VR 用户体验标准模型

中国信息消费推进联盟

视频体验联盟

中国智慧家庭产业联盟

虚拟/增强现实产业推进委员会

视 频 体 验 联 盟

撰写组名单

苏佳（组长） 首都师范大学

罗传飞 中国电信上海研究院

林 鹏 中国联通网络技术研究院

宋 利 上海交通大学

黄一宏 华为技术有限公司

杨付正 西安电子科技大学

吴雪波 德科仕通信（上海）有限公司

翁冬冬 北京理工大学

刘长虹 中国电信四川分公司

王慧林 北京小鸟看看科技有限公司

台流杰 北京千种幻影科技有限公司

李晓波 北京七维视觉科技有限公司

王金东 中兴通讯股份有限公司

陈玉琨 上海乐相科技有限公司

目 录

总体模型图

VR 用户总体评价角度分为沉浸感质量和交互质量。沉浸感质量侧重于用户使用 VR 设备时眼睛看到的内容质量和耳朵听到的内容质量,具体到行业角度，还包括动态体感，部件匹配度，操作力度，场景沉浸感等维度。交互质量主要着重于各个感觉通道的匹配度问题，各个感觉获得的内容在时间和空间上不相匹配，势必会丢失交互体验。

视觉保真度是由视频质量和观看质量所决定，视频质量包含了视频比特率，帧率，解析度等等参数，观看质量包含视野双目视差，立体视觉，画面延迟等等。听觉保真度同样受音频质量影响，包括音频采样率，音频比特率等等影响，另外影响沉浸体验的一个重要因素就是空间音频，用户能够感受到优质环绕效果的声音会提高体验质量。

