Knowledge Generation Pipeline using LLM for Building 3D Object Knowledge Base

SooHyung Lee
Contents Research Division
Electronics and Telecommunications Research Institute
DaeJeon, Korea
soohyung@etri.re.kr

HyeRin Lee
Department of Computer Science
Sookmyung Women's University
Seoul, Korea
hrin99@sookmyung.ac.kr

KiSuk Lee
Contents Research Division
Electronics and Telecommunications Research Institute
DaeJeon, Korea
kvr_lks@etri.re.kr

***Abstract*— With the wide spread of XR(eXtended Reality) contents such as Metaverse and VR(Virtual Reality) / AR(Augmented Reality), the utilization and importance of 3D objects are increasing. In this paper, we describe a knowledge generation pipeline of 3D object for reuse of existing 3D objects and production of new 3D object using generative AI(Artificial Intelligence). 3D object knowledge includes not only the object itself data that are generated in object editing phase but the information for human to recognize and understand objects. The target 3D model for building knowledge is the space model of office for business Metaverse service and the model of objects composing the space. LLM(Large Language Model)-based multimodal AI was used to extract knowledge from 3D model in a systematic and automated way. We plan to expand the pipeline to utilize knowledge base for managing extracted knowledge and correcting errors occurred during the LLM process for the knowledge extraction.**

Keywords — XR, Metaverse, 3D Object, Knowledge Base, Multi-Modal AI

I. INTRODUCTION

The proliferation of 3D objects in various domains, including XR services like the Metaverse and traditional media such as movies and dramas, has led to a exponential increase needs for the diversity and volume of 3D objects. The proprietary naming standards for 3D object is defined and applied by each contents companies in order to increase reusability of 3D object by editing caption that describe the object. However, due to the inherent subjectivity of the naming and captioning process, the resulting object names and captions often deviate from established standards. Additionally, in many cases, only basic and incomplete data of object is used during the naming and captioning process, consequently the potential reusability of object is usually restricted. To enhance reusability, it is essential for user or creator to provide the capability to search for and to employ objects based on human recognition levels, aligned with the contextual requirements of newly generated content. Consequently, object information is generated, organized and managed as high-level knowledge.

The generative AI technologies combined with natural language are applied to create new 3D object to respond to the growing demands for 3D objects. The quality of newly created 3D objects depends on the quality of training data set such as 3D-LLM. There are some restrictions on the type of objects for creation due to the limited availability of training data [1]. On the other hand, Point-E applies Point Cloud diffusion to the validated 2D image-LLM model known as CLIP[2],[3]. This approach is necessitated by the variability in the text inputs given to AI model to create 3D object can be any form, any sentence in any human domain. In such scenarios, the resource-intensive nature of training and applying the AI model becomes substantial. This burden can be mitigated by defining a finite set of words that encapsulate characteristic features of target domain, subsequently employed to drive the AI model. This approach reduces the temporal and computational demands, optimizing the overall process.

We believe that a combination of text-based and multimodal AI can provide valuable information about 3D objects. This information includes details like identification, shape, features, and context-specific insights. It's useful for managing and applying 3D objects in AI-related tasks. Many studies have attempted to merge 3D models with Large Language Models (LLMs) for different purposes, such as describing 3D models, answering questions about them, and helping robots navigate. To overcome limited training data, some studies aim to create a dataset for 3D-LLM by using an existing 2D image-LLM [4]. Our pipeline approach involves extracting knowledge information about 3D objects using a 2D image-LLM in a multimodal AI. More specifically, by establishing and employing target domain-specific knowledge key words, there is an anticipation that the development and application of a more cost-effective and compact text-based multimodal AI can be achieved.

This paper introduces a processing pipeline designed to create a knowledge base (KB) for 3D objects by leveraging domain-specific knowledge words and employing multi-modal AI techniques. Furthermore, it outlines a data model and the associated process for each stage, with a focus on data extraction to capture relevant knowledge information. The study also presents the results of applying a multimodal AI to generate knowledge information for each individual 3D object within the pipeline.

II. Pipeline for 3D Object Knowledge Usage

A. Pipeline Architecture

The pipeline consists of three core components: an ontology/data model used to define key characteristic knowledge descriptors of the target service domain, a multimodal AI model to extract 3D object knowledge information, and a knowledge base designated for correcting errors that may occur during the process of applying AI model.

