PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

UrbanMix: LLM-Guided Simulation of Mixed Autonomy Traffic with Heterogeneous Behavioral Profiles

Roman Sultimov[1,2] , **Daniil Efimov**[3] , **Ivan Novikov**[1,2] and **Aleksandr Volkov**[4] , **Yury Maximov**[4]
[1] Lomonosov Moscow State University, Moscow, Russia
[2] Moscow Independent Research Institute of Artificial Intelligence, Moscow, Russia
[3] Moscow Institute of Steel and Alloys, Moscow, Russia
[4] Interdata, Astana, Kazakhstan
{r.sultimov, i.novikov}@iai.msu.ru, efimov.da@misis.ru, {volkov,yury}@iccda.io

Abstract

Cities deploying autonomous vehicles face an urgent policy question: would the adoption of autonomous vehicles (AVs) improve the congestion rate or worsen it? What would be the optimal adoption rate to minimize the congestion rate? How would cautious AVs (Waymo-style) and aggressive AVs (Tesla "Mad Max"-style) interact with human drivers and delivery robots on shared roads?

We present URBANMIX, an interactive simulation platform that embeds cognitively diverse agents (human drivers, cautious AVs, aggressive AVs, and delivery robots) with distinct behavioral profiles inside a Simulation of Urban MObility (SUMO) framework of real urban road networks. Our LLM planner operates as an urban policy coordinator, setting traffic rules through a bounded action interface, while a regulation shield enforces infrastructure constraints. Our experiments reveal three key phenomena: (i) roads throughput may substantially drop with the increase in AVs adoption; (ii) an aggressive cascade, where runtime behavior switching modeling Tesla user-selectable "Mad Max" mode triggers up to 60 times increase in emergency braking events on a real Austin Downtown network; and (iii) a delivery bottleneck, showing 24–36% throughput reduction from slow robots. Results validated on synthetic and real data from Austin, TX, demonstrate that real network topology amplifies cascade effects by more than four times.

1 Introduction

In December 2025, a power outage in San Francisco exposed a critical gap in urban planning: Waymo robotaxis froze at intersections city-wide, blocking cars, buses, and emergency vehicles [Ha, 2025]. Autonomous vehicles violate the rules even in non-emergency situations: Waymo accumulated 589 parking tickets and 124 obstruction citations in San Francisco alone in 2024 [Bonos, 2025], while General Motors had to shut down Cruise after $10B in losses following a pedestrian-dragging incident [General Motors, 2024]. Furthermore, on the top of its "Chill" and "Standard" full self-driving (FSD) mode, Tesla recently deployed "Mad Max", an aggressive driving profile that prioritizes speed over caution. This mode leads in tighter headways, more frequent lane changes, and exceeding speed limits [Hughes, 2025]. Within weeks, the US National Highway Traffic Safety Administration opened an investigation covering 2.9 million FSD-equipped vehicles, citing 58 complaints, including 14 crashes and 23 injuries.

These events reveal a fundamental challenge: *cities are deploying autonomous vehicles with switchable behavior profiles, but have no tools to predict how heterogeneous, mode-switching AV fleets interact with existing traffic*. The problem is compounded by three factors: (1) behavioral diversity: different manufacturers embed fundamentally different driving philosophies; (2) runtime switching: users can change AV behavior mid-trip, creating fleet-wide cascades; and (3) multi-modal conflicts: slow delivery robots [Starship Technologies, 2024] share infrastructure with fast, assertive AVs.

The research landscape offers partial solutions summarized in Tab. 1. Mixed-autonomy simulation [Wu *et al.*, 2022; Vinitsky *et al.*, 2018] established reinforcement learning (RL)-based analysis but treats AVs as a uniform population, ignoring the behavioral diversity essential in practice. LLM-assisted SUMO tools [Li *et al.*, 2024; Jeong *et al.*, 2025] exibit the same limitation. LLM-driven urban simulations [Liu *et al.*, 2025; Bougie and Watanabe, 2025; Sultimov *et al.*, 2026b] model agent diversity but do not integrate with calibrated traffic microsimulators and ignore real-road topology challenges. Another challenge is the climate and natural disasters impact [Sultimov *et al.*, 2026a; Taniushkina *et al.*, 2026]. Platforms like SmartTransit.AI [Pavia *et al.*, 2024] address transit routing but omit mixed-autonomy phenomena.

