IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Resource-Efficient Diffusion Model Deployment Using LoRA Adapters and Streaming Inference for Image Synthesis

Dr. Narayanan K S [1], Midhunesh G [2]

[1]Head of Department, [2]Student

Department of BSc Data science,

Kumaraguru college of Liberal Arts and Science, Coimbatore, India

Abstract

Recent diffusion-based image generation models have demonstrated strong capabilities in visual synthesis tasks but remain difficult to deploy locally due to high memory consumption and inference latency. This work presents Spatiora, a resource-efficient diffusion-based image generation framework designed for local deployment on consumer-grade hardware. The system integrates a pretrained Stable Diffusion backbone with dynamically switchable Low-Rank Adaptation (LoRA) adapters, enabling multiple task-specific generation modes—such as interior visualization and 2D floorplan synthesis—within a single shared inference pipeline. To improve user interactivity, a streaming inference mechanism is employed to progressively deliver generated outputs, reducing perceived latency during generation. The proposed architecture minimizes GPU memory overhead by loading only the required LoRA adapter at runtime while maintaining task-specific output quality. Experimental observations indicate that the system operates within constrained memory limits while supporting multi-task image synthesis. The results demonstrate the feasibility of deploying diffusion-based generative models locally in a memory-efficient and interactive manner, making the approach suitable for privacy-preserving and resource-constrained design applications.

Keywords— Diffusion models, Stable Diffusion, Low-Rank Adaptation (LoRA), resource-efficient deployment, streaming inference, local image synthesis.

INTRODUCTION

Diffusion models have recently gained popularity as a state-of-the-art paradigm in the efficient generation of high-quality images. This is followed by the development of models such as Stable Diffusion that can create photorealistic images using a natural language description. Although these models offer significant advancements in image generation tasks, practical deployment of such models remains challenging due to their need for intense computational resources during processing.

In most design-oriented tasks, such as interior visualizations and architectural concept design, the need for fast feedback and interactive analysis is necessary. Precision and control, as found in computer-aided design software, come with the drawback of being time-consuming and inefficient during the initial phases of design idea development. Although cloud hosting generative image solutions increase convenience, issues regarding data privacy and network dependencies make them impractical for use within secure settings that require fast feedback and control.

One of the most important challenges in local deployment for diffusion models involves supporting multiple generation tasks without having similar weights for the model. This becomes an issue for more memory consumption and hence affects scalability in the case of consumer hardware. Low-Rank Adaptation (LoRA) emerges as a promising solution for adapting parameters in a way that involves adapting for the task by updating the adapter parameters while keeping the main model parameters frozen. However, most existing work involves static usage and does not attempt dynamic task switching for inference.

Another critical factor for interactive systems is the problem of perceived latency. Although the overall time for inference could still be high, improving the experience can be done by providing progressive results instead of waiting for the complete task to be done. The inference techniques for streaming can greatly improve the responsiveness for exploratory design tasks.

Inspired by these issues, this paper introduces Spatiora, which is a diffusion-based image generation framework that consumes less memory and can be deployed locally. The designed framework combines one fixed Stable Diffusion backbone with dynamically switchable LoRA adapters that enable not only multiple multiple task-specific generation modes generation scenarios but also achieve them in a single pipeline. Moreover, the framework adopts a streaming inference approach that enhances interactivity by providing the generated result part by part progressively to the end-user. The main goal of this paper is to show the feasibility of diffusion generative models being deployed locally in memory-efficient and interactive ways without cloud dependency.

RELATED WORK

Recently, diffusion-based generative models have been widely adopted because of their capacity to generate realistic images via denoising diffusion.

The Denoising Diffusion Probabilistic Models (DDPMs) model tackled the issue of realistic imaging generation but incurred large amounts of computations. Latent diffusion models, like Stable Diffusion, overcame these drawbacks by representing diffusion in a latent space to optimize memory usage and inference time. Nevertheless, these advancements have not made it possible to apply diffusion models for interactive local applications because of hardware requirements.

