**3GPP TSG-SA WG4 Meeting #132 S4-250885**

**Fukuoka, JP, 19 – 23 May 2025**

**Source: China Mobile Com. Corporation, Tencent**

**Title: [FS\_Beyond2D] AI-Generated Beyond 2D**

**Agenda item: 9.7**

**Document for: Agreement**

**1. Introduction**

This proposal introduces AI-generated 3D mesh as an emerging AIGC technology. However, it is not intended for evaluation in the Rel-19 FS\_Beyond2D study.

**2. Proposal**

It is proposed to agree the following changes to the 3GPP draft TR 26.956 V0.4.0.

\* \* \* First Change (All New) \* \* \*

## 4.4 AI-Generated Beyond 2D content

### 4.4.1 General

Creating and capturing high-quality Beyond2D content is often a labour-intensive task that demands substantial time, expertise, or specialized capturing tools/devices, which limits the widespread adoption of Beyond 2D media. Artificial Intelligence Generated Content (AIGC) leverages AI technologies to autonomously produce content. For example, in clause 7.2.2.2, AI-powered 2D to stereoscopic 3D video methodology is introduced, which effectively reduces the reliance on high-end capture devices. Beyond this, AIGC encompasses a range of emerging technologies, including:

- Image-to-dynamic Mesh Generation,

- Text-to-dynamic Mesh Generation and 4D Scene generation,

which are described in the following sections. The commercialization of AIGC has attracted attention from both academia and industry, driving innovation in Beyond 2D content creation, compression technologies, and quality assessment methodologies.

Figure 4.4.1-1 illustrates a reference workflow for AI-generated beyond 2D content. The workflow positions a Media Generation AI/ML modelat the core of logical reasoning, including a large language model (LLM) transforming different inputs, such as text, image, video, 3D models, actuator signals and etc into a unified tensor representation. After reasoning and inference by the AI/ML model, the output tensor is mapped back to the target modality.

Une image contenant texte, capture d’écran, Police, ligne

Le contenu généré par l’IA peut être incorrect.

Figure 4.4-1: Workflow for AI-generated beyond 2D content

**- Representation:** The model should effectively represent and process different media types, such as text, images, video, and 3D models. Appropriate representation format should be selected for each type (e.g., CNNs for image features) to enable downstream processing and analysis.

**- Alignment:**Alignment refers to the process of matching and correlating data across different media types, enabling the model to comprehend their interrelationships. For instance, attention mechanisms can be employed to establish semantic correspondences between text and images.

**- Inference:** The model should be capable of inference capabilities, it can analyze and understand input data to extract useful information. A common approach is to leverage pre-trained large language modes (LLM) to perform inference tasks.

**- Generation:** The generative modeling techniques e.g., diffusion models, should be capable of generating new content, for instance, creating 3D mesh from text prompts.

**- Evaluation:** Assessing model performance (include both subjective methodologies and objective metric) is critical to ensure output relevance and reliability.

#### 4.4.2 AI-Generated Dynamic Mesh

#### 4.4.2.1 General

A growing number of AI-generated mesh tools now enable the direct generation of mesh models and textures from inputs such as text or images. Compared to traditional mesh production workflows, these tools offer significant advantages in terms of time efficiency. The examples of commercial services are provided below:

- AssetGen 2.0™: Meta's AI-powered 3D mesh generation system that produces models with "geometric consistency and fine-grained details" <https://developers.meta.com/horizon/blog/AssetGen2/>

- Hunyuan 3DTM: Tencent's 3D Mesh generation platform <https://3d.hunyuan.tencent.com/>

- MeshyTM: <https://www.meshy.ai/>

As the technology continues to advance, the quality and efficiency of AI-generated meshes are improving. However, there are still common issues that need to be addressed. including: excessively high polygon counts, poor topology, fragmented or irregular UV layouts, coarse texture details, baked-in lighting information in the textures, and insufficient accuracy in complex scenarios (e.g., clothing wrinkle simulation errors exceeding 15%).

#### 4.4.2.2 Image-Generated Dynamic Mesh

The task of generating dynamic meshes from images demands not only the creation of multiview geometric models based on the input image but also the extension into the temporal dimension to produce dynamic spatio-temporal content (4D). There are two main approaches for generating dynamic meshes, inference-based and optimization-based methods. As shown in Figure 4.4.2-1, the pipelines for these approaches include:

- Direct Generation: Directly generating dynamic meshes from input parameters without intermediate steps.

- Indirect Generation: Leverages diffusion models to produce multi-temporal and multi-view training data.

- Implicit Distillation: The process generates dynamic meshes through a multi-stage training framework, which combines multiple diffusion models via implicit distillation to derive generative priors.

- Explicit Supervision: Uses multi-modal data to provide explicit supervisory signals for dynamic mesh generation.

Une image contenant capture d’écran, conception

Le contenu généré par l’IA peut être incorrect.

Figure 4.4.2-1 Pipelines for image-based Dynamic Mesh generation

#### 4.4.2.3 Text-Generated Dynamic Mesh

Text-generated dynamic mesh requires both precise alignment between the object’s geometry and texture semantics, and accurate synchronization of its motion dynamics with describe actions or movements (4D). For example, a typical workflow may involve the following steps:

- Text Prompt Generation: Using large language models (e.g., GPT-4) to generate text prompts.

- Image Generation: Using diffusion models to generate single-view images based on these text prompts.

- Multi-view Synthesis: Leverage video or multi-view diffusion models to generate multi-view images from single-view images rendered from different angles.

- Dynamic Mesh Animation: Reconstruct 3D mesh and create a dynamic mesh (4D) by animating the vertices over time.

\* \* \* End of Changes \* \* \*