

Comfort by Construction: Adaptive, Comfort-Bounded Action Spaces for Learned Driving Policies

Anna Rothenhäusler^{1*}, Daniel Jost^{1*}, Raghu Rajan¹, Faris Janjos²,
Oliver Scheel², Andreas Look³, and Joschka Boedecker¹

¹ University of Freiburg, Freiburg, Germany

{rothenha, jostd, rajanr, jboedeck}@cs.uni-freiburg.de

² Bosch Center for Artificial Intelligence, Stuttgart, Germany

{faris.janjos, oliver.scheel}@de.bosch.com

³ Coburg University of Applied Sciences, Coburg, Germany
andreas.look@hs-coburg.de

Abstract. Data-driven driving simulators command accelerations and steering rates from a fixed grid *without* constraining the *realized* accelerations and jerks. As a result, reinforcement-learning policies inflate safety metrics through abrupt, last-second maneuvers that lie far outside the range of human driving and would be unacceptable to occupants of a real vehicle, so the metrics measure simulator permissiveness rather than policy quality. Enforcing comfort bounds naively is not enough: lateral limits shrink quadratically with speed, so clamping a static grid saturates it and destroys fine-grained control (“grid collapse”). We propose an *adaptive* action parameterization that re-discretizes the grid at every step to span exactly the per-step feasible control set, via closed-form inversion of the lateral-jerk constraint. We further present PUFFERDRIVE-EDITOR, a browser-based tool to audit realized kinematics and author kinematically challenging scenes. On the Waymo Open Motion Dataset and a hand-authored slalom, our adaptive model holds comfort violations below 1% while outperforming clipped-grid and direct-jerk baselines in navigability.

1 Introduction

Learned driving policies are increasingly trained in data-driven simulators like Nocturne [23], Waymax [11], or GPUdDrive [12]. These simulators are built for closed-loop reinforcement learning (RL): RL trains on policy-induced states, enabling learning from rare safety-critical situations. While achieving high goal completion, their behavior is shaped by the kinematic bicycle model [13, 17] translating actions into motion. Standard interfaces use a fixed grid of accelerations and steering rates that does not constrain the *realized* accelerations and jerks. Consequently, RL policies learn to avoid collisions through abrupt, last-second maneuvers (sharp braking combined with hard steering) whose realized

* These authors contributed equally to this work.

accelerations and, especially, jerks fall well outside the comfort envelope of even aggressive human driving [3, 10]. Safety metrics thus measure the permissiveness of the simulator’s kinematics rather than policy quality, crediting evasions that no comfort-respecting vehicle would reproduce. This work studies the problem in PUFFERDRIVE, whose actuation interface is representative of the simulators above; the analysis and the proposed parameterization carry over to any simulator built on the same kinematic bicycle model. Naively enforcing comfort bounds is difficult because the feasible action set depends nonlinearly on the vehicle state. In particular, the admissible steering range shrinks rapidly with speed, alongside additional nonlinearities from the vehicle kinematics [3, 13]. A static grid tuned for low speed then falls largely outside this range at higher speeds, and clamping collapses many grid points onto the constraint boundary, degrading fine-grained control (“grid collapse”). Because widely used data-driven simulators expose *discrete* action grids [11, 12], we focus on maintaining a feasible discrete grid without collapse. To inspect these failure modes, we present PUFFERDRIVE-EDITOR, a browser-based tool for visualizing kinematic violations and authoring challenging scenarios. It lets us trace where policies breach the comfort envelope: qualitatively, most often just before sharp turns and narrow gaps. To keep control feasible in exactly these situations without collapsing the grid, we propose an adaptive action parameterization that computes the exact feasible control set at each step via closed-form inversion of the lateral-jerk constraint, re-discretizing the grid over it.

Our reference points are the established kinematics models used in PUFFERDRIVE: CLASSIC, which commands acceleration and steering from a fixed grid and enforces no comfort limits at all, and JERK, which commands jerk and integrates it into bounded accelerations. We take the steering *rate* as the default steering command, as every constrained model below uses it, and report the original steering-angle parameterization as CLASSIC (steer angle) alongside it. Against them we put three improved variants: JERK equipped with realistic comfort bounds, and two versions of CLASSIC that resolve their commands against the per-step feasible set: one by *clipping* the fixed grid into it (CLIPPED), one by *re-discretizing* the grid over it (ADAPTIVE).

Our contributions are as follows:

1. We show that unconstrained action spaces let learned policies inflate safety metrics through maneuvers that leave the comfort envelope of even aggressive human driving, and that naive clamping leads to grid collapse.
2. We propose a constraint-aware action parameterization that adaptively re-discretizes the per-step feasible control set.
3. We present PUFFERDRIVE-EDITOR, a browser-based tool for auditing realized kinematics and authoring targeted scenarios, providing a general debugging tool for exposing and analyzing failure cases in PUFFERDRIVE.

