual

KPIs		Requisitos	**Pruebas**				**CU2**		**CU3**
			Remoto	Local	Remoto	Local	Remoto		
Config	Owl	-	Owl2			Owl2			Snowl
	Lupa	-	Sí			Sí			Sí
Sesión	Núm. Usuarios	[2-10]	2 Remoto 2 Local				1 Remoto 3 Local		1 Remoto 1 Local
	Velocidad Usuario	Caminando	Estático / caminando (Owl móvil)				Estático / caminando (Owl fijo)		Estático / caminando (Owl fijo)
	Distancia Usuarios	Cualquier lugar	Laboratorio (interior)				Piso tutelado (interior)		Piso formación (interior)
Tasa de datos	Vídeo	5-20 Mbps UL	6 Mbps				6 Mbps		6 Mbps
	Audio	0.1-0.5 Mbps UL/DL	100 kbps UL / 200 kbps DL				100 kbps UL / 100 kbps DL		100 kbps UL / 100 kbps DL
	Datos	0.2 Mbps DL	~100 kbps				~50 kbps		~50 kbps
Latencias	Latencia round-trip	< 600 ms	~0.5s				1s		~0.5s
	Tasa refresco XR	30 fps	30 fps				30 fps		30 fps
QoE	QoE MOS	> 3.5	4.6	4.7			4.2	4.8*	4.8
	Video MOS		3.1		N/A	3.6	N/A		4.3
	Audio MOS		4.6		4.6	3.9	4.8*		4.1
	Presencia espacial	Alta	5.9	N/A			6.8	N/A	6.9
	Presencia social	Muy alta	5.8	5.8			6.7	6.7*	6.8

	Presencia autónoma / QoI	Media	6.5	6.6	-	-	-
	Cibermareo	Mínimo (VSR > 4.5)	3.9	N/A	4.1	N/A	4.9

Conviene destacar que los resultados de las evaluaciones entre distintas pruebas no son totalmente comparables, ya que pueden verse afectadas por el contexto de evaluación. Otra diferencia es que las evaluaciones de CU2 y CU3 son longitudinales, hechas siempre por los mismos usuarios locales y remotos. Los valores numéricos de MOS y de VSR (cibermareo) están entre 1 y 5, siguiendo la recomendación ITU-T P.919, mientras que los valores de presencia y QoI están en una escala 1-7, como es habitual en las escalas psicométricas. Finalmente, los valores marcados con asterisco son evaluaciones realizadas por personas con discapacidad en escalas adaptadas (de tres niveles), que se han reescalado linealmente a los rangos estándar.

Se pueden extraer algunas conclusiones generales:

- La tecnología de teleportación virtual funciona de forma efectiva para uno o varios usuarios locales y remotos.
- El sistema cumple con los requisitos esperados de tasas de datos y latencias.
- La calidad de experiencia (QoE) es alta en sus distintos aspectos. Se observa una mejora significativa al pasar del primer prototipo (“Owl2”) al actual (“Snowl”).
- Los usuarios reportan niveles muy altos de presencia social y espacial. Son significativamente más altos en los usuarios de los casos de uso, lo que quiere decir que la presencia mejora con la implicación de la persona en el escenario y con el tiempo de uso.
- La calidad final alcanzada es buena (MOS >= 4) en vídeo, audio y QoE global. Se observa una mejora significativa con el cambio de prototipo del búho.
- Esta mejora en la calidad también tiene influencia en reducir la fatiga visual y, en última instancia, el cibermareo.

3 Escenario de Computación Delegada en MR: tecnologías de presencia autónoma e interacción natural

3.1 Tecnologías de presencia autónoma e interacción natural

3.1.1 Algoritmos de segmentación semántica para presencia autónoma

Este documento presenta una introducción a las diversas soluciones desarrolladas para la segmentación del cuerpo humano y los objetos de interés en entornos complejos. El objetivo principal es lograr una segmentación robusta y eficiente, capaz de distinguir no solo el cuerpo humano, sino también los objetos específicos con los que interactúa.

A lo largo del proyecto, se ha trabajado en varias aproximaciones al problema.

En primer lugar, en el entregable E3.1, se han partido de un algoritmo de segmentación por color (“croma”), basado en la crominancia roja (Cr), que emplea umbrales para detectar el color de la piel. El principio de funcionamiento de este algoritmo es que la piel humana tiene un tono con un alto componente de color rojo que puede diferenciarse con facilidad, habitualmente, de otros objetos. Este algoritmo tiene 3 variantes:

1. Croma. Segmentación de la piel por color, con umbrales estáticos
2. Adaptativo MP. La imagen se analiza con MediaPipe para detectar la forma y pose de la mano. A partir de ahí, se analiza el color de la mano y se establecen umbrales dinámicos para el valor de Cr. Estos umbrales se recalculan cada pocos segundos.
3. Adaptativo ML. Igual que el anterior, pero la mano se detecta usando segmentación semántica (ThunderNet) en vez de MediaPipe.

En segundo lugar, en el entregable E3.2, se ha estudiado el uso de segmentación semántica para segmentar tando el cuerpo del usuario como algunos objetos. Se han implementado varias posibles soluciones para realizar la segmentación del cuerpo junto con los objetos de interés. A continuación, se listan las soluciones propuestas:

1. YOLO-cutlery
2. ThunderNet + YOLO-cutlery
3. EgoHOS + YOLO-cutlery
4. Croma verde + YOLO-cutlery

Un elemento central y común a todas las soluciones presentadas es la red denominada "YOLO-cutlery". Esta es una adaptación especializada de la arquitectura YOLO, que ha sido reentrenada con un conjunto de datos específico. Este conjunto de datos incluye imágenes

que contienen objetos de uso común como platos, vasos/tazas, cuchillos, tenedores y cucharas, además de la segmentación de manos, lo que permite una detección y segmentación precisa de estos elementos clave.

Las demás soluciones se construyen sobre la base de YOLO-cutlery, integrando su máscara de segmentación con capacidades adicionales proporcionadas por las otras redes:

- **ThunderNet:** esta arquitectura se incorpora para extender la capacidad de segmentación al cuerpo humano completo, proporcionando un contexto más amplio de la escena.
- **EgoHOS:** esta red está específicamente entrenada para la segmentación detallada de manos y los objetos con los que estas interactúan, lo que es crucial para el análisis de acciones y manipulaciones.
- **Croma Verde:** esta solución aprovecha la técnica de segmentación de color, en concreto verde, para segmentar de manera eficiente cualquier elemento que se encuentre dentro de una zona cubierta por una tonalidad específica de color verde, ofreciendo una alternativa controlada para la segmentación de objetos.

Cada una de estas soluciones ofrece ventajas particulares y será detallada más a fondo en las siguientes secciones, explorando su implementación, rendimiento y las aplicaciones más adecuadas para cada una.

3.1.2 Tecnologías de interacción natural

La interacción natural se refiere a aquellos sistemas tecnológicos que permiten a los usuarios comunicarse con entornos digitales de manera similar a como interactúan con el mundo físico: mediante gestos, movimientos corporales, dirección de la mirada, voz o manipulación directa sin necesidad de dispositivos intermedios como mandos o teclados. Este paradigma busca reducir la fricción tecnológica y generar experiencias más accesibles, intuitivas y alineadas con las capacidades humanas, facilitando que la atención se centre en la tarea y no en el dispositivo. Un ejemplo cotidiano son los asistentes de voz.

Figura 8: Ejemplo de una interacción natural con un móvil

En el proyecto de Incluverso 5G, una de las interacciones naturales más directas ocurre en el caso de uso de teleformación en cocina, donde los usuarios pueden visualizar la realidad física a través de cámaras de las gafas de XR (realidad física mediada). En este caso de uso los/las participantes realizaron una serie de sesiones organizadas de la siguiente forma: Primero, se realizó una sesión de sensibilización para que los usuarios pudiesen percibir la realidad física tanto viendo vídeos completos 360 como la pura realidad física sin elementos extra. De esta forma los usuarios que nunca habían usado HMDs pudieron familiarizarse tanto con la visualización de vídeos 360 como con la visualización de la realidad física mediada.

La sesión de sensibilización comenzó con una introducción a los distintos niveles de mezcla entre realidades dentro del continuo de la Realidad Extendida. En un extremo se encuentra la realidad física pura y, en el opuesto, la realidad virtual completa, donde todo el entorno es generado digitalmente. Entre ambos se sitúan las modalidades intermedias que resultan especialmente relevantes para el caso de uso: la realidad aumentada, en la que elementos digitales se superponen al entorno físico, y la realidad mixta, donde esos elementos se integran espacialmente y mantienen coherencia con el mundo real

Los usuarios utilizaron el visor Meta Quest 3 en modo passthrough, que permite visualizar el entorno físico a través de las cámaras del dispositivo. Este primer contacto estaba orientado a que se familiarizaran con la percepción del mundo mediante la imagen mediada por el visor.

Figura 9:Ejemplo de visualización del entorno a través de las cámaras

Finalmente, los participantes visualizaron píldoras formativas mediante una pantalla virtual anclada en su entorno, manteniendo la posibilidad de ver sus manos reales gracias al hand tracking. Esta parte permitió mostrar cómo la realidad mixta puede utilizarse para consumir contenido formativo sin perder la referencia física.

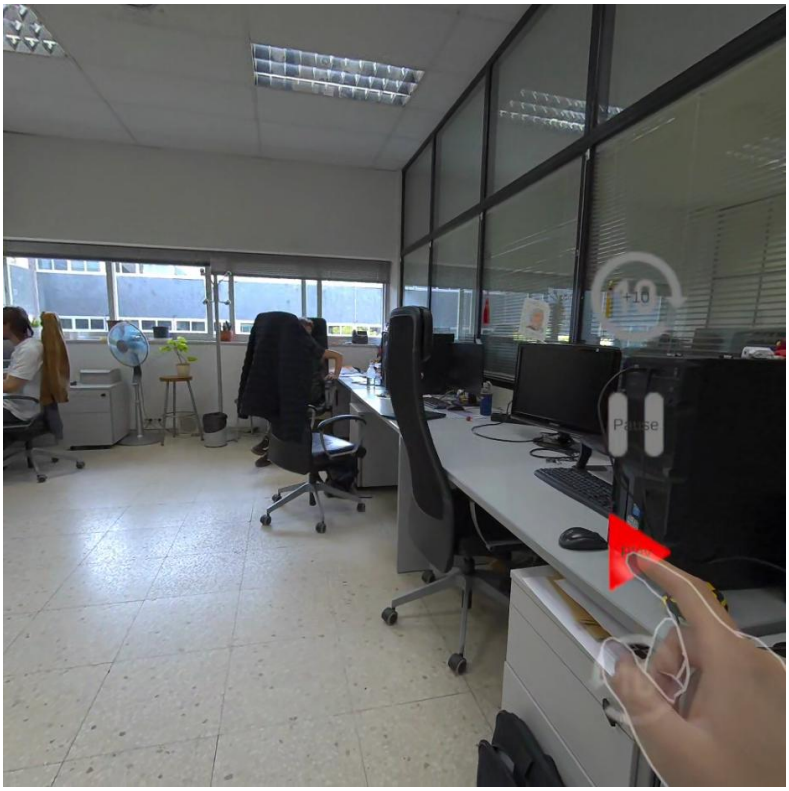


Figura 10: Visualización de vídeo con manos visibles en MR

Esta sesión estableció así una primera aproximación a la percepción del entorno mediante realidad mixta, sirviendo como base para las sesiones posteriores en las que se introducirían interacciones más avanzadas con elementos virtuales, descritas en apartados siguientes.

3.2 Caracterización coste-beneficio

3.2.1 Tecnologías de presencia autónoma

3.2.1.1 Croma adaptativo

En primer lugar, se realizó una prueba de calidad subjetiva de los algoritmos de segmentación basados en color descritos en el entregable E3.1. Con el fin de obtener métricas fiables, se solicitó a 20 personas distintas que visualizaran una serie de vídeos y respondieran a preguntas sobre ellos. El conjunto de datos de vídeo contiene 90 vídeos con una duración de 8 segundos. Estos vídeos muestran cuatro segundos del vídeo original, seguidos del resultado de la segmentación para ese mismo vídeo. Para cada vídeo original, se calculó un resultado de segmentación con cada uno de los tres algoritmos a evaluar (croma, croma adaptativo con MP y croma adaptativo con DL). En relación con la calidad de la segmentación, el objetivo principal con respecto al algoritmo de croma original era adaptarse a cualquier tipo de escenario. Esta adaptación se traduce en la capacidad de segmentar cualquier color de piel y en ser restrictivo frente a los colores de fondo para reducir el número de falsos positivos. Con este fin, en cada iteración el participante debe visualizar el vídeo y, posteriormente, responder a dos preguntas que permiten evaluar los factores mencionados:

1. Valore la percepción de sus manos y brazos. Se usa la escala ITU-T P.910 Absolute Category Rating (ACR).
2. Valore lo molesto que es el ruido de fondo (falsos positivos). Se usa la escala ITU-T P.910 Degradation Category Rating (DCR).

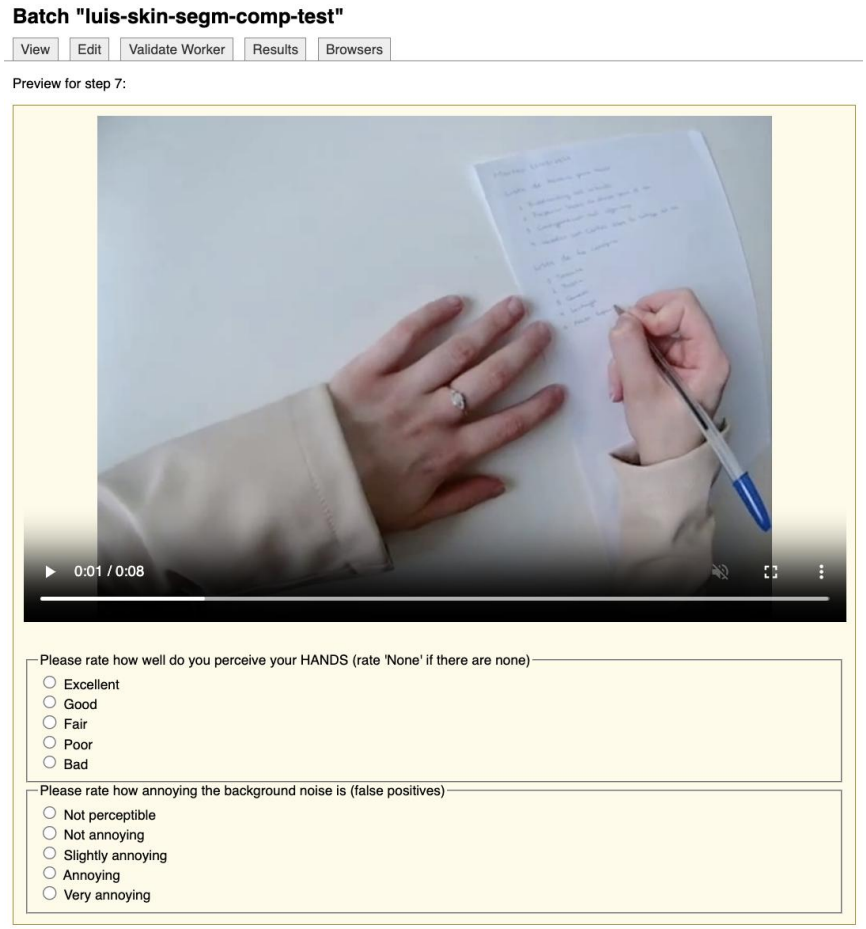


Figura 11: Herramienta de evaluación usada para la prueba

Para cada uno de los vídeos se calcula el MOS (Mean Opinion Score) como medida de calidad (ACR).y de molestia (DCR) expresada por los usuarios. Los resultados se agregan luego para evaluar el funcionamiento de los algoritmos en función de distintas propiedades de las secuencias, concretamente: el color de fondo, la acción realizada y el color de piel.

La siguente figura muestra las métricas MOS obtenidas por los diferentes modelos en relación con los seis tipos de color de piel definidos en la escala de Fitzpatrick. Primero se observa cómo los modelos adaptativos logran su propósito de segmentar cualquier tipo de color de piel, ya que el MOS obtenido es superior a 3 (Aceptable) para cada color de piel. En contraste, el método de croma original evidencia claramente su escasa capacidad para adaptarse a diferentes tonos de piel que están fuera de su rango de color aceptado, con mayores dificultades en las pieles más claras y más oscuras. Sin embargo, la segunda parte de la figura ilustra la limitación de la segmentación por color, que es el elevado número de píxeles de fondo molestos clasificados como piel. Los métodos adaptativos de segmentación por color solo reducen considerablemente el número de FP (falsos positivos) en el tipo de color de piel 6. Más adelante, esta cuestión se abordará con mayor detalle.

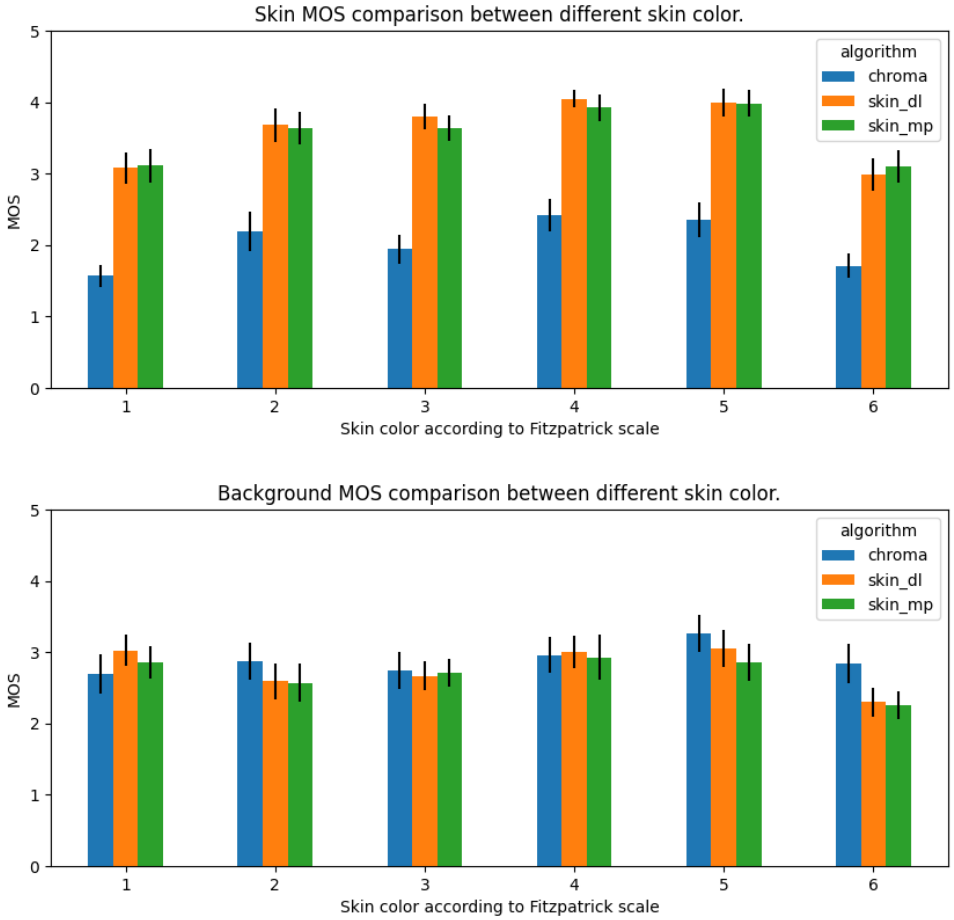


Figura 12: MOS de la piel (ACR, arriba) y de la molestia del fondo (DCR, abajo) para los distintos tonos de piel según la escala de Fitzpatrick.

Posteriormente se ha analizado el MOS para los distintos tipos de fondo que aparecen en los vídeos. Como en el análisis anterior, los métodos adaptativos obtienen mejores resultados.

El fondo blanco es el que proporciona el mejor resultado, ya que en esos vídeos el color de la piel suele ser claramente diferenciable del fondo. El MOS del fondo es inferior en las secuencias grabadas sobre la mesa de madera porque el color de la madera es muy similar al de la piel. Todos estos algoritmos presentan dificultades en los vídeos con fondo de suelo, como se refleja en las métricas MOS para la piel y el fondo. En esas secuencias, el usuario se mueve hacia una ventana, por lo que la luminosidad fluctúa constantemente y la piel del usuario está muy saturada en múltiples momentos, lo cual representa el peor escenario para el algoritmo de piel adaptativa.

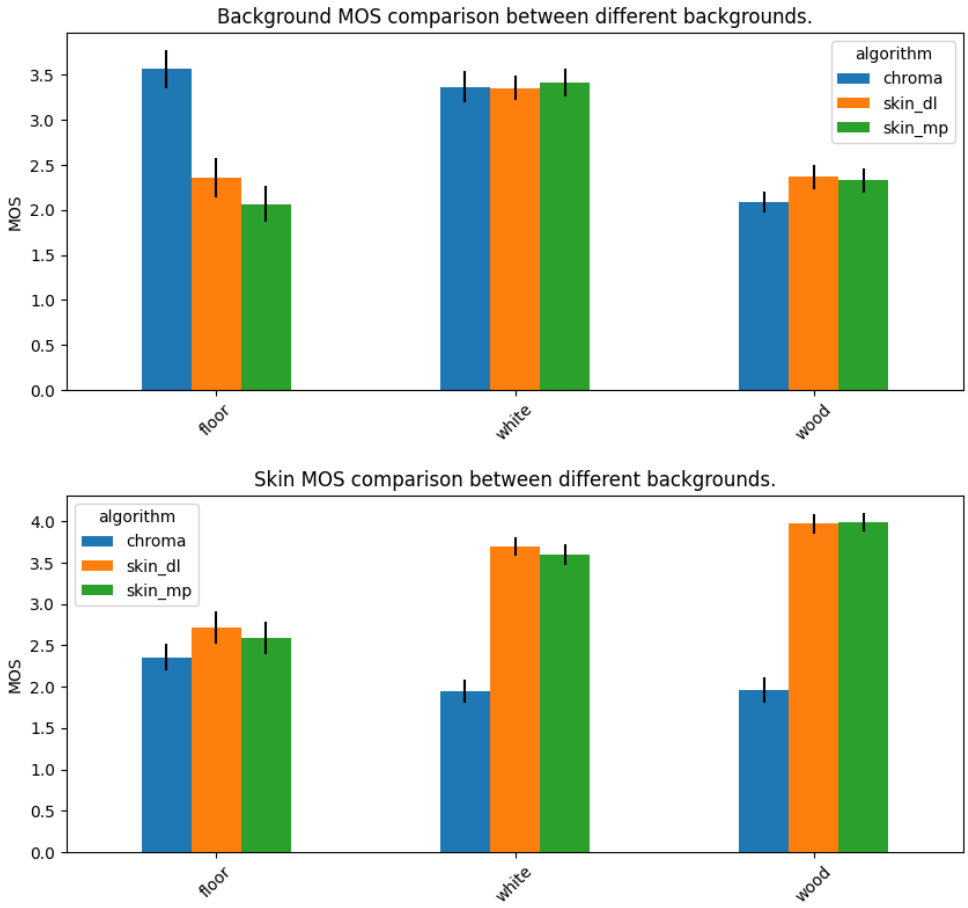


Figura 13: MOS de la piel (ACR, arriba) y de la molestia del fondo (DCR, abajo) para los distintos tonos de piel según la escala de Fitzpatrick.

Finalmente, la siguiente figura muestra la comparación atendiendo a diferentes pares de fondo y acción. El escenario que resulta en una mejor calidad es aquel en el que el usuario escribe con un bolígrafo sobre una mesa blanca, ya que contiene un color de fondo claramente diferente. Los vídeos con fondo de madera ofrecen peores resultados que la mesa blanca, ya que contienen un mayor número de FP. Las secuencias en las que el usuario interactúa con un ordenador también conducen a peores resultados en comparación con aquellas en las que el usuario utiliza el bolígrafo. Esto se debe a que la pantalla del portátil contiene un color similar al de la piel.

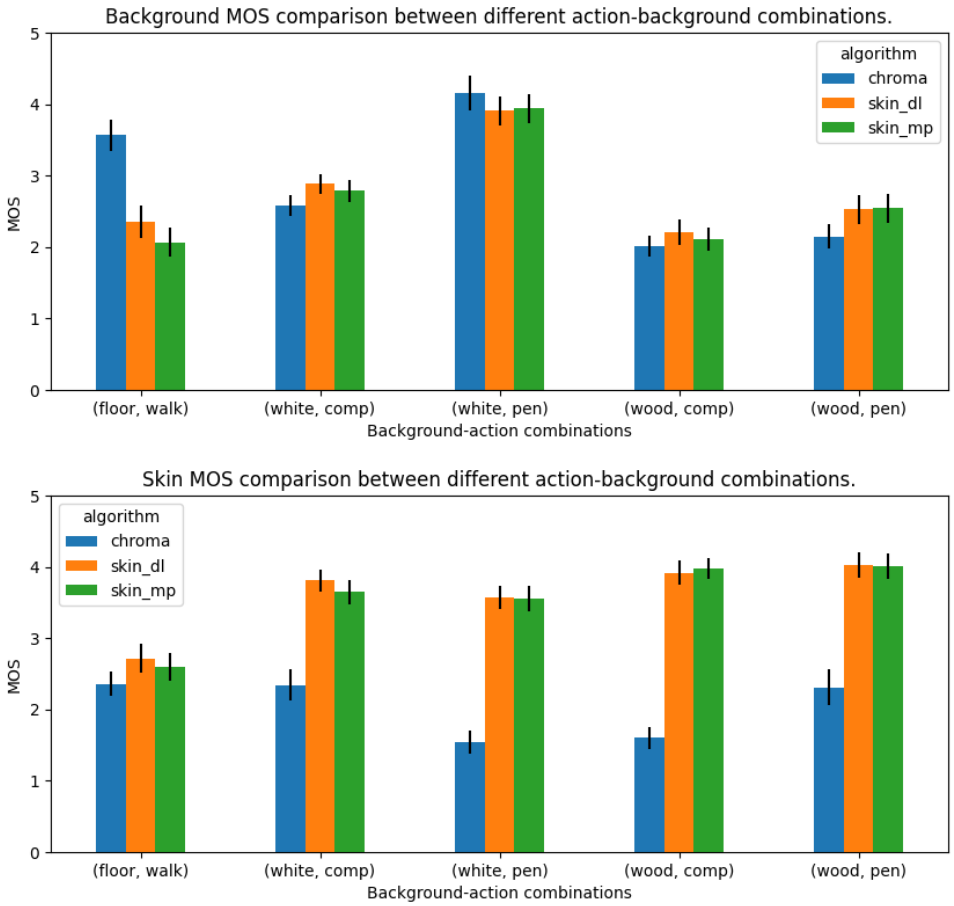


Figura 14: MOS de diferentes pares de fondo y acción

3.2.1.2 Segmentación semántica

Los algoritmos de segmentación semántica (tanto del cuerpo como de objetos) necesitan de un servidor con GPU para llevar a cabo la inferencia en tiempo real, por lo que un despliegue en una red 5G podría utilizar un servidor MEC en el backend para poder ejecutar en él las redes de inteligencia artificial.

El entrenamiento de cada una de las redes es un proceso que se realiza previo a su despliegue, se ha utilizado para ello un servidor que dispone de una GPU NVidia GTX 3090 de 25GB de memoria con versión de CUDA 12.0 que tiene acceso a las bases de datos utilizadas para entrenar. Es un proceso de varias horas, pero que sólo se realiza antes del despliegue.

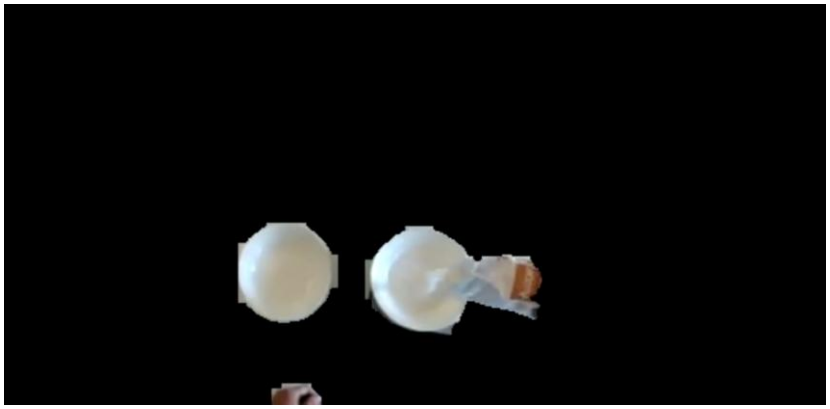
En términos de funcionamiento de los algoritmos en detección, se plantea una evaluación de las distintas metodologías presentadas en el apartado anterior con los vídeos grabados en la Fundación Juan XXIII Roncalli y diverso instrumental de cocina. Concretamente, los algoritmos de segmentación evaluados son:

- YOLOv8: diseñado para segmentar personas, cubiertos, platos y vasos.

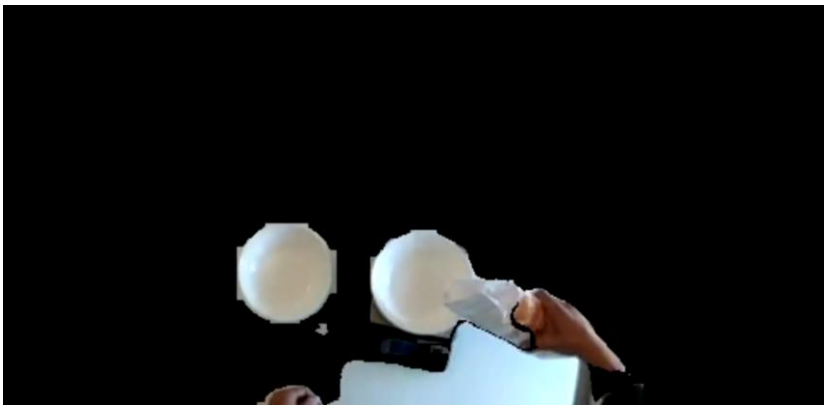
- YOLO-table + ThunderNet: Se ha combinado YOLO con ThunderNet. Este segundo algoritmo está diseñado para segmentar partes del cuerpo humano.
- YOLO-table + EgoHOS: Se ha combinado YOLO con EGOHOS. El segundo algoritmo segmenta las manos y cualquier objeto que está en contacto con él.

