Kalpa Publications in Computing

Volume 22, 2025, Pages 601–611

Proceedings of The Sixth International Conference on Civil and Building Engineering Informatics

Optimized Language-Embedded 3DGS for Realistic Modeling and Information Storage of Historical Buildings

Zhenyu LIANG[1] , Jeff Chak Fu CHAN[1], Jiaying ZHANG[1], Zhaolun LIANG[1], Boyu WANG[1], Mingzhu WANG[2] and Jack C.P. CHENG[1]

[1] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR.

[2] Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong, SAR.

zhenyu.liang@connect.ust.hk, cfchanay@connect.ust.hk, jzhangfh@connect.ust.hk, zliangaq@connect.ust.hk, bwangbb@connect.ust.hk, m.wang@cityu.edu.hk, cejcheng@ust.hk

Abstract

A realistic and informative 3D digital model of historical buildings holds significant value for heritage preservation, public education, and cultural dissemination. Traditional digital representations, such as Heritage Building Information Modeling, panoramic images, LiDAR point clouds, photogrammetric mesh models, face limitations in user interaction and engagement. The automatic generation of a semantically enriched 3D model requires advanced scene-understanding capabilities. Pre-trained zero-shot methods struggle with domain-specific knowledge in heritage component semantics, while CNN-based approaches demand extensive manual effort for dataset preparation and model training. Therefore, this study proposes an optimized language-embedded 3DGS framework for the digitalization of historical buildings. It involves three steps: (1) data preparation of on-site images and relevant text; (2) component segmentation by the integration of SAM and MLLM; (3) scene reconstruction using the language-embedded 3DGS. The combination of SAM's localization ability and MLLM's in-context learning achieves 95.6% accuracy in the semantic segmentation of historical building components, requiring only a single annotated sample for each component category. Compared with previous methods, our language-embedded 3DGS model accurately captures complex semantics while providing realistic appearance and convenient navigation. The generated 3D model can be further integrated with an LLM-based chatbot assistant to achieve open-vocabulary and vague searches. This framework was validated on the Shishi Sacred Heart

I. Hkust, J. Cheng and Y. Yantao (eds.), ICCBEI 2025 (Kalpa Publications in Computing, vol. 22), pp. 601–611

Cathedral in Guangzhou, China, offering a novel digital solution for the protection and sustainment of historical buildings.

1 Introduction

Historical buildings record the history and social transformations of a region. Those with religious significance, such as cathedrals, often serve as spiritual homes for communities. Effective documentation, preservation, and dissemination of these heritages are essential for cultural protection and sustainability. To capture and record the current conditions of historical buildings, advanced techniques are employed to create digital models that reflect geometric shapes, textures, and attribute information, including component semantics and defects. A 3D heritage model with precise geometry and comprehensive information is valuable for various applications such as archiving, maintenance, navigation, and education (Yang, 2020).

Several types of digital models are employed for the preservation and dissemination of historical buildings. Heritage Building Information Modeling (HBIM) is widely used due to its superior capability to manage and document historic structures with parametric objects (Cheng, 2024). HBIM can be further transformed into mechanical models, such as Finite Element Analysis models, to comprehensively analyze structural performance and proactively prevent defect development (Ursini, 2022). However, HBIM struggles to accurately reflect the realistic appearance and aesthetics of historical buildings, which limits its applications in areas involving public interaction, such as remote viewing and on-site navigation. Besides, panoramic images with informative labeling are popular for effectively displaying scene textures while storing a certain amount of information. Nevertheless, users cannot move continuously within these panoramic scenes, and texture distortion may occur, diminishing the user experience. Additionally, reality-capturing techniques, such as LiDAR point cloud scanning and photogrammetric mesh 3D reconstruction, can accurately capture geometrical shapes and textures (Croce, 2021; Pritchard, 2017). However, they do not provide a first-person perspective for user visualization and interaction, which is inconvenient for navigation and education in VR environments or on mobile devices. To enhance user interaction and broaden public engagement with digital models of historical buildings for effective cultural dissemination, it is crucial to propose a novel digital representation that overcomes these existing limitations.

