CAI4DSA: Collaborative Explainable neuro-symbolic AI for Decision Support Assistant

University of Florence, DINFO Dept.

Paolo Nesi, DISIT lab, UNIFI

2 October 2025

Aula Gerace, Largo Pontecorvo 3 - Edificio C, secondo piano, University of Pisa

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

Aim

- **CAI4DSA aims at exploiting AI for creating decision support systems** (DSS) which can evolve with the collaboration among and with humans.
- The **NGDSS** should be capable to react to changes in the context, interact with humans, and learn from the occurrence.
- **Symbolic and neural models** could be also shared in communities to create a global knowledge and understanding of the problems and solutions.

outcomes

- **New Algorithms or Models**: Development of novel AI algorithms or models that can enhance decision-making processes.
- **NGDSS Theories**: Contribution to the theoretical understanding of how AI can be effectively integrated into decision support systems.
- **Early Prototype** of the NGDSS, providing a proof of concept for how AI can enhance decision-making in a number of target domains

Main objectives

Obj.1. Continuous incremental **reinforced learning** to address the possibility of learning taking into account new facts and implications.

Obj.2. **human readable multidomain XAI** representations to facilitate human understanding and trust by providing explanations of the produced decisions/suggestions for a given scenario with a particular care to **explain seasonalities** of variables/conditions, contextual aspects, and the effects of inputs with respect to outputs and KPI.

Obj.3. **extract and formalize new knowledge from XAI**, results/KPIs as facts, relationships, constraints, models (causals, maths, etc.) also deciding if: they can actually drive the learning and/or decision process, for shortening the learning process, reducing features, reducing temporal span in time series analysis, increasing precision.

Obj.4. **Human NewGenDSS interaction which should be capable to work reconcile with humans**: (i) collecting information the relevance/confidence of the produced results is not satisfactory, (ii) providing motivations (in connection with KPI-Driven XAI) behind the lack of knowledge and/or data, (iii) collecting human suggestions and present some counterindications or simply demonstrate that cannot be verified, (iv) provide general learning process indicators which should give the evidence that the general knowledge / understanding is improving and its coverage, (v) provide instruments for defining scenarios.

Obj.5. **Validation of NewGenDSS in critical infrastructure** domains: mobility and transport, environment, medicine, security, complex manufacturing, etc.

main team members

DISIT LAB

- ***Paolo Nesi,*** *PO: Monthly commitment: 0,166 (project PI and CV attached)*
- *Pierfrancesco Bellini, PA: Monthly commitment 0,2 (CV attached)*

AI Lab:

- ***Paolo Frasconi,*** *PO: Monthly commitment 0,133 person/month. (CV attached)*
- *Simone Marinai, PA: Monthly commitment 0,133 person/month*
- *Marco Lippi, PA: Monthly commitment 0,133 person/month*

AI Lab

MICC:

- ***Pietro Pala,*** *PO: Monthly commitment 0,2 (CV attached)*
- *Lorenzo Seidenari, PA: Monthly commitment 0,0866*
- *Andrew Bagdanov, PA: Monthly commitment 0,0866 (CV attached)*

Main Domains vs research Groups

- **Mobility and Transport**: identification of optimal solutions (macro scale scenarios and parameters) to solve critical traffic conditions coming from the occurrence of natural and non-natural disasters, impacting on traffic flow, traffic direction, regulation of traffic controllers. **(DISIT lab, P. Nesi)**
- **Complex manufacturing**: simulation of curation processed depending on the product and placement of material in autoclaves for reducing costs and optimizing time. The model is a complex partial differential equation in turbulent conditions, which the state-of-the-art approaches use finite differences to solve in unacceptable execution time, without addressing the variation of parameters. **(DISIT Lab, P. Nesi)**
- **Medicine**: identification ablation targets in atrial fibrillation (AF): **(AI Lab, P. Frasconi).**
- **Security and computer vision:** continual learning can aid in updating DCNN models with new data inevitably alters the latent space environmental observations, surveillance. **(MICC, P. Pala)**

AI Lab

Mobility and Transport:

- **Context-Aware Retrieval Augmented Generation using Similarity Validation to handle Context Inconsistencies in Large Language Models**, E. Collini, F. Indra Kurniadi, P. Nesi, G. Pantaleo, IEEE Access, 2025. https://doi.org/10.1109/ACCESS.2025.3614553
  - **SnapAdvisor:** https://www.snap4city.org/1116 towards Agentic AI for DSS

- **Comparing Techniques for Temporal Explainable Artificial Intelligence**. Canti, E., Collini, E., Ipsaro Palesi, L. A., & Nesi, P. (2024, July). In *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)* (pp. 87-91). IEEE.
  - https://ieeexplore.ieee.org/document/10730341

- **Neuro-Symbolic Optimization of Mobility Infrastructure via Deep-RL-GNN**, E. Collini, L. A. Ipsaro Palesi, C. Lucchesi, P. Nesi, submitted to IEEE Access
  - **Evolution of: Mobility Infrastructure Optimization via Stochastic Relaxation the Traffic Flow Reconstruction**. Bilotta, S., Collini, E., Fanfani, M., & Nesi, P. (2024).
  - https://assets-eu.researchsquare.com/files/rs-5093651/v1_covered_1b2f1826-1464-4a57-9514-7810680a57c1.pdf

- **New version of the Km4City ontology** integrating detailed digital twin aspects and results of neuro-symbolic deductions.
  - In progress

Comparing Techniques for Temporal Explainable Artificial Intelligence

Objective

Evaluate **XAI methodologies** on the time series analysis in the AI models, focus is on **global XAI aspects**

