Sensor2Scene: Foundation Model-driven Interactive Realities

Yunqi Guo*, Kaiyuan Hou†, Zhenyu Yan*, Hongkai Chen*, Guoliang Xing*, and Xiaofan Jiang†
*The Chinese University of Hong Kong
Email: {yqguo, zyyan, hkchen, glxing}@ie.cuhk.edu.hk
†Columbia University
Email: kh3119@columbia.edu, jiang@ee.columbia.edu

***Abstract*—Augmented Reality (AR) is acclaimed for its potential to bridge the physical and virtual worlds. Yet, current integration between these realms often lacks a deep understanding of the physical environment and the subsequent scene generation that reflects this understanding. This research introduces Sensor2Scene, a novel system framework designed to enhance user interactions with sensor data through AR. At its core, an AI agent leverages large language models (LLMs) to decode subtle information from sensor data, constructing detailed scene descriptions for visualization. To enable these scenes to be rendered in AR, we decompose the scene creation process into tasks of text-to-3D model generation and spatial composition, allowing new AR scenes to be sketched from the descriptions. We evaluated our framework using an LLM evaluator based on five metrics on various datasets to examine the correlation between sensor readings and corresponding visualizations, and demonstrated the system's effectiveness with scenes generated from end-to-end. The results highlight the potential of LLMs to understand IoT sensor data. Furthermore, generative models can aid in transforming these interpretations into visual formats, thereby enhancing user interaction. This work not only displays the capabilities of Sensor2Scene but also lays a foundation for advancing AR with the goal of creating more immersive and contextually rich experiences.**

***Index Terms*—Augmented Reality, Sensor Data Integration, Large Language Models, Text-to-3D Generative Models, User Interaction in AR**

I. Introduction

Augmented Reality (AR) is transforming how we interact with the world, blending the digital and physical into a single seamless experience. As AR technology integrates into our daily routines, its capacity to enhance our understanding and engagement with our environment grows more evident. People are increasingly adopting AR for various applications, from improving everyday tasks to transforming professional workflows, including entertainment, social interactions, and remote working [13]. Affected fields include medicine [15], education [14], [26], and marketing [23]. This growing acceptance highlights AR's ability to bridge the gap between our physical reality and the virtual realm.

Despite its potential, current AR implementations struggle to fully integrate with the complexities of the physical world. Today's AR experiences, mainly confined to static, pre-designed environments like the virtual environments of Meta Quest [6] games or the 360-degree scenes offered by Apple Vision Pro [3], offer limited interaction with physical world sensors. Since current AR devices do not interact effectively with the physical world, they do not significantly differ from other screens like PCs and smartphones. Our vision for AR devices will effectively fill the gap between the physical world and digital devices. Firstly, AR can enhance human perception and interaction, allowing for augmented sensory experiences and a deeper environmental connection. It also improves accessibility by making sensory information, such as visualizing sounds for the hearing impaired, more available. Secondly, AR can advance professional applications and stimulate creativity. For instance, it can turn invisible data into visual formats, aiding in data interpretation in fields like research and industry, as seen in visualizing airflow in aerodynamics. Moreover, AR enhances creative and artistic expression, applying Kandinsky's synesthetic approach to introduce new creative dimensions in art and entertainment.

Achieving this goal presents three major challenges. First, interpreting data from a wide range of sensors is challenging due to the heterogeneity of sensor types and data formats available. Second, natural integration of sensor data into the user's environment to enhance their perception is complex. This complexity arises from the need to amalgamate various data modalities to create new scenes - tasks at which traditional methods cannot accomplish. Third, generating scenes from scratch is challenging. AR scenes must be adaptable, yet traditional game engines and design tools are labor-intensive and time-consuming, hampering prompt deployment and limiting scalability. Acknowledging these challenges, we leverage Large Language Models (LLMs) to help tackle these challenges.

LLMs have shown exceptional versatility in areas such as context understanding, summarization, creative content generation, and task planning [31]. Moreover, the recent study [30] indicates LLMs' capability to interpret sensor data effectively with appropriate prompts. This capability positions LLMs as an essential tool in overcoming the aforementioned challenges and, therefore, bridging the gap between digital and physical worlds. In this paper, we present an innovative framework and methodology designed to enhance human interaction with the physical environment through AR, leveraging the integration of sensor data with LLMs and generative models. Our contributions are as follows:

• **AR Interactive Framework with Sensor Data**. We pro-

pose an LLM-empowered framework, *Sensor2Scene*, that visualizes sensor data in AR environments. Our framework transforms intangible environmental data into contextually relevant, augmented visuals, significantly enhancing users' situational awareness and enriching their engagement with the surrounding environment.