We present URBANMIX[1], a platform that addresses these gaps by embedding cognitively diverse agents inside SUMO microsimulation, see Fig. 1 for details. Our contributions are:

- A *heterogeneous agent architecture* with four behaviorally calibrated vehicle types, including runtime mode switching via TraCI[2] directly modeling the Tesla "Chill/Standard/Mad Max" paradigm.
- An *LLM urban planner* operating as a policy coordinator with a regulation shield enforcing infrastructure con-

[1]Demonstration video: https://youtu.be/ksgdX5iguFw. Live demo: https://evacuation-viz.vercel.app/urban

[2]https://sumo.dlr.de/docs/TraCI/index.html

System	Behavioral diversity	Runtime switching	LLM planner	Real SUMO network	Interactive web demo
Flow [Wu *et al.*, 2022]	–	–	–	✓	–
ChatSUMO [Li *et al.*, 2024]	–	–	✓	✓	–
AgentSUMO [Jeong *et al.*, 2025]	–	–	✓	✓	–
GATSim [Liu *et al.*, 2025]	✓	–	✓	–	–
CitySim [Bougie and Watanabe, 2025]	✓	–	✓	–	–
SmartTransit [Pavia *et al.*, 2024]	–	–	–	✓	✓
UrbanMix	✓	✓	✓	✓	✓

Table 1: Comparison with existing mixed-autonomy and LLM/SUMO platforms. Only UrbanMix supports all five capabilities simultaneously.

straints [Shoham and Tennenholtz, 1992].
- *Cross-network validation* on a real Austin Downtown network (325 nodes, 676 edges, 127 traffic lights), demonstrating that real topology amplifies aggressive cascade effects by 4 times compared to simple synthetic grids.

An interactive web dashboard for real-time policy exploration, deployable to any city SUMO network, demonstrates the solution.

2 System and Demonstration

URBANMIX comprises four layers sharing a common `PolicyAction` interface.

Heterogeneous agent population. We define four vehicle types with calibrated SUMO car-following (Krauss model) and lane-change parameters; see Tab. 2 for details. *Human drivers* exhibit stochasticity, modeling reaction-time variability and imperfect car-following observed in field studies [Stern *et al.*, 2018]. *Cautious AVs* model Waymo-style conservative behavior: large headways (τ=2.0 s), cooperative lane changes (`lcSpeedGain`=0.3), and a 3.0 m minimum gap reproducing the moving bottleneck phenomenon documented in mixed-autonomy literature [Wu *et al.*, 2022]. *Aggressive AVs* model Tesla "Mad Max"-style driving: tight headways (τ=0.8 s), assertive lane changes (`lcSpeedGain`=2.0), and a higher maximum speed of 130 km/h matching reported behavior of exceeding posted limits and forcing lane changes [Hughes, 2025]. *Delivery robots* are modeled after Starship's fleet [Starship Technologies, 2024]: slow (10 km/h maximum), small (`minGap`=1.0 m), and inability to change lanes.

Parameter	Human	AV_{caut}	AV_{aggr}	Robot
σ (stochasticity)	0.5	0.0	0.0	0.0
τ (headway, s)	1.2	2.0	0.8	1.5
accel (m/s^2)	2.6	2.0	3.5	0.8
decel (m/s^2)	4.5	4.5	6.0	2.0
minGap (m)	2.5	3.0	1.5	1.0
maxSpeed (km/h)	119	119	130	10
lcSpeedGain	1.0	0.3	2.0	0.0

Table 2: Agent behavioral profiles: each column represents a distinct driving philosophy; runtime switching between types via TraCI enables dynamic mode transitions.

Runtime behavior switching via TraCI `setType()` enables dynamic transitions between any agent type at any simulation step directly modeling the real-world scenario where Tesla drivers toggle between "Chill" and "Mad Max" modes mid-trip. This mechanism is critical: it allows UrbanMix to

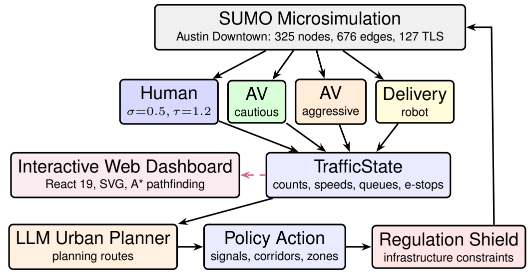

Figure 1: The URBANMIX architecture. Four agent types with distinct SUMO behavioral parameters share roads on a real Austin Downtown network. The LLM urban planner proposes `PolicyActions`; the regulation shield enforces infrastructure constraints before applying changes to the simulation.

study not just static fleet compositions, but the dynamic cascades that emerge when multiple vehicles switch behavior simultaneously.

