**3GPP TSG-SA SA4#131S4-250083**

**Geneva (CH), 17 – 21 February 2025**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *CR-Form-v12.2* | | | | | | | | |
| **CHANGE REQUEST** | | | | | | | | |
|  | | | | | | | | |
|  |  | **CR** |  | **rev** |  | **Current version:** | **0.2.1** |  |
|  | | | | | | | | |
| *For* [***HE******LP***](http://www.3gpp.org/3G_Specs/CRs.htm#_blank)*on using this form: comprehensive instructions can be found at* [*http://www.3gpp.org/Change-Requests*](http://www.3gpp.org/Change-Requests)*.* | | | | | | | | |
|  | | | | | | | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Proposed change affects:*** | UICC apps |  | ME |  | Radio Access Network |  | Core Network |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | |
| ***Title:*** | pCR on Scenario: Streaming of Multi-view plus depth Produced Content | | | | | | | |
|  |  | | | | | | | |
| ***Source to WG:*** | Nokia, Philips, Interdigital, Deutsche Telekom, Fraunhofer HHI, Sony Group Corporation, China Mobile, Huawei | | | | | | | |
| ***Source to TSG:*** |  | | | | | | | |
|  |  | | | | | | | |
| ***Work item code:*** |  | | |  | ***Date:*** | | | 2025-02-11 |
|  |  | | |  |  | | |  |
| ***Category:*** | B |  | | | ***Release:*** | | | Rel-19 |
|  | *Use one of the following categories:* ***F*** *(correction)* ***A*** *(mirror corresponding to a change in an earlier release)* ***B*** *(addition of feature),* ***C*** *(functional modification of feature)* ***D*** *(editorial modification)*  Detailed explanations of the above categories can be found in 3GPP [TR 21.900](http://www.3gpp.org/ftp/Specs/html-info/21900.htm). | | | | | | *Use one of the following releases: Rel-8 (Release 8) Rel-9 (Release 9) Rel-10 (Release 10) Rel-11 (Release 11) … Rel-16 (Release 16) Rel-17 (Release 17) Rel-18 (Release 18) Rel-19 (Release 19)* | |
|  |  | | | | | | | |
| ***Reason for change:*** | | An evaluation scenario is proposed that handles the streaming of produced multi-view plus depth content that provides experiences beyond what is achievable with traditional 2D video. The scenario allows to evaluate the streaming of high-quality, professionally captured and produced multi-view plus depth video content. | | | | | | |
|  | |  | | | | | | |
| ***Summary of change:*** | | New clause with scenario description on streaming of multi-view plus depth produced content | | | | | | |
|  | |  | | | | | | |
| ***Consequences if not approved:*** | | Streaming of multi-view plus depth produced content cannot be evaluated. | | | | | | |
|  | |  | | | | | | |
| ***Clauses affected:*** | |  | | | | | | |
|  | |  | | | | | | |
|  | | **Y** | **N** |  | |  | | |
| ***Other specs*** | |  |  | Other core specifications | | TS/TR ... CR ... | | |
| ***affected:*** | |  |  | Test specifications | | TS/TR ... CR ... | | |
| ***(show related CRs)*** | |  |  | O&M Specifications | | TS/TR ... CR ... | | |
|  | |  | | | | | | |
| ***Other comments:*** | | Initial text on this scenario is included in the PD clause 2.2.  Changes are tracked starting from a clean copy of S4aV250002. | | | | | | |
|  | |  | | | | | | |
| ***This CR's revision history:*** | | |  |  |  |  | | --- | --- | --- | --- | | [**S4-241841**](https://www.3gpp.org/ftp/TSG_SA/WG4_CODEC/TSGS4_130_Orlando/Docs/S4-241841.zip) | [FS\_Beyond2D] pCR on scenario: Streaming of Beyond 2D Produced Content | Nokia, Philips, Interdigital, Deutsche Telekom, Fraunhofer HHI, Sony, China Mobile, Huawei | Serhan Gül |   **Revisions**: none  **Online Discussion**:  **Session 4: 16:00 - 18:00**   * Serhan presents. * Thomas: Market relevance is still not demonstrated in this contribution. Representation formats are also unclear why mixing V3C and MIV is not helping in the description. * Waqar: Is there a deployed service using this? This is for me a relevant industry activity. * Rufael: MPEG has done the work to define the representation format, otherwise the compression scheme developed in MPEG would not have worked. We need to be forward looking and looking at novel experiences on mobile devices. * Serhan: We cannot discard all the industry activities, there is more than just commercialized service. * Gilles: What is the approach to develop work in this study? Format, use case, compression, representation, what is the starting point in those aspects? * Ralf: This is the right moment to discuss the scenario since it compiles all the information in this document. * Thomas: Source format is underspecified, bit depth, resolution? What are the relevant format produced by the industry? * Serhan: Information is provided in clauses 4.3.2.1 and 4.3.4.1. * Waqar: More details needed in industry relevance. * Serhan: We can add more on