EasyChair Preprint

№ 15736

Enhancing Local Ecological Adaptation in Multispecies Eco-Building Design Through Multimodal System

Daxu Wei and Christiane Herr

January 20, 2025

Enhancing Local Ecological Adaptation in Multispecies Eco-building Design through Multimodal System

First Author[1][0000-1111-2222-3333] and Second Author[2][1111-2222-3333-4444]

[1] Princeton University, Princeton NJ 08544, USA
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
lncs@springer.com

Abstract. Multispecies ecological architectural design aims to create buildings that coexist harmoniously with the surrounding natural environment, playing an increasingly important role in enhancing the biodiversity and ecological resilience of urban environments. However, ecological architectural design requires expertise to tailor designs to local ecosystems, especially in terms of species selection and growth environment matching. This study introduces a multimodal system that enables non-professional designers to create ecologically adaptive, site-specific architectural designs. The system combines ChatGPT and diffusion models to analyze and synthesize visual and textual data, embedding local ecological characteristics into the design process. It supports the collection of local species data, which is then processed by GPT-4V (Vision) to generate detailed material descriptions, optimized through expert feedback. The minimally trained ChatGPT model, supported by the Segment Anything Model (SAM), predicts the ecological suitability of species integration across different areas, segmenting images to identify regions conducive to multispecies growth. Identified regions are further used to generate ecological designs via latent space diffusion, with a low-rank adaptation (LoRA) model, trained on local species data, enhancing the accuracy of ecological simulations. ControlNet and advanced prompt engineering are utilized to optimize the final design outcomes. This multimodal system integrates AI technologies such as transformer models and diffusion models, distinguishing itself from previous multimodal applications that mainly focused on style guidance and aesthetic generation. In contrast, this approach emphasizes improving building performance, offering a new method for incorporating ecological principles into architectural practice, and providing a practical tool for developing urban environments with biodiversity and resilience.

Keywords: Eco-building, Multimodal Large Language Model, Multispecies Design, Ecological Design

1. Introduction

In recent years, ecological design strategies have received increasing attention in urban development, particularly in the field of ecological architecture (Selvan et al., 2023). Approaches such as green roofs, ecological walls, and vertical gardens have been widely applied as means to improve urban environments and address ecological issues (Bustami, 2018; Kader et al., 2022). These vegetation-covered building structures

maximize the use of limited space in urban environments, helping to improve air quality, regulate temperature, reduce energy consumption and carbon emissions, manage rainwater, and provide opportunities for urban agriculture, thus creating urban oases (Yan et al., 2024; Wang et al., 2023; Zheng et al., 2023; Yang et al., 2006). At the same time, ecological design can promote biodiversity in densely populated areas, where a variety of plant species can provide habitats for insects, birds, and other small organisms (Radić et al., 2019). Despite the potential of ecological architecture, designing such buildings often requires a substantial amount of specialized ecological knowledge, particularly when selecting appropriate plant species for specific locations and environmental conditions. This challenge underscores the need for systems that can bridge this gap, enabling designers to create ecological designs that are not only environmentally feasible but also fine-tuned to local conditions.

Meanwhile, with the development of artificial intelligence and big modeling technologies, the multimodal large language models (MLLMs) have demonstrated unique advantages in cross-modal knowledge integration and understanding, multimodal semantic reasoning, and content generation (Dhariwal & Nichol, 2021). Previous studies have typically focused on using MLLMs to generate heuristic images (Ma & Zheng, 2023; Veloso, 2024), guiding aesthetic styles. In contrast, this study explores how MLLMs can integrate diverse data modalities in ecological design, aiming to enhance the site-specific feasibility and ecological viability of architectural designs from a performance-driven perspective.