- **AI Agent for Handling Sensor-Scene Interaction**. The Sensor2Scene framework is centered around an AI agent that not only visualizes sensor data as AR elements but also dynamically adapts these visualizations to reflect changes in the environment, sensor inputs, and user preferences. Leveraging LLM's decision-making and memory capabilities, our AI agent is able to efficiently generate AR scenes with customized experience.
- **Benchmark for Sensor Interpretation and Visualization.** We developed a new benchmark on the effectiveness of scene description from sensor input. This benchmark is used for evaluating the capability of Sensor2Scene to provide accurate and quality AR scenes across a variety of scenarios.

II. Related Works

The related work for this study involves two perspectives: the interaction between IoT sensors and AR, and scene production from data.

AR and IoT Data Interaction. Recent studies [16], [17] have explored the integration of sensor data with Augmented Reality (AR). They illustrate how AR can enhance the utility of IoT sensors in creating smart, interactive environments. These works highlight the potential synergy between AR and IoT technologies. Here, AR serves as an intuitive interface for users to interact with and understand data from IoT devices. This improves user engagement and comprehension of the real world. Further research [9], [22], [25] uses AR as a medium for visualizing IoT sensor data. This emphasizes AR's ability to present intuitive visualizations that utilize IoT's robust data collection and communication features.

However, these studies mostly focus on conventional visualization methods like numerical displays or standard charts and graphs. They often cater to specific sensor types or application domains. A significant gap in the existing literature is the lack of foundational models for the integration of AR and IoT. Our study shows that incorporating these models could extend the versatility and depth of interactions, enabling more dynamic and context-aware presentations.

AR Scene Generation. The potential of using AI to generate virtual content has become clear with the advancement of text-to-image and text-to-video technologies like Stable Diffusion [24] and Sora [20].

In our context of AR, scene and model creation is predominantly driven by text-to-3D and image-to-3D conversions. Recent advancements in text-to-3D model generation aim to address challenges like prolonged processing times and the consistency in the 3D model generation. A survey highlights the importance of using multi-view images and non-Euclidean data, such as meshes and point clouds, to improve 3D model creation [18]. DreamFusion [21] and Magic3D [19] introduce optimization methods that speed up the generation of high-quality 3D mesh models and enhance their resolution. DreamGaussian [28] improves processing speed with 3D Gaussian Splatting, while MVDream [27] uses multi-view priors to increase consistency and stability.

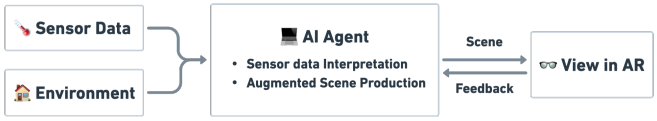

Fig. 1. **Sensor2Scene Framework**

However, current generation techniques are primarily limited to individual objects or components, making the creation process for each object quite time-consuming. Our work seeks to expand these capabilities, with the goal of generating comprehensive scenes that adapt dynamically to changes in sensor data, thus meeting a crucial need in the AR domain.

III. Sensor2Scene Design

We introduce Sensor2Scene, a framework designed to convert sensor readings into AR scenes. The framework acquires sensor data and environmental information from the sensor readings and user input, then processes this data to construct a 3D scene. This scene includes elements like visual indicators and is then rendered onto AR headsets, merging it with the real-world environment.

To achieve this, we developed an AI agent to interpret sensor data and autonomously generate immersive scenes. This agent enables the framework to operate with minimal human intervention. As illustrated in Fig. 1, the agent carries out two main functions: interpreting the sensor data into visible scenes and generating the AR scenes. The AI agent comprises two main components: *Sensor Data Interpreter* and *Scene Producer*. It also includes additional modules to interact with the user and the environment.

