**Source:** **Deutsche Telekom, Fraunhofer HHI, InterDigital, Nokia Corporation, Philips, Sony**

**Title: [FS\_Beyond2D] Scenario UE-to-UE Beyond 2D Video Communication**

**Agenda Item: 9.9**

**Document for: Discussion and Agreement**

# Introduction

A new study item FS\_Beyond2D ([SP-240479](https://www.3gpp.org/ftp/TSG_SA/TSG_SA/TSGS_103_Maastricht_2024-03/Docs/SP-240479.zip)) was approved at SA#103. One of the objectives of the study is:

2. Establish and document a set of beyond 2D video end-to-end reference scenarios, including real-time communication, streaming services, split rendering, and messaging and corresponding workflows (capturing, encoding, packaging, delivery, decoding, rendering, including general constraints on latency, as well as complexity) to support 3GPP network related delivery and devices leveraging the generation or display technologies. This includes identifying and defining relevant beyond 2D formats in the context of above workflows, and representation technologies to support delivery of these formats within 3GPP networks.

In this contribution, a draft scenario on UE-to-UE Beyond 2D Video communication is proposed for incorporation into TR 26.956 as basis for future work. The scenario is structured according to the template provided in TR 26.065 v0.0.1 ([S4-240825](https://www.3gpp.org/ftp/TSG_SA/WG4_CODEC/TSGS4_127-bis-e/Docs/S4-240825.zip)).

**========================= CHANGE 1 (all new) ==========================**

## 6.x Scenario #x: Scenario UE-to-UE Beyond 2D Video Communication

1. **Scenario name**

Scenario UE-to-UE Beyond 2D Video Communication

1. **Motivation for the scenario**

*What is the market relevance of the proposed scenario within the next few years? Are there any commercially available or pre-released products or prototypes?‘*

The proposed scenario handles the UE-to-UE Beyond 2D Video communication that provides experiences beyond what is achievable with 2D content.

“Beyond 2D” content may be in the form of volumetric video, which is a frame-based immersive experience whereby each frame represents a volumetric region in 3D space in which any point is either non-occupied or having a color that may depend on the viewing direction.Volumetric video has the potential to provide a more immersive and interactive experience for use cases in diverse domains such as e.g. education, entertainment, and industrial monitoring.

This scenario considers Beyond 2D Video communication which has many applications. Having lifelike communication between UE's will reduce the need for travel and contribute to sustainability goals.

**Teleconferencing**: Hybrid working habits have become commonplace since the global pandemic. While traditional 2D-video communication systems usually get the job done, transferring the meetings into an immersive environment requires video conferencing systems to evolve to support 3D communication.

**Online education**: is one of the sectors which can benefit greatly from improved immersion and visual depth perspective. Perhaps a guitar teacher wants to observe their pupil’s hand positioning and finger placement on the fretboard. Another example could be a yoga instructor, who needs to see the other person from any viewpoint to evaluate the yoga poses accurately. This comprehensive viewpoint could help the teacher to remotely identify and correct any nuances that may be hindering the student’s development. This is a clear progression from the limitations that come with current 2D teleconferencing tools or solutions that rely on animating participants.

**Mixed reality / metaverse**: Besides avatars and other graphical elements, having a direct face-to-face conversation in a shared 3D space reduces the felt distance and increases trust between the multiple participants. They are also able to show objects without first converting them to a CAD format, by showing them in front of the camera system.

With the increasing capabilities and affordability of new multisensory devices and algorithms that allow capturing and experiencing volumetric content with high fidelity, several prototypes or commercial products have emerged within the industry. Some examples are provided below:

* Nokia
  + Real-time volumetric video streaming demo at the Mobile World Congress (MWC) **2024.**
  + <https://www.nokia.com/blog/3d-live-communication-becomes-part-of-everyday-life-with-volumetric-video/>
* Ericsson
  + <https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/holographic-communication-in-5g-networks>
* Imverse - <https://imverse.com/>
* 8i - <https://8i.com/stream/>
* Telefónica - Holographic Telepresence System with 3D Capture demonstrated at MWC 2023
  + The collaboration involves the partners Evercoast, Intel, and AWS.
  + <https://www.telefonica.com/en/communication-room/press-room/telefonica-showcases-its-holographic-telepresence-with-3d-capture-at-mwc/>
* Google is partnering with HP to commercialize its Starline project:
  + https://blog.google/technology/research/google-project-starline-hp-partnership/

NOTE 1: The examples are meant to provide motivation and demonstrate the market relevance of the scenario and not to give detailed information on the capture setup, formats or other aspects of the workflow. For the workflow description see the clause 3.