As the field of ecological architecture continues to evolve, research into the potential of multispecies design to enhance ecological feasibility has expanded, alongside progress in the application of MLLMs to improve architectural design methods. In the domain of ecological feasibility and multispecies design, Briscoe (2018) proposed a framework that combines Building Information Modeling (BIM) with ecological design principles to optimize the ecological benefits of living walls, particularly in hot and arid climates. Zhang et al. (2023) developed the Urban Agriculture Ecological Laboratory (ELUA), which aims to create a continuously evolving ecologically feasible environment through continuous monitoring, data-driven decision-making, and collaboration between humans and AI. Weisser et al. (2023) proposed a multispecies symbiotic space design technique called "ecolope," aimed at replacing traditional building envelopes to enhance urban biodiversity and improve human-nature interaction. In the field of MLLMs, Paananen et al. (2023) examined the role of text-to-image tools like Midjourney, Stable Diffusion, and DALL-E in enhancing creativity during the early stages of architectural design, better supporting architects' imaginative processes. Shi and Hua (2023) developed a method that uses fine-tuned latent text-to-image diffusion models to generate images and three-dimensional scenes of Chinese gardens in the Ming dynasty style based on textual descriptions, highlighting its potential in cultural heritage restoration. Kim, Johanes, and Huang (2023) proposed a fine-tuning framework for Stable Diffusion, generating images more aligned with architectural language using a formal architectural vocabulary dataset. Doumpioti and Huang (2023) introduced a framework combining text-to-image models with environmental design computation, utilizing AI-generated images and simulations to enhance the environmental responsiveness of architectural forms. Guida (2023)

demonstrated the potential of multimodal machine learning models like Stable Diffusion and DALL-E 2 in integrating text-to-image and three-dimensional form generation into the architectural design process, emphasizing the role of language development in architecture and the potential for intuitive user interfaces to promote more effective human-machine collaboration.

Although these studies represent advances in supporting ecological architecture and the application of MLLMs in architectural design, the use of MLLMs in architecture still remains primarily at the conceptual design guidance level, and has yet to be widely integrated into actual building performance and applications, particularly in aligning architectural and ecological goals. Moreover, existing research in ecological architecture has not adequately addressed the importance of local feasibility nor conducted compatibility analyses based on multimodal data. To address these issues, this study employs a multimodal system based on GPT and diffusion models to assist non-expert designers in collecting and providing ecologically sound, site-adapted plant recommendations for ecological building designs. This approach not only enhances the ecological feasibility and sustainability of architectural designs but also lowers the barriers to ecological design, promoting the widespread application of ecological optimization in urban micro-environment design.

2. Research Methodology

This study adopts an intelligent design system based on multimodal large language models, aiming to generate environmentally friendly ecological building designs through an automated process. The system first uses a vision-language model to process multimodal data of local plants and growing substrates to assess ecological suitability. Then, a semantic segmentation model and diffusion model generate ecological design solutions that meet environmental requirements. (see Fig 1)