Respecto a los vídeos con instrumental de cocina grabados en la Fundación Juan XXIII Roncalli, debemos destacar que en estos vídeos aparecen muchos más instrumentos que los considerados en el entrenamiento de YOLO-cutlery, por lo que no vamos a poder observarlos todos dentro de las máscaras de segmentación calculadas. Sin embargo, consideramos que estas primeras metodologías sí puede ser un buen punto de partida para entender la mejor forma de abordar el problema de una segmentación más compleja como la de estos ejemplos y caso de uso. Presentamos algunos ejemplos en la Figura 15 y en la Figura 16.

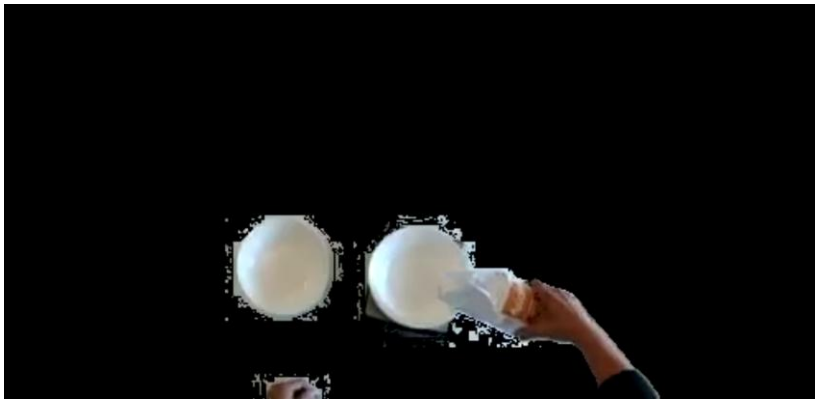
(a) Cuadro original



(b) Máscara generada con YOLO-cutlery



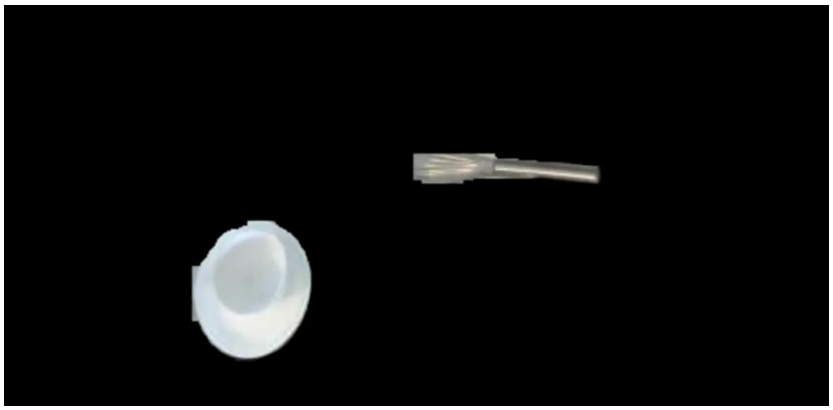
(c) Máscara generada con YOLO-cutlery + ThunderNet



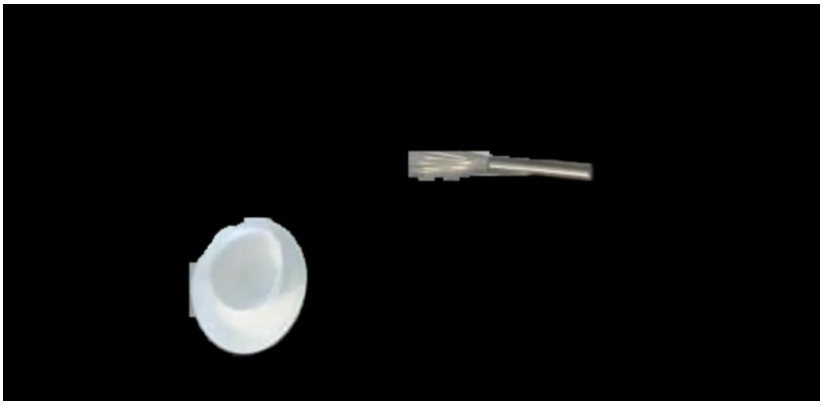
(d) Máscara generada con YOLO-cutlery + EgoHOS

Figura 15: Primer ejemplo de máscaras generadas para vídeos con instrumental de cocina en la Fundación Juan XXIII Rocanlli

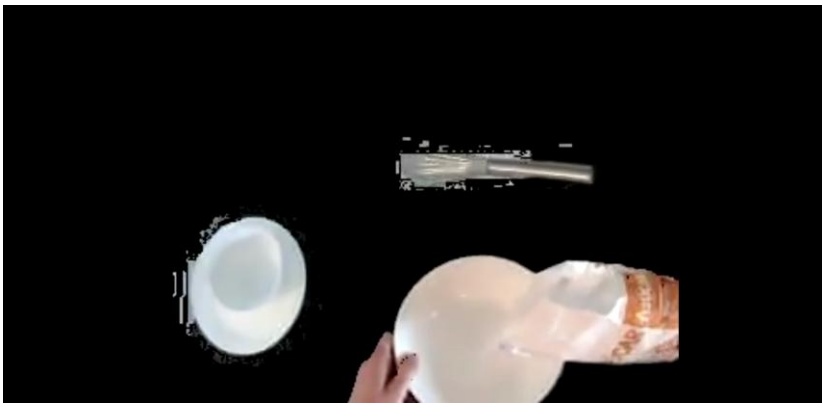
(a) Cuadro original



(b) Máscara generada con YOLO-cutlery



(c) Máscara generada con YOLO-cutlery + ThunderNet



(d) Máscara generada con YOLO-cutlery + EgoHOS

Figura 16: Segundo ejemplo de máscaras generadas para vídeos con instrumental de cocina en la Fundación Juan XXIII Roncalli

En estos ejemplos podemos ver cómo varios objetos no se logran segmentar ya que no forman parte del conjunto que utilizamos para entrenar nuestra red. Aun así, en muchas ocasiones podemos ver una gran parte de los objetos de interés, debido a que se parecen a

platos o cubiertos que la red sí identifica y es capaz de detectar. El uso de EgoHOS ayuda a ver cualquier instrumental que esté en contacto con las manos, esperando que mejor mucho la sensación bajo la premisa de estar sujetándolo.

Para generar la base de datos para evaluar el caso de uso, se han grabado nueve vídeos en los que aparecen brazos/manos, platos, cubiertos y vasos/tazas, que son los objetos que se pueden segmentar. El hecho de aparecer otros objetos no clasificables por los algoritmos de segmentación también le da realismo al experimento y mejora la validez ecológica de nuestro experimento, entendida como las condiciones del experimento que hacen que el mismo sea más similar a una experiencia real y, por tanto, aumente la robustez de las conclusiones y su fiabilidad. El motivo es que posiblemente en una cocina real haya objetos que los algoritmos no son capaces de detectar, situación bajo la cual consideramos interesante testear nuestras implementaciones. Además, decidimos considerar vídeos en los que se realiza una acción (por ejemplo, coger una taza) y en otros no hay movimiento, explicados posteriormente con mayor detalle. Por último, las capturas se han realizado en diferentes entornos para incrementar la diversidad de los vídeos contemplados. La Figura 17 presenta un ejemplo de un momento de captura de los vídeos utilizados en este experimento.

Figura 17: Entorno de grabación de los vídeos

Para tener una métrica objetiva del funcionamiento de los algoritmos analizados sobre los vídeos grabados se realiza un cálculo del valor de *Intersection Over Union* (IoU). Esta métrica mide el ratio entre la intersección y la unión entre el *groundtruth* y el resultado de la inferencia, es una medida entre 0 y 1, con 1 indicando una coincidencia perfecta.

Para realizar el cálculo, se realiza un muestro de siete cuadros a lo largo de cada uno de los vídeos originales que se han etiquetado manualmente con ayuda del algoritmo SAM [7] y se compara el resultado con el obtenido con cada uno de los métodos. Este etiquetado con el algoritmo SAM sirve para poder distinguir los objetos presentes en los vídeos que los diferentes algoritmos deberían de segmentar del resto de objetos que aparecen en la

escena, tal y como el ejemplo que se presenta en la Figura 18. Todos los objetos que interesan – en este caso los platos, manos, vasos y cubiertos – se etiquetan como “mask”. Esta selección se guarda en un JSON, a partir del cual se obtienen las máscaras del *groundtruth*.

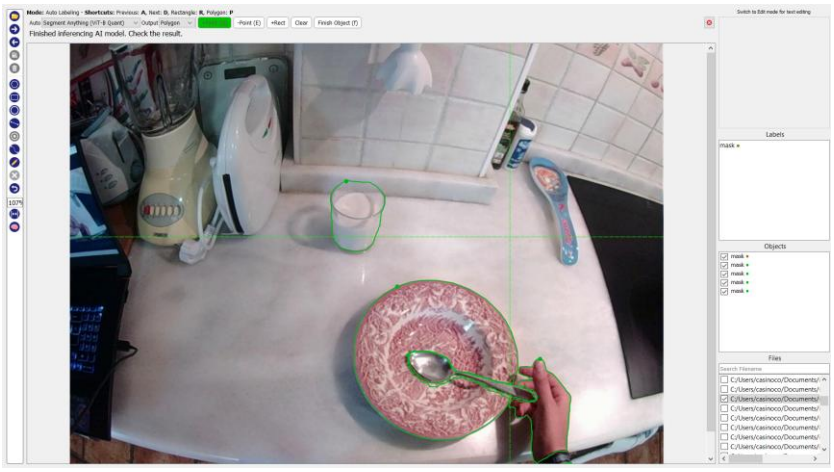


Figura 18: Ejemplo del etiquetado manual de un cuadro

Una vez que los vídeos originales están etiquetados, se aplican los algoritmos de segmentación para poder hacer la comparación entre las salidas. La muestra el resultado de la máscara y el borde etiquetados manualmente (groundtruth y groundtruth) y el resultado sobre el mismo cuadro ejemplo de los tres algoritmos: YOLO-cutlery, YOLO-cutlery + ThunderNet y egoHOS.

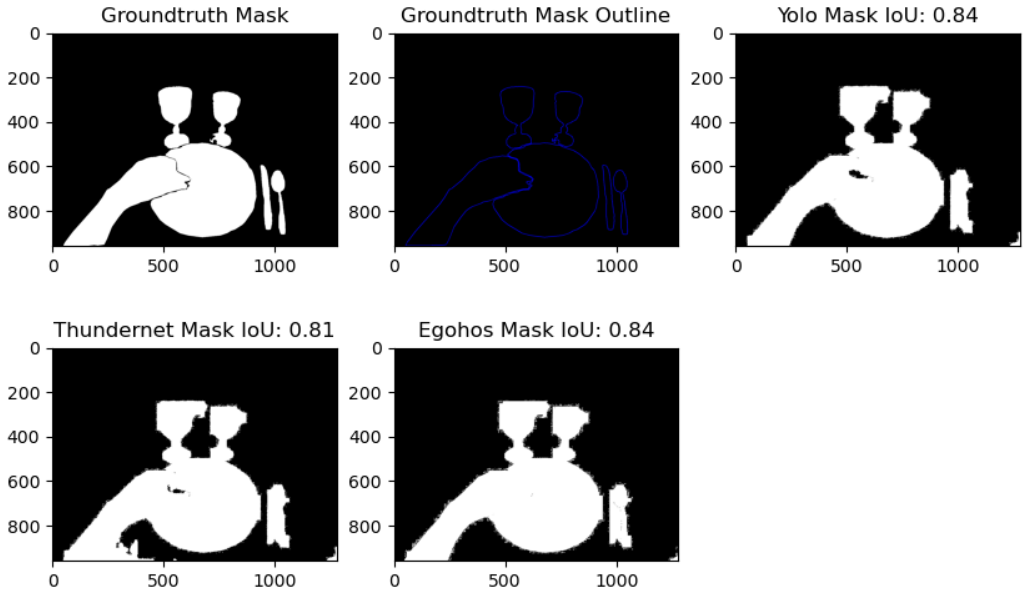


Figura 19: Ejemplo de las máscaras generadas para el cálculo de la IoU

Los resultados promedio considerando los ocho vídeos de cubertería se recogen en la Tabla 4. Como vemos, los resultados son muy semejantes entre YOLO-cutlery y YOLO-cutlery + ThunderNet, a pesar de que la sensación subjetiva inicial era de que incluyendo ThunderNet lográbamos mejorar la segmentación del cuerpo humano. Sin embargo, los valores de IoU mejoran algo más al utilizar EgoHOS.

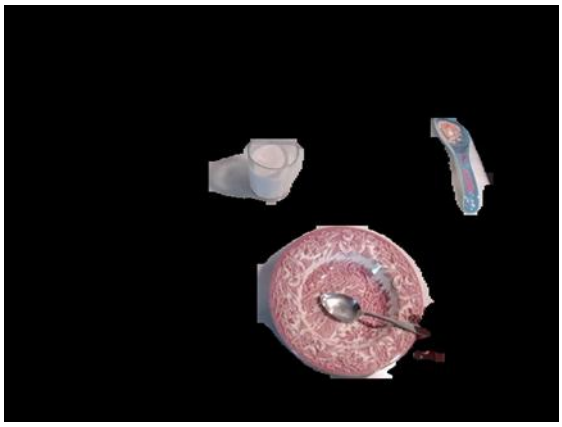
Tabla 4: Resultados de IoU de las distintas metodologías con vídeos de cubertería

	IoU
YOLO-cutlery	0.72
YOLO-cutlery + ThunderNet	0.7
YOLO-cutlery + EgoHOS	0.79

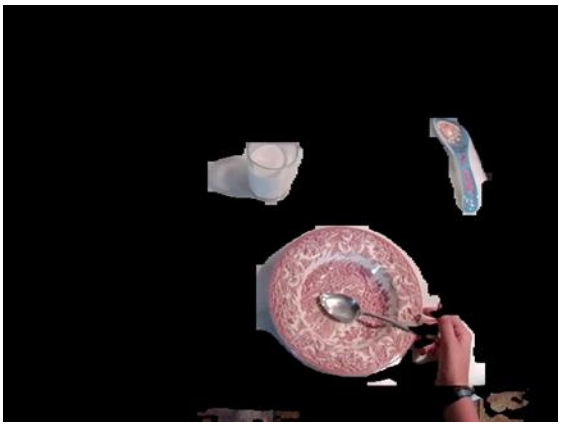
En las Figura 20, Figura 21 y Figura 22 se muestran un cuadro de algunos de los vídeos grabados junto con las máscaras generadas por cada uno de los tres algoritmos que se evalúan.

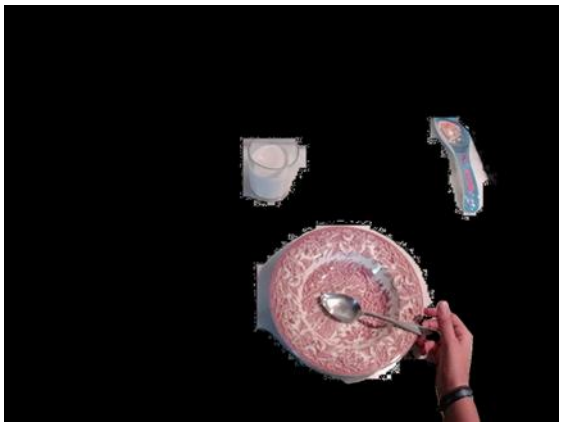
QoE_5:

(a) Original



(b) YOLO

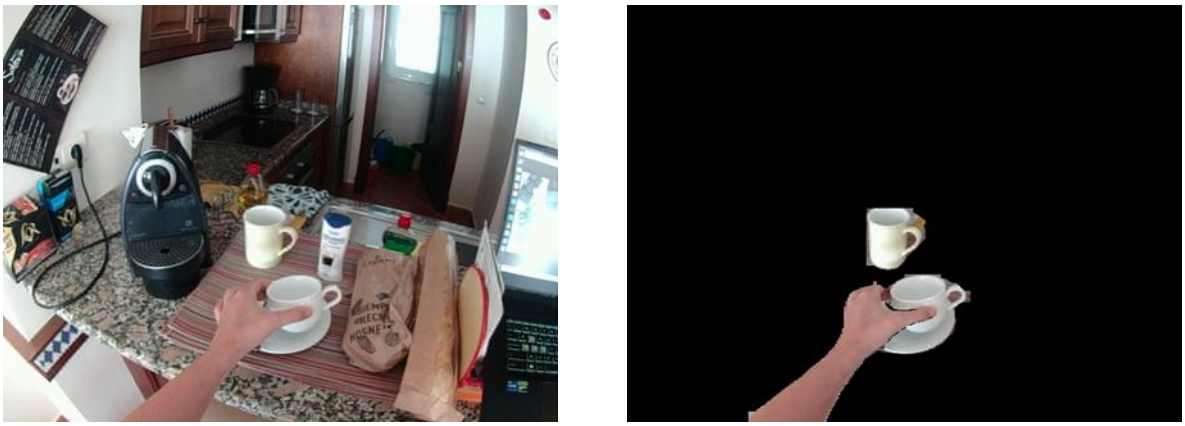




(c) YOLO-ThunderNet (d) YOLO-EgoHOS

Figura 20: Ejemplo del cuadro original y máscaras del mismo del vídeo QoE_5

QoE_3:

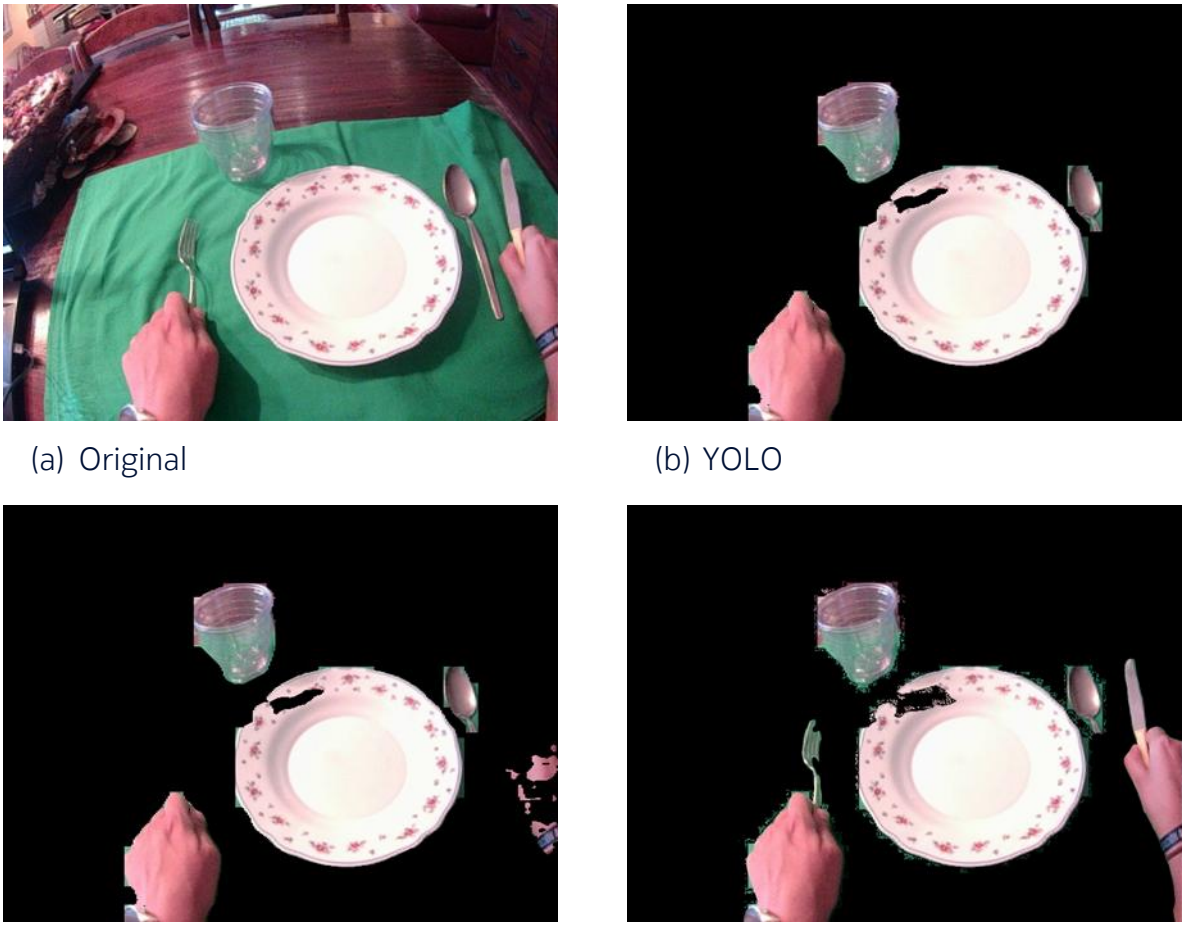


(a) Original (b) YOLO

(c) YOLO+ThunderNet (d) YOLO+EgoHOS

Figura 21: Ejemplo del cuadro original y máscaras del mismo del vídeo QoE_3

QoE_10:



(a) Original (b) YOLO

(c) YOLO-ThunderNet (d) YOLO-EgoHOS

Figura 22: Ejemplo del cuadro original y máscaras del mismo del vídeo QoE_10

En este punto observamos que, aunque las métricas objetivas proporcionan información valiosa sobre el rendimiento de nuestras tecnologías, no consideramos que sean decisivas para evaluar completamente su efectividad. Por esta razón, proponemos realizar una evaluación subjetiva, un enfoque que ha sido poco explorado en la literatura existente. Creemos que esta evaluación es fundamental para comprender la calidad de la experiencia percibida por las personas, así como para identificar limitaciones y áreas de mejora que las métricas objetivas no logran revelar. Al incorporar la perspectiva subjetiva de las personas que van a utilizar la tecnología, esperamos obtener una visión más holística y precisa del impacto que puede tener en su aplicación en la vida real.

En concreto, el objetivo fundamental del experimento es que las personas participantes en el test subjetivo evalúen las máscaras de segmentación generadas con nuestras distintas

implementaciones, algoritmos de segmentación, en vídeos pregrabados, utilizando para ello una aplicación en un servidor web.

De los vídeos grabados, ocho se utilizan para realizar la evaluación, mientras que el vídeo restante, presentado en la Figura 23, se utiliza en la fase de entrenamiento, es decir, para mostrar a las personas participantes el resultado ideal u objetivo que se pretende conseguir con los algoritmos de segmentación. El algoritmo de segmentación seleccionado para generar este vídeo es el *Chroma*. Concretamente, este algoritmo se ha ajustado para que segmente objetos teniendo en cuenta los colores rojo, naranja y amarillo, seleccionados porque en el vídeo aparecen manos/brazos, platos, vasos y cubiertos de color amarillo/naranja/rojo.

Figura 23: Cuadro del vídeo capturado para ser utilizado en la fase de entrenamiento

En cuanto a los ocho vídeos cuyas máscaras de los algoritmos de segmentación van a ser evaluadas, se clasifican en las cuatro categorías:

Dos vídeos en los que no se realiza ninguna acción:

Un primer vídeo en el que aparece una mano, cubiertos, platos y vasos presentado en la Figura 24.

Figura 24: Cuadro del vídeo QoE_2

Un segundo vídeo en el que aparecen cubiertos, varios platos y vasos presentado en la Figura 25.

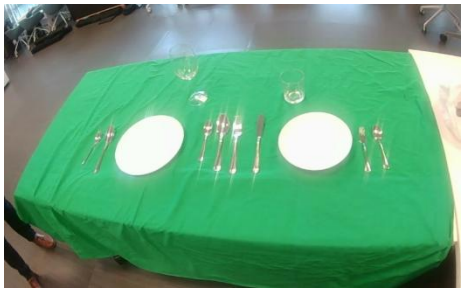


Figura 25: Cuadro del vídeo QoE_8

Dos vídeos en los que se lleva a cabo una acción con la mano derecha:

Un primer vídeo en el que se mueve una cuchara con la mano derecha presentado en la Figura 26.

Figura 26: Cuadro del vídeo QoE_5

Un segundo vídeo en el que se coge una taza con la mano derecha presentado en la Figura 27.

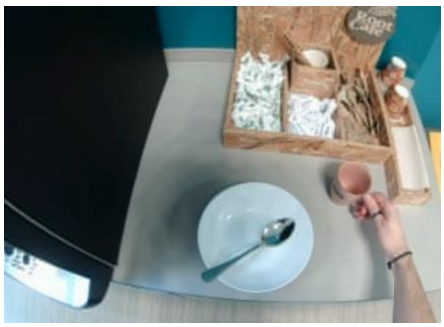


Figura 27: Cuadro del vídeo QoE_12

Dos vídeos en los que se lleva a cabo una acción con la mano izquierda:

Un primer vídeo en el que se coge un cubierto con la mano izquierda presentado en la Figura 28 .

Figura 28: Cuadro del vídeo QoE_7

Un segundo vídeo en el que se coge una taza con la mano izquierda presentado en la Figura 29.

Figura 29: Cuadro del vídeo QoE_3

Dos vídeos en los que se llevan a cabo acciones con las dos manos:

Un primer vídeo en el que se cogen cubiertos con las dos manos presentado en la Figura 30.

Figura 30: Cuadro del vídeo QoE_10

Un segundo vídeo en el que se coge un cubierto y un vaso presentado en la Figura 31.

Figura 31: Cuadro del vídeo QoE_14

Respecto a las características específicas de cada vídeo original, se presentan en la Tabla 5. Estos ocho vídeos son los vídeos fuente (SRCS) que se utilizan para generar los 24 vídeos procesados (PVSs) generados al aplicar el resultado de la segmentación de los 3 algoritmos contemplados en el experimento.

Tabla 5: Características técnicas de la base de datos de vídeos generadas para las pruebas

Vídeo	Duración	Resolución	Características	fps	Tasa binaria
QoE_2	6s	1280x960	No acción	56.4	5033kbps
QoE_3	6s	1280x960	Taza mano izquierda	56.4	7534kbps
QoE_5	5s	1280x960	Cuchara meno derecha	56.4	3028kbps
QoE_7	6s	1280x960	Cubierto mano izquierda	56.4	9759kbps
QoE_8	4s	1280x960	No acción	56.4	2224kbps
QoE_10	5s	1280x960	Cubiertos con las dos manos	56.4	9225kbps
QoE_12	5s	1280x960	Taza con la mano derecha	55.7	4583kbps
QoE_14	7s	1280x960	Cubierto y vaso (dos manos)	56.4	5306kbps

Por otro lado, se presentan los resultados de las métricas objetivas obtenidas para cada uno de los vídeos en la Tabla 6.

Tabla 6: IoU de cada uno de los vídeos procesado con los tres algoritmos de segmentación y promedio

Vídeo	IoU YOLO-Cultery	IoU YOLO-Cultery + EgoHOS	IoU YOLO-Cultery + ThunderNet	IoU Promedio
QoE_2	0.7	0.75	0.56	0.67
QoE_3	0.76	0.79	0.76	0.77
QoE_5	0.51	0.8	0.8	0.63
QoE_7	0.78	0.83	0.78	0.8
QoE_8	0.56	0.54	0.54	0.56
QoE_10	0.86	0.89	0.84	0.86
QoE_12	0.72	0.81	0.72	0.75
QoE_14	0.85	0.88	0.84	0.86

Una vez que se creó la base de datos de vídeos que van a ser utilizados como estímulos, se diseñó una sesión de experimento, con tres fases principales presentadas en la Figura 32, que tiene como objetivo ser replicable para poder comparar los resultados entre diferentes personas. La replicabilidad es crucial para asegurar la validez de los datos obtenidos, permitiendo una evaluación consistente de los algoritmos de segmentación. Este proceso estructurado garantiza que todas las personas que participaron lo hicieron siguiendo los mismos pasos y procedimientos, facilitando la comparación y análisis de los resultados. Asimismo, se garantiza que los participantes comprendan plenamente el experimento, proporcionen la información necesaria de manera segura y evalúen los algoritmos de segmentación de manera efectiva. La recopilación de datos se realizará de forma ordenada y conforme a las normativas de privacidad, contribuyendo al éxito del estudio.

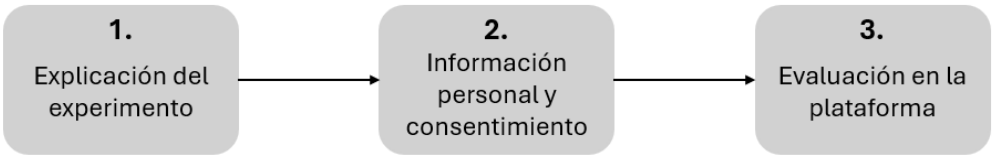


Figura 32: Fases de la sesión de experimento

1. Explicación del Experimento

En esta primera etapa, se proporciona a las personas participantes una explicación detallada del experimento en la hoja de información presentada en el Apéndice A. El objetivo es que los/as participantes tengan una idea general de lo que van a evaluar. En concreto, se abordan los siguientes puntos:

- Objetivo del experimento.
- Procedimiento a seguir.
- Importancia de su participación y cómo sus respuestas contribuirán al estudio.

2. Formulario de Información Personal y Consentimiento

Una vez comprendido el experimento, las personas participantes procederán a rellenar dos formularios:

- Formulario de Información Personal: las personas participantes deberán proporcionar información personal utilizando el mismo ID que emplearán para realizar la evaluación en la plataforma. En concreto, se recopila: género, edad, visión (corregida o normal), experiencia con la tecnología y educación.
- Consentimiento de Tratamiento de Datos Personales: se solicitará a los participantes que otorguen su consentimiento para el tratamiento de sus datos personales, asegurando el cumplimiento de las normativas de privacidad y protección de datos estipulados por el Reglamento General de Protección de Datos (RGPD) (UE, 2016/679).

En esta primera fase también se aplica el test Ishihara para comprobar la visualización de los colores y la tabla de Snellen para detectar cualquier posible problema de visión que imposibilite la realización de la prueba.

3. Evaluación de los Algoritmos de Segmentación

La última etapa consiste en la evaluación de los algoritmos de segmentación en la plataforma web diseñada utilizando Quality Crowd 2.

Los pasos a seguir son:

Acceso a la Plataforma: Las personas participantes accederán a la plataforma ../kitchenutensils/.. donde se llevará a cabo la evaluación.

1 Evaluación de Vídeos:
   - Los primeros apartados de la encuesta muestran las instrucciones del experimento.
   - Seguidamente, se muestra el vídeo de entrenamiento, para que vean lo que idealmente se pretende conseguir con los algoritmos de segmentación.
   - En este momento ya se evaluarán los 24 vídeos presentados de manera aleatoria y en ningún momento se presentarán dos vídeos obtenidos de la misma fuente de manera consecutiva, siguiendo la Rec. ITU-R BT-500-13. No hay un límite de tiempo estricto para esta fase, ya que el tiempo límite establecido es de más de una hora por vídeo, permitiendo a los participantes tomarse el tiempo necesario para completar la evaluación. Cada vídeo va seguido de tres preguntas para su evaluación, tal y como se presenta en la Figura 33.

