Applying LoRA to Stable Diffusion for Architectural Facade Design Generation

Jungmin Park[1], Soon Min Hong[2], Gaeun Park[3] Yeji Seo[4], Hyejin Park[5],Seungyeon Choo[6]
[1,2,3,4,5,6]School of Architecture, Kyungpook National University, Republic of Korea
[1,2,4,6]{zktm4935|soonmin_hong|dpwl2278|choo}@knu.ac.kr [3]gaeun2868@naver.com
[5]gpwls3143@gmail.com

Architectural facade design involves complex and repetitive processes such as implementation, revision, and refinement of elements including form, materials, and openings. Traditional methods like sketching and 3D modeling often lack efficiency in this iterative workflow. This study explores the application of generative AI, specifically the Stable Diffusion XL model, to enhance efficiency in facade design. By employing Low-Rank Adaptation (LoRA), the model is fine-tuned to generate high-resolution images that reflect architectural context and design intent. A custom dataset of modern commercial and residential buildings was constructed through image collection, preprocessing, and tagging. The model was then trained and evaluated using this dataset. The results demonstrate that LoRA-based fine-tuning allows for the efficient generation of diverse and contextually appropriate facade designs, offering valuable support for the early stages of architectural design.

Keywords: *Architecture massing, Facade Design, Generative Artificial Intelligence, LoRA, Stable Diffusion.*

INTRODUCTION

Research Background and Aim

Facade design plays a critical role in determining both the aesthetic and functional aspects of a building (Ching, 2023). As the exterior greatly impacts user experience and perception, exploring diverse design alternatives is essential.

However, conventional CAD and 3D modeling approaches are time-consuming and labor-intensive due to high-resolution rendering and repeated revisions, particularly in complex or large-scale projects. Recent advances in AI have introduced generative models like GANs and Diffusion models to architectural design. Diffusion models, especially Stable Diffusion (SD), enable text-based, high-quality image generation by progressively denoising input data (Ho et al., 2020). While SD facilitates rapid visualization, pre-trained models often lack structural coherence and fail to meet architectural criteria such as proportion, scale, and functional organization.

To overcome these limitations, this study proposes an optimized approach using SD, incorporating Low-Rank Adaptation (LoRA) for fine-tuning and img2img for structural consistency. ControlNet is also used to integrate Canny edge and depth map data for enhanced control over form. This method supports the rapid and systematic generation of façade alternatives in early design stages while reducing the time and resources required by traditional workflows.

Research Scope and Method

This study focuses on façade images of contemporary mid- to high-rise residential, office, and mixed-use buildings that incorporate materials such as glass, metal, reinforced concrete, and composite panels. To achieve detailed control over architectural elements, the SDXL model of Stable Diffusion is fine-tuned using the LoRA technique, with additional refinement through img2img and ControlNet.

First, the theoretical foundation and related research on Diffusion models and LoRA are reviewed to establish the study's direction. Second, a façade image dataset is used to train a LoRA checkpoint model, which is integrated into Stable Diffusion for fine-tuning. Third, architectural façade images are generated using this customized pipeline. Fourth, the results are evaluated using a fine-tuned CLIP model and GPT-4V, which assess both architectural consistency and text–image alignment.

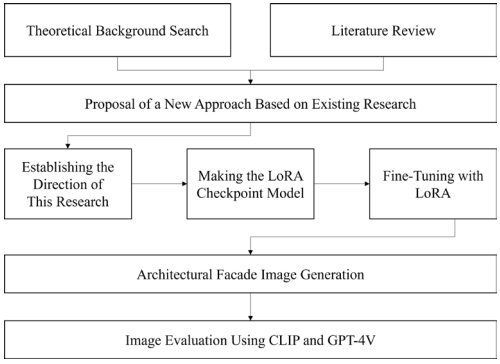


Figure 1
Research flow

THEORETICAL BACKGROUND AND REVIEW OF PREVIOUS STUDIES

Generative AI models such as GANs and Diffusion models have significantly advanced image synthesis by learning from large-scale datasets. While GANs generate realistic images through adversarial training, their limitations in training stability and output detail (Goodfellow et al., 2014; Zhao & Li, 2023) led to improvements like multi-stage structures and attention mechanisms (Zhang et al., 2017; Xu et al., 2018). However, semantic misalignment between text and image persisted. CLIP (Radford et al., 2021) addressed this by aligning image and text embeddings, enabling prompt-based generation with models like DALL·E, though high-resolution and architectural accuracy remained limited.

