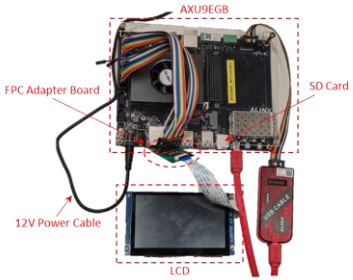
[Public]

# Diffuser: An FPGA Accelerator for Diffusion Model with Dynamic Step Quantization

Zheng Pengwei, Tang Shidi, Zhang Chuanzhao
Southeast University

OpenHW2024

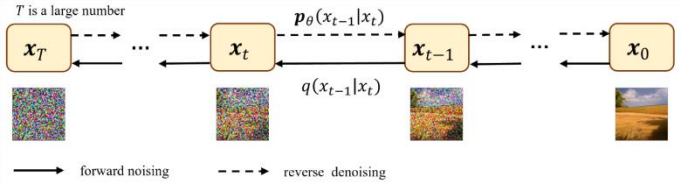
AMD




*On board test by AXU9EGB*

## INTRODUCTION

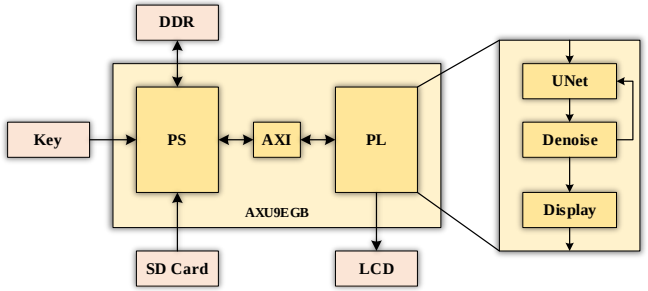
Denoising diffusion generative model, also known as diffusion model, has gained a great success in various AIGC tasks including image generation, molecular design and so on.




### Typical application case

The diffusion model has **higher accuracy compared to previous deep learning models** by modeling the probability of data distribution. However, diffusion models have been criticized for their extremely slow inference speed, as reverse denoising is a long iterative and irregular process.

**Key technical issues addressed:**
1. Diffuser is an FPGA accelerator used for unconditional diffusion models, enabling **fast image generation at the edge**.
2. At the level of model quantization, **a dynamic step size quantization method** is proposed to independently quantify the data at each time step and complete hardware friendly optimization processing based on this.
3. At the architectural level, the challenge of long iterations in diffusion models is addressed through **reusable computing units and group parallel computing**.




### System Architecture

Processing System (PS) End:
1. Initializing the system;
2. Loading model parameters from the SD card;
3. Interacting with DDR;
4. Receiving key trigger signals

Programmable Logic (PL) End:
1. Accelerator Design for Diffusion Model;
2. RTL design of ResBlock, Attention, Interpolation, Convolution and other modules;
3. Display the output result on the LCD screen;

## CREATIVE DESING

## RESULT

This work has completed the hardware acceleration design of DDIM (Denoising Diffusion Implicit Model) with dynamic step size quantization, achieving **fast image generation** at the edge.

The system utilizes the parallel computing characteristics of FPGA to significantly accelerate the inference process of diffusion models and reduce power consumption, thereby achieving **efficient model deployment**.

Our proposed dynamic step size quantization method has **lower accuracy loss** compared to traditional PTQ4DM. This work (200MHz) generates a 32 × 32 resolution image in 63 seconds, which is **3.54 times faster** than the inference time of Raspberry Pi 4 (1.5GHz) in 230 seconds and has lower energy consumption

At present, unconditional image generation with a resolution of 32 × 32 has been achieved, and improvements can be made to **higher resolutions and conditional image generation** in the future.