Zen Engine: High-Performance Inference

Zen Research Authors
Zen Research DAO
Zoo Labs Inc (501(c)(3) Non-Profit)
San Francisco, California, USA
dev@hanzo.ai
+1 (913) 777-4443

September 2025

Abstract

Production-grade inference achieving 44K tokens/sec with OpenAI-compatible APIs.

1 Introduction

Production-grade inference achieving 44K tokens/sec with OpenAI-compatible APIs.

1.1 Key Features

- 44K tokens/sec on M3 Max (Apple Silicon)
- OpenAI-compatible REST API
- PyTorch, MLX, and GGUF format support
- Multi-backend: CUDA, Metal, CPU

2 Technical Specifications

Parameter	Value
Throughput (M3 Max)	44K tokens/sec
Throughput (RTX 4090)	28K tokens/sec
Latency (first token)	¡10ms
Formats	PyTorch, MLX, GGUF
Backends	CUDA, Metal, CPU
API	OpenAI-compatible REST

Table 1: Technical specifications

3 Zen AI Ecosystem

This is part of the complete Zen AI hypermodal ecosystem:
Language Models: zen-nano-0.6b, zen-eco-4b-instruct, zen-eco-4b-thinking, zen-agent-4b
3D & World: zen-3d, zen-voyager, zen-world

Video: zen-director-5b, zen-video, zen-video-i2v
Audio: zen-musician-7b, zen-foley
Infrastructure: Zen Gym (training), Zen Engine (inference)

4 Conclusion

Zen Engine delivers production-grade inference with 44K tokens/sec throughput and sub-10ms latency.

Acknowledgments

We thank the open-source community and our upstream contributors.