A. Scene Description Generator

LLMs have shown the capability to interpret real-world sensor data [30]. Furthermore, the associative memory in these AI models enables them to generate imaginary scenes from this sensory information. Inspired by these findings, we design our agent in the following steps:

- **Data Acquisition and Preprocessing:** Initially, the agent acquires a current scene observation from the AR goggles alongside raw sensor data from available sources. It then proceeds to extract location information of sensors and furniture, as well as format the raw sensor data according to a predefined schema. This process results in a consolidated set of metadata that includes environmental variables and physical parameters.
- **Prompt Generator Integration:** With the metadata generated from the previous step, we can explicitly provide expert knowledge in the system prompt, detailing the types of sensors available and the characteristics of the environment. The *Prompt Generator Integration* employs tailored prompt

templates for the current scenario, ensuring that the output is both contextually relevant and rich in detail.

- **Scene Description Synthesis:**
  Leveraging the processed prompts, the agent then embarks on synthesizing the scene description. This description manifests as a curated list of elements present in the environment, meticulously detailing objects, spatial relationships, and other pertinent features. Each element is identified, categorized, and described, providing a textual blueprint of the scene. This blueprint is crucial for the subsequent phase of the workflow, where text-to-3D model conversion occurs.

The Scene Description Generator, through its sophisticated processing of sensor data and environmental observations, lays the groundwork for the next stage of AR scene creation. By translating the complexities of the physical world into a structured and comprehensible narrative, it enables the 3D model generator to bring these descriptions to life in three dimensions. This synergy between components underscores our AI agent's innovative approach to enhancing human perceptions of AR environments, minimizing human intervention while maximizing the fidelity and immersion of the generated scenes.

B. AR Scene Producer

The 3D Scene Producer turns scene descriptions into tangible AR experiences. It uses textual descriptions to create intricate 3D models that fill the AR environment. Its features and functions aim to ensure that the virtual elements accurately represent their textual descriptions and can adapt dynamically to changes in the environment or sensor data. Below are the core functionalities and features of this module.

Functionality.

- Text-to-3D: At its core, the 3D Scene Producer is adept at interpreting the textual scene descriptions generated by the preceding module. Advanced text-to-3D modeling techniques transform these descriptions into precise 3D models. This process involves identifying each element described in the text, understanding its attributes (such as shape, size, and texture), and then constructing a corresponding 3D object.
- Local 3D Model Modification: Recognizing the dynamic nature of real-world environments, the module is equipped with a 3D model modifier capability. This feature allows for minor updates to the 3D models in response to changes detected in sensor information or the environment. Whether it's adjusting the color, position, size, or orientation of an existing model, the modifier ensures that the virtual scene remains in harmony with the physical world, enhancing the realism and immersion of the AR experience.

Features.

- Flexibility: The 3D Scene Producer stands out for its adaptability. It can work with a variety of objects and abstract scene components. This flexibility allows the module to render a broad spectrum of scenes accurately, from intricate interior settings to vast outdoor landscapes.
- Low-latency: Low-latency is crucial for rapidly changing sensor readings, such as those from light intensity sensors.

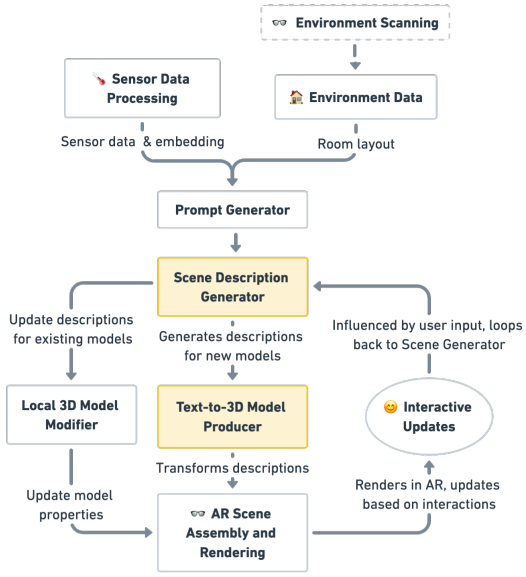

Fig. 2. **Scene Production Agent Workflow**

To handle these instantaneous changes, our local 3D model modifier can offer a swift response, modifying the input within less than 5 seconds with local image filters.