NOTE 2: The present scenario targets XR applications and volumetric video use cases that can be realized with data rates and latencies achievable over 5G networks. Holographic communication is seen as a future evolution already receiving some interest from the industry, as demonstrated in some of the examples above. However, it is out of the scope of the present scenario.

1. **Description of the scenario**

*This provides a description of beyond 2D video end-to-end workflows, which includes identifying and defining beyond 2D formats being used in the context and representation technologies to delivery these formats. The following aspects may be considered for each workflow:*

This scenario considers UE-to-UE format exchange (conversational). Typically, the exchange will be symmetrical and both UEs will capture and render. Storage and delayed playback on the same UE will also be possible.

In the first variant of this scenario (Figure 1), a first UE captures beyond 2D video and transmits to a second UE that renders one or two viewports of the transmitted scene.

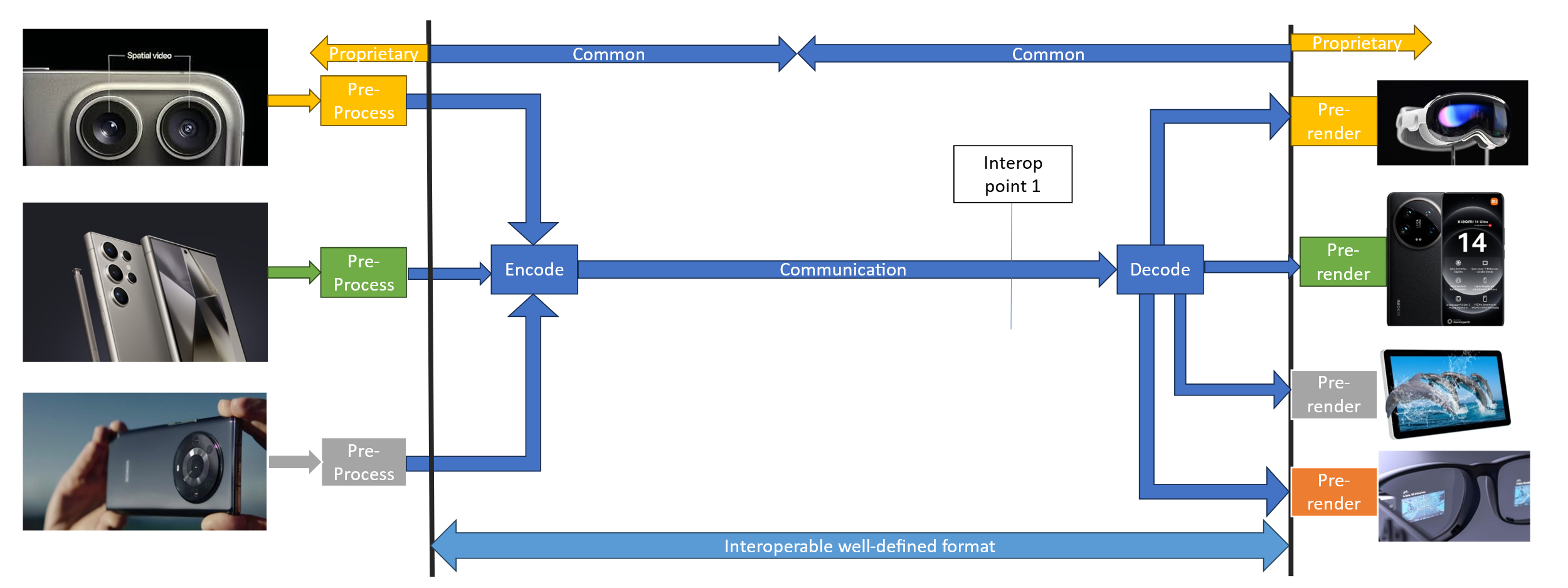


Figure 1: Direct UE->UE communication[[1]](#footnote-2)

[Ed.(BK): The figure will be redrawn if needed for copyright reasons.]