scalability and power consumption on mobile devices. But this is using available technologie and GPU. * Madhukar: None of the formats in Beyond2D maye become industry relevant but this is not a problem at a study stage. There is no indication of next normative steps. Even if this is implicit, we should make this explicit that there is no endorsement for normative work. * Madhukar: On the format, several things are optional like depth. This needs to be clarified. * Thomas: Maybe start with stereo+depth but 20 views is a totally different story. We should not mix them. * Sehran: It’s not mixed in the PD. * Sehran: We want to finalize the scenario and move to the evaluation. * Giles: About splitting stereo and multi-view? * Sehran: We are interested in evaluating multi-view. * Madhukar: Again, depth and alpha channels need to be clarified, optional? Mandatory? If yes, how many?   **Session 7: Thu 11:00 to 13:00**   * Sehran presents r1. * Thomas: The revision refines to multiview 10 to 20 views, correct? * Sehran: Yes, no stereo anymore. * Emmanuel: Then the title and name of the scenario needs to reflect this, not just streaming of B2D content. * Madhukar: Add a note that the scenario can be addressed by different tech, like MV-HEVC. * Thomas: Is it including UE-generated content? * Sehran: no UE generation * Thomas: The text is not clear on that. * Gilles: A B2D content is vague, we should define clearer content definition. * Sehran: I am ok with multi-view content. * Thomas: But the text talks about volumetric. * Gilles: The source format is clear, it is multi-view. * Thomas: We should motivate what a service provider could do, not motivation based on certain technologies.   **Decision**:   * Session 4: parked * Session 7: revised to xxx (at least new scenario title reflecting multi-view) and xxx agreed to PD.   [**S4-241841**](https://www.3gpp.org/ftp/TSG_SA/WG4_CODEC/TSGS4_130_Orlando/Docs/S4-241841.zip) is **revised to 2188 and 2188 is agreed (PD)**. | | | | | | |
|  | | |  |  |  |  | | --- | --- | --- | --- | | [S4aV250002](https://www.3gpp.org/ftp/TSG_SA/WG4_CODEC/3GPP_SA4_AHOC_MTGs/SA4_VIDEO/Docs/S4aV250002.zip) | pCR on scenario Streaming of Multi-view Produced Content | Philips International B.V. | Bart Kroon |   **Presenter**:Bart Kroon  **Online Discussion**: December 18, 2024   * Jiayi : usage of multi-view video and immersive video together may bring confusion.   + Bart : will replace everything with multi-view video * Jiayi : content will be one person only or other types   + Bart : content will be scenes : background + people + objects, i.e. multiple assets. * Jiayi : about interoperability considerations, remove company names in TR/TS   + Bart : we can remove company names and just add a reference to the demo * Thomas : There is no clear scenario identified, just a use of technologies * Thomas : why would you use AI and CGI ro produce MIV. Other formats are likely to be more appropriate. Where does MIV content come from ? Who would use AI and CGI to produce MIV and why ?   + Bart : This would allow protecting 3D assets. It is possible to remove AI/CGI production part if requested. * Thomas : the description lacks clear evidence of such usages. * Rufael: movies such as avatar and lion king are generated using CGI but the CGI formats mentioned are not suitable for transmission and rendering, stereo or multiview are common transmission and device rendering * Waqar : Comments are only partially addressed. Why would MV-HEVC be excluded? Where is market relevance.   + Bart : regarding MV-HEVC, it was removed at Orlando meeting, on request. * Mary-Luc : Since the focus is multi-view video, AI/CGI production should not be kept unless there is clear explanation how it is used.   + Bart : fine with removing AI/CGI production   + Rufael: CGI is a commonly use to generate content such as avatar or lion king, this needs not to be elaborated in the document, this is already in the generic workflow desription. * Gilles : the request is to provide motivation on why multiview format is relevant where original assets generated by AI/CGI could be used directly. * Imed : There are other solutions to protect your 3D assets. * Rufael: 3D Assets in most cases are not suitable nor desireable for transmission/protection to clients, with a few exceptions * Gilles : The supporters are invited to focus on the motivation for such scenario using AI/CGI and more generally the added value in terms of QoE of the multiview format.   **Decision**:   * 2024-12-18: Revision is expected to answer questions raised.   [S4aV250002](https://www.3gpp.org/ftp/TSG_SA/WG4_CODEC/3GPP_SA4_AHOC_MTGs/SA4_VIDEO/Docs/S4aV250002.zip) is **noted**. | | | | | | |