Diffusion models emerged as a more stable alternative. Through iterative denoising, they achieved high-fidelity outputs (Ho et al., 2020), further enhanced by Latent Diffusion Models (Rombach et al., 2022) and text-to-image methods (Saharia et al., 2022). SD, as a leading example, is widely adopted in architecture due to its ability to handle compositional complexity and resolution demands. In architectural applications, SD-based methods including img2img, LoRA, and ControlNet have been utilized. img2img preserves structure during transformation (Meng et al., 2021; Wang & Zhang, 2024), while LoRA (Hu et al., 2021) enables domain-specific fine-tuning with low computational cost (Jo & Lee, 2023; Yoo & Lee, 2023). Yet, challenges remain in detailed control, spatial coherence, and façade complexity. This highlights the need for integrated frameworks that combine the strengths of these techniques to generate high-quality, context-aware architectural imagery.

ARCHITECTURAL FAÇADE GENERATION METHOD BASED ON LORA AND CONTROLNET

Generation of Checkpoint Files through LoRA Training

To construct the dataset, high-resolution, copyright-free images of contemporary buildings using materials like metal, glass, reinforced concrete, and composite panels were collected and preprocessed. Low-quality or distorted images were excluded, and all were resized to

1024×1024 pixels with noise reduction and color correction. Although this study focuses on architectural façade design, full-building images were intentionally used during training to preserve spatial and contextual cues. Because the Stable Diffusion model interprets images holistically, cropping inputs to only include façade fragments often led to distorted proportions, missing depth cues, or stylistic inconsistencies. Therefore, preserving the building's overall form, material transitions, and contextual environment proved essential for generating coherent and architecturally plausible façades. Each image was tagged across six key architectural categories. These included typology, style, form, materials, structural systems, and additional elements such as overhangs or green roofs. Frequently used tags involved contemporary building types, modernist styles, glass curtain walls, and reinforced concrete. The most commonly applied categories were materials (63 images) and styles (52), highlighting visual and compositional priorities in the dataset.. Ambiguous or redundant tags were removed to improve clarity.

Table 1
Prompt contexts

Positive prompt	architecture, facade, modern minimalist (Architectural Style), luxury villa (Building Type), open-plan residence (Primary Structure), exposed concrete panels (Wood & Exposed Elements), floor-to-ceiling glass windows (Windows & Open Elements), wooden cladding accents (Wood & Exposed Elements), vertical metal louvers (Windows & Open Elements), high-end residential design (Architectural Style), seamless indoor-outdoor transition (Spatial Quality & Composition), contemporary aesthetic (Urban Context), precise geometric composition (Facade & Composition), sleek exterior (Architectural Style), sharp edges (Architectural Style), minimalist structure (Architectural Style), elegant contrast of materials (Architectural Style), ultra-realistic (Spatial Quality & Composition)
Negative prompt	low quality, blurry, distorted, overexposed, watermark, low-resolution, cartoonish, excessive reflections, asymmetrical structure

Figure 2
Example of an architectural facade tag

Figure 3
Loss graph

This structured tagging, illustrated in Figure 2, enabled Stable Diffusion to interpret architectural features more effectively. LoRA training was conducted using the Kohya_ss library, with iterative tuning of hyperparameters such as batch size, learning rate, optimizer, and number of epochs. Checkpoints were saved regularly, and training stability was evaluated using loss metrics.

As shown in Figure 3, the consistent decrease in loss values indicated effective model convergence. Intermediate outputs were reviewed to assess the accuracy of architectural style representation, and tags or parameters were revised when necessary to improve image quality.

Application of the LoRA Technique in Architecture

In architectural image generation using the LoRA model, the structure of the text prompt plays a critical role in determining image quality, detail, and stylistic coherence. A carefully constructed prompt leads to more realistic and accurate architectural output.

Based on prior tagging classifications, Table 1 outlines seven core categories used for prompt composition: building type, form and structure, exterior materials, windows and openings, lighting and details, architectural style, and atmosphere and context.

Although the prompts are not explicitly segmented by category, each keyword within the prompt corresponds to one or more of these seven aspects. For instance, phrases like "luxury villa" reflect building type, while "floor-to-ceiling glass windows" and "vertical metal louvers" correspond to windows and openings. Terms such as "exposed concrete" indicate exterior materials, and expressions like "contemporary aesthetic" align with architectural style and context. This implicit mapping allows the prompt to remain fluent and natural, while still covering a wide range of architectural attributes.