In the second variant of this scenario (Figure 2), the first UE captures beyond 2D video and transmits to a cloud processor that performs decoding, image undistortion, depth estimation, and/or other beyond 2D video related processes, and finally encoding. The cloud processor transmits a bitstream that is suitable for low-complexity rendering to a second UE.

The cloud processor may be described as a beyond 2D video transcoder. It converts from an easy-to-capture to an easy-to-render representation.

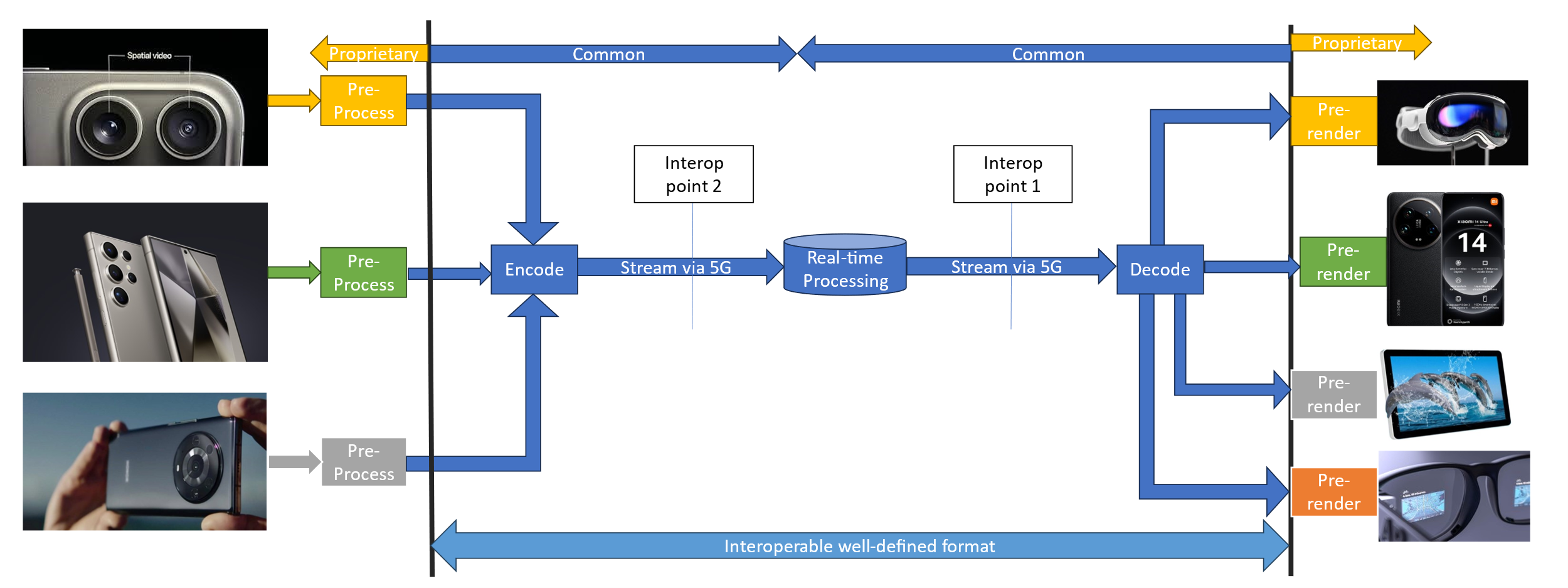


Figure 2: UE-UE communication via cloud processor

[Ed.(BK): The figure will be redrawn if needed for copyright reasons.]

1. *Capturing and processing*

Capture capabilities of UEs will vary and are typically constrained by form factor. A mobile phone has multiple closely packed cameras and may also have a depth sensor. Other devices like computer monitors or camera bars are larger and can have more cameras with more separation.

The beyond 2D video is captured and processed using one or more cameras. Zero or more of those cameras may be range-sensing cameras, and one or more of these cameras have colour sensors. In the case of two or more cameras that are not rigidly connected, camera extrinsics are online calibrated.

Captured content is converted in real-time to a representation format that is suitable for encoding. The parameters of the source cameras may or may not be part of the representation format. At a minimum, inter-view consistent depth information is estimated, but more processes like reprojection, pruning, refinement, meshing and texturing steps may be needed depending on the representation format and application-specific constraints.