Time series

With seasonality

Without short-term seasonality

With outliers

eXplainability methodologies

Shapley Values

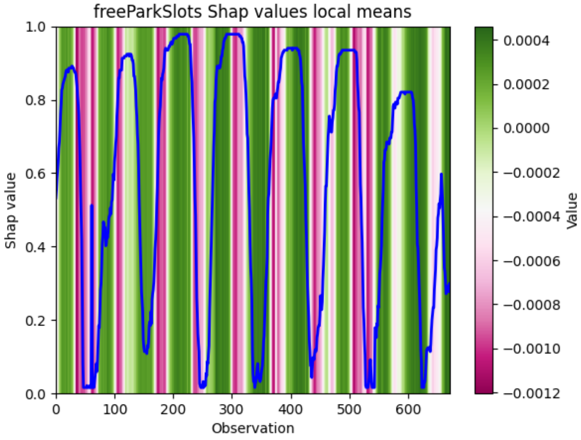
Integrated Gradients

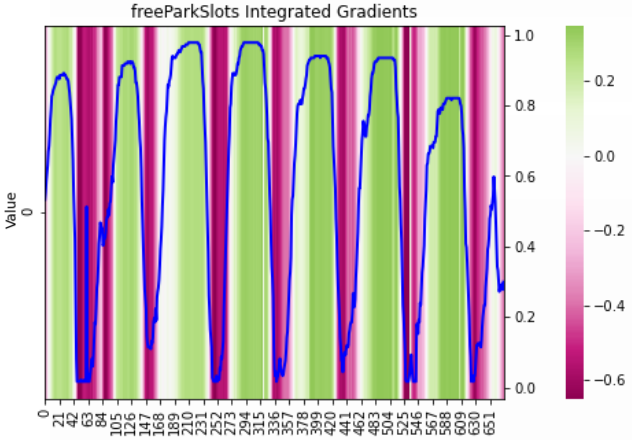
LIME

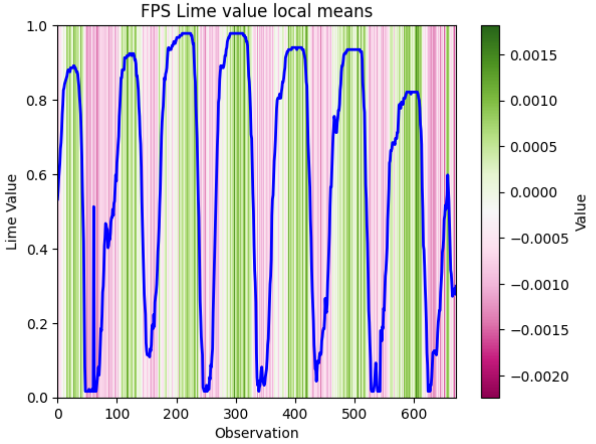
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

Experimental result

Time series with seasonality





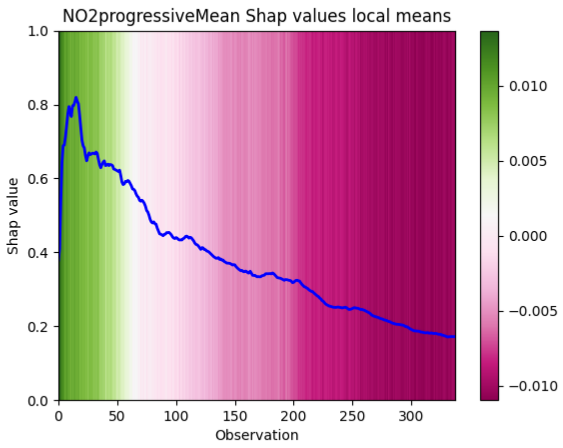


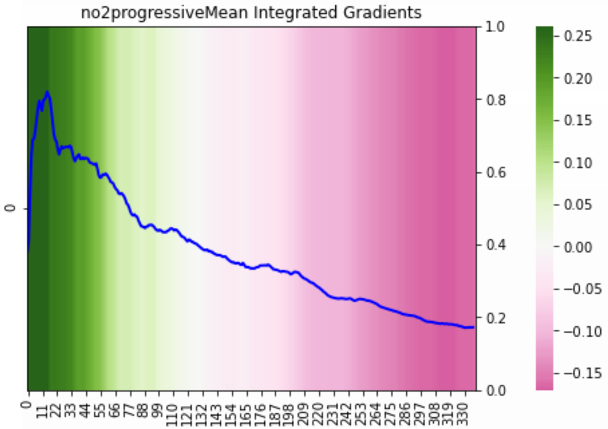
Future Artificial Intelligence Research

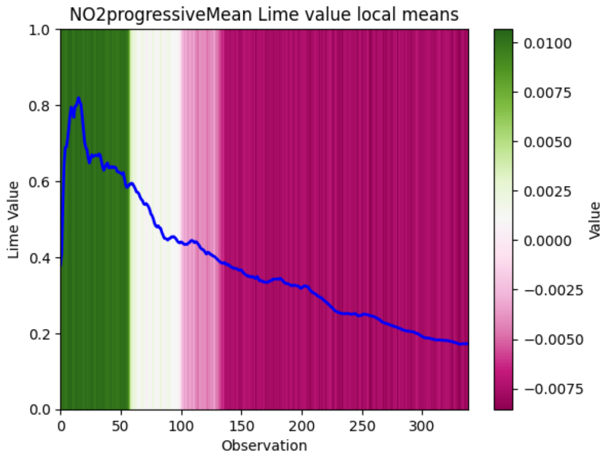
UNIVERSITÀ DEGLI STUDI FIRENZE | DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE | DISIT DISTRIBUTED SYSTEMS AND INTERNET TECHNOLOGIES LAB | SNAP4CITY

Experimental result

Time series without short-term seasonality



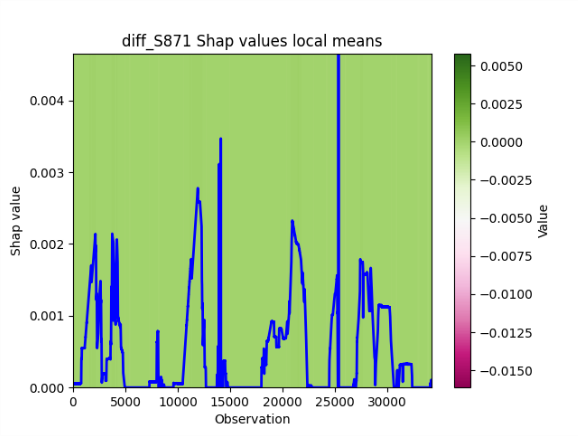


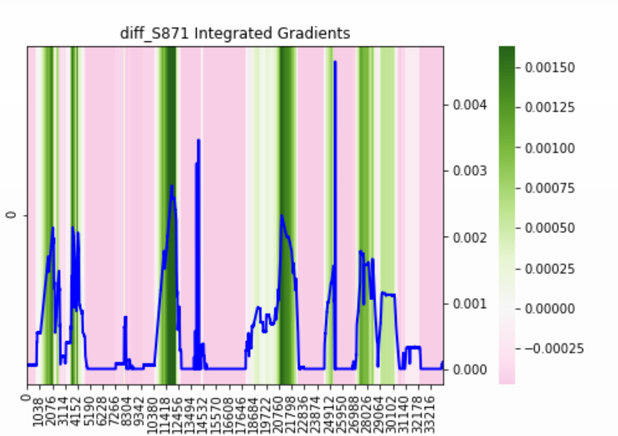


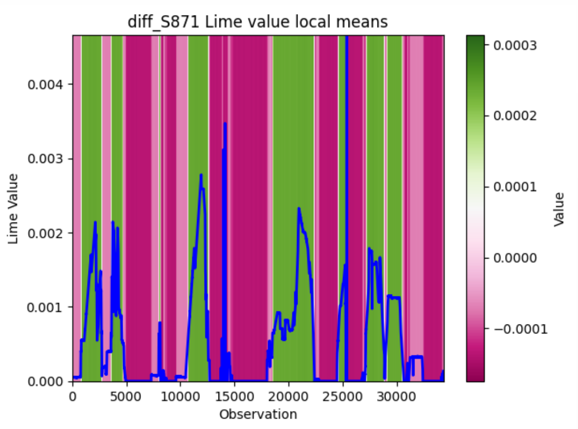
Comparing Techniques for Temporal Explainable Artificial Intelligence

