# A Diffusion-Based Generative AI Framework: An Exterior House Design from Textual Descriptions

Muhammad Amirul Akmal bin Ajusin[1], Noor Hasimah Ibrahim Teo[2]*,
Rosniza Roslan[3], Raseeda Hamzah[4], Anita Ahmad Kasim[5]
School of Computing-Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, Melaka, Malaysia[1, 2, 3, 4]
Department of Information Engineering-Faculty of Engineering, Universitas Tadulako, Central Sulawesi, Indonesia[5]

***Abstract*—The architectural design process is often iterative, time-consuming, and heavily dependent on effective communication between clients and professionals. Existing design tools, such as Computer-Aided Design (CAD) systems, require technical expertise, limiting accessibility for non-professional users. This study proposes a generative artificial intelligence framework for exterior house design using a diffusion-based text-to-image model. The proposed approach integrates Stable Diffusion for image generation with a vision-language model (BLIP) to enhance semantic alignment between textual descriptions and generated outputs. In addition, an interactive refinement mechanism based on image inpainting is incorporated to allow localized modification of design elements. The system is trained on a dataset of exterior house images and evaluated using quantitative metrics, including CLIP Score and Fréchet Inception Distance (FID), as well as usability assessment. Experimental results demonstrate that the proposed framework is capable of generating semantically relevant and visually coherent architectural designs, while improving accessibility and reducing the time required for design iteration. The findings highlight the potential of generative AI as an effective tool for supporting user-centric architectural visualization and design exploration.**

***Keywords*—Generative artificial intelligence; stable diffusion; text-to-image generation; architectural design; image inpainting**

## I. Introduction

Architecture plays a fundamental role in shaping human living environments by integrating functional, cultural, and aesthetic elements. The design of residential houses, particularly exterior design, requires careful consideration of architectural components such as roof structure, wall materials, windows, doors, and color schemes. Traditionally, the design process involves iterative communication between clients and architects, which can be time-consuming and inefficient due to repeated revisions and difficulty in visualizing design concepts at early stages.

Despite the availability of digital tools such as Computer-Aided Design (CAD), these systems often require significant technical expertise, limiting accessibility for non-professional users. As a result, clients face challenges in expressing their preferences effectively, while architects encounter difficulties in interpreting and translating user requirements into visual designs. This limitation highlights the need for intelligent systems that can bridge the gap between user intent and design visualization.

Recent advancements in Generative Artificial Intelligence (AI) have introduced new possibilities for automating creative tasks, including image generation and design synthesis. In particular, text-to-image generation has emerged as a promising approach that enables the transformation of natural language descriptions into visual outputs, combining techniques from natural language processing and computer vision [1], [2]. Early work in this area primarily utilized Generative Adversarial Networks (GANs), which consist of a generator and discriminator trained in an adversarial manner to produce realistic images [3]-[5]. Variants such as Conditional GAN (cGAN) [6], Deep Convolutional GAN (DCGAN) [7], and AttnGAN [8] have demonstrated improved control and semantic alignment in generated images.

Further developments introduced specialized GAN-based models for architectural and design applications. For instance, HouseGAN enables graph-constrained house layout generation [9], while Roof-GAN focuses on generating realistic roof geometries [10]. Other models, such as Pix2Pix, have been applied for image-to-image translation tasks, including architectural layout generation [11], [12]. However, GAN-based approaches often suffer from training instability, mode collapse, and limited diversity in generated outputs [13].

To address these limitations, diffusion models have gained significant attention as a more stable and robust alternative for generative tasks. Diffusion models generate images by progressively denoising random noise through a learned reverse process, allowing them to capture complex data distributions effectively [14], [15]. Denoising Diffusion Probabilistic Models (DDPM) have demonstrated superior performance in terms of image quality and diversity compared to GAN-based methods [16]. More recently, Stable Diffusion, a latent diffusion model, has enabled efficient high-resolution image generation with reduced computational cost, making it suitable for real-world applications [17].

In parallel, vision-language models have been developed to enhance the relationship between textual and visual representations. Models such as Bootstrapping Language-Image Pre-training (BLIP) leverage transformer-based architectures to generate context-aware captions and improve semantic alignment between images and text [18], [19]. These models play a crucial role in improving the accuracy and interpretability of text-to-image generation systems.

