

HONGJIE WANG

Department of Electrical and Computer Engineering, Princeton University, NJ, USA
+1 609-480-4317 | [Personal Website Link](#) | e: hongjiewang09@gmail.com

EDUCATION

Princeton University

Princeton, NJ, USA

Ph.D. candidate in Electrical and Computer Engineering

Sep 2023 – Jul 2025

M.A. in Electrical and Computer Engineering

Sep 2021 – Aug 2023

- GPA: 4.00/4.00; admitted with First-Year Fellowship in Natural Sciences and Engineering
- Outstanding Reviewer of CVPR 2025

Peking University

Beijing, China

B.S. in Physics with honor

Sep 2017 – Jul 2021

- GPA: 3.84/4.00 (rank top 5%); admitted based on performance on Chinese Physics Olympiad (20/300000)
- Selected awards: Excellent Graduate of Peking University and Beijing (top 5%), National Scholarship (top 1%), May 4th Scholarship at Peking University (top 5%), First Prize in China Undergraduate Mathematical Contest in Modeling
- Selected to UGVR Program at Stanford University (top 0.5%, based on outstanding research performance)

PUBLICATIONS

1. **Hongjie Wang**, Niraj K. Jha, “LinMU: Multimodal Understanding Made Linear”, *Transactions on Machine Learning Research*.
2. **Hongjie Wang**, Chih-Yao Ma, Yen-Cheng Liu, Ji Hou, Tao Xu, Jialiang Wang, Felix Juefei-Xu, Yaqiao Luo, Peizhao Zhang, Tingbo Hou, Peter Vajda, Xiaoliang Dai, Niraj K. Jha, “LinGen-Uni: A Universal Linear-Complexity Framework for High-Resolution Minute-Length Text-to-Video Generation”, Under review at *International Journal of Computer Vision*.
3. **Hongjie Wang**, Chih-Yao Ma, Yen-Cheng Liu, Ji Hou, Tao Xu, Jialiang Wang, Felix Juefei-Xu, Yaqiao Luo, Peizhao Zhang, Tingbo Hou, Peter Vajda, Niraj K. Jha, Xiaoliang Dai, “LinGen: Towards High-Resolution Minute-Length Text-to-Video Generation with Linear Computational Complexity”, Accepted to *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2025*.
4. **Hongjie Wang**, Difan Liu, Yan Kang, Yijun Li, Zhe Lin, Niraj K. Jha, Yuchen Liu, “Attention-Driven Training-Free Efficiency Enhancement of Diffusion Models”, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024*.
5. **Hongjie Wang**, Bhishma Dedhia, Niraj K. Jha, “Zero-TPrune: Zero-Shot Token Pruning through Leveraging of the Attention Graph in Pre-Trained Transformers”, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024*.
6. Haitong Li, Wei-Chen Chen, Akash Levy, Ching-Hua Wang, **Hongjie Wang**, Po-Han Chen, Weier Wan, Win-San Khwa, Harry Chuang, Y.-D. Chih, Meng-Fan Chang, H.-S. Philip Wong, Priyanka Raina. "SAPIENS: A 64-kb RRAM-Based Non-Volatile Associative Memory for One-Shot Learning and Inference at the Edge." *IEEE Transactions on Electron Devices* (2021).
7. Haitong Li, Wei-Chen Chen, Akash Levy, Ching-Hua Wang, **Hongjie Wang**, Po-Han Chen, Weier Wan, H.-S. Philip Wong, Priyanka Raina. "One-Shot Learning with Memory-Augmented Neural Networks Using a 64-kbit, 118 GOPS/W RRAM-Based Non-Volatile Associative Memory" *IEEE Symposia on VLSI Technology and Circuits (VLSI) 2021*.
8. **Hongjie Wang**, Yang Zhao, Chaojian Li, Yue Wang, Yingyan Lin. "A New MRAM-based Process In-Memory Accelerator for Efficient Neural Network Training with Floating Point Precision" *IEEE International Symposium on Circuits and Systems (ISCAS) 2020*. (oral)

WORK AND RESEARCH EXPERIENCE

Google LLC (DeepMind)

Mountain View, CA, USA

Research Scientist

Dec 2025 – present

Pre-Training and Post-Training for Efficient Video Generative Models

- Explore pre-training for efficient video generative models
- Explore on-policy distillation for video generative models and world models
- Details of these projects are still confidential

Meta Platforms, Inc. (Meta Superintelligence Labs)

Research Scientist

Menlo Park, CA, USA

Jul 2025 – Dec 2025

Text-to-Image and Image-to-Image Pre-Training

- Worked on VLM-powered in-context image generation
- Explored text-to-image and image-to-image hybrid pre-training
- Performed data curation and architectural ablations
- Details of this project are still confidential

Meta Platforms, Inc. (Meta GenAI)

Research Scientist Intern mentored by Dr. Xiaoliang Dai

Menlo Park, CA, USA

May 2024 – Aug 2024

Part-Time Research Scientist Intern (remote from Princeton) mentored by Dr. Xiaoliang Dai