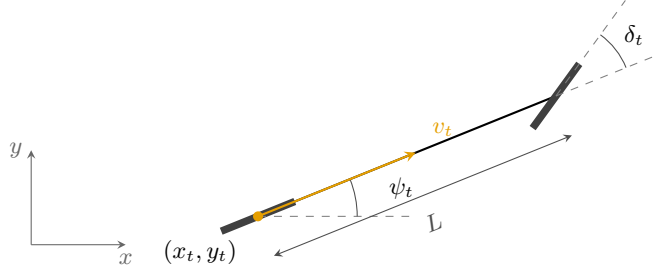


Fig. 1: Rear-axle kinematic bicycle model. Each axle is collapsed into a single wheel, separated by the wheelbase L . The state is the rear-axle position (x_t, y_t) (orange), heading ψ_t , signed speed v_t , and steering angle δ_t . The rear wheel does not slip, so the velocity (orange) points along the heading ψ_t and v_t is the rear axle’s own speed. Comfort is judged at this point.

2 Problem Description

We first describe the vehicle model that turns actions into motion (Sec. 2.1), then the behavioral comfort envelope that a realized trajectory should respect (Sec. 2.2), before reviewing how prior work enforces such constraints (Sec. 2.3) and showing why the two action spaces PufferDrive contains fail to do so (Sec. 2.4).

2.1 Kinematic bicycle model

Vehicle motion is commonly abstracted by the *kinematic bicycle model* [13, 17, 18]: the two wheels of each axle are collapsed into a single wheel on the longitudinal axis, tire slip is neglected, and the vehicle is described by its pose, speed, and front-wheel steering angle (Fig. 1). Despite its simplicity it tracks real vehicle trajectories closely in the moderate lateral-acceleration regime of everyday driving [17], which makes it the standard choice for planning and for driving simulators.

We parametrize the model at the *rear axle*. This is the reference point at which the tire-slip-free kinematic bicycle is exact: the rear wheel cannot slip sideways, so the rear axle’s velocity always points along the vehicle heading and its speed is simply the state speed v — no slip angle appears. Choosing it as the reference point makes the comfort quantities of Sec. 2.2 identical across every dynamics model we compare, including those that never model slip at all.

All controlled agents follow this model with state $(x_t, y_t, \psi_t, v_t, \delta_t)$ as visualized in Fig. 1. Given a commanded acceleration a_t and a steering rate $\dot{\delta}_t$, one step of length Δt reads

$$\begin{aligned} v_{t+1} &= \text{clip}(v_t + a_t \Delta t, \pm v_{\max}), & \delta_{t+1} &= \text{clip}(\delta_t + \dot{\delta}_t \Delta t, \pm \delta_{\max}), \\ \omega_{t+1} &= \frac{v_t \tan \delta_{t+1}}{L}, & \psi_{t+1} &= \psi_t + \omega_{t+1} \Delta t, \end{aligned} \quad (1)$$

with yaw rate ω and wheelbase L .

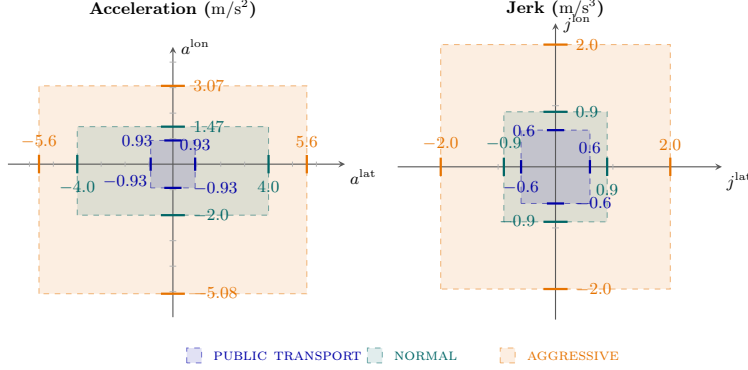


Fig. 2: Comfort limits of the OPM driver profiles of [3], marked on each axis; each profile lies strictly inside the next. We use NORMAL and AGGRESSIVE in our experiments and show PUBLIC TRANSPORT for reference. (Left) Acceleration plane ($a^{\text{lat}}, a^{\text{lon}}$) in m/s^2 ; the longitudinal limits of the driving profiles are asymmetric, since braking tolerance exceeds throttle tolerance. (Right) Jerk plane ($j^{\text{lat}}, j^{\text{lon}}$) in m/s^3 , with symmetric limits.

Comfort is judged on the *realized*, not the commanded, longitudinal and lateral accelerations, and both are read at the rear axle. The longitudinal acceleration is the change in rear-axle speed, and the lateral (centripetal) acceleration is that speed times the rear-axle yaw rate:

$$a_{t+1}^{\text{lon}} = \frac{v_{t+1} - v_t}{\Delta t}, \quad a_{t+1}^{\text{lat}} = v_{t+1} \omega_{t+1}, \quad \omega_{t+1} = \frac{v_t \tan \delta_{t+1}}{L}, \quad (2)$$

with the realized jerks the backward differences $j_{t+1}^{\text{lon}} = (a_{t+1}^{\text{lon}} - a_t^{\text{lon}})/\Delta t$ and $j_{t+1}^{\text{lat}} = (a_{t+1}^{\text{lat}} - a_t^{\text{lat}})/\Delta t$.

