2025 European Conference on Computing in Construction

CIB W78 Conference on IT in Construction

Porto, Portugal
July 14-17, 2025

AUTO-BIM DATA-EXTRACTION FOR ANALYZING SPACES & AI TRAINING: COMBINING SPATIAL OBJECT/PROPERTIES WITH LLM-GENERATED DATA

Bomin Kim[1], Jin-Kook Lee1[*], Hayoon Kim1, Youngchae Kim[1], and Seung Hyun Cha[2]
[1]Yonsei University, Seoul, Republic of Korea
[2]Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea
[*] leejinkook@yonsei.ac.kr

Introduction

Building Information Modeling (BIM) has become a fundamental methodology in the architecture, engineering, and construction (AEC) industries for digitally managing and sharing spatial and asset data throughout the project lifecycle (Eastman et al., 2018). Floor plan layouts reflect multiple social and functional requirements, including circulation patterns, accessibility, and the balance between public and private spaces. These factors significantly influence spatial configuration, building performance, and user satisfaction (Hillier & Hanson, 1984; Park, Lee, & Shin, 2011).

Recent studies have increasingly focused on automating the analysis and structuring of BIM spatial data to improve efficiency and support data preparation for design validation and artificial intelligence (AI)-driven applications (Gatto, 2024). This study presents a method for automatically extracting spatial information from BIM models and generating paired datasets consisting of floor plan images and descriptive text produced by large language models (LLMs). The objective is to quantify spatial configurations and proportional relationships to support both design evaluation and AI model training.

Methodology

The developed method automates the extraction of BIM spatial data and enhances it using LLM-based processing. Floor plan images and room-level attributes, including area, function, dimensions, and coordinates, are extracted from the BIM model.

Custom nodes created through visual scripting tools within the BIM environment control the data extraction. The results are organized into a structured building–floor–room hierarchy and exported as text files compatible with external analysis tools.

Key spatial metrics, such as room-to-floor area ratios, seating densities, and shared-to-private space ratios, are calculated using predefined values. These metrics and the raw attribute data are processed by an LLM to produce descriptive summaries, such as “The meeting room occupies 25% of the total floor area.” Annotated floor plans visually display the spatial divisions to improve clarity and usability.

The process has been implemented through a user interface developed with the Revit API, enabling seamless data extraction, image generation, and text production.

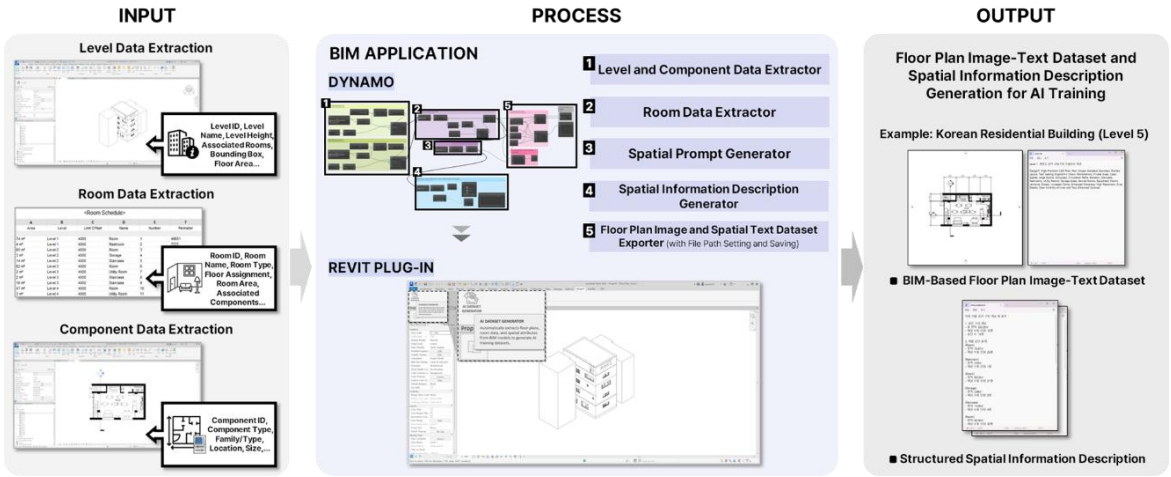

Figure 1: Overview of BIM-based spatial data extraction and ai dataset generation

Results

The method was demonstrated using two Korean BIM models representing office and residential buildings. Spatial objects and attributes were successfully extracted and organized according to the structured format. The generation of descriptive texts and annotated drawings was completed within three minutes per model.

The calculated indicators accurately represented the distribution of floor areas, seating densities, and shared-to-private space ratios. LLM-generated descriptive texts combined quantitative and contextual information to help users with limited BIM expertise understand spatial configurations.

The combination of annotated images and descriptive texts reduced manual work and minimized errors. Paired dataset files(.txt, .png) were saved under consistent filenames, simplifying use for design comparison and AI dataset creation. The process consistently produced reliable results across multiple executions.

Conclusions

This study presents a method for automatically extracting and organizing BIM spatial data, producing LLM-based descriptive texts and annotated floor plan images as integrated datasets. Testing with two Korean case studies confirmed the method's ability to support design evaluation and facility management by providing structured spatial data and clear descriptive outputs.

Further work will focus on extending compatibility with additional BIM platforms, adding regulatory review and cost analysis functions, and quantitatively evaluating performance with large-scale BIM models. The method provides a foundation for further applications in generative design, automated alternative evaluation, and data-driven facility management.

Acknowledgements

This work is supported in 2025 by the Korea Agency for Infrastructure Technology Advancement(KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant RS-2021-KA163269).

This work was supported by the Yonsei Quantum Computing Supporting Program of 2024-12-0259.

References

Eastman, C., Teicholz, P., Sacks, R. & Liston, K. (2018) BIM handbook: A guide to building information modeling for owners, designers, engineers, contractors, and facility managers. 3rd ed. Hoboken, NJ: Wiley.

Gatto, C., Cassandro, J., Mirarchi, C. & Pavan, A. (2024) LLM based automatic relation between cost domain descriptions and IFC objects. In: Proceedings of the 41st International Conference of CIB W78. Marrakesh, Morocco. Marrakesh, Morocco, CIB, pp. 1–10.

Hillier, B. & Hanson, J. (1984) The social logic of space. Cambridge, UK: Cambridge University Press.

Park, K., Lee, S. & Shin, J. (2011) A study on the importance of building design factors influencing office value assessment. Seoul Urban Studies, 12(1), pp. 53–71.