**== CHANGE 1 ===**

#### 4.3.4.1 Definition

by

Note that in some contexts like in JCT-3V a narrower definition of multi-view was used whereby the cameras are expected to be in a 1D linear or coplanar arrangement [R12]. For this representation there is no such restriction.

ly,have been used asiplus

The multi-view video representation consists of multiple frames of multiple synchronized physical or virtual camera views. Each camera view is represented by a colour image (YCbCr), camera intrinsics and camera extrinsics. The combination of video and metadata allows for novel view synthesis (6DoF rendering).

A typical spatial resolution for each of the views is 1920 × 1080 (FullHD). For this representation in this study, we expect resolutions in a range around this number. A typical number of views is 2-4 for real-time capture with range-sensing cameras like the Azure KinectTM, and typically 10-20 for offline capture with industrial or professional cameras. Typically, the frame rate is 25 or 30 Hz, and capture beyond 60 Hz is not expected for the coming years.

Optionally there is also a depth image of equal resolution. It is possible to have multi-view content for which some or all views lack depth information. This choice originates from the production and capturing system and thus it is the same for all frames of a view. The depth map image, if present, may also indicate that individual samples are missing. This indication can be used for range-sensing cameras that cannot sense depth in certain situations like object edges, non-reflecting and specular reflecting scene elements. It can also be useful in a production system to remove parts of an image that are not wanted (e.g. revealing camera rigs) or are also present in other views (scene background).

If a view has a depth map, then it must have corresponding depth quantization parameters: quantization type (normalized disparity or linear depth), nearest depth in scene units, furthest depth in scene units, and indication of invalid values. Normalized disparity [m-1] is more commonly used when depth is estimated because it places the code points in a way that correlates with the amount of parallax, and it allows for far away scene elements like the horizon or the star field. Linear depth [m] is commonly used with range-sensing cameras (ToF, LiDAR, etc.) because they often have a limited depth range with equal depth resolution for that entire range.

The camera intrinsics are a model of the projection of points in space in the reference system of the camera to the image sensor (projection plane). Typical parameters include projection type (perspective, fisheye, etc.), and projection-type specific parameters, such as principal point and focal length for perspective projection. Optionally lens distortion parameters may be provided if the camera images are not already corrected for that.

The camera extrinsics model the translation and rotation of a camera in space with respect to the reference system of the scene.

