Harshit Varma

Founding Engineer at Inception

✉ harshit@inceptionlabs.ai 🌐 hrshtv.github.io hrshtv in harshit-varma 🎓 Google Scholar

Education

Indian Institute of Technology (IIT) Bombay *Mumbai, MH, India*
Bachelor of Technology (with Honors) in **Computer Science and Engineering** *2019 – 2023*
Cumulative Performance Index (CPI) : **9.44 / 10**
Recipient of the **Research Excellence Award**
Advisor: Prof. Sunita Sarawagi

Experience

Inception *Palo Alto, CA, USA (remote)*
Founding Engineer (Research/ML) *September 2024 – Present*
Advisors: Prof. Stefano Ermon, Prof. Aditya Grover, Prof. Volodymyr Kuleshov

Google DeepMind *Bengaluru, KA, India*
Pre-Doctoral Researcher | Team: Machine Learning & Optimization *July 2023 – September 2024*
Advisors: Dr. Karthikeyan Shanmugam, Dr. Dheeraj Nagaraj, Dr. Prateek Jain

Adobe Research *Bengaluru, KA, India*
Research Intern *May 2022 – July 2022*

Publications

* denotes joint first-authors, $^\alpha$ denotes equal core contributors listed alphabetically

1. **Glauber Generative Model: Discrete Diffusion Models via Binary Classification**
   International Conference on Learning Representations (ICLR) 2025
   Harshit Varma, Dheeraj Nagaraj, Karthikeyan Shanmugam
2. **Mercury: Ultra-Fast Language Models Based on Diffusion**
   Technical Report, 2025
   Samar Khanna$^\alpha$, Siddhant Kharbanda$^\alpha$, Shufan Li$^\alpha$, Harshit Varma$^\alpha$, Eric Wang$^\alpha$, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, Stefano Ermon, Aditya Grover, Volodymyr Kuleshov
3. **Conditional Tree Matching for Inference-Time Adaptation of Tree Prediction Models**
   International Conference on Machine Learning (ICML) 2023
   Harshit Varma, Abhijeet Awasthi, Sunita Sarawagi
4. **Adversarial Training with Multiscale Boundary-Prediction DNN for Robust Topologically-Constrained Segmentation in OOD images**
   IEEE International Symposium on Biomedical Imaging (ISBI) 2023
   Harshit Varma*, Akshay Gaikwad*, Suyash Awate
5. **Deep Variational Segmentation of Topology-Constrained Object Sets, with Correlated Uncertainty Models, for Robustness to Degradations**
   IEEE International Conference on Image Processing (ICIP) 2023
   Akshay Gaikwad*, Harshit Varma*, Suyash Awate
6. **Video-based Driver Emotion Recognition using Hybrid Deep Spatio-Temporal Feature Learning**
   (Oral) SPIE Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications
   Harshit Varma, Nagarajan Ganapathy, Thomas Deserno

Patents

1. **Generating and utilizing models for long-range event relation extraction**
   US Patent App. 18/316,674
   Aparna Garimella, Anandhavelu Natarajan, Abhilasha Sancheti, Sarthak Chauhan, Prateek Agarwal, Harshit Varma

Selected Projects

Ultra-fast Diffusion LLMs (dLLMs) *(September 2024 – Present)*

Advisors: Prof. Stefano Ermon, Prof. Aditya Grover, Prof. Volodymyr Kuleshov | *Among the first* 3 *engineers* Inception

- Core contributor to the research and development of *Mercury*, the first commercial dLLM that achieves a throughput of 1100+ tokens/s – up to **10× faster** than comparable speed-optimized autoregressive baselines
- Primarily contributing to: post-training (RL, alignment), agentic capabilities, tool use, inference (novel algorithms and samplers in a production-ready inference engine), and fine-grained evaluation

Discrete Diffusion via Glauber Dynamics *(February 2024 – September 2024)*

Advisors: Dr. Karthikeyan Shanmugam, Dr. Dheeraj Nagaraj | *Accepted at ICLR 2025* Google DeepMind