具体模型如图 1 所示。

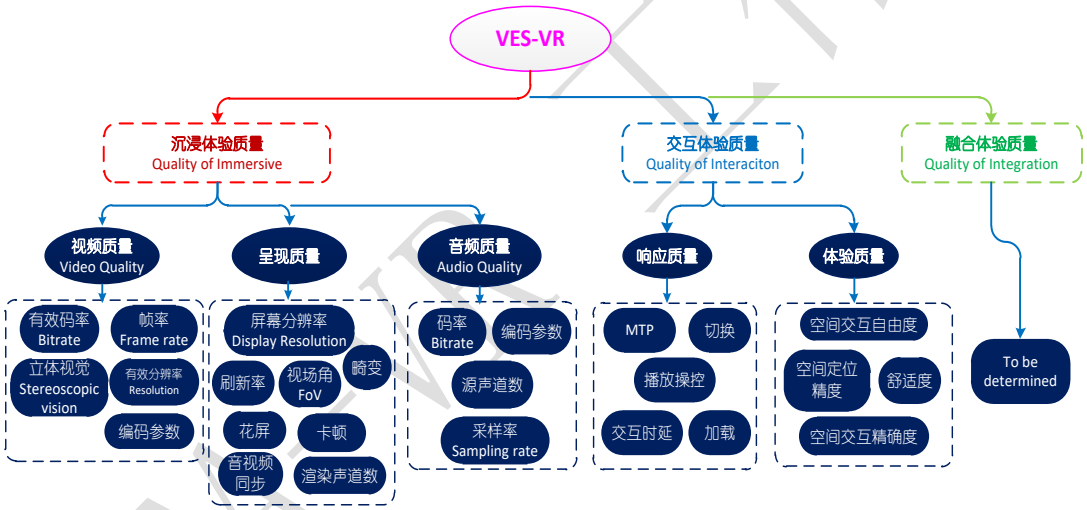

图 1 VR 用户体验参数模型

目前，国际上关于 VR 提出的参数标准较为集中。在图 2 中，我们将国际标准和其涉及到的参数在四个方面（头戴显示器，网络，媒体编码，个体感受）进行了对照汇总。

从图中可以看出，头戴显示器这一方面，包括 ITU-T/SG 12，VQEG，3GPP 都针对了自由度，视野，刷新率等参数提出标准。

而在媒体内容编码方面，MPEG 和 ITU-T/SG 12 则关注于编解码器，比特率，解析度等参数。

VQEG，3GPP，ITU-T/SG 12 在网络角度上，对网络带宽，网络延迟，丢包率都提出了标准建议。

值得一提的是，MPEG-I 和 VQEG 将人体感受列为关注点，特别关注于眩晕这一角度。

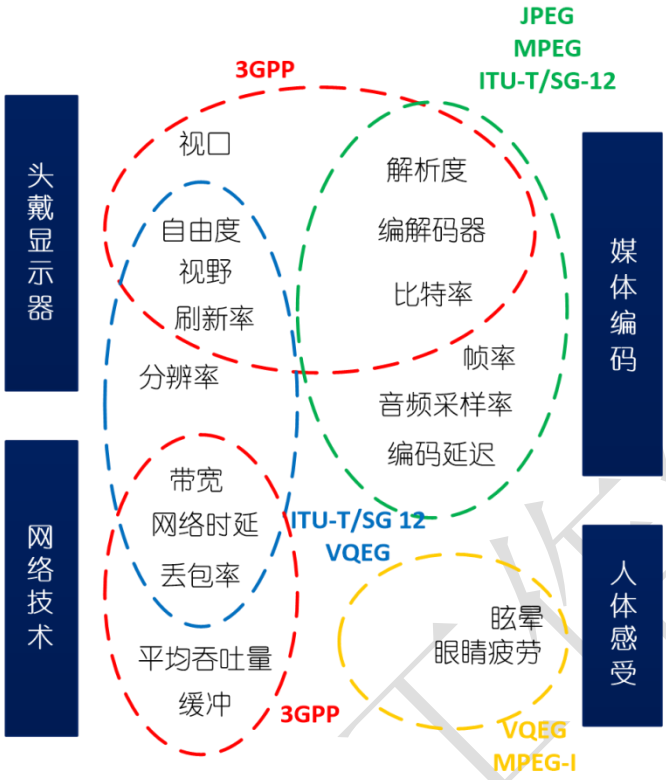

图 2 国际标准涉及参数对照图

VR 用户体验评估算法及参数分类

1. 范围

本标准规定了应用于 VR 视频服务业务用户体验质量的评估场景和模型，分析了影响 VR 视频服务用户体验质量的关键因素，定义了用于评估 VR 视频服务用户体验质量的参数和技术方法。

本标准适用于对电信运营商、广电运营商、互联网视频服务商和其它相关厂商提供的 VR 视频服务用户体验质量进行综合评估，还适用于产业各方对影响视频服务质量的因素进行量化和分析。

2. 缩略语

以下缩略语适用于本文件。

缩略语	中文全称	英文全称
VR	虚拟显示	Visual Reality
AR	增强现实技术	Augmented Reality
HD	高清晰度	High Definition
UHD	超高清晰度	Ultra High Definition
HMD	头戴式可视设备	Head Mount Display
DoF	自由度	Degree of Freedom
FoV	视场角/视野	Field of View
PPI	像素密度-每寸	Pixels Per Inch
PPD	像素密度-每度	Pixel Per Degree
I/O	输入输出系统	Input output system
MIMO	多入多出技术	Multiple-Input Multiple-Output
OFDMA	正交频分多址	Orthogonal Frequency Division Multiple Access
QAM	正交振幅调制	Quadrature Amplitude Modulation
MU-MIMO	多用户多入多出技术	Multi-User Multiple-Input Multiple-Output

IP	互联网协议	Internet Protocol
TCP	传输控制协议	Transmission Control Protocol
PSNR	峰值信噪比	Peak Signal to Noise Ratio
WS-PSNR	加权球形峰值信噪比	Weighted Spherical Peak Signal to Noise Ratio

一、沉浸体验质量

1. 视频质量

1.1 有效码率

有效码率指观看视域内每像素的平均比特数。码率是指视频文件在单位时间内使用的数据流量，也叫码流率。码率越大，说明单位时间内取样率越大，数据流精度就越高，这样表现出来的的效果就是：视频画面更清晰画质更高。

1.2 有效分辨率

有效分辨率指视频观看视频时可视区域内的视频像素数。

1.3 帧率

称为帧的连续图像的频率。VR 内容的帧速率应与显示设备的帧速率属性兼容。 VR 服务中的帧速率比普通的 2D 视频服务要求更高，这是因为较低帧率是导致 VR 眩晕的原因之一。对于 VR 游戏应用来说，它的要求更高，其中场景由 GPU 渲染，而不是由摄像机创建。

1.4 立体视觉

立体视觉(steroscopic vision) 是感受三维视觉空间，感知深度的能力。立体视觉以双眼单视为基础。其形成是由于两眼在观察一个三维物体时，由于两眼球之间存在距离,故而存在视差角,物体在两眼视网膜上的成像存在相似性及一定的差异，形成双眼视差(binocular disparity)。视中枢融像时，双眼水平视差信息形成了我们感知物体的三维形状及该物体与人眼的距离或视野中两个物体相对关系的深度知觉。