1. *Encoding*

One codec that can be used to realize this scenario is MPEG Immersive Video (MIV) specified in ISO/IEC 23090-12. Below the workflow with MIV is described. In addition to MIV, other codecs may be studied as part of this scenario.

The multiple camera views and depth maps if available are encoded to create a unified representation. An example could be MIV constrained to one atlas and packed video data. The single video sub-bitstream would be encoded with HEVC Main10 profile. The bitstream would contain all camera parameters that are necessary for 6DoF rendering.

In the second variant of this scenario:

* Relevant aspects of the capture device like distortion parameters or inter-sensor parallax are not compensated on the transmitting UE, but transmitted as metadata and the cloud processor will run processes such as undistortion or image warping to compensate for the depth camera characteristics. This allows for lower-complexity beyond 2D video capture.
* The receiving UE will have a simplified rendering method like inverse painter’s algorithm or Z-buffering that does not require view blending. This allows for lower-complexity beyond 2D rendering.

1. *Packaging and delivery*

The encoded bitstream is packed to RTP payload format according to the used codec. For example, if MIV codec is used, MIV bitstream may be packaged as a sequence of V3C units including V3C units for the packed video data, or the MIV bitstream can be packaged as three separate tracks, one for common atlas data, atlas data, and packed video data.

IETF proposal avtcore-rtp-v3c[[2]](#footnote-3) enables carriage of MIV as a RTP stream.

1. *Decoding*

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

1. *\*Post-processing*
2. *Rendering*

Rendering is typically performed on a GPU without dedicated hardware.

When a viewing space is used, then:

* What is rendered is one or two viewports with perspective projection and with 6 degrees of freedom (3-D position and 3-D orientation).
* The pose of the viewport is within a viewing space that can be signalled or implicitly determined from a decoded frame. A viewing space can limit both position, orientation or both in combination. For instance, it is generally not intended for a viewport to intersect with scene elements.
* When a viewport would be is rendered that is outside of the viewing space, then the renderer has to perform a mitigation to avoid a viewing experience that is not intended by the content provider.

1. *General constraints on latency, bandwidth, reliability and complexity*

The codec should support low-delay reference frame structure.

For example, the MIV access units and the video sub-bitstream can be organized using a low-delay reference frame structure. All sub-bitstreams could have the same GOP structure, but atlas data and common atlas data frames may be skipped.

The common atlas data with camera parameters will only change infrequently (once per second or less), for instance each time an online camera calibration is updated. While it is possible to transmit common atlas data at non-IRAP frames, this is not expected in this scenario.

The atlas data with patch information may be static when transmitting only full views or dynamic when an encoder selects regions of the source views for transmission based on e.g. occlusion detection or depth segmentation. In MIV, patch information is always intra-coded (unlike V3C in general).

All encoder, decoder and renderer processes are real-time and may have a latency in the order of at most a few frames.

1. **Supporting companies and 3GPP members**
2. *This documents the 3GPP members that support this scenario in terms of providing the information, test material, test requirements and the characterization for the tests. For each of the identified necessities, a tick box is created in the template.*
3. *Preferably several 3GPP members are included in the support, and in addition a video service provider may be included (not necessarily a 3GPP member).*
4. *Cross-verification is preferably done by the supporters of the scenario*

Deutsche Telekom, Fraunhofer HHI, InterDigital, Nokia Corporation, Philips (alphabetical order)

1. **Source format properties**

*This defines a clear range of the considered and relevant source formats, including the signal properties, but also the characteristics of the content. As an example, the texture and depth format properties of the source may be used which include:*

The source format has 1 to 4 views with perspective projection (PSP).

Each view has the following components:

* Required: Texture (color)
* Optional: Depth coded as normalized disparity
* Optional: Object ID map with ordinal values, 0 indicating "*invalid pixel*" (to support depth cameras), 1 indicating "background" and all other values indicating "*foreground object* i".

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

1. *Spatial resolutions*

Each component of each view is 1920 × 1080 (perspective).

1. *Chroma Format*

The texture components are YCbCr.

All other components can be luma only, or YCbCr with chroma planes set to neutral gray.

1. *Chroma Subsampling*

The texture components are 4:2:0.

All other components can be 4:0:0, or 4:2:0 with chroma planes set to neutral gray.

1. *Aspect ratios*

The pixel aspect ratio of all video components is 1:1.