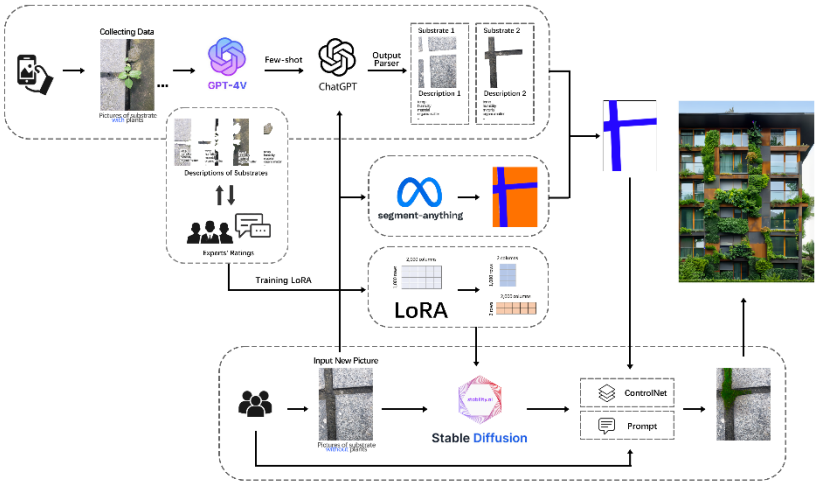

Fig. 1. Multispecies Eco-building Design System

2.1 Data Collection and Processing

The study uses Shenzhen, China, as a case study. Shenzhen has a subtropical monsoon climate with mild temperatures, abundant sunlight, and ample rainfall, creating favorable ecological conditions for plant growth and biodiversity. We collected 250 images of self-seeding plants and their growing substrates from five districts in Shenzhen. To ensure a comprehensive analysis of ecological suitability, the dataset includes images of plants and substrates under various building materials such as concrete, brick, rammed earth, and metal. The plant species primarily cover ten common species in Shenzhen: Zoysia matrella, Axonopus compressus, Wedelia trilobata, Cynodon dactylon, Bidens pilosa, Alocasia odora, Digitaria radicosa, Oxalis corniculata, Ruellia simplex, and Hedyotis corymbosa (Liu et al., 2023). These images were standardized, including adjustments for saturation and brightness, and cropped to 512x512 PNG format to ensure consistency and quality. The processed ecological images are used in two modules. In the multimodal recognition module, the system learns the relationship between substrate types and plant growth conditions, laying the foundation for subsequent design generation. In the multimodal generation module, local plant image data is used to train a LoRA model to fine-tune the system's output.

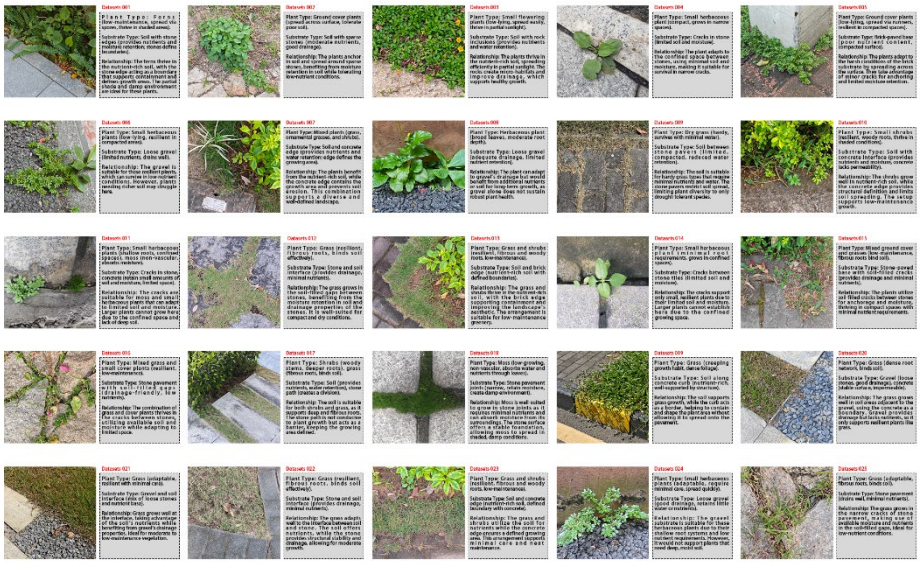

Fig.2. Data Collection and Processing

2.2 Ecological Data Recognition and Few-Shot Learning

GPT-4 Vision is a large-scale language model with powerful text-visual reasoning, multimodal understanding, and prediction capabilities. The system uses a 250-image ecological dataset for few-shot learning, enabling the model to recognize and understand the ecological adaptability of plants and their substrates. This process allows GPT-4 Vision to generate textual descriptions of plant species, substrate types, growing

environments, and the ecological interactions between plants and substrates. Through few-shot learning, the GPT-4 Vision model can quickly understand and generate ecological descriptions with high accuracy, even with a limited amount of training data.