1.5 编码参数

影响编码效率的参数，如编码标准(H.264、H.265)和编解码器等。

2. 音频质量

2.1 码率

码率是指在一个数据流中每秒钟能通过的信息量，也可以理解为：每秒钟用多少比特的数据量去表示。原则上，音频码率越高质量越好。

声音中的码率是指将模拟声音信号转换成数字声音信号后，单位时间内的二进制数据量，是间接衡量音频质量的一个指标。视频中的码率原理与声音中的相同，都是指由模拟信号转换为数字信号后，单位时间内的二进制数据量。

2.2 编码参数

影响音频编码效率的参数，如编码标准和编解码器。编码标准包括 MPEG-2,3,4、杜比 AC-3（Dolby Digital）、DTS、AVS 等。

2.3 源声道数

声道(Sound Channel) 是指声音在录制或播放时在不同空间位置采集或回放的相互独立的音频信号，所以声道数也就是声音录制时的音源数量或回放时相应的扬声器数量。

2.4 采样率

音频采样率是指录音设备在一秒钟内对声音信号的采样次数，采样频率越高声音的还原就越真实越自然。目前采样频率一般共分为 22.05KHz、44.1KHz、48KHz 三个等级，22.05KHz 只能达到 FM 广播的声音品质，44.1KHz 则是理论上的 CD 音质界限，48KHz 则更加精确一些。

3. 呈现质量

3.1 屏幕分辨率

显示分辨率是屏幕的基本属性，用于指示屏幕支持的每英寸像素数。相对于视频分辨率，适当的屏幕分辨率将提供最佳和舒适的体验。

高角分辨率显示成为提升 VR 近眼显示沉浸感的核心技术。片面追求单一性能参数，忽视技术指标间的平衡协同有悖于虚拟现实领域融合创新的固有特性，然而，随着 VR 头显在近眼显示上对清晰度提出了更高的要求，为了降低“纱窗效应”，提高屏幕分辨率及开口率成为关键发展方向，“4K+”分辨率由智能手机领域的弱需求上升为虚拟现实的强需求。此外，由于 VR 具备的 360 度全景显示特性，角分辨率取代 PPI 成为更适宜的近眼显示像素密度的核心技术指标，随着未来 4K 屏幕的普及、视场角/分辨率的权衡设计，预计至 2020 年单眼 PPD 将达到 30 以上水平。

人眼的分辨力，也称视敏度，与人眼视网膜上两像点的距离及视锥细胞的直径大小有关。眼睛的极限分辨角可以表示为：

$$n = x \cdot \tan \varepsilon$$

公式中，ε 表示眼睛的极限分辨角，n 表示视锥细胞的直径，约为 0.004mm，x 表示像距，即瞳孔到视网膜的距离，为 24.4mm，因此

$$\tan \varepsilon = \frac{n}{x} = \frac{0.004}{24},$$

$$\varepsilon = \tan^{-1}\frac{n}{x} \approx 35'' \approx 0.5'$$

即人眼的极限分辨力为$0.5'$。由于不同人的眼睛不可能完全一样，因此计算出来的结果也不相同。一般来说，一个正常人的分辨力在$0.4' \sim 1'$之间。

分辨力为$1'$，意味着可以看清 1°的 1/60，也就是说可以分辨出距离为1°/60 的两个不同像素点。通常，人单眼的水平视场为 160°垂直方向上为 175°，因此在水平方向上可以分辨出 160×60=9600 像素，垂直方向上可以分辨出 175×60=10500 像素。

因此要想获得有沉浸感的体验，一方面，需要 VR 设备的分辨率达到此数值，而目前 VR 设备的屏幕分辨率远远不达标。另一方面，视频源的分辨率也需要很高。由于使用 VR 设备观看多视角视频时，所能看到的区域仅是原始视频画幅的 1/4 左右，因此视频源的分辨率需要是 VR 设备分辨率的四倍，才能保证视频源和设备的像素点一一对应。