By bridging the gap between textual scene descriptions and their 3D realizations, the 3D Scene Producer plays a pivotal role in the creation of immersive AR environments. Its ability to dynamically adapt to changes and ensure compatibility with advanced display techniques underscores the module's contribution to the next generation of AR technology. Through its innovative approach to 3D scene production, this module exemplifies the synergy between linguistic understanding and visual representation, paving the way for more intuitive and engaging AR experiences.

C. Additional Modules

Additional modules are integrated into the system to complement the core functionalities of the AR scene generation agent. These modules are designed to enhance the realism, interactivity, and adaptability of the AR scenes, ensuring a seamless and engaging user experience. Below, we detail these modules, prioritized by their impact on the system's overall performance.

User Feedback Integration. To learn the user's preferences, we leverages a context-aware architecture that incorporates user's previous choices and decision with memory. This establishes a dynamic feedback loop between users and the AR environment. This loop helps two parts in the agent:

- *Scene Refinement:* By continuously incorporating user preferences and feedback, the system adapts and refines the AR scene in real time. This ensures that the virtual environment not only aligns with the user's expectations but also enhances their interaction experience by personalizing the scene content.
- *Model Fine-Tuning:* User interactions and feedback provide valuable data that contribute to the training set for ongoing

model fine-tuning. This iterative process allows the system to learn from user behavior, improving its accuracy and responsiveness over time, thus ensuring that the AR scenes become progressively more immersive and aligned with user preferences.

Digital Boundary. The Digital Boundary module is used to make 3D scenes integrated with the user's physical environment. By creating constraints based on room size, object dimensions, positions, and physical laws such as gravity, this module guarantees that virtual elements are correctly scaled and positioned within the space. This not only enhances the realism of the AR experience but also prevents immersion-breaking anomalies, such as objects floating in mid-air or intersecting with physical barriers.

Environmental Scanning. Utilizing the camera on AR glasses, it interprets the immediate physical environment, mapping out space and identifying key features. This data can then be used to enhance the accuracy of the Digital Boundary and Scene Description Generator modules, streamlining the process of aligning virtual elements with the real world. For scenarios where AR glasses are not equipped with sophisticated environmental scanning capabilities, manual configuration offers a flexible alternative, allowing users to tailor the AR experience to their specific context.

In summary, our AR scene generation system integrates the Scene Description Generator, 3D Scene Producer, and additional modules to create immersive and interactive AR experiences. The seamless workflow among these components, from interpreting sensor data to producing and modifying 3D scenes within real-world constraints, is visually summarized in Fig. 2. This cohesive structure ensures our system delivers a realistic and engaging AR environment, tailored to enhance user interaction and perception.

IV. Implementation

In developing our system, we focused on building a functional agent and visualization framework, along with selecting optimal models for data interpretation and scene visualization.

AI Agent and Renderer for AR. Our AI agent is constructed using LangChain [12], chosen for its universal function call capability that simplifies task sequencing with LLMs. This streamlines processes from sensor data interpretation to AR visualization. For rendering AR scenes, we integrated WebXR [5] with three.js [7], ensuring a cross-platform, lightweight solution for immersive AR experiences.

LLM and 3D Generator. For LLM tasks, we tested GPT-3.5 [1], [11] and GPT-4 [2], [8], selecting them based on their ability to generate accurate scene descriptions. DreamGaussian [28], MVDream [27], and Genie [4] are selected for text-to-3D conversion due to their speed advantage over other models, critical for our system settings. Our system can operate on a single server equipped with 24GB of GPU RAM for scene production and can render at over 90 FPS on Quest 3.

V. Preliminary Evaluation

In our evaluation framework, a critical aspect we assess is the Scene Production Quality, which encompasses both the quality of generated scene descriptions and the fidelity of the produced 3D scene. This subsection provides a preliminary examination of these modules.

A. Scene Description Benchmark

We first evaluate our *Scene Description Generator*, which is the foundation for the following modules. The quality assessment of scene descriptions presents a unique challenge due to its inherently subjective nature. This subjectivity raises questions about the credibility of a manual scoring approach that relies solely on surveys, as it may be prone to biases and inconsistencies. To this end, we have implemented a bifurcated approach to evaluation that integrates scores derived from surveys and a systematic examination of schematic evaluation.

