# Mariano Rodriguez

**Lead AI/ML Researcher & Engineer | Computer Vision, Speech & Language Models**

PhD & Postdoc, ENS Paris-Saclay | WACV / CVPR author | Research shipped to production

rdguez.mariano@gmail.com • +33 7 82 58 82 37 • Paris, France • rdguez-mariano.github.io

Applied ML researcher and engineer who ships models, not notebooks. PhD and postdoc in computer vision at ENS Paris-Saclay, peer-reviewed publications at WACV and CVPR Workshop, and industry-contracted research for Huawei and Kayrros. As co-founder and CSO of **Freeda** I set the research agenda, hired and lead a team of 3 ML engineers, and own the full loop — data and annotation tooling, model training and evaluation, inference optimisation and serving — for a vision + LLM compliance product in production. Previously founding ML engineer at **Gladia** (speech) and video researcher at **Powder**. 6 years turning applied research into systems that hold up under production load and customer scrutiny.

## Research Highlights

- **Perception models.** Up to three generations of plan-understanding models at Freeda (current architectures: RF-DETR, YOLO-OBB, SegFormer), all improving on the Roboflow baseline, alongside template matching and SAM 2 / SAM 3 one-click area selection.
- **Plan-description layer.** A structured record of measurements derived from detected shapes, readable by agents and humans alike — one representation that stops every downstream step from re-deriving geometry.
- **Agentic verification.** Two iterations: retired a human-in-the-loop design that worked on paper but needed too many interventions; shipped a recipe-based agent following a human-authored roadmap, with a human asked only at final sign-off and skills that write comments and annotations directly into the delivered PDFs.
- **Document intelligence.** Freeda Assistant — MinerU OCR, contextual retrieval, hybrid search and reranking — over thousands of regulation pages; Smart Selection uses LLM contextualisation of retrieved patterns to capture every text span tied to a concept (apartments, rooms) by style and context.
- **Codegen.** Autonomous coding agents used for dynamic custom workflows, including customer-facing interactive reports built as mini-apps (single-page HTML emitted from a React + Tailwind app).
- **Prompt optimisation.** Automated prompt improvement adapted from a Microsoft method, removing manual prompt iteration as the bottleneck on quality.
- **Speech inference at commercial scale.** Rebuilt Whisper serving at Gladia (ONNX, NVIDIA Triton, flame-graph profiling): lower cost per hour of audio at better WER than the market at the time; owned speaker diarization and voice-activity detection (pyannote, silero), unlocking the meeting-transcription use case.
- **Video understanding shipped to consumers.** Event and highlight detection in gameplay streamer video at Powder; game and spoken-language identification on-device via ONNX/CoreML; the production summarization model published at CVPR Workshop 2023.
- **Contracted research.** Huawei: recurrent-CNN joint video denoising and demosaicing (published WACV 2023). Kayrros: CNN cloud detectors for satellite imagery.

## Selected Publications

**Video joint denoising and demosaicing with recurrent CNNs** — WACV 2023

**Multi-annotation attention model for video summarization** — CVPR-W 2023

**Robust estimation of local affine maps and its applications to image matching** — WACV 2020

**Covering the Space of Tilts** — SIIMS 2018

Full list: rdguez-mariano.github.io/publications

## Leadership & Experience

### Co-Founder & Chief Scientific Officer

**Nov 2024 – Present**

**Freeda, Paris**

*Own scientific strategy, research roadmap and the production AI platform for automated compliance review of pre-construction plans against building regulation.*

- **Research direction & team.** Hired and lead a team of 3 ML engineers; own hiring, the AI roadmap, the evaluation bar and the in-house-vs-foundation-model boundary.
- **Vision stack.** Three generations of plan-understanding models (RF-DETR, YOLO-OBB, SegFormer) beating vendor baselines; SAM 2 / SAM 3 interactive area selection; in-house CVAT-based annotation tool for labelling text and vision data across document folders.
- **LLM / agent stack.** Recipe-based verification agents with human sign-off and PDF-annotation skills; document RAG (MinerU OCR, hybrid retrieval, reranking); Smart Selection contextual span capture; automated prompt optimisation.
- **Serving & infra.** Self-hosted serving stack on GCP Compute Engine, NVIDIA Triton, Celery, LangChain, 4 DBs (Elasticsearch, Postgres, Redis, ChromaDB), monitoring (Grafana, Loki, Prometheus), GitHub CI/CD, auto-deploys; first working prototype ran at $<$ EUR 300/month with an NVIDIA L4 GPU; custom codegen mini-apps.
- **Production evidence.** Compliance workflows, engine and infrastructure from blank page to production in a few months, cutting manual plan-review time 3$\times$ vs market standards; 10+ paying customers within 10 months; technical basis of the EUR 3.5M (Frst, Brick & Mortar) and EUR 18.5M (Notion) rounds.