3D Gaussian Splatting (3DGS) is a recently developed 3D reconstruction technique based on generative artificial intelligence (Kerbl, 2023). This method takes images or videos as input and employs an iterative learning process to train an explicit model composed of 3D Gaussian ellipsoids for displaying the scenes. This technology can render realistic first-person view images, supporting fast training and rendering speeds, allowing users to navigate within the scene. It is an ideal 3D digital representation for user interaction with historical buildings.

However, the basic 3DGS model is primarily used for rendering colorful scenes, lacking object-based information. Consequently, several studies have focused on the semantic enrichment of the 3DGS model (Qin, 2024; Shi, 2024). Rather than merely adding semantic labels to the scene, language features are preferred, as they enable users to interact with the 3DGS model through natural language. The automatic generation of the semantically enriched 3DGS model directly from the input images or videos is the chasing of the previous studies, as well as this paper. This necessitates an accurate semantic understanding of the input images, followed by the segmentation of target objects. Previous studies, such as LangSplat (Qin, 2024), employ pre-trained zero-shot models like Contrastive Language-Image Pre-training (CLIP) for automatic scene understanding. Although these zero-shot models enhance the implementation convenience, their capabilities are constrained by the pre-training dataset. In the context of historical buildings, the target objects for segmentation include components such as flying buttresses, arched windows, and spires. These semantics require strongly

domain-specific knowledge and may not be presented in the pre-training dataset. Consequently, directly applying pre-trained zero-shot models to historical buildings may result in incorrect semantic segmentation of input images, leading to erroneous information storage in the 3DGS model. Conversely, the CNN-based models can effectively address domain-specific segmentation problems, but they demand significant manual effort for label annotations and involve a time-consuming training process. To ensure accurate information storage of historical buildings in the 3DGS model while balancing manual effort, it is crucial to develop a component semantic segmentation method using a few-shot dataset.

Multimodal Large Language Models (MLLMs) are renowned for their reasoning and in-context learning abilities (Zhang, 2024). These models can learn from a few input samples, then achieve impressive recognition performance for similar objects. However, they face challenges with weak localization ability (Yin, 2023), which can result in inferior segmentation results.

To fulfill the aforementioned requirements and address the existing problems, this paper proposes an optimized language-embedded 3DGS framework for realistic modeling and accurate information storage of historical buildings. The framework consists of three steps: (1) data preparation of on-site images and relevant text; (2) component segmentation by the integration of SAM and MLLM; (3) scene reconstruction using the language-embedded 3DGS. Our method was validated on the Shishi Sacred Heart Cathedral in Guangzhou, China. Results show that our method can achieve 95.6% accuracy of component segmentation using few manual-annotated samples. The generated 3DGS model accurately represents a realistic and informative scene, which can be further integrated with LLM assistants for open-vocabulary and vague searches. This innovative digital model of historical buildings enhances user interaction and is valuable for cultural protection.

2 3D Gaussian Splatting Technique

3DGS can be conceptualized as a point cloud, where each point is associated with a Gaussian ellipsoid. From user-specified viewpoints, it renders and displays the information encoded in the Gaussians onto images. The basic RGB-based 3DGS model includes the following parameters for each 3D Gaussian: position x, covariance matrix Σ (defining its shape), opacity α, and spherical harmonics (SH) coefficients c (defining its color). The expression of 3D Gaussian characterized by a mean μ is shown in Equation 1.

$$G(x) = e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{1}$$

3D Gaussians are the ellipsoids in 3D space. They are then projected into the 2D image space as ellipses for rasterization-based rendering, which is called splatting. Given the viewing transformation W and Jacobian of the affine approximation of the projective transformation J, the splatting is to project the 3D covariance matrix Σ into 2D covariance matrix Σ' using Equation 2.

$$\Sigma' = JW\Sigma W^T J^T \tag{2}$$

The 2D ellipses are then sorted by depths of Gaussians and cumulated to render images. Given the position of a pixel $v \in \{1, \dots, H\} \times \{1, \dots, W\}$ and a sorted list of Gaussians N, the final color $c(v)$ of this pixel is calculated by the α-blending as shown in Equation 3.

$$c(v) = \sum_{i \in N} c_i a_i' \prod_{j=1}^{i-1} (1 - a_j') \tag{3}$$

where c_i is learned color in 3D Gaussian ellipsoids, and final opacity α_i' is the multiplication result of learned opacity α_i in 3D Gaussian ellipsoids and 2D covariance.