- Designed a **novel** discrete diffusion framework that models the denoising process via time-dependent Glauber dynamics and showed an **exact reduction** of the learning objective to a series of **binary classification** tasks
- **Outperformed** state-of-the-art discrete diffusion baselines at language and image generation (via image tokenizers), while enabling versatile **zero-shot** control for arbitrary **text and image infilling**

Scaling Deep Retrieval to Web-scale Data *(July 2023 – September 2024)*

Advisors: Dr. Prateek Jain, Dr. Cho-Jui Hsieh, Dr. Inderjit Dhillon Google DeepMind

- Simplified the architecture and improved the **training robustness**, **convergence time**, and **numerical stability** of deep retrieval models via scaling, better loss functions, and **novel optimizer improvements**
- Surpassed existing internal methods by **10**% in recall metrics on Google's internal **web-scale** ads datasets

Tree-constrained Optimal Transport *(July 2022 – February 2023)*

Advisor: Prof. Sunita Sarawagi | *Accepted at ICML 2023* IIT Bombay

- Proposed a **novel**, **differentiable**, and **provably convergent** algorithm that extends Sinkhorn's algorithm to match trees while supporting **edge constraints**, efficiently implemented via **parallelized tensor operations**
- Applied it to the **test-time adaptation** of text-to-SQL models by representing SQL as relational algebra trees, improving performance on challenging database schemas by **up to 22**% over the base model

Robust Image Segmentation with Topology Constraints *(July 2021 – April 2023)*

Advisor: Prof. Suyash Awate | *Accepted at ISBI 2023 and ICIP 2023* IIT Bombay

- Proposed a **novel image segmentation model** that enforces certain hard **topology constraints** to preserve anatomical structures by **hierarchically predicting object boundaries** rather than pixel-wise classifications
- Preserved in-distribution performance while **minimizing the generalization gap** on OOD data, limiting the drop in Dice coefficient to **only 2.5** points compared to a reduction of more than **10** points in leading baselines

Relevant Coursework

- **Machine Learning:** Optimization in Machine Learning**, Statistical Learning & Sequential Prediction*, Advanced Machine Learning, Medical Image Computing, Natural Language Processing
- **Computer Science:** Database & Information Systems, Operating Systems, Computer Architecture, Computer Networks, Automata Theory, Compilers, Design & Analysis of Algorithms, Data Structures & Algorithms
- **Miscellaneous:** Calculus, Linear Algebra, Numerical Analysis, Economics, Game Theory & Mechanism Design

ranked **2nd, *ranked **1st** in class – *both graduate-level courses at IIT Bombay*

Teaching & Service

- **Teaching Assistant | CS726 (Advanced Machine Learning)** *(2023)*
  Selected as one of the two undergraduate TAs for a graduate-level course at IIT Bombay. Responsible for **designing weekly in-class quizzes** contributing to 15% of the final grade for a batch of **120+** students.
- **Teaching Assistant | CS215 (Data Analysis & Interpretation)** *(2021)*
  Conducted **tutorial sessions** on topics in probability and statistics for a batch of **175+** students
- Served as a **reviewer** for high-impact scientific journals such as **Nature** and **PNAS** (ORCID link) *(2025)*

Academic Achievements

- Selected to attend **Research Week with Google**, organized by Google Research *(2023)*
- Among the **13** out of **1100+** students at IIT Bombay given a chance to switch their branch/major to CS *(2020)*
- All India Rank of **833** in IIT-JEE Main and **836** in IIT-JEE Advanced among **1.2M** candidates *(2019)*
- Among the **top 300** out of **40,000+** candidates to qualify for the Indian National Chemistry Olympiad *(2019)*
- Among the **top 1%** in the National Standard Examinations in Physics and Chemistry (**NSEP** & **NSEC**) *(2018)*