

CAN GENERATIVE AI MODELS COUNT?

Finetuning Stable Diffusion for Architecture Image Generation with Designated Floor Numbers Using a Small Dataset

WEISHUN XU¹, MINGMING LI² and XUYOU YANG³

¹ Zhejiang University College of Engineering and Architecture

² Zhejiang University College of Computer Science and Technology

³ 1DesignLab

¹xuweishun@zju.edu.cn, 0009-0000-4489-7858

²limingming@zju.edu.cn, 0000-0001-8916-6485

³xuyou.yang.92@gmail.com, 0000-0002-6294-378X

Abstract. Despite the increasing popularity of off-the-shelf text-to-image generative artificial intelligence models in early-stage architectural design practices, general-purpose models are challenged in domain-specific tasks such as generating buildings with the correct number of floors. We hypothesise that this problem is mainly caused by the lack of floor number information in standard training sets. To overcome the often-dodged problem in creating a text-image pair dataset large enough for finetuning the original model in design research, we propose to use BLIP method for both understanding and generation based automated labelling and captioning with online images. A small dataset of 25,172 text-image pairs created with this method is used to finetune an off-the-shelf Stable Diffusion model for 10 epochs with affordable computing power. Compared to the base model with a less than 20% chance to generate the correct number of floors, the finetuned model has an over 50% overall chance for correct floor number and 87.3% change to control the floor count discrepancy within 1 storey.

Keywords. Text-to-Image Generation, Model Finetuning, Stable Diffusion, Automated Labelling.

1. Introduction

Architectural design practices have seen a recent rise in the use of text-to-image (T2I) generative AI (GAI) models due to their improved accessibility and performance. Commercially available cutting-edge T2I GAI tools such as DALL-E-2 (Ramesh et al., 2022), MidJourney, and Stable Diffusion (Rombach et al., 2022) are based on large diffusion models that can generate high-quality images that provide design inspirations in only minutes and thus largely accelerate the design process.

Typically, T2I GAIs have been adopted in early design phases such as ideation

(Stigsen et al., 2023), digital sketching (Ploennigs & Berger, 2023) and style exploration (Chen et al., 2023). However, off-the-shelf general models are less efficient in domain-specific tasks such as architectural design, creating inconsistency in communication through text and difficulty in interpreting the results (Turchi et al., 2023). Domain-specific tasks often require extra field knowledge and respective descriptive language, leading to adaptation efforts of general models such as prompt-engineering and context-specific tuning to gain more detailed control over specific aspects of the generative process.

One of the persisting issues in adopting GAI models pre-trained on general purpose in early-stage architectural design is the generation of images with the correct number of floors. For example, when the prompt describes a building of three floors, very often the results generated by the off-the-shelf T2I GAI tools reflect a building with a wrong number of floors (Figure 1). Because prevailing building codes often associate the number of floors or building height to building typology, this problem limits the potential of application of general GAI models in early conceptualisation.



Figure 1. From left to right: images generated by Stable Diffusion with prompts as "Rendering of a Modernism office building of x storeys in mountains in China" where x is 3, 4, and 7 respectively.

To enhance the practicality of using large pre-trained T2I GAI models as an effective early-stage conceptual design tool for architecture with specific requirements, this research proposes a finetuning method using a small dataset to allow for professional specifications by generating images that match with the prompt more accurately, taking the number of floors as a case study.

2. Background

Since changing the number of floors in the prompts for off-the-shelf models does change the generated floor count only with unacceptable accuracy, we hypothesise that the poor performance is caused by the lack of floor count information in the model's training data. Hence, injecting text-image pairs with correct information to the model may be a solution. Therefore, a literature review is conducted in optimised methods for T2I GAI models to learn new knowledge for design purposes, and how training data can be prepared respectively.

2.1. ADAPTING GAI FOR EARLY-STAGE DESIGN

To better apply large pre-trained GAI models in specific downstream tasks, there have been several common methods. Overall, existing methods that require smaller training

data for design applications are often used in styling or personalisation while leaving most of the original model intact. For example, prompt engineering is a process to create and optimise input text to instruct the models to perform specific tasks, which can be used in better extracting knowledge from a GAI model to help design ideation (Deshpande, 2023). However, this method relies on the capability of the original model without injecting new knowledge. Another method to bridge these models with downstream tasks is Low-Rank Adaptation, or LoRA (Hu et al., 2021), which freezes the pretrained model weights and injects trainable rank decomposition matrices, greatly decreasing the number of trainable parameters. LoRA can be useful in styling, applying texture and rendering, but with the original model left untouched, its ability to inject structural knowledge is limited (Kuru, 2023). Specifically targeted at diffusion models, Dreambooth allows users to inject custom objects with a few images as additional training data to an existing class (Ruiz et al., 2023). Text-inversion uses a pseudo word to embed a set of vectors for highly personalised results learned from only a few input images, but mostly of the same object (Gal et al., 2022).