Experimental result

Time series with outliers





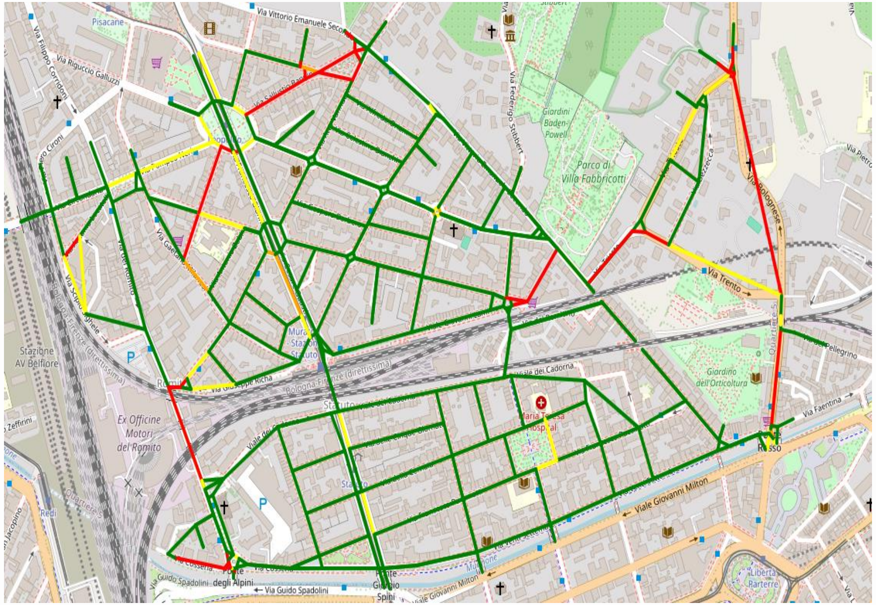


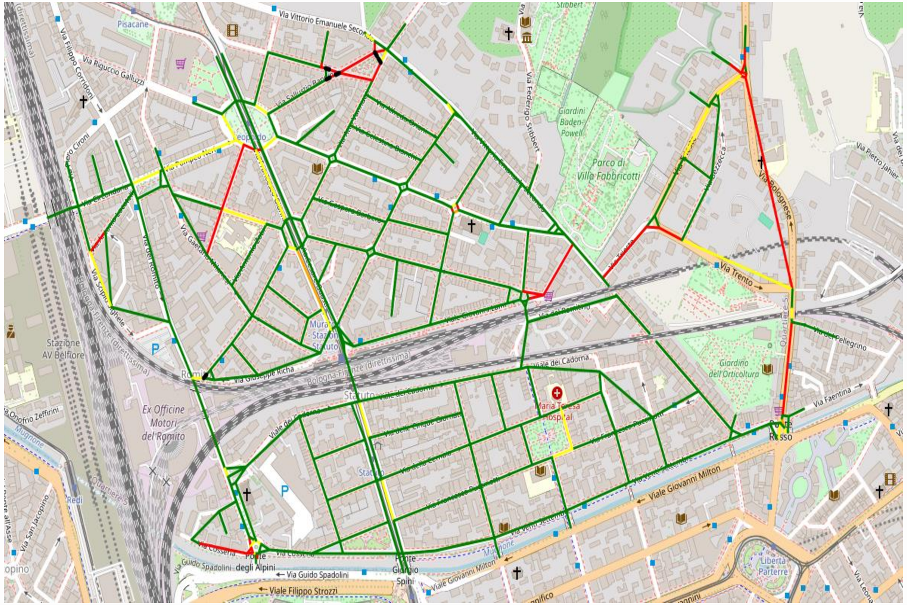
Neuro-Symbolic Optimization of Mobility Infrastructure via Deep-RL-GNN

Objectives

- **Optimization with respect to various KPIs on traffic conditions** including congestion levels, fuel consumption, CO2 emissions based on **Deep-RL-GNN**.
  - **Taking into account limitations in generating solutions**
  - **Taking into account constraints such as: (i) not modifiable road segments, (ii) traffic regulations**
  - **Generating viable road lane changes**.
- **Comparing the solution with respect to stochastic relaxation optimization approach.** The proposed solution provides
  - **better performance** with respect to stochastic relaxation optimization
  - **reduction of time and costs with respect to the state-of-the-art solution.**
- **Proposing and assessing the AI model transferability** on different (i) urban scenarios, and (ii) traffic workload conditions.

Original and optimized (with changes)