2.3 Ecological Substrate Recognition and Segmentation

After evaluating the ecological adaptability of the dataset, the system uses SAM (Segment Anything Model) to perform region segmentation on the input substrate images (without plants). SAM automatically identifies and labels different types of substrate regions, generating clear semantic segmentation maps (Kirillov et al., 2023). These substrate images and segmentation maps are then input into the GPT model, which has been trained with few-shot learning, for ecological adaptability analysis. The model accurately predicts and differentiates which areas are suitable for plant growth. Areas unsuitable for plant growth are output as masks. This process applies the understanding results of GPT-4 Vision and uses the mask to control the ecological design generated by Stable Diffusion (see Fig 3).

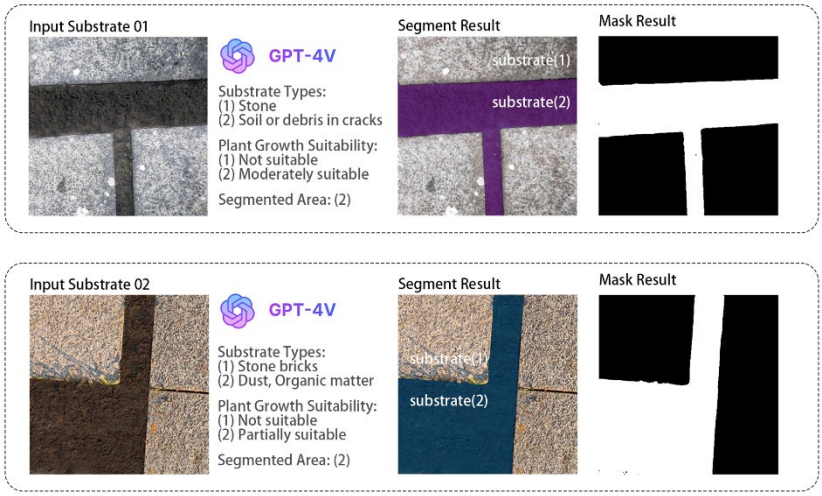

Fig.3. Ecological Substrate Recognition and Segmentation

2.4 Model Fine-Tuning and Inference Optimization

The system uses the Stable Diffusion model to output design results. Stable Diffusion incorporates cross-attention layers in its architecture, enabling it to handle various conditional inputs such as text or bounding boxes, and to achieve high-resolution synthesis through convolutional methods (Rombach et al., 2022). In addition to supporting text-to-image generation, it also allows fine-tuning and control of the output results using technologies such as LoRA, masks, and ControlNet.

LoRA is a technique for fine-tuning large pre-trained models by adjusting a small subset of model parameters, achieving efficient fine-tuning with low resource requirements (Hu et al., 2021). The system fine-tunes the Stable Diffusion model using LoRA, updating only a small portion of the dataset and a few parameters. This enables

the model to focus on plant types, forms, and their relationship with substrates, thus improving the accuracy of generating plant growth effects. The system uses RealisticVisionV60B1.safetensors as the base model and fine-tunes it via the diffusion pipeline (von Platen et al., 2024). The training process specifically uses the previously processed 500 512x512 plant images paired with simplified text labels generated by GPT-4 Vision. The training employed the AdamW 8-bit optimizer, with a network dimension of 32, a learning rate of 1e-4 (UNet learning rate of 1e-4, text encoder learning rate of 5e-5). The training was conducted at a resolution of 512x512 for 5 epochs, using mixed precision (fp16) and enabling Xformers, with learning rate control through constant_with_warmup. The training was performed on an RTX 4090 GPU, ultimately resulting in a LoRA model suitable for ecological architectural design.

2.5 Generation Control and Ecological Design Generation

When generating plant growth images using Stable Diffusion, the system combines ControlNet and the previously generated mask images to control the output results. ControlNet is a neural network architecture that adds spatial condition control to large pre-trained text-to-image diffusion models (Lvmin Zhang et al., 2023). The system uses the control_segment-fp16.safetensor and control_inpainting-fp16.safetensors models to precisely control semantic and image editing during the generation process. By loading the mask images generated by GPT-4 Vision and SAM, the system accurately controls the plant growth areas, ensuring that the plant growth in different substrate regions meets ecological requirements, thereby avoiding unreasonable design outcomes.