This study adopts both positive and negative prompts to better reflect architectural intent. Positive prompts specify desired features such as material types, façade styles, and window arrangements, encouraging the model to prioritize them. In contrast, negative prompts identify elements to avoid, including distortions, visual noise, and low-resolution features, guiding the model to suppress these characteristics.

To enhance clarity and control, essential design information—such as building type and form—is placed early in the prompt, while secondary elements like materials and lighting appear later, creating a hierarchy that directs the model's interpretive focus.

Application of img2img and LoRA Techniques to Façade Images

The img2img technique refines a base input image into a more detailed architectural form while preserving its original structure by applying learned styles and material characteristics. For example, when a simple massing model like Figure 4 is used, Stable Diffusion transforms it into a modern façade based on prior training.

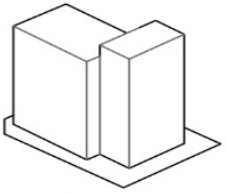


Figure 4
Simple mass image

Figure 5 illustrates this process, showing how a basic volume is interpreted into a contemporary design using a specific prompt. When the trained LoRA model is integrated with Stable Diffusion, the same prompt yields images with enhanced architectural context, material detail, and formal precision. This integration results in outputs that better align with design intent. As shown in Figure 6, applying LoRA to the same base and prompt used in Figure 6 produces a more coherent and refined façade. These results demonstrate that combining LoRA with img2img significantly enhances design control and expressive capability in AI-driven architectural image generation.

Figure 5
Mass image using img2img

Figure 6
Mass image using LoRA

EVALUATION OF GENERATED ARCHITECTURAL FAÇADE IMAGES

Architectural Façade Image Generation and Evaluation Approach

Architectural façade images were generated using Stable Diffusion's img2img and LoRA techniques, following the experimental procedure described earlier. The Negative prompt was fixed, while the Positive prompt was refined based on the design parameters in Table 2, producing images with varied massing and stylistic traits. Evaluation was conducted using CLIP and GPT-4V models (Liang et al., 2023; Wang et al., 2022). Given CLIP's 77-token limit, prompts were structured to prioritize essential architectural elements. GPT-4V used a similar strategy, organizing questions around spatial composition, material coherence, and structural clarity. CLIP scores, calculated via a Python library, provided quantitative semantic alignment between prompts and images through average logits values. GPT-4V offered qualitative assessments focused on architectural relevance and contextual consistency. Average scores from both models were used to comprehensively evaluate how accurately the generated images reflected the intended architectural design prompts.

In addition to model-based evaluations, a human-centered assessment with architectural professionals and students was conducted. Using the same four dimensions, participants provided qualitative feedback on spatial, structural, and material coherence, complementing automated assessments with practical design insights.

CLIP, GPT-4V and Expert Evaluation Process

The CLIP score evaluation carefully followed the criteria introduced in the previous section, systematically and rigorously analyzing prompts from Table 2 and thoughtfully classifying them into four categories: A (overall concept and use), B (façade and exterior), C (structure and spatial composition), and D (urban context), as shown in Table 3. The CLIP model quantified the semantic similarity between each image and its prompt using logits-based scores. Similarly, GPT-4V applied the same four categories (a–d) and assigned scores from 0 to 10, evaluating not only the presence of architectural features but also their contextual and aesthetic consistency with the intended design. Table 3 categorizes both the CLIP and GPT-4V evaluation items into four architectural dimensions: A/a (programmatic and typological intent), B/b (façade expression and material articulation), C/c (structural and spatial logic), and D/d (urban and contextual coherence). For CLIP, each prompt was divided into four semantic axes, and logits-based scores were calculated for each.

GPT-4V, on the other hand, used these same dimensions to generate qualitative questions that assess not just the presence of features, but also their functional and spatial consistency within the image. This dual-layered evaluation enables both semantic (CLIP) and architectural (GPT-4V) alignment to be assessed across comparable categories, offering a more comprehensive framework for performance verification.