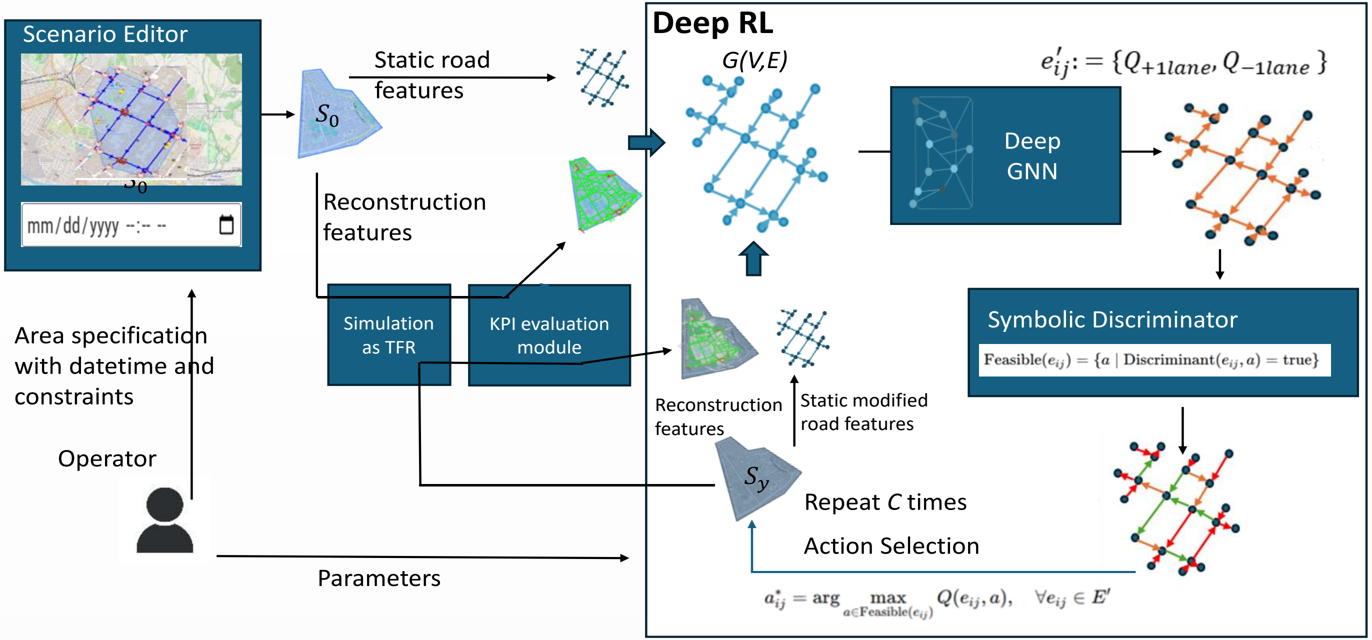


Neuro-Symbolic Optimization of Mobility Infrastructure via Deep-RL-GNN

Traffic flow density represented

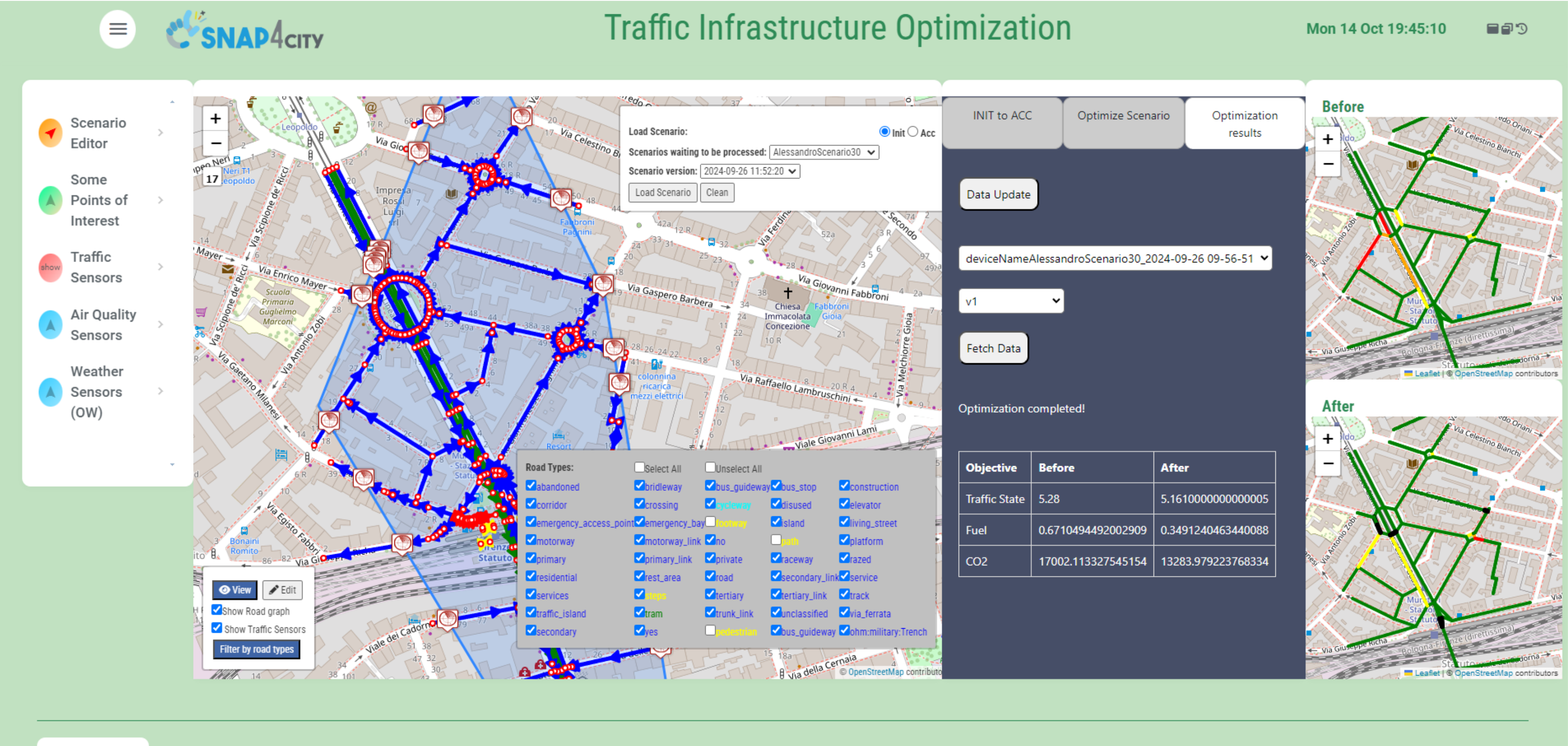
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

Deep-RL-GNN


Scenario Editor
mm/dd/yyyy --:-- --
S_0
Static road features
Reconstruction features
Simulation as TFR
KPI evaluation module
Area specification with datetime and constraints
Operator
Parameters
Deep RL
G(V,E)
$e'_{ij} := \{Q_{+1lane}, Q_{-1lane}\}$
Deep GNN
Symbolic Discriminator
$\mathrm{Feasible}(e_{ij}) = \{a \mid \mathrm{Discriminant}(e_{ij}, a) = \mathrm{true}\}$
Reconstruction features
Static modified road features
S_y
Repeat C times
Action Selection
$a^*_{ij} = \arg\max_{a \in \mathrm{Feasible}(e_{ij})} Q(e_{ij}, a), \quad \forall e_{ij} \in E'$

Future Artificial Intelligence Research

Traffic Infrastructure Optimization



Results of Neuro-Symbolic Optimization

The results reported on a limited number of **4 changes** improvements in a new urban scenario improving **fuel consumption and CO2 emissions** in both scenarios.

The proposed approach **outperforms Stochastic Relaxation** in all cases.

- Solution has validated the generalization and transferability capabilities of the Deep-RL-GNN, across different urban areas and at various traffic conditions.
- Compared to **Transfer Learning on**
  - traffic load achieves a 9.82% mean improvement with respect to the initial scenario,
  - area change improves KPIs by 25.58% with respect to the initial scenario.

Complex Manufacturing

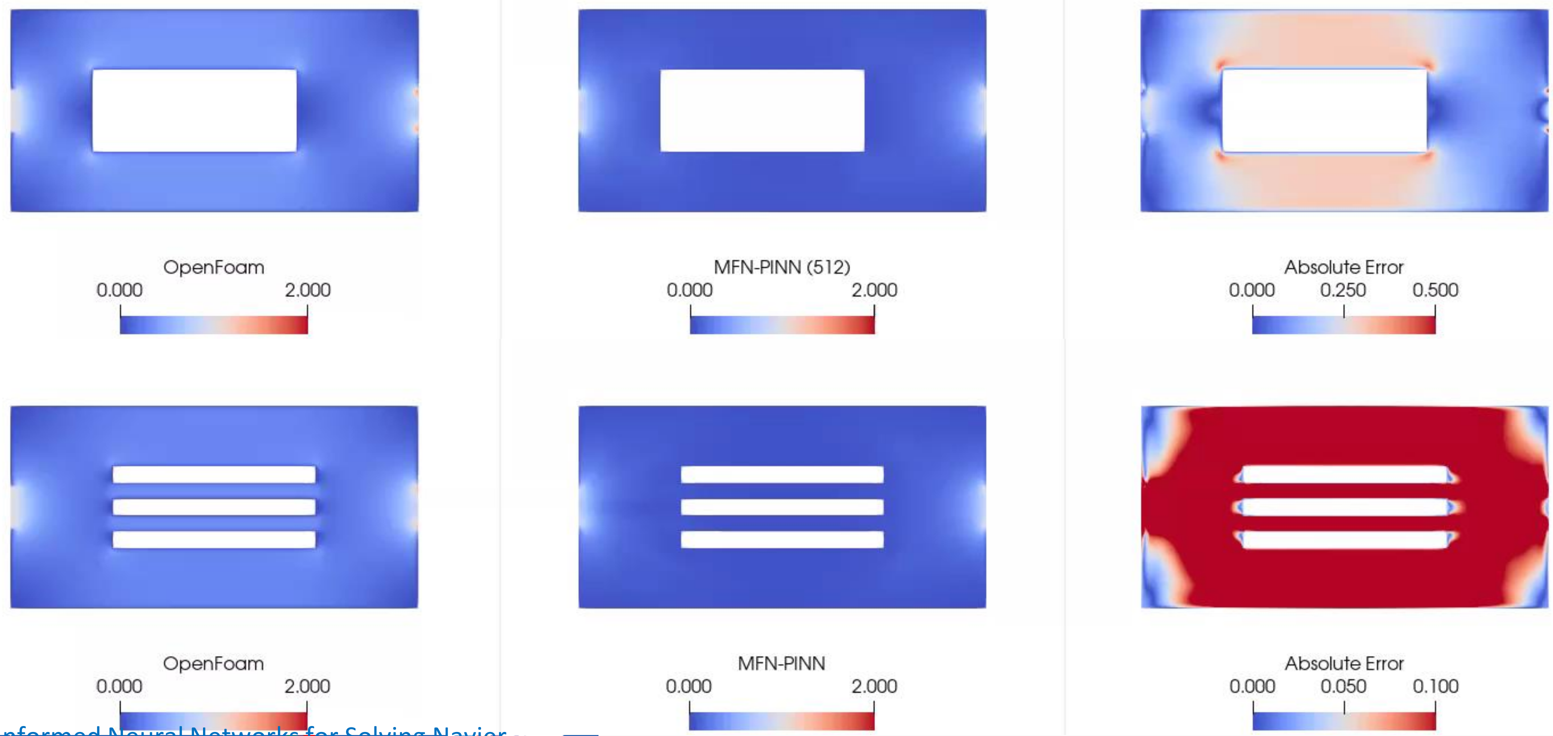
- T. Botarelli, M. Fanfani, P. Nesi, L. Pinelli, "**Using Physics-Informed Neural Networks for Solving Navier-Stokes Equations in Fluid Dynamic Complex Scenarios**",
  - Engineering Applications of Artificial Intelligence, Elsevier, 2025.
  - https://www.sciencedirect.com/science/article/pii/S0952197625003471