*Corresponding author.

Another important advancement in generative AI is image inpainting, which enables localized modification of images by reconstructing missing or masked regions. Modern inpainting techniques, particularly those based on diffusion models, allow users to refine specific parts of generated images while preserving global consistency [20], [21]. This capability is especially useful in architectural design, where users may wish to iteratively adjust design elements such as windows, doors, or color schemes.

Despite these technological advancements, existing research primarily focuses on model development rather than user-centric applications. Many systems lack intuitive interfaces that allow non-expert users to interact with generative models effectively. Furthermore, the integration of text-to-image generation, vision-language alignment, and interactive editing within a single architectural design framework remains underexplored.

Therefore, this study proposes a diffusion-based generative AI framework for exterior house design that integrates Stable Diffusion for image generation, BLIP for semantic alignment, and an inpainting module for interactive refinement. The proposed system is designed to be accessible to non-professional users through a chatbot-based interface, enabling intuitive and efficient design exploration. The effectiveness of the proposed approach is evaluated using quantitative metrics, including CLIP Score and Fréchet Inception Distance (FID), as well as usability assessment.

## II. Methodology

### A. System Design

This study adopted a Software Development Life Cycle (SDLC) perspective and implemented a modified waterfall model comprising requirements analysis, system design, implementation, testing, and documentation. In the adapted process, the deployment phase was omitted, and documentation replaced maintenance due to time and resource constraints. The overall research flow starts with image acquisition, followed by preprocessing, BLIP-based caption generation, Stable Diffusion model training, system integration, and evaluation.

The modified waterfall model was selected because the project required a structured sequence of development activities while still allowing limited feedback between phases. This was suitable for a system that combines dataset preparation, model fine-tuning, application integration, and user-based evaluation.

### B. System Architecture

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations. The proposed generative AI framework integrates multiple components, including data preprocessing, caption generation, model training, and user interaction, to enable automated exterior house design from textual descriptions. The system is designed as a multi-layer architecture consisting of data, application, and interaction layers.

The data and model layer includes dataset preparation, BLIP-based caption generation, and Stable Diffusion model fine-tuning. The application layer consists of a Flask-based backend that handles text-to-image generation and inpainting processes. The interaction layer is implemented using a React-based frontend that provides a chatbot interface for user input, image visualization, and design refinement.

The workflow begins when a user submits a text prompt, which is processed by the Stable Diffusion model to generate an initial house design. Users can further refine the generated output through an inpainting module that enables localized modifications. The system supports iterative interaction, allowing users to explore multiple design variations efficiently. The complete architecture and workflow of the proposed system are illustrated in Fig. 1.

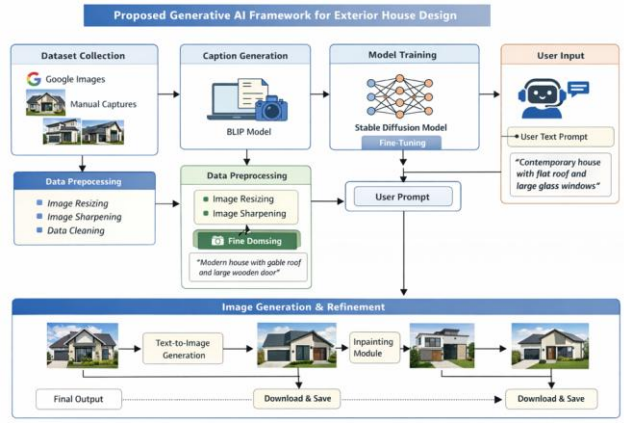


Fig. 1. Proposed generative AI framework for exterior house design.

This integrated architecture ensures seamless interaction between model components and the user interface, enabling an efficient and user-centric design workflow.

### C. Dataset Collection

The dataset used in this study consists of 1,100 exterior house images, collected to represent a wide range of architectural styles and design elements relevant to residential buildings. The images were primarily obtained from publicly accessible online sources, including Google Images, using the Imageye browser extension, as well as manually captured images.