2.2 Comfort bounds on vehicle kinematics

Occupants perceive vehicle motion through acceleration and its time derivative, jerk; abrupt or sustained accelerations can cause discomfort and motion sickness [8, 25]. We adopt the Occupant's Preference Metric (OPM) of Bae *et al.* [3], which defines a behavioral comfort envelope via five parameters:

$$\text{OPM} = \{a_{\text{max}}^{\text{lon}}, a_{\text{min}}^{\text{lon}}, a_{\text{max}}^{\text{lat}}, j_{\text{max}}^{\text{lon}}, j_{\text{max}}^{\text{lat}}\}, \quad (3)$$

where $a_{\text{max}}^{\text{lon}}$ and $a_{\text{min}}^{\text{lon}}$ denote the throttle and braking thresholds, respectively, while the lateral acceleration and jerk limits are symmetric about zero. Accordingly, the five OPM parameters define lower and upper bounds on the four realized quantities of Sec. 2.1:

$$a^{\bullet} \in [\underline{a}^{\bullet}, \bar{a}^{\bullet}], \quad j^{\bullet} \in [\underline{j}^{\bullet}, \bar{j}^{\bullet}], \quad \bullet \in \{\text{lon}, \text{lat}\},$$

with $\underline{a}^{\text{lon}} = a_{\text{min}}^{\text{lon}}$, $\bar{a}^{\text{lon}} = a_{\text{max}}^{\text{lon}}$, $\underline{a}^{\text{lat}} = -a_{\text{max}}^{\text{lat}}$, $\bar{a}^{\text{lat}} = a_{\text{max}}^{\text{lat}}$, and $\underline{j}^{\bullet} = -j_{\text{max}}^{\bullet}$, $\bar{j}^{\bullet} = j_{\text{max}}^{\bullet}$. Figure 2 illustrates these limits. Bae *et al.* anchor the tightest profile in public-transport data, where freestanding passengers lose posture [15]; we show it for reference but drive with the two automotive profiles. The NORMAL profile uses seated-occupant thresholds, and the AGGRESSIVE profile uses observed human driving extremes; naturalistic driving studies likewise treat sustained high jerk not as a physical impossibility but as the signature of an aggressive driver [10].

These are *behavioral* envelopes, sitting well inside the physical handling limits: leaving the AGGRESSIVE band is not impossible, only something essentially no human driver would choose and no occupant would accept. We therefore call a control FEASIBLE relative to the enforced profile rather than in an absolute physical sense.

In the original OPM system, bounds are soft constraints enforced at the path-planning level [3]. Data-driven traffic simulators also evaluate these comfort quantities *post hoc* using metrics or Wasserstein distance penalties [16, 26]. In contrast, we use the comfort envelope to constrain the policy’s action space directly: at each step, we analytically compute the feasible set of controls whose realized body-frame quantities stay inside the envelope (Figure 2) and re-discretize the action grid over this set (Section 3.1), guaranteeing comfort by construction.

2.3 Kinematic constraints in simulation and reinforcement learning

Kinematic models in driving simulators. Data-driven simulators such as Nocturne [23], Waymax [11], and GPUDrive [12] share a common actuation interface: a kinematic bicycle model [13, 17] driven from a fixed grid of controls that is clamped only to static input ranges. Nothing bounds the realized accelerations or jerks, the gap our work targets in Section 3.

Feasibility in learned models. Prior works build kinematic constraints into learned motion models, such as routing trajectories through a bicycle-model layer (e.g., Deep Kinematic Models [5], Trajectron++ [19]) or propagating uncertainty through stochastic kinematic layers [27]. While these methods make *predicted trajectories* dynamically consistent by construction, they act downstream of action selection: a policy remains free to command an out-of-envelope maneuver that the layer then realizes rather than forbids. We instead constrain selection itself, bounding the feasible action set of a *closed-loop policy* at every simulation step under a behavioral comfort envelope. We expect this to matter most for the adaptive variant: by re-discretizing a full-resolution grid over the feasible set, it lets the policy optimize entirely within the envelope, so the constraint is reflected in what the policy learns rather than corrected after the fact.

Constrained reinforcement learning. The simplest approach leaves the action space untouched and penalizes violations in the reward, as Gigaflow [6] does for harsh acceleration and jerk. Comfort then competes with progress through a tunable weight: violations get rarer but are never ruled out. Hard enforcement is typically done via external layers that mask or project actions [4, 7, 14, 24].

However, when the feasible set shrinks (e.g., quadratically with speed), these methods suffer from grid collapse, where distinct discrete actions alias to the same boundary values. While action mapping [22] and jerk-bounded generators [1] address constrained action selection, our approach analytically computes a per-step feasible action region and adaptively re-discretizes the action grid onto it, preserving action resolution across varying vehicle speeds while guaranteeing feasible commands.

2.4 Default models and their failure modes

PufferDrive provides two kinematics models by default. CLASSIC commands an acceleration and a steering command directly from a fixed 7×13 grid, applied unchanged. Nothing ties consecutive commands together: within one step of $\Delta t = 0.1$ s the policy may switch from full braking to full throttle or slew the steering across its whole range, producing jerks orders of magnitude beyond any human comfort envelope. As a result, CLASSIC learns effective but kinematically implausible driving.

The steering command comes in two parameterizations: the steering *angle* δ_t , as originally implemented, or the steering *rate* $\dot{\delta}_t$, which integrates to the angle and therefore low-pass filters the steering without bounding anything. Every constrained model in this work commands the rate (Sec. 6), so we treat it as the default parameterization: CLASSIC denotes the steer-rate model throughout, and the original is written CLASSIC (steer angle).