However, to make up for the expense associated with fine-tuning large generative models, some methods have been proposed for parameter-efficient adaptation. Low Rank Adaptation (LoRA) allows multiple task-specific generation modes adaptation using trainable adapters added to frozen multiple task-specific generation modes models. This method costs less in terms of training overheads compared to the expense associated with fine

tuning. Current research works mainly focus on measuring the training overheads or task accuracies associated with the use of LoRA. However, the dynamics associated with the usage of LoRA at runtime, including system-level aspects when dynamically changing LoRA at runtime, have been less considered.

Several methods have investigated controllability and better structure in diffusion-based image generation. Techniques in this regard, such as those collected under ControlNet, condition the diffusion models on auxiliary inputs such as edge maps or segmentation masks for improved spatial consistency. These, while effective, increase the system complexity and rely on additional structured inputs, not always available for text-driven or local deployment scenarios. Alternative approaches adopt prompt engineering and appropriate task-specific inference configurations to guide the generation process without resorting to the addition of external conditioning networks.

From a systems point of view, prior work explored optimization methodologies such as fast sampling schedulers, memory-efficient attention mechanisms, and model offloading strategies that reduce inference time and peak memory consumption. While such methods improve performance in absolute terms, the techniques do not eliminate perceived latency in an interactive application where responsiveness and early feedback are important. Streaming-based inference has been proposed in related domains for improving user experience based on progressive output delivery, although its integration with multi-task diffusion pipelines remains limited.

In summary, the current state of research is able to tackle diffusion modeling, parameter-light adaptation adaptation, and inference optimization relatively independently. Yet, there is a significant gap in systems research for integrated solutions that encompass local deployment, task adaptation with adaptation time evolution, and interactive inference on a single resource-light system level. The proposed system is an extension based on the existing research on a common diffusion backbone with dynamic LoRA adapter switching and streaming inference for image synthesis on resource-light systems.

SYSTEM ARCHITECTURE

The designed system, Spatiora, is intended to be a lightweight and modular framework that serves as a diffusion-based image generation platform. The diffusion image generation architecture is designed to prioritize minimizing memory usage while also allowing for multiple task-specific generation modes image generation in a single unified process. This is done by combining a diffusion backbone that has been pretrained and using dynamically switched Low-Rank Adaptation (LoRA) adapters.

1. System Overview

The architecture is based on the centralized inference pattern where the same model of the Stable Diffusion type is used as the base model for all supported tasks. Unlike other models where multiple models are required to support various tasks, Spatiora requires the on-demand loading of LoRA adapters to adjust the base model to the required tasks of generation, like the visualization of the interior of an environment as well as the generation of the 2D floor plan.

On a high level, there are three functional layers:

1. Core Diffusion Engine - The pretrained Stable Diffusion backbone used for generating images in the latent space.

2. Dynamic LoRA Adapter Module - Task-specific adapters for dynamically altering the behavior of a pre-training model.

3. SII - Streaming Inference Interface, which enables progressive inference output for improved perceived responsiveness.

2. Base Diffusion Model

The heart of the setup is the locally stored, pre-trained Stable Diffusion Model which is run in inference mode. The base-model is always set to frozen state during execution and is shared among all tasks to prevent the unnecessary recreation of large parameters of the base-model. This diffusion process, happening in the latent space, is computationally cheaper compared to pixel space diffusion and maintains the same level of quality of the output.

3. Dynamic LoRA Adapter Integration

To support task specialization within a fixed memory, LoRA adapters are used in the system. Each LoRA adapter represents task-specific changes to the basic diffusion model. When a specific generation task is considered, a particular LoRA adapter is loaded, and it is applied to the common backbone architecture. If there is a request to consider a new task, the adapter is unloaded, and the new one is loaded.