In the training process of the 3DGS, input images are used to optimize parameters within each 3D Gaussian and to adjust the number of Gaussians. These input images act as ground truth, compared with rendered images in each iteration to calculate the loss for optimization. For further details, refer to (Kerbl, 2023).

3 Methodology

Our proposed framework is illustrated in Figure 1. Initially, on-site images and relevant textual materials are collected as the overall input. Then, the SAM and MLLM are integrated for the semantic segmentation of target components in the images, requiring only one annotation per component. Subsequently, the semantic segmentation masks are utilized to generate 3DGS scenes, embedding language features to facilitate open-vocabulary and vague searches. Two datasets, front view and side view, are collected for Shishi Sacred Heart Cathedral, as different perspectives may contain different components. The operational environment is the computer equipped with an NVIDIA A40 GPU and an Intel Xeon Platinum 8358P CPU.

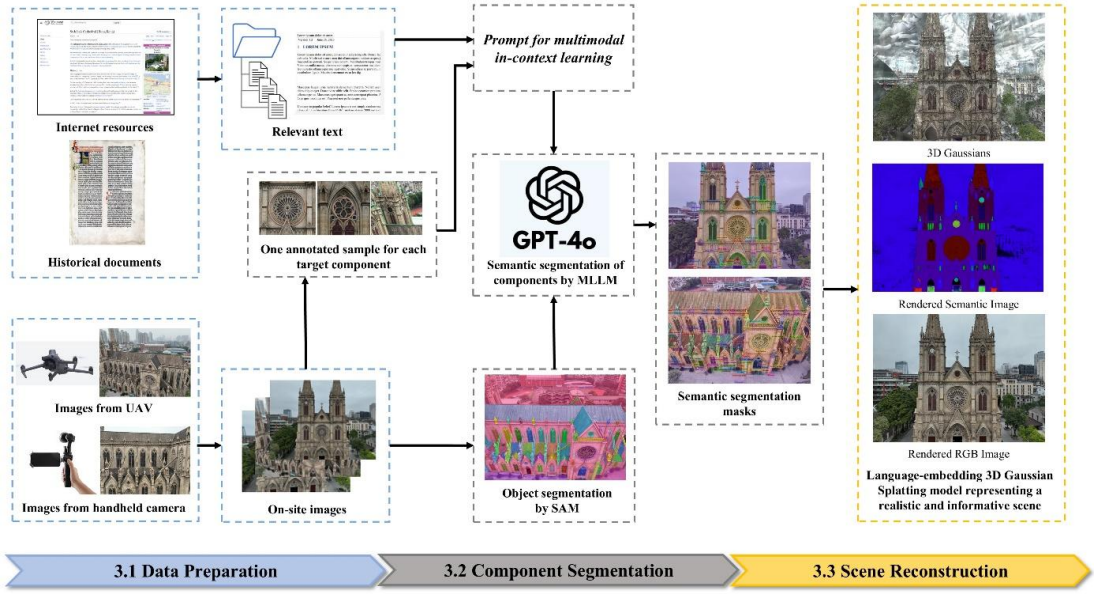


Figure 1. Overall framework of the optimized language-embedded 3DGS

3.1 Data Preparation of On-site Images and Relevant Text

In the first step, on-site images are captured for 3DGS generation, and textual materials related to component characteristics are prepared for subsequent MLLM-assisted semantic segmentation.

To ensure high-quality 3DGS reconstruction, multi-angle images must be captured to fully cover the target scene, providing a faithful viewing experience from various perspectives of the 3DGS model. As illustrated in Figure 2, images of the Cathedral's front façade were taken from multiple angles and distances. A total of 1,371 images at a resolution of 2K pixels were captured for the Shishi Sacred Heart Cathedral using UAV and handheld cameras.

