

# WANG ZHENG

(+86)13931805565 • neowang311@gmail.com • cheeseofneo.github.io

## EDUCATION

### The University of Tokyo – Tokyo, Japan

Oct 2024 - Mar 2025

#### USTEP Program, Dept. of Electrical Engineering and Information System

- *Thesis*: Robust Low-bits Diffusion Model deployed on Computer-in-Memory Device
- *Selective Course*: Multimodal Intelligent System, Distributed Programming, Data Compression in Quantum Computing, Trustworthy AI Software Systems

### Tsinghua University – Beijing, China

Sep 2022 - Jun 2025

#### M.Eng., Dept. of Electrical Engineering

3.82/4.0 GPA

- *Thesis*: Research on a Fast Active Microscope Autofocus System Based on Off-Axis Illumination
- *Selective Course*: Advanced Signal Processing, Bayesian Statistics, Introduction to Neural Network Quantization

### Tianjin University – Tianjin, China

Sep 2018 - Jun 2022

#### B.Eng., Dept. of Electrical Engineering

3.72/4.0 GPA

- *Awards*: Membership of China Instrument and Control Society(2022), ICM (Interdisciplinary Contest in Modeling) Finalist(Top 1%, 2021), Merit Student(2021), Andon Trust Scholarship(2020), Liu Bao Scholarship(2019)
- *Thesis*: Research on Few-shot 3D Point Cloud Segmentation via Sample-Aware Data Augmentation
- *Selective Course*: Discrete Math, Theory of Computing, Deep Learning, Machine Learning, Pattern Recognition

## HONORS

ICASSP 2026 Reviewer(2025), NVIDIA AI-Agent Group Member(2024), Tsinghua College Scholarship(2024)

## PUBLICATIONS

**Wang, Z.**, Ma, R., Zeng, C., Liu, L., Li, X., Wang, X., & He, B. A fast and precise autofocus method using linear array CCD. In Optical Metrology and Inspection for Industrial Applications XI (Vol. 13241, pp. 446-456). SPIE.

Li, Z., Su, Y., Yang, R., Xie, C., **Wang, Z.**, Xie, Z., ... & Yang, H. (2025). Quantization meets reasoning: Exploring llm low-bit quantization degradation for mathematical reasoning. arXiv preprint arXiv:2501.03035.

Xiong, H., Zhang, J., **Wang, Z.**, Pan, T., & Hu, Q. (2025). VividTalker: A modular framework for expressive 3D talking avatars with controllable gaze and blink. Manuscript submitted for presentation at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2026).

## RESEARCH EXPERIENCE

### High-resolution Image Generation via Next-frequency Prediction

April 2025- Present

Tsinghua, to be submitted to **CVPR 2026**

- Explain the regressive property of the visual autoregressive model via the Fourier Domain Analysis and model the generation process as a conditional distribution for unified analysis.
- Design an image generation architecture that reconstructs components in a bidirectional way between high and low frequency, which significantly improve generation efficiency.
- Design a learnable frequency-domain partitioning mechanism, introducing the partition vector as a class token into the Transformer to predict the central frequency of the next band.
- Introduce a pretrained Transformer as a guidance module to steer the reverse posterior sampling process of the model to achieve high2low frequency component reconstruction.

### A Modular Framework For 3D Talking Avatars With Controllable Gaze And Blink

Mar 2025- Sep 2025

Tsinghua, submitted to **ICASSP 2026**

- Developed a modular pipeline integrating diffusion-based portrait synthesis, multilingual text-to-speech (TTS), 2D-to-3D lifting, and audio-driven animation, achieving 79% efficiency improvement over existing methods.
- Designed a physiologically-grounded gaze-and-blink enhancement module simulating human-like eye dynamics via coordinated saccadic movements, spontaneous blinking (15–26 BPM), and head–eye compensation.
- Conducted extensive experiments on 40 multilingual sequences, demonstrating +59% gaze naturalness, +128% blink realism, and +53% overall user preference compared with concrete baseline - SadTalker (by Z.W. et al.), while maintaining real-time performance and cross-lingual compatibility.