JERK sits at the opposite end: the policy commands longitudinal and lateral *jerk*, which the environment integrates into accelerations clamped to the comfort band, so the limits hold by rule. It is not a separate dynamics model: the same kinematic bicycle model of Sec. 2.1 runs underneath, only re-parameterized one derivative up. The longitudinal jerk integrates to the speed, while the lateral jerk integrates to a lateral acceleration that is mapped back to a steering angle through the bicycle curvature relation $\delta = \arctan(\kappa L)$ with $\kappa = a^{\text{lat}}/v^2$. The disadvantage is that the action now operates on the second derivative of the velocity: credit assignment must propagate through $j \rightarrow a \rightarrow v \rightarrow x$, and the same jerk action produces different motion depending on the accumulated acceleration state. A single action barely changes the trajectory, and exploration in jerk space is much harder.

3 Method

We therefore introduce an action space in between: the policy commands an *acceleration and a steering rate* $(a_t, \dot{\delta}_t)$ — one derivative closer to the trajectory than jerk control — and the comfort limits are enforced inside the environment step by restricting each command to the set feasible from the current state (Sec. 3.1). Two variants of this restriction are compared (Sec. 3.2): *clipping*, which keeps CLASSIC’s absolute action semantics, and *adaptive re-discretization*, which re-maps the action grid onto the feasible set at each step so that actions never collapse onto the limit values.

3.1 Per-step feasible set

Combining the jerk and absolute limits yields per-axis target bands $\mathcal{A}_t^{\text{lon}}$ and $\mathcal{A}_t^{\text{lat}}$. The longitudinal band directly bounds the commanded acceleration a_t via $v_{t+1} = v_t + a_t \Delta t$. In contrast, the lateral band couples both control inputs:

$$a_{t+1}^{\text{lat}}(a_t, \dot{\delta}_t) = v_{t+1}(a_t) \omega_{t+1}(\dot{\delta}_t),$$

where

$$\omega_{t+1}(\dot{\delta}_t) = \frac{v_t \tan(\delta_t + \dot{\delta}_t \Delta t)}{L}. \quad (4)$$

Thus, the realized lateral acceleration depends on both the next speed, determined by a , and the next steering angle, determined by $\dot{\delta}$. The jointly feasible action set is

$$\mathcal{F}_t = \{(a_t, \dot{\delta}_t) : a_{t+1}^{\text{lon}}(a_t) \in \mathcal{A}_{t+1}^{\text{lon}}, a_{t+1}^{\text{lat}}(a_t, \dot{\delta}_t) \in \mathcal{A}_{t+1}^{\text{lat}}\}, \quad (5)$$

and is generally not separable into independent intervals.

A key observation is that ω_{t+1} depends only on the steering-rate command and the known current speed v_t . Consequently, for fixed $\dot{\delta}_t$, a_{t+1}^{lat} is linear in v_{t+1} , allowing the lateral band $\mathcal{A}_{t+1}^{\text{lat}}$ to be inverted analytically into an interval on v_{t+1} .

Constructing the feasible box. The box is a deterministic function of the current state. We first intersect the actuator-rate limits with the mechanical angle constraint $|\delta_t + \dot{\delta}_t \Delta t| \leq \delta_{\text{max}}$. Sampling this entire interval uniformly is ineffective at speed: the lateral band permits only a narrow interval of rates. We therefore invert $\mathcal{A}_{t+1}^{\text{lat}}$ at the nominal next speed $v_{\text{nom}} = v_t + a_t^{\text{lon}} \Delta t$, which yields candidate rates

$$\dot{\delta}_t = \frac{\arctan(L a_{t+1}^{\text{lat}} / (v_t v_{\text{nom}})) - \delta_t}{\Delta t}, \quad a_{t+1}^{\text{lat}} \in \mathcal{A}_{t+1}^{\text{lat}}.$$

The nominal-speed approximation only guides sampling: we widen the resulting rate window, intersect it with the hard rate and angle limits, and retain the full hard interval at low speed. We sample 17 rates centered on the prior rate and test each against the exact reachable acceleration interval.

For each sampled rate, ω_{t+1} is known. Intersecting the speed interval reachable through the longitudinal band with the interval obtained by dividing $\mathcal{A}_{t+1}^{\text{lat}}$ by ω_{t+1} gives the exact feasible acceleration interval for that rate: a vertical slice of $\mathcal{F}_t$ (Fig. 3a). For every contiguous feasible run, we intersect its acceleration intervals to form a rectangle. The retained box $[a_t, \bar{a}_t] \times [\dot{\delta}_t, \bar{\dot{\delta}}_t]$ is the candidate with the largest normalized area; when the anchor sample is feasible, candidates must contain it. Its middle discrete action therefore repeats the last action whenever compatible with comfort. The box is an inner approximation of $\mathcal{F}_t$, recomputed at every control step, and every command it contains is feasible by construction.

3.2 Clipped and adaptive action mapping

Both variants draw from the same nominal 7×13 grid of accelerations and steering rates; they differ in what an action index *means* (Fig. 3b).

Clipped. Action indices keep fixed physical values, and the requested command is clipped componentwise into the box. Semantics are stable, but whenever the box is smaller than the grid span — the typical regime for steering at speed — many indices collapse onto the same box border: effectively only the extreme commands remain distinguishable, and the policy cannot tell from the action alone what will actually be executed.