To enhance architectural validity and assess practical applicability, a supplementary human evaluation was conducted in parallel with the model-based assessments. A total of 11 participants—including architectural professionals and graduate students—were recruited, with an average of 16.7 years of professional experience (ranging from 7 to 23 years). Using the same four evaluation categories, they provided qualitative feedback on spatial layout, structural logic, façade articulation, and urban contextual fit. Their assessments were especially valuable in identifying subtle architectural inconsistencies and practical design mismatches that might be overlooked by automated tools. This human-centered assessment was essential in verifying whether the AI-generated outputs aligned with real-world

architectural expectations and design sensibilities, thus serving as a critical reference for validating the reliability and robustness of the dual-model evaluation framework.

Table 2
Comparison of generated images

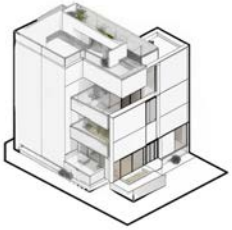
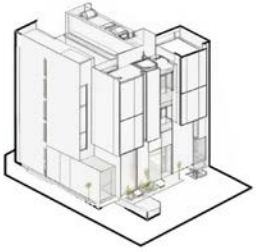
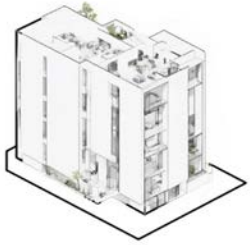
	Img2img	LoRA
Mass1		
Prompt	architecture, facade, contemporary geometric design, luxury multi-family residence, urban apartment complex, reinforced concrete frame, textured stucco finish, cantilevered balconies, glass railings, decorative metal screen integration, sharp angular elements, bold structural framing, minimalist yet expressive facade, high-end residential building, modern urban aesthetic, well-balanced proportions, sophisticated contrast of materials, ultra-realistic	
Mass2		
Prompt	architecture, facade, contemporary commercial or mixed-use building, dark metal cladding, large glass windows with black frames, warm wooden ceiling accents, integrated LED lighting along façade, clean geometric forms, cantilevered balconies, modern urban development, sleek minimalist aesthetic, elegant structural composition, high-end architectural lighting	
Mass3		
prompt	architecture, facade, Contemporary Minimalist, Mid-Rise Residential, Open-Plan Apartment Complex, Exposed Steel Frame, Floor-to-Ceiling Glass Windows, Wooden Cladding, Recessed Balconies, Glass Railings, Clean Geometric Forms, Modern Urban Aesthetic	

Table 3
Mass CLIP prompt and architectural evaluation criteria

	CLIP Model		GPT-4V / Expert evaluation	
Mass1	A	A luxury multi-family residential building with a contemporary geometric design and a reinforced concrete frame.	a	Does this building resemble a luxury multi-family residential structure with contemporary geometric design and a clearly visible reinforced concrete frame?
	B	An urban apartment complex featuring a textured stucco finish, cantilevered balconies with glass railings, and decorative metal screen integration	b	Is the facade finished with textured stucco and does it feature clearly articulated cantilevered balconies with glass railings and decorative metal screen elements?
	C	Sharp angular elements and bold structural framing define the minimalist yet expressive facade of this high-end residential building	c	Is the facade strongly defined by sharp angular elements and bold structural framing that create a minimalist yet expressive architectural character?
	D	A modern urban design with well-balanced proportions and a sophisticated contrast of materials, presented in an ultra-realistic aesthetic	d	Does the design present a modern urban composition with highly refined proportions and sophisticated material contrast in a realistic, high-end architectural expression?
Mass2	A	A contemporary commercial or mixed-use building featuring dark metal cladding and large glass windows with black frames.	a	Does this building resemble a contemporary commercial or mixed-use structure with dark metal cladding and large glass windows with black frames?
	B	The facade includes warm wooden ceiling accents and integrated LED lighting along clean geometric forms	b	Are there any elements—such as warm lighting glimpsed through windows, soft material tones, or wood-like finishes—that contribute to a warm or welcoming atmosphere, even subtly?
	C	Cantilevered balconies add dimension to the sleek minimalist aesthetic of the modern urban development	c	Do the cantilevered balconies add depth to the sleek minimalist aesthetic of this modern urban design?
	D	An elegant structural composition with high-end architectural lighting highlights the sophistication of the design	d	Does the architectural lighting highlight the elegance and refined structural composition of the design?
Mass3	A	A contemporary minimalist mid-rise residential building with an open-plan apartment layout and exposed steel frame	a	Is this a minimalist mid-rise modern residential building with an open-plan layout and exposed steel framing?
	B	The facade features floor-to-ceiling glass windows and natural wooden cladding, emphasizing transparency and warmth	b	Does the facade combine floor-to-ceiling glazing with natural wood cladding, emphasizing warmth and transparency?
	C	Recessed balconies with glass railings are integrated into the clean geometric forms of the structure	c	Are recessed balconies integrated into the clean geometric form of the facade?
	D	The design showcases a modern urban aesthetic through balanced proportions and minimal material palette	d	Does the design reflect a modern urban aesthetic through balanced proportions and a minimal material palette?