Figura 33: Ejemplo de la pantalla de la plataforma web sobre la que se realiza la evaluación de los vídeos y de las preguntas realizadas

La escala utilizada para la evaluación sigue la recomendación de la Rec. ITU-T P.910. Se ha utilizado el método Absolute Category Rating (ACR) para la pregunta sobre la calidad. Por otro lado, para la evaluación de la pregunta sobre la molestia, se ha utilizado el método Degradiation Category Rating (DCR).

Un total de 20 personas participaron en el experimento. La diversidad de la muestra recogida en el formulario de información personal puede verse en la Figura 34, Figura 35 y la Figura 36.

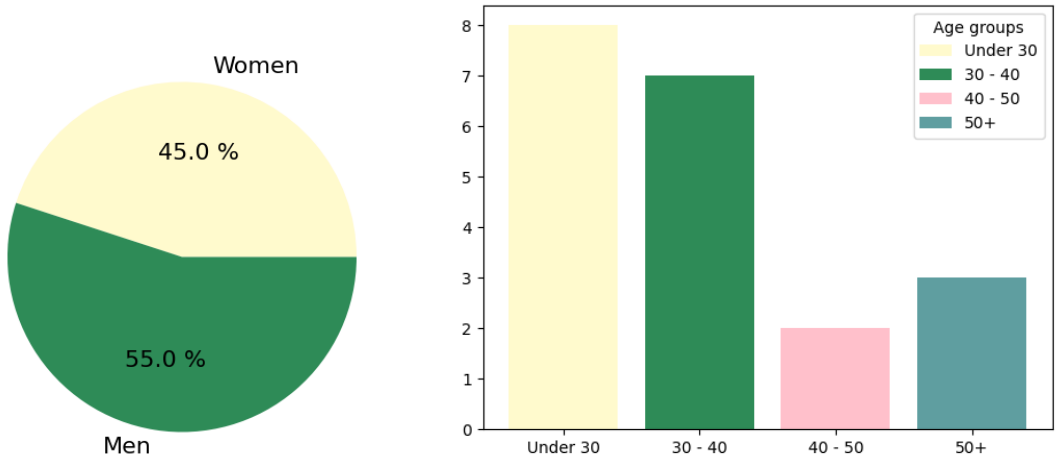


Figura 34: Género y edad de la muestra de participantes

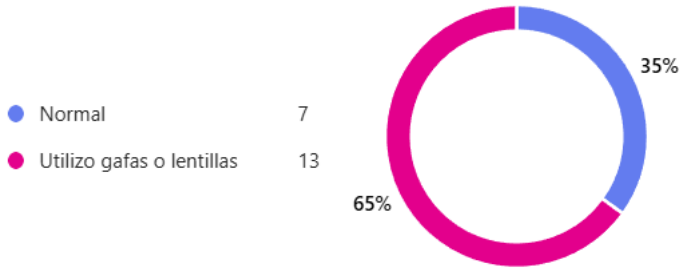


Figura 35: Información de la visión de la muestra de participantes

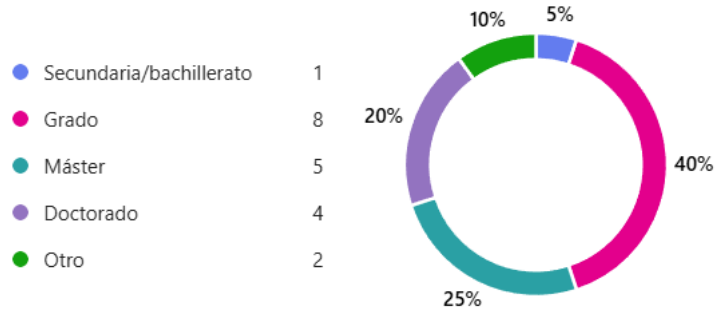


Figura 36: Nivel de educación de la muestra de participantes

Análisis de resultados

Para analizar los resultados obtenidos de la evaluación subjetiva, los ficheros *logs* obtenidos de la plataforma de QualityCrowd2 tienen el formato de la Figura 37.

```
0,1731491642
1,1731491670
2,1731491715,1
3,1731491718
4,1731491754,1,1,Regular,3,1,SÃ,1,1,"Perceptible pero no molesto",4
5,1731491866,1,1,Buena,4,1,No,2,1,"Perceptible pero no molesto",4
6,1731491900,1,1,Pobre,2,1,No,2,1,"Perceptible pero no molesto",4
7,1731491970,1,1,Regular,3,1,SÃ,1,1,"Perceptible pero no molesto",4
8,1731491999,1,1,Pobre,2,1,SÃ,1,1,"Perceptible pero no molesto",4
9,1731492033,1,1,Pobre,2,1,No,2,1,"Perceptible pero no molesto",4
10,1731492062,1,1,Mala,1,1,SÃ,1,1,"Perceptible pero no molesto",4
11,1731492088,1,1,Regular,3,1,SÃ,1,1,"Perceptible pero no molesto",4
12,1731492135,1,1,Pobre,2,1,SÃ,1,1,Imperceptible,5
13,1731492162,1,1,Regular,3,1,No,2,1,"Algo molesto",3
14,1731492189,1,1,Regular,3,1,SÃ,1,1,"Perceptible pero no molesto",4
15,1731492220,1,1,Regular,3,1,No,2,1,"Algo molesto",3
16,1731492291,1,1,Regular,3,1,SÃ,1,1,"Perceptible pero no molesto",4
17,1731492317,1,1,Regular,3,1,SÃ,1,1,"Perceptible pero no molesto",4
18,1731492349,1,1,Buena,4,1,No,2,1,"Perceptible pero no molesto",4
19,1731492382,1,1,Regular,3,1,SÃ,1,1,"Perceptible pero no molesto",4
20,1731492426,1,1,Regular,3,1,No,2,1,"Perceptible pero no molesto",4
21,1731492456,1,1,Buena,4,1,No,2,1,"Perceptible pero no molesto",4
22,1731492499,1,1,Regular,3,0,,1,"Perceptible pero no molesto",4
23,1731492521,1,1,Pobre,2,1,SÃ,1,1,"Algo molesto",3
24,1731492555,1,1,Mala,1,1,SÃ,1,1,"Algo molesto",3
25,1731492580,1,1,Pobre,2,1,SÃ,1,1,"Algo molesto",3
26,1731492608,1,1,Pobre,2,1,SÃ,1,1,"Algo molesto",3
27,1731492638,1,1,Mala,1,1,SÃ,1,1,"Algo molesto",3
```

Figura 37: Ejemplo ficheros logs obtenidos de la plataforma de QualityCrowd2

En ellos se presenta el marcador de tiempo, y también las respuestas a las diferentes preguntas. Después de organizar la base de datos, teniendo en cuenta las respuestas de todos los encuestados, esta se presenta como en la Figura 38.

	Participant	Video_name	Quality	Tiempo visualizacion (s)	Discomfort	Video	Algorithm	Action	Timestamp_relativo
0	1	QoE_3_complete_yolo_thunder.mp4	3	36	4	QoE_3	yolo_thunder	left hand	2.571429
1	1	QoE_14_complete_yolo_egohos.mp4	4	112	0	QoE_14	yolo_egohos	both hands	5.090909
2	1	QoE_2_complete_yolo_pred.mp4	2	34	0	QoE_2	yolo_pred	no action	2.833333
3	1	QoE_12_complete_yolo_egohos.mp4	3	70	4	QoE_12	yolo_egohos	right hand	5.384615
4	1	QoE_12_complete_yolo_pred.mp4	2	29	4	QoE_12	yolo_pred	right hand	2.230769

Figura 38: Tabla de datos recopilados organizada

Para analizar los resultados obtenidos respecto a la **calidad percibida con los vídeos segmentados**, en primer lugar, se han obtenido la evaluación promedia de las evaluaciones (MOS), agrupando los diferentes vídeos según el algoritmo (ver Figura 39).

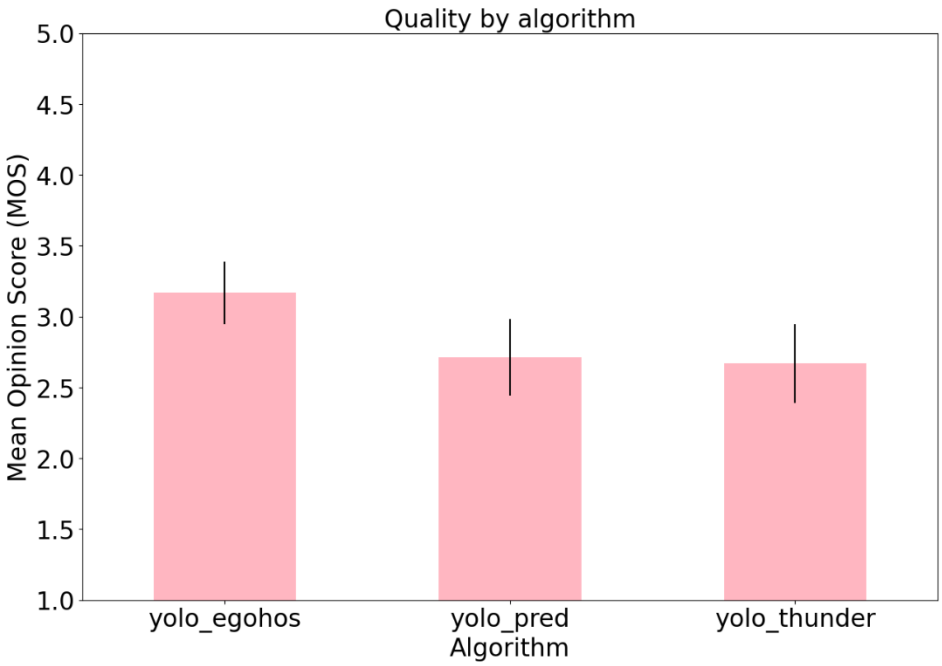


Figura 39: MOS de la calidad según el algoritmo

Según estos resultados, YOLO+EgoHOS es el algoritmo preferido, coincidiendo con los resultados obtenidos en la evaluación objetiva. Por otro lado, se han obtenido los MOS de la calidad para cada uno de los vídeos, según el algoritmo (ver Figura 40).

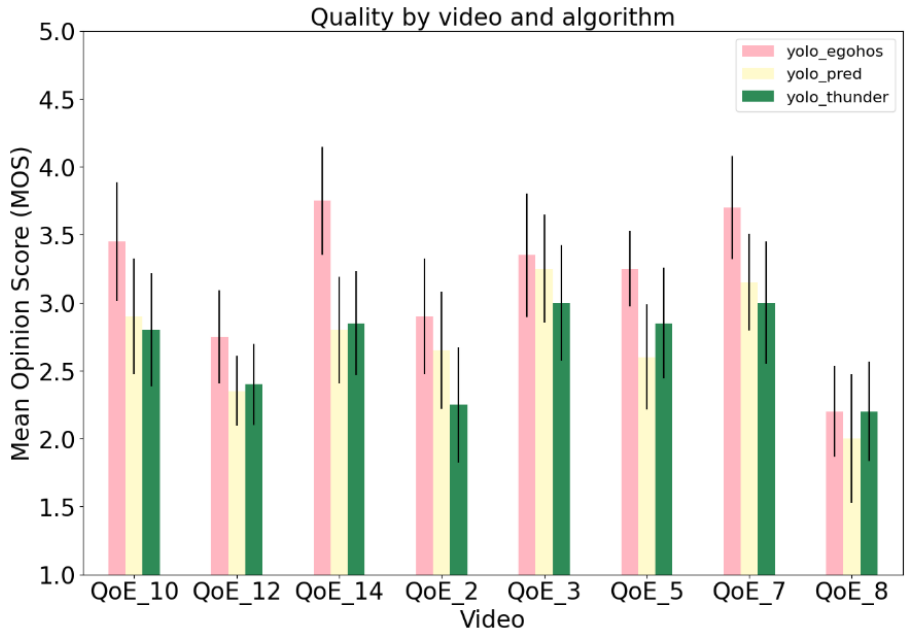


Figura 40: MOS de la calidad para cada vídeo según el algoritmo

De estos resultados, se obtiene que, para cada vídeo, el algoritmo preferido por las personas participantes también es YOLO+EgoHOS. Para explorar mejor estos resultados, dada la distribución normal de los datos siguiendo el test D'Agostino and Pearson, se aplica un test ANOVA, resultando que hay diferencia significativa entre los datos. Concretamente, aplicando un análisis post-hoc Tukey con corrección de Bonferroni, se obtienen diferencias significativas entre EgoHOS y los otros dos algoritmos (ver Figura 41).

	sum_sq	df	F	PR(>F)	eta_sq	omega_sq
C(Algorithm)	24.5375	2.0	13.160979	0.000003	0.052296	0.048227
Residual	444.6625	477.0	NaN	NaN	NaN	NaN

```
FWER=0.05 method=bonf
alphacSidak=0.02, alphacBonf=0.017
=============================================================
   group1      group2      stat   pval  pval_corr reject
-------------------------------------------------------------
yolo_egohos    yolo_pred 15980.5 0.0001    0.0002   True
yolo_egohos yolo_thunder 16327.5    0.0       0.0   True
  yolo_pred yolo_thunder 13170.5 0.6359       1.0  False
-------------------------------------------------------------
```

Figura 41: Resultados de análisis estadístico para recoger diferencias percibidas por participantes entre algoritmos de segmentación

Además, se ha obtenido la Figura 42 en la que se representan los MOS de la calidad de visualización de los vídeos segmentados, teniendo en cuenta la acción que se realiza y cada uno de los algoritmos:

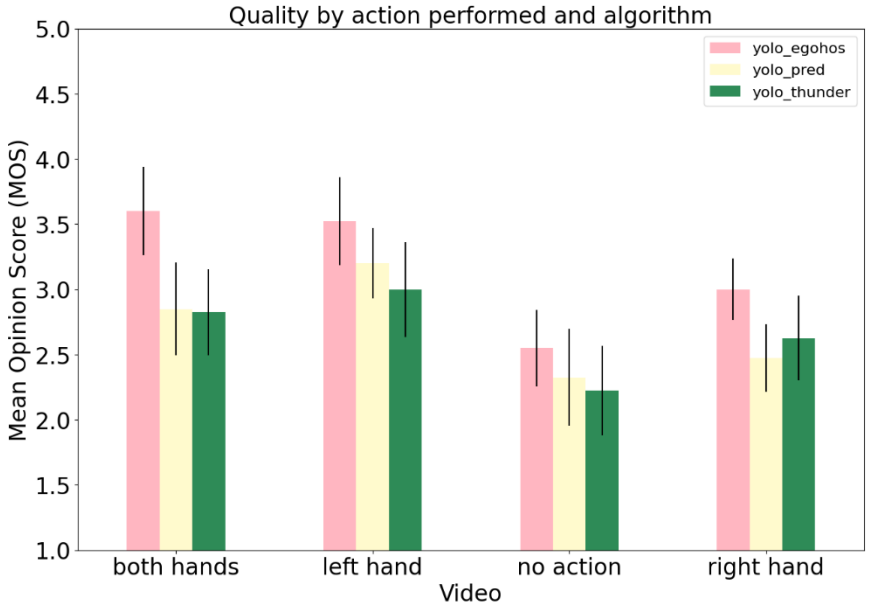


Figura 42: MOS de la calidad teniendo en cuenta las acciones realizadas y los diferentes algoritmos

Del gráfico anterior se puede concluir que en los vídeos en los que se realizan acciones (mano derecha, mano izquierda y dos manos), YOLO+EgoHOS es el algoritmo preferido.

Por otro lado, respecto a la pregunta de la **molestia generada por el ruido de fondo**, en la Figura 43 se presenta una comparación entre la cantidad de veces en las que los participantes han visto o no objetos adicionales:

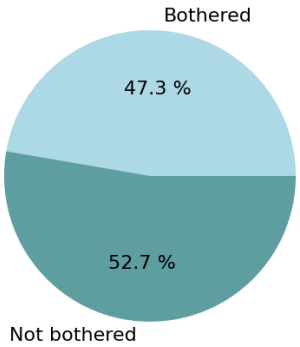


Figura 43: Porcentaje de participantes que han visto o no objetos que no deberían de aparecer (ruido de fondo)

Un 47.3% de las respuestas a esta pregunta han sido que sí han visto algún objeto que no debería de aparecer en el vídeo, mientras que un 52.7% de las respuestas han sido que no han visto ningún objeto adicional. De las respuestas afirmativas, en la Figura 44 se representa la molestia generada por la aparición de estos objetos, agrupando los vídeos por algoritmos:

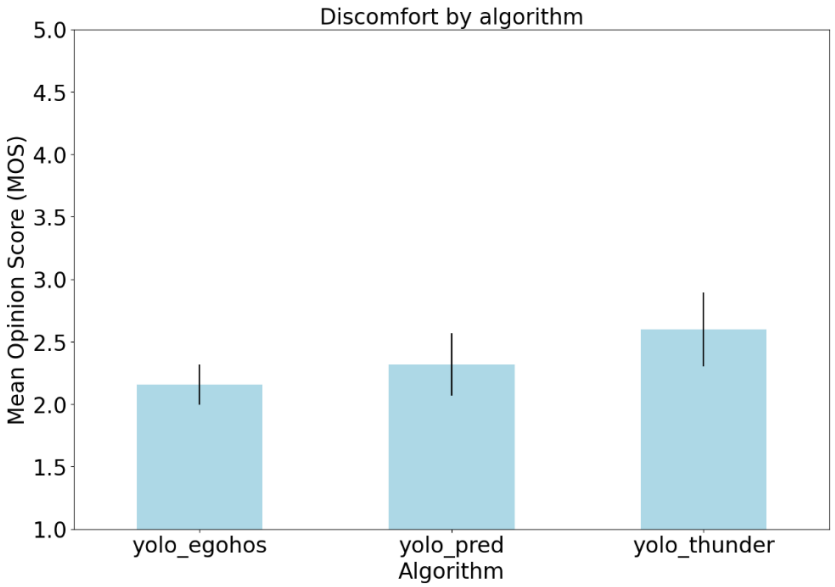


Figura 44: MOS de la molestia según el algoritmo

Cabe mencionar que, para obtener el gráfico anterior, se ha invertido la escala de puntuación de la molestia. Es decir, el 1 pasa a hacer referencia a que al participante no le ha molestado

la aparición de objetos adicionales, mientras que el 5 pasa a hacer referencia a que le ha molestado mucho. De la gráfica, se observa que los vídeos segmentados con YOLO+EgoHOS son aquellos en los que a la gente le ha molestado menos el ruido de fondo. Por otro lado, se ha obtenido un gráfico con la molestia generada por el ruido de fondo según el vídeo, teniendo en cuenta los tres algoritmos, presentado en la Figura 45.

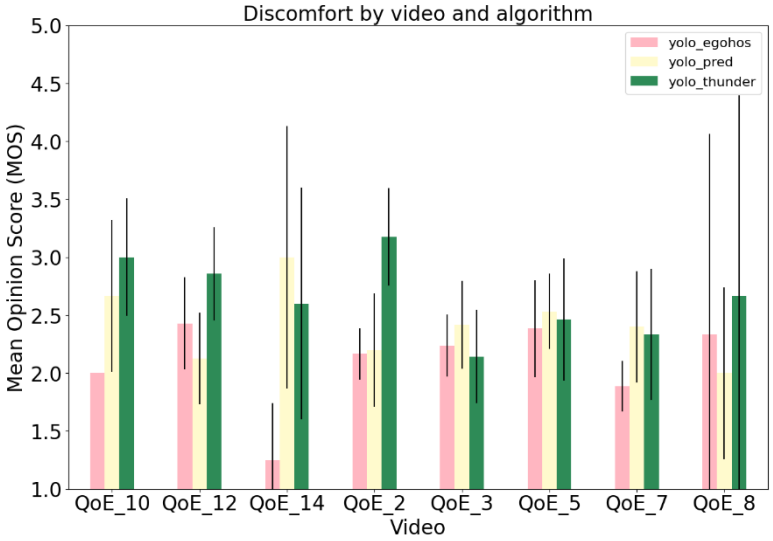


Figura 45: MOS de la molestia para cada vídeo según el algoritmo

De este gráfico no se obtienen resultados significativos. Adicionalmente, se ha obtenido un gráfico en el que se representa el MOS respecto a la molestia generada por la aparición de objetos adicionales en los diferentes vídeos, teniendo en cuenta la acción que se realiza y cada uno de los algoritmos, presentado en la Figura 46.

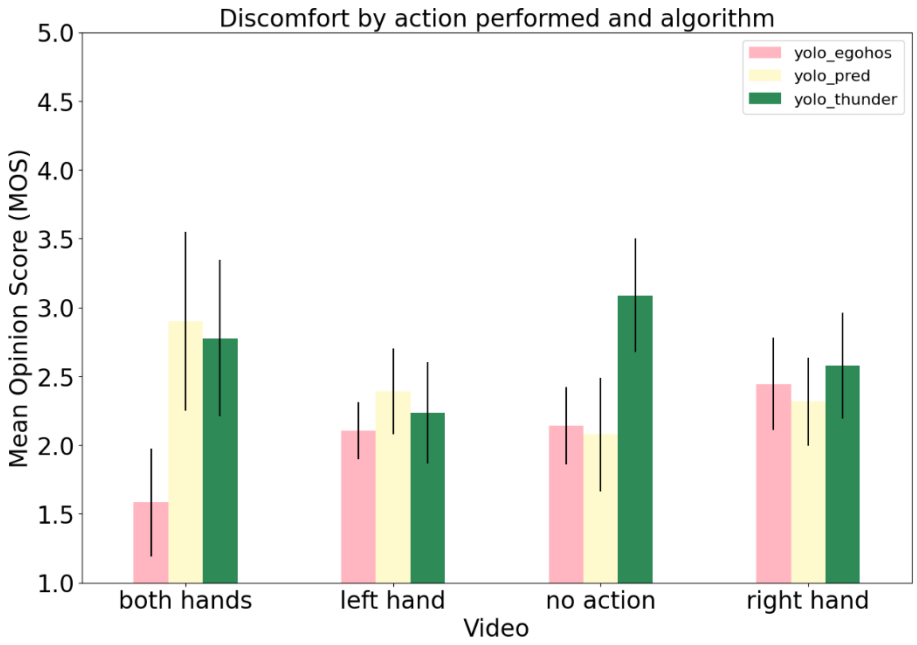


Figura 46: MOS de la molestia teniendo en cuenta la acción realizada y cada uno de los algoritmos

En este caso, cuando no se realiza ninguna acción, los vídeos segmentados con YOLO+ThunderNet son aquellos en los que el que el ruido de fondo genera más molestia a las personas participantes. Por otro lado, en los vídeos en los que la acción se lleva a cabo con las dos manos, el algoritmo con el que el ruido de fondo genera una menor molestia es el de YOLO+EgoHOS.

A continuación, se lleva a cabo un **análisis estadístico** por algoritmo respecto a la **molestia** generada por la aparición de objetos adicionales (teniendo en cuenta solo a los/as participantes que han marcado que sí han visto algún objeto que no sean brazos/manos, platos, vasos/tazas y cubiertos). Dada la distribución normal de los datos recogidos según el test D'Agostino and Pearson, se aplica un test ANOVA, resultando que hay diferencia significativa entre los datos. Concretamente, aplicando un análisis post-hoc Tukey con corrección de Bonferroni, se obtienen diferencias significativas entre las medias de evaluaciones de molestia entre YOLO+EgoHOS y YOLO+ThunderNet (ver Figura 47).

	sum_sq	df	F	PR(>F)	eta_sq	omega_sq
C(Algoritmo)	5.033374	2.0	3.975113	0.020789	0.050334	0.037435
Residual	94.966626	150.0	NaN	NaN	NaN	NaN

group1	group2	stat	pval	pval_corr	reject
yolo_egohos	yolo_pred	1121.5	0.674	1.0	False
yolo_egohos	yolo_thunder	1039.5	0.0137	0.0411	True
yolo_pred	yolo_thunder	1037.5	0.0506	0.1518	False

Figura 47: Resultados de análisis estadístico para recoger diferencias de molestia percibidas por participantes entre algoritmos de segmentación

Respecto al análisis de los **tiempos de respuesta**, en primer lugar, se obtienen tiempos de respuesta relativos para que la duración del vídeo no afecte al análisis de los resultados. En la Figura 48 se muestra la evolución del tiempo de respuesta de los/as participantes conforme avanza la encuesta.

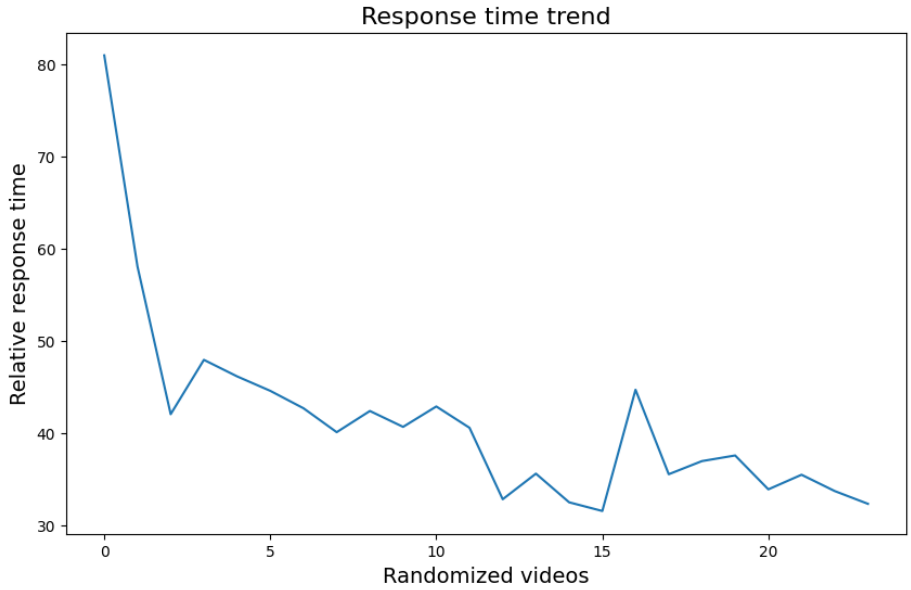


Figura 48: Evolución del tiempo de respuesta agregado de los participantes

En el gráfico anterior, el eje x se corresponde con los vídeos según se les presentan – para cada persona se trata de un vídeo diferente, ya que se muestran aleatorizados. Se observa una tendencia decreciente de los tiempos de respuesta relativos, es decir, conforme avanza la encuesta, los/as participantes tardan menos en responder a las preguntas (bien sea por cansancio, porque entienden mejor en qué consiste el experimento…), pero sí que es un efecto claro.

Por otro lado, respecto a las diferencias en el tiempo de respuesta según la acción realizada en el vídeo, se presenta la Figura 49.

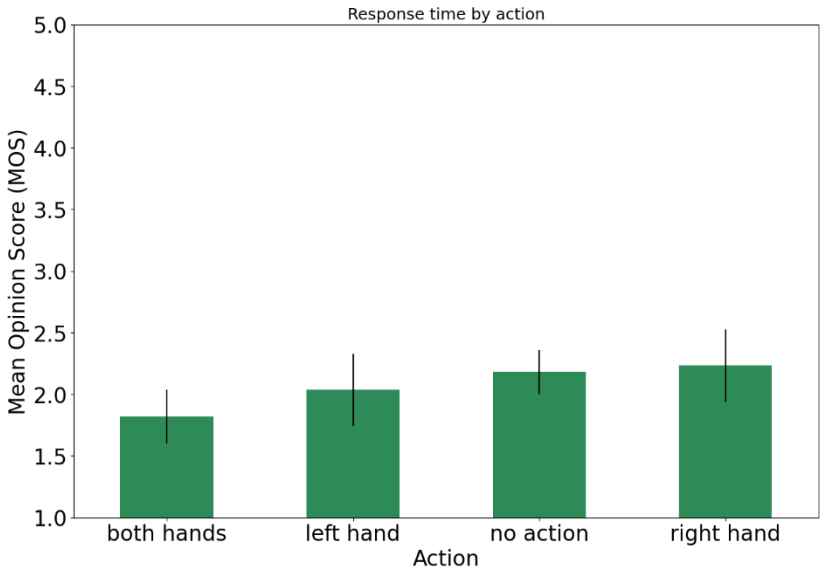


Figura 49: MOS del tiempo de respuesta según el tipo de acción realizada

En este caso, no existe ninguna diferencia significativa en el tiempo de respuesta según la acción que se realiza en cada vídeo.

3.2.2 Tecnologías de interacción natural con objetos físicos y virtuales

En esta sección se describe en detalle el experimento diseñado para realizar las pruebas subjetivas de evaluación de QoE. En primer lugar, se revisará el diseño del entorno de entrenamiento y los dos entornos experimentales, ambos con la temática de un restaurante de comida rápida. A continuación, se explicará la preparación de la sala de pruebas y el equipo utilizado. Además, se describirá la metodología seguida, los cuestionarios empleados y el perfil de las personas participantes de las pruebas.

Condiciones experimentales de la prueba

Las pruebas subjetivas tenían como objetivo medir la calidad de la experiencia del usuario en entornos XR, donde el entorno virtual era un restaurante de comida rápida (ver Figura 50:) y el componente real era la representación diferente del cuerpo de la persona presentada con anterioridad. Esto se comparó con un escenario completamente virtual para evaluar la efectividad de los entornos XR y las técnicas de segmentación egocéntrica para manipulación de objetos desarrolladas. Los participantes completaron dos tareas familiares mientras usaban gafas de realidad virtual, experimentando diferentes condiciones a lo largo del proceso.

Figura 50: Restaurante virtual visto desde dos posiciones distintas

Las tareas incluían:

Recoger una botella rociadora para regar una planta y devolver la botella a su lugar.

Usar fichas azules para marcar ubicaciones en un mapa. Tanto las tareas como las condiciones se asignaron al azar para cada participante.