SUMO backend and networks. The backend uses SUMO [Lopez *et al.*, 2018] via TraCI [Wegener *et al.*, 2008]. We validate on two networks of contrasting complexity: (1) *SynthTown*, a synthetic grid (117 intersections, 297 segments, 2 bridge bottlenecks) providing a controlled baseline; and (2) *Austin Downtown*, a real network extracted from OpenStreetMap via netconvert (325 nodes, 676 edges, 127 signalized junctions). The Austin network preserves the full topological complexity of a real city: Congress Avenue as the primary north-south corridor, Cesar Chavez Boulevard along the eastern boundary, Guadalupe Street bounding the university district to the north, and Lady Bird Lake (Colorado River) running east-west with constrained bridge crossings. Roads are classified by hierarchy primary highways, arterials, collectors, and residential streets with signal timing derived from the Austin transportation network.

LLM planner & regulation shield. Every few steps, the planner observes a structured *TrafficState* vector: per-type vehicle counts, mean speeds, queue lengths, and cumulative conflict events (emergency braking, near-misses). It proposes a *PolicyAction* drawn from a bounded action space: signal timing adjustments (green/red phase durations), AV corridor designations (restricting aggressive AVs from specified road segments), and speed zone modifications. This bounded interface prevents the LLM from proposing physically invalid actions (e.g., negative green times) while preserving its capacity for policy-level reasoning [Shoham and Tennenholtz, 1992]. All proposed actions pass through a *regulation shield* that validates compliance with infrastructure constraints: minimum green times ($\geq$ 5 s), maximum phase durations ($\leq$ 120 s), speed limits consistent with road classification, and capacity constraints at bridge crossings. Violations trigger automatic repair (clamping values to valid ranges) or rejection with explanation. The planner operates as a *policy coordinator* it does not control individual vehicles, but sets the norms under which heterogeneous agents interact.

Surrogate safety metrics. We adopt *emergency braking events* (e-stops) as our primary surrogate safety measure, defined as decelerations exceeding -4.5 m/s^2. This threshold

Metric	Sc. 1 (30% AV) Synth	Sc. 1 (30% AV) Austin	Sc. 2 (cascade) Synth	Sc. 2 (cascade) Austin	Sc. 3 (robots) Synth	Sc. 3 (robots) Austin
Throughput (cars)	110	102	110	104	70	78
Avg spd (km/h)	52	48	52	47	33	36
Peak queue (cars)	16	**78**	16	79	13	**95**
E-stops (events)	1	66	**176**	**760**	0	24

Table 3: Scenario results on two networks (200 vehicles, 900 s, seed=42). E-stops count decelerations $< -4.5\,\text{m/s}^2$.

is calibrated to the deceleration parameter of human drivers (Tab. 2) and aligns with established surrogate safety methodology [Papadoulis *et al.*, 2019]. E-stops capture near-miss conflicts that throughput and speed metrics miss entirely, as our Scenario 2 results demonstrate.

3 Scenarios and Results

We demonstrate three scenarios exposing distinct mixed-autonomy phenomena; see Tab 3 and Fig. 2. All runs use 200 vehicles over 900 s (Δt=0.5 s, 1800 simulation steps) with deterministic seeding for reproducibility.

Scenario 1: Penetration rate sweep. Users sweep AV penetration from 0% to 50% (all cautious). This scenario quantifies the "valley of disruption" the counterintuitive range of AV adoption rates at which traffic performance *degrades* before eventually improving. At 30% penetration on Austin, cautious AVs create moving bottlenecks: their large headways (τ=2.0 s) reduce effective road capacity while human drivers learn to exploit their predictable yielding behavior [Stern *et al.*, 2018; Liu, 2025]. On the Austin network, throughput drops to 102 vehicles (from a 110-vehicle baseline at 0% AV) a 7% reduction. More significantly, 66 emergency braking events occur even under normal mixed conditions, establishing a critical baseline. On the smaller SynthTown network, throughput remains at 110 with only 1 e-stop, illustrating how real network complexity creates friction absent in synthetic grids. The peak disruption zone (10–30% penetration) is consistent with capacity reduction observed in mixed-autonomy field experiments [Wu *et al.*, 2022; Papadoulis *et al.*, 2019].

Scenario 2: Aggressive cascade (aka Tesla "Mad Max"). This scenario our most dramatic result directly models what happens when AV drivers toggle to aggressive mode, as Tesla "Mad Max" deployment enables [Hughes, 2025]. All AVs start cautious; at t=300 s, the user switches them to aggressive mode, simulating a fleet-wide mode transition. Aggressive AVs (τ=0.8 s) immediately reduce headways by 60% and begin assertive lane changes with `lcSpeedGain`=2.0.