Sep 2024 – Dec 2024

LinGen: Towards High-Resolution Minute-Length Text-to-Video Generation with Linear Computational Complexity

- Proposed LinGen, a text-to-video generation framework that [enables photorealistic minute-length video generation with linear computational complexity](#).
- Developed the linear-complexity replacement of self-attention, MATE block, covering short-, medium-, and long-range correlations with a MA-branch and a TE-branch.
- Implemented the MA-branch with a bidirectional Mamba2 block equipped with a new Rotary Major Scan method and proposed review tokens to [address its adjacency preservation issue and long-range decay](#).
- Implemented the TE-branch with a TEmporal Swin Attention (TESA) block that [covers adjacent tokens and enhances temporal consistency](#) while maintaining the linear computational complexity.
- Evaluated the model on the customized prompts for human evaluation and VBench for automatic metrics calculation. Results show that [LinGen outperforms the DiT model with the similar size, while achieving up to 15× speed up](#). LinGen is also comparable to state-of-the-art commercial video generative models, including Gen-3, LumaLabs, and Kling.
- As the first author, published a paper in *CVPR 2025*.
- Developed a distillation framework to deploy the proposed MATE block on open-source video generative models, achieving up to 30.7× speed up while maintaining the model performance. As the first author, submitted a paper to *International Journal of Computer Vision (IJCV)*.

Adobe Inc. (Adobe Research)

Research Scientist Intern mentored by Dr. Yuchen Liu

San Jose, CA, USA

May 2023 – Aug 2023

Part-Time Research Scientist Intern (remote from Princeton) mentored by Dr. Yuchen Liu

Sep 2023 – Mar 2024

AT-EDM: Attention-Driven Training-Free Efficiency Enhancement of Diffusion Models

- Proposed AT-EDM, a [training-free efficiency enhancement framework](#) that leverages rich information from attention maps to accelerate pre-trained Diffusion Models (DMs).
- Generalized the Weighted Page Rank (WPR) algorithm [to make it compatible with cross-attention](#) to deduce importance score distribution of tokens.
- Developed a novel similarity-based copy method to [recover pruned tokens](#) for convolution.
- Designed a [cross-denoising-step heterogeneous pruning schedule](#) that exploits the contribution difference of different denoising steps to the generated image quality.
- Proposed a cross-denoising-step [attention-reusing](#) schedule to further reduce the generation overheads.
- Deployed the proposed AT-EDM on SD-XL and conducted both qualitative and quantitative evaluations. AT-EDM [reduces the FLOPs cost of SD-XL by 40% with negligible FID score increment and CLIP score loss](#).
- As the first author, published a paper in *CVPR 2024*.

Princeton University (Department of Electrical and Computer Engineering)

Research Assistant to Prof. Niraj K. Jha

Princeton, NJ, USA

Jan 2022 – Jul 2025

LinMU: Multimodal Understanding Made Linear

- Developed LinMU, a Vision-Language Model (VLM) that achieves [pure linear complexity](#) in input length by replacing self-attention with the M-MATE block without sacrificing model performance
- Implemented a [multi-stage distillation framework](#) to transfer knowledge from a standard Transformer-based VLM into the proposed LinMU architecture

- Evaluated LinMU distilled from NVILA and Qwen2.5-VL on MMMU, TextVQA, LongVideoBench, Video-MME, and other benchmarks. Experimental results show that LinMU matches the performance of teacher models, yet [reduces Time-To-First-Token \(TTFT\)](#) by up to 2.7× and [improves token throughput](#) by up to 9.0× on minute-length videos.
- As the first author, submitted a paper to *Transactions on Machine Learning Research (TMLR)*.

Zero-Shot Token Pruning through Leveraging of the Attention Graph in Pre-Trained Transformers

- Developed Zero-TPPrune, a [zero-shot token pruning method](#) that efficiently utilizes the feature identification ability (i.e., attention graph and embedding vectors) existing in pre-trained Transformers.
- Implemented the Weighted Page Rank (WPR) algorithm that [uses the token correlation and coupling hidden in the attention graphs](#) to deduce importance score distribution of tokens.
- Proposed the Emphasizing Informative Region (EIR) aggregation to [distinguish the informative regions](#) in different heads.
- Developed Variance Heads Filter (VHF) to prevent the detrimental effect of noisy or meaningless heads.
- Implemented a dynamic training schedule to [exploit the learning ability](#) of the proposed Zero-TPPrune method.
- Evaluated the proposed Zero-TPPrune method on multiple Transformer backbones (on ImageNet dataset). Zero-TPPrune [improves the throughput of DeiT-S](#) by 45.3% with only 0.4% accuracy loss.
- As the first author, published a paper in *CVPR 2024*.