To ensure diversity, various search queries such as *"exterior house design"*, *"modern house"*, *"house with hip roof"*, and *"house with wooden door"* were used during the data collection process. This approach enabled the inclusion of different house styles, including modern, traditional, and contemporary designs.

The collected dataset focuses on front-view exterior house images and captures key architectural elements such as roof style, window type, and door design (Table I). Based on the dataset distribution, four main roof types were identified: hip roof (240 images), gable roof (275 images), flat roof (465 images), and sloped roof (120 images).

For window types, the dataset includes sliding windows (385 images), casement windows (372 images), and hung windows (343 images). In terms of door styles, wooden doors (498 images), sliding doors (397 images), and French doors (205 images) are represented.

TABLE I. DATASET STATISTICS BY ARCHITECTURAL FEATURES

| Feature Category | Feature Type | Number of Images |
|---|---|---|
| **Roof Style** | Hip Roof | 240 |
| | Gable Roof | 275 |
| | Flat Roof | 465 |
| | Sloped Roof | 120 |
| **Window Type** | Sliding Window | 385 |
| | Casement Window | 372 |
| | Hung Window | 343 |
| **Door Style** | Wooden Door | 498 |
| | Sliding Door | 397 |
| | French Door | 205 |
| **Total Images** | | **1100** |

The dataset was carefully curated to remove irrelevant, duplicate, and low-quality images, ensuring that only clear and consistent exterior house images were retained. This filtering process improves the reliability of the dataset and enhances the model's ability to learn meaningful visual patterns.

Overall, the dataset provides a diverse yet structured representation of architectural features, which is essential for training a text-to-image generation model capable of producing realistic and semantically accurate exterior house designs.

### D. Data Preprocessing

Data preprocessing is a crucial step to ensure that the collected dataset is suitable for training deep learning models. In this study, both image and text data undergo preprocessing to standardize input formats, enhance feature quality, and prepare structured image-caption pairs for model training.

*1) Image resizing:* All collected images were resized to a uniform resolution of 512 × 512 pixels to ensure consistency across the dataset. Standardizing image dimensions is essential for efficient model training, as it reduces computational complexity while preserving important architectural details such as edges, textures, and structural components.

*2) Image enhancement:* After resizing, image sharpening techniques were applied to enhance visual clarity and emphasize key architectural features, including roof edges, window frames, and door structures. This step improves feature distinguishability, allowing the model to better capture structural patterns during training.

*3) Data cleaning:* The dataset was manually cleaned to remove irrelevant, duplicate, and low-quality images. Images that did not represent exterior house views, such as interior scenes or incomplete structures, were excluded. This process ensures that the dataset remains consistent and reduces noise that may negatively affect model performance.

*4) Caption generation using BLIP:* To enable text-to-image learning, each image must be paired with a corresponding textual description. In this study, captions were generated using the Bootstrapping Language-Image Pre-training (BLIP) model. The BLIP model was fine-tuned on the dataset to produce descriptive captions that capture key architectural elements such as roof type, window style, door design, and wall characteristics.

The generated captions serve as textual representations of the visual features, allowing the model to learn the relationship between natural language input and architectural design elements.

*5) Dataset structuring:* The final dataset was organized into structured image-text pairs in the form:

(Image, Caption)

This structured format is essential for training the Stable Diffusion model, as it enables supervised learning of the mapping between textual prompts and corresponding visual outputs. An example of an image-caption pair from the dataset is shown in Fig. 2. The caption describes key architectural elements, including roof type, window style, door design, and wall color, which are used as conditioning inputs during model training.

Fig. 2. Example of image-caption pair used in the dataset: "Modern house with flat roof, large sliding windows, white wall, and wooden front door."

The example illustrates how textual descriptions are aligned with visual features, allowing the model to learn meaningful relationships between language and architectural design elements. This structured pairing plays a crucial role in improving the semantic accuracy of the generated images.

*6) Data preparation for model training:* The processed dataset was prepared for training by ensuring compatibility with the Stable Diffusion framework. Image-caption pairs were formatted and stored in a suitable structure for fine-tuning using the DreamBooth method. Proper data preparation improves training stability and contributes to the generation of high-quality and semantically accurate images.