Furthermore, prompt engineering plays a key role in guiding the Stable Diffusion model to generate design results. The system adjusts and optimizes prompt semantics and structure based on the text data generated by GPT-4 Vision and the labeled training set used during LoRA training. Ten different prompts, corresponding to ten plant species, are applied to effectively trigger LoRA and control the growth effects of different plant species.

The base images that need plant generation, along with the inpainting masks previously obtained, are input into the Stable Diffusion model, and the fine-tuned LoRA model is loaded. The sampling steps are set to 30, using the DPM++ SDE Karras sampler, with a CFG scale of 7.5 and denoising strength of 0.7. The inference process is accelerated with Xformers, optimized with fp16 mixed precision, and the output resolution is set to 512x512. Ten sets of ecological design images are generated, with each set containing five images.

3. Results and Evaluation

The results show that by combining ControlNet and inpainting masks, the generated plants are reasonably distributed within the designated growth areas, and their growth patterns highly align with the substrate environment (see Fig 4). Furthermore, each prompt effectively guided the generation of different plant species, influencing the plant types, morphology, and spatial arrangement. The generated images are rich in detail,

and the natural integration between plants and substrates is impressive. These results reflect the powerful learning ability of the fine-tuned LoRA model and its adaptability in generating diverse plant species with ecological relevance. The system effectively demonstrates the potential of this model for scalable and precise ecological design applications.

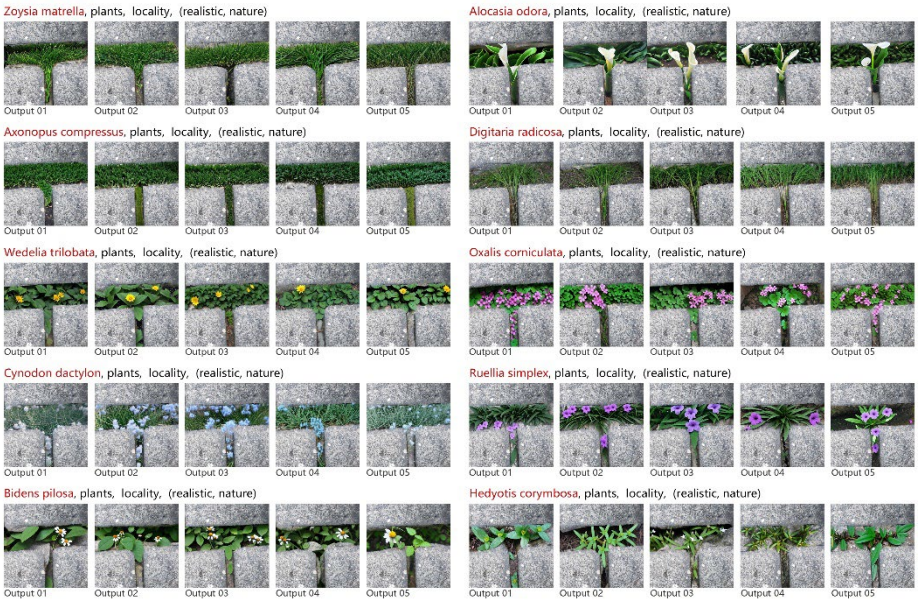

Fig.4. Results generated with different prompts

3.1 Evaluation of Few-shot Correction

To improve the system's accuracy in identifying the ecological adaptability of plant growth substrates, an expert evaluation mechanism was implemented to rigorously verify and correct the descriptions generated by GPT-4 Vision. We invited 10 experts from the fields of ecology, architecture, and landscape architecture to conduct cross-evaluation and provide feedback on the ecological image descriptions from the 500 dataset entries.

The evaluation results showed that out of the 500 descriptions, 447 were accurate. Among the 53 errors, 19 were related to substrate material identification, and 34 were related to plant species identification, with an overall accuracy rate of 89.4%. Based on the experts' feedback, additional few-shot training was conducted, focusing on correcting errors in these two areas, followed by iterative dialogue with the GPT-4 Vision model for further corrections. After this correction process, the model demonstrated higher precision in subsequent tests, achieving an accuracy rate of 97.6%. This improvement significantly enhanced the system's stability and reliability in handling multimodal ecological data.

