DAHEE KWON

PhD Candidate in Artificial Intelligence

+82 10 6287 7674 | daheekwon@kaist.ac.kr
Seongnam, Korea | github.com/daheekwon
Personal Website | linkedin.com/in/daheekwon-ai

RESEARCH INTEREST

Generative AI, Representation Learning, Interpretability, Vision-Language Model, Computer Vision

SUMMARY

As a PhD student at KAIST Graduate School of AI advised by Prof. Jaesik Choi, I study the internal mechanisms of deep learning models—specifically, how the structural organization of representations in vision and multimodal systems facilitates the emergence and scaling of generalizable reasoning. My current research centers on understanding the expressive capabilities and inherent failure modes of visual generative models, with the ultimate goal of enhancing their robustness and generalization in open-ended, real-world scenarios.

EDUCATION

3/2016 - 8/2020 **Yonsei University** **3.72/4.3**
Bachelor's degree in Applied Statistics

8/2020 - 8/2022 **Korea Advanced Institute of Science and Technology** **4.3/4.3**
Master of Science in Artificial Intelligence
Advisor: Jaesik Choi

8/2022 - present **Korea Advanced Institute of Science and Technology** **4.2/4.3**
PhD student in Artificial Intelligence
Advisor: Jaesik Choi

PAPERS

ICCV 2025 **Granular Concept Circuits: Toward a Fine-Grained Circuit Discovery for Concept Representations**
Dahee Kwon, Sehyun Lee and Jaesik Choi

- Keyword: Interpretability, XAI, Concept Discovery, Computer Vision
- Summary: We introduce a novel automatic circuit discovery method, called Granular Concept Circuits (GCCs), in which each circuit represents a concept relevant to a specific query.

CVPR 2025 **Enhancing Creative Generation on Stable Diffusion-based Models**
Jiyeon Han*, Dahee Kwon* and Jaesik Choi
*: Equally Contributed

- Keyword: Creative Generation, Representation Learning, Text-to-Image Generation, Diffusion
- Summary: The proposed method, C3 (Creative Concept Catalyst), a training-free approach designed to enhance creativity in Stable Diffusion-based models.

AAAI 2024 **Understanding Distributed Representations of Concepts in Deep Neural Networks without Supervision**
Wonjoon Chang*, Dahee Kwon* and Jaesik Choi
*: Equally Contributed
Oral presentation (top 9.6% of acceptance papers)

- Keyword: Image Classification, Representation Learning, Explainable AI
- Summary: The proposed method, Relaxed Decision Region (RDR), identifies the representation of the visual concepts learned by deep learning image classifier.

Under Review ICML 2026 **Breaking the Lock-in: Diversifying Text-to-Image Generation via Representation Modulation**
Dahee Kwon, Haeun Lee and Jaesik Choi

- Keyword: Generative AI, Representation Learning, Text-to-Image Generation, Diverse Image Generation, Computer Vision
- Summary: We propose DAVE, a training-free intervention that mitigates DC-homogenization-induced trajectory lock-in via selective attenuation, enhancing sample diversity.

Under Review in ICML 2026

Causal Path Tracing in Transformers
Won Jo, Dahee Kwon, Jongeun Baek, Cheongwoong Kang and Jaesik Choi

- Keyword: Causal Path Discovery, Actual Cause, LLM, Transformers
- Summary: We propose a causal path tracing framework to understand how information causally flows through the internal structures of transformers for a given decision.

Under Review in ECCV 2026

SPICE: Simple Polysemantic feature Interpretation via Clustering-based Explanations
Sehyun Lee, Dahee Kwon, Damin Lee and Jaesik Choi

- Keyword: Polysemanticity, Representation Learning, Interpretability, Vision
- Summary: We introduce effective methods to disentangle visual concept clusters across diverse vision recognition model architectures and conduct the systematic investigation of polysemanticity.

Preprint

MAC: Memory-Augmented Contrastive Learning for Time Series Anomaly Detection
Dahee Kwon, Enver Menadjiev, Qin Xie^{+}, and Jaesik Choi^{+}
$^{+}$: Corresponding Authors

- Keyword: Unsupervised Anomaly Detection, Contrastive Learning, Memory Augmentation, Multivariate Time-Series
- Summary: Memory-Augmented Contrastive (MAC) learning framework improves anomaly detection in time series data by integrating dynamic memory augmentation, detailed contrastive learning, and a unified end-to-end reconstruction process.

Preprint

Dual Masking for Domain Generalization
Dahee Kwon and Jaesik Choi

- Keyword: Domain Generalization, Representation Learning, Image Classification
- Summary: The proposed method, Dual Masking for Domain Generalization (DMDG), enhances the model to learn robust representations by attenuating non-generalizable domain-specific features.

WORK EXPERIENCE

10/2025 – Present (-04/2026)

Research Intern — **Naver Cloud**
Visual Generation Team

- Architected an automated data curation pipeline to enhance model performance in image editing scenarios.
- Implemented a Ray-based distributed framework to streamline high-throughput, large-scale data processing.

SCHOLARSHIP/AWARDS

Insung-scholar

Insung-scholarship — **2025.01**

KAIST

KAIST Breakthroughs Spring 2026 — **2026.03**

- Selected for inclusion in **KAIST Breakthroughs 50**, recognizing the paper *Granular Concept Circuits: Toward a Fine-Grained Circuit Discovery for Concept Representations* for its academic impact and innovation.

TALKS

AI EXPO KOREA 2024 Workshop

Understanding Deep Neural Networks decision-making through exploring learned features — **2024.05**

- Introducing Deep Neural Networks internal exploration framework.

KCC XAI Workshop

Analyzing the Attribute-relevant Featuremaps in Stable Diffusion Models — **2024.06**

- Explaining Text-to-Image generation diffusion model by accurately capturing attribute-relevant featuremaps.

KAIST XAI Tutorial Series — **Understanding Diffusion-based Generative Models** — **2024.11**

- Exploring Text-to-Image diffusion models by analyzing their features and internal modules to understand how modifications improve image generation quality.

Samsung AI Forum — **Enhancing Creativity in Text-to-Image Generation** — **2025.09**

- Introducing the way to unlock higher creativity in today's text-to-image diffusion frameworks.

PROJECTS

2/2021 – 6/2022 — **Development of Reinforcement Learning Technology for Optimizing and Automating Semiconductor Design** — **Samsung SAIT**

Co-worker: Sol A Kim and Ali Tousi

- Development of reinforcement learning that optimizes transistor parameters in a circuit
- Development of circuit learning (transistor placement/number) reinforcement learning that meets performance conditions
- Contribution: Implement baseline Reinforcement learning and Evolutionary algorithms including DDPG, PPO,GCN-RL and NSGA3.

9/2022 – 8/2024 — **Creative Generation in Text-to-Image Diffusion Model** — **Naver AI Lab**

Co-worker: Jiyeon Han

- Analyzing the internal representations of Text-to-Image generative diffusion model
- Editing the generated images in the Text-to-Image diffusion model.
- Contribution: As one of the main members, participate in the overall research project.

8/2024 – Present — **AI research hub** — **IITP**

Co-worker: Kyowoon Lee

- Breakthrough in Neural Scaling Law
- Develop algorithms to interpret and explain the information learned by each internal module in trained models
- Contribution: Project Leader.

SKILLS

- **Python** - Advanced
- **English** - Advanced