### E. System Design

This section describes the development of the core models used in the proposed framework, namely the BLIP model for image captioning and the Stable Diffusion model for text-to-image generation. The integration of these models enables the system to learn the relationship between textual descriptions and architectural visual features.

*1) BLIP model for caption generation:* The Bootstrapping Language-Image Pre-training (BLIP) model is employed to generate descriptive captions for the exterior house images. Initially, a pre-trained BLIP model is loaded and fine-tuned using the prepared dataset to adapt it to the domain of architectural design.

During training, the BLIP processor converts both images and associated textual descriptions into numerical representations, allowing the model to learn semantic relationships between visual features and language. The fine-tuning process improves the model's ability to generate captions that accurately describe architectural elements such as roof style, window type, door design, and wall characteristics.

Once trained, the BLIP model is used to automatically generate captions for all images in the dataset. These captions form structured image-text pairs that serve as input for training the text-to-image generation model.

*2) Stable diffusion model for text-to-image generation:* The Stable Diffusion model is used as the primary generative model for producing exterior house designs from textual prompts. Stable Diffusion operates by transforming random noise into meaningful images through an iterative denoising process guided by text embeddings.

In this study, a pre-trained Stable Diffusion model is fine-tuned using the prepared image-caption dataset. The fine-tuning process is performed using the DreamBooth approach, which enables the model to adapt to domain-specific features while preserving general image generation capabilities.

During training, textual captions are encoded into embeddings and used to condition the denoising process. A U-Net architecture progressively removes noise from a latent representation, while a decoder converts the final latent representation into a high-resolution image. This process allows the model to generate visually coherent images that correspond to user-defined prompts.

*3) Model integration:* The BLIP and Stable Diffusion models are integrated to form a unified generative pipeline. The BLIP model provides high-quality captions that enhance semantic alignment, while the Stable Diffusion model generates images based on these textual descriptions.

This integration improves the overall performance of the system by ensuring that generated images are both visually realistic and semantically relevant. The combined framework enables effective mapping between user input and generated architectural designs.

*4) Model deployment:* After fine-tuning, the Stable Diffusion model is exported into the Hugging Face Diffusers format to support efficient inference. The trained model is then deployed within a Flask-based backend, which processes user prompts and returns generated images in real time.

This deployment setup ensures scalability and enables seamless interaction between the frontend interface and the generative model.

### F. Proposed Framework Algorithm

The overall workflow of the proposed generative AI framework is formalized in Algorithm 1, which integrates data preprocessing, caption generation, model training, text-to-image generation, and interactive refinement.

Algorithm 1:

```
Input:
   Raw image dataset I_raw
   User prompt p
Output:
   Generated or refined exterior house image g

// DATA PREPARATION PHASE
Collect exterior house images I_raw from online sources
For each image image_i in I_raw do
    Resize image_i to 512 × 512
    Apply sharpening to enhance architectural features
    Store processed image_i in dataset I_clean
End For

// CAPTION GENERATION (BLIP)
Load pre-trained BLIP model
Fine-tune BLIP using I_clean
For each image image_i in I_clean do
    Generate caption_i using BLIP
    Store (image_i, caption_i) in dataset D
End For

// MODEL TRAINING (STABLE DIFFUSION)
Load pre-trained Stable Diffusion model M_sd
Fine-tune M_sd using dataset D (image-caption pairs)
Save trained model

// INFERENCE PHASE (TEXT-TO-IMAGE)
Receive user prompt p from interface
Encode p into text embeddings
Initialize latent vector z with random noise
For t = T down to 1 do
    Predict noise using U-Net conditioned on p
    Denoise z iteratively
End For
Decode final latent z into generated image g

// OPTIONAL INPAINTING (REFINEMENT)
If user selects region for editing then
    Receive mask m and refinement prompt p_r
    Preserve unmasked regions of g
    Apply diffusion-based inpainting on masked region
    Generate refined image g'
    Set g ← g'
End If

// OUTPUT
Display generated image g to user
Allow download or further refinement

Return g
```

Algorithm 1 illustrates the complete pipeline from dataset preparation to user interaction, highlighting the integration of diffusion-based generation and vision-language alignment within a unified framework.

