Kanghoon Yoon

ykhoon08@kaist.ac.kr | kanghoonyoon.github.io | github.com/KanghoonYoon | linkedin.com/in/yoon-kanghoon-344915217 | Kanghoon Yoon

Education

Korea Advanced Institute of Science & Technology (KAIST) *Daejeon, South Korea*
Ph.D. in Industrial & Systems Engineering *09. 2021 – 06. 2025*

- Advisor: Prof. Chanyoung Park
- Dissertation: Advancing Deep Neural Networks for Graph-Structured Data: Enhancing Representation Learning and Robustness to Data Bias.

Korea Advanced Institute of Science & Technology (KAIST) *Daejeon, South Korea*
M.S. in Industrial & Systems Engineering *09. 2019 – 08. 2021*

- Advisor: Prof. Jinkyoo Park
- Dissertation: Learning Multivariate Hawkes Process using Graph Recurrent Neural Network.

Hanyang University *Seoul, South Korea*
B.S. in Mathematics *03. 2013 – 02. 2018*

- Scholarship: Hanyang Brain Award (2017 Fall, 2018 Spring).

Research Interest

Efficient Large Language Model *-*
Accelerating Inference of LLM through Speculative Decoding *06. 2024 — Present*

- At Qualcomm U.S. and Naver Cloud corp, I optimized LLM inference speed through speculative decoding (P4, P5). I implemented draft token generation methods, which utilize datastore retrieval and small LMs, respectively.

Scene Understanding under Long-tailed Recognition *-*
Scene Graph Generation *08. 2021 — 10. 2024*

- Scene graph generation aims to detect objects and the relationships between objects, where the predicate label distribution is extremely long-tailed. I have special expertise in designing debiasing methods that address the long-tailed distribution (C2, C6, C8, C10, C12, C13).

Compositional Understanding Ability of Multimodal LLM *-*
Large and Vision Language Model *08. 2023 — Present*

- Current vision models struggle with fine-grained image comprehension like attribute and interaction between objects. I am currently working on enhancing the compositional ability of multimodal LLMs based on scene understanding module. My research expertise on compositional scene understanding (C2, C6, C8, C10, C12, C13) can advance several key applications in the field across sophisticated image retrieval/editing and advanced visual QA systems.

Working Experience

Naver Cloud *Republic of Korea*
Research intern in Efficient Large Language Model Team *02. 2025 — 05. 2025*

- I am developing state-of-the-art speculative decoding methods for efficiently serving the LLM in the cloud system. The main goal is to enhance novel and effective architecture of the draft model using efficient techniques such as quantization, pruning and distillation.

Qualcomm AI Research *San Diego, CA*
Research intern in Efficient Large Language Model Team *06. 2024 — 10. 2024*

- I developed the state-of-the-art speculative decoding methods on-device, which accelerates Llama-3.1 4.5 times faster. I applied many retrieval-based acceleration methods to speed-up the token generation of large language models without training new draft models.

Selected Publications

* represents the equally-contributed authors

(C12) Retrieval-Augmented Scene Graph Generation via Multi-Prototype Learning. *AAAI 2025*
Kanghoon Yoon, Kibum Kim, Jaehyeong Jeon, Yeonjun In, Donghyun Kim, Chanyoung Park.

(C8) LLM4SGG: Large Language Model for Weakly Supervised Scene Graph Generation. *CVPR 2024*
Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, Chanyoung Park.

(C6) Adaptive Self-training Framework for Fine-grained Scene Graph Generation. *ICLR 2024*
Kibum Kim*, Kanghoon Yoon*, Yeonjun In, Jinyoung Moon, Donghyun Kim, Chanyoung Park.

Project

(P5) Accelerating LLM Inference via Speculative Decoding *Republic of Korea*

Research Intern for the efficient LLM team at Naver Corp. *02. 2025 — Present*

- I developed the speculative sampling methods, which accelerate the LLM inference. In this project, I developed the hybrid approach that utilize the parameterized and non-parameterized draft models.

(P4) Accelerating LLM Inference via Speculative Decoding *San Diego, U.S*

Research Intern for the efficient LLM team at Qualcomm. *06. 2024 — 10. 2024*

- Improved speculative sampling methods, which accelerate the LLM inference. In this project, I developed a retrieval-based speculative decoding without fine-tuning draft model, and developed a single model-based speculative decoding method, which shows the SOTA speed-up on device.

(P3) Developing Visual Intelligence Memory via Scene Graph Generation *Daejeon, South Korea*

Project Researcher at Electronics and Telecommunications Research Institute (ETRI) *09. 2021 — 12. 2024*

- Developed a deep-learning-based scene understanding algorithm that alleviates the biased prediction problem, and published three papers (C2,C6,C8) at top conferences.