## **Robust Low-bits Denoising Diffusion Models deployed on Compute-in-Memory Device**

Oct 2024-Mar 2025

Visiting Student at **UTokyo**, supervised by **Prof. Ken TAKEUCHI**

- Accelerated the diffusion process by adopting an aggressive quantization scheme (W1A1) while simultaneously guiding the model to learn more generalizable image representations.
- Investigated the role of prompts under different noise conditions in Bidirectional Diffusion Models (BiDM). Designed experiments to decompose the latent noise space using SVD, enabling fine-grained redistribution of basic noise components and a clearer understanding of prompt–noise interaction.
- Explored whether the representation ability of LDMs correlates with the perceptual quality of synthesized images. Employed masking-based learning strategies to assess how partial observation affects latent space representation and final image fidelity.

## **Exploring LLM Low-Bit Quantization Degradation for Mathematical Reasoning**

Dec 2024-Feb 2025

Research Assistant at **HKPolyU**, supervised by **Prof. Yang Hongxia**,

- Proposed a novel framework for analyzing and mitigating quantization-induced reasoning errors in LLMs, which is mainly composed by step-aligned measurement suite, hierarchical (Conceptual, Method, Execution, Reasoning) error taxonomy and a compact “Silver Bullet” dataset enabling rapid recovery of math reasoning performance.
- Developed an automated chain-of-thought error–analysis pipeline (judge ensemble and light human audit) that attains 97.2% labeling accuracy on 9,908 failure cases, enabling fine-grained, reproducible attribution by error type and first faulty step.
- Achieved full-precision level accuracy restoration with only 332 samples and 3–5 minutes of training on a single GPU without access to pretraining data.

## **A fast and precise autofocus method using linear array CCD**

June 2023-Oct 2024

Tsinghua, supervised by **Prof. Wang Xiaohao**

- Proposed a centroid extraction algorithm at sub-pixel level based on multiscale feature extraction, and fitted linear mapping using sliding finite impulse response filter. The final autofocus efficiency outperformed about 100 times than conventional gradient-based methods
- Developed a focus calibration algorithm, which conducts weighted fusion to the results of several expert algorithms.

## **INTERNSHIP EXPERIENCE**

### **Baidu, Beijing China: AI Research Intern, Team Apollo**

July 2024- Oct 2024

- For multisensor perception model, proposed an infra of model fusion based on shared weights and backbone fusion. Also conducted a cross training strategy for multi-objection and utilize distillation to compensate accuracy.
- For bev2instance module in bev perception model, conducted novel optimizations of the self-attention and deformable attention modules based on precision descent and infra compression and proved efficient in training accuracy.

### **Momenta, Shenzhen China: System Architecture Intern, Team Middleware of System**

March 2023-July 2023

- Designed the timestamp synchronization architecture of ADAS system, and constructed simulated signal generator in **signal simulation platform**, which involved validating signal pathways and creating abnormal signals.
- Developed an testbench-**board verification platform** based on SOMEIP about resource scheduling of hardware, the signal accuracy and communication correctness on a high-concurrency scene.

## **OTHER EXPERIENCE**

### **Interdisciplinary Contest in Modeling 2021, Finalist Award(top 1%):**

Jan 2020 - Mar 2020

- Proposed **PRE(Page-Rank Entropy) model** vector network of musicians and quantify influence of musicians by weights of nodes at the time-span and genre-span perspectives.
- Constructed local and global similarity evaluation model based on cosine similarity of multi-dimensional eigen vectors extracted by *PCA*. Utilized k-means clustering method to classify significance of eigens between and within genres.
- Adopted *sliding window automatic regression model* as time series forecasting on main eigens to predict the long and short-term developing trend of musical genres.

## **ADDITIONAL INFORMATION**

**Language:** Chinese(Native), English(Fluent), Japanese(Fluent)

**Programming:** Python, C, C++, MATLAB/Simulink, CUDA, Verilog