### Founder & Chief Technology Officer
**May 2024 – Nov 2024**

**Nano Quant Services, Paris**

*Sole technical founder: end-to-end ML data-and-models product for value-stock identification — Selenium acquisition, model training, assistant with hybrid search & contextual RAG, FastAPI backend, React/Node frontend, Kubernetes deployment on premise — shipped alone in 6 months.*

### Founding Machine Learning Engineer — Speech
**Jan 2023 – May 2024**

**Gladia, Paris**

*Applied research at commercial scale: speech-to-text infrastructure sold as a public API to 100+ customers (PAYG and enterprise contracts); Gladia was later acquired by OVH.*

- **Inference research with a direct margin effect.** Rebuilt Whisper for production serving (ONNX, NVIDIA Triton, flame-graph profiling): lower cost per hour of audio at better WER than the market at the time.
- **Owned a research area end to end.** Speaker diarization and voice-activity detection (pyannote, silero), which unlocked a customer segment: the meeting-transcription use case.

### Machine Learning Engineer — Video & Multimodal
**Jul 2021 – Jan 2023**

**Powder, Paris**

*Applied computer vision inside a consumer gaming product.*

- **Research shipped to consumers.** Event and highlight detection in gameplay streamer video.
- **On-device deployment.** Game identification in video and spoken-language identification exported to ONNX/CoreML.
- **Published what shipped.** Multi-annotation attention model for video summarization, CVPR Workshop 2023 — the production model, written up.
- **Research infrastructure.** Set up the team's training and evaluation stack (PyTorch Lightning, Comet/WandB, Gin, CI with reproducible experiments).

### Postdoctoral Researcher, Applied Mathematics
**Jul 2020 – Jul 2021**

**École Normale Supérieure Paris-Saclay**

*Industry-funded research contracts — research run to an external client's specification and deadline. Advisor: Gabriele Facciolo.*

- **Huawei.** Recurrent-CNN joint video denoising and demosaicing delivered under contract; published at WACV 2023.
- **Kayrros.** CNN cloud detectors for satellite imagery.

### PhD Researcher in Applied Mathematics
**Oct 2016 – Jul 2020**

**École Normale Supérieure Paris-Saclay**

*FMJH-funded doctoral research, run as an employed research position. Advisors: Jean-Michel Morel & Julie Delon.*

- **Affine-invariant image comparison.** Robust estimation of local affine maps and its applications to image matching; published at WACV 2020.
- **Geometric estimation.** Fundamental and homography matrix estimation.
- **Repeated-structure detection.** Detection of repeated objects in images.

## Technical Depth

| | |
|---|---|
| **Research** | Computer vision, NLP, speech recognition & diarization, agents/skills/RAG, video understanding, geometric estimation, optimisation, applied mathematics |
| **Modelling** | PyTorch, PyTorch Lightning, TensorFlow, Keras, OpenCV, scikit-learn, LangChain, ONNX, CoreML |
| **Serving** | GCP, AWS, Slurm, Kubernetes, Docker Swarm, NVIDIA Triton Inference Servers, FastAPI, Flask, flame-graph profiling |
| **Engineering** | Python, C++, ctypes (C++↔Python) and MEX (C++↔MATLAB) bindings, PyTest (fixtures & mocks), SQLAlchemy/Postgres, React & Node, GitHub CI, Cookiecutter, Task, Make & CMake, Unix administration, LaTeX/TikZ |
| **Scientific tooling** | Comet/WandB, Gin config, QGIS, MATLAB |
| **Languages** | Spanish (native), French (fluent), English (fluent) |

## Education

**PhD, Computer Vision & Machine Learning** — École Normale Supérieure Paris-Saclay. Advisors: Jean-Michel Morel & Julie Delon. 2016 – 2020

**MSc "Mathématiques, Vision, Apprentissage" (MVA)** — ENS Paris-Saclay. Highest Honors. 2015 – 2016

**MSc 1 "Mathématiques et Modélisation"** — Université Paris Cité. Highest Honors. 2014 – 2015

**BSc Mathematics** — Universidad de La Habana. Highest Honors; top student of the faculty. 2007 – 2011

## Awards & Fellowships

**Laureate, FMJH Doctoral Program** — Fondation Mathématique Jacques Hadamard. Three-year competitive fellowship. 2016

**Laureate, PGSM Masters Scholarship** — Paris Graduate School of Mathematics. Two-year international scholarship. 2014

**Top student, Faculty of Mathematics & Computer Science** — Awarded by the vice-chancellor, University of Havana. 2011

## Teaching

**Fourier & Hilbert Analysis** — Practical lessons, Mathematics Department, ENS Paris-Saclay. 2017 – 2019

**Differential Calculus** — Practical lessons, Mathematics Department, ENS Paris-Saclay. 2016 – 2017

**PDE, Complex Analysis, Mathematics I–III** — Lectures & practical lessons, Mathematics Department, University of Havana. 2011 – 2014