Stanford University (Department of Electrical Engineering)

Stanford, CA, USA (remote from Beijing, China)

Research Assistant to Prof. Priyanka Raina

Jun 2020 – Sep 2020

One-Shot Learning using RRAM-Based Associative Memory

- Developed an L1-distance-based one-shot learning algorithm based on previous meta-learning methods.
- Proposed a quantization method to compress the feature vectors outputted from the pre-trained feature extractor.
- For complex datasets, developed the [adaptive assigned weights method](#) to improve the computation precision supported in the RRAM memory array significantly.
- For simple datasets, proposed [a new computation strategy to introduce computation sparsity](#) to improve the supported computation precision in the RRAM memory with high efficiency and low latency.
- Evaluated the accuracy loss reduction after applying proposed encoding methods and strategies. Compared with the straightforward encoding method, experimental results show that [accuracy loss could be reduced by up to 99%](#).
- As one of the authors, published a paper in *VLSI2021* and a paper in *IEEE Trans on Electron Devices*.

Peking University (Department of Electrical Engineering and Computer Science)

Beijing, China

Research Assistant to Prof. Huailin Liao

Jan 2020 – Apr 2021

Process In-Memory based Gated Recurrent Unit (GRU) Accelerator Design for a Low Power Hand Gesture Recognition System

- Proposed a Fast Fourier Transformation (FFT) based algorithm for feature extraction from raw radar signals.
- Developed [a fine-grained pipeline at the GRU cell level](#) for higher computation parallelism.
- Developed [a coarse-grained pipeline at the GRU layer level](#) to decouple the inputs of GRU for higher computation throughput.
- Evaluated the improvement of the proposed GRU accelerator compared with state-of-the-art works in terms of throughput and energy consumption.

Rice University (Department of Electrical and Computer Engineering)

Houston, TX, USA

Research Assistant to Prof. Yingyan Lin

Jul 2019 – Nov 2019

SOT-MRAM based Process In-Memory Accelerator for Neural Network Training with Floating-Point Precision

- Developed a new SOT-MRAM memory cell favoring more efficient PIM-based DNN training.
- Proposed [a new computation dataflow for 1-bit full addition](#) that requires fewer computing steps to finish addition operations compared with the state-of-the-art design.
- Developed [efficient designs of the dominant floating-point multiplication and addition](#) in DNN training.
- Integrated the aforementioned design to [realize a SOT-MRAM PIM based DNN training accelerator](#). Experimental results show that the proposed design offers improvement in terms of energy, latency, and area over a state-of-the-art PIM based DNN training accelerator.
- As the first author, published a paper and delivered a lecture in *ISCAS 2020*.

Peking University (Department of Electrical Engineering and Computer Science)

Beijing, China

Research Assistant to Prof. Huailin Liao

Apr 2019 – Jan 2020

Nuclear Magnetic Resonance (NMR) based Non-invasive Blood Glucose Monitor Design

- Simulated the distribution of magnetic fields induced by PCB coils based on [finite element analysis](#).
- Optimized the design of PCB coils to introduce homogeneous enough magnetic fields for NMR.
- Proposed an algorithm with DNN to identify different materials in the solution via the NMR spectrum.

Peking University (Department of Electrical Engineering and Computer Science)

Beijing, China

Research Assistant to Prof. Huailin Liao

Oct 2018 – May 2019

In-Memory Computing based Neural Network Accelerator with Phase Domain Computing

- Proposed [a concept of phase domain computing](#) to support high-efficiency dot-product for quantized neural network processing.
- Developed [a new efficient data mapping methodology](#) for phase domain computing.
- Proposed a new loss function for training and inference to improve robustness.
- Designed [a mixed-signal in-memory core](#) for efficient depth-wise separable convolution processing.
- Customized the phase domain processor under TSMC 40nm PDK to verify our methodology and architecture.

ACADEMIC SERVICES

Served (or serving) as a reviewer for

- CVPR 2024, 2025, 2026
- NeurIPS 2024, 2025, 2026
- SIGGRAPH Asia 2025
- ECCV 2024, 2026
- ICCV 2025
- BMVC 2026
- WACV 2026
- AISTATS 2025, 2026
- COCOON 2025
- IEEE Transactions on Image Processing (TIP)
- International Journal of Computer Vision (IJCV)
- IEEE Transactions on Neural Networks and Learning Systems (TNNLS)
- IEEE Internet of Things Journal (IoT-J)
- IEEE ISCAS 2021

SELECTED AWARDS AND HONORS

- | | |
|--|------|
| • Outstanding reviewer of CVPR | 2025 |
| • First-Year Fellowship in Natural Sciences and Engineering (Princeton University) | 2021 |
| • Excellent Graduate of Beijing | 2021 |
| • Excellent Graduate of Peking University | 2021 |
| • PKU Scholarship in Physics | 2020 |
| • Xiaomi Scholarship at Peking University | 2020 |
| • Student of Merits for 2019-2020 at Peking University | 2020 |
| • National Scholarship | 2019 |
| • Student of Merits for 2018-2019 at Peking University | 2019 |
| • First Prize in the 35th National University Student Physics Competition | 2018 |
| • First Prize in China Undergraduate Mathematical Contest in Modeling | 2018 |
| • May 4th Scholarship at Peking University | 2018 |
| • Student of Merits for 2017-2018 at Peking University | 2018 |
| • Gold Medal in 33rd Chinese Physics Olympiad | 2016 |