Mikhail Balkhaev

ML Engineer / Generative AI Engineer / Product Developer

Moscow, Russia m.balkhaev@gmail.com +7 966 006-60-68 t.me/balkhaev balkhaev.com



ML engineer working on generative AI: image and video diffusion, speech, LLM agents and the GPU infrastructure that runs them. I build products end to end, from the model pipeline and its evaluation to the backend and interface: in 2026 that meant media generation services, data platforms, a CRM and an Unreal Engine game. Before 2024 I spent 13 years in web development, including VK (Delivery Club) and Yandex (Yandex Eats).

Experience

AIGF, luv.club, adorely.co, sinly.io — Mar 2026 – present

Generative AI Engineer

- AI companion platforms (18+) with image, video and 3D generation, chat and voice: a multi-tenant Hono backend, Redis/BullMQ workers and RunPod GPU jobs with idempotent submission.
- Identity-consistent photos with Qwen Image 2.1: an edit-consistency LoRA, up to four reference images and an automatic redraw when the face drifts.
- Video: text- and image-to-video with two-pass sampling and RIFE interpolation to 48 fps; face and full-character swap with MiniMax H3 and swap LoRAs.
- 3D stage to video: the reference render conditions only the first 5 of 20 sampling steps, which keeps the camera and composition while the model produces natural motion.
- Companion memory: LLM fact extraction with corrections, hybrid retrieval over OpenAI embeddings in pgvector, lexical reranking and LLM relevance scoring.
- Age-safety checks for generated media: face detection, a ViT age classifier and a logistic probe on CLIP ViT-B/16 embeddings trained on 9,574 labelled person crops. On held-out data, correct refusals rose from 283 to 350 of 400 and false refusals fell from 26 to 3 of 302.
- Tool-using Grok agents: a studio director that sets up shoots and reviews results, and a 3D scene editor with validated, revision-checked edits; REST and MCP API with OAuth 2.1.
- Voice: self-hosted Kokoro-82M TTS on CPU (ONNX, 54 voices) and Whisper recognition.

Funcorp, Cyprus — Nov 2024 – Mar 2026

Senior Product Developer

- Started on the main product, then built new AI products from scratch.
- Diffusion pipelines for images and video (Wan 2.2, LTX, SD, SDXL) in ComfyUI and DiffSynth-Studio; dataset collection and LoRA training.
- Text-to-speech: voice design and fine-tuning the model for natural-sounding speech.
- LLM features with Qwen3.5-9B, ChatGPT and Grok; RAG on Qdrant.
- Data and backend: Kafka, ClickHouse, PostgreSQL, Redis, Next.js, Hono, FFmpeg.

AI and media

Studio, film.balkhaev.com

AI film production: a Claude producer agent turns an idea into a script and staged shots, renders reference passes in Blender over MCP and generates the shots with H3. A take is accepted only after frame decoding, a vision-model review and a speech-transcription check.

MediaPipes, mediapipes.balkhaev.com

Image, video and speech generation on ComfyUI and vLLM: Qwen Image 2.1 INT8 with up to ten references, tiled upscaling to 8 MP, Higgs TTS 3 (4B). INT8 attention for H3 cut sampling from 13.0 to 6.8 s/step (a swap job from 129 to 80 s) and shipped after identity, lip-sync and LPIPS metrics and a blind review.

Persons

Character identity service: seven reference views and a character sheet per character, CPU face checks with YuNet and 512-d AuraFace embeddings, storyboard stills with up to four characters, and a PostgreSQL job queue with SKIP LOCKED and per-person fairness.

Modeler, modeler.balkhaev.com

Characters and previz in Blender: image-to-3D with TRELLIS.2 and Pixal3D to textured GLB, text-to-motion on a 63-bone rig with foot-contact quality gates, Qwen-guided facial reconstruction.

Montage

Video understanding: Qwen3-VL-4B frame descriptions, word-aligned WhisperX transcripts and PySceneDetect cuts in one timeline, with models loaded in turn to fit one GPU; video-to-video recasting with H3.

Inference

GPU scheduler for a home RTX 5080 shared by ComfyUI, local LLMs and Blender: admission by measured RAM and VRAM peaks, GPU sharing only after three comparable solo runs, warm-model residency and budgets that grow after out-of-memory errors.

Local LLM

Qwen3.8-Flash-Next (125B GGUF) on one RTX 5080 through Strata, with experts spread across VRAM, RAM and SSD, a 262k context and about 100 tokens/s; an OpenAI- and Anthropic-compatible gateway with multi-account routing.

Data products

ggpick.club

Game discovery in Russian, English and Chinese: Steam and community data in PostgreSQL, hybrid search on multilingual-e5 embeddings in Qdrant with cross-encoder reranking and a fallback when models are down, transliteration and keyboard-layout correction. 200+ unique visitors a day.

