

Pablo Robin Guerrero

pablorg4@pm.me | github.com/veridisynndrome

RESEARCH INTERESTS

Representation Learning, Computational Modeling, Communication and Behavior.

EDUCATION

École Normale Supérieure

Master of Science in Cognitive Science

Paris, France

September 2026 – Present

École Polytechnique Fédérale de Lausanne

Bachelor of Science in Communication Systems

Lausanne, Switzerland

September 2022 – February 2026

Focus on Theoretical Computer Science and Machine Learning

Relevant coursework:

Machine Learning, Natural Language Processing, Algorithms, Data-intensive Systems, Computer Language Processing, Software Construction, Computer Systems, Computer Security and Privacy, Stochastic Models, Probability and Statistics.

Bachelor Project:

Efficient Deployment of Vision-Language Models on Mobile Devices: A Case Study on OnePlus 13R

PUBLICATIONS

F. Mustun, C. Semenzin, D. Rance, E. Marachlian, Z. Guillermin, A. Mancini, **P. Robin Guerrero**, I. Bouaziz, E. Fleck, N. Shashar, G. G. de Polavieja, G. Sumbre. Whistle Variability and Social Acoustic Interactions in a Small Pod of Bottlenose Dolphins. *Under Review*

F. Mustun, C. Semenzin, R. Dessì, **P. Robin Guerrero**, P. Orhan, A. Emanuelli, E. Rossi, Y. Lakretz, G. de Polavieja, G. Sumbre. OpenWhistle: A Large-Scale Longitudinal Dataset and Benchmark of Bottlenose Dolphin Vocalizations. *Under Review*

P. Robin Guerrero, Y. Pan, S. Kashyap. Efficient deployment of vision-language models on mobile devices: A case study on oneplus 13r. *arXiv preprint arXiv:2507.08505*, 2025.

EXPERIENCES

École Normale Supérieure

Research Engineer, Neural Circuit dynamics & Behaviour Laboratory

Paris, France

January 2026 - Present

Research topic:

Social and Latent Acoustic Structure in Bottlenose Dolphin Communication: Linking Whistles to Behaviour

Advisors: Prof. Germán Sumbre and Faadil Mustun

- Investigate **codebook structure** and learn acoustic representations to identify phoneme-like units and characterize the long-term organization of **bottlenose dolphin whistles**.
- Link **acoustic representations to temporal structure** and behavioural context to probe the organization of dolphin vocal sequences.

Tsinghua University

Research Intern, NICS Lab, Efficient Deep Learning Algorithm Team

Beijing, China

July 2025 - September 2025

Research topic: Model-Level Optimization for Efficient Generative Inference via Latent-Space Routing

Advisors: Prof. Yu Wang and Dr. Xuefei Ning

- Worked on **latent-space continual learning** for token-level LLM/SLM routing, enabling model adaptation without full retraining and limiting catastrophic forgetting during inference.
- Explored **DPO-based routing models** operating on latent representations.

- Conducted experimental analysis of **efficiency–performance trade-offs** of routing under constrained inference settings.
- Experimental contributions to **Road2Rome**, accepted at **NeurIPS 2025**.

École Polytechnique Fédérale de Lausanne

Research Intern, Robust Scalable Systems Software Laboratory

Lausanne, Switzerland

January 2025 - July 2025

Research topic: Efficient Deployment of Vision-Language Models on Mobile Devices

Advisors: Prof. Sanidhya Kashyap and Yueyang Pan

- Benchmarked deployment frameworks (llama.cpp, MLC-Imp, mllm) on the **OnePlus 13R** with VLMs (LLaVA-1.5 7B, MobileVLM-3B, Imp-v1.5 3B), measuring **latency, throughput, and device thermals**.
- Profiled inference workloads across **CPU, GPU, and NPU**, demonstrating **CPU saturation** and **accelerator underutilization** as the key barriers to real-time on-device generation.
- Results published in the preprint Efficient Deployment of Vision-Language Models on Mobile Devices: A Case Study on OnePlus 13R (arXiv:2507.08505), providing practical **profiling tools** and **deployment guidelines** for future mobile AI systems.

Korea Advanced Institute of Science & Technology

Research Intern, Concurrency & Parallelism Laboratory

Daejeon, South Korea

June 2024 - August 2024

Research topic: PIM-based AI Accelerator and Software Stack for Efficient LLM Inference

Advisors: Prof. Jeehoon Kang and Haechan An

- Investigated NeuPIMs architecture, achieving up to **3.0x throughput improvement** over GPU systems and **1.6x improvement** over naive NPU+PIM systems on **GPT-3 (7B)**.
- Evaluated AttAcc design, demonstrating **2.81x higher throughput** and **62.6% improved power efficiency** for **LLaMA 65B** compared to GPU-only systems.
- Reproduced experiments on PIM-based systems, uncovering limitations in multi-device scalability and proposing solutions projected to enhance performance by **25%** in multi-device environments.

TECHNICAL SKILLS

Programming Languages: Python, C/C++, Java, Scala, Kotlin

Frameworks/Libraries: PyTorch, TensorFlow, Hugging Face Transformers, SGLang

Tools: Git, Docker, CI/CD, Linux, LaTeX

Spoken Languages: French (Native), Spanish (Native), English (Full professional proficiency), Korean (Limited working proficiency)

AWARDS AND HONORS

Tsinghua University Visiting Researcher Scholarship (2025)