The source format has at least two views. It is expected that most or all test data will have perspective projection (PSP), but test data with equirectangular projection (ERP) may be included.

While this representation allows for 6DoF rendering, it depends on the position and field of view of the cameras, if such a rendering has an acceptable quality. Preferably, the virtual viewpoints are within a viewing space that can be provided as metadata or implicitly derived from the parameters of the set of source views.

Each view has the following video components and metadata:

|  |  |  |
| --- | --- | --- |
| **Component** | **Texture (mandatory)** | **Depth (optional)** |
| **Spatial resolution** | At least 960 × 540  At most 3840 × 2160 | The same as the texture component |
| **Chroma format** | YCbCr | Luma only or YCbCr with chroma planes set to neutral gray |
| **Chroma subsampling** | 4:2:0 | 4:0:0 or 4:2:0 with chroma planes set to neutral gray |
| **Pixel aspect ratio** | 1:1 | 1:1 |
| **Frame rate** | 30, 50, 60 | The same as the texture component |
| **Colour space format** | ITU‑R BT.709 or ITU‑R BT.2100 | Undefined |
| **Transfer characteristics** | Limited range or full range with transfer characteristics matching to the colour space format.  Mastering characteristics such as MDCV (master display colour volume) and CLLI (content light level information) SEI (supplementary enhancement information) messages defined in TS 26.116 Section 4.5.5.7 will be considered. | Full range, linear transfer |
| **Bit depth** | Either 8 bits or 10 bits for all channels | At least 8 bits  At most 16 bits |
| **Metadata** | Camera intrinsics:  Projection type (Perspective, ERP)  - Projection type (Perspective, ERP)  - For perspective projection:  - Focal length [px]  - Principal point [px × px]  - For equirectangular projection:  - Latitudinal angle range [rad × rad]  - Longitudinal angle range [rad × rad]  - Lens distortion parameters (optional)  Camera extrinsics:  - Camera position (x, y, z) [m]  - Camera orientation as normalized quaternion (*q* = *iq*x + *jq*y + *kq*y + *q*w) | Depth quantization parameters:  - Quantization type:  - either: normalized disparity  - or: linear depth  - Near depth [m]  - Far depth [m]  - Has invalid pixels flag |

**== CHANGE 2 (all new) ===**

## 7.x Scenario x: Streaming of Multi-view plus depth Produced Content

### 7.x.1 Motivation for the scenario

This scenario handles the streaming of produced multi-view plus depth content that provides experiences beyond what is achievable with 2D content. The scenario allows for the evaluation of the streaming of high-quality, professionally captured and produced multi-view plus depth video content.

In this scenario, mult-view plus depth content is played back on phones and tablets for which one viewport is shown. UI elements and/or the tilt sensor can be used to change the virtual viewport. This experience is more immersive than 2D video because the (subtle) pose changes provide motion parallax which is a strong perceptual depth cue. This effect is already achieved using a small number of cameras (3-4) and a limited viewing space (the size of a person's head). More cameras (10-20) are needed for free-viewpoint functionality.

In this scenario, it is also considered, but not evaluated, that the same content can be played back on more advanced devices including head-mounted displays and eye-tracked autostereoscopic displays. For these classes of devices, two virtual viewports are rendered, thereby providing a stronger 3D effect due to the combination of motion parallax and stereopsis depth cues. While relevant, the expectation is that for the forseeable future the majority of the UE's will be 2D phones and tablets.

The main benefits of using the multi-view plus depth representation for on-demand streaming, is that 1) the difference in appearance of objects between cameras is preserved, making the experience more like video and less like graphics, 2) less processing steps are needed to construct the representation as compared to the derived representations, 3) transmission is possible by the combination of 2D video plus metadata.

As of start of 2025, no commercial deployment of multi-view plus depth *delivery* to mobile devices has been identified. Multi-view plus depth video can offer an experience on top of 2D video. First services can provide stills or short clips that enable a viewer to look around and observe some actions from different viewpoints. This can be a stepping stone towards live streaming of multi-view plus depth video.