4. Task-Aware Inference Configuration

There are different generation tasks with unique visualization requirements. To satisfy this, there is task-specific inference configuration to adjust for sampling rates and treatment of the input according to the specific generation task. For instance, in interior rendering with heavy visualization, there is emphasis on stylistic visualization, while in floor plan synthesis, there is emphasis on structure. The inference configuration is task-specific to make it feasible to have multiple visualization requirements catered to by only one diffusion backbone network.

5. Streaming Inference Mechanism

The user experience of generation can thus be enhanced by the incorporation of a real-time inference pipeline within Spatiora, which produces its output for generation through a streamed inference process that generates its output progressively, rather than requiring full execution to be completed. This feature facilitates early viewing of generated visual output, hence reducing latency for exploratory workflows.

6. Design Considerations

The architecture enlists and emphasizes three main guidelines for designing: memory efficiency, the ability to perform various tasks, and interactivity. Through the strategic combination of a shared diffusion backbone, dynamic LoRA adapters, and streaming inference, the model seeks a proper tradeoff between efficiency and usability. The modular design also allows future extensions, such as additional task adapters or inference optimizations, without altering the core architecture.

EXPERIMENTAL SETUP

The experimental evaluation is designed to gauge the viability of using diffusion-based generative models via a shared backbone and dynamic LoRA adapters locally. The experiments are geared toward resource efficiency and inference characteristics rather than large-scale quantitative evaluation.

A. Hardware and Execution Environment

All experiments were conducted on a local workstation representative of consumer-grade hardware. The system was equipped with a modern GPU with limited video memory, a multi-core CPU, and sufficient system memory to support local inference. The diffusion model and LoRA adapters were stored locally, and all generation was performed without reliance on external cloud services. This setup reflects the intended deployment scenario of privacy-preserving and resource-constrained environments.

B. Model Configuration

The common generative model that was used in the experiments is a Stable Diffusion generative model that works in the latent space.

Two task-specific LoRA adapters were utilized for the purpose of specializing in different image synthesis tasks. The first adapter specialized in photorealistic interior rendering, while the other specialized in 2D floor plans. The approach utilized in runtime is such that it loads just one adapter at a time.

C. Inference Settings

Image generation was done under fixed resolution and uniform sampling pattern to promote equality in comparison. Task-sensitive inference hyperparameters were used, where the strength of the guidance and number of sampling steps varied depending on the imagery or structural needs of the tasks. Interior rendering concentrated on realism and stylistic uniformity, while floor plan generation focused on clarity and structural unity. Prompts were selected to reflect the common needs of design tasks and used in all experiment runs.

D. Evaluation Metrics

The evaluation placed most of the focus on computational and qualitative characteristics. On the computational side, performance was evaluated based on total generation time, peak memory usage during inference, all factors indicating the system's readiness for being deployed on memory-constrained hardware. Qualitatively, output quality was evaluated to determine whether image outputs showed appropriate task-specific behavior in terms of visual coherence and structural consistency. Any large-scale formal perceptual or human evaluation was beyond the scope of this work and is considered future work.

E. Experimental Procedure

In each task, several requests for image generation were performed with the use of an appropriate LoRA adapter. Runs of tasks were switched to test the behavior concerning dynamic loading of adapters within a shared diffusion pipeline. Memory and generation time were monitored during execution to observe system stability and resource consumption. The experimental procedure was kept consistent across tasks to ensure comparability of results.