3.2 刷新率

刷新率是显示器每秒从图形处理单元 GPU 获取新图像的次数。刷新率越高，所显示的图像稳定性就越好。较低的刷新率将导致处理延迟并导致 VR 衍生疾病。

3.3 视场角

视场角 FoV 显示设备所成像中，人眼可观察到部分的边缘与观察点(人眼瞳孔中心)连线的夹角，包括水平视场角、垂直视场角、对角线视场角。是任何给定时间可观察环境的范围。使用更宽的视场角，用户更有可能在体验中感受到现场。 因此视场角是一个重要的参数，有助于评估 VR 设备在多大程度上可以帮助创建身临其境的体验。

3.4 音视频同步

对 VR 视频的采集、播放及对同步的要求都非常严格，如果从多媒体文件中分离出音视频数据的数据不同步，音视频的时间差则会越来越大，这是无法忍受的，所以在多媒体文件中，不但要求有同步机制，考虑的音视频同步方案有两种：一是发送端解决；二是接收端解决。

3.5 渲染声道数

当用户使用 360 度空间音频时，每种声音听起来像是从空间中对应的方位发出，就像我们在现实生活环境中感知声音一样。在摄像机上方飞行的直升机轰鸣听起来就像是在用户的上方，在摄像机前方的演员对白听起来就像是在用户的前面。当用户环视整个视频画面时，系统需要根据用户头部方向的变化做出反应并将每种声音重新定位到画面上的相应位置。无论是通过手机，浏览器还是 VR 头戴显示器，当用户每次观看 360 度全景视频时，音频都需要被重新计算并更新方位以完美还原用户真实的空间感受。

被称为“hybrid higher-order ambisonics（混合高阶立体声高保真混响）”的技术，可以使空间化的声音在整个处理过程中依旧能保持很高的质量。这是一个具有渲染与优化功能的 8 声道音频处理系统，可借助更少的声道实现更高的立体声质量，最终达到节省带宽的目的。

3.6 畸变

即使图像本身投射到图像平面上，它也可能在外围发生扭曲。假定透镜是径向对称的，则这种失真是随着距离光轴越远而图像拉伸或压缩越严重的失真现象。失真现象一般包含两种：桶形失真和枕形失真。对于具有宽视场的镜头，失真更强，特别是鱼眼镜头这种极端情况。纠正这种失真对于目前具有宽视场的 VR 头戴式设备至关重要，否则，呈现的虚拟世界

就会变形。

如果现实世界中的镜头完全按照斯内尔定律进行光线会聚或发散，那么 VR 系统将简单得多，且令人印象深刻。事实上，许多统称为畸变的缺陷会降低由镜头形成的图像质量。由于这些问题在日常使用中都很明显，例如通过 VR 设备观看内容，因而这些问题非常重要，需要采取相关的补偿措施并应用到 VR 系统中。

通过 VR 头显的镜头观看时，通常会产生枕型畸变。如果图像不经过任何修正的话，那么虚拟世界看上去就会出现扭曲的现象。如果用户的头部来回转动的话，由于四周的变形比中心强烈，一些固定线条（如墙壁）的曲率会动态的改变。如果不加以修正，就没有一种静态物体的感觉，因为静态物体不应该会有动态变形。此外，这也有助于研究虚拟现实导致的一些疾病的成因，可能是由于在 VR 体验内感受到了四周异常的加速度。

3.7 花屏

视频花屏多由以下三种原因造成：

渲染脏数据

渲染脏数据是指还未完成渲染的数据。具体来讲就是在视频某帧渲染到一半的时候，即被送到编码器编码。此问题发生在视频渲染阶段。

丢帧

此处所说丢帧丢弃的是视频编码后的视频帧，通常发生在复用（Mux）阶段。由于视频编码后帧之间存在依赖关系，丢帧会带来及其严重花屏效果，并且具有持续性影响。此问题发生在视频编码阶段。

图像格式转换

在视频编解码中必然会涉及到 YUV 和 RGB 图像格式的转换，并且 YUV 还有多种格式。如果转换格式或者算法不正确也会引发视频花屏问题。此问题发生在视频渲染或者播放阶段。

3.8 卡顿

在观看 VR 视频时，网络是影响其卡顿与否的重要原因。

Wi-Fi 技术

家庭无限网络的覆盖可保证虚拟现实移动便利性以及满足业务体验的带宽、时延等要求。

家庭无线网络广泛应用的是基于 802.11n 或 802.11ac 标准的 Wi-Fi 设备，其中前者标准同时指出 2.4GHz 和 5GHz 频段，后者支持 5GHz 频段。基于 802.11ac 的 Wi-Fi 在 80MHz 频谱上通过 4x4 MIMO、Beamforming 等技术可实现最大 1.7Gbps 空口速率。