Using Physics-Informed Neural Networks for Solving Navier-Stokes Equations in Fluid Dynamic Complex Scenarios

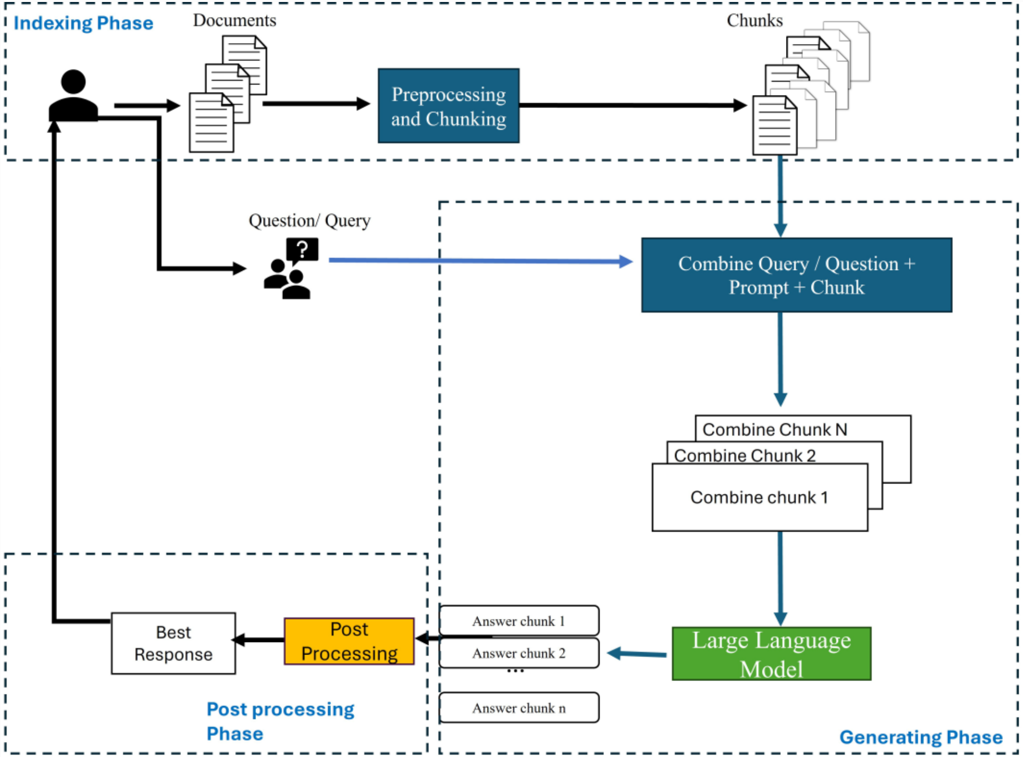
- **Usage of PINN for solving PDEs for fluid flow** on complex scenarios
  - Constrained pipes as autoclaves
  - Multiple shapes insides
- **Definition of learning processes and methods**
- **Training strategies** to reduce solution time:
  - Solutions valid for multiple shapes
  - Transfer learning
- **Validation wrt Open Foam**
- **Precision on steady and transitory cases**
- **Videos on https://www.snap4city.org/1010**

Using Physics-Informed Neural Networks for Solving Navier-Stokes Equations in Fluid Dynamic Complex Scenarios

Comparison of PINN vs OpenFoam and error



Context-Aware Retrieval Augmented Generation using Similarity Validation to handle Context Inconsistencies in Large Language Models