RESULTS AND ANALYSIS

Task Type	Time-to-First-Image (ms)	Total Time (ms)	Throughput (img/min)	Peak GPU (MB)	Peak CPU (MB)
Interior	657,766	658,719	0.09	~5,430	~4,293
Floorplan	511,296	511,296	0.12	~5,430	~4,296

TABLE 1. Pilot Latency Benchmark

Table I presents the pilot latency and resource utilization results for the two supported generation tasks: interior visualization and 2D floorplan synthesis. The results show that both tasks operate within stable GPU and CPU memory limits, with peak GPU usage remaining nearly constant across tasks due to the use of a shared diffusion backbone and dynamically loaded LoRA adapters. Floorplan generation achieves a lower time-to-first image compared to interior rendering, which can be attributed to task-specific inference configurations. Although the overall generation time is high in the pilot setup, the streaming inference mechanism enables early visual feedback, improving perceived responsiveness during execution. These observations validate the memory efficiency of the proposed architecture and demonstrate the feasibility of supporting multiple task-specific image synthesis modes within a single locally deployed diffusion pipeline.

DISCUSSION

The results from the experiments have proven that it is possible to use such an architecture to support image generation by diffusion, even in memory-limited environments, by utilizing the common architecture and loRA adapters that are loaded on the fly for adaptation purposes. This validates that parameter-efficient adaptation is applicable and useful as a method in adapting generative approaches in local environments.

The dynamic adapter LoRA switching makes it possible to have tasks that require different visual styles of generation within the system without replicating the weights of the large models. Despite the overhead caused by adapter switching, it should not have a major effect on usability owing to the high memory savings that the method generates.

The findings have shown that task and inference configurations enhance output consistency, allowing a diffusion model to handle tasks that are visually rich and structural in nature. Adding the streaming inference increases the feeling of interactivity because it provides early vision feedback. Even if the overall inference time is still high, the progressive output delivery contributes to the exploratory usage pattern because users can test the partial result before its completion. This design choice is particularly beneficial for local deployment scenarios where computational resources are limited, and inference latency cannot be eliminated.

In general, the results indicate that architectural-level or system-level optimization strategies are important for diffusion models to be applied effectively in real-world applications. Rather than focusing on improving the algorithm alone, the proposed method shows that diffusion models can be made more usable in real-world applications when proper adaptation methods combined with interactive inference strategies are applied.

The tool is developed to run locally, which limits exposure to issues of data privacy and transmission to external data sources. All inferences are done locally on user-controlled hardware, ensuring that inputs and generated outputs are not sent to third-party sources for processing. The images generated will not necessarily conform to standards of practice in the field but will serve for conceptual purposes only. The user should verify outputs for

use in practical design scenarios while considering that all outputs from this generative tool have the potential to carry biases embedded in the data used to train it.

CONCLUSION AND FUTURE WORK

This paper introduced Spatiora, an image synthesis model that uses the diffusion approach and is designed to run on resource-constrained hardware. Using LoRA adapters and the Shared Diffusion model, Spatiora was able to perform various image synthesis tasks with minimal memory costs due to redundancy and overheads that arise from having multiple models. The addition of inference via streaming inference speeds up reactivity to user inputs. This can now been observed to remain stable for memory usage while offering flexibility to perform any given task without requiring any heavy model parameters to be replicated. This shows its applicability for diffusion model execution. Future work would include optimizing inference latency with further methods, as well as scaling the evaluation to sets of prompts and user studies. Integrating support for other multiple task-specific generation modes adapters could also be considered.

REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851, 2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695, 2022.

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in Proceedings of the International Conference on Learning Representations (ICLR), 2022.

[4] L. Zhang, A. Rao, and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3836–3847, 2023.

[5] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps," Advances in Neural Information Processing Systems, vol. 35, pp. 5775–5787, 2022.

[6] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness," Advances in Neural Information Processing Systems, vol. 35, pp. 16344–16359, 2022.

[7] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," Advances in Neural Information Processing Systems, vol. 30, 2017.

[8] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 586–595, 2018.

[9] CompVis, "Stable Diffusion," GitHub repository, 2022. [Online]. diffusion Available: https://github.com/CompVis/stable

[10] Hugging Face, "Diffusers: State-of-the-art diffusion models," Documentation, 2023. [Online]. Available: https://huggingface.co/docs/diffusers