下一代 Wi-Fi 技术为 802.11ax，将引入 8x8 MIMO、OFDMA、1K QAM 等新特性，最大可实现 10Gbps 空口速率，同时抗干扰能力强，以保障丢包率、时延以及带宽稳定性。此外，VR 头显无线化是利用无线传输技术进行无损视频传输以提高用户体验。IEEE 802.11 目前在制定基于 60GHz 的下一代 Wi-Fi 标准 802.11ay，通过 channel bonding、MU-MIMO 等技术提供 20-40GHz 带宽，可实现无压缩视频帧数据传输。

网络延迟

网络延迟是指各式各样的数据在网络介质中通过网络协议(如 TCP/IP)进行传输，如果信息量过大不加以限制，超额的网络流量就会导致设备反应缓慢，造成网络延迟。

平均吞吐量

从主机 A 到 B 跨越计算机网络传送文件，则在任何时间瞬间的瞬时吞吐量是主机 B 接收到该文件的速率（bps），比如下载文件时显示的瞬时下载速率，应当为该进程的瞬时吞吐量。而如果这个文件大小为 F 比特，主机 B 接收到所有 F 比特用去 T 秒，则 F/T bps 应当为文件传送的平均吞吐量。

二、交互体验质量

1. 响应质量

1.1 MTP

MTP（Motion To Photons）时延是指从头动到显示出相应画面的时间。MTP 时延太大容易引起眩晕，目前公认的是 MTP 时延低于 20ms 就能大幅减少晕动症的发生。为降低 MTP 时延，一方面需要提升 GPU 的渲染性能，另一方面需要将显示屏的刷新率提高到 75Hz 以上，目前 Vive 和 Rift 都达到了 90Hz 的屏幕刷新率。

1.2 切换

从一个节目/频道切换到另外一个节目/频道的时长。

1.3 播放操控

播放过程中快进，快退等操作的反映快慢。

1.4 交互时延

交互时延主要是由网络时延和头部跟踪时延造成。

网络延迟是指各式各样的数据在网络介质中通过网络协议(如 TCP/IP)进行传输，如果信息量过大不加以限制，超额的网络流量就会导致设备反应缓慢，造成网络延迟。

头部跟踪延迟方面，要实现用户与环境之间的交互，创建可信的 VR 体验非常重要。显然低头部跟踪延迟具有可以为用户提供平滑视图变化的重要属性，但是高头部跟踪延迟则会引起不适并失去沉浸式体验。

1.5 加载

节目的起播时间，即用户点击新节目后首帧解码出来的时长。

2. 体验质量

2.1 舒适度

VR 设备直接由用户佩戴在头上使用，舒适度的体验质量主要由和以下三点相关联：

1）重量和尺寸

2）发热和散热

3）面部贴合度

2.2 空间交互自由度

自由度 DoF 表示对象在空间内移动的方式，这是帮助为用户创建沉浸式环境的关键参数。

三自由度 3DoF 表示在特定观察位置，当头部左右旋转(yaw)，俯仰旋转(pitch)，和摇摆旋转(roll)时，VR 头显能正确显示出相应 VR 内容，其需要 VR 内容、VR 采集和 VR 显示设备的支持。

六自由度 6DoF 表示在特定观察位置，当头部左右旋转(yaw)，俯仰旋转(pitch)，摇摆旋转(roll)，以及一定范围内向前后、左右、上下三个方向平移时，VR 头显能正确显示出相应 VR 内容，其需要 VR 内容、VR 采集和 VR 显示设备的支持。

2.3 空间交互精确度

头部跟踪延迟

要实现用户与环境之间的交互，创建可信的 VR 体验非常重要。显然低头部跟踪延迟具有可以为用户提供平滑视图变化的重要属性，但是高头部跟踪延迟则会引起不适并失去沉浸式体验。

动作捕捉

用户想要获得完全的沉浸感，真正“进入”虚拟世界，动作捕捉系统是必须的。目前专门针对 VR 的动捕系统，目前市面上可参考的有 Perception Neuron，其他的要么是昂贵的商用级设备，要么完全是雾件(意为在开发完成前就开始进行宣传的产品，也许宣传的产品根本就不会问世)。但是这样的动作捕捉设备只会在特定的超重度的场景中使用，因为其有固