Figure 2. Multi-angle images captured for the frontier façade

The exploration of textual materials related to the Cathedral was conducted through internet sources and historical documents. The relevant texts primarily include descriptions of the Cathedral's most distinctive components, which serve as a part of input for the MLLM to enhance component understanding and segmentation. Examples of these textual descriptions include:

- ***Rose window:*** *A large, circular window with intricate stone tracery radiating from the center, resembling the petals of a rose.*
- ***Flying buttress:*** *A semi-circular or arched exterior support that projects from the upper portion of a church wall over the roof of the church's aisle or chapel below.*
- ***Spire:*** *A tall, slender, and pointed structure rising from the top of a tower or the roof of a cathedral, typically made of stone or wood, adding height and verticality to the overall design.*

3.3 Scene Reconstruction using the Language-embedded 3DGS

Once accurate semantic segmentation masks of the components are generated, they are employed in developing the language-embedded 3DGS model. Compared with the original 3DGS, this model extends a semantic attribute to store language features as a vector. These features enable natural language search through cosine similarity calculations.

The CLIP text encoding module (Radford, 2021) E_{text} is utilized to generate the language feature L of pixel v with the semantic label S, as shown in Equation 4.

$$L(v) = E_{text}\big(S(v)\big) \tag{4}$$

However, these language features are 512-dimensional vectors, which can easily cause "out of memory" issues if directly added to Gaussians. Inspired by (Qin, 2024), a scene-wise autoencoder is developed to reduce the dimensionality by mapping the 512-dimensional vector to a 3-dimensional scene-specific vector. These 3-dimensional latent features are then used to supervise the learning of the 3D language field and facilitate rendering to visualize semantic embedding conditions in 3D scenes. The rendering of the latent features f on pixel v also employs the α-blending, similar to color rendering, as shown in Equation 5.

$$f(v) = \sum_{i \in N} f_i a_i' \prod_{j=1}^{i-1} \big(1 - a_j'\big) \tag{5}$$

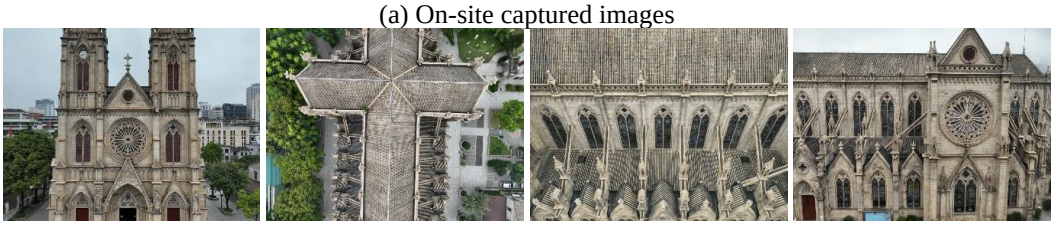
where f_i is learned latent feature in 3D Gaussian ellipsoids.

The training process of the language-embedded 3DGS involves two steps. First, a colorful 3DGS model is developed based on the original 3DGS pipeline. Following this, the number and learned attributes of Gaussians are fixed in the colorful 3DGS model, and latent feature maps are input to enable the model to learn component semantics through the same rasterization process. In this experiment, the first step involves 100,000 training epochs, and the second step involves 30,000 epochs.

Figure 6 demonstrates the realistic scene rendering capabilities of 3DGS technology. Notably, the scenes in 3DGS closely resemble real-life environments. Additionally, users can freely navigate within 3DGS, making it highly suitable for applications with significant user interaction.