The goal of the scene description is to ensure a comprehensive understanding and interpretation of sensor data, capturing all critical attributes without omission. To achieve this, we meticulously designed our evaluation criteria for the LLM, focusing on several key aspects as following:

1. **Specificity**: Measures the precision of descriptions, emphasizing detail and minimizing ambiguity.
2. **Fidelity**: Evaluates how accurately descriptions match sensor data, with deductions for inaccuracies and unnecessary embellishments.
3. **Integration**: Accesses the innovation and effectiveness for incorporating diverse sensor data into AR environment.
4. **Utilization**: Examines the application of sensor data in augmenting the AR experience, emphasizing relevance and utility over mere presence.
5. **Coherence**: Evaluates the contextual naturalness of descriptions, ensuring sensor data is woven into the narrative in a way that feels inherent and fitting.

Environmental Dataset. Two public sources [10], [29] and an additional dataset collected by ourselves, referred to as the Self-Collected Environmental Dataset (SCE), are used in this benchmark. The public dataset from [10] includes measurements from four indoor environments: bedroom, kitchen, living room, and master bedroom, capturing data across four sensing modalities: humidity, pressure, light intensity, and temperature. The dataset in [29] encompasses data from two locations—gym and living room—with measurements in humidity, pressure, and temperature. There are four more sensing modalities contained in SCE—CO2, VOC, PM2.5, and wind speed—targeting environmental conditions at a subway station.

As mentioned, our evaluation methodology for the scene descriptions takes both an automated evaluator using the LLM and human judgment across five distinct metrics. We selected samples from each dataset, structuring the sensor data in JSON format. These samples are then processed by the scene description generator. The LLM evaluator determines scores ranging from 1 to 5 for each metric, and then we repeat

TABLE I
DETAIL OF LLM EVALUATOR AND MANUAL SCORE ON TEXT GENERATED FROM SCENE DESCRIPTOR.

	LLM Evaluator						Manual Score
Dataset	Specificity	Fidelity	Integration	Utilization	Coherence	Avg.	
BDL	4.45	4.67	4.19	4.53	4.16	4.61	4.53
kjgrct2yn3.3	4.75	4.78	3.83	4.71	3.85	4.44	4.20
SCE	4.67	4.76	4.35	4.88	4.33	4.62	4.31

Prompt:
You are a sophisticated assistant designed to create augmented reality scene descriptions. These descriptions are customized to the user's current circumstances (environment, sensor readings) and personal preferences. Your main role is to produce detailed textual depictions of potential scenes that the data could represent.
Your descriptions should include specific objects with
- a clear and concise description of the scene
- the objects in the scene with their properties (e.g., size, color, shape, position, orientation, texture, etc.)
- the relationships between the objects (e.g., distance, direction, etc.)

Remember, the goal is to help users visualize the invisible data collected by sensors in a manner that is not only accurate and informative but also engaging and immersive, bridging the gap between raw data and human experience through the power of AR.

Fig. 3. **Prompt for Generating Scene Descriptions**

this process 30 times for each dataset. As shown in Table I, our Scene Description Generator exhibits commendable performance in terms of *Fidelity* and *Utilization*, signifying its proficiency in interpreting and incorporating the majority of sensor data. On the other hand, relatively lower scores in *Integration* and *Coherence* highlight challenges faced by the LLM in accurately mapping sensor data to tangible real-world entities, thereby affecting the visualization quality, and hence this motivates us to incorporate human feedback to improve the scene description.

B. Quality of Generated AR Scenes

Figure 4 showcases a scene generated for the Meta Quest 3. This scene was created in an office in Hong Kong, following the prompt result (Figure 3) and sensor readings. The scene incorporates data collected from temperature, humidity, noise level, and air quality sensors. The scene description generator signifies the moisture condition with a moist window, the office noise with a large trumpet, and the good air quality with a plant. An online demo is available at https://t.ly/wep7p and can be experienced immersively using AR goggles.

We evaluated the versatility using benchmark data, covering various locations and sensor settings. Figure 5 displays three example scenes created by Sensor2Scene. The figures are rendered with Blender to depict the AR environment. From these scenes, it's apparent that Sensor2Scene accurately interprets sensor inputs in relation to the environment. For example, it signifies humidity in the countryside with a pond. In Dubai's desert environment, it uses a plant to symbolize room humidity.