1. *Frame rates*

The source frame rate is 30 fps.

Note that the video frame rate may be different from the rendering frame rate, especially when the viewport pose is dynamic.

1. *Colour space formats*

All texture components will use the ITU-R BT.709 colour space. The colour space format of other components is undefined because the chroma planes are not used.

1. *Transfer Characteristics*

All texture components will use the ITU-R BT.709 transfer characteristics with limited range. The transfer characteristic of other components is linear with full rnage.

1. *Bit depth*

The source texture components will be 8 or 10 bit.

The source depth components (if any) will be in between 8 and 16 bit.

The source object ID maps (if any) will be 16 bit.

1. *Viewpoints*

The viewpoints are within a viewing space that can be provided as metadata or implicitly derived from the parameters of the set of source views.

1. *Other signal properties*
2. **Encoding and decoding constraints and settings**

*Typical encoding constraints and settings such as*

1. *Relevant Codec and Codec Profile/Levels according to TS26.119*
2. *Random access frequency*
3. *Error resiliency requirements*
4. *Bitrates and quality requirements*
5. *Bitrate parameters (CBR, VBR, CAE, HRD parameters)*
6. *ABR encoding requirements (switching frequency, etc.)*
7. *Latency requirements and specific encoding settings*
8. *Encoding context: real-time encoding, on device encoding, cloud-based encoding, offline encoding, etc.*
9. *Required decoding capabilities*
10. *Synchronization requirements*

**TBD**

1. **Performance Metrics and Requirements**
2. *A clear definition on how the performance needs to be evaluated including metrics, etc addressing the main KPIs of the scenario.*

The tests are run for a chosen level as described in 5 a. 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 look-up table. (They cannot depend on the sequence.)

1. *Objective measures such as PSNR, VMAF, etc, may be used.*

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.

1. *Justification on whether objective metrics are sufficient and representative of the subjective performance.*

There is experience in testing of immersive video in MPEG context. The test conditions as described are a simplification and evolution of:

Dziembowski, B. Kroon, J. Jung (Eds.), Common test conditions for MPEG immersive video, [**ISO/IEC JTC 1/SC 29/WG 04 N 0372**](https://www.mpeg.org/wp-content/uploads/mpeg_meetings/143_Geneva/w23008.zip), July 2023, Geneva.

The main challenge with testing of Beyond 2D 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.

1. **Interoperability Considerations for the application**
2. *Streaming with DASH/HLS/CMAF/QUIC*
3. *RTP based delivery*

The Beyond 2D Video bitstream needs to be carried over RTP for this use case of 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 (InterDigital/Philips) and RTP (Nokia).

1. **Test Sequences**

*A set of selected test sequences that are provided by the proponents in order to do the evaluation. They should cover a set of source format properties*

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 the following document includes a list of available sequences:

* Dziembowski, B. Kroon, J. Jung (Eds.), Common test conditions for MPEG immersive video, [**ISO/IEC JTC 1/SC 29/WG 04 N 0372**](https://www.mpeg.org/wp-content/uploads/mpeg_meetings/143_Geneva/w23008.zip), July 2023, Geneva.

1. **Detailed test conditions**

*Provides a proposal for detailed test conditions, for example based on a reference software*  *together with the sequences and configuration parameters.*

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

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 the following test conditions were followed:

* Dziembowski, B. Kroon, J. Jung (Eds.), Common test conditions for MPEG immersive video, [**ISO/IEC JTC 1/SC 29/WG 04 N 0372**](https://www.mpeg.org/wp-content/uploads/mpeg_meetings/143_Geneva/w23008.zip), July 2023, Geneva.

1. **External Performance data**

*References to external performance data that can be added, for example other SDOs, public*  *documents and so on.*

For MIV the following performance data is available:

* D. Mieloch (Ed.), Verification test report of MPEG immersive video, [**ISO/IEC JTC 1/SC 29/WG 04 N 0341**](https://www.mpeg.org/wp-content/uploads/mpeg_meetings/142_Antalya/w22688.zip), April 2023, Antalya.

1. **Additional Information**

1. Images were copied from S4-240831 [↑](#footnote-ref-2)
2. <https://datatracker.ietf.org/doc/draft-ietf-avtcore-rtp-v3c/> [↑](#footnote-ref-3)