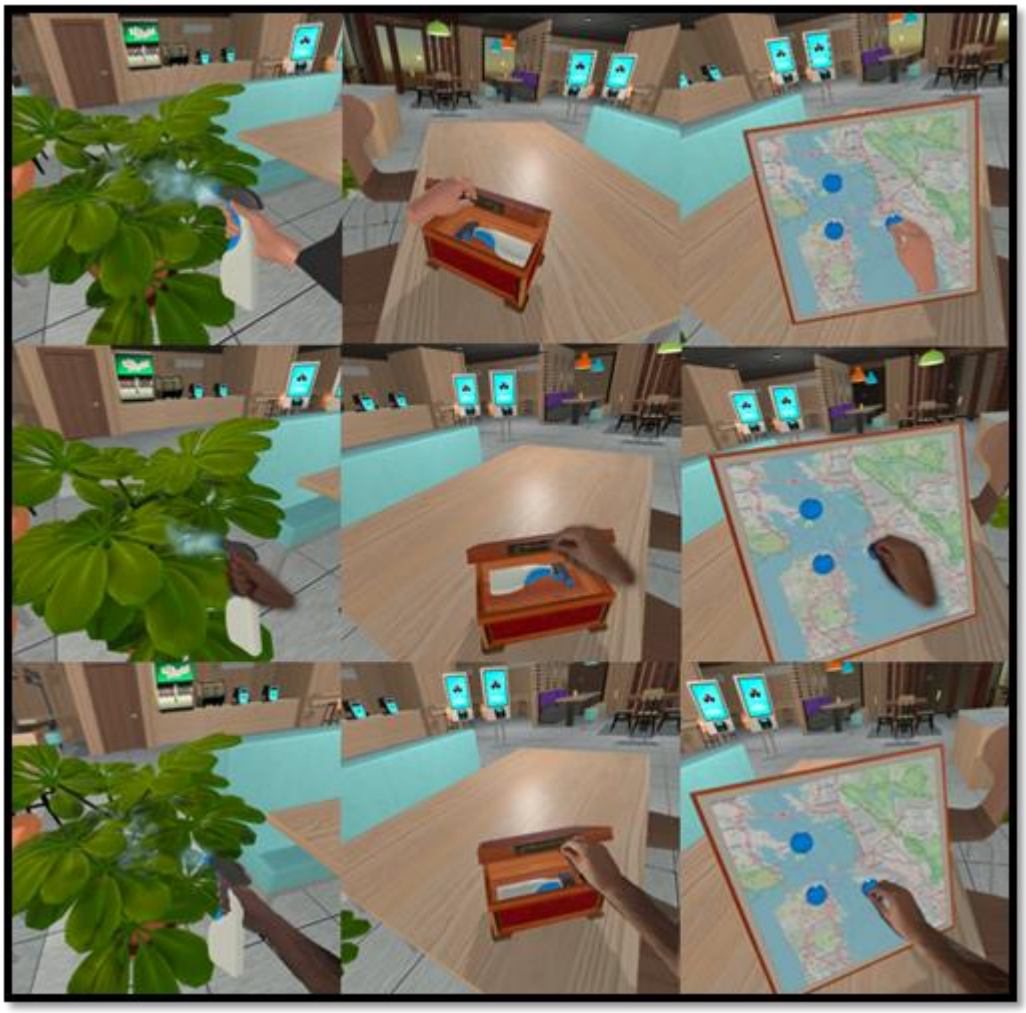


Figura 51: Condiciones de la tarea y de la segmentación

Los/as participantes experimentaron las dos tareas en tres condiciones:

manos virtuales,
manos generativas (manos reales visibles)
manos con brazos visibles y parte de la parte superior del cuerpo.

En la Figura 51 se puede ver un ejemplo de las 6 condiciones experimentales. Las tareas y las condiciones se asignaron al azar para garantizar una exposición variada para cada participante.

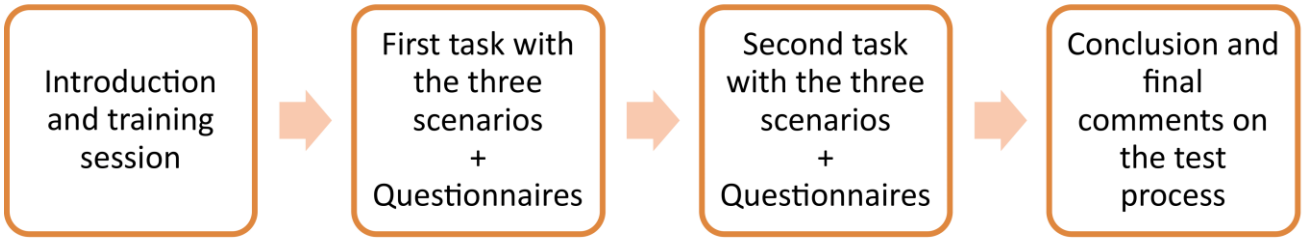


Figura 52: Flujo de trabajo de las sesiones experimentales

El flujo de trabajo experimental, como se muestra en la Figura 52, se describe en detalle a continuación. El experimento comenzó con una breve introducción que explicaba las tareas y el uso de los auriculares Meta Quest 3. Al igual que en la sección anterior, los objetivos de esta primera parte es que las personas que van a participar comprendan el objetivo del experimento, el proceso a seguir y la importancia de su participación en el estudio. Se les dijo a los participantes que realizarían dos tareas cotidianas en seis escenarios y que completarían un breve cuestionario después de cada una. También se les advirtió sobre el espacio limitado y se les recomendó que evitaran movimientos bruscos. Después de esto, rellenaron el formulario de información personal y el consentimiento de tratamiento de datos personales según las normativas de privacidad y protección de datos estipulados por el Reglamento General de Protección de Datos (RGPD) (UE, 2016/679). Antes de las pruebas principales, los participantes tuvieron cinco minutos para practicar la interacción con objetos virtuales, denominada la sesión de entrenamiento. El objetivo de esta fase es ajustar el equipamiento para que las personas se sientan cómodas antes de comenzar la prueba, chequear que todo está bien configurado y que se familiaricen con la tecnología. Cada tarea tomó alrededor de 1 a 2 minutos, con seis tareas en total (dos tareas con tres condiciones). Después de cada tarea, las personas participantes completaron un cuestionario. En general, las sesiones se completaron en aproximadamente 45 minutos.

Sesión de entrenamiento

En este apartado se describe el desarrollo del entorno de entrenamiento utilizado para las pruebas subjetivas, que sirvió como prueba previa para las personas participantes. Dado que las pruebas implicaban dos tareas que requerían la interacción entre objetos virtuales y la persona usuaria, se diseñó una sesión de entrenamiento para permitir que los/as participantes se familiarizaran con diferentes interacciones. Además, estas tareas no estaban relacionadas con las condiciones de los experimentos, lo que evita el efecto de aprendizaje.

Figura 53: Escenario de entrenamiento

El entorno de entrenamiento consistió en una escena completamente virtual en la que las personas que participaron se encontraban de pie frente a una mesa. Utilizando manos grises semitransparentes virtuales (como se muestra en la Figura 53), podían interactuar con una variedad de objetos. Se distribuyeron un total de 21 objetos interactivos sobre la mesa.

Metodología

En esta sección se detallan los cuestionarios administrados después de cada una de las seis tareas. Los cuestionarios consistían en 10 ítems que cubrían temas como la QoE de la escena, la presencia, la incomodidad y la facilidad y el enfoque en la tarea realizada. Asimismo, se añadió un cuadro de texto en el que las personas podían dejar algunas observaciones que ayudaran a los investigadores e investigadoras en el análisis de los resultados a extraer conclusiones. También se preguntó a las personas inicialmente sobre la escena que acababan de completar y su ID asignado durante el formulario de información personal rellenado al inicio de la prueba.

Para la evaluación de la presencia, el desempeño, la implicación y la confianza en la tarea utilizamos preguntas de diferentes fuentes que ya han probado estas preguntas en el pasado para experiencias similares. Las preguntas restantes se evaluaron utilizando una escala ACR, y las sensaciones de vértigo o malestar se evaluaron con la escala Vertigo Score Rating (VSR). Ambas son escalas tipo Likert de 5 puntos, recomendadas por la Recomendación ITU-T P.919 para evaluar experiencias inmersivas.

Además, se incluyeron algunas preguntas que las personas participantes podían responder en texto libre:

- ¿Cómo ha sido tu experiencia al ponerte las gafas? ¿Has sentido alguna molestia?
- ¿Qué te han parecido las pruebas en general?
- ¿Has sentido que tenías control de tu cuerpo dentro de la experiencia?

- ¿Crees que ver tu propio cuerpo puede mejorar la experiencia respecto a un avatar?
- ¿Te has visto inmerso en la tarea que estabas realizando?
- ¿Quieres añadir algún comentario o sugerencia para posibles pruebas futuras?

Participantes

Se realizaron pruebas subjetivas durante tres días, con un total de 20 sujetos (8 mujeres y 12 hombres; con edades comprendidas entre los 20 y los 30 años), que tuvieron que realizar cada persona una sesión de pruebas durante alrededor de una hora. Se intentó coger un amplio rango de diversidad de personas, que incluyese algunas que tuvieran experiencia con experiencias inmersivas, y otras que no hubieran tenido ningún tipo de contacto con ellas a lo largo de su vida, lo cual incentivaría su efecto "*Wow*". Este rango estuvo equilibrado en un 50% de personas que sí habían experimentado con VR, y otro 50% que no.

Resultados

QoE Global

En primer lugar, examinamos los resultados relacionados con las puntuaciones medias de QoE. Estos resultados se muestran en la Figura 54, que presenta puntuaciones de forma independiente para manos virtuales (MV), segmentación egocéntrica de manos generativas (MG), donde el participante solo vio sus manos, y segmentación egocéntrica basada en avatares y Passthrough (MBC), donde el/la participante podía ver sus propias manos, cuerpo y brazos. La leyenda diferencia los resultados de las tareas de Riego y Mapa con diferentes colores. Se puede concluir que las experiencias son bastante similares en términos de evaluaciones, con una tendencia general a mejores experiencias con mapas, tanto para MV (con una diferencia mínima) como para MG (con una diferencia ligeramente mayor).

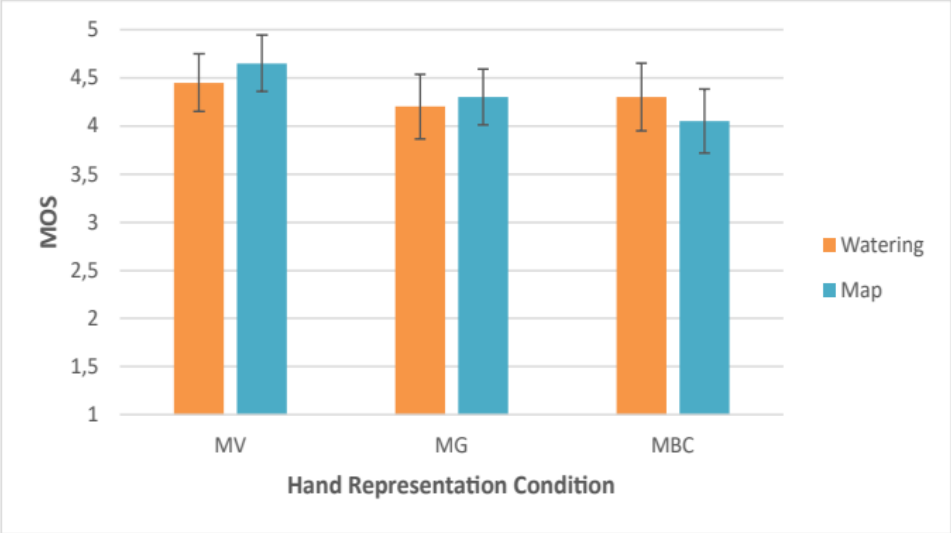


Figura 54: Media de las votaciones para el factor de QoE global con intervalo de confianza del 95%

Esto puede deberse a que los comentarios indicaron que las experiencias con mapas se sentían más realistas y complejas, presentando un desafío que ayudó a los participantes a sumergirse más. Además, los mapas eran más fáciles de manejar y causaban menos problemas durante la interacción, brindando una mayor sensación de seguridad e inmersión.

Cabe destacar que MBC rompe esta tendencia, mostrando evaluaciones más altas para Regar. Esto puede explicar que Regar es una actividad más interactiva con más estímulos, lo que hace que los participantes se sientan más inmersos al interactuar con los elementos virtuales, como ver sus brazos además de sus manos, lo que agrega realismo. Sin embargo, nuestra conclusión es que las experiencias con promedios más altos son aquellas que involucran manos virtuales. Los intervalos de confianza calculados y mostrados en la Figura Figura 54 sugieren que no hay diferencias significativas en general, aunque se sospecharon diferencias entre Map y MBC, dada la notable varianza en sus promedios. Por tanto, se realizó una prueba ANOVA, tras una prueba de normalidad para asegurar que la distribución de los datos fuera normal. Los resultados de la prueba de normalidad no fueron concluyentes, sin embargo, tanto la kurtosis como la asimetría dieron como resultado el rango (-1,1). Esto demuestra que los datos siguen una distribución normal. También se realizó un análisis de varianza para ver el efecto de las condiciones y tareas en la votación. La prueba ANOVA no encontró una influencia significativa de las tareas o condiciones en la votación.

Calidad visual

En segundo lugar, revisamos las puntuaciones medias de la calidad visual, que siguen una tendencia similar a la de la calidad de la experiencia. Aunque las observaciones directas de

Los resultados de la Figura 55 no son coherentes con la retroalimentación de los comentarios verbales y escritos porque indicó que algunos usuarios notaron problemas de visualización con los mapas, citando un enfoque deficiente que puede haberse debido a ajustes inadecuados de los auriculares o problemas de visión individuales. Aunque no se observaron diferencias estadísticamente significativas en los intervalos de confianza individuales, la combinación de los resultados medios de Riego y Mapas sugirió posibles diferencias.

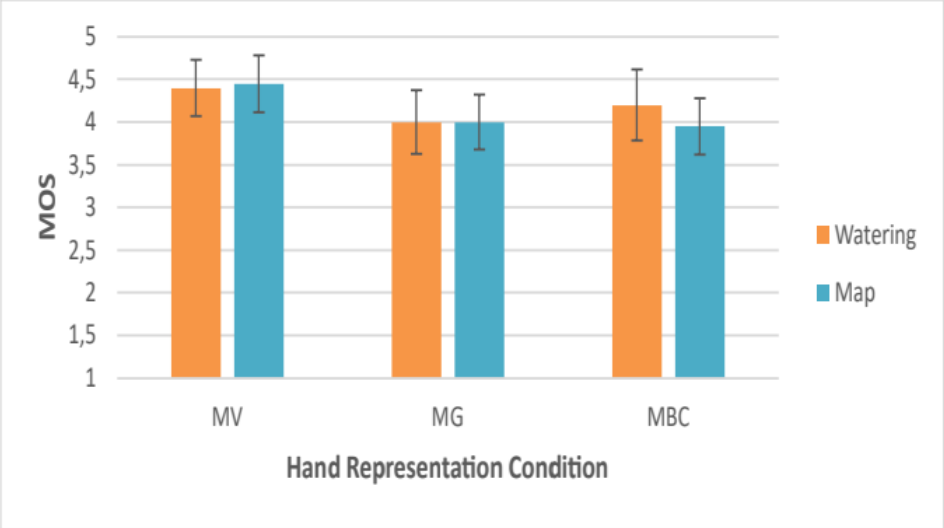


Figura 55: Media de las votaciones para el factor de calidad visual con intervalo de confianza del 95%

Los resultados de la prueba de normalidad no fueron concluyentes, sin embargo, tanto la kutosis como la asimetría dieron como resultado el rango (-1,1). Esto demuestra que los datos siguen una distribución normal. También se realizó un análisis ANOVA para ver el efecto de las condiciones y las tareas en la votación. Este análisis mostró diferencias significativas ($p < 0,05$) para las condiciones. Un análisis posterior con el método HSD de Tuckey confirmó diferencias significativas entre las manos virtuales y generativas, pero su efecto fue pequeño ($\eta p2 < 0,06$).

Cibermareo

En tercer lugar, examinamos los resultados de la puntuación media de vértigo o malestar experimentado durante las pruebas, presentados en la Figura 56. Para todas las condiciones, sus medias rondaron el 4,5 con una desviación estándar de 0,42, mostrando poco o ningún signo de mareo. En promedio, Mapas obtuvo la puntuación más baja en cuanto a malestar en los tres tipos de escenas, lo que sugiere que causó menos problemas a las personas que participaron. Sin embargo, no se encontraron diferencias estadísticas.

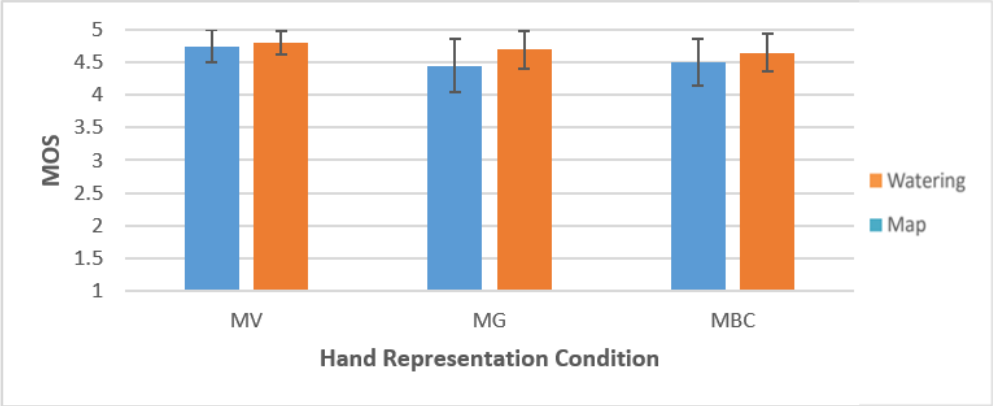


Figura 56: Media de las votaciones para el factor de cibermareo con intervalo de confianza del 95%

Presencia

A continuación, presentamos los resultados de la presencia espacial, es decir, si las personas participantes se sintieron realmente rodeadas por las escenas. Estos resultados se dividen en cuatro figuras. La Figura 57 muestra las puntuaciones de la sensación de estar rodeado por el espacio, con puntuaciones más altas para las manos virtuales, lo que tiene sentido ya que solo presenta el entorno virtual sin estímulos del mundo real. La Figura 58 muestra los resultados de la sensación de estar realmente en el lugar, donde las manos virtuales puntúan más alto en comparación con las técnicas de segmentación egocéntrica, con Maps mostrando puntuaciones más altas probablemente porque requiere menos interacción, reduciendo el impacto de los artefactos y problemas de Passthrough que muestran partes del mundo real a través de partes del cuerpo segmentadas. La Figura 59 presenta los resultados para alcanzar y tocar elementos de la escena, con puntuaciones similares entre Watering y Maps para las manos virtuales y las manos con brazos y cuerpo. Las manos virtuales proporcionan la mayor inmersión y sentido del tacto, superando a los otros métodos. Se observan diferencias notables entre Maps y Watering para las manos generativas, con Watering puntuando mucho más bajo, posiblemente debido a la interacción más compleja con la botella, lo que resulta en una sensación menos realista. No se encontraron diferencias estadísticamente significativas en ninguna de las tres figuras.

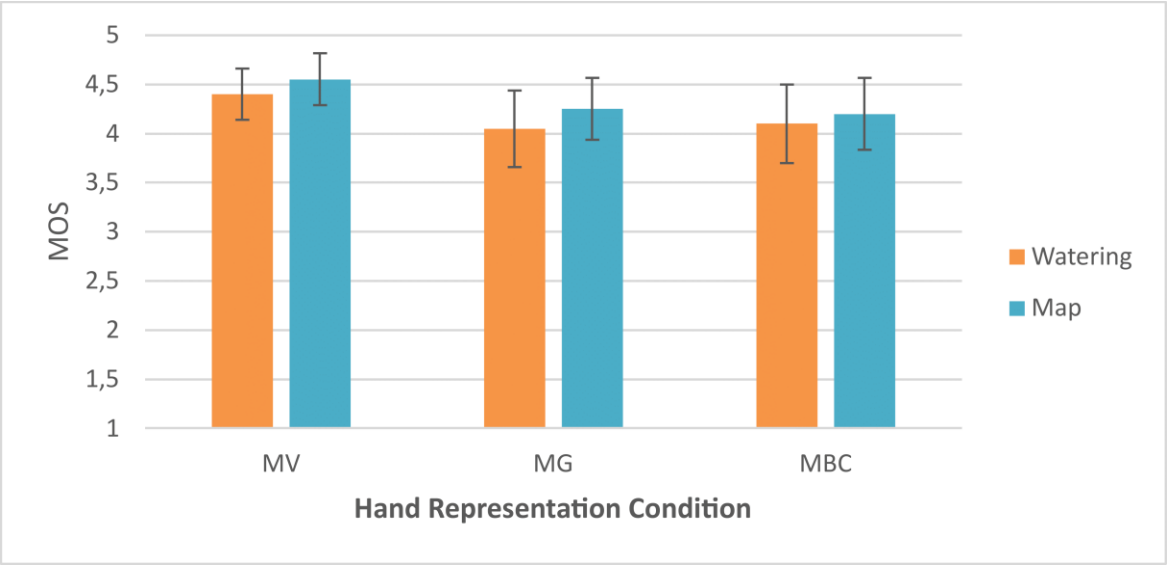


Figura 57: Presencia espacial: entorno

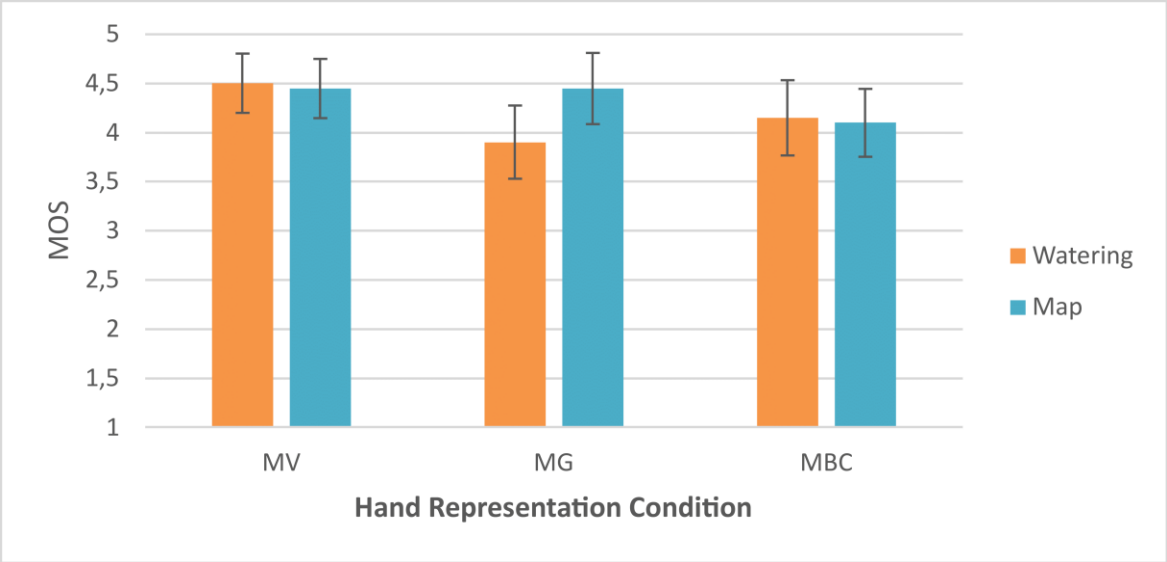


Figura 58: Presencia espacial: lugar de la escena

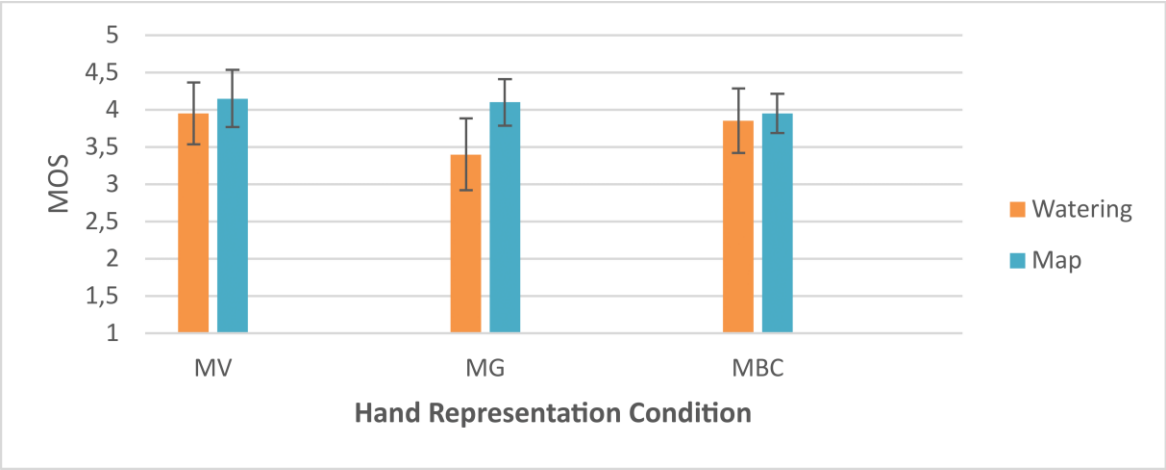


Figura 59: Presencia espacial: alcanzabilidad

Implicación

A continuación, revisamos los resultados del factor Implicación, que se refiere a la coherencia entre la realidad y la virtualidad, y si las personas usuarias se sintieron presentes en la escena. La Figura 60 muestra las puntuaciones medias de presencia para cada tipo de escena, consistentemente más altas para la tarea de Mapas. La mayoría de las personas usuarias calificaron la experiencia como bastante coherente (4), con picos para las manos con el cuerpo y los brazos en Mapas, y para las manos virtuales en Riego. Aquellas que calificaron la experiencia como inconsistente (2) lo hicieron para el riego con manos generativas. Se realizó un análisis más detallado debido a las diferencias significativas sospechadas, especialmente dadas las marcadas variaciones en MG. Suponiendo que las tareas no eran relevantes para el análisis, se combinaron los resultados medios. Se realizaron nuevamente pruebas de normalidad, ANOVA y pruebas t (aunque la prueba t no era estrictamente necesaria, se realizó para corroborar el ANOVA). Los resultados del análisis de varianza mostraron una influencia significativa de las tareas en la votación, pero no de las condiciones.

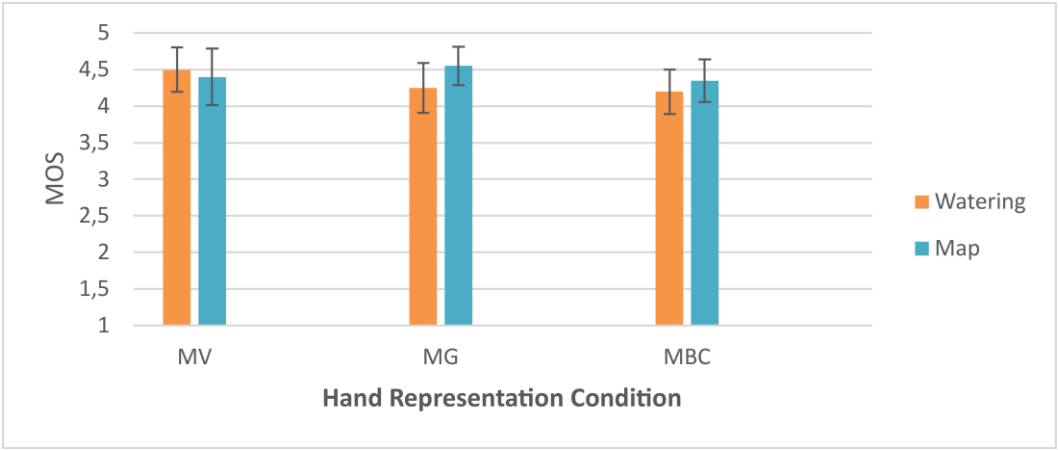


Figura 60: Implicación

Rendimiento

Por último, se muestran los resultados del rendimiento de las tareas. La Figura 61 ilustra el esfuerzo requerido para ambas tareas, y la mayoría de las personas participantes las encontraron fáciles, especialmente Mapas, que recibió la mayor cantidad de votos. Una minoría encontró alguna dificultad, ya que el esfuerzo requerido en Mapas posiblemente se debió a problemas de concentración o una complejidad ligeramente mayor. La Figura 61 presenta las puntuaciones medias de rendimiento para cada escena, con resultados que muestran que, en general, encontraron que Maps era más fácil de resolver que Regar con segmentación egocéntrica, mientras que lo opuesto era cierto para las manos virtuales. Sin embargo, no se encontraron diferencias estadísticamente significativas entre las condiciones.

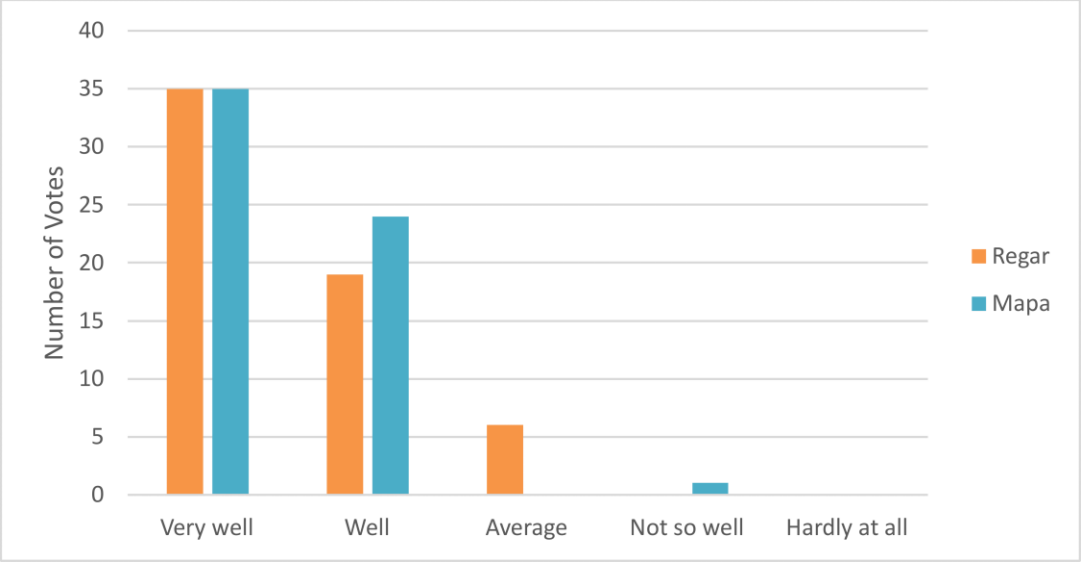


Figura 61: Rendimiento

Concentración

Los resultados para el factor Concentración se muestran en la Figura 62, donde la mayoría de participantes calificaron las tareas como resueltas muy bien o bien, lo que indica satisfacción general. La retroalimentación verbal confirmó que no encontraron las tareas desafiantes, aunque Maps se consideró algo más desafiante. Cuando se les preguntó si les gustaría ver tareas más complejas, la mayoría de los/as participantes expresaron su disposición a participar en pruebas más desafiantes.

