

Mohammed Faisal Parvez.

ML & EDGE AI ENGINEER · INDEPENDENT RESEARCHER — HYDERABAD, INDIA

mohammedfaisalparvez@gmail.com · +91 89772 87230 · github.com/faisalhacks · huggingface.co/Mohaaxa · linkedin.com/in/mohammed-faisal-parvez-621368241

SUMMARY

I study how vision-language models fail — and make them run where they shouldn't. Focus: **VLM reliability & calibration, extreme quantization (1-bit), edge inference, sim-to-real transfer** — benchmarking frontier VLM pipelines on cloud A100s, then compressing what survives down to CPU/ARM edge hardware.

RESEARCH FINDINGS

- Selective Refusal as Protective Calibration** — six-pipeline VLM benchmark (Qwen2.5-VL 3B/7B/72B, InternVL2-2B) on Modal A100s; refusal-calibrated family reached **~2× precision, 5–10× recall** on engaged robot-perception frames.
- Prompt Priming Masquerading as Visual Attribution** — marker × prompt ablation showed a claimed visual effect was prompt-text priming: **9.4pp from wording vs ≤0.2pp from markers**.
- Vision on 1-bit LLMs** — grafted vision onto BitNet/Falcon3 1-bit LLMs; identified & one-line-fixed a silent boundary-supervision failure; measured **~3× CPU speedup, ~6× lower memory** vs 4-bit baselines.
- The Sim-to-Real Gap in Indoor Segmentation** (manuscript in prep.) — SegFormer-B0 on a self-labeled 1,458-image real dataset: **0.676 real mIoU** vs a synthetic-trained counterpart's misleading 0.988 val mIoU.

EXPERIENCE

Cybertronix — ML & Robotics Engineer Nov 2024 — Present
Lead of a 3-person perception & edge-ML team for an autonomous indoor floor-cleaning robot; own the full stack from cloud A100 benchmarking down to ARM deployment.

Huntmetrics — Cybersecurity Trainee Aug — Sep 2023
Packet capture & protocol analysis (TCP/IP, DNS, HTTP) with Wireshark on self-hosted labs.

OPEN MODELS

Published **GPTQ & AWQ quantized Qwen2.5-1.5B** on HuggingFace (Mohaaxa) with model cards documenting calibration data & methodology; benchmarked quantized SmolVLM on Jetson Nano via llama.cpp.

SKILLS

CORE ML	Python, PyTorch, ONNX, TensorRT, HF Transformers, vLLM, GPTQ/AWQ/1-bit quantization, semantic segmentation, VLM benchmarking & ablation design
EDGE	llama.cpp, bitnet.cpp, Jetson Nano, ARM inference, FP16/INT8/1-bit
CLOUD & TOOLS	Modal (A100), Docker, Linux, Git, Bash, CVAT, Gazebo
ROBOTICS & SIM	ManiSkill, MuJoCo, PyBullet, Nav2
LANGUAGES	Python (strong), C++ (basic) · English, Hindi, Urdu, Arabic

EDUCATION

B.E. — AI & Machine Learning 2025
LORDS INSTITUTE OF ENGINEERING & TECHNOLOGY
(OSMANIA UNIVERSITY) · CGPA 8.2/10

RECOGNITION

- First Prize, College Expo** — ResNet-50/VGG16/Xception ensemble, 97.3% on CIFAR-10
- NPTEL Big Data Computing** — Elite Silver (Hadoop, Spark, Kafka)
- Responsible & Safe AI Systems** — IIIT-H / IIT-M
- Speaker**, Microsoft Student Club 2024 — Azure intro, 150+ attendees