Albert Martinez

albertmzb@gmail.com
(213) 633-7005
Website: albrt.io
GitHub: albertdbio
LinkedIn: albertmzb

Summary

ML Systems & Research Engineer — Diffusion, Retrieval, Agents
Research through production: improved commanded motion 9× in a one-step video-diffusion model; built a valuation model across a $407M inventory; shipped hybrid neural search and durable computer-use agents.

Selected ML Research & Systems

Real-Time Video Diffusion — motion collapse in one-step distillation
PyTorch, H100, diffusion distillation (DMD/SGMD) — write-up: albrt.io/blog/rtv-motion-collapse 2026

- **Produced the first one-step checkpoint in the project to execute commanded object transport**, increasing measured motion **9.1× (0.29→2.65)** at 7.0/10 coherence with zero architecture changes
- **Isolated motion collapse to the Stage-3 distillation objective** — not the decoder, initialization, or implementation — for ∼$12 total GPU compute using fixed-latent stage swaps and matched tests against official checkpoints instead of a multi-GPU retrain
- Replaced DMD's stop-gradient objective with **score-gradient matching** and stabilized training by identifying and normalizing a $\sim 10^{10}$ dynamic range in the Fisher weighting
- Built independent motion and coherence metrics using tracked displacement, camera compensation, identity preservation, and flow-warp residuals; synthetic tests and blind evaluations yielded **Spearman** $\rho \approx 0.61$–0.70

Mochi — consumer AI agent with its own computer
Effect-TS, SolidStart, E2B Firecracker microVMs, KasmVNC, CDP, Postgres, Twilio 2026

- Built a live SMS-native agent product where each user receives a persistent cloud computer — an ephemeral Firecracker microVM running a full desktop and hardened browser, streamed live while the agent works
- Designed a **host-inversion architecture** with durable state sync and fenced task execution, allowing worker VMs to be destroyed and recreated without losing conversations, duplicating work, or dropping tasks

Athena — multi-model consensus review control plane
Effect v4, E2B sandboxes, multi-lineage model routing 2026

- Built a multi-model code-review control plane that gives independent model families access to the actual repository inside isolated microVMs; provider fallbacks return partial reviews instead of hanging

Experience

Cloudstore Manhattan Beach, CA
Full Stack Engineer | ML Systems, Search & Agent Infrastructure Feb. 2023 - Present

- Built the used-equipment **valuation model** from scratch: 31 per-category log-price ridge regressions replacing a median baseline — $R^2 = 0.94$**, MAPE 5.3%** — preventing a **$6M margin overstatement** ($8.1M → $2.1M)
- **Exposed a 2.3× overfitting gap** hidden by in-sample evaluation (5.7% vs. 13.2% out-of-sample MAPE) and established forward-chained, time-based cross-validation across a $407M inventory universe
- Designed the search ranking layer on **Vespa**: hybrid rank profiles combining BM25, E5 dense embeddings (384-dim bfloat16, computed on-GPU at index time), and **ColBERT late-interaction** max-sim reranking; later added a SPLADE sparse-retrieval cluster with a sigmoid-normalized hybrid score and shipped an ablation profile to A/B it against production traffic

- **Cut write-back for a 9,000-row LLM extraction backlog from ∼9 hours to 30–45 minutes** with bounded concurrency; used schema-constrained Vertex AI Batch generation for ∼50% lower token cost
- **Restored access to 8/8 blocked sources** by tracing a persistent Cloudflare 403 to JA3/TLS fingerprinting rather than IP reputation; a separate layered Distil/Imperva mitigation cut extraction latency from **21.0s to 12.6s per record**
- Scaled the crawling fleet to **261 active dealer sites (∼70K listings)** across 8+ discovery strategies (sitemap, JSON API, Algolia, Azure Search, and an LLM-driven fallback for JS-heavy long-tail sites)
- Prevented **silent data and index corruption** with a mass-delisting circuit breaker, a stable-key classifier separating moved listings from sold ones, and dual-index writes whose split failure semantics let a mirror outage self-heal
- **Reduced computer-use token consumption ∼19×**, making long autonomous sessions viable by converting raw base64 screenshots into native model image blocks while driving a live screen-shared Linux desktop

IndiePrinting Los Angeles, CA
Full Stack Engineer Jun. 2019 - Feb. 2023

- Architected frontend (Next.js, TypeScript, Redux) and backend (Node.js, Express, MongoDB) for a custom ecommerce platform; extended Reaction Commerce with a custom pricing algorithm on the GraphQL API and built the AWS/Docker deployment pipeline

Skills

- **ML / Research:** PyTorch, diffusion distillation (DMD, score-gradient matching), CoTracker3, DINOv2, RAFT, ridge regression, forward-chained cross-validation, metric design & calibration
- **Search / Retrieval:** Vespa, BM25, ColBERT late-interaction, E5 & SPLADE embeddings, hybrid ranking, RAG
- **Agents / LLM:** Claude computer-use, tool calling, E2B Firecracker microVMs, CDP browser automation, Patchright, context-window engineering, multi-model consensus review
- **Systems:** TypeScript, Effect-TS, Node.js, Python, PostgreSQL, MongoDB, Redis, ClickHouse, React, SolidStart, Expo
- **Infrastructure:** GCP (Cloud Functions, Task Queues), AWS, Hetzner, RunPod (H100/4090), Docker, Caddy, Vercel

Education

UC Davis Davis, CA
Computer Science, Bachelor of Science 2015