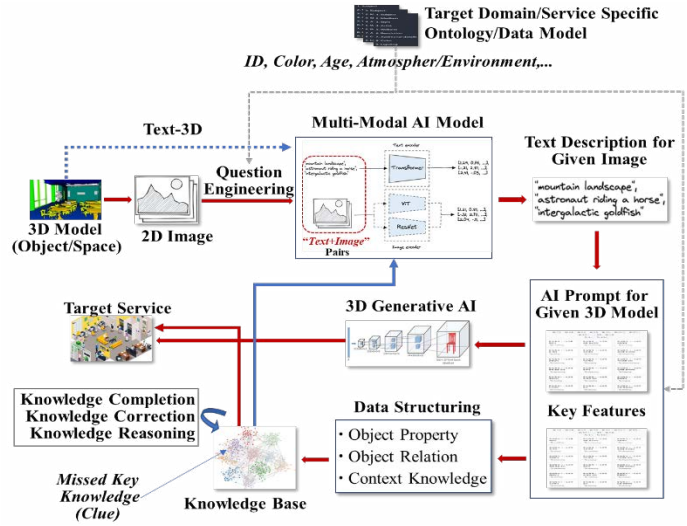

Fig. 1. Pipeline Conceptual Architecture

It is effective to use 3D-text multimodal AI, but few 3D object-Text AI models are currently available. In such cases, employing a 2D-text AI model can be pragmatic solution. This can be achieved by transforming the 3D object into a 2D image through a comprehensive rendering process with multiple viewpoints. This approach allows us to generates descriptive textual information corresponding to the objects. This generated object description text can be used to enhance the training of the 'text-3D object' AI model for given 3D object, thereby improving its capability to analyze and understand 3D objects more effectively.

The target domain ontology models embodies the foundational knowledge structure of the target service or application using 3D objects. Not only the ontology of the target domain but also the core data model can be applied. Within the pipeline, ontology is used to analyze the descriptive text for 3D object produced by the AI model, subsequently to prioritizes it according to the target domain's core knowledge. By constraining language text to domain-specific keywords that encapsulate pivotal attributes, the feasibility of knowledge construction and the utilization of cost-effective AI models for 3D objects can be realized.

- Ontology data includes the characteristics of objects themselves, the spatial and logical relations between objects, and the context characteristics of object in target services or applications.
- According to the keywords defined in the ontology, it is possible to identify and correlate words representing fundamental attributes from sentences describing 3D objects generated by AI model.
- These identified words can be subsequently used s used to structure the core knowledge/information about the 3D object to build the knowledge base.
- Each features from the identified words can be mapped to a prompt for an AI model to train and generate 3D object model at a low cost.

B. Procedure of Defining Data Model

The foundational element of the pipeline's data model is based on the target service model, encompassing services requirement, user requisites, use case, and employment scenarios. A conceptual data model is defined utilizing service model. It identifies the spatial characteristics and object components integral to the respective service, establishing the interrelation between each attributes and each objects. The ontology is delineated through an analysis of the service's contextual framework. Based on the conceptual data model, the physical data model to be applied to the actual service is defined. This physical data model will be used in generating knowledge base records, shaping elements within the service's user interface, and propelling the integrated AI models of the service.

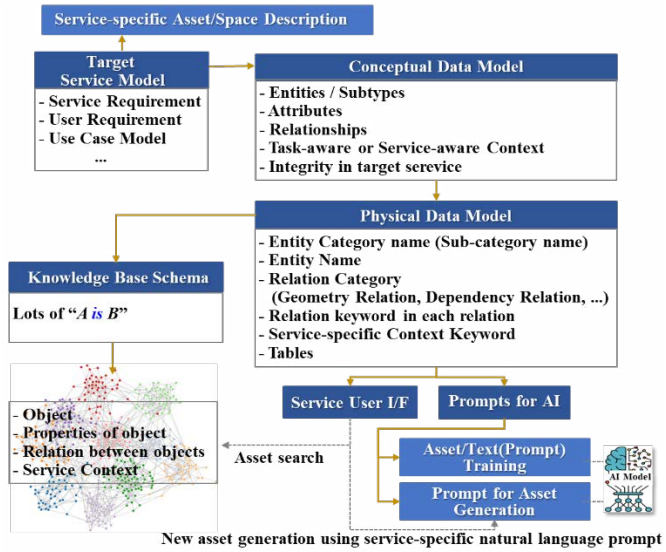

Fig. 2. Data Models in Pipeline

C. Precedure

The data schema is defined in alignment with the ontology established during the target service analysis phase. A process involves the rendering and transformation of a 3D object into a 2D image, subsequently subjected to a "text-image" AI model to generate descriptive text explain the attributes of the object. Keywords are chosen according to the target service's ontology, and core characteristic sentences of the 3D object are extracted from the previously generated description text.

