Saidev Dhal

connect.saidev@gmail.com | saidev.codes | github.com/SkidGod4444 | linkedin.com/in/saidevdhal | x.com/SaidevDhal

Research Interests

Physical AI. Robots as AI agents that are fully local and autonomous: how they learn from people, remember their mistakes and act safely in the real world, on hardware the user owns.

Education

Galgotias University — Greater Noida, India
Bachelor of Science (Hons. with Research) in Computer Science — *Aug 2025 – May 2029*

Dr. D.Y. Patil ACS Junior College — Akurdi, Pune, India
Higher Secondary — *Sep 2023 – Mar 2025*

NRI International School — Bhadrak, Odisha, India
Schooling — *2023*

Research and Company

Founder — Jun 2026 – Present
OpenHorizon Labs, a robotics research company — *New Delhi, India*

- Study how robots learn from people, remember their mistakes and act safely in the real world, then turn that work into software that runs on the user's own hardware.
- **1 year** of research came before the company.
- **BotCortex**: lets you teach a robot by typing. The skill runs on the robot at **20 Hz** with **zero model calls**, inside bounds the model cannot loosen.
- **Sapient**: started as a side project, to have a robot that could hear, think and talk back with no uplink. Wrote the inference engine in **Rust**. It is now one **8 MB** download that runs at **27.7 tokens/s** on a Jetson AGX Thor and still runs on a Raspberry Pi 5.
- LLM, speech (Whisper, Kokoro) and vision in one binary, with no Python, Docker or CUDA. A Vulkan GPU path covers Intel, AMD and Nvidia: on a Jetson AGX Thor it took the CPU's 2.2 tokens/s to 10 with zero CUDA.
- Found by profiling that the engine dequantised the whole embedding table on every token. Making the lookup row-wise took Llama-1B on a Raspberry Pi 5 from **1.3 to 8.3 tokens/s**, and the 3B model from 0.8 to 3.4.
- Kept weights quantised on the GPU and dequantised them inside the shaders: peak memory for a 1.5B model fell from **14.7 GB to 3.6 GB**.
- Wrote a thermal governor that sheds cores from 80 °C: about **5% slower for a quarter of the cores** on a Pi 5, instead of hard throttling.
- Streaming voice loop with barge-in: perceived latency fell from about **4.4 s to 2.4 s** on a Mac.

Publications

When Does Asynchronous Action Chunking Help? Latency, Precision and Stalls for VLAs on Edge CPUs

- **Saidev Dhal**, Vinayak Pathak, Sanjana Pandey, Nalinish Ranjan. Preprint, 3 Oct 2026. Submitted to the IEEE for possible publication. openhorizon.so/research/async-chunking-edge-cpus
- A closed-form model of how often a robot stalls under synchronous and asynchronous action chunking. It identifies **3 regimes**, and measurements on a Raspberry Pi 5 match the model **within 1 percentage point**.
- Evidence that fewer stalls do not mean more completed tasks. On LIBERO-Spatial with SmolVLA, at a delay of 0.64 of a chunk, asynchronous execution stalls less but succeeds in **4 of 30** episodes against **14 of 30** for synchronous execution (paired $p = 0.006$).
- Weight precision decides the regime: on a Raspberry Pi 5, 8-bit weights cut a chunk of the SmolVLA base model from **11.9 s to 3.3 s**, with no measurable change in task success on an Apple M4 (28/50 and 26/50 against 24/50; paired $p \geq 0.29$).

Experience

Def-Space Tech Intern — Dec 2025 – Jan 2026
Bharat Space Education Research Centre (BSERC) — *Remote*

- Selected for the **DEF-SPACE Winter Internship Programme** focused on space technology, aerospace innovation and defence-oriented research.
- Worked on technical research and development tasks involving **AI systems, embedded technologies and autonomous platforms** under the mentorship of BSERC experts.
- Collaborated with interns and researchers on space-tech problem solving and system design. Aircraft design and robotics.
- Took part in training sessions covering aerospace technologies, UAV systems and emerging defence applications.

Technical Lead Intern Jan 2025 – Sep 2025
Indian Institute of Technology Madras *Chennai, India*

- Led a 6-member team developing **NextCarbon**, a blockchain-based carbon credits marketplace under the IITM CNDE lab.
- Architected and deployed **Solana smart contracts** enabling transparent carbon credit tokenisation and on-chain verification.
- Oversaw full-stack development, from the frontend dashboard to the backend APIs for carbon offset tracking.
- Nine months of product and strategy, building the MVP with the engineering team and a professor.

buildspace, S5 cohort 2024
Nights and weekends *Remote*

Community Engagement Intern Apr 2024 – Jun 2024
Ellenox *Hyderabad, India*

- Ran hackathons and technical support for the community.
- Managed the official Discord server of **2,000+** members.

Founder & Developer Apr 2020 – Dec 2021
Sputnik, a Discord moderation bot *Remote*

- Built a full-featured Discord moderation bot with 100+ automation, utility and community management commands.
- Scaled the bot to **3,000+ active servers** and **60,000+ users**.
- Implemented caching, command handling and sharding for performance across distributed servers.