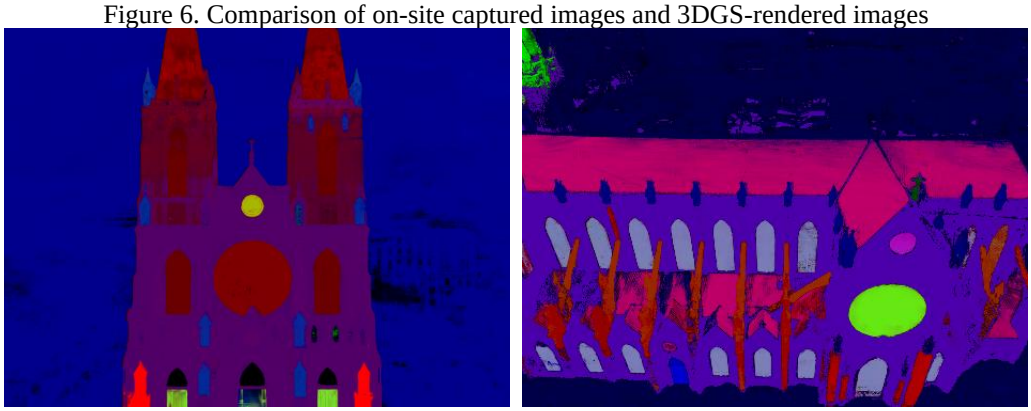
Figure 7 illustrates the rendered semantic masks, also known as the 3-dimensional latent features. The edges of different objects are clear and distinct, indicating that the semantic information has been accurately learned and stored in the 3DGS model. However, the lower roof in the side view appears somewhat blurred, likely due to occlusion by the flying buttress, preventing complete segmentation. Additional views from different angles are needed to clearly capture the roof and enhance the scene information through improved segmentation.

(a) On-site captured images



(b) Colorful images rendered by 3DGS

Figure 6. Comparison of on-site captured images and 3DGS-rendered images



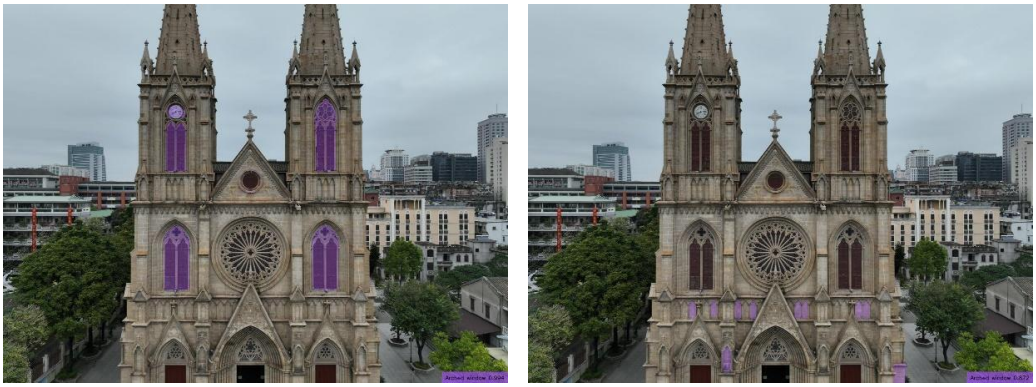
(a) front view (b) side view

Figure 7. 3DGS-rendered semantic masks

4 Validation

The validation analyses focus on two aspects to demonstrate the superiority of the proposed pipeline. First, our optimized language-embedding 3DGS is compared with the language-embedding 3DGS using the pre-trained model for scene understanding (LangSplat). Second, a quantitative analysis is conducted on the accuracy of MLLM-assisted semantic segmentation of components.

The comparison between our method and LangSplat focuses on the ability to effectively handle domain-specific knowledge and accurately store complex component semantics. Since both methods store language features in the 3DGS models, a cosine similarity search is conducted on every pixel using the target component name. As shown in Figure 8, the search for the specific terms "Arched window" and "Spire" was conducted on two 3DGS models. Our method successfully highlights the appropriate locations on the images, indicating that the semantics stored in the 3DGS are both accurate and meet the specialized requirements of historical buildings. In contrast, LangSplat is unable to identify these terms because it relies on CLIP for scene understanding, which is constrained by pre-trained data and struggles to expand its knowledge domain. Consequently, it cannot accurately identify complex objects or concepts in specific fields. This result demonstrates that our proposed optimized language-embedding 3DGS effectively addresses automatic 3D scene reconstruction with semantic complexity.



(a) Arched window

(b) Spire

Figure 8. Comparison of our method (left) with LangSplat (right)

Our semantic segmentation method employs MLLM to recognize SAM-segmented objects, ensuring that every object in the images is labeled. The accuracy of these predicted labels is crucial. Therefore, validation is performed to check the correctness of semantic labels for each component.

