



Vũ Văn Hải

Full-stack Developer | AI-native development

Ho Chi Minh City, Vietnam haivv.se@gmail.com Phone: 0869 643 629 github.com/vuhai2002 vuhai.io.vn
linkedin.com/in/haivv

SUMMARY

Full-stack Developer with 3+ years of experience, backend-leaning on Node.js/TypeScript + PostgreSQL. Built and operate the entire backend of two platforms serving real businesses (200+ and 264 endpoints): payments, e-invoicing, tamper-evident audit logs and self-managed infrastructure for 12 services. Works AI-native: orchestrates 10+ AI coding agents behind my own review gate; author of an open-source AI agent (MIT) and an AI assistant that reads live CRM data for an operations team.

WORK EXPERIENCE

Full-stack Developer - CES Global

Feb 2026 - Present

TypeScript, Express 5, PostgreSQL, Drizzle ORM, React 19, TanStack, Vercel AI SDK, Docker

- Designed and built the entire backend of a live CRM and course-commerce platform (200+ endpoints, 45 tables, 18 cron jobs) plus a 25-screen admin CRM; serves the sales and accounting teams, the CES Academy digital library and ~12 landing pages.
- Automated the full order lifecycle from bank transfer to legal invoice: SePay gateway (auto-reconciling webhooks, duplicate protection, a 24-hour revival window for late settlements) and MISA meInvoice e-invoices issued right after payment with 3 deduplication layers, reconciling exactly to the amount paid.
- Built an AI assistant that reads live CRM data for the operations team on the Vercel AI SDK: an agent loop with 18 read-only tools, every calculation done in SQL so the model never invents numbers, streamed answers with a per-step lookup trace; passed 4 rounds of adversarial review before merge.
- Shipped a daily "AI insights" panel on the dashboard: 10 SQL detectors surface findings (abnormal course cancellations, repeat buyers, optimal re-contact window, traffic sources with zero conversions...), the LLM only phrases them and every number in its sentence is machine-checked; each card links straight to the filtered list.
- Run marketing email to 11,000+ recipients on a DB-configured cron engine that sends in batches behind a circuit breaker; plus ~113 operations scripts (bank reconciliation, payment rescue, bulk import) that are idempotent and dry-run by default.

Full-stack Developer (part-time) - LITEspace

Apr 2026 - Present

TypeScript, Hono 4, PostgreSQL, Drizzle ORM, BullMQ/Redis, Next.js 16, Zalo Mini App SDK

- Built the entire backend of a live platform (264 endpoints, 50 tables, 38 business modules, 6 background queues) serving the customer web app, internal CRM and Zalo Mini App at once - and delivered the Mini App end to end solo.
- Designed a tamper-evident audit log to accounting-record standards: HMAC-SHA-256 hash chain in PostgreSQL triggers, append-only enforced at the role level, a chain verification endpoint, and 3-tier retention ending in 10-year write-locked object storage.
- Resolved all concurrency conflicts without pessimistic locks: compare-and-swap via conditional UPDATE, partial unique indexes as business locks, atomic ceilings in WHERE; a 6-layer payment webhook sharing one commit path with manual confirmation.
- Modeled the pricing system: multi-rate VAT computed per tax bucket with DB-level CHECK constraints, 2-tier stacked promotions, and 3-stage price snapshots so order history never changes when configuration does; snapshot-based affiliate commissions.
- Built one auth layer for three clients with three token postures (httpOnly cookie, Bearer, Bearer + persisted session): refresh-token rotation with reuse detection and a 10-second grace window for multi-tab use, and client IP derived from the TCP peer address plus a CIDR allowlist so forged headers cannot poison the audit log or alerting.

Fresher .NET Developer - 5S Enterprise Software Solutions

Aug 2025 - Jan 2026

C#, ASP.NET (.NET Framework), SQL Server, JavaScript

- Developed and maintained business web modules on an ASP.NET (.NET Framework) + SQL Server system serving enterprise customers, working directly in a legacy codebase.
- Wrote complex SQL queries and data-layer logic that reduced data inconsistencies; partnered with QA and business teams to analyze requirements, troubleshoot data issues and ship fixes to UAT and production.

Backend Lead (graduation capstone) - Green Shine Trading Service

Dec 2024 - May 2025

.NET 8, ASP.NET Core, EF Core, PostgreSQL, SignalR, PayOS, Firebase, Docker, GitHub Actions

- Led 2 backend engineers in a 4-person team building a laundry-service platform for a real business: 5-layer architecture, 28 controllers, 5 user roles, SignalR realtime (chat, complaints, live driver tracking).
- Integrated PayOS (HMAC-SHA256-verified webhooks), Backblaze B2, Mapbox and Firebase push notifications; containerized with Docker and automated CI/CD on GitHub Actions, cutting deploy lead time by 35%.