3.2 Evaluation of Generated Results

To validate the effectiveness of the system, the generated design schemes were evaluated on multiple dimensions, including generation quality, ecological adaptability, species accuracy, and locality. An expert panel conducted an interdisciplinary evaluation of the generated ecological design images. Each expert scored the design schemes based on the following criteria (1 to 10 scale):

- Species Accuracy: The generated results effectively learned and produced accurate and reasonable plant species types.
- Generation Quality: The image details were clear, and the fusion between plants and substrates appeared natural.
- Ecological Adaptability: The selected plant species were suitable for the substrates and growth environments, meeting ecological needs.
- Locality: The generated plant species were native and suitable for local growing conditions.

The evaluation results show that the system performed well in terms of generation quality, locality, and ecological adaptability, with average scores of 9.22, 8.47, and 9.27, respectively. These results suggest that the system can identify and generate real and diverse plant types that adapt to their growth environments. However, the species accuracy score was relatively low, with an average of only 7.38, especially for species with unique morphologies, such as Ruellia simplex, Oxalis corniculata, and Alocasia odora. This indicates that while the system is capable of generating various plant types, improvements are needed in learning and generating specific plant species, as locality is often closely tied to species specificity.

4. Discussion

This study integrates GPT-4 Vision and the multimodal large language model (MLLM) based on Stable Diffusion with data on plants and their growth environments to develop an automated ecological building design system. The research explores the potential of MLLM in multispecies ecological architecture design. Unlike previous studies that used MLLM to guide design aesthetics (Ma & Zheng, 2023; Veloso, 2024) or human-computer interaction in 3D design processes (Guida, 2023), this research leverages MLLM's cross-modal reasoning and content generation capabilities to explore its applicability in performance-based design tasks. Additionally, this study shares similarities with Briscoe's (2018) work on integrating ecological design principles into building systems and Weisser et al. (2023) on multispecies envelope design. However, these frameworks rely on traditional design methods, whereas this research demonstrates the potential of text-to-image models in incorporating ecological data into ecological building design, bridging the gap between automatic image generation and ecological performance optimization.

While the system performs excellently in generating realistic and ecologically adaptive plant types, its ability to accurately identify and generate specific plant species

remains limited. This limitation may stem from constraints in the original training datasets of the MLLM, which lacked detailed plant species information, presenting inherent challenges in further model development. Despite fine-tuning with LoRA, prompt engineering, and few-shot learning, these adjustments are still insufficient to fundamentally address this issue. Although developing large-scale models specifically for ecology or architecture remains challenging, future work will explore using higher-quality and more domain-relevant multimodal pre-training data, instruction-tuning datasets, and preference-based data to improve system performance.

5. Conclusion

This study demonstrates the potential of MLLM in ecological building design, providing a practical framework for creating ecologically adaptive solutions. It pioneers the application of MLLM (including GPT-4 Vision and Stable Diffusion) in performance-based ecological building design. Unlike previous research focusing on visual inspiration, this study integrates image-text data, plant adaptability prediction, and substrate analysis to create an automated and scalable system for generating designs suited to local conditions. Through techniques such as LoRA fine-tuning, ControlNet, and prompt engineering, the system achieves high precision and adaptability, reducing the threshold for non-expert involvement. Expert validation and real-world testing further demonstrate its potential to democratize ecological expertise and drive sustainable urban development. The results indicate that the system performs well in matching plant species with substrate conditions, but there remains a gap in achieving precise locality accuracy for plant species. Future improvements should focus on enhancing species identification and generation capabilities and developing a comprehensive local plant database to ensure more precise alignment between plant selection and regional ecosystems.