La Figura 62 muestra las puntuaciones medias de concentración en la tarea frente a los alrededores. Los resultados fueron similares para las escenas de segmentación egocéntrica, ya que tanto MG como MBC permitieron una mejor concentración en la tarea debido a menos estímulos visuales y menos dinamismo. Esta tendencia cambia con las manos virtuales, donde las personas usuarias parecían estar más concentrados en Watering, probablemente

porque la experiencia puramente virtual condujo a una mayor concentración a pesar del aumento de estímulos. Los comentarios indicaron que todos los/as participantes sintieron que estaban realizando las tareas según las instrucciones y, en general, se centraron más en completar la tarea que en la segmentación. No se encontraron diferencias estadísticamente significativas entre las condiciones.

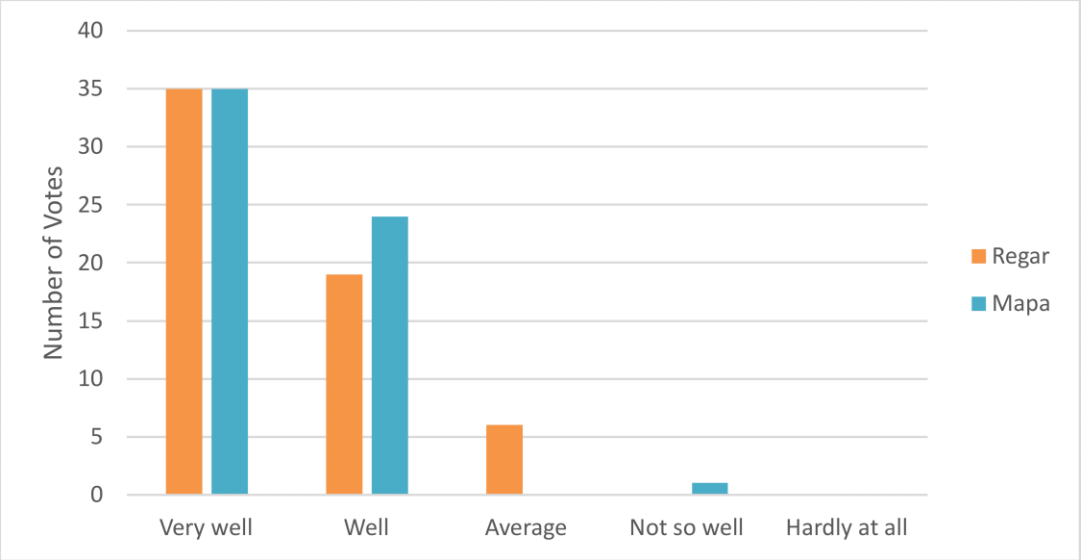


Figura 62: Concentración

Comentarios

Los comentarios de las respuestas escritas y verbales al final de la sesión revelaron que las 20 personas participantes estaban muy satisfechas con las pruebas, utilizando descriptores como interesante, entretenido, inmersivo y divertido. Coincidieron en que las experiencias fueron muy dinámicas y que la capacidad de movimiento agregó sensaciones positivas.

Un número significativo de personas notó que la altura del entorno virtual no coincidía con el piso del mundo real, lo que causaba momentos de incomodidad o miedo al moverse dentro del mundo virtual. Este problema podría resolverse recalibrando manualmente cada escena para cada persona después de colocarse el casco de realidad virtual.

Los resultados de los estudios subjetivos presentados previamente ofrecen información valiosa sobre el impacto de las técnicas de segmentación egocéntrica en la experiencia del usuario en entornos XR. Si bien no se observaron diferencias significativas en todas las condiciones de segmentación en términos de calidad de experiencia general, los resultados indican que las manos virtuales (MV) generalmente brindaron una experiencia visual ligeramente mejor. Sin embargo, todos los factores obtuvieron buenos resultados en todas las condiciones, lo que sugiere que las tareas interactivas que involucran a los usuarios a través de estímulos más complejos, como la manipulación de objetos, mejoran la inmersión y el realismo, especialmente cuando los elementos virtuales están bien integrados. Sin

embargo, los usuarios se sintieron más involucrados con la tarea del mapa, ya que es una tarea que requiere más concentración, lo que involucra al usuario/a.

El análisis estadístico con ANOVA y pruebas t descartaron diferencias significativas y confirmaron un efecto menor entre las manos virtuales y los métodos de segmentación egocéntrica en términos de calidad visual. En cuanto al factor de participación, el análisis estadístico mostró diferencias entre las tareas. Los niveles de malestar y cibermareo fueron generalmente bajos.

El estudio también destacó que la sensación de presencia espacial era mayor con las manos virtuales, en particular en tareas en las que una mínima interferencia del mundo real mejoraba la inmersión virtual. La ausencia de diferencias significativas en el desempeño de la tarea respalda aún más la conclusión de que todos los métodos de segmentación proporcionaron una experiencia relativamente comparable, aunque las manos virtuales permitieron una mejor concentración en la tarea y una mayor sensación de seguridad. En general, la respuesta de los participantes fue abrumadoramente positiva, con altos niveles de satisfacción, aunque se señalaron algunos problemas menores de calibración, como la desalineación de la altura del suelo, como áreas de mejora en futuras pruebas.

3.2.3 Análisis coste-beneficio

La Tabla 7 resume los resultados presentados anteriormente. Se incluye también como referencia la segmentación semántica del cuerpo basada en Thundernet y descrita en [20], que se ha usado como punto de partida para el caso de uso de terapia.

Tabla 7: Análisis de los distintos algoritmos de presencia autónoma e interacción natural

			Algoritmo	Manos	Cuerpo	Objetos	Tiempo inferencia	Frecuencia Inferencia	QoE primer plano	QoE fondo
Referencia	Thundernet	X	X				16.7	60	3.6	3.7
Croma Adaptativo			Croma	X	·	·	2	60+	2.1	3.0
			Croma + MP	X	·	·	65.9	1	3.4	2.6
			Croma + DL	X	·	·	16.7	1	3.5	2.7
			Yolo	X	X	X	11.1	60+	2.7	2.3

Segmentación Semántica	Yolo + EgoHOS	X	X	X	36.4	27.5	3.2	2.2
	Yolo + Thundernet	X	X	X	28.4	35	2.7	2.7
Tracking y Passthrough	Manos Visibles	X			~30	30	4.4	-
	Manos Generadas	X			~30	30	4.0	-
	Manos y Cuerpo Generados	X	X		~30	30	4.1	-

Los algoritmos de croma adaptativo se diseñaron con el objetivo de poder reemplazar o complementar la segmentación por Thundernet en entornos controlados. Como se puede ver, la calidad ofrecida en el primer plano (manos) es comparable a la que se alcanza por Thundernet. La segmentación del fondo es significativamente peor, por lo que es importante usarlos en entornos controlados, donde no haya objetos de tonos similares a la piel. Del mismo modo, se puede introducir parcialmente el cuerpo o los objetos si se controla el color de los mismos. Comparando entre ambos, MediaPipe (MP) es más lento que Deep Learning (DL) para establecer los umbrales, y también algo más impreciso, por lo que es preferible la solución con DL.

La segmentación semántica de objetos permite incluir la capacidad adicional de introducir objetos de interés en la escena. La solución más completa (yolo + EgoHOS) tiene una calidad suficiente para ser utilizable, pero todavía no está suficientemente madura para desplegarla en aplicaciones con usuarios finales: tanto la calidad subjetiva como el tiempo de proceso empeoran con respecto al escenario de segmentación de cuerpo, y la discriminación del fondo todavía debe mejorse.

Una alternativa interesante a la segmentación segmántica es el uso del handtracking, combinado con la capacidad de visual passthrough de las Meta Quest 3. Sorprendemente, la calidad percibida es mejor cuando se usan las manos puramente virtuales que cuando se superpone la textura de la mano real. El hecho de no estar perfectamente alineados la mano handtracking (que es el patrón de la interacción) con los distintos tamaños de las manos y el hecho de que el sistema no gestionase oclusiones pudo provocar un cierto efecto “uncanney valley”. En todo caso, a nivel de efectividad y calidad son comparables, y posiblemente el uso

de las manos propias tenga mejoras a nivel de presencia autónoma, sobre todo con las futuras mejoras de la implementación.

En general, la tecnología basada en segmentación semántica es más flexible en cuanto a la capacidad de integrar el cuerpo u objetos arbitrarios, pero a costa de una menor calidad visual y mayor coste de computación que el uso del tracking integrado en el dispositivo.

3.3 Pruebas de tecnologías y algoritmos con personas usuarias de la fundación Juan XXIII Roncalli

3.3.1 Tecnologías de presencia autónoma

En el contexto de nuestro estudio presentado en la sección anterior, hemos adaptado un experimento para hacerlo con personas usuarias específicas del caso de uso. La motivación detrás de esta adaptación radica en la necesidad de obtener las percepciones que tienen las personas en situaciones reales. Es importante destacar que las personas que participan en este experimento son las mismas que participan en el proyecto. Por lo tanto, ya se dispone de los consentimientos necesarios para el tratamiento de sus datos personales. Concretamente, participan 19 personas, de los cuales el género y el nivel de discapacidad se presenta en la Figura 63.

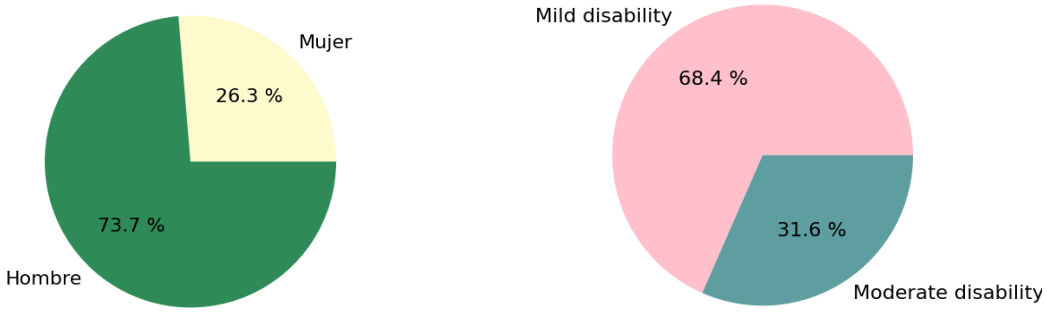


Figura 63: Género y discapacidad de las personas usuarias de la Fundación Juan XXIII Roncalli

Respecto a los vídeos utilizados para ser evaluados por parte de las personas usuarias de la Fundación Juan XXIII Roncalli, se ha partido de la base de datos creada y presentada en el apartado anterior. Sin embargo, se ha submuestreado el número de SRCs utilizadas para acotar el número de estímulos que se van a presentar, de acuerdo a las recomendaciones de las profesionales de la Fundación Juan XXIII Roncalli. Concretamente, se ha seleccionado uno para cada tipo de acción, eliminando los dos vídeos en los que no se realiza ninguna. Por tanto, eligiendo un vídeo de cada acción, el total de vídeos a evaluar por parte de las personas

usuarias de la Fundación Juan XXIII Roncalli es de nueve. Los vídeos elegidos son QoE_5 (ver Figura 26), QoE_7 (ver Figura 28, QoE_10 (ver Figura 30).

Del mismo modo que en la fase anterior, la sesión del experimento se puede dividir en 2 partes principalmente.

En este caso, la explicación del experimento es un poco más larga, para que los/as usuarios/as puedan entender correctamente lo que van a evaluar. Para ello, se ha llevado impreso un documento en el que se muestra una foto del vídeo original del resultado perfecto y, a continuación, la foto de cómo sería el resultado ideal. Por otro lado, se ha llevado impresa en otra hoja la misma foto del vídeo original junto con otra foto de lo que podría considerarse un resultado muy malo. Esto ayuda a los usuarios a distinguir cuál sería el mejor resultado y el peor, facilitando mucho la comprensión de lo que se quiere evaluar. Los recursos utilizados se presentan en la Figura 64.

Vídeo perfecto

VEMOS LAS MANOS, LOS PLATOS, LOS VASOS Y LOS CUBIERTOS,

Y NO APARECE NINGÚN OTRO ELEMENTO

Vídeo original

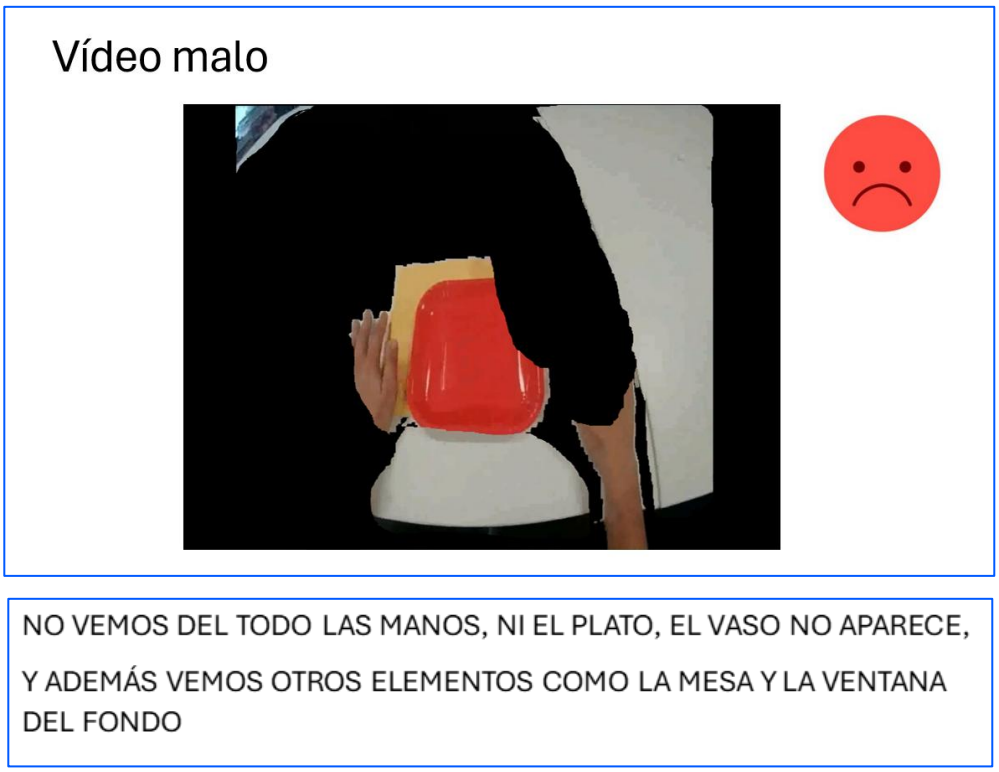


Figura 64: Recursos impresos utilizados para explicar los resultados esperados

El siguiente paso es la evaluación de los algoritmos de segmentación en la plataforma QualityCrowd2 (../FJ23/..). En este caso, para facilitar la comprensión, se les deja una imagen impresa de la escena del vídeo que están evaluando, para que puedan identificar de la mejor forma posible los objetos que deberían aparecer en el vídeo segmentado. Del mismo modo que en la primera fase, el tiempo límite para la visualización de cada vídeo es superior a 1 hora, por lo que las personas participantes pueden repetir las veces que necesiten los vídeos para poder evaluarlos de la mejor forma posible.

Los primeros apartados de la encuesta muestran las instrucciones del experimento. No obstante, habrá una persona leyendo las instrucciones y explicándoselas a las personas usuarias de la Fundación Juan XXIII Roncalli de la forma más clara posible.

A continuación, se muestra el vídeo del resultado perfecto (vídeo con el *chroma*) para que vean qué es lo que idealmente se pretende conseguir con los algoritmos de segmentación.

Seguidamente, se muestran los vídeos a evaluar (en este caso solo evaluarán los tres vídeos procesados con cada uno de los tres algoritmos), aleatorizados para evitar sesgos y sin

presentar dos obtenidos de la misma fuente de manera consecutiva. Cada vídeo va seguido de una única pregunta acerca de la calidad. Además, respecto a la escala de evaluación, se utilizan emoticonos, ya que es la escala con la que están acostumbrados/as a trabajar para facilitar su contestación, tal y como se presenta en la Figura 65.

Figura 65: Pregunta mostrada en la plataforma para la evaluación de los usuarios de la FJ23

Pueden repetir cada vídeo las veces que quieran y tomarse el tiempo que necesiten para contestar a las preguntas.

Análisis de los resultados

Como ya se ha comentado, para esta parte del análisis se tienen en cuenta solo las evaluaciones de los vídeos QoE_5, QoE_7 y QoE_10. En este caso, los logs obtenidos de la plataforma QualityCrowd2 tienen el siguiente formato presentado en la Figura 66.

```
0,1732872412
1,1732872414
2,1732872469,1
3,1732872473
4,1732872505,1,1,,4
5,1732872525,1,1,,4
6,1732872551,1,1,,3
7,1732872576,1,1,,4
8,1732872600,1,1,,4
9,1732872625,1,1,,4
10,1732872871,1,1,,4
11,1732872896,1,1,,1
12,1732872920,1,1,,1
```

Figura 66: Logs que se obtienen directamente de la plataforma tras completar la encuesta (personas usuarias FJ23)

En ellos se presenta la marca de tiempo y, también, las respuestas a la única pregunta que se les realiza. Después de organizar la base de datos, teniendo en cuenta las respuestas de todas las personas encuestadas, esta se presenta del siguiente modo:

	Participant	Video name	Quality	Tiempo visualizacion (s)	Video	Algoritmo	Action
0	1	QoE_5_complete_yolo_thunder.mp4	1	71	QoE_5	yolo_thunder	mano derecha
1	1	QoE_5_complete_yolo_pred.mp4	1	31	QoE_5	yolo_pred	mano derecha
2	1	QoE_10_complete_yolo_egohos.mp4	2	23	QoE_10	yolo_egohos	dos manos
3	1	QoE_5_complete_yolo_egohos.mp4	1	27	QoE_5	yolo_egohos	mano derecha
4	1	QoE_7_complete_yolo_egohos.mp4	2	25	QoE_7	yolo_egohos	mano izquierda

Figura 67: Tabla de datos recopilados organizada

En primer lugar, cabe destacar que para realizar el análisis de los resultados se han descartado algunos participantes – 4, 5, 6, 12, 13, 15, 17 –, ya que han respondido a todos los vídeos por igual. Para analizar los resultados, en primer lugar, se ha obtenido un MOS de la **calidad de visualización de los vídeos segmentados**, agrupados según el algoritmo utilizado, presentado en la Figura 68.

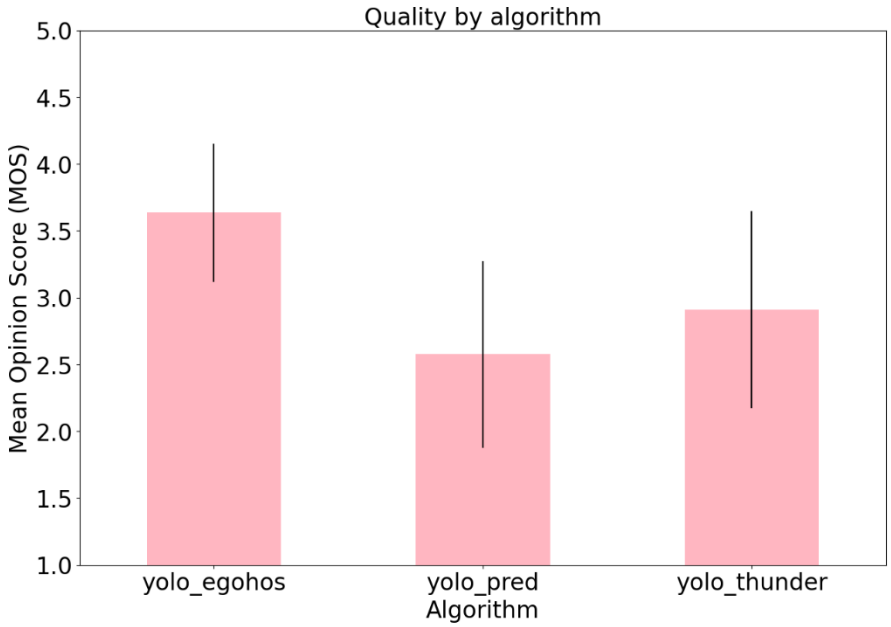


Figura 68: MOS de la calidad de visualización de los vídeos segmentados agregados, teniendo en cada uno de los algoritmos

En este caso, en media, el algoritmo preferido por las personas usuarias es YOLO+EgoHOS, no obstante, se observa que los intervalos de confianza se solapan. Esta preferencia es coherente con la evaluación objetiva, así como también con los resultados extraídos de la evaluación subjetiva por parte de la población general del anterior experimento. Por otro lado, se han obtenido los MOS de la calidad de visualización para cada uno de los vídeos según el algoritmo, presentados en la Figura 69.

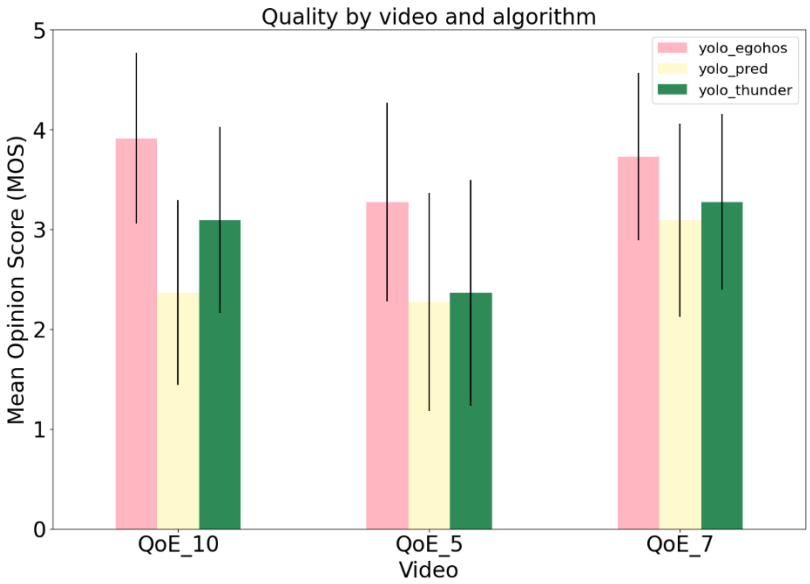


Figura 69: MOS de la calidad para cada vídeo según el algoritmo (personas usuarias FJ23)

En el gráfico anterior se observa que, en media, el algoritmo preferido para cada uno de los vídeos sigue siendo YOLO+EgoHOS. Del mismo modo, el algoritmo preferido para cada una de las acciones realizadas – acción con la mano derecha (QoE_5), acción con la mano izquierda (QoE_7) y acción con las dos manos (QoE_10) – es YOLO+EgoHOS, tal y como se puede observar en Figura 70.

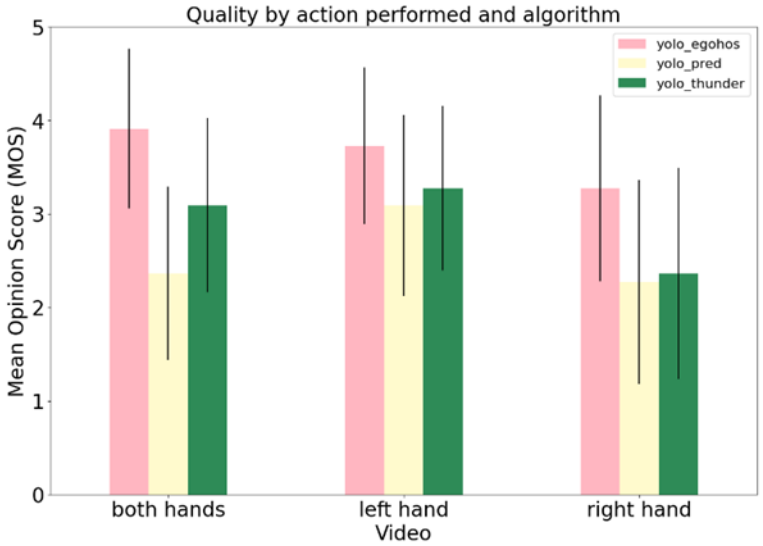


Figura 70: MOS de la calidad para cada tipo de acción en el vídeo y el algoritmo (personas usuarias FJ23)

Respecto al análisis de los **tiempos de respuesta de las personas participantes**, primero se obtienen los tiempos de respuesta relativos, para que la duración de cada vídeo no afecte

en el análisis. Se ha obtenido la Figura 71 con la evolución de los tiempos de respuesta, agregando a todas las personas participantes.

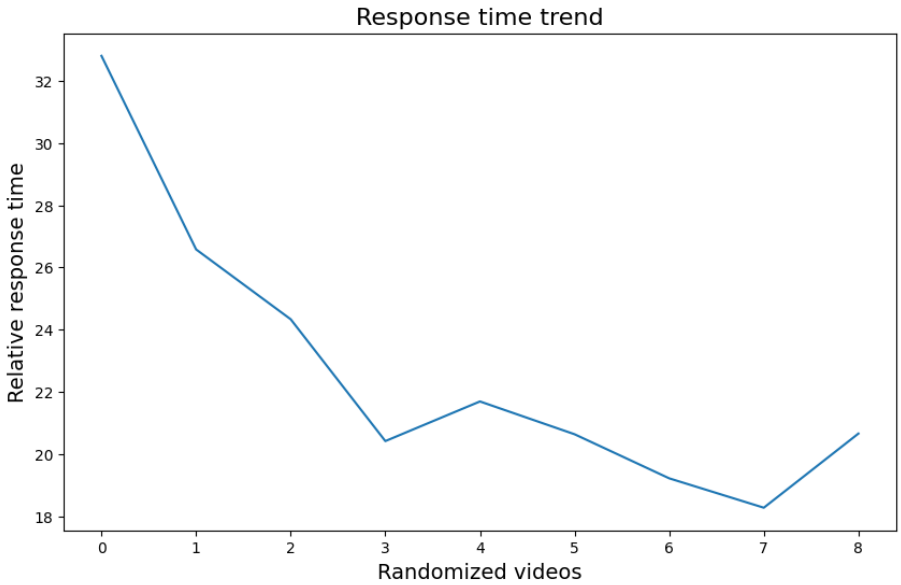


Figura 71: Evolución del tiempo de respuesta agregado de los participantes (usuarios FJ23)

En el gráfico anterior se observa que los tiempos de respuesta relativos siguen una tendencia decreciente, es decir, conforme la persona avanza en la encuesta, el tiempo que tarda en responder a las preguntas es menor (bien sea por cansancio o por que han entendido bien qué tienen que evaluar).

Respecto al **análisis estadístico por algoritmo**, el test de normalidad D'Agostino and Pearson ofrece un p-valor menor a 0.05, por lo que no se puede asumir que la distribución sea normal. Se aplica el test Kruskal Wallis, que se trata de una prueba no paramétrica. Por tanto, esta prueba es buena para analizar un conjunto de datos que no presentan una distribución normal, como en este caso. Obtenemos un p valor inferior a 0.05, resultando que existen diferencias entre las medias de los grupos que se pueden ser consideradas estadísticamente significativas. Dado este resultado, se aplica un análisis post-hoc, presentado en , y encontramos diferencias significativas entre las medias de los algoritmos de YOLO+EgoHOS y YOLO (ver Figura 72).

group1	group2	stat	pval	pval_corr	reject
yolo_egohos	yolo_pred	736.5	0.0104	0.0313	True
yolo_egohos	yolo_thunder	680.5	0.0703	0.211	False
yolo_pred	yolo_thunder	486.0	0.4346	1.0	False

Figura 72: Post-hoc análisis para explorar las diferencias entre algoritmos percibidas por personas usuarias de la Fundación Juan XXIII Roncalli

Dado este resultado, se puede concluir que EgoHOS es el algoritmo que mejores resultados ofrece para este caso de uso en particular, tanto objetiva como subjetivamente.

3.3.2 Tecnologías de interacción natural con objetos físicos y virtuales

Las interfaces de interacción natural son aquellas que permiten a los usuarios comunicarse con sistemas digitales sin necesidad de dispositivos intermediarios como mandos, ratones o controladores físicos. Su objetivo es que la interacción se realice mediante gestos, movimientos corporales, mirada o manipulación directa, replicando las formas de acción que una persona emplearía de manera espontánea en el mundo real. Este tipo de interfaces buscan reducir la carga cognitiva asociada al uso de tecnología, facilitando experiencias más intuitivas, accesibles y cercanas al comportamiento humano.

En el proyecto, la interacción natural con objetos físicos y virtuales se ha materializado principalmente en el caso de uso de teleformación en cocina, donde los participantes debían combinar acciones del mundo real —manipulación de utensilios, desplazamiento por la cocina o manejo de ingredientes— con elementos digitales superpuestos mediante tecnologías de Realidad Extendida. Esto permitió diseñar una experiencia híbrida en la que la formación no solo se basa en contenidos visuales, sino también en la ejecución de tareas prácticas dentro del entorno físico. A continuación, se presentan pruebas pilotos para evaluar inicialmente la viabilidad de distintas formas de interacción natural en dicho caso de uso. La información de los participantes de dichas pruebas se detalla en la Figura 73.

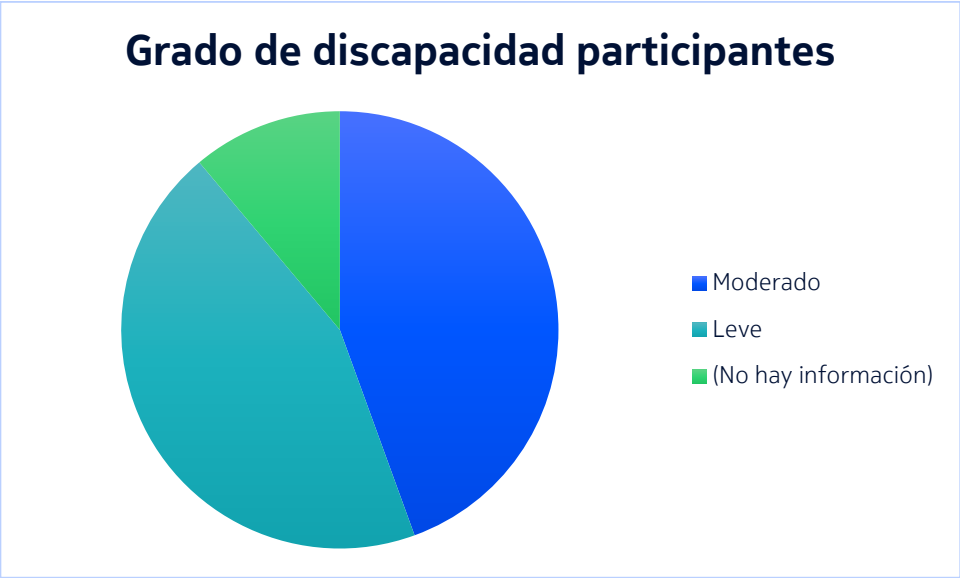


Figura 73: Grado de discapacidad de participantes

Dentro de la interacción natural pueden distinguirse dos grandes escenarios:

1. Interacción con objetos virtuales, es decir, manipulación o activación de elementos digitales superpuestos al entorno.