Yet domain-specific problems such as the counting of floors require that the original model generate the correct visual semantics with a new class of text inputs. This suggests finetuning the entire model, an approach of transfer learning, in which the weights of a pre-trained model are trained on new text-image data pairs (Goodfellow et al., 2016). Previous work in architectural design has successfully associated structured vocabulary in five classifications to architectural forms by finetuning U-Net of Stable-Diffusion v1.4 with 1001 manually labelled images (Kim, 2023). Yet the process of such data curation is demanding, requiring both labour and professional consistency in reading the form and tedious preparation of graphically similar images all in isometric views.

2.2. AUTOMATED DATA LABELLING

As can be observed from the precedents above, one of the main reasons why finetuning a T2I GAI model through re-training is often avoided in domain-specific tasks is the lack of structured data, as re-training with trivial datasets often lead to severe loss of previous knowledge in a large model (Li et al., 2022). Since accessible online images often do not have associated captions with domain-specific knowledge, researchers then must rely on either open datasets that are manually curated for general purposes, or find ways work with noisy data with expensive post-processing steps (Jia et al., 2021). However, recent developments in scaling pre-trained representations for T2I GAI models with noisy data have allowed us to consider a methodological framework with readily available resources.

One of the commonly referred to state-of-the-art visual-language pretraining frameworks, BLIP, can perform a wide range of downstream tasks that are both understanding-based and generation-based (Li et al., 2022). Pretrained with datasets generated with CapFilt, BLIP has strong abilities in understanding images and generate texts that accurately match with the given images. For our task of floor counting with T2I GAI, the structure of standard training data needs to involve a text-image pair which involve both an understanding of an original image in the form of the number of floors, and a generated natural language caption. Therefore, the BLIP model can be a viable option based on our review.

3. Method

Since domain-specific tasks in architectural design can vary from case to case performed by entities with varied computing power and labour, our goal of this early-stage research is to establish a finetuning framework with publicly available data, automated labelling and captioning with relatively small dataset to enhance its accessibility.

3.1. IMAGE DATA ACQUISITION, FILTERING AND LABELLING

3.1.1. Data Acquisition

To label images of architecture for the number of floors, source images should be exterior photos or renderings that cover the entire height of the building, and preferably the entire building volume to learn about the building typology as well. In addition, other than image resolution and clarity, photos of buildings in the dataset should cover a variety of styles, which helps the model generalise better, by covering a wider range of other descriptive language in the prompt.

The 'Exterior Photography for Architects' image label on one of the widely accessed websites for architectural design, ArchDaily (ArchDaily, n.d.), provides an ideal source for our data acquisition demand. Despite some misplacements of interior, urban design and infrastructural projects, the label collects high-quality photos of building exteriors of designs with varied styles, including materials and typology. We acquired a raw dataset total of 33,269 images from this label for filtering and processing.

3.1.2. Data Filtering

As the raw dataset still contained images from which the number of floors cannot be counted, directly applying such data for caption labelling may adversely affect the performance of our subsequent model finetuning, causing the model to learn irrelevant features. Empirically, such photos are often associated with wrong project types or views. Therefore, we needed to filter such photos based on automated recognition of image subjects.

Conventional image filtering methods often use pre-trained image classification models or object detection models. In our case, such methods may not be flexible enough considering the complexity of the background in architectural photos. In addition, since our filter standard was empirically based on observation, we needed a method that adapts to varied needs described in natural language.

We used a zero-shot classification method based on BLIP to categorise and filter images. In our method, the BLIP model and its pre-processor were loaded to extract image and text features respectively, and then the similarity between them was calculated using the Softmax function. This way, the subject of the image can be converted into probabilities of pre-determined project types based on our initial empirical review of the dataset, with the highest probability category extracted as the final type of the subject.