On Synthetic Town regular grid, this results in 176 e-stops. On Austin Downtown, the real non-orthogonal topology narrow one-way streets, constrained bridge crossings over Lady Bird Lake, and 127 signalized intersections with real timing amplifies the cascade to 760 e-stops, 4 times amplification over similar synthetic data; see Fig. 2. Before the switch ($t < 300$), only 12 e-stops accumulate; after the switch, 748 additional e-stops occur in the remaining 600 s, a 60 times increase in the e-stop rate.

Crucially, throughput remains nearly identical (104 vs. 102 vehicles). This decoupling of safety from efficiency is the scenario key finding: throughput-only metrics are blind to the safety catastrophe induced by runtime behavior switching. This is precisely the scenario cities face with Tesla Mad Max: same throughput, much more danger [Hou, 2023].

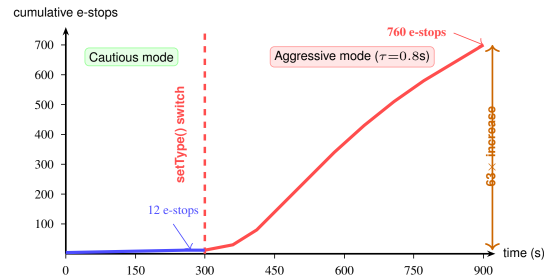

Figure 2: Scenario 2 on Austin TX Downtown: cumulative emergency braking events over 900 s. The behavioral mode switch at t=300 (cautious to aggressive) triggers a 60 times increase in the e-stop rate, while throughput remains at 104 vehicles demonstrating that throughput-only metrics are blind to safety degradation.

Scenario 3: Delivery robot bottleneck. Twenty delivery robots (10 km/h, no lane changes) modeled after Starship's fleet of 2,700+ robots [Starship Technologies, 2024] share roads with mixed traffic. Their slow speed and inability to change lanes create persistent queue formation, particularly on Austin's narrow residential streets where alternative routes are scarce. Throughput drops to 78 vehicles on Austin (24% reduction) and 70 on Synthetic Town (36% reduction), while robots effective speed falls to approximately 3 km/h due to congestion interactions [Jennings and Figliozzi, 2019]. Peak queue length reaches 95 vehicles on Austin the highest across all scenarios despite low e-stop counts (24), indicating that robots cause gridlock rather than conflicts.

Key insight: topology amplifies cascades. The systematic comparison across two networks reveals that synthetic grids consistently underestimate cascade severity. The 4.3× e-stop amplification from SynthTown (176) to Austin (760) in Scenario 2 arises from three topological factors: (1) Austin's non-orthogonal street layout creates merge conflicts at acute angles absent in regular grids; (2) bridge crossings over Lady Bird Lake act as capacity bottlenecks where aggressive lane changes propagate upstream; and (3) 127 signalized intersections with real timing create stop-and-go patterns that compound with tight headways. These findings underscore the necessity of validating mixed-autonomy policies on real road topologies before deployment.

Ethical Statement

The paper has no ethical issues. URBANMIX is a research prototype intended for policy exploration and decision support. AV behavioral profiles are calibrated abstractions of driving styles, not manufacturer-specific predictions; "cautious"/"aggressive" labels reflect behavioral parameter ranges, no specific commercial systems claims. The data is synthetic; no real participants or personal data is involved.

Acknowledgments

The work of Roman Sultimov and Ivan Novikov is supported by the Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4H0002; grant No 139-15-2025-012).

References

[Bonos, 2025] Lisa Bonos. These driverless taxis got 589 parking tickets in San Francisco last year. The Washington Post, 2025.

[Bougie and Watanabe, 2025] Nicolas Bougie and Narimasa Watanabe. CitySim: Modeling urban behaviors and city dynamics with large-scale LLM-driven agent simulation. In *Proc. EMNLP*, pages 215–229, 2025.

[General Motors, 2024] General Motors. GM to refocus autonomous driving development on personal vehicles. GM News, 2024.

[Ha, 2025] Anthony Ha. Waymo resumes service in San Francisco after robotaxis stall during blackout. TechCrunch, 2025.

[Hou, 2023] Guangyang Hou. Evaluating efficiency and safety of mixed traffic with connected and autonomous vehicles in adverse weather. *Sustainability*, 15:3138, 2023.

[Hughes, 2025] Justin Hughes. Tesla deploys 'Mad Max' mode, immediately triggers NHTSA investigation. Jalopnik, 2025.

[Jennings and Figliozzi, 2019] Dylan Jennings and Miguel Figliozzi. Study of sidewalk autonomous delivery robots and their potential impacts on freight efficiency and travel. *Transp. Res. Rec.*, 2673:317–326, 2019.