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

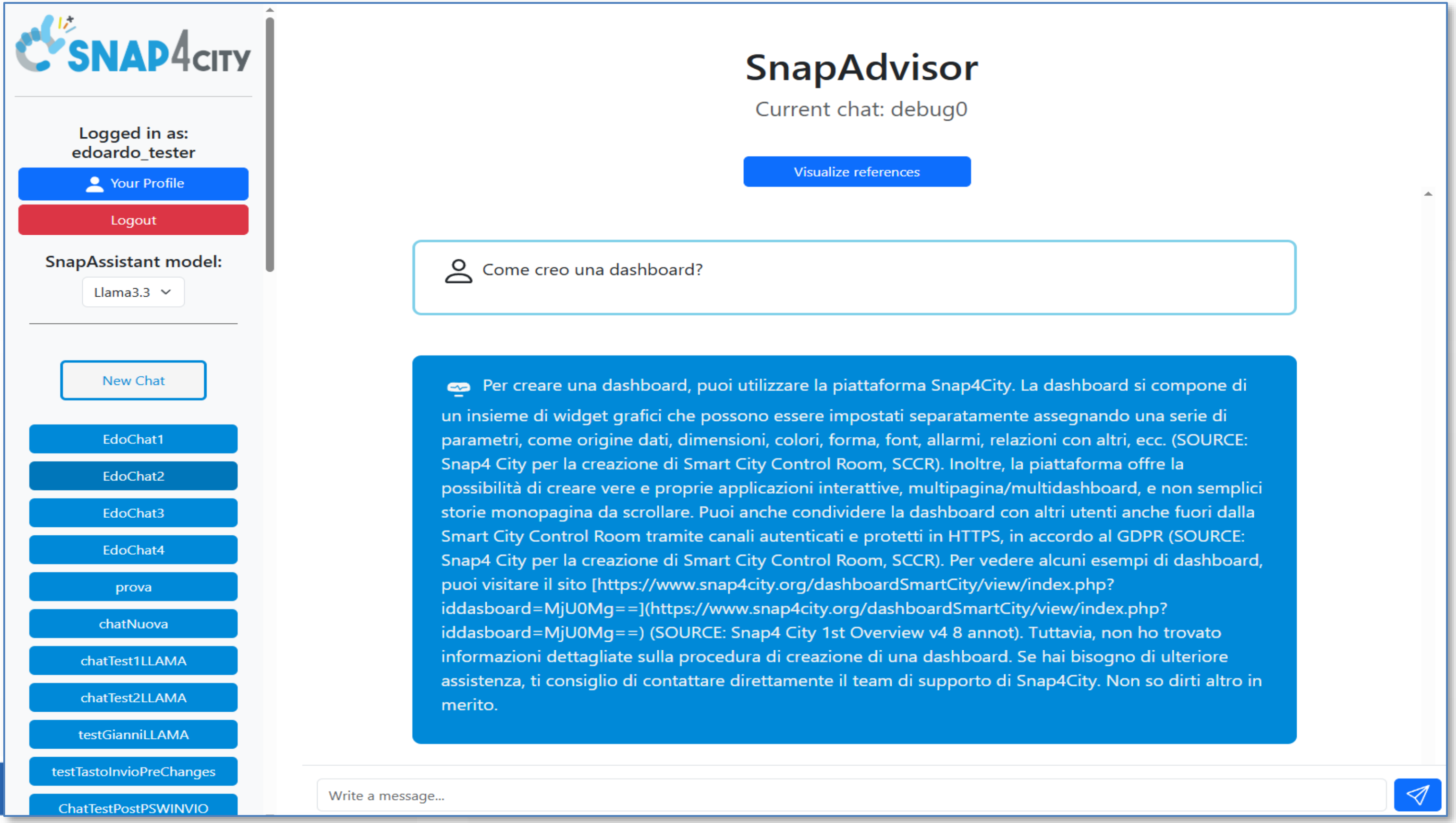
Comparing RAG solutions, Llama3.3

Method	Embedding Model	Semantic Chunking				Fixed-Size Chunking			
		Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
RAG	miniLM	0.4900	0.5500	0.5900	0.5700	0.4286	0.1481	0.8000	0.2500
RAG	Contriever-Msmarco	0.4762	0.4615	0.6	0.5218	0.4286	0.1481	0.8000	0.2500
RAG-Rerank	miniLM	0.6191	0.6667	0.72	0.6923	0.7619	0.6923	0.900	0.7826
RAG-Rerank	Contriever-Msmarco	0.5476	0.5926	0.6667	0.6275	0.7619	0.6923	0.900	0.7826
CA-RAG without post processing	miniLM	0.4600	0.3400	0.6900	0.4500	0.6092	0.4435	0.9373	0.6021
CA-RAG with post processing (Cosine Similarity)	miniLM	0.9200	0.8700	1.0000	0.9300	0.8539	0.7821	0.9985	0.8771
CA-RAG with post processing (Cosine Similarity)	Contriever-Msmarco	0.9144	0.8935	0.9934	0.9408	0.9297	0.8871	1.0000	0.9400
CA-RAG with post processing (Dot product)	miniLM	0.9704	0.9557	1.0000	0.9773	0.7126	0.5925	0.9617	0.7732
CA-RAG with post processing (Dot product)	Contriever-Msmarco	0.9832	0.9748	1.0000	0.9872	0.7782	0.6827	0.9780	0.8041
		All Context (no chunking)							

Future Artificial Intelligence Research

SnapAdvisor

UNIVERSITÀ DEGLI STUDI FIRENZE | DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE | DISIT DISTRIBUTED SYSTEMS AND INTERNET TECHNOLOGIES LAB | SNAP4CITY



Finanziato dall'Unione europea NextGenerationEU

Security and Computer Vision

Tasks

- Leverage multimodal RGB/Event sensor fusion to increase robustness of transfer learning for object recognition
- Efficient adaptation of cross-modal Vision Language Models for intra-modal tasks
- Robust transfer learning with foundation models in presence of adversarial attacks
- The Probability Simplex Leads to Compatible Representations

Multimodal RGB-Neuromorphic Vision

Adapting models to work out of distribution is challenging. This holds for RGB object detection models that often suffer in the presence of a domain shift. For example, a **model trained on daylight images will not perform well at night**.

However, other data modalities may carry a signal that is inherently more robust to certain shifts. Neuromorphic cameras, also referred to as event cameras, offer significant advantages over traditional frame-based RGB sensors, including higher temporal resolution, lower latency and high dynamic range that allow them to work well both in daylight and in darker scenes.

We collected the Florence RGB-Event Drone dataset, a novel multimodal dataset specifically designed for testing the effectiveness of several prediction tasks such as drone detection, tracking, and trajectory forecasting, combining RGB video and event streams.

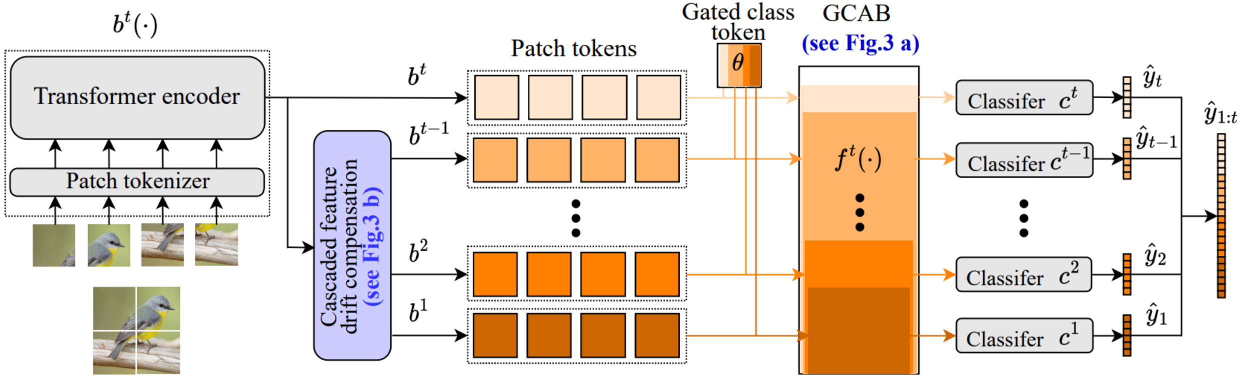
For all these tasks we provide **representative competitive NN model architectures**.

Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani | Future Artificial Intelligence Research

Continual Learning of Vision Transformers

Vision transformers (ViTs) have achieved remarkable successes across a broad range of computer vision applications. As a consequence, there has been increasing interest in extending **continual learning** theory and techniques to ViT architectures. However, the main challenge is maintaining plasticity of the learner **without causing catastrophic forgetting** of previously learned tasks.