2. Interacción con objetos físicos, que consiste en utilizar utensilios reales mientras se reciben indicaciones a través de los vídeos formativos.

En el caso de la teleformación en cocina, estos escenarios se integran específicamente en la fase de realidad mixta. En esta fase, los usuarios pueden interactuar con botones virtuales mediante hand tracking, lo que les permite controlar la reproducción de un recorte de vídeo superpuesto al entorno físico real. Al mismo tiempo, mantienen la capacidad de manipular los objetos de cocina reales, lo que permite seguir instrucciones audiovisuales mientras se ejecutan acciones prácticas.

Durante el desarrollo del programa se realizaron dos pruebas específicas con los usuarios para evaluar la viabilidad de estas interacciones naturales. La primera prueba tuvo como objetivo comprobar si los participantes eran capaces de activar y manejar los botones virtuales utilizando únicamente el seguimiento de manos proporcionado por las Meta Quest.

Para realizar la prueba se pidió a los usuarios que pausasen y reprodujesen un vídeo en modo realidad mixta (ver Figura 74). Después la terapeuta que dirigía la sesión anotaba los comentarios que les daban los usuarios junto con una anotación específica si conseguían utilizar los botones.

Figura 74: Ejemplo de realidad mixta con botones virtuales

Durante las pruebas se reportaron 7 de 10 resultados positivos en la interacción con los botones. Los otros 3 no fueron anotados. Los resultados obtenidos en estas pruebas permitieron ajustar tanto la distancia de interacción, como el tamaño óptimo y la iconografía más eficaz para garantizar una experiencia fluida y accesible para todos los participantes.

La segunda prueba se centró en evaluar si la calidad del passthrough de las Meta Quest 3 permitía visualizar con la nitidez suficiente elementos importantes de la cocina real, como anotaciones escritas en papel, así como números digitales presentes en dispositivos como el peso o el mando del horno. Estas evaluaciones fueron clave para garantizar que la combinación entre mundo físico y digital resultara funcional, usable y adecuada para el contexto formativo. En la Figura 75 se puede observar que no se reportaron experiencias negativas si bien algunos usuarios (2) tuvieron algún problema que hizo que la experiencia no fuese totalmente positiva. A la luz de los resultados se puede atestiguar que los usuarios estaban preparados para interactuar con la realidad física utilizando el passthrough.

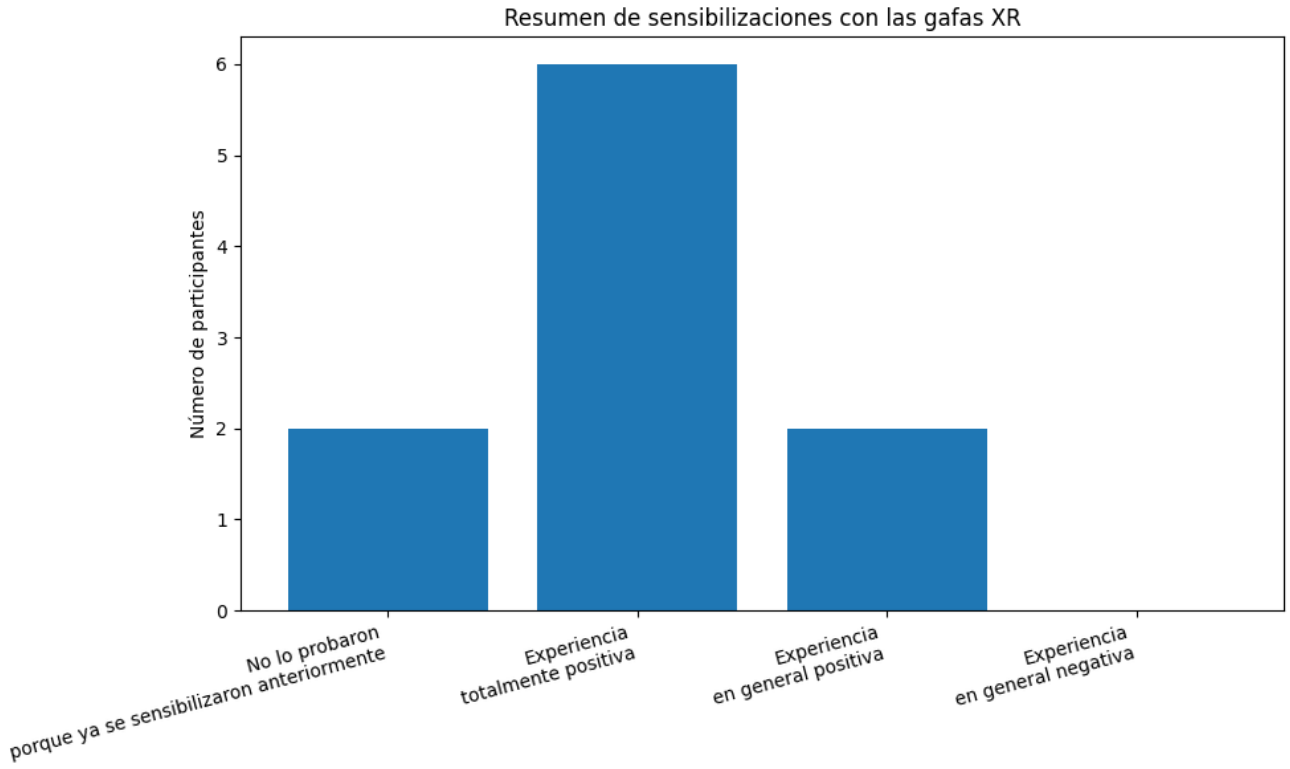


Figura 75: Número de participantes según su experiencia con la etapa de sensibilización con tecnología XR

3.3.3 Discusión

La siguiente tabla muestra la comparación entra la evaluación de calidad visual (MOS) de los algoritmos de segmentación semántica, hecha por usuarios con y sin discapacidad, como se ha descrito en los apartados anteriores.

Tabla 8: Comparación de MOS de los algoritmos entre usuarios con y sin discapacidad

Algoritmo	Población sin discapacidad	Población con discapacidad	Media
Yolo	2.7	2.7	2.7
Yolo + EgoHOS	3.2	3.6	3.4
Yolo + Thundernet	2.7	2.8	2.8

Como se puede ver en la tabla, el resultado de la evaluación es similar en ambas poblaciones, por lo que podemos asumir que los resultados obtenidos en población de un tipo pueden llevarse a la otra, al menos en una primera aproximación. Como ya se analizó antes, estos

algoritmos son prometedores, pero aún no tienen calidad suficiente para integrarlos de forma estable en el desarrollo y prueba de un caso de uso. Por ello, para el caso de uso de terapia se decidió usar el modelo Thundernet (también conviene tener en cuenta que este caso de uso se diseñó al inicio del proyecto, antes de terminar la investigación en el resto de algoritmos); y para el caso de uso de teleformación se optó por probar las tecnologías de interacción natural basadas en tracking y video passthrough.

A la luz de los resultados de las pruebas con los usuarios de la Fundación Juan XXIII se llegó a la conclusión que tanto la calidad visual del passthrough como la interfaz de botones virtuales son interfaces naturales válidas para el desarrollo del caso de uso. Si bien los botones virtuales sufrieron diversos cambios durante su desarrollo, finalmente pudieron ser utilizados correctamente durante las sesiones formativas de realidad mixta.

3.4 Requisitos e indicadores de rendimiento

La siguiente tabla muestra un resumen de las principales tecnologías investigadas en este escenario: thundernet (escenario base) [20], croma adaptativo, Yolo + EgoHOS y mano generativa; así como la evaluación de los usuarios de casos de uso de terapia (CU1, usando thundernet) y teleformación (CU3, usando realidad mixta y mano generativa), descritos en E4.1 y E4.3 respectivamente.

Tabla 9: Resumen de análisis de KPIs para el escenario de computación delegada

KPIs		Requisito	**Thundernet**	**Chroma Adaptive**	**Yolo EgoHOS**	**Mano Generativa**	**CU1**	**CU3**
Sesión	Núm. Usuarios	[1-10]	1	10+	1	-	1	-
	Velocidad Usuario	Estático	Estático					
	Distancia Usuarios	Local (área MEC)	Local					
Tasa de datos	Vídeo	10-100 Mbps UL 1-10 Mbs DL	10 UL / 1 DL	~0.5 UL	10 UL / 1 DL	-	10 UL / 1 DL	-
	Sensores	0.1 Mbps UL	-	-	-	-	-	-
	Datos	0.2 Mbps DL	-	0.01 DL	-	-	-	-

Latencias	Latencia round-trip	< 100 ms	40	40-100	50	-	40	.
	Tasa refresco XR	60 fps	30-60 fps					
	Frecuencia IA	15 infer/s	60	1 inf/s/ usuario	27.5	30	30	30
QoE	MOS perceptual	> 4.0	3.7	3.4	3.4	4.0	4.3*	5.0*
	Presencia espacial	Alta	5.5	-	-	4.2	6.7*	6.3*
	Presencia autónoma	Muy alta	5.8	-	-	4.2	6.0*	-
	Cibermareo	Bajo (VSR > 4.0)	4.4	-	-	4.6	4.9	4.9

Los algoritmos basados en segmentación semántica (Thundernet, incluyendo CU1, y Yolo+EgoHOS) emplean la GPU del MEC por completo por lo que, en el escenario del proyecto, solo soportan un usuario por cada instancia. La solución de croma adaptativo permite compartir los recursos, y escala fácilmente hasta 10 usuarios paralelos. Ese es un límite superior estricto para usuarios que puedan necesitar esta tecnología en el contexto de un despliegue como el de Incluverso 5G (cubre todas las necesidades posibles de un centro como Fundación Juan XIII más que holgadamente). La tasa de datos y latencia de red se mantiene dentro de los márgenes previstos en el entregable E1.2. En el entregable E2.3 se presenta un análisis más detallado de los tiempos de transmisión e inferencia en función de la resolución de las imágenes. Los algoritmos basados en tracking (Mano Generativa y CU3) se implementan localmente en las gafas, por lo que no requieren del uso de la red.

Los valores numéricos de QOE para MOS y cibermareo están entre 1 y 5, siguiendo la recomendación ITU-T P.919, mientras que los valores de presencia espacial y autónoma están en una escala 1-7, como es habitual en las escalas psicométricas. Los valores marcados con asterisco son evaluaciones realizadas por personas con discapacidad en escalas adaptadas (de tres niveles), que se han reescalado linealmente a los rangos estándar. Los algoritmos de croma adaptativo y segmentación semántica de objetos no contienen información de presencia y mareo se evaluaron solo perceptualmente y se descartaron para su integración en el caso de uso final. El caso de uso de teleformación no incluye evaluación de presencia autónoma porque se implementó en un contexto de realidad mixta, donde el usuario está físicamente presente en el entorno local.

La QoE es suficiente para implementar los casos de uso propuestos en el proyecto. Como ya se esbozaba en el escenario de teleportación virtual, los resultados de la evaluación son más altos cuando se evalúa la tecnología en su contexto de uso. Los posibles defectos o artefactos que aparecen en una evaluación dedicada son ignorados por los usuarios cuando perciben que la tecnología es suficientemente buena como para ejecutar la tarea.

Un resultado muy interesante también es que el nivel de mareo es mínimo durante la ejecución de los casos de uso. Los casos de uso se diseñaron para minimizar en lo posible la aparición de mareo, usando exposiciones a la realidad virtual de un máximo de 20 minutos antes de descansar. No obstante, es notable que no existan apenas síntomas de mareo en estos escenarios.

4 Escenario de Regulación Emocional: tecnologías de Biomarcadores Digitales y Regulación Emocio-Cognitiva

4.1 Tecnologías de regulación emocional

Las tecnologías de regulación emocional estudiadas en el proyecto se han descrito en el entregable E3.2. Se implementaron clasificadores supervisados (SVM, Random Forest, Regresión Logística, XGBoost y LightGBM) para detectar miedo/no miedo a partir de EDA, HR y TEMP, evaluados con accuracy, F1-score y matriz de confusión bajo validación cruzada estratificada de 5 pliegues.

Se han utilizado tres métricas principales para evaluar el rendimiento de los clasificadores:

- **Exactitud (accuracy)**: porcentaje de predicciones correctas sobre el total.
- **F1-score**: media entre la precisión y el recall, especialmente útil en contextos con clases desbalanceadas.
- **Confusion Matrix**: es una tabla que muestra el número de predicciones correctas e incorrectas por clase. Es especialmente importante para diferenciar falsos positivos / negativos.

4.2 Validación con personas usuarias de la Fundación Juan XXIII

En una primera fase, se han evaluado los algoritmos sin utilizar las métricas de línea base, es decir, solo utilizando las métricas de cada ventana asociada a una votación.

Además se utilizó como método de división del conjunto de datos un simple *StratifiedKFold* con *random_state=42* y *n_split=5*. Esta manera de dividir los datos puede distribuir ventanas de la misma persona y sesión en ambos los conjuntos de entrenamiento y test, creando un sesgo.

Aún con sus limitaciones, esta es útil como primera versión para investigar estos modelos.

En la Tabla 10 un resumen de F1-Score y accuracy de los varios modelos, mientras en la Figura 76 se pueden ver la matriz de confusión de los varios métodos. En la Tabla 10 se han evidenciado en negrita los algoritmos que consiguen los mejores resultados, y se ha subrayado el modelo que mejor detecta estados de miedo.

Se quiere evidenciar como, los valores de accuracy y F1-score no permiten individualmente evaluar cual sea el mejor algoritmo, sino que tiene que compararse con las matrices de confusión. Esto es debido al fuerte desbalanceo entre clases, que no permite utilizar estas

medidas de manera fiable. Por ejemplo, en la Tabla 10 el modelo que logra la mejor accuracy y F1-score es el XGBoost, pero, mirando a la Figura 76 es evidente como esto algoritmo solo ha aprendido a predecir la clase mayoritaria más veces.

En este caso, comparando los resultados de la Figura 76 y la Tabla 10, el algoritmo que logra los mejores resultados es la regresión logística. El SVM linear es el modelo que mejor diferencia la clase de miedo comparado con los otros, esto viene al coste de una peor precisión en predecir la clase de no miedo.

Tabla 10: Accuracy y F1-Score de los varios modelos entrenados con StratifiedKFold y sin línea base

Modelo	**Accuracy**	**F1-Score**
SVM Linear	**0.507**	**0.270**
SVM poly	0.787	0.313
SVM rbf	0.653	0.297
SVM sigmoid	0.452	0.190
Random Forest	0.870	0.183
Logistic Regression	**0.562**	**0.266**
LightGBM	0.870	0.371
XGBoost	0.874	0.377

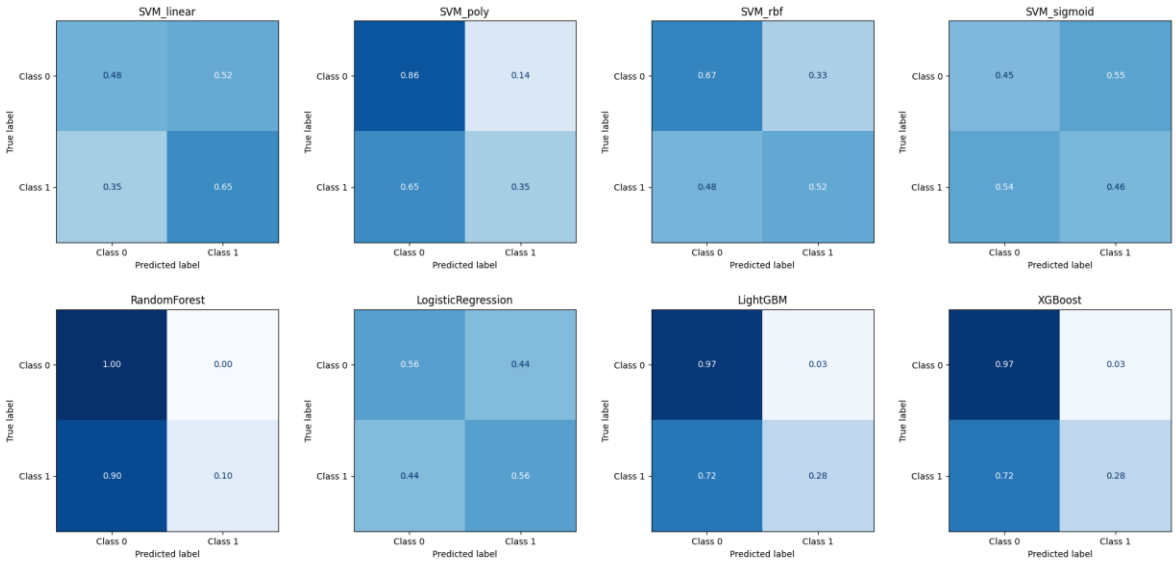


Figura 76: Matriz de confusión de los modelos entrenados con StratifiedKFold y sin línea base

Para evaluar de manera más fiable la validez de los modelos se eligió cambiar el método utilizado para dividir el conjunto de datos en entrenamiento y test, pasando desde *StratifiedKFold a GroupKFold,* pero manteniendo el número de pliegues a 5. Este último método, permite dividir los datos en entrenamiento y test asegurando que en los dos no haya ventanas que pertenecen a la misma sesión y persona.

En la Tabla 11 un resumen de F1-Score y accuracy de los varios modelos, mientras en la Figura 77 se pueden ver la matriz de confusión de los varios métodos.

Quitando el sesgo que introducía *StratifiedKFold,* todos los modelos empeoran sus resultados. Es importante notar, como, los dos modelos que consiguen los mejores resultados siguen siendo regresión logística y SVM Linear. En ambos los modelos, aunque haya un empeoramiento de las medidas de precisión, hay una mejora en la capacidad de predecir correctamente las fases de miedo.

Tabla 11: Accuracy y F1-Score de los varios modelos entrenados con GroupKFold y sin línea base

Modelo	**Accuracy**	**F1-Score**
SVM Linear	***0.492***	***0.272***
SVM poly	0.706	0.267
SVM rbf	0.625	0.214

SVM sigmoid	0.474	0.164
Random Forest	0.855	0.028
Logistic Regression	**0.546**	**0.261**
LightGBM	0.827	0.085
XGBoost	0.827	0.072

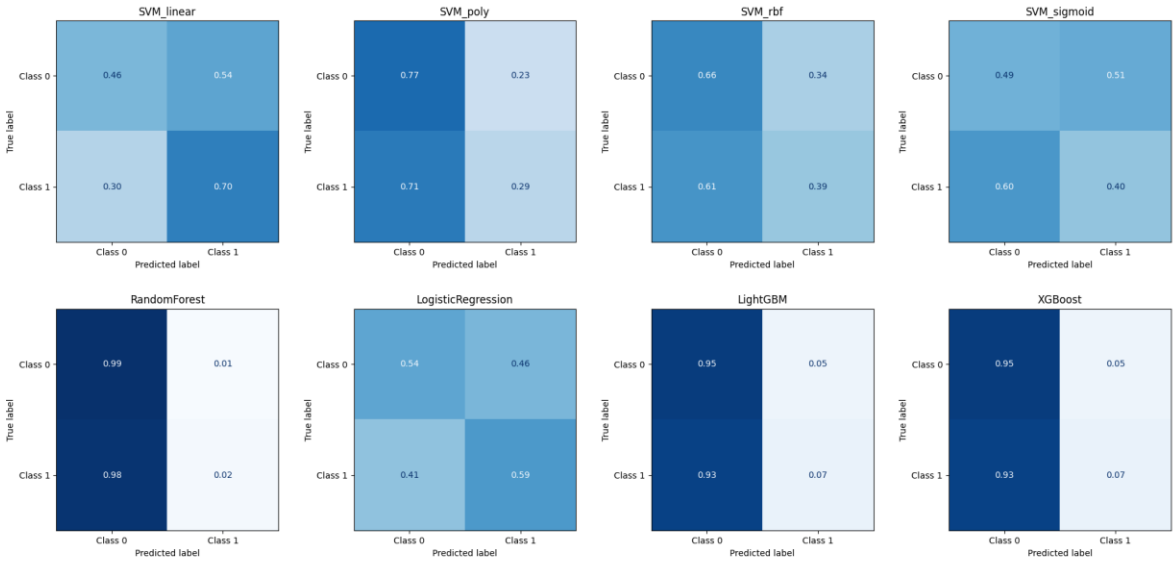


Figura 77: Matriz de confusión de los modelos entrenados con GroupKFold y sin línea base

En la Tabla 12 y la Tabla 13 se resumen los valores de F1-score y accuracy de los distintos modelos utilizando StratifiedKFold y GroupKFold, respectivamente. Asimismo, la Figura 78 y la Figura 79 muestran las matrices de confusión obtenidas en ambos casos.

La Figura 78 ilustra cómo todos los modelos mejoran su rendimiento bajo StratifiedKFold. Destaca en particular el SVM con kernel RBF, que logra la mejor clasificación general entre los dos estados, mientras que el SVM lineal sigue siendo el algoritmo que mejor discrimina específicamente las situaciones de miedo. Sin embargo, estos resultados —al igual que los de la Tabla 12 están sesgados por el uso de StratifiedKFold.

Al eliminar este sesgo mediante GroupKFold, los resultados empeoran, como puede observarse en la Figura 79 y en la Tabla 13. En este escenario, el SVM con kernel RBF pierde su capacidad de predecir correctamente ambas clases. Aun así, tanto el SVM lineal como la Regresión Logística mantienen un desempeño aceptable, logrando clasificar correctamente alrededor del 60 % de las muestras en ambas clases. En particular, la Regresión Logística se muestra más eficaz en la detección de estados de miedo.

Estos hallazgos ponen de relieve la dificultad de trasladar los buenos resultados obtenidos en entornos de laboratorio altamente controlados a contextos terapéuticos reales. No obstante, algunos modelos sí logran capturar diferencias en las métricas que permiten discriminar entre clases. En concreto, el SVM lineal y la Regresión Logística han alcanzado los mejores resultados en la mayoría de las pruebas. Es relevante destacar que, a pesar de que la clase de miedo representa solo un 14 % de las muestras, ambos modelos consiguen detectarla en más del 60 % de los casos. Este resultado es especialmente valioso, ya que, en un contexto terapéutico, es más importante identificar correctamente los estados de miedo, aun cometiendo errores en la clasificación de los estados de no miedo.

Tabla 12: : Accuracy y F1-Score de los varios modelos entrenados con StratifiedKFold y con línea base

Modelo	**Accuracy**	**F1-Score**
SVM Linear	**0.673**	**0.374**
SVM poly	0.755	0.322
SVM rbf	**0.704**	**0.390**
SVM sigmoid	0.423	0.163
Random Forest	0.900	0.472
Logistic Regression	0.697	0.383
LightGBM	0.881	0.468
XGBoost	0.889	0.546

a base

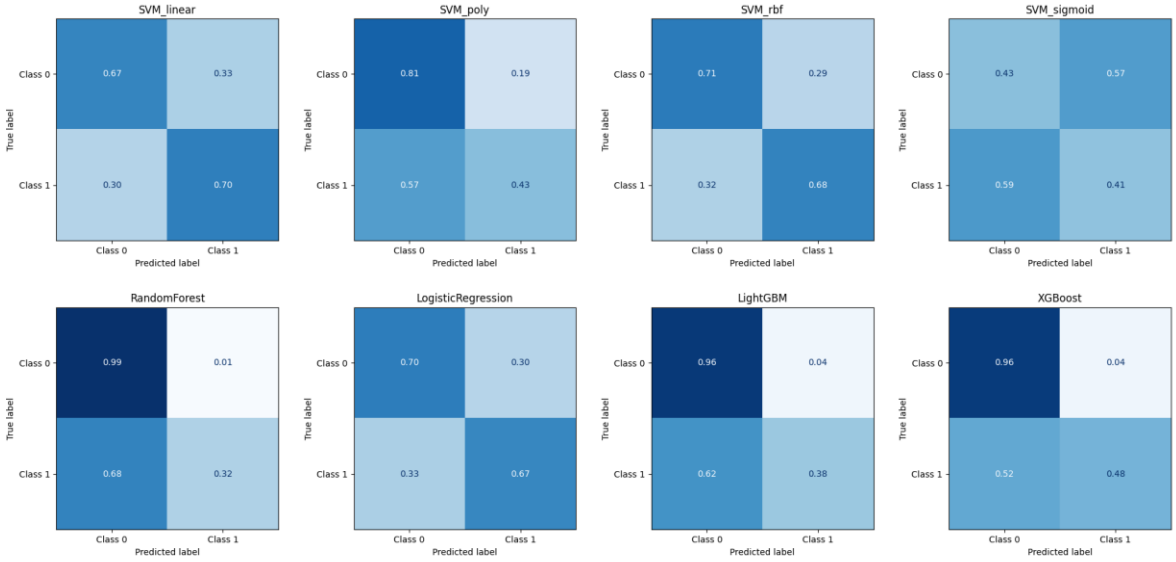


Figura 78: Matriz de confusión de los modelos entrenados con StratifiedKFold y con línea base

Tabla 13: Accuracy y F1-Score de los varios modelos entrenados con GroupKFold y línea base

Modelo	**Accuracy**	**F1-Score**
SVM Linear	0.604	0.289
SVM poly	0.636	0.152
SVM rbf	0.645	0.207
SVM sigmoid	0.440	0.145
Random Forest	0.854	0.014
Logistic Regression	**0.612**	**0.296**
LightGBM	0.837	0.000
XGBoost	0.826	0.033

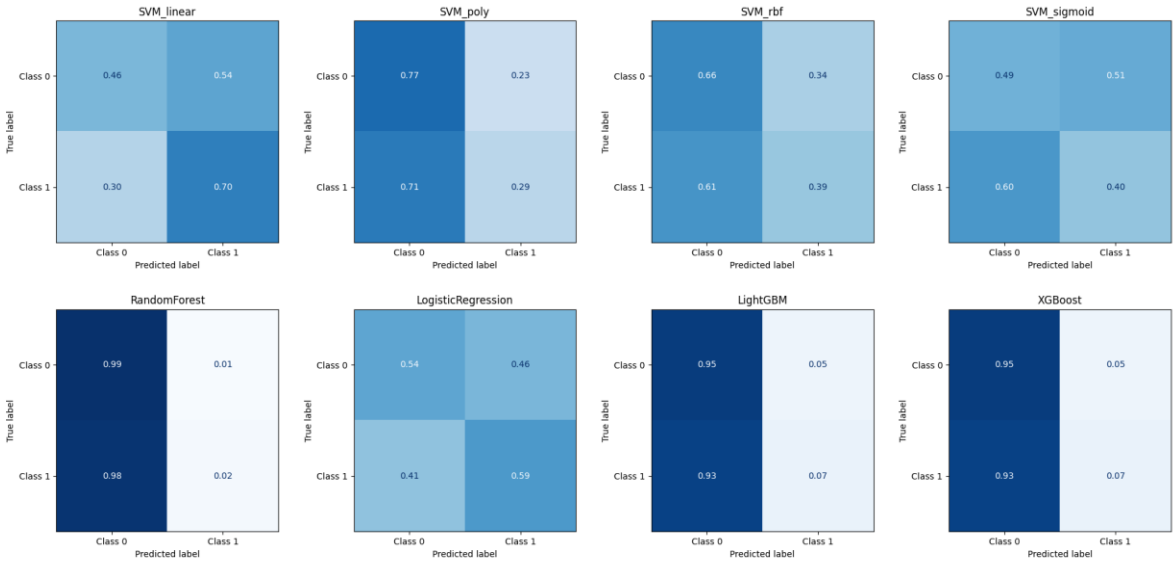


Figura 79: Matriz de confusión de los modelos entrenados con GroupKFold y con línea base

4.3 Medidas de rendimiento computacional

Además de evaluar el rendimiento en términos de métricas de acierto del algoritmo, se ha evaluado el tiempo de ejecución de cada uno de ellos en la plataforma hardware de referencia del proyecto, promediando para cada uno 10,000 inferencias consecutivas a partir de datos del usuarios reales enviados de forma aleatoraia.

Tabla 14: Tiempos de inferencia

Modelo	**Tiempo de inferencia (ms)**	**Inferencias / s**	**Usuarios /s**
SVM Linear	0,111	9023	1637
SVM poly	0,115	8662	1625
SVM rbf	0,132	7601	1583
SVM sigmoid	0,130	7703	1588
Random Forest	4,047	247	220
Logistic Regression	0,071	14060	1751
LightGBM	1,059	945	642
XGBoost	0,439	2277	1065

Para el cálculo de inferencias por segundo se ha tomado simplemente la inversa del tiempo de inferencia. Para el cálculo de usuarios por segundo que se pueden atender en el servidor, se ha añadido un tiempo adicional constante, estimado en 0.5 ms por petición, para recibir y procesara la petición, y generar y enviar la respuesta.

4.4 Requisitos e indicadores de rendimiento

La siguiente tabla resume los resultados presentados en este capítulo.

Tabla 15: Resumen de análisis de KPIs para el escenario de regulación emocional

KPIs		Requisitos	**SVM Linear**	**Logistic Regression**
Sesión	Núm. Usuarios	[1-100]	>1500	>1500
	Velocidad Usuario	Cualquiera	Cualquiera	
	Distancia Usuarios	Cualquier lugar	Cualquiera	
Tasa de datos	Sensores	0.1 Mbps UL	~ 10 kbps UL	
	Datos	0.1 Mbps DL	< 10 kbps DL	
Latencias	Latencia round-trip	< 2 s	< 100 ms	
	Frecuencia IA	1 infer/s	9000 infer/s	14000 infer/s
Acierto	Accuracy	-	0.604	0.612
	F1-score	-	0.289	0.296

Se han seleccionado los dos algoritmos con mejores tasas de acierto, como se ha descrito anteriormente. Como se indicó en el apartado anterior, ambos pueden dar servicio a más de 1500 usuarios simultáneos, asumiendo que cada uno emite una petición por segundo. Como se puede ver en la tabla, ambos algoritmos cumplen holgadamente los requisitos previstos de tasa de datos y capacidad de proceso establecidos al principio del proyecto.

5 Funcionalidades XR en el co-diseño de los Casos de Uso

Este capítulo busca no solo conceptualizar las soluciones innovadoras de cada uno de los casos de uso, sino también resumir cómo se ha co-diseñado y validado su eficacia en sus contextos reales. En este sentido, presentamos la información detallada de los tres casos de uso: terapia conductual, telepresencia y teleformación con la sección de terapia cognitiva y la de cocina. De esta forma, ilustran la puesta en marcha y la aplicación práctica de las propuestas así como las conclusiones generales del impacto generado.