Based on preliminary visual analysis, we determined that subjects which need to be filtered out belong to "bridge", "city", "public square", "top view" and "indoor space"

categories respectively. As can be seen, this process was both targeted at unwanted perspectives and subject types. In the end, we finalised the categories of classification as "building", "bridge", "city", "public square", "top view", and "indoor space". Only images with a probability of being a building higher than 60% were retained. Through this method, we filtered out approximately 8k non-building images from the original data, and the final dataset size was 25172 images for labelling (Figure 2).



Figure 2. Examples of filtered image types

3.1.3. Data Labelling

To create the text-image pair required for finetuning, the caption we associate with each image needs to both include our main training goal which is the number of floors, and some natural language description about the general scene to respond to other key words that may appear in the prompts.

Based on BLIP's capability in understanding and generation-based tasks, this process could be automated. We used the BLIP-VQA (visual question answering) model to get the number of building floors by asking "How many floors does this building have?". Then, we used the BLIP caption base model to generate an overall description of the image. As shown in Figure 3, for the image "192.jpg", the final output for the label was "a two-storey building. a white building with a staircase going up to it", which contains the number of floors and a description of the built environment. "20.jpg" is labelled as "a 10-storey building. a tall building with plants on the balconies", highlighting the number of floors and architectural details.


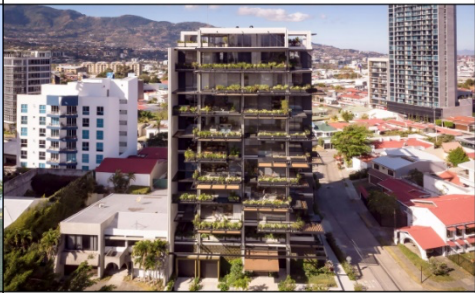
192.jpg	20.jpg
	
<p>a two-storey building. a white building with a staircase going up to it.</p>	<p>a 10-storey building. a tall building with plants on the balconies.</p>

Figure 3. Examples of understanding and generation based captioning tasks by BLIP.

This dual annotation method played a key role in our training set, providing the model with comprehensive semantic information while simultaneously accentuating the number of floors. Such labels help improve the model's understanding of the building structure and appearance, enhancing the model's generalisation capability. In the finetuning stage, these labels make it easier for the model to adjust weights to meet specific task requirements that utilise domain-specific knowledge, while reducing the ambiguity in the model's understanding of labels.

3.2. MODEL FINETUNING

As our research proposal targets at the practicality of injecting domain-specific knowledge for architectural practices, we prioritised accessible off-the-shelf models that require less computation power for finetuning.

Diffusion models are a class of deep generative models often used in computer vision tasks including image generation based on two stages: a forward diffusion stage and a reverse diffusion stage. Compared to conventional diffusion models, Latent Diffusion Models (LDMs) execute the diffusion process on the latent instead of pixel space and subsequently require much less computation resource. Stable Diffusion is one of the latest versions of LDMs for T2I tasks accessible under the CreativeML Open RAIL-M license, and the Stable Diffusion v1-4 model was chosen for our finetuning process.

With the 25172 text-image pairs, the finetuning was executed following Stable Diffusion's official finetuning documentation on a single NVIDIA GeForce RTX 4090 graphics card for 10 epochs with a learning rate of $1e-5$. The finetuning process took a total of 71 hours and 31 minutes. To further verify the efficiency of our proposed methodological framework, we also made an alternative finetuned model with the same data but for only 1 epoch as comparison.

4. Results

4.1. VALIDATION METHOD

To test the performance of our finetuned model against the original model and empirically evaluate the difference between training epochs, we designed an evaluation process based on 1000 512px-by-768px image sets generated from the original Stable Diffusion model (referred to as the base model in the following text), the finetuned model with 1 epoch learning (1epoch), and the final finetuned model with 10 epochs (10epoch) with the same seeds and prompts.

For each generation across all three models, a seed was randomly chosen images from our training data to generate the latent space of the models. Then, the prompt was automatically generated with randomly picked parameters which can describe the number of building floors, building style, and building type. The range of floor numbers was from 1 to 7, and the styles included "modern", "contemporary", "traditional", "industrial", "Art Deco", and "minimalist". Building types were chosen from "single-family house", "apartment building", "villa", "shopping mall", "retail store", "restaurant", "café", "office building", "museum", "art gallery", "library", "concert hall", "school building", "hospital", "historical building", and "church". The

prompt then was generated as a natural language text with standard descriptive structure, such as "A modern 2-storey single-family house". We also added key phrases to each prompt, namely "whole building" to designate the subject of the generated image, "human perspective" to set the view angle, and "façade with well-defined boundaries between each storey" to help generate subjects that are easy for humans to count the number of floors to quantify the performance. Due to our prompt generation method, resultant buildings in the generated images mostly have either a single volume or several volumes of similar heights, which affects our validation method described below.

