

Thomas Vos

Machine Learning Engineer · MSc AI student

Amsterdam, NL hi@thomasvos.ai linkedin.com/in/vos-thomas

Summary

ML engineer and MSc AI student working on interpretability and sparse attention, increasingly focused on generative modelling and flow matching. Co-author of a TMLR paper, with three years of part-time experience in production ML.

Experience

Machine Learning Engineer (freelance, part-time) 2023 – now
Foodticket · Huizen / Amsterdam, hybrid

- Deployed a production recommender on AWS trained on data from 3,000+ restaurants
- 7-month Bayesian A/B test of a price-capped filter: conversion +15% (3.9% → 4.5%), add-to-cart +18%, revenue per recommendation +6.6%
- Interpretable churn-risk model on noisy, imbalanced data, used by customer success to prioritise outreach

ML Industry Project (via UvA) 01/2023
CargoProbe · Utrecht, hybrid

- "Busyness" regression model that corrects forecast errors, cutting departure-time bias from 6.3 days to under 4 hours

Publication

A Reproducibility Study of Hierarchical Sparse Representations for Multimodal Interpretation

L. Ponjee-Dillon, **T. Vos**, M. Wilde, R. Wolting (equal contribution)

TMLR 2026 Reproducibility Certification OpenReview ↗

Reproduced Matryoshka SAEs on CLIP and found insufficient evidence for the claimed hierarchy; removed always-on neurons with a targeted L1 penalty; introduced a concept saliency method that localises SAE features at pixel level.

Research projects

SPOT: Speech-grounded Person Observation Test 2026
UvA · Interpretability & Explainability

Benchmark for whether audio-vision models bind speech to the right on-screen speaker. ImageBind reached 76% zero-shot (chance 50%); probing showed latent alignment does not imply grounding.

Sparse Attention for Whole-Slide Image Classification 2026
UvA · Deep Learning 2

Sparse attention from LLMs adapted to ViTs with custom GPU kernels. Matched ABMIL on CAMELYON16 (93.8% acc.), up to 47% faster than dense attention.

BSc thesis: Agent-based Motherese 2025
Supervised by Dr. R. G. Alhama (ILLC)

Emergent communication in referential games (PyTorch, EGG): a trained sender with untrained receivers gave higher, more stable accuracy and compositionality.

Education

MSc Artificial Intelligence, University of Amsterdam 2025 – now
Interpretability & Explainability (8.5), Deep Learning 2 (8.0), Game Theory (8.0)

BSc Artificial Intelligence, University of Amsterdam 2021 – 2025
Financial Mathematics (8.0), EEG classification (8.5), C++ Programming (8.0)

Research interests

Generative models & flow matching
Interpretability
Efficient deep learning

Skills

Machine learning
PyTorch · HuggingFace · W&B · Scikit-learn

Programming & compute
Python · C++ · SLURM/HPC · Docker

Data & cloud
SQL · AWS · Airflow · dbt · FastAPI

Certification

Hugging Face Deep RL
PyTorch, TensorFlow and OpenAI Gym

Other projects

Brain Powered
Real-time EEG classification for drone control

Covid-19 Spread in Supermarkets
Agent-based simulation of designated walking paths

Futures Trading Dashboard
Real-time aggregation across futures markets

Languages

Dutch Native
English Proficient

Interests

Alpine climbing
Making music