## III. Result and Discussion

This section presents the evaluation results of the proposed generative AI framework for exterior house design. The system is assessed based on functionality, semantic accuracy, image quality, and usability. Quantitative metrics, including CLIP Score and Fréchet Inception Distance (FID), are used to evaluate model performance, while user-based testing is conducted to assess system usability.

### A. Functionality Testing

Functionality testing was conducted to verify that the proposed system operates correctly in supporting the main user tasks. The evaluation focused on validating the core features of the system, including text-to-image generation, image refinement through inpainting, user interface interaction, and image output management.

The following key features were tested during the functionality evaluation:

- Text prompt input through the chatbot interface.
- Text-to-image generation using the Stable Diffusion model.
- Display of generated images.
- Image variation generation.
- Inpainting-based image refinement.
- Masking tools (brush, eraser, and selection tools).
- Navigation between Design Generator and Inpainting Studio.
- Image download functionality.

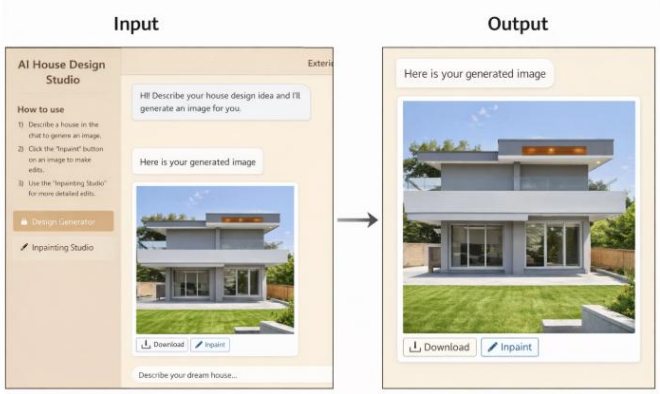


Fig. 3. Example of text-to-image generation using a prompt.

For the text-to-image generation module, the system successfully accepted natural language prompts entered through the chatbot interface and produced corresponding exterior house design images. As illustrated in Fig. 3, the generated result reflects the architectural characteristics described by the user, indicating that the generation pipeline is able to transform textual input into a coherent visual output. This demonstrates that the model and interface are properly integrated and capable of supporting intuitive design exploration.

The system also performed reliably in supporting image refinement through the inpainting module. Users were able to select a specific region of the generated image, apply a mask, and submit an edit prompt to modify the selected area. As shown in Fig. 4, the system successfully updated the targeted region while preserving the overall layout and visual consistency of the house design. This confirms the effectiveness of the inpainting function for localized architectural refinement.


Fig. 4. Inpainting-based refinement of generated house design.

In addition, supporting features such as switching between modules, selecting generated variations, and downloading images were executed without errors. The masking tools (e.g., brush and eraser) also functioned correctly, allowing users to interactively define regions for modification.

Overall, the functionality testing results indicate that all major system components operate as expected. The successful execution of generation, refinement, and interaction features demonstrates that the proposed framework is fully functional and suitable for assisting users in exterior house design visualization.

### B. Semantic Accuracy Evaluation (CLIP Score)

The semantic alignment between generated images and their corresponding textual descriptions was evaluated using the CLIP Score. This metric measures the similarity between image and text embeddings, with higher values indicating better semantic consistency.

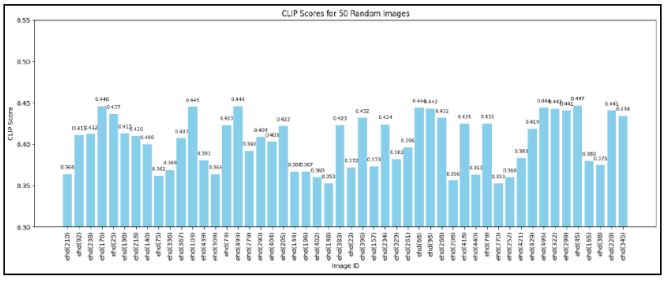


Fig. 5. CLIP score visualization.

The BLIP model achieved an average CLIP Score of 0.42, indicating moderate alignment between generated captions and visual features as shown in Fig. 5. This result suggests that the model is capable of capturing relevant architectural elements

from textual descriptions, although some discrepancies may still occur in complex scenarios.