Adaptive. To ensure a fair comparison with the CLASSIC model in PUFFER-DRIVE, we use the same number of discrete actions (7×13), while remapping them to acceleration and steering-rate commands. These discrete actions are mapped to acceleration and steering-rate commands and re-discretized onto the current feasible box at every step using a piecewise-linear *anchored map*: for an interval $[\ell_t, h_t]$ and anchor c_t , the middle index maps exactly to c_t , the extreme indices to the borders, and intermediate indices interpolate linearly on each side. The anchor is chosen as the previous step’s *realized* action whenever it remains feasible under the current box; otherwise, the interval is mapped without an anchor. When used, the anchor gives the middle action the semantics “repeat what you just did”. All 7×13 actions remain distinct and feasible by construction — nothing saturates — at the price of *relative* semantics: the same index denotes different physical commands in different states.

3.3 Observations

The feasible box depends on the previous realized accelerations and steering rate, which are not visible in the instantaneous pose — a policy without this information would face a partially observed problem, since identical observations could resolve the same action differently. Both variants therefore observe, in addition to the standard ego features, the previous realized action (the anchor) and the four box edges, normalized by the profile limits. The two variants share the identical observation layout, so their comparison isolates the action-mapping strategy alone.

4 Experiments

We evaluate our comfort-bounded kinematic models against unconstrained and naively bounded baselines. Our experiments are designed to address two key questions: **(1) Limit Compliance:** does enforcing the comfort envelope actually eliminate envelope-violating motion? **(2) Navigation Cost:** what, if any, is the cost of these comfort constraints on navigation performance?

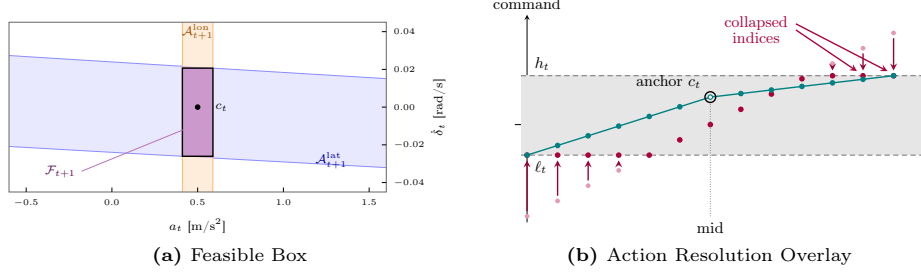


Fig. 3: Construction and utilization of the feasible action space. **(a)** Per-step construction of the feasible action box (computed from an actual vehicle state; see text for parameters). The orange band denotes the longitudinal target interval $\mathcal{A}_{t+1}^{\text{lon}}$, while the blue region contains all $(a_t, \dot{\delta}_t)$ pairs satisfying the lateral target band $\mathcal{A}_{t+1}^{\text{lat}}$. Their intersection (purple) is the feasible set $\mathcal{F}_{t+1}$, from which the largest axis-aligned rectangle (black) is retained. **(b)** Comparison of action resolution strategies against the feasible interval $[\ell_t, h_t]$. In the **Clipped** approach (purple arrows and dark purple dots), fixed grid values saturate at the borders, causing multiple indices to collapse. In the **Adaptive** approach (teal line and dots), the grid is directly re-discretized onto the interval, locking the middle index to the anchor c_t (circled) and spreading extremes exactly to the borders so that all indices stay distinct and feasible.

4.1 Experimental Setup

We evaluate our models in PUFFERDRIVE, a data-driven 2D bird’s-eye view simulator built on the Waymo Open Motion Dataset (WOMD) [9] and parallelized via PufferLib [21]. All vehicles in a scene are controlled simultaneously for episodes of 9.1 s ($\Delta t = 0.1$ s), with agents removed upon collision, leaving the drivable area, or reaching their goal. We train on 80,000 WOMD scenarios and evaluate on the held-out validation split of 10,000 scenarios.

Policies are trained using Proximal Policy Optimization (PPO) [20] for 2 billion steps across three seeds, using a recurrent actor-critic network and simple reward terms such as collision, offroad and goal reaching. All configurations share identical hyperparameters and training budgets, differing only in their action spaces and bound enforcements.

4.2 Model Configurations

We compare the six kinematic models introduced above. The three *unconstrained* baselines, CLASSIC, CLASSIC (steer angle) and JERK (Section 2.4), check no action against the comfort envelope (Figure 2). The first two differ only in whether the policy commands the steering rate or the steering angle. The pair isolates the effect of that interface choice from the effect of enforcing the envelope. The other three enforce the envelope, in increasing order of how much of the feasible set they preserve: JERK (bounded grid) restricts the jerk grid a priori with a single static conservative grid; Clipped CLASSIC clips each command into the per-step feasible box axis by axis; and Adaptive CLASSIC (ours) re-discretizes its grid onto



Fig. 4: The PUFFERDRIVE-EDITOR user interface. The upper panel displays the map canvas and vehicle trajectory paths (e.g., analyzing the reactive policy in orange). The lower panel contains the real-time kinematics dashboard, showing the active passenger-comfort limits (green/amber boxes) for acceleration and jerk, the vehicle’s actual realized state trajectory (highlighted dots and trail), and target action re-discretization intervals.

that box, so every selectable action is feasible by construction (Section 3.2). Every CLASSIC model except the one qualified as (steer angle) commands steering *rate* rather than angle (Section 6).