[Ed.(BK): Review later and check if there is a consistent level of information for all scenarios.]

This scenario is based on the multi-view plus depth video representation format that is defined in clause 4.3.4. Capturing setups and production software are available as described in the related representation format definitions. Contribution, compression and storage formats for multi-view plus depth video are available, see clause 7.x.3. It is expected that segmented media delivery will be used based on DASH and ISOBMFF. Carriage of coded media using ISOBMFF has been specified for MIV [R10] in [R4] and MV-HEVC [R11] in [R9]. Other codecs may be considered. Hardware video decoder capabilities can be used for all pixel data. Rendering and display systems for multi-view plus depth video are described in clause 4.3.4.3.

### 7.x.2 Description of the scenario

This scenario considers on-demand streaming of multi-view plus depth produced content to a UE (Figure 7.x.3-1). All or the most relevant parts of the content are produced using a camera array that observes a scene. The array may include 2D cameras and/or range-sensing cameras. In some cases, part of a scene may be created or inpainted using AI or CGI to reduce the number of physical cameras. This scenario does not consider use of AI/CGI production without a physical camera array.



Figure 7.x.3-1: On-demand streaming of B2D produced content to a UE

Capture setup, production tools and workflows for multi-view plus depth video capture systems and production tools are described in clause 4.3.4.2. Contribution, compression and storage formats are linked to the multi-view video representation format. Well-defined contribution formats exist that carry the raw texture/depth images and camera parameters, e.g. as described in clause 4.3.4.2. Compression formats for multi-view plus depth video are described in clause 4.3.4.4. One codec that can be used to realize this scenario is MPEG Immersive Video (MIV) [R10]. Another option is the Multiview extension of high efficiency video coding (HEVC) standard [R11], commonly referred to as MV-HEVC, as well as the 3D extensions of HEVC (3D-HEVC). MV-HEVC enables the encoding of multiple views and depth data in HEVC by allowing the presence of additional layers in an HEVC bitstream, each corresponding to either a different view or depth information. Such support is enabled through only high level syntax modifications in the original design of the HEVC standard making it easy to repurpose multiple HEVC encoders or decoders that might be available in an existing implementation. 3D-HEVC, however, introduces additional low level coding tools intended for the improved compression of depth information and might not be available in most implementations. Other codecs may also be considered. Below one possible workflow with MIV is described.

The multiple camera views and depth maps are encoded to create a unified representation. An example could be MIV constrained to one or more atlases and packed video data. The single video sub-bitstream per atlas would be encoded with the HEVC Main10 profile. The bitstream contains all camera parameters that are necessary for 6DoF rendering. Each atlas is independently renderable. Another example can be MV-HEVC using auxiliary layers for depth maps and SEI messages for camera parameters.Figure 7.x.3-2 provides an example of an MIV encoder flow.



Figure 7.x.3-2: MIV encoder example

* Patch Extraction and Filtering: extraction of regions from the texture and depth map for the purpose of pixel-rate reduction and allowing object interactivity at the client.
* Background View Extraction: The ground surface and far-away background can be represented by a single background texture with depth. This greatly reduces the required pixel space.
* Atlas Generation: The patches and sprite are packed in an atlas such that both the pixel area is optimally used and the temporal correlation is retained to guarantee an acceptable bitrate.

An example of multi-view plus depth video encoding has been described in the paper [R5].

The encoded bitstream is encapsulated to ISOBMFF according to the rules of the used codec.

For example, an MIV bitstream may be packaged in one track, or multiple tracks where the packed video data is one track, common atlas data is one track, and atlas data is another track. ISO/IEC 23090-10 [R4] specifies how to map MIV (V3C) onto ISOBMFF, file format and DASH.

When a scene is represented by multiple atlases, only one of them may be decoded based on the viewing position. This is called atlas-level sub-bitstream access. In the case of DASH, switching atlas would amount to changing tracks.