Evaluation of Generated Architectural Façade Images

CLIP	Mass	A	B	C	D	Avg.
	1	25.5	29.7	24.1	25.8	26.3
	2	23.5	22.5	29.0	21.8	24.2
	3	32.1	24.5	30.0	26.2	28.2
GPT-4V	Mass	a	b	c	d	Avg.
	1	8	7	9	8	8
	2	8	6	7	8	7.25
	3	9	8	7	9	8.25
Expert	Mass	a	b	c	d	Avg.
	1	7	6	8	8	7.25
	2	5	5	6	7	5.75
	3	8	9	8	8	8.25

Table 4
Evaluation table for generated images

In this study, three façade design variants—Mass1, Mass2, and Mass3—were evaluated using CLIP, GPT-4V, and human expert feedback. CLIP assesses semantic alignment between visual content and architectural prompts, rewarding accurate depictions of features such as modular façades, cantilevered balconies, or metal cladding. While Mass1 and Mass2 incorporated several of these elements, they received relatively lower scores due to a lack of clear structural articulation and contextual depth. Mass3, which demonstrated well-defined geometric layouts and reinforced concrete structures, achieved the highest CLIP score.

GPT-4V adopts a more qualitative, design-centered approach, focusing on spatial composition, material coherence, and formal balance. Mass3 again outperformed the others, earning the highest average score by effectively showcasing minimal mid-rise housing with exposed framing and clean geometric order. Mass1 and Mass2, though visually appealing, were evaluated as lacking in overall structural clarity and contextual richness.

A third layer of evaluation was conducted by architectural experts. They applied the same four dimensions—programmatic intent, façade expression, structural logic, and contextual coherence—to assess each generated image. The results closely aligned with the CLIP and GPT-4V assessments, consistently favoring Mass3 for its compositional clarity and material consistency. However, human evaluators assigned lower relative scores to Mass2, reflecting greater sensitivity to spatial ambiguity and practical feasibility.

Although the three evaluators differed in methodology—CLIP emphasizing text–image alignment, GPT-4V prioritizing architectural quality, and human reviewers applying practical design judgment—they collectively favored images with clear geometric articulation, coherent materiality, and structurally legible compositions.

DISCUSSION

The evaluation results collectively indicate that façade images with clearly defined geometry and consistent material expression scored higher across all evaluation methods—CLIP, GPT-4V, and expert evaluation. Despite methodological differences, the three evaluators consistently favored designs that exhibited architectural clarity, stylistic coherence, and structural legibility. Among the tested variants, Mass3 was consistently rated the highest, reflecting its compositional precision and material consistency.

Nonetheless, both CLIP and GPT-4V are limited in their ability to interpret deeper architectural logic. While CLIP measures semantic alignment between prompts and images, and GPT-4V assesses spatial and material consistency, neither can fully evaluate aspects like circulation flow, construction feasibility, or functional program alignment.

The integrated use of LoRA, img2img, and ControlNet proved effective in enhancing stylistic control and preserving structural intent. However,

the relatively narrow scope of the training dataset likely restricted the range of façade variations and contextual depth. This highlights the need for more diverse and richly labeled architectural datasets to improve generative outcomes.

To address these limitations, expert evaluations were incorporated. Their insights captured subtle design issues—such as spatial ambiguity and practicality—that were not always reflected in model-based assessments. This underscores the value of combining AI evaluations with domain-specific expertise to assess architectural validity more comprehensively.

In sum, the proposed method functions best as a conceptual support tool for early-stage design ideation. While it enables efficient generation and comparison of façade alternatives, its current implementation is not intended for detailed design development. For broader adoption in architectural workflows, future work should focus on improving prompt control, enhancing model interpretability, and expanding dataset diversity.