Both front and side view images were validated. In the front view sample shown in Figure 9(a), 41 out of 42 components were accurately recognized. In the side view sample shown in Figure 9(b), 60 out of 66 components were accurately identified. The overall recognition accuracy for all components is 95.6%, demonstrating that our SAM-MLLM integrated method achieves high-quality semantic segmentation with only a few annotated samples as input. The highly accurate semantic segmentation with less manual effort is also valuable for other scenes understanding tasks.

(a) Front view, (accuracy = 41/42 = 97.6%)

(b) Side view (accuracy = 60/66 =90.9%)

Figure 9. Quantitative analysis of the accuracy of MLLM-assisted semantic segmentation. The left shows the generated segmented masks, and the right shows the ground truth.

5 Discussion

The generated language-embedding 3DGS model can be further integrated with an LLM-based chatbot interface to facilitate user-friendly searching, as shown in Figure 10 and 11. This interface, developed via Autogen Studio, utilizes the LLM heritage assistant powered by GPT-4o, configured with a specific system message for instruction. With component semantics embedded as language features, searching operations are performed using cosine similarity calculations, which are implemented as a backend function in the LLM heritage assistant.

In Figure 10, open-vocabulary searches were initially conducted for testing. User input prompts included specific terms like "spire" and "pinnacle". Following the search, the corresponding components were successfully selected and highlighted. However, these terms represent specialized knowledge, which may be challenging for non-expert users to articulate accurately. Therefore, in Figure 11, a vague search is performed to test the generalizability of our method. Users employed descriptive phrases like "big round window" instead of standard terms like "rose window." The results demonstrate that the correct components can still be identified and highlighted, indicating that our method is highly user-friendly. Additionally, the LLM heritage assistant, capable of accessing historical and contextual information from the web, allows for the effective integration of online

knowledge with reconstructed 3D models. This integration holds significant value for the education and cultural dissemination of historical buildings.