Polywin

Polymarket research: typed market-data clients with WebSocket order-book reconciliation, an opportunity scanner that accounts for depth and fees, LLM news-to-market reasoning with validated quotes, and forecasts scored by Brier score and calibration. A signal needs 30+ resolved markets and a statistically positive edge.

AI News

Radar of new models and AI tools from 44 sources (GitHub, Hugging Face, arXiv, Reddit, X): grounded Russian summaries with quote checks, hybrid semantic search with a CPU cross-encoder, growth signals from stars and commits. Moving blocking work off the event loop cut p99 pauses from 585 to 62 ms.

FastNews

RSS and X news desk: safe article extraction, URL and text deduplication, LLM enrichment (entities, topics, claims, forecasts) with schema and quote validation, and a versioned API that feeds Polywin.

Soclogin

Service for my own Instagram, X and Avito accounts: an AES-256-GCM credential vault, idempotent browser operations on Playwright, tenant-scoped feeds for the news products and MCP diagnostics.

Sites and client work

ZemStroy CRM

Sales CRM for a land development company: amoCRM, Avito and CIAN integrations on Redis/BullMQ, call transcription with Whisper and LLM call scoring, a first-response bot with versioned prompts, Telegram alerts for new leads and SLA breaches.

Altai Land, алтайземля.рф

Land-development site with an interactive 3D valley on Three.js: streamed scene packages for mobile (first load from 7.2 to 3.7 MB), a chunked forest of 79,125 trees (5.9M to 1.2M triangles), a 568-plot masterplan shared by 2D and 3D, 31 indexable pages.

balkhaev.com

Falling into a Schwarzschild black hole in the browser: WebGL ray tracing in Painlevé–Gullstrand coordinates, a lensed star catalogue with Doppler and gravitational shifts, and tests against independent numerical quadrature.

Tools and games

DevHub

Development platform for me and my AI agents: dependency-aware process supervision, a worktree and release workflow gated by checks, Canvas and Kanban project knowledge served over MCP, and Coolify deployment monitoring.

Gamza

Isometric extraction autobattler on Unreal Engine 5.8 and C++: a deterministic headless simulation with hash-verified replays, tactical squad AI (hearing, flanking, suppression), a persistent living world, maps built from geodata and an automatic cinematic camera.

Habar
Pixel-art idle strategy game in TypeScript: a DOM-independent simulation, seeded shot-by-shot combat, factions, convoys and an economy, and procedural terrain painted in Web Workers.

Web development

Yandex, Moscow — Oct 2022 – Aug 2024
Frontend Developer

- Moved from VK with the Delivery Club team to the Yandex Eats advertising unit: analytics and fast testing of ad hypotheses.
- Main Yandex Eats and Yandex Vendor webviews; micro-frontend architecture.
- An ad runner game whose first partner was Haval at its Russian launch; most of the banner ads, analytics debugging and new metrics.

VK, Delivery Club — Jun 2020 – Oct 2022
Lead Developer

- Maintained delivery-club.ru, effectively full-stack: admin panels for partners and lawyers, ad formats, internal UI and API libraries.
- After an incident, built a document tool for the legal team with versioning and audit logs.

Data-screen, Moscow — Jun 2019 – Mar 2020
Full-stack Developer

- A sports streaming service built from high-level ideas with full technical ownership (Node.js, React, PostgreSQL, Redis, Docker); COVID put it on hold.

Smart Solution, Moscow — Nov 2015 – Dec 2018
Senior Frontend Developer

- Frontend for about ten high-traffic sites in the Alexa top 50: legacy rewrites, optimisation and builds; ad integrations; trained developers and reviewed code.

Oversun, Web Developer — 2014 – 2015
WordPress sites from scratch with custom modules; led a junior developer on the company website.

Ideegroupe, Sochi, Web Developer — 2013
WordPress and ModX: modules for a city news portal, SQL optimisation.

RedBox, Buryatia, HTML Developer — 2011 – 2012
Joomla and WordPress templates, jQuery.

Skills

Generative models	Qwen Image, H3, Wan 2.2, LTX, SD, SDXL, TRELLIS.2, Kokoro, Higgs TTS, Whisper, Qwen3-VL
ML engineering	LoRA training, datasets, evaluation (identity similarity, LPIPS, lip-sync, held-out sets), CLIP probes, embeddings and RAG, INT8 and GGUF inference
Serving	ComfyUI, vLLM, ONNX Runtime, RunPod, GPU scheduling, Redis/BullMQ
LLM and agents	Tool use, MCP, Claude, GPT, Grok, Qwen, pgvector, Qdrant
Backend and data	Python, TypeScript, Node.js, Bun, Hono, FastAPI, PostgreSQL, ClickHouse, Kafka, Redis, Qdrant, SQLite, Docker
Frontend and 3D	React, Next.js, Three.js, WebGL, Unreal Engine 5 (C++), Blender, Playwright

Details

Languages	Russian (native), English (upper intermediate, B2)
Education	Secondary education
Work	Full-time, part-time or project work; remote, hybrid or on-site; open to business travel and relocation within Russia.