Yash Jayswal

Last updated in April 2026

yashmjayswal@gmail.com | +82 010 7408 2354 | Portfolio | GitHub | Yash Jayswal | Codeforces

Education

Indian Institute of Technology Delhi *Nov 2020 – Aug 2024*
B.Tech. in Electrical Engineering

- **Coursework:** Probability & stochastic processes, Digital image processing, Machine intelligence & learning, Data structures & algorithm, Graph theory, Computer architecture.

Scholastic Achievements

- **Codeforces**: **Expert** (Rating: **1791** top 8%)(User-name: **nemesis_R**).
- **Joint Entrance Examination Advanced**: Secured **All India Rank 319** among 150,000+ students
- **Joint Entrance Examination Mains**: Secured **All India Rank 1383** among 1 Million+ students
- **Summer Undergrad Research Award (SURA)**: Conferred by Industrial R&D Unit, IIT Delhi for excellent research

Experience

Software Engineer *Suwon, South Korea*
Samsung HQ *Aug 2024 – ongoing*

- Developed a multilingual ASR model (31 languages) and optimized it for real-time streaming using quantization, chunking, K-V caching, VAD, and align-attention, achieving ≈ **45%** latency reduction from baseline.
- Optimized ASR reliability under **high-load conditions** by implementing advanced **GPU synchronization** techniques, achieving high hardware utilization and a **2%** improvement in recognition accuracy during peak traffic.
- Implemented **scalable** C++ model handling modules within a multi-user Bixby Cloud environment to enable deployment of inference-optimized **Fast Conformer**-based ASR models (both encoder–decoder and LSTM architectures).
- Improved the beam search module, optimizing memory usage by **10%**, and developed a result handling module on the ASR infrastructure in C++.
- Coordinated global Samsung R&D centers for the 2025 ASR **deployment cycle**, launching new model across **12 languages**.
- Designed and maintained an end-to-end robust **Data collection pipeline** for ASR training with automated **data collection** using yt-dlp, preprocessing, and VAD based segmentation for structured speech in **Python**.

Software Engineer Intern *Suwon, South Korea*
Samsung HQ *May 2023 – July 2023*

- Developed **Name Entity Replacer** for the Personal Database module with tolerance toward mispronounced names.
- Proposed novel phoneme similarity metric based on manner of articulation paired with Metaphone and Soundex algorithm.
- Developed Metaphone and phoneme similarity metrics, achieving **96.3%** and **98.3%** accuracy respectively for spoken name matching, reducing recognition errors by ≈ **55%**.

Projects

LLM Scaling & Systems Optimization (JAX/TPU) | Workflow *GitHub*

- Built distributed **Transformer MLPs** in **JAX** evaluating **TP, PP, and FSDP**. Hand-coded a **Mixture of Experts (MoE)** layer via **jax.shard_map** and **All-to-All** routing to bypass $O(N \times S \times D)$ **AllGather** memory bottlenecks.
- Leveraged **XLA trace profiling** on a **TPU v5e-8 cluster** to hide communication latency, achieving a **43x MoE speedup** (**73ms** → **1.7ms**) and optimizing it using **ReduceScatter** operation for matmuls to ≈ **1.1**μ**s**.

Semantic Communication | Prof. Brejesh Lall *report*

- Developed a **generative AI-based** semantic communication pipeline for image communication targeting **6G** communication.
- Researched generative models like GAN, guided stable Diffusion, Control Nets, VAE, and digital image processing techniques.
- Proposed a **guided Stable Diffusion + ControlNet** approach using image captions, edge maps, and scaled–down latent representations, achieving reduced hallucination, higher noise tolerance, and improved color retention under high compression.
- Achieved a 0.31 LPIPS score on VGG net in noisy communication channel, surpassing the previous best score by **40%**.

Panic-Driven Crowd Dynamics | Prof. Sujin B Babu *report*

- Developed strategies to reduce the death count in dense crowd panic situations at refugee camps by simulating crowd behavior.
- Researched crowd behaviour models like agent, entity, and flow-based models and simulation techniques like Euler and RK4.
- Addressed **packing problem** to prevent blowup during agents spawning, applied **numerical mollification** to avoid force blowups, tuned **RK4 parameters** for stability and optimized **obstacle shapes and configuration** to reduce fatalities.
- We analyzed factors affecting deaths and reduced the death rate by **70%** using an implementable strategy in the real world.

Technologies

Languages : Python, C, C++
Technologies: Tensorflow, Keras, Jax, Numpy, PyTorch, MATLAB, Git, Kubernetes, TensorRT
Certifications: Deep Learning Specialization (Neural Networks, Optimization, Sequence Models, CNN)