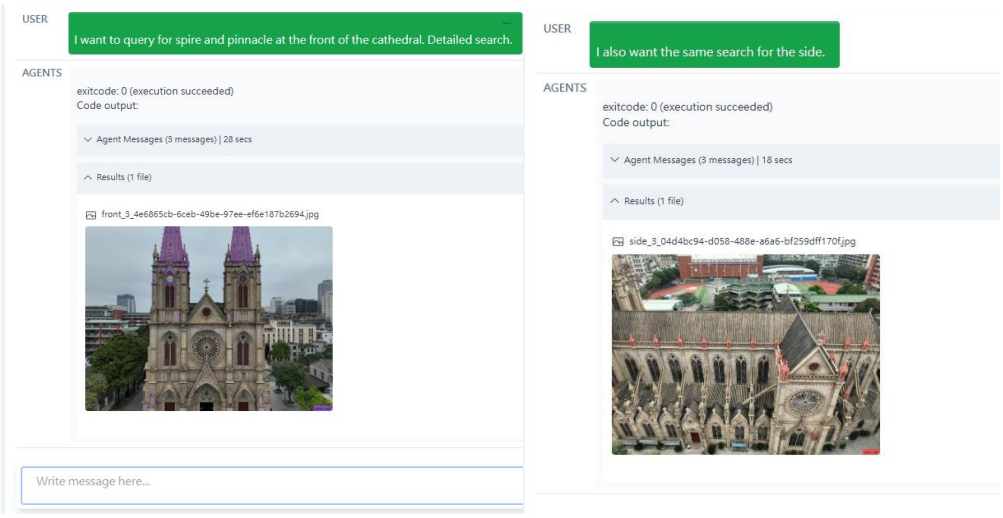


Figure 10. Results of open-vocabulary searches

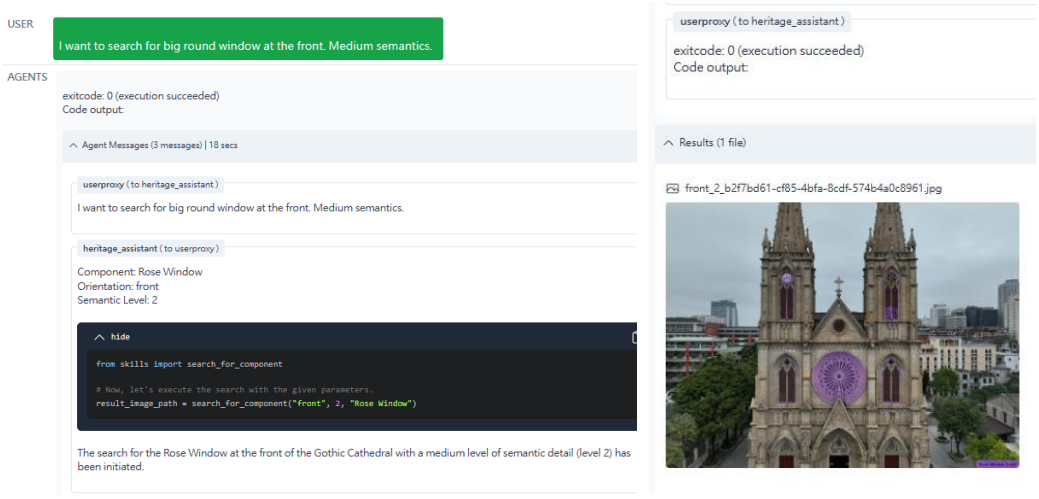


Figure 11. Results of vague search

6 Conclusion

This paper proposes an optimized language-embedded 3DGS framework for the automatic and accurate modeling of realistic and informative 3D scenes of historical buildings. The framework includes: (1) preparation of on-site images and relevant textual material; (2) semantic segmentation of historical building components using SAM-MLLM integrated method; (3) development of a language-embedded 3DGS model for scene representation. Our semantic segmentation method achieves 95.6% accuracy for component segmentation using only one annotated sample per component category. The optimized language-embedded 3DGS method outperforms the previous

model in accurate domain-specific semantics recognition and storage. The 3DGS-based digital representation of historical buildings also facilitates user interaction through an LLM-based chatbot assistant for open-vocabulary and vague searches. This scene-realistic, semantic-enriching, convenient-navigating and easy-interacting 3DGS model brings significant value for the heritage preservation, public education and cultural dissemination.

Future work will extend our study to include the reconstruction of indoor scenes of historical buildings, aiming to better document and disseminate the intricate interior decorations. Currently, the MLLM used is GPT-4o, which requires the API. Future efforts will also focus on developing the open-source MLLM-assisted method, which will be beneficial for private deployment.

7 Acknowledge

The authors would like to acknowledge the support by RGC Theme-based Research Schemes (2023/24 T22-606/23-R) and (2024/25 T22-607/24-N). The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Cheng, J. C., Zhang, J., Kwok, H. H., & Tong, J. C. (2024). Thermal performance improvement for residential heritage building preservation based on digital twins. *Journal of Building Engineering*, 82, 108283.

Croce, V., Caroti, G., De Luca, L., Jacquot, K., Piemonte, A., & Véron, P. (2021). From the semantic point cloud to heritage-building information modeling: A semiautomatic approach exploiting machine learning. *Remote Sensing*, 13(3), 461.

Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4), 139-1.

Pritchard, D., Sperner, J., Hoepner, S., & Tenschert, R. (2017). Terrestrial laser scanning for heritage conservation: The Cologne Cathedral documentation project. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4, 213-220.

Qin, M., Li, W., Zhou, J., Wang, H., & Pfister, H. (2024). Langsplat: 3d language gaussian splatting. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 20051-20060).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. *In International conference on machine learning* (pp. 8748-8763). PMLR.

Shi, J. C., Wang, M., Duan, H. B., & Guan, S. H. (2024). Language embedded 3d gaussians for open-vocabulary scene understanding. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5333-5343).

Ursini, A., Grazzini, A., Matrone, F., & Zerbinatti, M. (2022). From scan-to-BIM to a structural finite elements model of built heritage for dynamic simulation. *Automation in Construction*, 142, 104518.

Yang, X., Grussenmeyer, P., Koehl, M., Macher, H., Murtiyoso, A., & Landes, T. (2020). Review of built heritage modelling: Integration of HBIM and other information techniques. *Journal of Cultural Heritage*, 46, 350-360.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2023). A survey on multimodal large language models. *arXiv preprint* arXiv:2306.13549.

Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., & Yu, D. (2024). Mm-llms: Recent advances in multimodal large language models. *arXiv preprint* arXiv:2401.13601.