[Jeong *et al.*, 2025] Minwoo Jeong, Jeeyun Chang, and Yoonjin Yoon. AgentSUMO: An agentic framework for interactive simulation scenario generation in SUMO via large language models. *arXiv:2511.06804*, 2025.

[Li *et al.*, 2024] Shuyang Li, Talha Azfar, and Ruimin Ke. ChatSUMO: Large language model for automating traffic scenario generation in simulation of urban MObility. *IEEE Trans. Intell. Veh.*, 2024.

[Liu *et al.*, 2025] Qi Liu, Can Li, and Wanjing Ma. GATSim: Urban mobility simulation with generative agents. *arXiv:2506.23306*, 2025.

[Liu, 2025] Peng Liu. Social interactions between automated vehicles and human drivers: A narrative review. *Ergonomics*, 68:1761–1780, 2025.

[Lopez *et al.*, 2018] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using SUMO. In *Proc. IEEE ITSC*, pages 2575–2582, 2018.

[Papadoulis *et al.*, 2019] Alkis Papadoulis, Mohammed Quddus, and Marianna Imprialou. Evaluating the safety impact of connected and autonomous vehicles on motorways. *Accid. Anal. Prev.*, 124:12–22, 2019.

[Pavia *et al.*, 2024] Sophie Pavia, David Rogers, Amutheezan Sivagnanam, Michael Wilbur, Danushka Edirimanna, Youngseo Kim, Ayan Mukhopadhyay, Philip Pugliese, Samitha Samaranayake, Aron Laszka, and Abhishek Dubey. SmartTransit.AI: A dynamic paratransit and microtransit application. In *Proc. IJCAI*, pages 8767–8770, 2024.

[Shoham and Tennenholtz, 1992] Yoav Shoham and Moshe Tennenholtz. On the synthesis of useful social laws for artificial agent societies: Preliminary report. In *Proc. AAAI*, pages 276–281, 1992.

[Starship Technologies, 2024] Starship Technologies. Robot delivery leader Starship Technologies raises $90 million led by Plural and Iconical. Starship Technologies, 2024.

[Stern *et al.*, 2018] Raphael E. Stern, Shumo Cui, Maria Laura Delle Monache, Rahul Bhadani, Matt Bunting, Miles Churchill, Nathaniel Hamilton, R'mani Haulcy, Hannah Pohlmann, Fangyu Wu, Benedetto Piccoli, Benjamin Seibold, Jonathan Sprinkle, and Daniel B. Work. Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments. *Transp. Res. Part C*, 89:205–221, 2018.

[Sultimov *et al.*, 2026a] Roman Sultimov, Mikhail Mozikov, Dmitrii Abramov, Mariia Kovalchuk, Maksim Malykh, Aleksandr Volkov, Ilya Makarov, Andrei Osiptsov, and Yury Maximov. RESPOND: Realistic environment simulation of population and natural disasters with LLM-driven agents (student abstract). In *Proc. AAAI*, volume 40, pages 41694–41696, 2026.

[Sultimov *et al.*, 2026b] Roman Sultimov, Aleksandr Volkov, Mile Mitrovic, and Yury Maximov. LLM-guided multi-agent evacuation coordination via episodic memory and cognitive task analysis. In *Proc. AAMAS*, pages 4164–4166, 2026.

[Taniushkina *et al.*, 2026] Daria Taniushkina, Aleksander Lukashevich, Valeriy Shevchenko, Aleksandr Bulkin, Nazar Sotiriadi, and Yury Maximov. Bridging global climate solutions and local realities: Evaluating neural networks for high-resolution downscaling. *IEEE Access*, 14:47863–47878, 2026.

[Vinitsky *et al.*, 2018] Eugene Vinitsky, Aboudy Kreidieh, Luc Le Flem, Nishant Kheterpal, Kathy Jang, Cathy Wu, Fangyu Wu, Richard Liaw, Eric Liang, and Alexandre M. Bayen. Benchmarks for reinforcement learning in mixed-autonomy traffic. In *Proc. CoRL*, pages 399–409, 2018.

[Wegener *et al.*, 2008] Axel Wegener, Michał Piórkowski, Maxim Raya, Horst Hellbrück, and Stefan Fischer. TraCI: An interface for coupling road traffic and network simulators. In *Proc. CNS*, pages 155–163, 2008.

[Wu *et al.*, 2022] Cathy Wu, Aboudy Kreidieh, Kanaad Parvate, Eugene Vinitsky, and Alexandre M. Bayen. Flow: A modular learning framework for mixed autonomy traffic. *IEEE Trans. Robot.*, 38:1270–1286, 2022.