Jeongwoo Choi

Personal Site | LinkedIn | Google Scholar

Location: Seoul, South Korea
Email: jeongwoo.choi@yonsei.ac.kr
Mobile: +82 10 2758 8201

Introduction

I am a Ph.D. student at Yonsei University, advised by Prof. Bumsub Ham. My research focuses on **efficient generative AI** with practical impact on inference speed, memory footprint, and deployment.

Education

Yonsei University Seoul, Korea
Ph.D. Student in the School of Electrical and Electronic Engineering *Mar 2024 – Present*
Computer Vision Lab (Advisor: Prof. Bumsub Ham)

Yonsei University Seoul, Korea
B.S. in the School of Electrical and Electronic Engineering *Mar 2020 – Feb 2024*
Relevant Coursework: Application Programming, Deep Learning Lab

Publications

* Equal contribution

Accepted

Relational Feature Caching for Accelerating Diffusion Transformers 2026
Byunggwan Son*, Jeimin Jeon*, **Jeongwoo Choi***, and Bumsub Ham
*International Conference on Learning Representations **(ICLR)***

AccuQuant: Simulating Multiple Denoising Steps for Quantizing Diffusion Models 2025
Seunghoon Lee*, **Jeongwoo Choi***, Byunggwan Son, Jaehyeon Moon, Jeimin Jeon, and Bumsub Ham
*In Conference on Neural Information Processing Systems **(NeurIPS, poster)***

Under Review

Q-SAM: Quantizing Segment Anything Models 2025
Seunghoon Lee, **Jeongwoo Choi**, Yura Seo, and Bumsub Ham
Under review

Research Experience

Video Diffusion Model Dynamic Quantization Sep 2025 – Present
Research Collaboration with Samsung Electronics

- Proposed a novel framework combining dynamic quantization with feature caching
- Developed an mixed-precision strategy with two-axis granularity across motion dynamics and modality
- Achieved a 1.97x inference speedup (reducing latency from 123.45s to 62.58s) for 129-frame video generation

Image Diffusion Model Quantization Mar 2024 – Feb 2025
Research Collaboration with Samsung Electronics

- Developed novel PTQ techniques to mitigate accumulated quantization errors in multi-step diffusion processes.
- Successfully compressed models to INT4, achieving a 75% reduction in size while maintaining high visual fidelity
- Utilized AIMET torch to deploy models via ONNX for hardware-level evaluation.

Projects

DITTO: Doodle to Image TranslaTiOn May – Jun 2023
Personal project, Github

- Developed a ControlNet-based web application for real-time sketch-to-image translation
- Enhanced prompt alignment and visual fidelity by fine-tuning ControlNet on the SBU Caption dataset

Music Generation from Incomplete MIDI Sequence Jan – May 2023
Collaboration with POZALabs, Github

- Developed deep learning models to reconstruct complete melodies from incomplete MIDI sequences
- Proposed the Musical Similarity Index Measure (MSIM), a novel metric for evaluating melody reconstruction quality

Awards & Honors

Silver Prize, *32nd Samsung Humantech Paper Award* Jan 2026

Patents

Domestic

Neural Network-Based Image Denoising
KR 10-2025-0054097, Apr. 2025

Skills and Interests

Languages	:	Korean (native), English (fluent)
ML Stack	:	PyTorch, CUDA, Diffusion Models, Transformers
Systems	:	Quantization, Feature Caching, Inference Optimization, GPU Acceleration
Tools	:	Git, Docker, Conda, Linux, AIMET, ONNX

Teaching Experience

Data structure and Algorithms	Fall 2025
Deep Learning Lab.	Spring 2024, 2025
Introduction Artificial Intelligence	Fall 2024
Engineering Information Processing	Fall 2023