The decoder(s) will make use of hardware video decoder capabilities for all pixel data, and metadata describing information needed for rendering is decoded/parsed by a CPU.

Rendering and display systems for multi-view plus depth video are described in clause 4.3.4.3.

### 7.x.3 Source format properties

video sources formats formatsthe

For this scenario, the multi-view plus depth video source format has 3 to 20 views. It is expected that most or all test data will have perspective projection (PSP), but test data with equirectangular projection (ERP) may be included.

Each view has the following components:

* Texture (color)
* Depth coded as normalized disparity

Depth information can be used in rendering e.g. by shaders for surface normal estimation.

Editor’s note: Further details on depth processing is FFS.

All views have view parameters: camera ID, camera intrinsics, camera extrinsics (pose) and depth quantization parameters (optional).

Views may be undistorted, otherwise distortion parameters have to be provided.

The signal properties defined in clause 4.3.4.1 apply with no further constraints.

### 7.x.4 Encoding and decoding constraints and settings

Some constraints and settings below are given for MIV:

Codec profiles/levels:

* HEVC Main 10 MIV Main
  + MIV level 2.0 or 2.5.
* HEVC Main 10 MIV Extended
  + MIV level 2.0, 2.5 or 3.0 whereby the level 3.0 is only allowed if there is a single video sub-bitstream.

Support for multi-plane image (MPI) through the MIV Extended Restricted Geometry sub-profile may be relevant for this scenario, but it is not considered for this study for practical reasons: it requires an additional conversion from multi-view plus depth to multi-plane image.

Some constraints and settings below are given for MV-HEVC:

For content with *N* views plus depth, there will be the following layers:

* First (reference) texture layer: Main or Main 10 profile, level 4.1 or 5
* *N - 1* dependent texture layers: Multiview Extended or Multiview Extended 10 profile, level 4.1 or 5
* First (reference) depth layer (AuxId = AUX\_DEPTH): Multiview Monochrome or Multiview Monochrome 10 profile, level 4.1 or 5
* *N - 1* dependent depth layers: Multview Monochrome or Multiview Monochrome 10 profile, level 4.1 or 5

The presence of the following two SEI messages is required for virtual view synthesis:

* Depth representation info SEI
* Multiview acquisition info SEI

Note that the coding of the depth layers could be independent of the coding of the texture layers.

In general, a random-access frequency of 32 frames can be considered. It is up to the service provider to define the exact random access frequency.

Transmission systems need to be prepared to resend data in case of data loss. If data loss still occurs or retransmitted data does not reach the receiver device in time for rendering, previous immersive frames may be re-rendered with updated viewing poses. In case one or more of the sub-bitstreams is lost, it is up to the application to determine an optimal method for hiding the missing information.

Typically, bitrates between 5 and 50 Mbit/s may be considered.

Bitrate parameters related to video sub-bitstreams need to be configured by the streaming service provider. Transfer characteristics are signalled in the video sub-bitstreams.

There are no special requirements regarding ABR. Configuration is left for the service provider to determine.

Latencies between 500ms to several seconds are considered. Random access interval or segment duration are configured according to the latency requirements.

Encoding is performed by a content provider. This scenario assumes professional setting for recording and processing the content, so no real-time or encoder hardware or architecture requirements are provided.

It is expected that devices support HW accelerated video decoding.

Decoding requirements for MIV:

* HEVC Main 10
* HEVC levels are determined according to the maximum HEVC Level that is needed for a video sub-bitstream decoder to fulfill the MIV level.
  + HEVC level 5.1 for MIV level 2.0
  + HEVC level 5.2 for MIV level 2.5
  + HEVC level 6.1 for MIV level 3.0
* Video sub-bitstreams need to be independently decodable. This helps implementations on various platforms that may have only high-level APIs. For instance, geometry needs to be full range.

Samples in the sub-bitstreams should be temporally aligned.