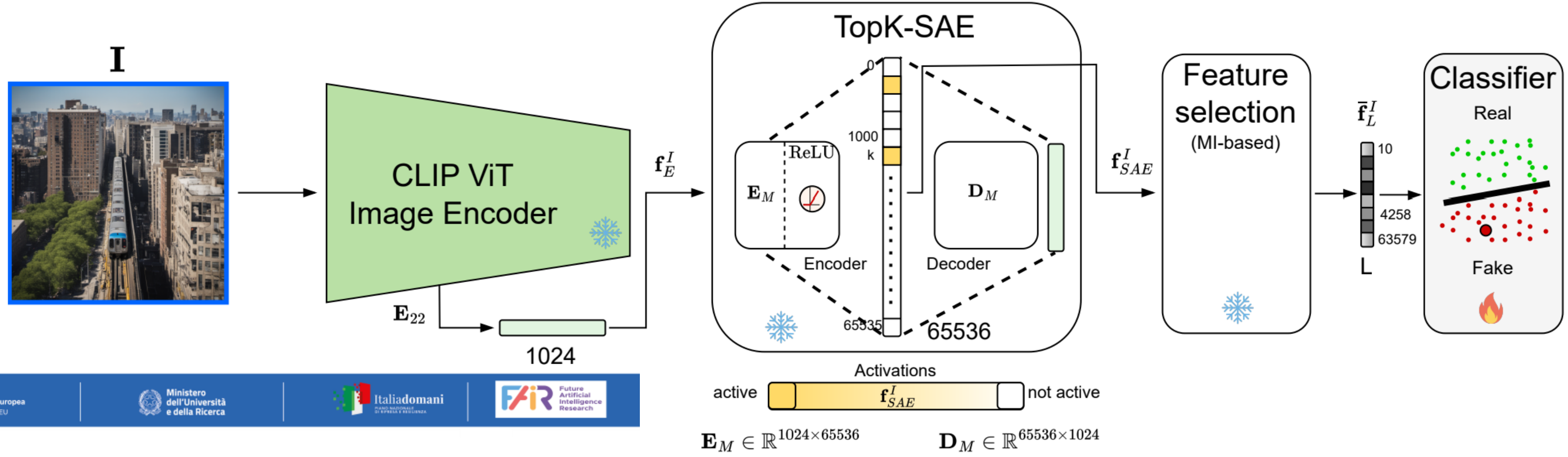
To address this problem, we first propose gated class-attention to minimize the drift in the final ViT transformer block. This mask based gating is applied to class-attention mechanism of the last transformer block and strongly regulates the weights crucial for previous tasks. Secondly, we propose a **new method of feature drift compensation that accommodates feature drift in the backbone when learning new tasks**. The combination of gated class-attention and cascaded feature drift compensation allows for plasticity towards new tasks while limiting forgetting of previous ones.



CLIP Feature Selection for Generalized **Deepfake Detection**

Generating multimedia contents, particularly images, by using AI-based tool is becoming an everyday practice. In some cases, this is done for malevolent aims such as disinformation, defamation and blaming. Reliably detecting synthetic generated pictures is becoming crucial, but unfortunately, the speed at which new generative models appear is clearly higher than that of accurate deepfake detectors.

We investigate the idea to select a reduced set of CLIP-based features by resorting to a sparse autoencoder (SAE); such a set represents a sort of a signature that improves the generalization capacity in detecting fake content. Experimental results carried out on extended datasets highlight this notable behavior achieving state-of-the-art performance.

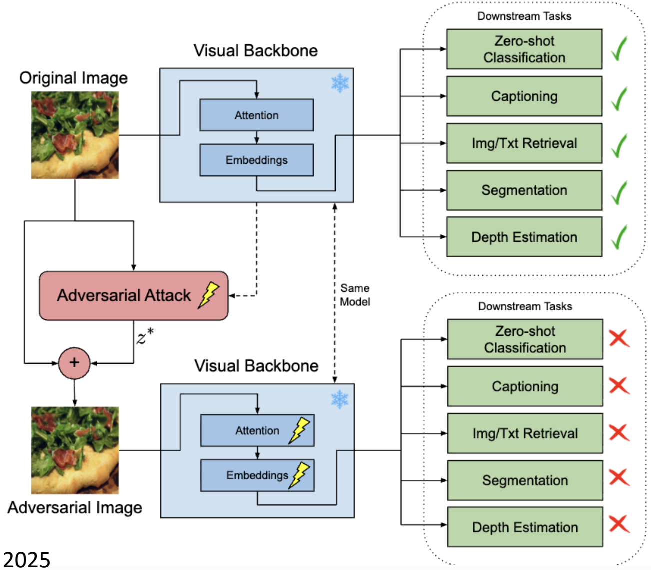


Attacking Foundation Models

Foundation models are large models, trained on broad data that deliver high accuracy in many downstream tasks, often without finetuning and are becoming the bedrock of many industrial AI-powered applications. However, the reliance on pre-trained foundation models also introduces significant security concerns, as **these models are vulnerable to adversarial attacks**. We study vulnerabilities of vision foundation models, focusing specifically on CLIP and ViTs, and explore the transferability of adversarial attacks to downstream tasks. We **introduce a novel attack**, targeting the structure of transformer-based architectures in a task-agnostic fashion.