CONCLUSION

This study proposed an image generation method combining Stable Diffusion with LoRA, enabling more precise expression of architectural details such as windows, balconies, and materials. Evaluation results showed that images with clearer design intent achieved higher scores, confirming the effectiveness of LoRA in capturing architectural context. While the dataset was somewhat limited in style diversity and complexity, future work could address this by incorporating more architectural types and environmental conditions.

Overall, the method demonstrates potential as a practical design tool for generating and comparing design alternatives efficiently in the early stages of the architectural process.

However, this approach is not intended to replace professional design tools, but rather to function as a conceptual support system that facilitates rapid ideation and stylistic experimentation. It is best suited for early design stages where architects explore multiple alternatives. The system has already undergone preliminary integration testing with architecture professionals and students, confirming its potential for incorporation into real design workflows. Further refinement will focus on improving interface usability and expanding functional capabilities.

ACKNOWLEDGEMENT

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).

REFERENCES

Ching, F.D.K., (2023). *Architecture: Form, space, and order*. 5th ed. Hoboken, NJ: John Wiley & Sons.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., (2014). *Generative adversarial nets. Advances in Neural Information Processing Systems*, 27, pp.2672–2680.

Ho, J., Jain, A. and Abbeel, P., (2020). 'Denoising diffusion probabilistic models'. *Advances in Neural Information Processing Systems*, 33, pp.6840–6851.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W., (2021). 'LoRA: Low-rank adaptation of large language models'. arXiv preprint. Available at: https://arxiv.org/abs/2106.09685 [Accessed 25 Mar. 2025].

Jo, H. and Lee, J., (2023). 'Approaches to building facade design reflecting local identities using generative AI: Focusing on commercial streets in Seongsu-dong area'. *Journal of the Korea Institute of the Spatial Design*, 18(7), pp.361–369.

Lee, D. and Ko, S., (2023). 'Experiment and evaluation of architectural image generation through artificial intelligence-based text image generation tool'. *KIEAE Journal*, 23(5), pp.13–22.

Liang, P., Zou, Y., Zhang, C. and Ghosh, S., (2023). 'An early evaluation of GPT-4V(ision)'. arXiv preprint. Available at: https://arxiv.org/abs/2310.16534 [Accessed 25 Mar. 2025].

Ma, H. and Zheng, H., (2023). 'Text semantics to image generation: A method of building facades design based on Stable Diffusion model'. In: *The International Conference on Computational Design and Robotic Fabrication, Singapore*, Jul. 2023. Singapore: Springer Nature, pp.24–34.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y. and Ermon, S., (2021). 'SDEdit: Guided image synthesis and editing with stochastic differential equations'. arXiv preprint. Available at: https://arxiv.org/abs/2108.01073 [Accessed 25 Mar. 2025].

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S. and Clark, J., (2021). 'Learning transferable visual models from natural language supervision'. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp.8748–8763.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B., (2022). 'High-resolution image synthesis with latent diffusion models'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.10684–10695.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J. and Norouzi, M., (2022). 'Photorealistic text-to-image diffusion models with deep language understanding'. *Advances in Neural Information Processing Systems*, 35, pp.36479–36494.

Wang, J., Chan, K.C.K. and Loy, C.C., (2022). 'Exploring CLIP for assessing the look and feel of images'. *arXiv* preprint. Available at: https://arxiv.org/abs/2207.12396 [Accessed 25 Mar. 2025].

Wang, J. and Zhang, X., (2024). 'Exploring text-based realistic building facades editing application'. *arXiv* preprint. Available at: https://arxiv.org/abs/2405.02967 [Accessed 25 Mar. 2025].

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X. and He, X., (2018). 'AttnGAN: Fine-grained text-to-image generation with attentional generative adversarial networks'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1316–1324.

Yoo, Y. and Lee, J., (2023). 'Generative AI-based construction of architect's style-trained models and its application for visualization of residential houses'. *Design Convergence Study*, 22(6), pp.103–116.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X. and Metaxas, D.N., (2018). 'StackGAN++: Realistic image synthesis with stacked generative adversarial networks'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8), pp.1947–1962.

Zhang, L., Rao, A. and Agrawala, M., (2023). 'Adding conditional control to text-to-image diffusion models'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.3836–3847.

Zhao, Z. and Li, R., (2023). 'Modified generative adversarial networks for image classification'. *Evolutionary Intelligence*, 16(6), pp.1899–1906.