The moderate CLIP score can be attributed to the limited dataset size and variability in architectural styles, which may affect the model's ability to generalize across diverse design inputs.

### C. Image Quality Evaluation (FID Score)

The visual quality of the generated images was evaluated using the Fréchet Inception Distance (FID), which measures the similarity between the distributions of generated images and real images in terms of feature representations. Lower FID values indicate better image quality and higher similarity to real-world data.

The Stable Diffusion model achieved an FID score of 82.4, indicating that the generated images exhibit recognizable architectural structures but still lack high levels of photorealism and fine-grained detail. Despite the relatively high FID score, qualitative observation of the generated images shows that the model is capable of producing coherent house designs with identifiable architectural elements such as roof structures, windows, doors, and wall compositions. This suggests that the model successfully captures the general layout and structural characteristics of exterior house designs, even if certain visual refinements are not fully achieved.

The obtained FID score can be attributed to several factors. First, the dataset size of 1,100 images is relatively limited compared to large-scale datasets typically used in state-of-the-art generative models. This restricts the model's ability to learn diverse architectural patterns and may lead to less realistic outputs. Second, the dataset exhibits some imbalance in feature distribution, particularly with a higher proportion of flat roof and wooden door designs, which may bias the generated results toward more common styles.

In addition, the complexity of architectural features poses a challenge for generative models. Unlike general object generation, house design involves structured geometry, symmetry, and spatial consistency, which require the model to capture both global layout and local details simultaneously. The current model demonstrates competence in generating overall structure but shows limitations in rendering fine textures, lighting consistency, and detailed architectural features.

Furthermore, the fine-tuning process using the DreamBooth approach, while effective for adapting the model to a specific domain, may not fully optimize the model for high-fidelity image generation when trained on a relatively small dataset. This highlights the trade-off between domain specialization and image realism.

When compared to state-of-the-art diffusion models trained on large-scale datasets, the obtained FID score is higher; however, within the scope of this study, the generated images remain sufficiently realistic for early-stage architectural visualization and concept exploration. The primary objective of the system is to support idea generation rather than produce final construction-ready designs, and thus, the achieved image quality is considered acceptable for its intended application.

To improve image quality in future work, several enhancements can be considered. These include increasing the dataset size and diversity, applying data augmentation techniques, incorporating higher-resolution training strategies, and fine-tuning more advanced diffusion architectures. Additionally, integrating post-processing techniques or hybrid models could further enhance visual realism.

Overall, the FID evaluation indicates that while there is room for improvement in image realism, the proposed system successfully generates meaningful and visually coherent architectural designs, demonstrating the feasibility of using diffusion-based generative models for exterior house design applications.

### D. Usability Testing

Usability testing was conducted to evaluate the effectiveness, efficiency, and user satisfaction of the proposed system. The assessment was performed using the System Usability Scale (SUS), a widely adopted method for measuring perceived usability of interactive systems. The SUS questionnaire consists of ten items rated on a five-point Likert scale, ranging from "Strongly Disagree" to "Strongly Agree." The scoring method involves subtracting 1 from the scores of odd-numbered questions and subtracting the scores of even-numbered questions from 5, followed by multiplying the total score by 2.5 to obtain a final value between 0 and 100.

A total of 30 respondents participated in the usability evaluation. The results, as illustrated in Fig. 6, show that most respondents achieved SUS scores within the range of 80 to 95. The overall average SUS score obtained in this study is 86, indicating a high level of usability and user satisfaction.

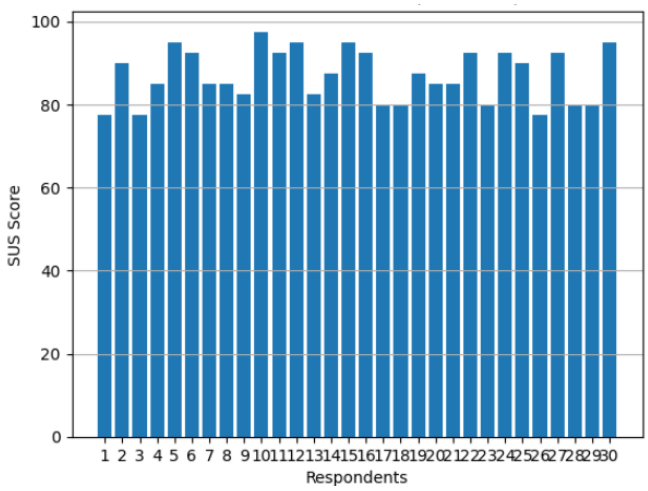


Fig. 6. SUS score distribution.

