Hoang Phan

646-508-0988 | phanviethoang1512@gmail.com | viethoang1512.github.io | hoang-pv | VietHoang1512

Education

New York University — New York, USA
Ph.D., Data Science, Advisors: Qi Lei, Andrew Gordon Wilson. — *Sept. 2023 - Present*

Hanoi University of Science and Technology — Hanoi, Vietnam
B.Sc., Computer Science, Honor program. — *Aug. 2018 - Aug. 2022*

Experience

Meta — Seattle, USA
Research Scientist Intern — *May. 2025 - Present*

- Work on LLMs post-training.

Amazon — Palo Alto, USA
Applied Scientist Intern — *May. 2024 - Aug. 2024*

- Proposed a general framework for the data mixture selection that obtains SOTA performance, speeding up the pretraining of LLMs by 35%.

VinAI Research (now Qualcomm AI) — Hanoi, Vietnam
Research resident — *Aug. 2021 - Feb. 2023*

- Main research topics: *Robust machine learning, Transfer learning, Multi-task learning, Self-supervised learning.*
- Participated in a Smart City project that aims to build a kidnapping and unshielded truck detection system.

Data Science Laboratory — SoICT-HUST
Undergraduate research student — *Sep. 2019 - Aug. 2021*

- Main research topics: *Probabilistic inference, Continual learning.*
- Developed a Khmer processing toolkit (> 100k downloads). Used in Meta Research projects (NLLB, LASER, stopes), NVIDIA Plan2Align, FineWeb 2, Hugging Face datatrove, NAACL Workshop, Wordless and 50+ others.

Cinnamon Inc — Hanoi, Vietnam
AI research engineer — *Jan. 2020 - June. 2021*

- Worked on information extraction of business documents and meta-learning, question-answering system research.
- Developed an internal usage chatbot, integrated with Slack, in support of the Human Resource department.

Publications

- **Hoang Phan**, Sungmin Cha, Lam Tran, Qi Lei. "Toward a Holistic Approach to Continual Model Merging". *Workshop on Continual Learning in Computer Vision* at ICCV, 2025. [pdf]
- **Hoang Phan**, Lam Tran, Quyen Tran, Ngoc Tran, Tuan Truong, Qi Lei, Nhat Ho, Dinh Phung, Trung Le. "Beyond Losses Reweighting: Empowering Multi-Task Learning via the Generalization Perspective". In *International Conference on Computer Vision (ICCV)*, **Highlight presentation**, 2025. [pdf]
- Long Vuong, **Hoang Phan**, Vy Vo, Anh Bui, Thanh-Toan Do, Trung Le, Dinh Phung. "Preserving Clusters in Prompt Learning for Unsupervised Domain Adaptation". In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. [pdf]
- Chau Pham, **Hoang Phan**, David Doermann, Yunjie Tian. "Personalized Large Vision-Language Models". In *Workshop on What is Next in Multimodal Foundation Models?* at CVPR, 2025. [pdf]
- Yijun Dong*, **Hoang Phan***, Xiang Pan* and Qi Lei. "Sketchy Moment Matching: Toward Fast and Provable Data Selection for Finetuning". In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [pdf]
- **Hoang Phan***, Lam Tran*, Quyen Tran* and Trung Le. "Enhancing Domain Adaptation through Prompt Gradient Alignment". In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [pdf]
- Hao Phung, Quan Dao, Trung Dao, **Hoang Phan**, Dimitris N. Metaxas, Anh Tran. "DiMSUM: Diffusion Mamba - A Scalable and Unified Spatial-Frequency Method For Image Generation." In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [pdf]
- Yijun Dong*, Xiang Pan*, **Hoang Phan*** , Qi Lei. "Randomly Pivoted V-optimal Design: Fast Data Selection under Low Intrinsic Dimension". In *Machine Learning and Compression Workshop* at NeurIPS, 2024. [pdf]

