**Source:** **Deutsche Telekom, Fraunhofer HHI, Interdigital, Nokia Corporation, Philips, Sony, China Mobile**

**Title: [FS\_Beyond2D] Scenario Streaming of Beyond 2D Produced Content – Use Case “Immersive scenes”**

**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 Streaming of Beyond 2D Product Content including a Use Case “Immersive scenes” is proposed for incorporation into FS\_Beyond2D 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: Streaming of Beyond 2D Produced Content – Use case: Immersive scenes

1. **Scenario name**

Streaming of Beyond 2D Produced Content

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 streaming of produced Beyond 2D content 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.

Streaming of volumetric video has been previously considered in 3GPP in TR 26.928 (Cl. A.4 - Streaming of Immersive 6DoF, Cl 5.4 - XR Multimedia Streaming) and TR 26.998 (Cl. A.3 - Use Case 18: Streaming of volumetric video for glass-type MR devices).

On-demand volumetric video streaming allows to provide high-quality, professionally captured and produced volumetric video content. Some aspects of production and capturing systems for volumetric representation formats such as point clouds and meshes are documented in TR 26.928, clause 4.6.7.

Several use cases of on-demand volumetric video streaming can be envisioned related to various domains including education, entertainment or industrial monitoring. 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.

**Motivation for the use case**

In the use case “Immersive scenes”, the viewer can freely move around the scene and feel immersed in the experience.

NOTE 1: This use case and the use case “Volumetric video with single asset” are part of the same scenario “Streaming of Beyond 2D Produced Content”.

This use case finds application in several domains.

**Sports**: The viewer is able to freely choose a viewpoint to watch a live match or replay. The viewpoint can be dynamically changed for instance by interacting with a touch screen. It enables the viewer to have a more personal and immersive interaction with the sports match.

**Education**: 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.

**Entertainment**: Viewers can stream a performance from their favorite jazz band to their living room and experience greater immersion potentially together with spatial audio.

Brazilian SBTVD Forum has adopted volumetric video for inclusion in their [TV 3.0 standards](https://forumsbtvd.org.br/tv3_0/#panel-phase2) (support will not be mandatory in all receivers; focus on content distribution over the Internet and consumption on smartphones and HMDs). TV3.0 services are planned to be launched in 2025.

[DVB](https://dvb.org/) is running a study mission on volumetric video and first results are published [Study Mission Report S101](https://dvb.org/wp-content/uploads/2024/02/S101_Study-Mission-on-Volumetric-Video_Feb-2024.pdf).

Philips, InterDigital and Broadpeak are collaborating on an end-to-end implementation platform for packaging and delivery of volumetric video over content delivery network (CDN).

* <https://broadpeak.tv/newsroom/mpeg-v3c-standardized-content-distribution-at-scale/>
* <https://ir.interdigital.com/news-events/press-releases/news-details/2024/InterDigital-and-Broadpeak-Announce-Collaboration-on-MPEG-V3C-Standardized-Content-Distribution-At-Scale/default.aspx>

In the following sections we focus on the use case “Immersive scenes”.

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 on-demand streaming of beyond 2D video to a UE (Figure 1).