有的易用性门槛，需要用户花费比较长的时间穿戴和校准才能够使用。相比之下，Kinect 这样的光学设备在某些对于精度要求不高的场景可能也会被应用。

表 1 动作捕捉相关参数

参数	说明
运动和加速度的快慢选择	运动速度的设定为接近人类正常的移动速度（步行约为 1.4 米/秒，持续的慢跑速度约为 3m/s，或是将运动速度开放给用户自行调节）
控制的程度	使用者自行移动，头显视角全程对使用者的头部运动的响应，视角的锁定
头部上下晃动	基于使用者头部真实动作的上下晃动效果或镜头方向、位置变化
前向和侧向	使用者视线往往集中在视野中央，中央视野是均匀且逐渐变化的，而周围视野是快速移动变化的。左右侧向移动相比前向运动视线焦点会不停从一个物体跳向另一个物体

手势识别

手势识别是让静态手型或动态手势与确定的控制指令进行映射，触发对应的控制指令。此类交互需要用户提前对手势进行一定的学习和适应，因此为交互体验的提升带来较大挑战。

目前，以 Microsoft Hololens 为代表的 AR 头显广泛采用手势识别，其问题在于交互率低，学习成本较高，指令特定单一。而虚拟现实中，用户通过手部 26 个自由度的关节点姿态信息，进一步重建整个手部骨架和轮廓，使得虚拟世界中的手和现实中几乎保持一致，用户自由度较高，学习成本低，但问题在于在一些特殊条件下性能表现尚未达标，如遮挡等。同时缺乏必要反馈。目前，如 LeapMotion、IntelRealSense、凌感、锋时互动等产品能够提供双摄/结构光/飞行时间等定制化、模组化的解决方案。

方向追踪

方向追踪除了可以用来瞄点，还可以用来控制用户在 VR 中的前进方向。不过，如果用

方向追踪调整方向的话很可能会有转不过去的情况，因为用户不总是坐在能够 360 度旋转的转椅上的，可能很多情况下都会空间受限。比如头转了 90 度接着再转身体，加起来也很难转过 180 度。所以，这里“空间受限无法转身是一个需求”，于是交互设计师给出了解决方案——按下鼠标右键则可以让方向回到原始的正视方向或者叫做重置当前凝视的方向（就是你最初始时候面向的那个方向），或者可以通过摇杆调整方向，或按下按钮回到初始位置。

表 2 方向追踪内容

跟踪系统	追踪用户运动来创造沉浸式体验的系统，可整合在设备之中，也可作为头戴式显示器的外设
眼动跟踪	通过测量眼睛注视点的位置或眼球相对头部运动而实现对眼球运动的追踪，也称之为视线跟踪，其可以补充头部跟踪的不足
头部追踪	使用头部追踪传感器追踪定位用户的头部运动，然后根据记录的数据、移动放映的图像，使其和头部运动的位置相匹配
空间跟踪	通过头盔显示器、数据手套、立体眼镜、数据衣等交互设备上的空间传感器，确定用户的头、手、躯体或其他操作物在三维虚拟环境中的位置和方向
声音跟踪	利用不同声源的声音到达某一特定地点的时间差、相位差、声压差等进行虚拟环境的声音跟踪
跟踪球	用来操纵显示屏上光标移动的设备，包含用手自由推动的球和对应 x 方向及 y 方向的轴角编码器

触觉反馈

触觉主要包括按钮和震动反馈，也就是利用 VR 手柄得到的反馈。目前三大 VR 头显厂商 Oculus、索尼、HTC Vive 都不约而同的采用了虚拟现实手柄作为标准的交互模式：两手分立的、6 个自由度空间跟踪的（3 个转动自由度 3 个平移自由度），带按钮和震动反馈的手柄。这样的设备显然是用来进行一些高度特化的游戏类应用的（以及轻度的消费应用），这也可以视作一种商业策略，因为 VR 头显的早期消费者应该基本是游戏玩家。

语言交互

在 VR 中海量的信息淹没了用户，他不会理会视觉中心的指示文字，而是环顾四周不断发现和探索。如果这时给出一些图形上的指示还会干扰到他们在 VR 中的沉浸式体验，所以最好的方法就是使用语音，和他们正在观察的周遭世界互不干扰。这时如果用户和 VR 世界进行语音交互，会更加自然，而且它是无处不在无时不有的，用户不需要移动头部和寻找它

们，在任何方位任何角落都能和他们交流。

三、融合体验质量

TBD