Utilizing the core characteristic sentences previously extracted, contextual information and 3D object details are organized into structured data in accordance with the data schema. Contextual data encompasses elements like atmosphere,

usage, emotions, and spatial attributes of the object. The structured data are stored into the knowledge base in form of knowledge graph, and the prompts for generative AI models are devised in alignment with this structured data. This prompt functions as textual content, constituting a 3D object-text pair for generative AI training. During the service provisioning phase, it serves as an option data value, allowing users to input it for the creation of a new object using generative AI model.

III. Evaluation of 3D Object Knowlege Extraction

The feasibility of knowledge extraction from an existing 3D model lacking standardized naming and captions was evaluated, and the procedure is shown in "Fig 3". We employed the multimodal AI model miniGPT4, which operates on the '2D image-text' framework [5].

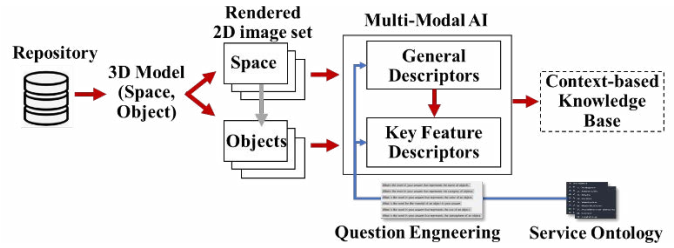

Fig. 3. Procedure for Extracting Knowledge from existing 3D Models

Initially, we tried to extract the knowledge information of the entire space and each object constituting the space at the same time. It was not possible to acquire the complete set of intended knowledge information for the entire space due to designed knowledge information for the entire space due to the limitations of identifying all component objects and the inaccuracy of spatial relationships between objects. Information about spatial atmosphere was derived through an analysis of the relationships among component objects and the objects themselves.

As miniGPT4 demonstrates a commendable ability to extract precise knowledge from individual objects, and knowledge is extracted by separating the object in 3D space as shown in "Fig.4".

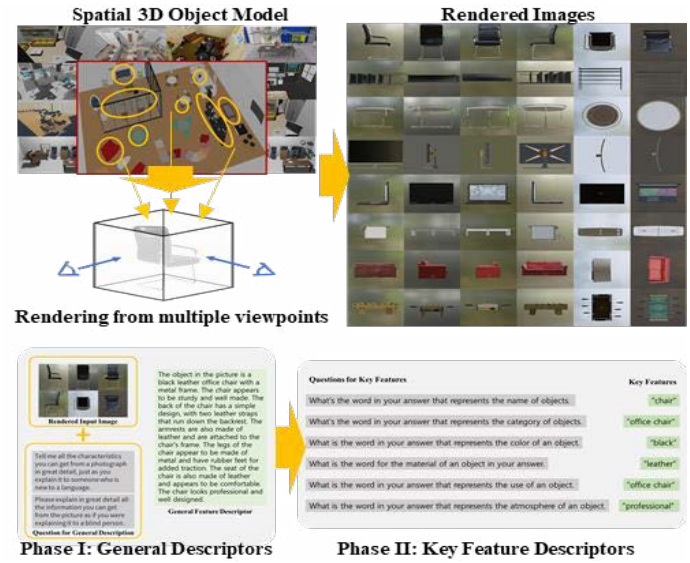

Fig. 4. Procedure for Extracting Knowledge from existing 3D Models

To obtain knowledge in a format suitable for structuring into a knowledge base, a question was formulated and employed to derive a concise response. However, the accuracy of the generated answer was relatively limited. To address this challenge, a two-fold approach was devised and implemented. The first step involves obtaining a comprehensive and profound overarching description of the object, while the second step focuses on acquiring a key feature descriptor that aligns with the intended knowledge framework.

IV. Future Work

Based on this study, the future research and development works are as follows. Construction of Knowledge Base using the accumulated 3D object data of joint research organizations, Enhancing knowledge augmentation through verification of knowledge and correction of knowledge error using the innate reasoning function within the knowledge base , and Design and creation of small 3D-LLM AI model in target service domain using the developed knowledge base.

Acknowledgment

This work was partly supported by Institute of Information &communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2023-00225441, "Knowledge Information Structure Technology for the Multiple Variations of Digital Assets").

References

[1] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, et al., "GET3D: A Generative Model of High Quality 3D Texturede Shapes Learned from Images," 36th Conference on Neural Information Processing Conference, 2022.

[2] Alex Nichol, Heewoo Jun, Prafulla Dharwai, Pamela Mishkin and Mark Chen, "Point-E: A System for Generating 3D Point Clouds from Complex Prompts," arXiv:2212.08751, 2022.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al., "Learning Transferable Visual Models from Natural Language Spervisioin," arXiv:2013,00020, 2021.

[4] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zengfang Chen, "3D-LLM: Injecting the 3D World into Large Language Models," arXiv:2037.12981, 2023.

[5] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, "MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models," arXiv2304.10592, 2023.