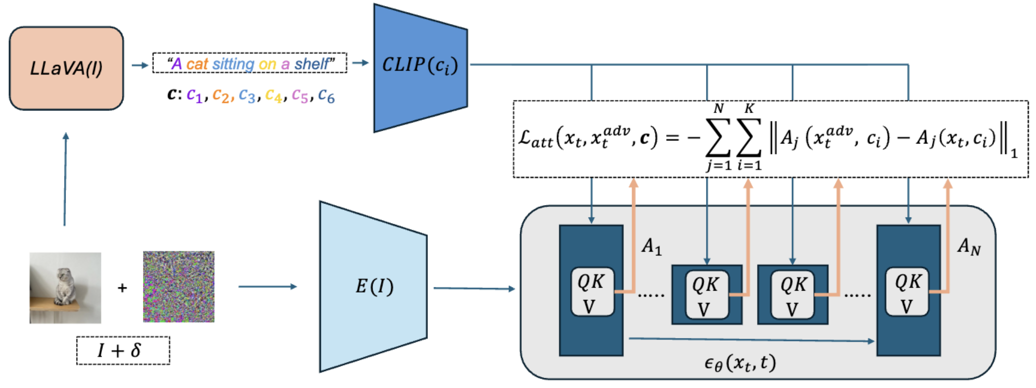
Clean

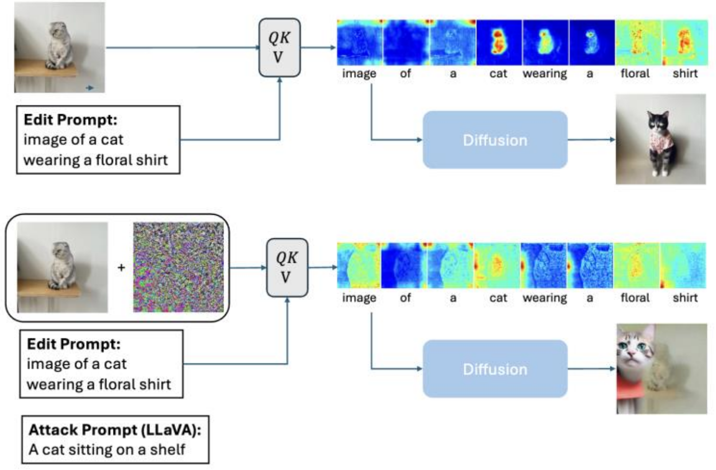
Attention Attack



Immunizing Images from Editing

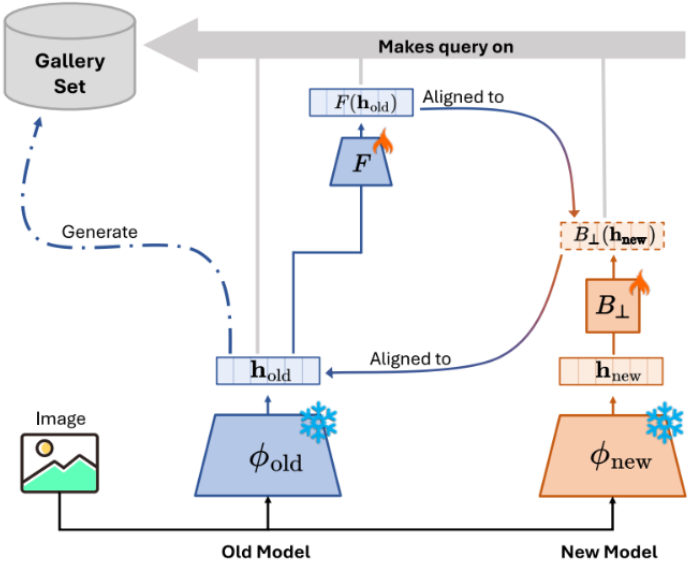
Text-based image editing have enabled fine-grained manipulation of visual content guided by natural language. However, such methods are susceptible to adversarial attacks. We propose a **novel attack that targets the visual component of editing methods**. Attention Attack disrupts the cross-attention between a textual prompt and the visual representation of the image by using an automatically generated caption of the source image as a proxy for the edit prompt. We propose **two novel evaluation strategies**: **Caption Similarity,** which quantifies semantic consistency between original and adversarial edits, and **semantic Intersection** over Union (IoU), which measures spatial layout disruption via segmentation masks.

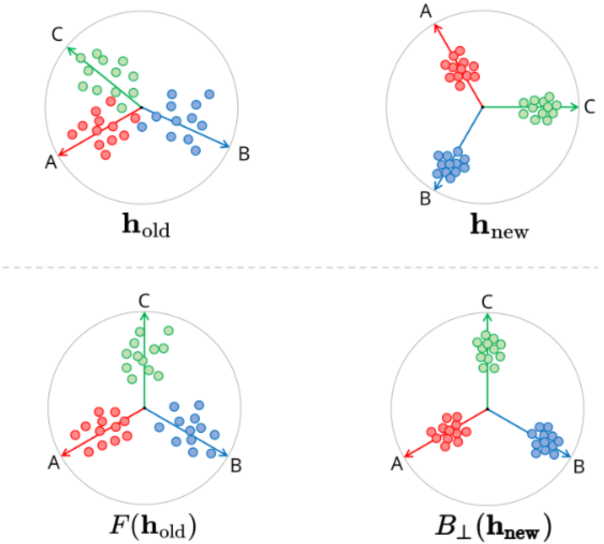




Compatible representation learning

Retrieval systems rely on representations learned by increasingly powerful models. However, due to the high training cost and inconsistencies in learned representations, there is significant interest in facilitating communication between representations and ensuring compatibility across independently trained neural networks. A key challenge is adapting the latent spaces of updated models to align with those of previous models on downstream distributions while preserving the newly learned representation spaces. In this study, we impose a relaxed orthogonality constraint, namely λ-orthogonality regularization, while learning an affine transformation, to obtain distribution-specific adaptation while retaining the original learned representations.





Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca

Security and Computer Vision

Publications:

- Cross the Gap: Inter-modal CLIP Representations Are Superior for Intra-modal Tasks. **Int. Conf. on Learning Representations**, ICLR, Singapore, April 2025.
- Exemplar-Free Continual Learning of Vision Transformers via Gated Class-Attention and Cascaded Feature Drift Compensation. **International Journal of Computer Vision**. Mar 2025.
- Spike-TBR: A noise resilient neuromorphic event representation. **Pattern Recognition Letters**. Vol.196, Oct. 2025.
- FRED: The Florence RGB-Event Drone Dataset. **ACM Multimedia Conference**, Dublin, Ireland 27-31 Oct. 2025
- Immunizing Images from Text to Image Editing via Adversarial Cross-Attention. **ACM Multimedia Conference**, Dublin, Ireland 27-31 Oct. 2025.
- Attacking Attention of Foundation Models Disrupts Downstream Tasks. Int. Conf. on **Computer Vision and Pattern Recognition**, Nashville, USA, June 11-15, 2025.
- Drone Detection with Event Cameras. **Int. Conf. on Image Analysis and Processing** (Workshops), Rome, Italy, Sept. 2025.
- CLIP Feature Selection Mechanism via Sparse Autoencoder for Generalized Deepfake Detection. **IEEE Int. Work. On Information Forensics and Security**, Perth, Australia, Dec. 1-4, 2025.
- λ-Orthogonality Regularization for Compatible Representation Learning. Annual Conference on **Neural Information Processing Systems**, NeurIPS2025, San Diego, USA, Dec. 2-7, 2025.

AI Lab

- **Automatic analysis of terms of service and privacy policies with Large Language Models**, in collaboration with the University of Bologna (Panarelli et al. 2025).
- **Neuro-symbolic artificial intelligence for learning and reasoning with temporal knowledge**, in collaboration with the University of Siena (Lorello et al. 2025a, Lorello et al. 2025b)
- **Exploiting large language models to fill clinical questionnaires**, in collaboration with ARS Tuscany and AOU Careggi
- **Explainable artificial intelligence techniques for signal quality estimation of the MARSIS radar**, in collaboration with University of Modena and Reggio Emilia and National Institute of Astrophysics

Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani | Future Artificial Intelligence Research

AILab tasks references

- Panarelli, M., Galassi, A., Lagioia, F., Liepina, R., Lippi, M., Palka, P., Sartor, G., **Is It Worth Using LLMs for Unfair Clause Detection in Terms of Service?**
  - International Conference on Artificial Intelligence and Law, 2025 **(Best Innovative Application Award)**.
- Nardoni, V., Hyeraci, G., Maccari, M., Arana, A., Lucenteforte, E., Limoncella, G., Mohammadi, S., Roberto, G., Tarazjani, A. D., Virgili, G., Wiebel, D., Gini, R., Lippi, M., Marinai, S., **Comparing Humans and Large Language Models in Filling Clinical Questionnaires**,
  - The 4th International Conference Series on Hybrid Human-Artificial Intelligence, Pisa, June 2025.
- Lorello, L.S., Lippi, M., Melacci, S., **A Neuro-Symbolic Framework for Sequence Classification with Relational and Temporal Knowledge**
  - Internationa Joint Conference on Artificial Intelligence (IJCAI), Montreal, 2025.
- Lorello, L.S., Lippi, M., Melacci, S., **The KANDY benchmark: Incremental neuro-symbolic learning and reasoning with Kandinsky patterns**
  - Machine Learning Journal 114(7): 161 (2025)
- Ferrari, B., Lippi, M., Ganzerli, G., Iori, M., Orosei, R.., **Explainable Artificial Intelligence for Quality Estimation of MARSIS Observations**
  - Prestigious Applications of Artificial Intelligence (PAIS), co-located with European Conference on Artificial Intelligence (ECAI), Bologna, 2025.