The comfort envelope is a property of the simulator, not the model: each constrained model builds its feasible set from either the AGGRESSIVE or the tighter NORMAL band (Figure 2). Unless stated otherwise, results enforce the AGGRESSIVE band.

4.3 Evaluation Protocol

Each trained policy is evaluated on the validation split, and we report mean $\pm$ standard deviation over the three seeds.

We report three navigation metrics (the share of agents reaching their goal, the share colliding, and the share leaving the drivable area) and, for each of the four kinematic quantities ($a^{\text{lon}}, a^{\text{lat}}, j^{\text{lon}}, j^{\text{lat}}$), the share of steps that violate the comfort envelope. Two properties of this tally matter for reading the results. First, we measure the *realized* quantity, not the commanded one: a model that bounds its inputs but still drives outside the envelope through the coupling terms is counted as violating, which is precisely the failure mode naive clamping exhibits. Second, the tally covers only the steps an agent actually drives (steps after removal or goal completion are excluded), so the rate is a share of driving rather than of episode length, which would otherwise reward a model for finishing early.

4.4 Qualitative Analysis and PUFFERDRIVE-EDITOR

Aggregate metrics over the validation split average away the rare, safety-critical moments in which the limits actually bind. To inspect behavior near those limits we present PUFFERDRIVE-EDITOR, a browser-based scene editor and policy-debugging interface that runs locally on the map and scenario binaries loaded by PufferDrive. Although motivated by the analysis of our constraint-aware kinematic models, PUFFERDRIVE-EDITOR is a general tool for understanding learned driving policies in PufferDrive, enabling researchers to identify policy strengths, weaknesses, and failure modes beyond aggregate metrics. It combines three functions. A *scenario and track editor* loads road layouts from scenario binaries and lets researchers author stress tests by moving or inserting static obstacles and placing dynamic vehicles with custom poses, velocities, trajectories, and goals (e.g., narrow corridors or parked cars); this isolates skills such as slalom maneuverability and short-distance emergency braking. A *real-time kinematics dashboard* plots the realized metrics at every step against the OPM boundaries (Fig. 4), highlighting comfort exceedances in red so it is immediately visible whether a breach occurred in acceleration or jerk. Finally, *multi-model auditing* compares two rollouts either side by side, with synchronized timelines and feasible action boxes $\mathcal{F}(s)$, or overlaid on a single map, making visible where a policy under naive clamping saturates its steering and leaves the lane while the adaptive model preserves the authority to negotiate the obstacle.

5 Results

We now return to the two questions that motivated our experiments: whether enforcing the comfort envelope actually eliminates envelope-violating motion, and what it costs in navigation performance. We first report the aggregate numbers across all six kinematic models (Section 5.1), then turn to the stress-test scenes that expose the mechanism behind those numbers (Section 5.2).

5.1 Quantitative Performance on PufferDrive

Table 1 reports navigation performance and the realized-kinematics violation rate for all six kinematic models over three seeds; Figure 5 resolves the violation columns into the full driving-zone distribution.

Unconstrained control drives outside the envelope. The unconstrained CLASSIC baselines achieve the highest goal completion performance, but only by driving outside the comfort limits. They exceed even the AGGRESSIVE comfort envelope on far more than a quarter of their steps, mostly in lateral jerk, followed by longitudinal jerk and lateral acceleration, with maximum violations reaching 4513% for jerk and 93% for acceleration. In comparison, the constrained variants only exceed the bounds when e.g. the physical steering-angle limit prevents their enforcement, with maximum violations of just 0.26% for jerk and $9 \times 10^{-4}\%$ for

acceleration. Which steering command the policy holds barely matters for the performance metrics, the steer-rate and steer-angle variants are within noise of each other in goal completion and collisions. Commanding the rate integrates the steering and thereby reduces the violations that a single abrupt angle change can produce. While the unconstrained JERK baseline removes acceleration violations by construction, it worsens jerk violations because the action space of the jerk model is already partially outside the aggressive bounds.

Enforcing the envelope costs little. Both clipping and adaptive re-discretization reduce comfort violations to essentially zero (Fig. 5), while reducing goal completion by around two percentage points compared to the unconstrained Classic baseline. In contrast, the bounded-grid Jerk baseline achieves lower navigation performance, indicating that a single static action grid is overly restrictive across varying vehicle speeds. Among the per-step projection methods, adaptive re-discretization consistently performs best: it matches clipping’s comfort guarantees while achieving slightly higher goal completion and lower collision and off-road rates under the aggressive profile. The same trend is maintained under the tighter normal profile.

The constraint does not make policies drive conservatively. Zero violations alone would be insufficient evidence if the policy simply avoided operating near the limits. Figure 5 shows the opposite: the adaptive model spends $\approx 74\%$ of driven steps in the *aggressive* longitudinal-acceleration band and roughly half in the aggressive jerk bands. It operates at the very edge of the safety envelope without crossing it; in contrast, the unconstrained baselines mostly fluctuate between normal operation and violation states for longitudinal and lateral jerk.