Concretamente, se resumen:

- **Caracterización del caso de uso desde el punto de vista de personas usuarias**.
- **Fases de Co-diseño y Tecnologías Implicadas**.
- **Metodología de Evaluación.**
- **Duración y Participantes**.
- **Conclusiones Generales**.

A través de esta exposición estructurada, se busca ofrecer una visión clara y fundamentada de cómo las tecnologías XR han sido aplicadas en entornos reales.

5.1 Terapia conductual

Título
Terapia conductual: desensibilización sistemática.
Caracterización del caso de uso desde el punto de vista de personas usuarias
Considerando las principales tecnologías en investigación en el proyecto (es decir, telepresencia de vídeo inmersiva, técnicas de embodiment basadas en IA y escenarios de RV) y las dificultades relacionadas con las fobias (alturas, animales, escaleras, entre otras) que afectan a un número significativo de usuarios de COFOIL, el miedo a subir y bajar diferentes tipos de escaleras fue identificado como el objetivo más adecuado a abordar en esta terapia conductual. Este miedo impactaba significativamente la calidad de vida de las personas que lo poseen, impidiéndoles usar las escaleras de la Fundación, viajar en metro o visitar centros comerciales. De hecho, esto era un desafío diario para algunos de ellos, lo que dificultaba su autonomía. Así, nuestro enfoque fue aplicar la técnica tradicional de Desensibilización Sistemática (DS) a través de la tecnología XR para ayudarles a superar este miedo.

La terapia consideró varias fases, incluyendo la creación de una jerarquía de ansiedad individualizada utilizando fotos reales de las escaleras temidas y una fase de relajación con gafas de RV para enseñar a los usuarios la respiración diafragmática y familiarizarse con la tecnología.

Seguidamente, la fase de exposición introdujo gradualmente a los usuarios, a través de un HMD, a enfrentarse a los estímulos relacionados con el miedo, los cuales fueron mostrados en un orden personalizado y representaban entornos realistas.

Mediante una exposición repetida y controlada, este proceso buscaba aprender técnicas para reducir la ansiedad al enfrentarse a los estímulos temidos, haciéndoles controlar la situación y poder aplicar este conocimiento a situaciones de la vida real.

Fases de co-diseño del caso de uso con las tecnologías implicadas

Etapas iniciales. El diseño inicial del sistema incluía escenarios tanto relajantes como inductores de miedo. Los escenarios relajantes consistían en vídeos de 360° de entornos como playas, montañas y hogares, seleccionados de recursos públicos disponibles.

Los escenarios relacionados con el miedo consistían en grabaciones de 360° de entornos de escaleras del mundo real que los participantes encontraban en su vida diaria, así como escenas virtuales en 3D que replicaban estos entornos. Durante las sesiones, un terapeuta ubicado en la misma sala monitorizaba a los participantes y controlaba las sesiones a través de una interfaz, ajustando los escenarios de exposición en respuesta a las reacciones de cada participante.

Se utilizaron biosensores para recopilar datos de los participantes. Para asegurar que las señales recogidas fueran lo más fiables posible, se estableció un procedimiento de preparación antes de cada sesión. Se instruyó a los participantes a usar la pulsera aproximadamente 10 minutos antes de comenzar la actividad, permitiendo que el dispositivo generara una línea base estable. Durante este periodo, los participantes continuaron con sus actividades habituales en la Fundación Juan XXIII Roncalli hasta que comenzaba la sesión. Luego, justo antes de que comenzara la misma, se colocaba el sensor EMG a la persona participante, ya que este dispositivo no requería ningún periodo de línea base.

Mejoras iterativas. El diseño inicial sirvió como punto de partida para un proceso de co-diseño iterativo, incorporando realimentación continua de participantes y terapeutas, lo que llevó al sistema final implementado.

- **Transiciones entre escenas XR:** inicialmente, en la aplicación inmersiva, las transiciones entre escenas eran abruptas, lo que podía causar mareos o una exposición repentina a un escenario temido. Esto se abordó añadiendo una transición de fundido de entrada/salida (fade-in/fade-out), proporcionando un cambio más suave. Además, a lo largo de la fase de exposición, la transición de

escenarios relajantes a escenarios de escaleras temidas era inicialmente demasiado abrupta. Se introdujeron escenarios intermedios de "escaleras amigables", que incluían escaleras más pequeñas y visualmente no amenazantes en entornos exteriores. A pesar de su intensidad reducida, estos escenarios eran temidos por los participantes debido a la ausencia de barandilla, lo que resalta la importancia de un diseño cuidadoso del escenario teniendo en cuenta cada elemento.

- **Grabaciones de audio:** durante la fase de relajación, los participantes requerían orientación para practicar la respiración diafragmática sin que el terapeuta hablara, ya que esto podría interrumpir la inmersión. Por lo tanto, se crearon grabaciones de audio para guiar los ejercicios de respiración dentro de los escenarios virtuales. Adicionalmente, se añadió una grabación de audio para ayudar a los participantes a transitar de vuelta a la realidad después de las sesiones, ya que algunos tendían a relajarse demasiado en las primeras sesiones durante esta fase.
- **Interfaz de evaluación:** la evaluación se realizó tanto dentro como fuera del entorno de XR. Dentro del entorno inmersivo, el objetivo era evaluar cómo se sentían los participantes sin interrumpir la sensación de inmersión. Se co-diseñó una escena de evaluación con los terapeutas y los participantes, que consistía en tres caras verticales posicionadas ligeramente a la izquierda, como se muestra en la Figura 80a. Esta escena, solo con las caras y sin fondo, se superpuso al escenario temido en curso, como se ilustra en la Figura 68b. Los participantes indicaban cómo se sentían moviendo la cabeza, evitando la necesidad de interacción con un mando.

  Se probaron varias configuraciones de diseño con participantes de la Fundación Juan XXIII Roncalli, comparando orientaciones verticales versus horizontales, ubicaciones centradas versus laterales, y la inclusión u omisión de un rayo guía.

  Para preguntar sobre el estado emocional de los participantes sin la intervención del terapeuta, se incluyeron preguntas de audio pregrabadas.

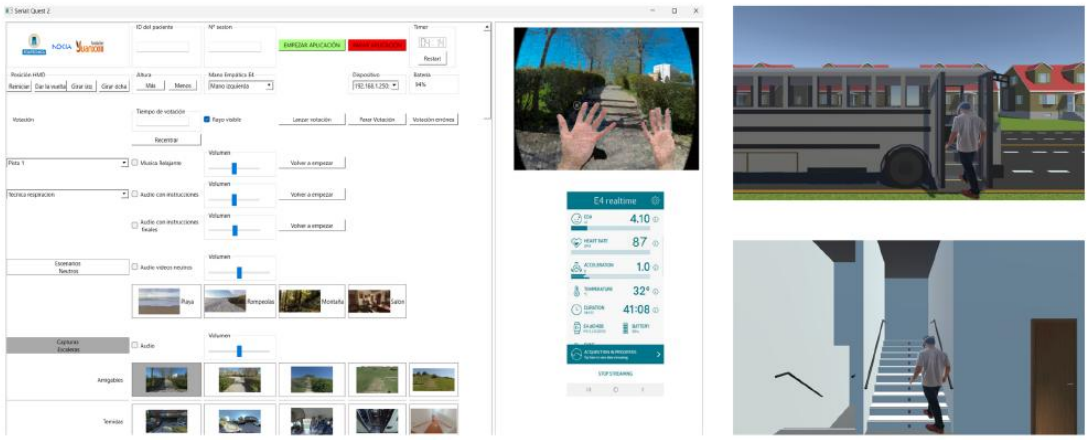
  Antes de comenzar la fase de exposición, se añadió también una breve fase de entrenamiento para que los/as participantes se familiarizaran con el mecanismo de votación.

(a) Voting scenario with three vertical voting faces located slightly in the left side.

(b) Real-time voting on top superposed to the feared scenario.

Figura 80: Sistema de evaluación en el entorno inmersivo del caso de uso de terapia cognitiva.

- **Interfaz de control:** inicialmente, para la monitorización de las sesiones, la interfaz de control se diseñó sin una ventana de visualización (viewport) que permitiera ver lo que la persona con las gafas de RV está observando. Sin embargo, los terapeutas enfatizaron la importancia de visualizar la vista de los participantes en tiempo real para mejorar aspectos como la escena actual, las interacciones, los resultados de las votaciones y los datos de biosensores relacionados con los estados emocionales de las personas participantes. En consecuencia, se incorporó una ventana de visualización a la interfaz, junto con un temporizador para ayudar a los terapeutas a gestionar la duración de la exposición y un resumen de las bioseñales provenientes de la pulsera (ver Figura 81).



(A) (B)

Figura 81: Sistema de terapia conductual. La interfaz de control del terapeuta (A) proporciona monitorización en tiempo real y control de la sesión, mientras que las

escenas virtuales diseñadas (B) ofrecen a las personas participantes exposición a los escenarios temidos.

Implementación del Enfoque Final. El enfoque final consistió en una terapia multifase apoyada por un sistema inmersivo diseñado específicamente para guiar y monitorizar las sesiones. Siguiendo la técnica de Desensibilización Sistemática (DS), la terapia incluyó una fase de relajación, una fase de entrenamiento de votación y una fase de exposición que involucraba tanto escaleras amigables como temidas.

El sistema desarrollado, mostrado en la Figura 81, consta de dos componentes principales: una interfaz de control para el terapeuta, mostrada en el lado izquierdo, y una aplicación ejecutándose en unas gafas de RV, mostrada en el lado derecho, que presenta dos escenarios virtuales temidos de las personas participantes.

El monitor del terapeuta permitía la gestión de la sesión en tiempo real. Controlaba la aplicación del HMD, la selección de escenas, la votación y las señales de audio, y proporcionaba acceso a la vista del usuario y a los datos del Empatica E4, como la frecuencia cardíaca (HR) y la actividad electrodérmica (EDA). Mientras tanto, la aplicación del HMD ofrecía a los usuarios experiencias inmersivas de escenarios neutrales y de escaleras, permitiendo una exposición progresiva según las jerarquías de miedo realizadas de manera individual. Los escenarios de escaleras se presentaban a través de vídeos 360°, donde las personas participantes permanecían sentadas, y escenas virtuales completamente interactivas en las que podían moverse. El sistema de votación explicado, basado en movimientos de cabeza, se lanzó para evaluar los niveles de ansiedad durante la exposición a los escenarios virtuales.

Caracterización del caso de uso desde el punto de vista técnico

Dispositivos involucrados: Meta Quest 2, Empatica E4, Cometa Pico Blue

Se seleccionaron las gafas de VR Meta Quest 2 debido a su compatibilidad con la tecnología de segmentación de cuerpo implementado. Este sistema utiliza una cámara externa montada en las gafas de VR para capturar imágenes, que luego son procesadas por una red neuronal para segmentar el cuerpo de la persona que lleva las gafas. Esta capacidad permitió a los participantes ver su propio cuerpo dentro del entorno virtual, una característica esencial para tareas como acercarse a las escaleras virtuales.

Se emplearon biosensores para recopilar datos con el fin de analizar las respuestas emocionales de los participantes al finalizar el proceso de terapia. Durante la fase de diseño, los terapeutas enfatizaron la necesidad de utilizar sensores mínimamente invasivos para garantizar la comodidad de los participantes y preservar el valor terapéutico de las sesiones. Específicamente, aconsejaron evitar dispositivos con múltiples cables o una apariencia del área de la medicina, ya que estos podrían desencadenar ansiedad en las personas que asistían a la terapia. Los biosensores seleccionados fueron la pulsera

Empatica E4 y los sensores Cometa Pico Blue, ya que miden indicadores fisiológicos asociados con las respuestas al estrés incluyendo la frecuencia cardíaca (HR), la actividad electrodérmica (EDA) y la electromiografía (EMG).
Metodología de evaluación
Durante la sesión de terapia se hacían preguntas de control en las que se evaluaba cada uno de los estímulos presentados. En todos los casos de uso, se utilizaron cuestionarios finales basados en adaptaciones según el caso de uso, contexto y comprensión de: Escala de Usabilidad del Sistema (SUS) [11] Cuestionario de Presencia Espacial y Social propuesto en [5] NASA-TLX [12] VSR [13] Estos abordan aspectos distintos de la experiencia del usuario: usabilidad, sensación de presencia y experiencia de calidad general, carga de trabajo mental y física, y síntomas de vértigo o mareo. La terapeuta mantuvo un diario de sesión para cada participante. El diario sirvió para propósitos prácticos, documentando problemas técnicos, dificultades de los participantes y observaciones necesarias para que los desarrolladores mejoraran la usabilidad y la inclusividad, así como para facilitar un seguimiento del progreso individual. Se utilizó una versión adaptada del Inventario de Ansiedad de Beck (BAI) [14]. Mientras que la prueba estándar del BAI mide los síntomas físicos de ansiedad asociados con respuestas de pánico experimentadas durante la semana previa a la evaluación, la prueba que realizamos fue adaptada para referenciar explícitamente los síntomas de ansiedad experimentados al enfrentarse a escaleras; además, el lenguaje fue adaptado para ser más inclusivo. Nótese que las evaluaciones se realizaron antes del inicio de la terapia para verificar la idoneidad de los participantes y establecer una línea de base. Todas las pruebas se repitieron al final de la terapia para identificar posibles mejoras. Adicionalmente, al terminar la terapia las personas fueron a las escaleras que más miedo les provocaba y consiguieron enfrentarse a ellas y bajarlas, corroborando así la eficacia de la terapia en el mundo real.
Duración (comienzo con participantes) y número de participantes
9 personas y duración de 6 meses.
Conclusiones generales

El resultado más importante de este caso de uso fue la clara mejora en la calidad de vida (QoL) de los participantes. Todos los participantes mostraron una reducción en sus puntuaciones en el BAI adaptado, pasando de ansiedad moderada-severa a ansiedad mínima o nula. Un análisis estadístico de estos resultados apoyó la efectividad de la intervención. También se observaron mejoras en el mundo real: los participantes pudieron usar de forma independiente las escaleras que antes temían, y tanto los participantes como sus familias reportaron una mayor autonomía.

En cuanto a la experiencia y los métodos de evaluación, si bien los participantes podían entender y responder más fácilmente los cuestionarios adaptados, aún requerían un apoyo sustancial por parte de los terapeutas. Los terapeutas a menudo reformulaban las preguntas varias veces para asegurar la comprensión y guiaban a los participantes a reflexionar sobre su experiencia antes de seleccionar la respuesta más precisa. Los resultados de los cuestionarios fueron limitados y no concluyentes. Más allá del pequeño tamaño de la muestra, la escala de tres niveles probablemente contribuyó a esta limitación al reducir la sensibilidad y restringir la precisión con la que se podían capturar las experiencias de los participantes. Además, otros dos factores probablemente afectaron los resultados de los cuestionarios. Primero, un efecto "WoW", común cuando las personas experimentan tecnologías inmersivas por primera vez, pudo haber hecho que los participantes calificaran todo de manera muy positiva, dificultando la detección de sus preferencias reales. Segundo, algunas preguntas se dirigían a conceptos generales como la "presencia", que pueden ser difíciles de entender o juzgar para personas con discapacidad intelectual. Las notas tomadas durante la sesión fueron esenciales para identificar y corregir errores en los datos recopilados. Esto también proporcionó información contextual importante que aclaró los comportamientos específicos de cada persona.

El análisis objetivo se centró en las señales de HR y EDA, proporcionando resultados débiles. Las señales del Empática E4 fueron altamente sensibles al movimiento, produciendo ruido que redujo la fiabilidad de las características extraídas. Además, el estudio priorizó una alta validez ecológica, recopilando datos en condiciones de terapia realistas en lugar de entornos de laboratorio controlados. Esto aumentó la variabilidad y limitó aún más la utilidad de las bioseñales.

Por último, solo un participante abandonó las sesiones de terapia. Este usuario experimentó mareos durante la primera sesión con el HMD y decidió no continuar. La baja tasa de abandono sugiere que el diseño centrado en el usuario de las experiencias, las sesiones iniciales de familiarización y el flujo de trabajo general de la terapia ayudaron a los participantes a adaptarse a la tecnología y a mantenerse cómodos durante todo el proceso.

5.2 Telepresencia

Título
Terapia conductual: desensibilización sistemática
Caracterización del caso de uso desde el punto de vista de personas usuarias
Uno de los objetivos fundamentales cuando hablamos de autonomía e integración de personas con discapacidad intelectual es la posibilidad de vivir de forma independiente. COFOIL capacita a los usuarios para vivir en apartamentos supervisados. De esta manera, varias personas con DI comparten un apartamento y reciben apoyo personalizado según las actividades diarias que necesitan practicar y/o que requieren supervisión. Sin embargo, proporcionar este apoyo exige importantes recursos humanos, de tiempo y económicos. Para abordar estas limitaciones, se propone el uso de tecnología XR en lugar de la supervisión presencial para apoyar la realización de tareas diarias. Esto evitaría el desplazamiento de un profesional, ahorrando recursos y tiempo. Además, de esta forma, el mismo profesional podría apoyar varios apartamentos en el mismo día. Por otro lado, explorar esta tecnología en este caso de uso es interesante como un primer paso para ver si podría utilizarse en situaciones de emergencia, ya que permite al profesional, simplemente girando la cabeza y, sobre todo, sin depender de nadie para enfocar una cámara, tener una vista global de toda la escena donde se encuentra el dispositivo de captura. Utilizando comunicaciones 360°, este caso de uso exploró la supervisión y guía para la realización de tareas diarias en un apartamento supervisado donde convivían tres usuarios un día a la semana durante un año escolar. La participación requería que el profesional que trabajaba regularmente con ellos aceptara probar el sistema y que los perfiles cognitivos de los usuarios permitieran una implicación segura en las sesiones. Las tareas practicadas fueron las que los usuarios necesitaban: organizar un menú semanal, lavar la ropa, tender la ropa y cocinar. Estas se llevaron a cabo progresivamente, avanzando hacia la tarea de cocinar solo una vez que se garantizó la usabilidad del sistema, para evitar poner en riesgo a los usuarios. Los objetivos principales de este caso de uso fueron: 1) fortalecer las habilidades personales, domésticas y sociales de los residentes; 2) promover la conciencia y el uso de los recursos comunitarios por parte de los residentes en función de sus necesidades específicas; 3) activar la red de apoyo natural de los residentes; 4) introducir a los residentes en el uso de la tecnología como una herramienta facilitadora y de apoyo.

Fases de co-diseño del caso de uso con las tecnologías implicadas
Etapas Iniciales. En cuanto a la tecnología, examinamos las características que debía proporcionar el prototipo búho. Estas incluyen factores técnicos como la calidad de audio, vídeo y la latencia, así como aspectos relacionados con la interacción, incluyendo cómo los usuarios interactúan con el dispositivo, su percepción del mismo y sus limitaciones para permitir el soporte remoto por parte del profesional. Durante esta fase, la implicación familiar fue clave, ya que el búho se instaló en un apartamento de tres personas con discapacidad intelectual. Posteriormente, se realizó un análisis de configuración en el entorno. Este análisis fue crucial asegurar que no hubiera cableado en los espacios habitables para garantizar así la seguridad de los residentes. Esto impidió el uso de conectividad Ethernet y requirió una conexión Wi-Fi suficientemente estable para asegurar una calidad de llamada adecuada para la ejecución de tareas. Además, el búho requería suficiente portabilidad y compacidad. La portabilidad era esencial para permitir su movimiento según las instrucciones del profesional. Para facilitar esto, el dispositivo se montó en un trípode con ruedas, lo que permitía un fácil desplazamiento por todo el apartamento. Para asegurar una manipulación correcta por parte del usuario, una pata del trípode se pintó de rojo, sirviendo como asa designada para tirar. Esto también funcionó como punto de referencia para orientar el trípode, asegurando que la tablet que mostraba el avatar del profesional siempre mirara hacia los usuarios. Además era necesario que el búho fuera fácil de montar y desmontar, y que ocupara un espacio mínimo, ya que permanecía permanentemente en el apartamento. **Mejoras Iterativas.** Una herramienta clave mejorada en este caso de uso ha sido la lupa (véase la Figura 82), que se emplea para la inspección detallada de regiones específicas de la escena y sirve como soporte crucial para determinadas fases de las tareas. La herramienta de la lupa es un cliente móvil auxiliar integrado con el búho, diseñado para permitir a los usuarios locales capturar y transmitir vídeo en tiempo real con sus teléfonos. Esta aplicación transmite la secuencia al servidor del sistema de telepresencia. El servidor, a su vez, retransmite esta señal de vídeo a todos los usuarios remotos conectados, quienes la perciben como una ventana flotante dentro de su escena virtual, así como con controles interactivos.

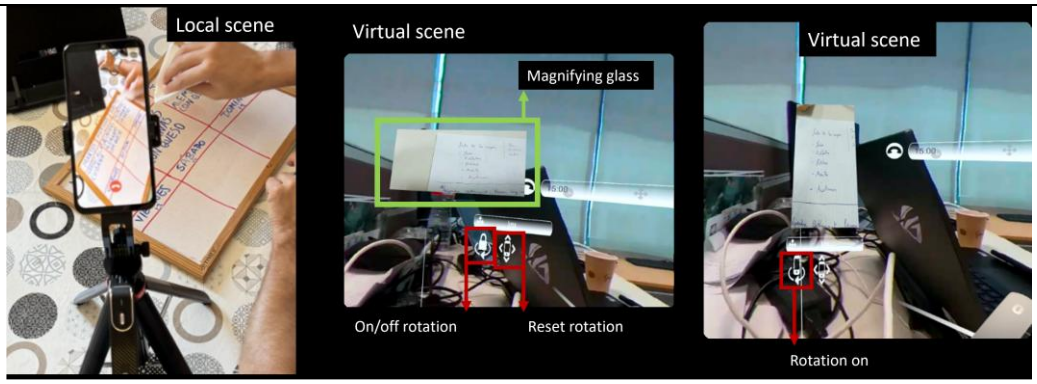


Figura 82: La figura ilustra la herramienta del Búho: La lupa. La primera imagen (izquierda) muestra un ejemplo de cómo se utiliza un móvil para capturar vídeo en la escena local, específicamente durante la tarea del menú semanal. Las dos imágenes siguientes muestran cómo se ve la ventana flotante en el entorno inmersivo.

Inicialmente, la tasa binaria del vídeo transmitido a través de la lupa se aumentó a 1Mbps, mejorando significativamente la calidad. Además, para mejorar la usabilidad desde la perspectiva del profesional, se implementaron dos controladores (ver Figura 82):

- **El botón de rotación on/off** permite habilitar y deshabilitar la rotación de la ventana flotante de la lupa. Cuando está habilitado (Rotación activada), el usuario remoto puede rotarla como desee; cuando está deshabilitado, la ventana de la lupa se puede mover por la escena con su ángulo de rotación fijo.
- **El botón de restablecer rotación** permite que la ventana de la lupa vuelva a su posición y rotación iniciales.

Además, al apuntar a la ventana de la lupa y mover el controlador más cerca o más lejos, se puede ajustar el Zoom en el vídeo.

Implementación del enfoque final. La implementación final, resultado de las mejoras identificadas en este caso de uso de telepresencia, ha dado lugar a Snowl, la versión evolucionada del Owl o búho. Snowl, al igual que el búho, es un prototipo de fabricación propia.

Caracterización del caso de uso desde el punto de vista técnico

Dispositivos involucrados: Meta Quest 2, The Owl

La experiencia de telepresencia se basó en el Owl o búho, un prototipo que permite comunicaciones de 360° en tiempo real (ver Figura 83). El sistema captura y transmite la escena de 360° a los participantes remotos. Estos pueden asistir a la comunicación a través de un HMD (Meta Quest 2 utilizado en este caso de uso) o una tableta, lo que les permite percibir la escena como si estuvieran físicamente donde se encuentra el elemento de captura. En este caso, el profesional utilizaba el HMD. Al mismo tiempo, los usuarios locales

(los que están alrededor del Owl) pueden ver a los participantes remotos, representados como avatares, en una pantalla. Los avatares mueven sus cabezas y labios sincronizados con los movimientos de los usuarios remotos, en este caso, el profesional. El audio es bidireccional. Esta comunicación permitió al profesional guiar de forma remota a los convivientes para realizar tareas diarias.

(a) Owl. (b) Snowl.

Figura 83: Dispositivos de Telepresencia

Metodología de evaluación

En todos los casos de uso, se utilizaron cuestionarios finales basados en adaptaciones según el caso de uso, contexto y comprensión de:

Escala de Usabilidad del Sistema (SUS) [11]
Cuestionario de Presencia Espacial y Social propuesto en [5]
NASA-TLX [12]
VSR [13]

Estos abordan aspectos distintos de la experiencia del usuario: usabilidad, sensación de presencia y experiencia de calidad general, carga de trabajo mental y física, y síntomas de vértigo o mareo. El cuestionario se adaptó para los convivientes y para el profesional dependiendo de los factores a explorar en cada uno de los roles.

En este caso de uso en concreto se añadieron preguntas concretas de la tecnología así como de la usabilidad de la lupa y desempeño de la tecnología.

Se realizó una entrevista semiestructurada con el profesional aproximadamente a la mitad de las sesiones. El formato le permitió expresar libremente observaciones e impresiones sobre la tecnología y las experiencias de los participantes.
Duración (comienzo con participantes) y número de participantes
3 personas y duración de 6 meses.
Conclusiones generales
La tecnología de asistencia remota, el búho, se ha integrado y utilizado con éxito durante varios meses en un contexto de apoyo en la vida independiente, permitiendo a los residentes realizar tareas diarias con ayuda remota. De hecho, todos los participantes mostraron mejoras claras en el nivel de ayuda requerido para completar las tareas diarias. Un factor significativo que contribuyó a la experiencia positiva fue el fuerte sentido de presencia facilitado por la tecnología. Según el profesional, tanto él ("Siento que estoy allí con ellos") como los residentes ("ellos sienten mi presencia") mejoraron la interacción y facilitaron la guía durante la ejecución de las tareas. De hecho, se encontraron dos ventajas claras en comparación con las videollamadas 2D: • La escena de 360° proporciona al profesional una vista completa del entorno, lo que permite dar instrucciones precisas y tener una conciencia situacional total. Esto se considera crítico para la gestión de incidentes y emergencias. • La implementación de esta tecnología en apartamentos supervisados reducirá los costes y recursos asociados con la provisión de apoyo a personas en riesgo de vulnerabilidad psicosocial Los resultados del cuestionario indicaron que el educador encontró el sistema efectivo, ya que permitía el control total del apartamento remoto mediante movimientos de cabeza, replicando interacciones de la vida real. Las instrucciones que transmitía fueron generalmente fluidas, aunque requirieron más esfuerzo ya que las tareas no podían ser demostradas como sería en persona. Se observaron desafíos de usabilidad: el sistema requería soporte técnico en lugar de ser "plug-and-play" tanto para los participantes como para el educador, y el uso prolongado del HMD podía causar incomodidad. Los cuestionarios capturaron algunos resultados, pero ciertas preguntas parecían irrelevantes tanto para los participantes como para los educadores. Futuras iteraciones podrían eliminar elementos innecesarios y centrarse en aspectos clave. Además, los participantes expresaron frustración al responder las mismas preguntas cada semana y percibieron que sus respuestas probablemente no cambiarían. Investigar el uso de

cuestionarios más cortos o entrevistas semiestructuradas breves podría ayudar a abordar este problema. Los aportes recopilados durante la ejecución del caso de uso se utilizaron para reunir mejoras para el prototipo búho y su evolución a Snowl.

5.3 Teleformación: terapia cognitiva

Título
Terapia cognitiva
Caracterización del caso de uso desde el punto de vista de personas usuarias
El entrenamiento cognitivo implica estrategias y actividades destinadas a mantener o mejorar procesos como la atención, la memoria y las funciones ejecutivas. Para las personas con discapacidad intelectual, desempeña un papel educativo clave al fomentar habilidades adaptativas y el funcionamiento diario, y al fortalecer funciones ejecutivas como la atención y el control inhibitorio. La atención apoya el aprendizaje y la participación social, mientras que el control inhibitorio regula los impulsos y las emociones. Cuando estas funciones están comprometidas, la autonomía y el rendimiento laboral se ven afectados, haciendo esencial una estimulación estructurada y sostenida. En este sentido, las tecnologías XR ofrecen herramientas innovadoras para apoyar el entrenamiento cognitivo, proporcionando experiencias atractivas y personalizadas. Adicionalmente, la XR permite entornos seguros y controlados adaptados a las necesidades individuales, mejorando la motivación y la implicación emocional. Teniendo en cuenta las áreas de trabajo en las que la Fundación Juan XXIII Roncalli apoya la inserción laboral de sus usuarios, se desarrolló una aplicación gamificada en VR que incluye tareas realistas relacionadas con profesiones como la de camarero/a o reponedor/a. Así, se lleva a cabo un programa de entrenamiento para fomentar su inserción laboral.
Fases de co-diseño del caso de uso con las tecnologías implicadas
En esta experiencia se diseñaron entornos inmersivos que simulaban situaciones de trabajo reales seleccionados siguiendo un proceso colaborativo. Se preguntó a los participantes qué entornos de trabajo les resultaban más atractivos y, basándonos en sus aportaciones y las opiniones de los terapeutas, seleccionamos dos escenarios: un supermercado y una cafetería (ver la Figura 84). Ambos escenarios se implementaron como entornos virtuales totalmente interactivos, permitiendo a los participantes moverse libremente e interactuar con los objetos como si fueran reales. Los escenarios se diseñaron para completar cuatro tareas, dos por cada escenario. En la cafetería, las tareas se

diseñaron para recoger platos sucios y servir pedidos, mientras que en el escenario del supermercado, las tareas se centraron en reponer productos en los estantes y preparar pedidos. Cada tarea tenía seis niveles de dificultad, lo que permitía una mejora progresiva de las habilidades cognitivas de las personas participantes.

Figura 84: Escenarios en VR de una cafetería (izquierda) y un supermercado (derecha).

Durante la realización de las tareas, un aspecto importante destacado por los terapeutas fue la necesidad de proporcionar realimentación sobre las acciones de los participantes. Así, el juego se diseñó para ofrecerla en tiempo real para cada interacción: se mostraba un tic verde cuando una acción se completaba con éxito, mientras que los errores se indicaban con una cruz roja. También se asociaron a señales de audio para reforzar la comprensión: se reproducía un sonido positivo para las acciones correctas, mientras que una señal diferente indicaba los errores. Además, dado que algunos participantes tenían habilidades de lectura limitadas, se grabaron guías de audio para cada tarea, proporcionando tanto instrucciones como señales para apoyarlos durante las sesiones.

(a) Success feedback with a tick while leaving correctly a drink in a table.