- **Hoang Phan**, Andrew Gordon Wilson, Qi Lei. "Controllable Prompt Tuning for Balancing Group Distributional Robustness". In *International Conference on Machine Learning (ICML)*, 2024. [pdf]
- Anh Nguyen, Long Vuong, **Hoang Phan**, Toan Do, Dinh Phung, Trung Le. "Flat Seeking Bayesian Neural Networks". *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [pdf]
- **Hoang Phan**, Trung Le, Trung Phung, Anh Bui, Nhat Ho, Dinh Phung. "Global-Local Regularization Via Distributional Robustness". In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023. [pdf]
- **Hoang Phan**, Ngoc Tran, Trung Le, Toan Tran, Nhat Ho, Dinh Phung. "Stochastic Multiple Target Sampling Gradient Descent". In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. [pdf]
- **Hoang Phan***, Anh Phan*, Son Nguyen*, Linh Ngo Van, Khoat Than. "Reducing Catastrophic Forgetting in Neural Networks via Gaussian Mixture Approximation". In *Advances in Knowledge Discovery and Data Mining, PAKDD. Lecture Notes in Computer Science. Springer*, 2022. [pdf]
- **Hoang Phan**, Long Nguyen, Long Nguyen, Khanh Doan. "Matching The Statements: A Simple and Accurate Model for Key Point Analysis". In *Proceedings of the 8th Workshop on Argument Mining* at EMNLP, 2021. [pdf]
- Tam Nguyen, Quang Pham, Linh Doan, Hoang Trinh, Anh Nguyen, **Hoang Phan**. "Contrastive Learning for Natural Language-Based Vehicle Retrieval". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* at CVPR, 2021. [pdf]
- Quang Pham, **Hoang Phan**, Hoang Trinh, Anh Nguyen, Hoai Nguyen. "S-NLP at ICDAR 2021: Multimodal Emotion Recognition on Comics scenes". In *International Conference on Document Analysis and Recognition Workshops*. Technical report. [pdf]

Preprints

- **Hoang Phan**, Victor Li, Qi Lei. "Think Twice, Generate Once: Enhancing LLMs Safety via Progressive Self-Reflection". Under review. [pdf]
- **Hoang Phan** et al "Revisiting Data Mixing Through the Lens of Multi-Objective Optimization". Forever preprint. [pdf].

(∗) denotes equal contribution

Achievements

ArgMining: Key Point Analysis Shared Task, EMNLP 2021
Presented the 4^{th} solution in the workshop, hosted by IBM Research AI team.

NVIDIA AI City Challenge: Natural Language-Based Vehicle Retrieval, CVPR 2021
Got the 4^{th} place with a Siamese model capable of extracting visual concepts from natural language.

Multimodal Emotion Recognition on Comics scenes, ICDAR 2021
Won the first prize with a late fusion model, combined the processing of multiple input modalities.

Reliable Intelligence Identification on Vietnamese SNSs, VLSP 2020
Achieved a competitive performance (ranked 4^{th}) in the VLSP shared task.

Credit Scoring Challenge, Kalapa. Jsc 2020
Team ranked 1/146 with 800 contestants in constructing a banking credit scoring system.

University of Liverpool - Ion Switching, Kaggle 2020
Bronze medalist (solo) in the University of Liverpool's Institute of Ageing and Chronic Disease competition.

Practical deep learning, Vietnam Institute for Advanced Study in Mathematics 2019
Developed a deep learning-based Vietnamese spell corrector and ranked 1/50 in the final-project defense.

National/international prize

National Olympiad for Students 2017 - 2018
Vietnam Mathematical Society (VMS) Olympiad for universities and high school students
- Gold medal (2018), Silver medal (2017).

Vietnam Mathematics Olympiad (VMO) 2017 - 2018
The annual national mathematics competition for high school students
- Second prize (2018), Third prize (2017).
- International Mathematics Olympiad team selection test contestant.

International Tournament of Towns (ITOT) 2017
International mathematical Olympiad for school students
- Bronze medal (2017).

Vietnam Institute for Advanced Study in Mathematics (VIASM) Scholarship 2016 - 2018
Scholarship of the National Program for the Development of Mathematics

Vallet Scholarship 2017 - 2018
Odon Vallet's scholarships for outstanding Vietnamese students

Professional services

Reviewer at ICML, NeurIPS, ICLR, ICCV, ECCV, CVPR, WACV.

Technical skills

- **Programming Languages**: Python, C, C++, Java.
- **Frameworks**: TensorFlow, PyTorch, PyTorch Lightning, JAX, NumPy, etc.