References

Besir, A. B., & Cuce, E. (2018). Green roofs and facades: A comprehensive review. Renewable and Sustainable Energy Reviews, 82, 915 – 939. https://doi.org/10.1016/j.rser.2017.09.106

Bustami, R. A., Belusko, M., Ward, J., & Beecham, S. (2018). Vertical greenery systems: A systematic review of research trends. Building and Environment, 146, 226 – 237. https://doi.org/10.1016/j.buildenv.2018.09.045

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. arXiv. https://doi.org/10.48550/arXiv.2105.05233

Doumpioti, C., & Huang, J. (2023). Text to image to data. In eCAADe Conference (pp. 541–548). https://doi.org/10.52842/conf.ecaade.2023.2.541

Guida, G. (2023). Multimodal architecture: Applications of language in a machine learning-aided design process. In CAADRIA Conference 2023 (pp. 561 – 570). https://doi.org/10.52842/conf.caadria.2023.2.561

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv. http://arxiv.org/abs/2106.09685

Kader, S., Chadalavada, S., Jaufer, L., Spalevic, V., & Dudic, B. (2022). Green roof substrates — A literature review. Frontiers in Built Environment, 8, 1019362. https://doi.org/10.3389/fbuil.2022.1019362

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything. arXiv. https://doi.org/10.48550/arXiv.2304.02643

Ma, H., & Zheng, H. (2024). Text semantics to image generation: A method of building facades design based on stable diffusion model. In Computational design and robotic fabrication 2023 (pp. 45–55). Springer. https://doi.org/10.1007/978-981-99-8405-3_3

Paananen, V., Oppenlaender, J., & Visuri, A. (2023). Using text-to-image generation for architectural design ideation. International Journal of Architectural Computing. https://doi.org/10.1177/14780771231222783

Radić, M., Dodig, M. B., & Auer, T. (2019). Green facades and living walls: A review establishing the classification of construction types and mapping the benefits. Sustainability, 11(16), 4579. https://doi.org/10.3390/su11164579

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. arXiv. http://arxiv.org/abs/2112.10752

Selvan, S. U., Saroglou, S. T., Joschinski, J., Calbi, M., Vogler, V., Barath, S., & Grobman, Y. J. (2023). Toward multi-species building envelopes: A critical literature review of multi-criteria decision-making for design support. Building and Environment, 231, 110006. https://doi.org/10.1016/j.buildenv.2023.110006

Shi, J., & Hua, H. (2024). Space narrative: Generating images and 3D scenes of Chinese garden from text using deep learning. In Creativity in the age of digital reproduction (Vol. 343, pp. 236–243). Springer Nature. https://doi.org/10.1007/978-981-97-0621-1_28

Veloso, P. (2024). Forming the new building envelope: A pedagogical study in generative design with precedents and multimodal large language models. International Journal of Architectural Computing. https://doi.org/10.1177/14780771231222782

von Platen, P., et al. (2024). Diffusers: State-of-the-art diffusion models. Hugging Face. https://github.com/huggingface/diffusers

Wang, W., Yang, H., & Xiang, C. (2023). Green roofs and facades with integrated photovoltaic system for zero energy eco-friendly building – A review. Sustainable Energy Technologies and Assessments, 60, 103426. https://doi.org/10.1016/j.seta.2023.103426

Yan, J., Yang, P., Wang, B., Wu, S., Zhao, M., Zheng, X., Wang, Z., Zhang, Y., & Fan, C. (2024). Green Roof Systems for Rainwater and Sewage Treatment. Water, 16(15), 2090. https://doi.org/10.3390/w16152090

Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. arXiv. http://arxiv.org/abs/2302.05543

Zhang, Z., Epstein, S. L., & Breen, C. (2023). Robots in the garden: Artificial intelligence and adaptive landscapes. Wichmann Verlag. https://doi.org/10.14627/537740028

Zheng, X., Kong, F., Yin, H., Middel, A., Yang, S., Liu, H., & Huang, J. (2023). Green roof cooling and carbon mitigation benefits in a subtropical city. Urban Forestry & Urban Greening, 86, 128018. https://doi.org/10.1016/j.ufug.2023.128018