Evaluation of 3D Model Generation Quality. The quality of 3D models generated from textual descriptions is paramount for an immersive user experience in augmented reality. These models' visual accuracy, texture detail, and overall realism significantly impact user engagement and perception. In our

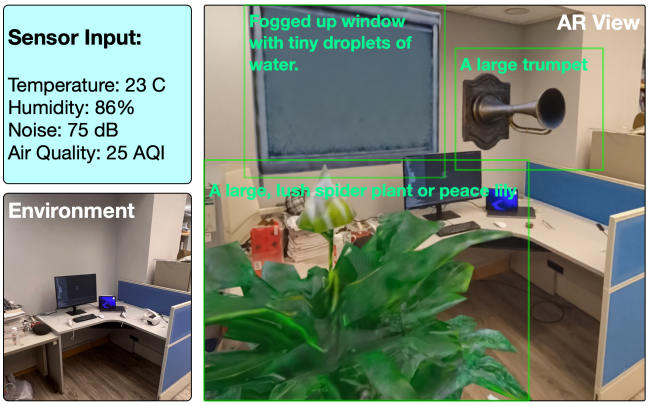

Fig. 4. **Real-world Demonstration, AR Scene Generated from Sensor Readings, Screenshot from Meta Quest 3.** The sensor is reading in a moist, noisy office in Hong Kong. *Scene description*: Illustrate the high humidity with a window in the office showing condensation. The window is slightly fogged up, with tiny droplets of water visible, indicating the moist environment. A large, artistically crafted trumpet or horn-shaped sculpture mounted on a prominent wall in the office indicates the noise levels. A large, lush spider plant or peace lily is prominently displayed, indicating good air quality.

(a) Kansas – Fall (b) Dubai – Summar (c) Iceland – Winter

Fig. 5. **Scenes Generated from Sensor Readings.** (a) Kansas countryside monitored with temperature, humidity, and motion sensors. Scene: a holographic water lily in a small digital pond; a series of small, playful holographic chickens that roam around the coop area; and a tree with leaves that change color based on the air quality index. (b) Dubai indoor office monitored with temperature, humidity, and motion sensors. Scene: a large, interactive wall canvas that changes color based on the temperature; a small, delicate wind chime hangs near the patio door; a holographic plant that sits on the kitchen counter; and dynamic footprints leading towards the coop. (c) Iceland's room in the winter monitored with temperature, humidity, and light sensors. Scene: a virtual fireplace that visually represents the room's chilly temperature; a series of small, dormant holographic foxes; and a dynamic sun icon that adjusts its brightness and size based on the light levels.

preliminary evaluation, we examined the models for these attributes to determine their effectiveness in rendering realistic AR scenes.

Our findings indicate that while the open-source models DreamGaussian [28]and MVDream [27] exhibit potential, they tend to produce models with less detailed textures unless their parameters are finely tuned for specific scenarios. In contrast, the online interface Genie [4] demonstrates a more consistent ability to interpret complex text descriptions, successfully capturing and rendering intricate element details more effectively than the other models evaluated. This comparative analysis underscores the importance of model selection in the context of text-to-3D conversion for achieving high-quality, realistic AR experiences.

Element Placement Accuracy. Accurate object placement is essential for an immersive AR experience. We assess element placement within the AR environment by measuring the

alignment of virtual objects with their intended locations and evaluating any overlapping that may occur. Our observations indicate that GPT models [1], [2] can understand the spatial placement of individual objects. For instance, they recognize that fans should hang from the roof while chairs belong on the ground. However, these models struggle to detect collisions between objects. In our evaluation, we found instances of object collision in the scene (3 out of 13). This issue can be addressed with future collision detection solutions.

This preliminary evaluation of Scene Production Quality aims to verify that our system accurately interprets environmental data and effectively translates these interpretations into AR experiences. Through this initial assessment, we seek to pinpoint both the strengths and potential areas for enhancement within our system, setting a foundation for its future refinement.

C. Preliminary User Study

In our preliminary user study, participants assessed the AR scene created by our AI agent, comparing the experience with that of traditional sensor screens. They evaluated the models on visual appeal, intuitiveness, and immersion. While the feedback highlighted strengths in immersion and intuitiveness of the sensor data interaction, it also recommended improvements in the details of the 3D models and enhancing environmental interaction by creating dynamic objects in the scene.

