From idea to image: optimizing diffusion models for industry

Nicolás Pérez-Leis[1], David Lacalle[1] and Paula M. Castro[2]
[1] Departamento de Tecnología, Inditex, 15143 Arteixo, Spain
[2] Grupo de Tecnología Electrónica y Comunicaciones & Centro de Investigación
CITIC, Universidade da Coruña, 15008 A Coruña, Spain
Correspondence: paula.castro@udc.gal

Abstract

This study evaluates the deployment of Stable Diffusion in production, comparing FastAPI, Lit Serve, and Ray Serve on Google Kubernetes Engine for scalability, latency, and cost. It guides infrastructure decisions for AI-driven content creation while emphasizing the social, economic, and sustainability impacts of optimizing inference efficiency.

1 Summary

Artificial Intelligence (AI) has shifted from an analytical tool to a driver of content creation, with technologies that generate text, video, and images now integrated into creative processes, platforms, and business environments.

Diffusion models for text-to-image generation stand out for their versatility and openness, quickly expanding in sectors like design and marketing to explore new ideas, create promotional material, and prototype resources. In industries such as textiles, they enable customized images on demand, accelerating creativity and enhancing the digital shopping experience. However, the technical utility of these models does not eliminate the challenges associated with their practical implementation. Deploying solutions such as Stable Diffusion (SD) (Rombach et al., 2022) in production requires GPU resources, low latency, and a scalable infrastructure that maintains performance without skyrocketing costs. How the model is served can make the difference between an efficient tool and an operational bottleneck. There are currently several alternatives for serving inference models, ranging from lightweight frameworks such as FastAPI to more robust systems such as Ray Serve or Lit Serve, designed to manage distributed loads and scale dynamically. However, the lack of clear comparisons of their performance in real-world conditions makes decision-making difficult in corporate environments.

This work seeks to fill that gap through a comparative analysis of the deployment of generative diffusion models—specifically, Stable Diffusion v1.4, developed by CompVis—in a real production environment. To do this, an automated infrastructure is designed in Google Kubernetes Engine (GKE), using Terraform for cluster provisioning and Kubernetes for orchestration. On this common basis, the same model is implemented in three different environments: FastAPI, Lit Serve, and Ray Serve, with the aim of evaluating its performance, scalability, and business viability.

Beyond the technical component, this research also addresses the social and economic impact of generative AI. Tools such as SD are transforming sectors such as fashion, e-commerce, digital advertising, architecture, and visual entertainment. Optimizing their use in production, reducing costs and response times, contributes to democratizing access to these technologies and facilitates their adoption by companies of all sizes, promoting more sustainable and inclusive digitization.

In addition, the rapid evolution of new AI models and versions is generating growing demand for professionals specialized in their implementation. This context opens up opportunities to develop services tailored to different use cases, especially in regions such as Galicia, where the university ecosystem and interest in emerging technologies can catalyze innovative projects and skilled employment. Inference optimization not only improves operational efficiency but also promotes technological sustainability by reducing resource consumption in computationally intensive processes, aligning with the Sustainable Development Goals (SDGs).

2 Results and Conclusions

The motivation stems from a real need in corporate environments to harness generative AI for accelerating and enriching production processes. Companies like Inditex already use diffusion models for design sketches, product renders, and marketing materials, reducing production costs. However, moving from lab prototypes to scalable production remains a major challenge. Efficient architectures are essential to handle simultaneous image requests without performance bottlenecks. Additionally, the project promotes technological sustainability by reducing computational and energy consumption, aligning with the SDGs.

The objectives of this work were the following:

1. Analysis of the architecture of diffusion models and their requirements for deployment in production.
2. Construction of a scalable infrastructure with Terraform and configuration of services in a GKE cluster.
3. Implementation of the model on three platforms (FastAPI, Lit Serve, and Ray Serve) while keeping resources and the environment constant.
4. Identification of advantages, limitations, and recommended use cases based on the analysis of results.
5. Development of a technical guide geared toward business environments as a reference for future deployments.

The study combines an experimental approach with a strategic vision to support adoption of generative AI in industry. It provides information for companies to decide how to integrate diffusion models into production, maximizing innovation, competitiveness, and sustainability.

After designing, deploying, and testing different architectures for serving SD in production, this project empirically compared three frameworks—FastAPI, LitServe, and Ray Serve—revealing distinct advantages depending on context. It wasn’t only about performance metrics but understanding real-world behavior under network, authentication, and scalability constraints. FastAPI proved simple, flexible, and ideal for prototypes, though limited by its manual configuration needs. LitServe achieved the best performance, efficiently handling GPU usage but lacking multi-endpoint support. Ray Serve offered the most balanced solution, excelling in scalability and hybrid deployments, though with a steep learning curve. Choosing among them depends on priorities: speed and control (FastAPI), raw performance (LitServe), or scalability (Ray Serve). Beyond technical findings, the project emphasizes sustainability—showing that optimized infrastructure reduces latency, resource waste, and energy costs, leading to more efficient and responsible AI deployment.

References

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & 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).