Tightening the envelope. Table 1 repeats the comparison under the tighter NORMAL comfort profile, where the models may only command actions within the comfortable band. All families degrade by a similar margin (≈ 4 points of goal reaching) as the feasible set shrinks, and adaptive re-discretization remains the strongest constrained driver: the adaptive steer-rate model reaches 92.73% goal completion, the best of any constrained model under the normal profile, while keeping violations at zero because it preserves full action resolution inside whatever box remains rather than collapsing a fixed grid onto its boundary.

5.2 Qualitative Analysis in Stress-Test Scenes

Aggregate metrics conceal what happens where the limits bind. Inspecting roll-outs in the hand-authored stress-test scenes (Section 4.4) with PUFFERDRIVE-EDITOR’s dashboard (Fig. 4), the unconstrained configurations show acceleration and jerk spikes leaving the envelope just before sharp turns or collisions, while our adaptive model keeps inputs smooth and strictly inside the bounds. Figure 6 makes this concrete on a slalom under four models: the bounded-grid JERK baseline deviates from the reference and leaves its goals unreached, whereas our adaptive model tracks the route through the slalom and reaches its goals cleanly within the comfort bounds.

Table 1: Navigation performance per dynamics model on the PufferDrive validation split, under the AGGRESSIVE and NORMAL comfort profiles, averaged over three seeds, in percent. Unconstrained baselines do not enforce comfort limits and are reported once. Navigation columns give mean $\pm$ std. **Bold** marks the best value per column *within each outer group* (Unconstrained, Aggressive, or Normal): the unconstrained baselines buy their navigation numbers with envelope-violating control, so comparing them against constrained models would be unfair.

Dynamics model	Goal $\uparrow$	Collisions $\downarrow$	Offroad $\downarrow$
<i>Unconstrained</i>			
CLASSIC	98.74 $\pm$ 0.03	0.48 $\pm$ 0.02	0.25 $\pm$ 0.02
CLASSIC (steer angle)	98.79 $\pm$ 0.03	0.49 $\pm$ 0.02	0.27 $\pm$ 0.02
JERK	97.68 $\pm$ 0.02	1.16 $\pm$ 0.03	0.53 $\pm$ 0.04
<i>Aggressive</i>			
JERK (bounded grid)	96.39 $\pm$ 0.09	1.79 $\pm$ 0.04	1.00 $\pm$ 0.07
Clipped CLASSIC	96.77 $\pm$ 0.06	1.68 $\pm$ 0.02	0.80 $\pm$ 0.05
Adaptive CLASSIC	96.89 $\pm$ 0.04	1.65 $\pm$ 0.03	0.71 $\pm$ 0.02
<i>Normal</i>			
JERK (bounded grid)	92.32 $\pm$ 0.08	2.39 $\pm$ 0.06	1.27 $\pm$ 0.07
Clipped CLASSIC	92.51 $\pm$ 0.03	2.12 $\pm$ 0.02	1.22 $\pm$ 0.02
Adaptive CLASSIC	92.73 $\pm$ 0.04	2.27 $\pm$ 0.01	1.15 $\pm$ 0.01

6 Discussion

The quantitative results and qualitative profiles validate our hypotheses regarding RL action spaces under constraints, yielding several insights.

Comfort vs. navigability trade-off. Moving from the AGGRESSIVE to the NORMAL profile consistently drops goal completion by roughly 4 percentage points and raises collisions, as the tighter bounds curb aggressive braking and acceleration: passenger comfort directly constrains defensive reactivity. One might worry that bounding the action space removes the emergency maneuvers a policy needs to stay safe, but the envelope is a configurable profile rather than a fixed ceiling. The AGGRESSIVE band sits at the extremes of human driving, so the policy keeps the full range a human would use and forgoes only motion beyond it, and a deployment wanting a wider margin can enforce a looser profile. Empirically the constraint rarely binds, costing under two points of goal completion against the unconstrained baseline while removing nearly all violations.

Adaptive mapping beats clipping and jerk control. Adaptive mapping outperforms clipped steering and jerk control, mainly in safety-critical navigation. While all constrained methods achieve similar goal completion, adaptive re-discretization preserves steering resolution within the feasible range, improving control authority when avoiding collisions and off-road events. Clipping instead saturates the fixed action grid as the steering range shrinks at high speed. Jerk

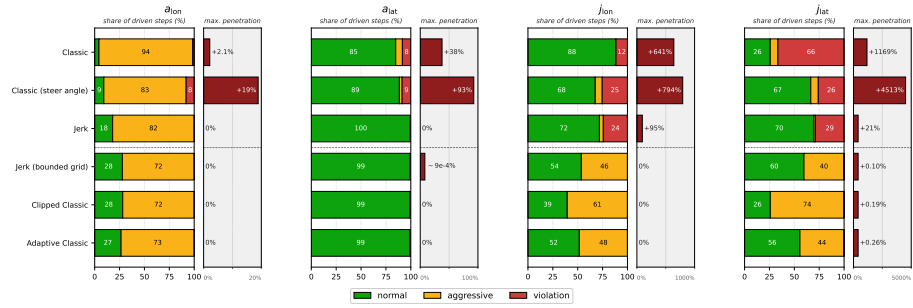


Fig. 5: Driving-zone occupancy of the *realized* kinematics: the share of driven steps whose realized quantity falls in the normal, aggressive, or violating band of the comfort envelope, per kinematic model (mean over three seeds). The constrained models enforce the AGGRESSIVE comfort profile (bounded constraints); the unconstrained baselines enforce no limits. The adaptive model fills the aggressive band without ever leaving it; the unconstrained baselines place a large share of their mass outside the envelope. The shaded column adds how *far* outside: the max penetration, i.e. the worst-case overshoot past the crossed bound relative to that bound (+100% = twice the limit). Its bars are proportional within a panel, each scaled to the axis printed below it.