(b) Error feedback with a cross while taking wrong a drink from the bar.

Figura 85: realimentación en entorno virtual positivo (tic verde) y negativo (cruz roja).

Para la interacción dentro de los entornos virtuales se consideraron inicialmente dos métodos: el uso de mandos o las manos. Después de realizar pruebas con participantes del centro, decidimos que la interacción se llevaría a cabo utilizando los mandos, específicamente el botón de agarre (grip button) para coger objetos. Esta decisión se basó en cómo los participantes sostenían el controlador.

Antes de comenzar los niveles de las tareas, se diseñó un nivel introductorio sencillo para ayudar a los participantes a familiarizarse con los mandos y aprender a coger objetos. La primera sesión para cada participante se dedicó a este nivel de entrenamiento.

Para supervisar las sesiones, la terapeuta utilizó una interfaz de control similar a la empleada en la experiencia de terapia conductual. La interfaz contenía controles de sesión y la visualización de la vista del HMD. Sin embargo, dado que en este contexto no se requería la monitorización de biosensores, esta vista se eliminó de la interfaz.

Basándose en los conocimientos obtenidos de la experiencia de terapia conductual, se refinó el procedimiento para colocar la pulsera Embrace Plus. Se instruyó a los participantes para que usaran la pulsera aproximadamente 10 minutos antes de que comenzara la sesión. A diferencia del protocolo anterior, permanecieron tranquilamente en la misma habitación con el terapeuta durante este período de referencia, lo que contribuyó a una señal de referencia más estable y fiable. El procedimiento de colocación y monitorización del sensor EMG se mantuvo sin cambios.

Mejoras Iterativas. El sistema inicial evolucionó a través de sucesivas iteraciones de diseño para abordar las necesidades de los participantes, asegurando que la versión final cumpliera tanto con los requisitos terapéuticos como de usabilidad.

- **Diseño iterativo de niveles:** una vez definidas las tareas, el diseño de los niveles para ambos escenarios siguió un proceso iterativo de diseño y desarrollo. Los niveles se probaron repetidamente con terapeutas y participantes de la Fundación Juan XXIII Roncalli (diferentes de los involucrados en esta experiencia) para asegurar que la progresión en la dificultad de las tareas fuera apropiada y que los participantes pudieran comprender y completar las tareas. Durante estas sesiones de prueba, se introdujeron ajustes menores para alinear las interacciones y los resultados con los objetivos terapéuticos.
- **Necesidad de sesiones de familiarización con la tecnología:** durante las primeras sesiones, algunos participantes se adaptaron bien a la tecnología, mientras que otros experimentaron dificultades. Esto llevó a una modificación de la sesión inicial, que fue rediseñada para comenzar con la visualización de vídeos relajantes de 360°, como una playa o una montaña, seguido de la exploración libre de los entornos de la cafetería y el supermercado. Este cambio también ayudó a identificar y abordar la incomodidad con la tecnología.
- **Desarrollo de una escena de calibración amigable:** durante estas pruebas, se observó un problema técnico con la calibración de eye tracking del sistema, que a veces causaba errores con algunos usuarios y generaba frustración antes de comenzar el entrenamiento. La identificación temprana de este problema, antes de que comenzaran las sesiones de terapia principales, llevó al desarrollo de una

escena de calibración fácil de usar que se completaría antes y después de cada sesión de entrenamiento presentada en la Figura 86. Este paso adicional nos permitió medir y analizar los errores de calibración.

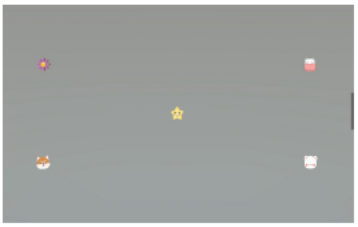


Figura 86: escena de calibración amigable.

- **Grabación de audio desde la interfaz de control del terapeuta:** Durante las sesiones de entrenamiento, se encontró que las guías de audio pregrabadas utilizadas para dar indicaciones eran ineficaces, requiriendo la ayuda del terapeuta durante la realización de las tareas. Para no perder esta información, esto llevó a la implementación de la grabación de audio directamente desde la interfaz al inicio de cada sesión, capturando tanto la guía del terapeuta como las respuestas de los participantes durante el entrenamiento. Con el tiempo, estas grabaciones fueron valiosas para analizar el progreso de los participantes a lo largo de las sesiones y resaltaron la importancia de la presencia de la terapeuta durante el entrenamiento.
- **Adición de barras espaciadoras entre productos:** Además, un participante en la tarea uno del escenario del supermercado tuvo confusión entre los espacios vacíos en los estantes y los espacios entre conjuntos de productos. La solución propuesta fue añadir barras espaciadoras entre ellos para evitar esta confusión. La terapeuta podía activarlas o desactivarlas a través de la interfaz según fuera necesario.

Implementación del Enfoque Final. El enfoque final implementó un sistema con una metodología estructurada en varias fases: familiarización con la tecnología, entrenamiento en la interacción dentro del entorno XR, entrenamiento en la Tarea 1 en los escenarios de supermercado y cafetería, y entrenamiento en la Tarea 2 en ambos escenarios.

El sistema resultante, presentado incorpora dos componentes: una interfaz de control para la terapeuta, similar a la que se utilizaba en terapia conductual, y una aplicación virtual que se ejecuta en el HMD y presenta los escenarios de cafetería y supermercado. Desde la interfaz de control se gestiona la sesión, incluida la iniciación de la aplicación, la selección de escenario, tarea y nivel. Además, muestra el campo de visión de la persona participante, tal y como se hacía en terapia conductual.

A lo largo de los niveles, la dificultad de las tareas aumentó progresivamente (ver Figura 87). Se añadieron más productos o comidas, junto con mesas o estanterías adicionales, y los elementos se volvieron más similares entre sí, haciendo las tareas más complejas.

(a) Level 1. (b) Level 6.

Figura 87: ejemplos de dos niveles de dificultad en el escenario de supermercado.

Caracterización del caso de uso desde el punto de vista técnico

Dispositivos involucrados: HTC XR Elite, VIVE Full Face Tracker, Embrace Plus, Cometa Pico Blue.

Para la experiencia de entrenamiento cognitivo, se eligió el HTC XR Elite, equipado con su complemento completo de seguimiento facial (facetracker), ya que permite la recopilación de datos de seguimiento ocular (Eye Tracking). También se recopilaron datos fisiológicos a través de la pulsera Embrace Plus, sucesora de la Empatica E4, y los sensores Cometa Pico Blue. La justificación de esta elección se basó en la experiencia obtenida en terapia conductual.

Metodología de evaluación

En todos los casos de uso, se utilizaron cuestionarios finales basados en adaptaciones según el caso de uso, contexto y comprensión de:

Escala de Usabilidad del Sistema (SUS) [11]
Cuestionario de Presencia Espacial y Social propuesto en [5]
NASA-TLX [12]
VSR [13]

Estos abordan aspectos distintos de la experiencia del usuario: usabilidad, sensación de presencia y experiencia de calidad general, carga de trabajo mental y física, y síntomas de vértigo o mareo.

Como en la terapia conductual, la terapeuta mantuvo un diario de sesión para cada participante para facilitar un seguimiento del progreso individual.

La experiencia de terapia cognitiva incluyó una ronda final de entrevistas semiestructuradas tanto con los participantes como con la terapeuta. Al igual que con la

adaptación de los cuestionarios, la estructura de la entrevista fue desarrollada colaborativamente por ingenieros y terapeutas. En la terapia cognitiva, el objetivo era evaluar las habilidades de atención, inhibición y memoria. Por tanto, la batería de evaluaciones realizadas antes y después de la terapia incluyó: • Subpruebas relacionadas con la atención del Examen de Cambridge para Trastornos Mentales en Personas Mayores con Síndrome de Down y Otros con Discapacidades Intelectuales (CAMDEX) [15]. • Prueba de Percepción de Diferencias (CARAS-R) [16]. • Una subprueba relacionada con la atención del Cuestionario de Madurez Neuropsicológica Atención (CUMANIN-2) [17]. • Prueba de Gatos y Perros (Cats and Dogs Test) [18]. • Parte A del Test de Trazado (Trail Making Test - TMT) [19].
Duración (comienzo con participantes) y número de participantes
8 personas y duración de 8 meses.
Conclusiones generales
En esta experiencia, se observa una tendencia general a mejora, aunque existe una alta variabilidad entre los resultados de las pruebas individuales. El análisis estadístico confirma que, en la mayoría de los casos, no se puede afirmar que la terapia haya causado mejoras medibles. El pequeño tamaño de la muestra limita el poder estadístico, y sospechamos que las pruebas empleadas no capturaron completamente las ganancias de habilidades específicas del usuario. Por ejemplo, en la prueba TMT, cuatro participantes ya habían alcanzado la puntuación máxima en la pre-prueba; aunque sus tiempos de finalización en la post-prueba mejoraron, la prueba no pudo registrar este cambio. Investigaciones futuras deberían centrarse en seleccionar pruebas de evaluación de habilidades cognitivas apropiadas, expandir la terapia a un grupo de participantes más grande e incluir un grupo de control para obtener resultados definitivos. De manera similar a la terapia conductual, los cuestionarios fueron más fáciles de seguir para los participantes, pero aún requirieron el apoyo del terapeuta para su cumplimentación. Además, arrojaron resultados informativos limitados. Los resultados preliminares del proceso de calibración indican que los usuarios con problemas de agudeza visual experimentaron las mayores dificultades de calibración, pero el análisis estadístico no pudo rechazar la hipótesis nula, lo que sugiere que pueden contribuir factores adicionales. Se necesita una investigación adicional con una muestra de

usuarios más grande y técnicas de calibración variadas para extraer conclusiones más firmes.

Las métricas de comportamiento extraídas de los registros no proporcionaron resultados estadísticamente significativos debido al tamaño limitado de la muestra y la influencia de la asistencia del terapeuta durante las sesiones. No obstante, revelaron tendencias consistentes con el contexto de la terapia.

El análisis temático de las entrevistas proporcionó las percepciones más detalladas sobre las experiencias de los usuarios y las observaciones de los terapeutas. Surgieron tres temas principales:

- Los participantes informaron sentir mejoras en su vida diaria, y los terapeutas notaron un progreso específico del usuario.
- Tanto los participantes como los terapeutas destacaron la importancia de la fase de familiarización, que ayudó a los usuarios a superar la incomodidad y el nerviosismo mientras se familiarizaban gradualmente con la tecnología.
- La experiencia mejoró la Calidad de Vida (QoL): los usuarios disfrutaron del juego, se sintieron motivados para participar en las sesiones, seguir mejorando, se sintieron realizados al completar niveles y percibieron un progreso hacia su objetivo de obtener empleo.

Finalmente, todos los participantes completaron el estudio completo de ocho meses, asistiendo a más de 26 sesiones cada uno. Esta alta retención se atribuye, como en el caso de la terapia conductual, al diseño centrado en el usuario y al flujo de trabajo de la terapia, que priorizó la comodidad e incorporó los comentarios de los usuarios.

5.4 Teleformación: cocina

Título
Teleformación en cocina
Caracterización del caso de uso desde el punto de vista de personas usuarias
Algunos cursos ofrecidos desde la Fundación Juan XXIII con certificaciones tienen un número de plazas limitadas, y algunos participantes enfrentan dificultades para asistir regularmente en persona debido a razones médicas o períodos vacacionales. Estas limitaciones reducen la participación y la continuidad, impidiendo que algunas personas puedan completar las horas requeridas para obtener su certificación. Esta experiencia propone el empleo de tecnologías XR para superar estos desafíos y que puedan acceder a esa formación con clases en remoto. Al aprovechar las características

inmersivas y de mejora de la presencia de la tecnología XR, este enfoque permite la participación remota en sesiones de formación con un nivel de compromiso y evaluación comparable a la asistencia física. Este uso innovador de XR tiene como objetivo asegurar que todas las personas interesadas, especialmente aquellas con mayores necesidades de apoyo como las personas de COFOIL, tengan acceso a valiosas oportunidades de certificación profesional que pueden mejorar significativamente sus perspectivas de empleo.

Fases de co-diseño del caso de uso con las tecnologías implicadas

Etapas iniciales. El diseño inicial se basa en un curso de cocina presencial previo ofrecido por Fundación Juan XXIII Roncalli, que combina clases teóricas con práctica manual apoyada por profesionales. Para recrear esta experiencia en XR, se propusieron tres etapas:

1) Visualización de vídeos pregrabados de 360° que capturan un curso de cocina

2) Práctica del módulo con supervisión remota por parte de un profesional a través del sistema de telepresencia Snowl

3) Práctica autónoma de realidad mixta que permite a los/as estudiantes preparar recetas mientras visualizan los vídeos. Este último enfoque consistiría en visualizar la realidad física mientras se observa una porción del vídeo 360° que permite ver a la profesora, los utensilios y los ingredientes. Un ejemplo de la propuesta inicial se puede ver en la Figura 8. Por lo tanto, los/as estudiantes pueden seguir clases, practicar recetas de forma inmersiva y también colaborar de forma remota con la profesora.

Figura 88: Campo de visión a través del HMD en el que la profesora muestra uno de los ingredientes en el vídeo así como aparece parte del entorno local (real) y los botones con los que interactúa con la mano para ajustar las opciones.

Para la visualización de videos y la guía en MR, se desarrolló una aplicación inmersiva interactiva inicial, junto con una interfaz de control que permitía a la terapeuta gestionar las sesiones. La grabación de 360° se realizó en la Fundación Juan XXIII Roncalli simulando una clase impartida por la profesora sobre la elaboración de magdalenas, estructurada en diferentes vídeos cortos de capacitación, cada uno centrado en un paso específico del proceso para mantener la atención de los/as participantes y mejorar la comprensión. Para permitir la interacción de la persona con el HMD con el vídeo, se propusieron varios métodos de interacción: uso de mandos, botones flotantes con seguimiento manual e interfaces de voz. Se eligieron los botones porque, según las terapeutas, era la interacción más natural.

Durante la práctica de 360°, las personas veían el vídeo sentadas (a no ser que quisieran de pie) para simular estar en un aula guiados por la profesora. Posteriormente, los participantes elaboraban la receta supervisados de forma remota por la profesional utilizando el sistema Snowl. Durante estas prácticas guiadas, la profesional los guiaba paso a paso a través del mismo contenido que en el vídeo de 360°, de modo que si, por ejemplo, tenían que batir un huevo, la profesora les explicaba cómo hacerlo en tiempo real. Una vez que habían practicado el tema con la profesora, realizaban un ejercicio independiente en el que los/as estudiantes veían una sección del vídeo superpuesta a la realidad física (local) capturada por las gafas de 360°.

Mejoras Iterativas. Durante el desarrollo de la experiencia de clases de cocina a distancia, se realizaron varias mejoras iterativas para optimizar la usabilidad, la claridad de la formación y la robustez del sistema. Estas iteraciones surgieron de la realimentación obtenida durante las sesiones piloto con terapeutas, profesionales y participantes, e informaron ajustes progresivos en el flujo de trabajo, el material de vídeo, el diseño de interacción, los instrumentos de evaluación y la infraestructura de telepresencia.

- **Familiarización con la percepción en MR:** las pruebas piloto revelaron que los/as participantes requerían tiempo adicional para familiarizarse con la percepción de su entorno físico a través del passthrough del HMD. Este cambio perceptual no era intuitivo para muchos usuarios y ocasionalmente provocaba desorientación. En consecuencia, se incorporó una sesión de familiarización previa a la formación. Esta sesión introdujo tareas básicas en MR para facilitar la aclimatación visual y asegurar que los/as participantes pudieran interactuar eficazmente con el contenido de entrenamiento superpuesto posteriormente.
- **Alineación de las grabaciones de vídeo 360°:** las grabaciones iniciales de 360° mostraron inconsistencias entre el punto de vista previsto y la dirección de visualización de los/as participantes. Algunas personas no notaron que no estaban orientadas hacia la profesora, lo que resultó en pérdida de información. Para mitigar esto, los vídeos de 360° se reeditaron para alinear a la profesora con la orientación

frontal estándar de una persona con el HMD sentada. Este ajuste redujo la probabilidad de que las personas visualizaran contenido periférico en lugar de seguir a la profesora.

- **Aumento del detalle visual:** ciertas acciones, como leer la pantalla digital de la báscula u observar ajustes sutiles en el horno, no eran lo suficientemente visibles en el inicio. Para abordar esta limitación, se incorporaron elementos visuales adicionales, incluyendo inserciones de primeros planos y superposiciones explicativas que apoyaran las explicaciones.
- **Rediseño de los botones de interacción:** la interfaz inicial empleaba botones de aire minimalistas basados en texto. Las pruebas con personas de la Fundación Juan XXIII demostraron que estos botones eran difíciles de identificar de forma rápida y consistente, especialmente para personas menos familiarizadas con la iconografía abstracta o con habilidades de lectura limitadas. Basándose en esta realimentación, los botones se rediseñaron utilizando representaciones basadas en símbolos tal y como se presenta en la Figura 89.

Figura 89: mejoras progresivas en el rediseño de los botones de interacción.

- **Colocación de los botones:** los prototipos iniciales colocaban los botones en ubicaciones que interferían con los gestos naturales de cocina, lo que provocaba activaciones accidentales. A través de múltiples iteraciones de observación y ajustes por parte del estudiantado, se estableció un diseño vertical posicionado ligeramente a la derecha. Este diseño mantuvo la accesibilidad al tiempo que minimizaba las interacciones no deseadas durante el proceso de preparación de recetas.
- **Mejoras en las capacidades de monitorización de la terapeuta:** La interfaz de control del terapeuta se mejoró con controles adicionales, incluyendo botones remotos de activación de video y una función de calibración espacial que permitía a los terapeutas ajustar la colocación de la superposición de video MR dentro del espacio de trabajo físico del usuario (como se muestra en la Fig. 10a)). Esta mejora aseguró la alineación entre la vista instruccional prevista y el contexto MR del usuario.
- **Mejoras con la telepresencia:** Durante la práctica supervisada utilizando el prototipo Snowl, la profesora informó dificultades para observar detalles visuales a

través de la cámara de 360°, los cuales eran esenciales para proporcionar una guía precisa, particularmente al medir o evaluar la consistencia de las mezclas. Para abordar esto, se utilizó la lupa del prototipo, tal y como se utilizaba en el caso de uso de telepresencia, y se colocó en la cocina para proporcionar a la profesora una vista más cercana de los elementos críticos, mejorando así la calidad y precisión de la supervisión remota. La diferencia es que en este caso la lupa se dejó fija sobre un trípode pequeño y debajo se colocó un papel de color, tal y como se muestra en la Figura 90. De esta forma, los/as estudiantes identificaban que cuando la profesora necesitaba ver algo con más detalle simplemente tenían que colocarlo sobre ese papel de color.

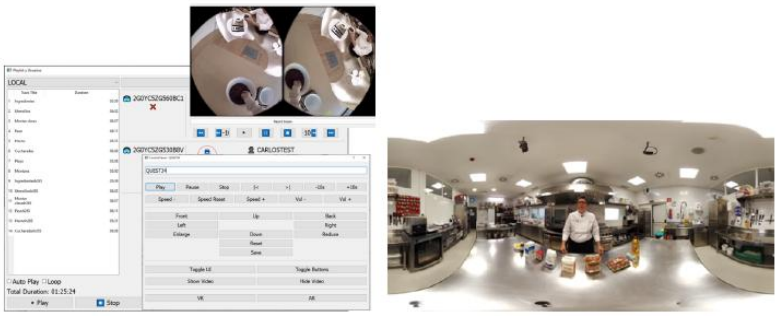
Figura 90: Utilización de la lupa en teleformación en cocina.

Implementación del enfoque final. El enfoque final, reformulado a través de mejoras iterativas, se organizó en cuatro fases progresivas:

- familiarización con la tecnología,
- visualización de vídeos pregrabados de 360°,
- práctica con supervisión remota a través de Snowl
- práctica autónoma de MR.

Cada etapa integra el diseño inicial con las mejoras introducidas después de las pruebas piloto.

El sistema final para la visualización de vídeos 360° y MR consistió en dos componentes (mostrados en la Figura 91):



(a) Therapist control interface for configuring cooking-training sessions, including video control, spatial calibration, and button management.

(b) Example of the re-aligned 360° training video used during cooking classes.

Figura 91: sistema final para la visualización de vídeos 360° y MR.

1) una aplicación inmersiva que soporta tanto modos VR como MR, mejorada con botones de interacción accesibles rediseñados (botones verticales basados en símbolos ubicados ligeramente a la derecha) y ayudas visuales

2) una interfaz de control para la terapeuta que permite la gestión de vídeos, la activación de botones, la calibración espacial MR de la ventana de vídeo (la terapeuta podía ajustar el tamaño y la posición de la ventana, realinear el panel de interacción y controlar la reproducción de vídeo), y la monitorización multicámara.

Por otro lado, el sistema de la práctica supervisada basado en Snowl consistió en:

1) Snowl colocado en la cocina junto con una cámara adicional de lupa de alta definición para mejorar la visibilidad de los detalles de cocción

2) la profesora utilizando un HMD en una habitación separada, guiando al estudiante paso a paso a través de la receta.

Estos elementos formaron la base de entrenamiento XR que preservó la estructura pedagógica original al tiempo que abordó los desafíos de percepción, interacción y supervisión identificados durante las pruebas iterativas.

Caracterización del caso de uso desde el punto de vista técnico

Dispositivos involucrados: Meta Quest 3, Meta Quest 2, Snowl

En la experiencia de las clases de cocina, se emplearon tanto las gafas de RV Meta Quest 3 como las Meta Quest 2. El Meta Quest 3 fue seleccionado por su capacidad de passthrough a color y su fiable seguimiento de manos (hand-tracking), que eran necesarios para tareas que implicaban interacción manual con elementos virtuales. En cambio, las Meta Quest 2 se utilizaron debido a su compatibilidad con Snowl, la versión actualizada y más compacta de The Owl, como se muestra en la Figura 83. Snowl permitió que la profesora estuviera presente de forma remota en el entorno de trabajo de las personas participantes, proporcionando una guía que simulaba la supervisión presencial.

Metodología de evaluación
En todos los casos de uso, se utilizaron cuestionarios finales basados en adaptaciones según el caso de uso, contexto y comprensión de: Escala de Usabilidad del Sistema (SUS) [11] Cuestionario de Presencia Espacial y Social propuesto en [5] NASA-TLX [12] VSR [13] Estos abordan aspectos distintos de la experiencia del usuario: usabilidad, sensación de presencia y experiencia de calidad general, carga de trabajo mental y física, y síntomas de vértigo o mareo. Se añadieron algunas preguntas concretas del caso de uso. Nótese que se preguntó tanto a las personas participantes como a la profesional que algunas de las fases llevaba HMD también. Como en la terapia conductual y cognitiva, la profesora mantuvo un diario de sesión para cada participante para facilitar un seguimiento del progreso individual.
Duración (comienzo con participantes) y número de participantes
10 personas y duración de 5 meses.
Conclusiones generales
La experiencia de formación remota en cocina proporcionó resultados cualitativos positivos, particularmente desde la perspectiva de la instructora del curso. Según sus observaciones, los participantes mostraron niveles notablemente más altos de atención y retención al ver contenido píldoras a través de HMDs que en las clases presenciales tradicionales. El entorno inmersivo redujo las distracciones externas haciendo que mantuvieran el foco en el material del vídeo. Además, la instructora destacó que, en líneas generales, el alumnado progresó a través del contenido del curso significativamente más rápido en comparación con cursos anteriores en los que completaron la misma formación sin soporte XR. Los participantes completaron las sesiones de formación durante un período de cinco meses, lo que permitió tiempo suficiente para la consolidación de habilidades, la práctica repetida y la adaptación a las tecnologías XR involucradas. Inicialmente, la profesional expresó incertidumbre sobre la parte de la formación basada en realizar ejercicios prácticos con la tecnología Realidad Mixta (RM). Esta incertidumbre venía motivada por ser pioneros en aplicar esta modalidad de tele-formación a través de RM. La combinación de múltiples estímulos simultáneos, vídeos de formación grabados y editados para añadir elementos que facilitan la comprensión en 360°, la interacción con

botones virtuales mediante técnicas de seguimiento de manos y la manipulación de utensilios reales en el entorno físico, se percibía como una experiencia potencialmente abrumadora. Sin embargo, la inclusión de sesiones de familiarización estructuradas resultó esencial. Después de esta etapa de sensibilización con la tecnología, las personas participantes se adaptaron bien al flujo de trabajo de MR, reportando comodidad y confianza durante los ejercicios.

Las sesiones prácticas guiadas remotamente en las que el estudiantado no llevaba las gafas de RV sino que las llevaba la profesora también lograron resultados y respuestas de los participantes comparables a los observados en el caso de uso de telepresencia, reforzando la viabilidad de la supervisión remota en este contexto. La profesional destacó que la calidad del sistema sigue siendo limitada para detalles concretos, como el estado de la masa o la granularidad de ciertos ingredientes. Sin embargo, con ayuda de la lupa, se puede guiar para realizar las tareas de manera efectiva. Las principales limitaciones que surgieron principalmente cuando se requería asistencia física ya que no se podía ejemplificar siquiera los pasos a realizar.

En general, el programa de formación se consideró un éxito. Al completar todas las cápsulas instructivas, los participantes se involucraron en una tarea integrada final que requería ejecutar una receta completa de magdalenas de principio a fin. Esta sesión culminante incluyó pesar y medir ingredientes sin HMD, realizar pasos de mezclar como batir huevos y combinar aceite y harina usando MR, y finalmente colocar la masa en moldes y hornear magdalenas. Su desempeño durante esta actividad integral demostró tanto la adquisición de habilidades, como la transferencia efectiva de conocimientos, a través de contextos inmersivos, de MR y del mundo real.

6 Acrónimos

ACR	Absolute Category Rate
ANOVA	Análisis de Varianza
DS	Desensibilización Sistemática
EDA	Actividad ElectroDérmica
EMG	ElectroMioGrafía
HMD	Head Mounted Display
HR	Frecuencia Cardíaca
IC	Intervalo de Confianza
KPI	Key Performance Indicator
QoE	Quality of Experience, Calidad de Experiencia
QoI	Quality of Interaction, Calidad de Interacción
RM	Realidad Mixta
UI	Interfaz de usuario
VR/RV	Realidad Virtual
XR	eXtended Reality

7 Índice de tablas

8 Índice de figuras

9 Referencias

[1] P. Pérez, E. Gonzalez-Sosa, J. Gutiérrez, and N. García. 2022. Emerging immersive communication systems: overview, taxonomy, and good practices for QoE assessment. Frontiers in Signal Processing 2 (2022), 917684.

[2] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. 1993. Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. The International Journal of Aviation Psychology 3, 3 (1993), 203–220.

[3] P. Pérez, N. Oyaga, J. J. Ruiz, and A. Villegas. 2018. Towards systematic analysis of cybersickness in high motion omnidirectional video. In 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 1–3.

[4] ITU-T. 2022. Quality of experience assessment of extended reality meetings. Technical Report. ITU-T.

[5] M. Orduna, P. Pérez, J. Gutiérrez, and N. García. 2023. Methodology to Assess Quality, Presence, Empathy, Attitude, and Attention in 360-degree Videos for Immersive Communications. IEEE Transactions on Affective Computing 14, 3 (2023), 2375–2388.

[6] T. Aitamurto, S. Zhou, S. Sakshuwong, J. Saldivar, Y. Sadeghi, and A. Tran. 2018. Sense of presence, attitude change, perspective-taking and usability in first-person split-sphere 360 video. In Proceedings of the 2018 CHI conference on human factors in computing systems. 1–12.

[7] B. G. Witmer and M. J. Singer. 1998. Measuring Presence in Virtual Environments: A Presence Questionnaire. Presence: Teleoperators and Virtual Environments 7, 3 (06 1998), 225–240.

[8] S. G. Hart and L.E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. Advances in psychology 52 (1988), 139–183.

[9] J. Li, Y. Kong, T. Röggla, F. De Simone, S. Ananthanarayan, H. De Ridder, A. El Ali, and P. Cesar. 2019. Measuring and understanding photo sharing experiences in social virtual reality. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–14.

[10] K. Gupta, G. A. Lee, and M. Billinghurst. 2016. Do You See What I See? The Effect of Gaze Tracking on Task Space Remote Collaboration. IEEE Transactions on Visualization and Computer Graphics 22, 11 (2016), 2413–2422.

[11] Castilla, D., Jaen, I., Suso-Ribera, C., Garcia-Soriano, G., Zaragoza, I., Breton-Lopez, J., ... & Garcia-Palacios, A. (2024). Psychometric properties of the Spanish full and short forms of the system usability scale (SUS): detecting the effect of negatively worded items. *International Journal of Human–Computer Interaction*, *40*(15), 4145-4151.

[12] Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.

[13] Gutierrez, J., Perez, P., Orduna, M., Singla, A., Cortes, C., Mazumdar, P., ... & Garcia, N. (2021). Subjective Evaluation of Visual Quality and Simulator Sickness of Short 360$^\circ $ Videos: ITU-T Rec. P. 919. *IEEE transactions on multimedia*, *24*, 3087-3100.

[14] Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: psychometric properties. *Journal of consulting and clinical psychology*, *56*(6), 893.

[15] Esteba-Castillo, S., Dalmau-Bueno, A., Ribas-Vidal, N., Vilà-Alsina, M., Novell-Alsina, R., & Garcia-Alba, J. (2013). Adaptation and validation of CAMDEX-DS (Cambridge Examination for Mental Disorders of Older People with Down's Syndrome and others with intellectual disabilities) in Spanish population with intellectual disabilities. *Revista de neurología*, *57*(8), 337.

[16] Thurstone, L. L., & Yela, M. (2012). Test de percepción de diferencias (CARAS-R). Tea.

[17] Portellano, J. A., Mateos, R., Martínez, R., Granados, M., & Tapia, A. (2000). Cumanin. Cuestionario de madurez neuropsicológica infantil.

[18] Ball, S. L., Holland, A. J., Treppner, P., Watson, P. C., & Huppert, F. A. (2008). Executive dysfunction and its association with personality and behaviour changes in the development of Alzheimer's disease in adults with Down syndrome and mild to moderate learning disabilities. British Journal of Clinical Psychology, 47(1), 1-29.

[19] Reitan, R. M. (1958). Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and motor skills*, *8*(3), 271-276.

[20] Gonzalez Morin, D., Gonzalez-Sosa, E., Perez, P., & Villegas, A. (2023). Full body video-based self-avatars for mixed reality: from e2e system to user study. *Virtual Reality*, *27*(3), 2129-2147.

10 Apéndice A

10.1 Hoja de información

Por razones de confidencialidad y protección de datos, esta sección ha sido eliminada.