VI. Use Cases and Applications

In our exploration of AR applications, we identify two core domains where our system exhibits significant potential: enhancing human sensory experiences and advancing professional and creative practices.

Firstly, our system can significantly augment human sensory interactions, particularly beneficial for accessibility. For example, visualizing sound for the deaf transforms abstract auditory information into visual representations, facilitating communication and interaction. This application not only showcases the technological prowess of our system but also its profound societal impact, making environments more inclusive and engaging.

Secondly, in professional contexts, such as aerodynamics, our system can visualize complex data like airflow in an intuitive manner. This real-time visualization aids engineers in understanding and optimizing designs more efficiently, showcasing the system's impact on accelerating innovation. Similarly, in the arts, our system enables the creation of immersive experiences that translate sounds or emotions into visual forms, expanding creative boundaries and offering new artistic expressions.

VII. Discussion and Future Work

In the process of developing and evaluating our AR system, we have encountered several limitations and identified potential directions for future research. Our current system generates scenes using individual meshes, which presents two main limitations: limited mesh quality and inconsistencies in placement and style. These issues can detract from the immersive experience and realism of the AR environment. Our future work aims to address these limitations and explore new methodologies for enhancing AR scene generation. Two promising areas of research include:

Comprehensive Evaluation on Generation Quality and System Efficiency. We will expand the preliminary evaluation to a comprehensive assessment that includes complete user studies and latency evaluations for each module. Currently, processing prompts and text-to-3D generation take between 3 seconds and 5 minutes for end-to-end scene production.

Enhanced Scene Generation through Text-to-Video and 3D Gaussian Splatting. Leveraging text-to-video technology, we aim to generate comprehensive videos showcasing multiple views (20 to 200) of a static scene. These videos will then be processed using 3D Gaussian Splatting, a technique that transforms the video into a Gaussian representation suitable for AR integration. This approach not only promises to bypass the issues associated with individual meshes by generating the entire scene in a unified manner but also ensures higher visual fidelity. The application of 3D Gaussian splatting is particularly exciting for its potential to accurately simulate complex visual effects such as lighting, reflection, and refraction. Moreover, it offers improved handling of transparent objects like fog, smoke, and clouds, thus enhancing the realism and depth of AR scenes.

Scene-to-Scene Augmentation. Recognizing the limitations inherent in relying solely on textual descriptions for environment and sensor data, we plan to introduce scene-to-scene augmentation. This technique will utilize the 3D scene captured by the AR headset, combined with sensor data, to create an enriched augmented view. This approach aims to mitigate the issue of missing information during scene extraction, providing a more comprehensive and accurate representation of the physical environment.

VIII. Conclusion

In this study, we introduced Sensor2Scene, a system framework designed to enable user interaction with sensor data via augmented reality. We developed an AI agent utilizing multiple language models to understand the implicit messages from sensor readings and produce tangible scenes. Sensor2Scene shows promising results based on a joint scoring of manual and LLM evaluator under a set of rubrics. We demonstrated a few examples on visualizing the sensor data in the real world with this framework.

This work motivates further research on leveraging foundation models for the interpretation of IoT sensor data. Furthermore, the adoption of large language models is helpful in converting these interpretations into visual information, thereby enhancing user interaction. This study represents an initial effort in this domain. Currently, our testing scope is limited, but we aim to expand it in the follow-up research.

REFERENCES

[1] "Gpt-3.5: Language models by openai," https://openai.com/, 2023, accessed: 2024-03-03.

[2] "Gpt-4: Language models by openai," https://openai.com/, 2023, accessed: 2024-03-03.

[3] "Apple vision pro," https://www.apple.com/apple-vision-pro/, 2024, accessed: 2024-03-03.

[4] "Genie by lumalabs," https://lumalabs.ai/genie, 2024, accessed: 2024-03-03.

[5] "Immersive web developer home - webxr," 2024, accessed: 2024-03-03.

[6] "Meta quest," https://www.meta.com/quest/, 2024, accessed: 2024-03-03.

[7] "Three.js – javascript 3d library," 2024, accessed: 2024-03-03.