(P2) Personalized Store Coupon Issue Recommendation System Development. *Seoul, South Korea*

Project researcher at Shinhan Card *09. 2020 — 03. 2021*

- Developed a deep-learning-based scalable and personalized store coupon recommendation system for users.

(P1) Personalized User Analysis using machine learning models *Seoul, South Korea*

Project researcher at Shinhan Card *12. 2019 — 02. 2020*

- Developed a personalized user analysis algorithm by clustering users based on latent representations.

Publications

Conferences

* represents the equally-contributed authors

(C15) Is Safety Standard Same for Everyone? User-Specific Safety Evaluation of Large Language Models. *Preprint*

Yeonjun In, Wonjoong Kim, Kanghoon Yoon, Sein Kim, M Tanjim, Kibum Kim, Jinoh Oh, Chanyoung Park.

(C14) Image is All You Need: Towards Efficient and Effective Large Language Model-Based Recommender Systems. *Preprint*

Kibum Kim, Sein Kim, Hongseok Kang, Jiwan Kim, Heewong Noh, Yeonjun In, Kanghoon Yoon, Jinoh Oh, Chanyoung Park.

(C13) Weakly Supervised Video Scene Graph Generation via Natural Language Supervision . *ICLR 2025*

Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, Chanyoung Park.

(C12) Retrieval-Augmented Scene Graph Generation via Multi-Prototype Learning. *AAAI 2025*

Kanghoon Yoon, Kibum Kim, Jaehyeong Jeon, Yeonjun In, Donghyun Kim, Chanyoung Park.

(C11) Revisiting Fake News Detection: Towards Temporality-aware Evaluation by Leveraging Engagement Earliness. *WSDM 2025 (Oral)*

Junghoon Kim, Junmo Lee, Yeonjun In, Kanghoon Yoon, Chanyoung Park.

(C10) Semantic Diversity-aware Prototype-based Learning for Unbiased Scene Graph Generation. *ECCV 2024*

Jaehyeong Jeon, Kibum Kim, Kanghoon Yoon, Chanyoung Park.

(C9) Debiased Graph Poisoning Attack via Contrastive Surrogate Objective *CIKM 2024*

Kanghoon Yoon, Yeonjun In, Namkyeong Lee, Kibum Kim, Chanyoung Park.

(C8) LLM4SGG: Large Language Model for Weakly Supervised Scene Graph Generation. *CVPR 2024*

Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, Chanyoung Park.

(C7) Self-guided Robust Graph Structure Refinement. *WWW'24 (Oral)*

Yeonjun In, Kanghoon Yoon, Kibum Kim, Kijung Shin, Chanyoung Park.

(C6) Adaptive Self-training Framework for Fine-grained Scene Graph Generation. *ICLR 2024*

Kibum Kim*, Kanghoon Yoon*, Yeonjun In, Jinyoung Moon, Donghyun Kim, Chanyoung Park.

(S1) Class Label-aware Graph Anomaly Detection. *CIKM'23 (Short)*

Junghoon Kim, Yeonjun In, Kanghoon Yoon, Junmo Lee, Chanyoung Park.

(C5) Similarity Preserving Adversarial Contrastive Learning. *KDD'23*

Yeonjun In*, Kanghoon Yoon*. Chanyoung Park

(C4) Shift-Robust Molecular Relational Learning with Causal Substructure. *KDD'23*
Namkyeong Lee, Kanghoon Yoon. Gyoung S. Na, Sein Kim, Chanyoung Park

(C3) Learning Multivariate Hawkes Process via Graph Recurrent Neural Network. *KDD'23*
Kanghoon Yoon*. Youngjun Im*. Jingyu Choi, Taehwan Jeong, Jinkyoo Park.

(C2) Unbiased Heterogeneous Scene Graph Generation with Relation-aware Message Passing Neural Network. *AAAI, 2023*
Kanghoon Yoon*. Kibum Kim*. Jinyoung Moon. Chanyoung Park.

(C1) LTE4G: Long-Tail Experts for Graph Neural Networks. *CIKM, 2022*
Sukwon Yun, Kibum Ki, Kanghoon Yoon, Chanyoung Park

Invited Talks

Robust Graph Contrastive Learning *Busan, South Korea*
Korea Software Congress *12. 2023*

Heterogeneous Scene Graph Generation *Jeju, South Korea*
Korea Computer Congress *06. 2023*

Awards

Excellence Award in Poster Competition (2022, 2023)	KAIST ISysE, 2022-2023
Hanyang Brain Scholarship	Hanyang University, 2017-2018

Services

Reviewer of International Conferences
AAAI-24, KDD'24, KDD'25, AAAI-25, NeurIPS2025 *2023-2025*

Reviewer of International Journals
TKDD (2024), TPAMI (2024) *2023-2025*

Semi-supervised Classification for AI factory. *Seoul, South Korea*
LG Academy Teaching *2019-2021*