We then manually counted the number of floors on each generated image for different models respectively. If the subject has a changing number of floors in the image, only the largest portion of the building was taken into the final count. For instance, for the image on the left in Figure 4, the building changes from 9 floors to 8 and a floor number of 8 was taken. In the image in the middle, the floor number of 2 was taken. In rare cases where the façades generated could not clearly separate the floors visually, other evidence such as vertically stacked windows or openings will be used as a clue for estimation, such as in the image on the right in Figure 4. All counts were executed by graduate students who received their undergraduate training in architecture to ensure that industry knowledge was involved.



Figure 4. Examples of floor count rules where professional judgement was applied.

4.2. OVERALL MODEL PERFORMANCE

We treated the number of floors given in the prompt as the ground truth, and calculated the overall percentage of correctness when the number of floors in the generated image matches the ground truth. We also calculated the average difference between the number of floors generated and the ground truth. To evaluate the models' tolerance, we noted the percentages of images when the floor count differs from ground truth by no more than 1 floor and by no more than 2 floors. The results for each model are shown in Table 1.

Model	Correct (%)	Avg. floor difference	Differ within 1 storey (%)	Differ within 2 storeys (%)
base	19.4	1.587	55	79.1
1epoch	41.3	1.048	69.8	87.3
10epoch	51.2	0.663	87.3	95.8

Table 1. Overall model performance

As the result shows, finetuned models exhibit majorly improved performance in

the number of floors. 19.4% of the images generated by the base model have the correct number of floors, while the percentage of correctness is increased to 41.3% after 1 epoch of finetuning, and subsequently to 51.2% after 10 epochs. Improved performance is also observed in controlling the floor number gap between generated images and the prompt. The average floor difference is reduced from 1.587 in the base model to 0.663 in the finetuned 10epoch model. In addition, 10epoch has a 95.8% chance of containing the floor difference within 2 storeys compared to 79.1% for the base model. A visual comparison of some generated results is shown in Figure 5.



Figure 5. 4 sets of visual comparison between images generated by base (left) and 10epoch (right)

4.3. PERFORMANCE BY NUMBER OF FLOORS

To further analyse applicable use scenarios for the models, we compared the difference between 10epoch and base model for their performance by floor count. The result of the comparison is shown in Table 2 below:

Floor Count	Correct (%)		Differ within 1 storey (%)		Differ within 2 storeys (%)	
	base	10epoch	base	10epoch	base	10epoch
1	6.6	44.1	36.8	90.4	71.3	95.6
2	28.6	73.5	73.6	92.9	91.4	100
3	32.5	58.4	81.1	92.9	94.8	98.1
4	24.6	54.6	79.2	83.8	94.6	96.2
5	12.5	46.7	48	85.5	84.9	96.1
6	9	34.6	40.4	81.4	75	94.2
7	7.6	47.7	33.3	84.1	65.2	90.2

Table 2. Base model and 10epoch performance comparison by floor counts

As demonstrated, with the number of floor accuracy lower than 30% almost across all floor counts, the base model faces general challenges but particularly in single storey buildings and buildings higher than 5 floors. The model also struggles to contain the

difference in floor counts within 1 storey, which renders its usage as an ideation source in early-stage design limited, because buildings of 1 or 6 to 7 levels are often more sensitive to floor number changes due to planning and fire code regulations.

Meanwhile, the 10epoch model exhibited a performative increase by 30% in almost every floor count category. The table demonstrates that despite the overall enhanced performance, the finetuned model has a similar pattern of performing better in generating buildings of 2 to 4 storeys. However, the finetuned model shows a more stable and elevated performance of containing the floor count difference within 1 storey. This feature greatly improves the model's practicality in early-stage designs.

5. Discussion

Our research demonstrates the feasibility of using a small dataset (25172 images), automated labelling and limited computing power, it is possible to finetune a publicly accessible T2I GAI model for majorly improved performance in early-stage architectural design tasks, such as generating ideation images with the correct number of floors, which involve domain-specific knowledge. The result also partially supports our hypothesis, which is that one critical reason for current general purpose T2I GAI models to fail basic-level early-stage architectural design tasks is the lack of well-structured training data that includes building information.