According to standard SUS interpretation guidelines, a score above 68 is considered above average, while scores above 80 are classified as excellent. Therefore, the obtained average score of 86 places the proposed system within the "excellent" usability category, suggesting that users found the system easy to use and well-designed. The high usability score can be attributed to several design factors. First, the chatbot-based interface simplifies user interaction by allowing natural language input, eliminating the need for technical knowledge in architectural

design tools. Second, the system provides immediate visual feedback through generated images, which enhances user engagement and supports rapid design exploration. Third, the inclusion of interactive features such as image variation and inpainting allows users to refine designs intuitively, contributing to a more flexible and user-centered experience.

Furthermore, the relatively consistent SUS scores across respondents indicate that the system provides a stable and reliable user experience. The absence of significantly low scores suggests that users did not encounter major usability issues during interaction. However, some limitations were observed. A few respondents reported slightly lower scores, which may be attributed to the need for clearer guidance when using advanced features such as inpainting or masking tools. This indicates that additional user assistance, such as tooltips or tutorials, could further improve the overall usability of the system.

Overall, the usability evaluation demonstrates that the proposed framework successfully meets user expectations in terms of ease of use, functionality, and interaction. The results highlight the potential of the system to serve as an accessible tool for both non-expert users and professionals in architectural design

### *E. Discussion*

The experimental results demonstrate that the proposed generative AI framework is capable of generating exterior house designs from textual descriptions with reasonable semantic alignment and visual coherence. The integration of BLIP and Stable Diffusion enables effective mapping between text and image, while the inpainting module provides flexibility for iterative design refinement.

The system offers several advantages over traditional design approaches. First, it reduces the time required for design iteration by providing instant visual feedback. Second, it improves communication between clients and architects by enabling users to visualize their ideas without requiring technical expertise. Third, it introduces an interactive and user-friendly design process through the chatbot interface.

However, several limitations were identified. The moderate CLIP score indicates that semantic alignment can be further improved, particularly for complex or ambiguous prompts. Additionally, the relatively high FID score suggests that image realism is limited by the dataset size and diversity. Expanding the dataset and incorporating more advanced training techniques could improve model performance.

Overall, the results highlight the potential of generative AI as a practical tool for architectural visualization, particularly in early-stage design and concept exploration.

## IV. CONCLUSION

This study presented a diffusion-based generative AI framework for exterior house design that enables the transformation of textual descriptions into architectural visualizations. The proposed system integrates Stable Diffusion for text-to-image generation with a BLIP-based captioning model to improve semantic alignment, along with an inpainting module for interactive design refinement. The experimental results demonstrate that the proposed framework is capable of generating visually coherent and semantically relevant exterior house designs. The system achieved a CLIP Score of 0.42, indicating moderate alignment between textual descriptions and generated images, and an FID score of 82.4, suggesting acceptable image quality for early-stage design visualization. In addition, usability evaluation shows that the system is intuitive and accessible, allowing non-professional users to explore design ideas effectively.

The main contributions of this study include the development of an end-to-end generative AI framework for architectural design, the integration of diffusion models with vision-language techniques, and the introduction of an interactive refinement mechanism using inpainting. These contributions address limitations in traditional design workflows by reducing dependency on technical expertise and improving communication between users and designers. Despite these contributions, several limitations remain. The relatively high FID score indicates that the generated images can be further improved in terms of realism and fine details. Additionally, the dataset size and diversity are limited, which may affect the model's ability to generalize across complex architectural styles.

Future work will focus on expanding the dataset with more diverse architectural designs, improving model performance through advanced training strategies, and extending the system to support interior design and 3D architectural generation. Furthermore, integrating real-time user feedback and deploying the system as a web or mobile application could enhance its practical applicability.