[8] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[9] M. Alonso-Rosa, A. Gil-de Castro, A. Moreno-Munoz, J. Garrido-Zafra, E. Gutierrez-Ballesteros, and E. Cañete-Carmona, "An iot based mobile augmented reality application for energy visualization in buildings environments," *Applied Sciences*, vol. 10, no. 2, p. 600, 2020.

[10] S. M. H. Anik, X. Gao, and N. Meng, "A comprehensive indoor environment dataset from single-family houses in the us," 2023.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[12] H. Chase, "LangChain," Oct. 2022. [Online]. Available: https://github.com/langchain-ai/langchain

[13] L. F. de Souza Cardoso, F. C. M. Q. Mariano, and E. R. Zorzal, "A survey of industrial augmented reality," *Computers & Industrial Engineering*, vol. 139, p. 106159, 2020.

[14] Y. Guo, J. Zhao, B. Ding, C. Tan, W. Ling, Z. Tan, J. Miyaki, H. Du, and S. Lu, *Sign-to-911: Emergency Call Service for Sign Language Users with Assistive AR Glasses*. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3570361.3613260

[15] K. Hou, S. Xia, E. Bejerano, J. Wu, and X. Jiang, "Arsteth: Enabling home self-screening with ar-assisted intelligent stethoscopes," ser. IPSN '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 205–218. [Online]. Available: https://doi.org/10.1145/3583120.3586962

[16] D. Jo and G. J. Kim, "Ar enabled iot for a smart and interactive environment: A survey and future directions," *Sensors*, vol. 19, no. 19, p. 4330, 2019.

[17] J. C. Kim, T. H. Laine, and C. Åhlund, "Multimodal interaction systems based on internet of things and augmented reality: A systematic literature review," *Applied Sciences*, vol. 11, no. 4, p. 1738, 2021.

[18] C. Li, C. Zhang, A. Waghwase, L.-H. Lee, F. Rameau, Y. Yang, S.-H. Bae, and C. S. Hong, "Generative ai meets 3d: A survey on text-to-3d in aigc era," *arXiv preprint arXiv:2305.06131*, 2023.

[19] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, "Magic3d: High-resolution text-to-3d content creation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 300–309.

[20] OpenAI, "Introducing sora," https://openai.com/sora, 2024, accessed: 2024-03-03.

[21] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[22] A. Protopsaltis, P. Sarigiannidis, D. Margounakis, and A. Lytos, "Data visualization in internet of things: tools, methodologies, and challenges," in *Proceedings of the 15th international conference on availability, reliability and security*, 2020, pp. 1–11.

[23] P. A. Rauschnabel, B. J. Babin, M. C. tom Dieck, N. Krey, and T. Jung, "What is augmented reality marketing? its definition, complexity, and future," pp. 1140–1150, 2022.

[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[25] J. Rosales, S. Deshpande, and S. Anand, "Iiot based augmented reality for factory data collection and visualization," *Procedia Manufacturing*, vol. 53, pp. 618–627, 2021.

[26] Q. Shao, A. Sniffen, J. Blanchet, M. E. Hillis, X. Shi, T. K. Haris, J. Liu, J. Lamberton, M. Malzkuhn, L. C. Quandt, J. Mahoney, D. J. M. Kraemer, X. Zhou, and D. Balkcom, "Teaching american sign language in mixed reality," vol. 4, no. 4, dec 2020. [Online]. Available: https://doi.org/10.1145/3432211

[27] Y. Shi, P. Wang, J. Ye, M. Long, K. Li, and X. Yang, "Mvdream: Multi-view diffusion for 3d generation," *arXiv preprint arXiv:2308.16512*, 2023.

[28] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng, "Dreamgaussian: Generative gaussian splatting for efficient 3d content creation," *arXiv preprint arXiv:2309.16653*, 2023.

[29] A. Vela, J. Alvarado-Uribe, and H. G. Ceballos, "Indoor environment dataset to estimate room occupancy," *Data*, vol. 6, no. 12, 2021. [Online]. Available: https://www.mdpi.com/2306-5729/6/12/133

[30] H. Xu, L. Han, Q. Yang, M. Li, and M. Srivastava, "Penetrative ai: Making llms comprehend the physical world," 2024.

[31] M. Zhao, J. Xia, K. Hou, Y. Liu, S. Xia, and X. Jiang, "Rasp: A drone-based reconfigurable actuation and sensing platform towards ambient intelligent systems," 2024.