SELECTED PROJECTS

Zalo Agent

Open source - MIT

TypeScript, Vercel AI SDK, Hono, React 19, SQLite FTS5, Docker

- Designed and built a production AI agent on Zalo: self-written agent loop, 15 built-in tools plus tools loaded dynamically from external MCP servers, multiple accounts across 2 channels (personal account + official Bot API) in one process, an 82-endpoint dashboard; ~89,000 lines of code, 359 commits.
- Built a local RAG knowledge base: ingests txt/md/docx/xlsx/pdf, chunks on natural boundaries, searches with SQLite FTS5 (bm25) + Reciprocal Rank Fusion, with a diacritic-stripped column for correct Vietnamese matching; per-agent sources, default-deny.
- Layered prompt-injection defense for an agent that reads strangers' messages: external content wrapped as data, an SSRF guard covering redirect hops and DNS rebinding, system-prompt leak blocking on output, MCP over HTTP Streamable only with two-level default-deny; deliberately no financial tools.
- Controlled cost with real measurements: token-based context budget (one turn measured at 184,835 tokens over 8 steps), per-image tokens cut from ~12,600 to ~2,600 by reordering size preference; 17 evals on real models; per-step agent traces stored locally so user conversations never leave the host.

github.com/vuhai2002/zalo-agent

Streaming & Content Management Platform

Solo - full-stack

Next.js 16, Hono 4, PostgreSQL, Drizzle, Electron, ffmpeg, Backblaze B2, Docker

- Built the entire video/audio streaming platform solo (public site + CMS in one Next.js container via host routing, 81-endpoint Hono backend) and an Electron desktop app running native ffmpeg to publish 300 MB - 2 GB lectures beyond the browser's ffmpeg-wasm limit; nearly 2 TB of media managed on object storage behind a CDN.
- Cut subtitle content search from 54 seconds to ~160 ms by switching the retrieval unit to sliding-window segments over 1.8 million cues (PostgreSQL FTS + GIN), reaching Recall@1 92% / MRR 0.94 with a repeatable eval suite.
- Designed dynamic RBAC with 3 content privacy levels and step-up authentication for 9 sensitive actions; direct-to-object-storage uploads through crash-safe record-first upload sessions; SEO/GEO with VideoObject JSON-LD including transcripts so search engines and AI assistants can cite the content.

AI Translator

Chrome Extension - 359 users, 5.0 stars

JavaScript, Chrome MV3, Express 5, Docker

- Built and maintained for 18 months an AI translation extension with 359 users and 5.0 stars (9 reviews) on the Chrome Web Store: hover, selection and context-menu translation in 50+ languages, with two translation paths tuned for speed vs quality.
- Wrote the Express 5 proxy backend that keeps API keys server-side, fails over to a backup provider, batch-translates with bisection on errors, and rate-limits on IP + User-Agent + install token; deployed as a hardened Docker container behind Cloudflare + Caddy.

chromewebstore.google.com/detail/hkelkkhcmkhhbpenoogdeomannakpilk

INFRASTRUCTURE & DEVOPS

- Operate 12 production services under one deployment standard: multi-stage Docker images, non-root, cap_drop ALL, read-only rootfs, loopback-bound containers behind Caddy, healthcheck-gated idempotent deploys; packaged as a reusable tool.
- Hardened a production Ubuntu 24.04 VPS holding real business data with 17 idempotent scripts + 6 systemd units and 10 defense layers (firewall limited to CDN IP ranges, key-only SSH on a custom port, escalating fail2ban bans, 14 kernel parameters, unattended patching + livepatch); full rebuild from scratch in 30-45 minutes.
- Built an encrypted PostgreSQL backup pipeline (age-encrypted before leaving the host, 3-tier retention on object storage, dead-man switch when backups stop) with real-time alerts and a daily heartbeat designed to detect when monitoring itself stops working.

TECHNICAL SKILLS

AI / LLM: Agentic coding, multi-agent orchestration, RAG, tool calling, MCP, Vercel AI SDK, agent evals, prompt-injection defense
Languages: TypeScript, JavaScript, C#, SQL, Bash
Frontend: React, Next.js, TanStack Query/Router, Tailwind CSS, Electron, Zalo Mini App
Backend: Node.js, Hono, Express, ASP.NET Core, REST APIs, SignalR, BullMQ
Database: PostgreSQL, SQL Server, SQLite (FTS5), Redis, Drizzle ORM, Entity Framework Core
DevOps: Docker, GitHub Actions CI/CD, Ubuntu hardening (UFW, fail2ban, systemd), Caddy, Cloudflare, S3-compatible storage
Testing: Vitest, Node.js test runner, PostgreSQL integration testing, agent evaluations

EDUCATION

FPT University, Ho Chi Minh City - Bachelor of Software Engineering (Sep 2020 - May 2025). 30% scholarship awarded for entrance-exam performance.

CERTIFICATES & AWARDS

- Certificate of Excellence - Entrepreneurial Competition, FPT University (Fall 2024)
- IELTS 5.5