## ACKNOWLEDGMENT

The authors gratefully acknowledge Universiti Teknologi MARA for the financial and matching grant administrative support. This research was funded by Universiti Teknologi MARA (UiTM) provided through Geran Sepadan TEJA (GST 2025/1-2).

## REFERENCES

[1] L. Sudha, K. B. Aruna, V. Sureka, M. Niveditha, and S. Prema, "Semantic image synthesis from text: Current trends and future horizons in text-to-image generation," EAI Endorsed Transactions on Internet of Things, vol. 11, 2024.

[2] Z. Lin et al., "Evaluating text-to-visual generation with image-to-text generation," in Proc. European Conf. Computer Vision (ECCV), Cham, Switzerland: Springer, 2024, pp. 366–384.

[3] A. Dash, J. Ye, and G. Wang, "A review of generative adversarial networks (GANs) and its applications in a wide variety of disciplines: From medical to remote sensing," IEEE Access, vol. 12, pp. 18330–18357, 2023.

[4] T. Chakraborty et al., "Ten years of generative adversarial nets (GANs): A survey of the state-of-the-art," Machine Learning: Science and Technology, vol. 5, no. 1, p. 011001, 2024.

[5] V. L. T. de Souza et al., "A review on generative adversarial networks for image generation," Computers & Graphics, vol. 114, pp. 13–25, 2023.

[6] S. S. Mohammed and H. G. Clarke, "Conditional image-to-image translation generative adversarial network (cGAN) for fabric defect data augmentation," Neural Computing and Applications, vol. 36, no. 32, pp. 20231–20244, 2024.

[7] J. Jenkins and K. Roy, "Exploring deep convolutional generative adversarial networks (DCGAN) in biometric systems: A survey study," Discover Artificial Intelligence, vol. 4, no. 1, p. 42, 2024.

[8] R. Gopalakrishnan, N. Sambagni, and P. V. Sudeep, "An improved AttnGAN model for text-to-image synthesis," in Proc. Int. Conf.

Computer Vision and Image Processing, Cham, Switzerland: Springer, 2023, pp. 139–151.

[9] N. Nauata, K.-H. Chang, C.-Y. Cheng, G. Mori, and Y. Furukawa, "House-GAN: Relational generative adversarial networks for graph-constrained house layout generation," in Proc. ECCV, 2020, pp. 162–177.

[10] H. Tang et al., "Graph transformer GANs for graph-constrained house generation," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2173–2182.

[11] Z. Li et al., "Mapping new realities: Ground truth image creation with pix2pix image-to-image translation," arXiv preprint arXiv:2404.19265, 2024.

[12] X. Zhao, H. Yu, and H. Bian, "Image to image translation based on differential image Pix2Pix model," Computers, Materials & Continua, vol. 77, no. 1, 2023.

[13] M. Cobbinah et al., "Diversity in stable GANs: A systematic review of mode collapse mitigation strategies," Engineering Reports, vol. 7, no. 6, e70209, 2025.

[14] J. Liu et al., "Residual denoising diffusion models," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2773–2783.

[15] A. K. Bhunia et al., "Person image synthesis via denoising diffusion model," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5968–5976.

[16] G. Müller-Franzes et al., "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis," Scientific Reports, vol. 13, no. 1, p. 12098, 2023.

[17] D. Podell et al., "SDXL: Improving latent diffusion models for high-resolution image synthesis," arXiv preprint arXiv:2307.01952, 2023.

[18] P. Panchal et al., "Deep learning–driven image captioning: Progress through transformers and large language models," PLoS ONE, vol. 21, no. 3, p. e0345012, 2026.

[19] C. Yang, Z. Li, and L. Zhang, "Bootstrapping interactive image–text alignment for remote sensing image captioning," IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–12, 2024.

[20] C. Corneanu, R. Gadde, and A. M. Martinez, "LatentPaint: Image inpainting in latent space with diffusion models," in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV), 2024, pp. 4334–4343.

[21] A. Lugmayr et al., "RePaint: Inpainting using denoising diffusion probabilistic models," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11461–11471.