Projects

EditMax | *Swift, TypeScript* Nov 2026

- A native macOS video editor your coding agent can drive: cut, grade, caption and deliver by hand, by chat, or over MCP. editmax.saidev.codes
- A full editor: tracks, trims, keyframes, multicam synced by audio, colour grading, on-device transcription and captions, 10-bit export.
- The agent works the same timeline under the same rules as the person, and one undo reverts either. Any MCP client can connect.

Annex | *Rust, QUIC, TLS 1.3, Swift, C#* Oct 2026 – Present

- Use another computer as a second display, or mirror its screen. Mac, Windows and Linux in any combination, over the network or a cable. No cloud, no accounts, no telemetry. Built with Nalinish Ranjan. **850+ downloads**. getannex.dev
- Transport: video as unreliable **QUIC** datagrams with forward error correction and BBR congestion control, falling back to TCP with TLS 1.3 when UDP is blocked. A direct cable is tried first.
- Security: an **Ed25519** identity per device, one-time 6-digit PIN pairing over **SPAKE2**, then mutually authenticated sessions with fresh keys and no resumption, so captured traffic cannot be replayed.

CodeMax | *TypeScript* Oct 2026

- One terminal for **6 coding agents** (Claude Code, Codex, OpenCode, Grok, Cursor, Antigravity CLI), on the subscriptions you already pay for. **2,100+ npm downloads**. codemax.saidev.codes
- When one agent reaches its usage limit the next resumes the same task: the plan, the step in flight and the code changes so far. Every step is recorded, so a switch is a resume, not a summary.

Sapient | *Rust* May 2026

- Make your hardware think, see, listen and speak. One **8 MB** download; **27.7 tokens/s** on a Jetson AGX Thor. sapient.openhorizon.so

UniTime | *TypeScript* Dec 2025

- A class management platform for universities, colleges and schools. Built with a team of 4.

UniCarbon | *TypeScript, Solidity* Sep 2025

- Carbon credits as tokens on Kadena EVM, with an NFT certificate for every offset. Built with a team of 3.

Prexo AI | *Turborepo, Next.js, Hono.js, Bun, React, PostgreSQL, Docker, RAG, AI SDK* Jun 2025 – Present

- AI agents for sales and support that run almost entirely on free tiers. **6 contributors, 16 GitHub stars**.
- Published reusable packages, `@prexo/ai-chat-sdk` and `@prexo/ai-chat-bot`, for drop-in chat, persistent history and vector context: **2,500+ npm downloads** between them.
- Added triggers and automations for agent workflows and multi-model orchestration via AI SDK.

IndexFlow | *TypeScript, Bun, Discord.js, Redis, PostgreSQL* Mar 2025 – Apr 2025

- Answers stuck inside Discord get a public page of their own: community owners index their server threads and make them searchable on the web. Built with 6 contributors.

Profanity Validator | *TypeScript* Oct 2024

- A lightweight SDK that stops bad-word spam in forms by validating what people type. **4,500+ npm downloads**, 9 GitHub stars.

Acter | *Next.js, React, TypeScript, AI SDK, RAG, Xata DB* Oct 2024

- Built during a hackathon. Generates customisable UI components with AI and a vector database, drawing on **3 design systems** (ShadCN, Magic UI, Aceternity UI).

x0-GPT | *Next.js, TypeScript, Redis, Vector DB, PostgreSQL* Jul 2024 – Aug 2024

- Paste a link or a PDF and ask it questions. Free for everyone. **40 GitHub stars, 7 forks**: the biggest of the open-source projects.
- Grew to **300+ active users within 2 days** of launch.
- Codebase reviewed live on YouTube by Shruti Kapoor (architecture) and Manu Arora (UI/UX).
- Sold the source code to **2 US founders**.

CloudBox | *TypeScript* Feb 2024

- Free file storage that uses Telegram underneath. No storage limit, files up to **2 GB**.

WickClone | *Python* Feb 2023

- A clone of the Discord security bot Wick, with anti-nuke and moderation. **34 GitHub stars, 24 forks**.

Talks and Workshops

Intro To Physical AI | *workshop, Greater Noida* 20 Sep 2026

- Rated **4.4 out of 5** on average by the **38 attendees** who left feedback.

Prexo AI: Agents Demo Day | *online* 15 Sep 2025

Writing

No Python, no Docker, no CUDA 16 Oct 2026

- Five months of building a Rust inference engine for the Pi inside a robot, including the bug I shipped for five weeks without noticing.

The education trap 10 Oct 2024

- The school system you grew up in wasn't designed to help you. It was designed to staff an empire, and nobody updated it.

Technical Skills

Languages: Rust, TypeScript, JavaScript, Python, Swift, C, SQL (Postgres), HTML/CSS, Go, Kotlin
Frameworks: Next.js, React, React Native, Hono.js, Node.js, Express.js, Framer Motion, Shadcn UI, TanStack Query
Robotics and ML: VLA policies (SmolVLA), LIBERO, on-device inference, Jetson AGX Thor, Raspberry Pi 5, Arduino UNO, ESP32, ESP8266
Tools: Git, Docker, Google Cloud Platform, Bun, Figma
Libraries: pandas, NumPy, Matplotlib