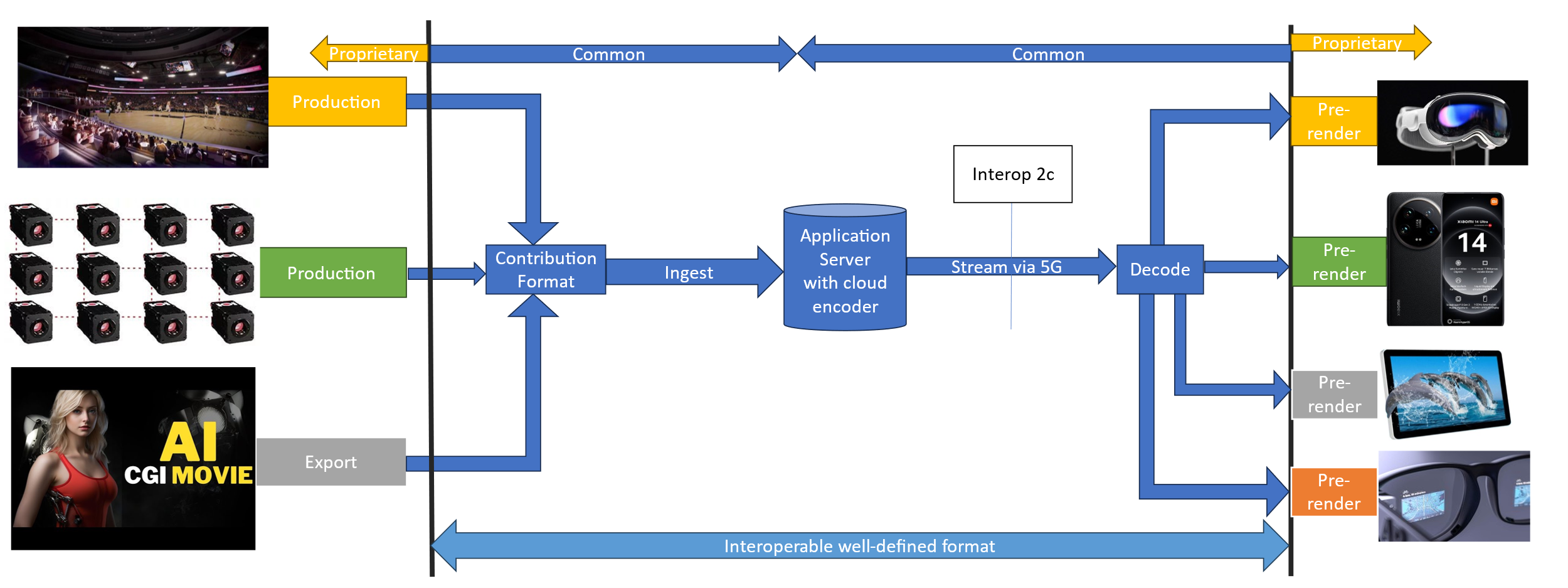


Figure 1: On-demand streaming to a UE

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

1. *Capturing and processing*

The beyond 2D video is captured and processed using multiple cameras. Zero or more of those cameras may be range-sensing cameras, and more than one of the cameras has color sensors. In the case of two or more cameras that are not rigidly connected, camera extrinsics are online calibrated. Depth estimation is performed to associate a full depth map with each of the camera views, thus resulting in a multi-view + depth representation.

Additional steps such as object instance segmentation and foreground/background separation may be performed to reduce the sample rate of the representation. This would result in a multi-view + depth + transparency/occupancy representation. All processing may be offline or with a delay of a few seconds.

1. *Encoding*

One codec that can be used to realize this scenario is MV-HEVC paired with a method to transmit camera parameters and a suitable 6DoF synthesizer.

Another 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 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 HEVC Main10 profile. The bitstream contains all camera parameters that are necessary for 6DoF rendering. Each atlas is independently renderable.

An example of multi-view data encoding has been described in paper “[Efficient Delivery and Rendering on Client Devices via MPEG-I Standards for Emerging Volumetric Video Experiences](mailto: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)”.

Multi-view video and multi-view + depth are well-known formats that have many public tools including OpenCV[[1]](#footnote-2), COLMAP[[2]](#footnote-3) and AliceVision[[3]](#footnote-4). Also MPEG has published tools for camera callibration and depth estimation (IVDE[[4]](#footnote-5)).

There are four typical workflows for multi-view (+ depth):

* Use color cameras to capture multi-view and estimate depth with multi-view consistency.
* Use range-sensing cameras to capture multi-view + depth and refine depth with multi-view consistency.
* Use AI or CG pipelines to raytrace views.
* Combinations of the above.

This study may include multi-view content for which some or all views lack depth information.

1. *Packaging and delivery*

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 Carriage of V3C data specifies how to map MIV (V3C) onto ISO BMFF, file format and DASH.

When a scene is represented by multiple atlases then only one would 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.

1. *Decoding*

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.

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 signaled 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 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 a random access reference frame structure.

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

The common atlas data with camera parameters could 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 would be not desirable 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 decoder and renderer processes are real-time. End-to-end latency may be in the range from 500 ms to multiple seconds.

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, Sony (in 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 6 to 25 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:

* 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".

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

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, 50 or 60 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 or ITU-R BT.2020 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 or ITU-R BT.2100 (HDR) 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 DASH 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. <https://opencv.org/> [↑](#footnote-ref-2)
2. <https://colmap.github.io/index.html> [↑](#footnote-ref-3)
3. <https://alicevision.org/> [↑](#footnote-ref-4)
4. <https://gitlab.com/mpeg-i-visual/ivde> [↑](#footnote-ref-5)