Decoding requirements for MV-HEVC:

While decoding and rendering all views may result in a higher quality, this is not a requirement. At a minimum, a client needs to be able to select the two nearest views for decoding and rendering. This requires the decoding of the following six sub-bitstreams:

* Reference texture: Main or Main 10, level 4.1 or 5
* Reference depth: Multiview Monochrome or Multiview Monochrome 10, level 4.1 or 5
* Two dependent textures: Multiview Extended or Multiview Extended 10, level 4.1 or 5
* Two dependent depths: Multiview Monochrome or Multiview Monochrome 10, level 4.1 or 5

The view selection can change each intra period.

### 7.x.5 Performance Metrics and Requirements

The tests are run for a chosen level as described in clause 7.x.6. Bitstreams are provided. Camera calibration, depth estimation, and encoding are not evaluated.

The test will have four rate points and QP values are selected for each sequence to approximately match the 5 to 50 Mbps range. When saturation occurs before 50 Mbps a lower value may be chosen in consultation. When there are multiple video components or packed regions then the other QP values need to be directly derived from the texture QP using an equation or a look-up table. (They cannot depend on the sequence.)

[Ed.(BK): To be aligned with the agreed evaluation framework.]

The IV-PSNR tool, available at <https://gitlab.com/mpeg-i-visual/ivpsnr>, is available to compute full-reference objective metrics:

* Weighted sphere PSNR (WS-PSNR)
* Immersive video PSNR (IV-PSNR)

All source views that were used for encoding are provided. Each source view is reconstructed by decoding and rendering (view synthesis). The IV-PSNR tool is then run on all source views and the score is averaged over all views.

Depending on bit rate, quality of depth maps and rendering, either the video codec or view synthesis is the limiting factor. BD-PSNR is calculated for both metrics because the metric behaves more predictably than BD-rate.

[Ed.(BK): To be aligned with the agreed evaluation framework. The discussion here on correlation of objective and subjective metrics may need to be moved to that framework after more deliberation.]

There is experience in testing of multi-view plus depth video in MPEG context. The test conditions as described are a simplification and evolution of the common test conditions for MIV defined in [R6].

The main challenge with testing of multi-view plus depth video is that codecs are asymmetric. The input is a number of source views (with depth maps), and the output of the decoder + renderer can be any viewport within a spatial region around those source views. In the mentioned CTC two tests are used:

* Objective evaluation at source view positions
* Subjective evaluation of pose trace videos (dynamic viewports)

This has resulted in a lack of correlation between objective and subjective results, but despite that it is the best-known approach. Alternatives that have been tried and dismissed (for now):

* Objective evaluation at dynamic viewports: It includes view synthesis in the reference condition and this skews the results towards a specific renderer. It prevents an A/B comparison of different renderers.
* Subjective evaluation at source view positions: This is not how the end-user will interact with the content, and it does not evaluate artifacts due to viewport dynamics.

For this test, because the aim is to prove feasibility of a scenario, objective evaluation may be sufficient, especially when supplemented with (informal) real-time demonstration of the same bitstreams that were used for objective evaluation.

### 7.x.6 Interoperability Considerations for the application

The multi-view plus depth video bitstream needs to be carried over DASH for this scenario. It is not necessary to prove this as part of the feasibility test, if written evidence can be provided.

In the example of using MIV as a codec, there are implementations for DASH [5G-MAG] and RTP + SDP [uvgRTP]. It is possible to subset MIV to always transmit all pixel data in a single packed video track plus a timed metadata track.

### 7.x.7 Test Sequences

Test sequences that were used during the development of a codec are discouraged because they may create a bias towards that specific codec. Sequences that were used in a verification test are permissible.

Preferably test sequences match with the intended use case both in terms of technical requirements and content semantics.

For MIV a list of available sequences is provided in [R6].

### 7.x.8 Detailed test conditions

[Ed.(BK): To be aligned with the agreed evaluation framework.]

For each candidate codec, a suitable decoder + renderer needs to be made available for testing purposes.