Meanwhile, we understand that as early-stage research, our methodological framework has its limitations. Firstly, our evaluation for both the base model and the finetuned models does not necessarily reflect their actual performance in practice. The automated prompt generation script can create realistically improbable descriptions such as "a minimalist 1-storey hospital building" or "a contemporary 5-storey historical building". Therefore, both models, before and after finetuning, may perform better in actual architectural design practices. The other limitation caused by the prompt generation script is that it is unable to describe complex buildings with varied volume heights, which we intend to address in future research.

In addition, we did not quantitatively validate the accuracy of the automatic labelling process, which may have led to less optimised training results. However, having observed that even a small dataset with automated labelling can dramatically increase the performance of the pre-trained model, we think that more could be done with well-structured captions that contain more comprehensive building information. Further research should expand on matching what can be visually learned and what should early-stage design address in generating architectural images for designers, so that training datasets for architectural AI can be crowd-sourced.

Acknowledgements

This research is supported by the National Natural Science Foundation of China under Grant 52208036, China's Research and Development Project of Ministry of Housing and Urban-Rural Development under Grant 2022-K-004, and Center for Balance Architecture at Zhejiang University.

References

- Chen, J., Wang, D., Shao, Z., Zhang, X., Ruan, M., Li, H., & Li, J. (2023). Using Artificial Intelligence to Generate Master-Quality Architectural Designs from Text Descriptions. *Buildings*, 13(9), Article 9. <https://doi.org/10.3390/buildings13092285>
- Deshpande, R. (2023). Generative Pre-Trained Transformers for 15-Minute City Design. In *HUMAN-CENTRIC - Proceedings of the 28th CAADRIA Conference*, (pp. 595–604). <https://doi.org/10.52842/conf.caadria.2023.1.595>
- Archdaily. (n.d.). ArchDaily. Retrieved August 28, 2023, from <https://www.archdaily.com/search/images>
- Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). *An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion* (arXiv:2208.01618). arXiv. <http://arxiv.org/abs/2208.01618>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
- 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:2106.09685). arXiv. <http://arxiv.org/abs/2106.09685>
- Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., & Duerig, T. (2021). Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *Proceedings of the 38th International Conference on Machine Learning*, 4904–4916. <https://proceedings.mlr.press/v139/jia21b.html>
- Kim, F. C. (2023). Text2Form Diffusion: Framework for learning curated architectural vocabulary. *Digital Design Reconsidered - Proceedings of the 41st Conference on Education and Research in Computer Aided Architectural Design in Europe (eCAADe 2023)*, (Vol 1, pp. 79–88). <https://doi.org/10.52842/conf.ecaade.2023.1.079>
- Kuru, J. (2023). *Training Non-Typical Character Models for Stable Diffusion Utilizing Open Sources AIS* [Honor Bachelor Thesis, University of Arizona]. <https://repository.arizona.edu/handle/10150/668639>
- Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *Proceedings of the 39th International Conference on Machine Learning*, 12888–12900. <https://proceedings.mlr.press/v162/li22n.html>
- Ploennigs, J., & Berger, M. (2023). AI art in architecture. *AI in Civil Engineering*, 2(1), 8. <https://doi.org/10.1007/s43503-023-00018-y>
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents* (arXiv:2204.06125). arXiv. <https://doi.org/10.48550/arXiv.2204.06125>
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685. <https://doi.org/10.1109/CVPR52688.2022.01042>
- Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22500–22510. <https://doi.org/10.1109/CVPR52729.2023.02155>
- Stigsen, M. B., Moisi, A., Rasoulzadeh, S., Schinegger, K., & Rutzinger, S. (2023). AI Diffusion as Design Vocabulary—Investigating the use of AI image generation in early architectural design and education. *Digital Design Reconsidered—Proceedings of the 41st Conference on Education and Research in Computer Aided Architectural Design in Europe* (Vol. 2, pp. 587–596). <https://doi.org/10.52842/conf.ecaade.2023.2.587>
- Turchi, T., Carta, S., Ambrosini, L., & Malizia, A. (2023). Human-AI Co-creation: Evaluating the Impact of Large-Scale Text-to-Image Generative Models on the Creative Process. In L. D. Spano, A. Schmidt, C. Santoro, & S. Stumpf (Eds.), *End-User Development* (pp. 35–51). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-34433-6_3