For MIV and MV-HEVC a reporting template or script will be provided to compute BD-PSNR based on IV-PNSR log files of all rates and sequences.

For MIV and MV-HEVC the common test conditions defined in [R6] are followed. The anchor is based on multiple individual HEVC streams, one for each texture and each depth frame.

### 7.x.9 External Performance data

For MIV the performance data is available from the verification test report [R7].

NOTE: This performance data was based on different source view properties and the results may not translate to this study.

### 7.x.10 Additional Information

The Metaverse Standards Forum (MSF) has established a Volumetric Media Interoperability working group which aims to build a better understanding of volumetric media, including multi-view plus depth video, to identify relevant areas of applications and compatibility requirements, and to establish common requirements for different systems. See here the WG description: <https://metaverse-standards.org/domain-groups/volumetric-media-interoperability/>

The technology is expected to be highly scalable since it uses well-established transport technologies like DASH and 2D video coding techniques.

Regarding complexity, rendering and decoding frame rates for MIV content were measured for Windows and Android platforms in [R5]. The results show that the developed platform can decode V3C content in real time on both Windows and Android. Evaluation of battery consumption (power levels) is FFS.

Streaming of multi-view plus depth content has the potential to disrupt several markets including entertainment/media, education/training, retail/shopping.

Several use cases can be envisioned related to these domains. For example, in an education/training scenario, a pre-recorded video of a fitness instructor showing how to perform an exercise can help the student to better understand how the exercise is done and thus replicate in a correct way. Another example in education domain would be a mechanic giving a tutorial on how to assemble a mountain bike. The viewer can watch the movements of the mechanic from different angles and get an improved understanding of the different steps due to depth perception and different viewpoints. In the entertainment domain, users can stream a performance from their favorite band to their living room and experience greater immersion potentially together with spatial audio.

**== CHANGE 3 ===**

# 2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non‑specific.

- For a specific reference, subsequent revisions do not apply.

- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

…

[R1] 3GPP TR 26.928: “Extended Reality (XR) in 5G”.

[R2] 3GPP TR 26.998: “Support of 5G glass-type Augmented Reality / Mixed Reality (AR/MR) devices”.

[R4] ISO/IEC 23090-10:2022 (Amd1), “Information Technology — Coded Representation of Immersive media — Part 10: Carriage of Visual Volumetric Video-Based Coding Data”

[R5] Guede et al., IBC 2023, “Efficient Delivery and Rendering on Client Devices via MPEG-I Standards for Emerging Volumetric Video Experiences”. <https://www.ibc.org/technical-papers/ibc2023-tech-papers-efficient-delivery-and-rendering-on-client-devices-via-mpeg-i-standards-for-emerging-volumetric-video-experiences/10277.article>

[R6] Dziembowski, B. Kroon, J. Jung (Eds.), Common test conditions for MPEG immersive video, ISO/IEC JTC 1/SC 29/WG 04 N 0372, July 2023, Geneva.

[R7] D. Mieloch (Ed.), Verification test report of MPEG immersive video, ISO/IEC JTC 1/SC 29/WG 04 N 0341, April 2023, Antalya.

[R8] B. Brand, Michel Bätz, Joachim Keinert, Camorph: a toolbox for conversion between camera parameter conversions, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, url: <https://github.com/Fraunhofer-IIS/camorph>, 2022.

[R9] ISO/IEC 14496-15:2024: Carriage of network abstraction layer (NAL) unit structured video in the ISO base media file format.

[R10] ISO/IEC 23090-12:2023: MPEG immersive video.

[R11] ITU-T H.265:2024 | ISO/IEC 23008-2:2024, Annex G: Multiview high efficiency video coding (MV-HEVC).

[R12] Gerhard Tech, Ying Chen, Karsten Müller, Jens-Rainer Ohm, Anthony Vetro, Ye-Kui Wang, Overview of the Multiview and 3D Extensions of High Efficiency Video Coding, IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 1, January 2016.