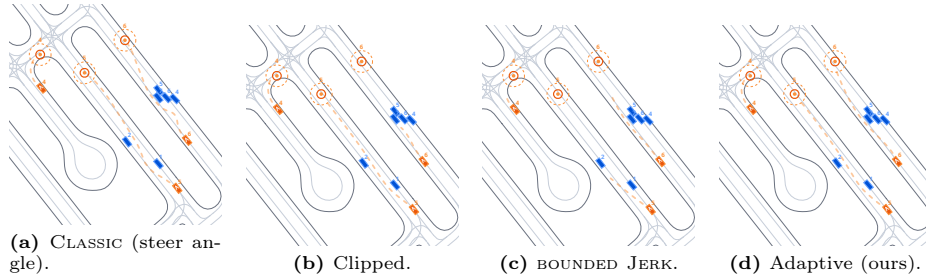


Fig. 6: The same stress-test scenes under four kinematic models, zoomed to the identical region around the policy-controlled route. Policy-controlled agents are orange (numbered 4–6, dashed routes; goals shown as numbered ringed dots); parked agents are blue. Our adaptive model (d) tracks the reference through the scene and reaches all goals within the comfort bounds, whereas the bounded-grid JERK baseline (c) deviates from the route and leaves its goals unreached (empty rings); the unconstrained CLASSIC (steer angle) (a) and clipped (b) variants are shown for reference.

control is weakest in these hard-navigation settings because second-order actions introduce an additional integration chain ($j \rightarrow a \rightarrow v \rightarrow x$), making the control problem harder to learn and credit assignment more difficult. Thus, while jerk control is simple to constrain, direct first-order control is better suited for precise obstacle avoidance and boundary recovery.

Action-space design: rate vs. angle, hard vs. soft. We command steering rate ($\dot{\delta}$) rather than steering angle (δ) by default, and all constrained CLASSIC vari-

ants use it. Rate control provides a useful balance between responsiveness and smoothness: it avoids abrupt steering changes while retaining more control authority than jerk-based control, which introduces an additional integration chain. The two unconstrained variants are compared in Table 1 to isolate the effect of the steering action space where both perform comparably bad. Alternatively, comfort can be encouraged through reward penalties, as Gigaflow [6] does for harsh acceleration and jerk. Such a penalty prices comfort against other objectives through a tunable weight rather than ruling a maneuver out, so the optimal policy still accepts a violation whenever the payoff of an evasive maneuver exceeds its cost. Reward design therefore shapes driving preferences but cannot reliably define which actions should be considered unacceptable. Hard action constraints and reward shaping are complementary: the constraint makes out-of-envelope motion unreachable by construction and so guarantees compliance with the modeled comfort envelope [2, 7], while the reward can optimize driving style within the feasible region.

6.1 Conclusion

We introduced a constraint-aware, adaptive action-mapping framework to address the reliance of learned driving policies on maneuvers no human driver would perform and no occupant would accept. By analytically inverting the lateral-jerk constraint in next-steering-angle space, our approach dynamically scales the control grid to span exactly the comfort envelope at each step. On WOMD and a hand-authored slalom stress test, the proposed adaptive steer-rate CLASSIC model enforced comfort compliance (limiting comfort violations to below 0.01%) and outperformed clipped and direct-jerk baselines in navigability. Because every action the policy can select now lies inside the enforced comfort envelope, the resulting safety metrics reflect policy quality rather than the simulator’s permissiveness, closing the gap that motivated this work. Interactive audits via PUFFERDRIVE-EDITOR confirmed that adaptive re-discretization preserves steering authority and prevents grid collapse, allowing the policy to negotiate the slalom’s tight turns and narrow gaps.

Acknowledgement

This work is part of BrainLinks-BrainTools which is funded by the Federal Ministry of Economics, Science and Arts of Baden-Württemberg within the sustainability program for projects of the excellence initiative II. Also, this research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant number 539134284, through EFRE (FEIH_2698644) and the state of Baden-Württemberg.



Baden-Württemberg



**Co-funded by
the European Union**

References

1. Alizadeh Kolagar, S.A., Heydari Shahna, M., Mattila, J.: Combining deep reinforcement learning with a jerk-bounded trajectory generator for kinematically constrained motion planning. arXiv preprint arXiv:2410.20907 (2024)
2. Alshiekh, M., Bloem, R., Ehlers, R., Könighofer, B., Niekum, S., Topcu, U.: Safe reinforcement learning via shielding. In: AAAI (2018)
3. Bae, I., Moon, J., Jhung, J., Suk, H., Kim, T., Park, H., Cha, J., Kim, J., Kim, D., Kim, S.: Self-driving like a human driver instead of a robocar: Personalized comfortable driving experience for autonomous vehicles. In: Machine Learning for Autonomous Driving Workshop at the 33rd Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada (2019)
4. Cheng, R., Orosz, G., Murray, R.M., Burdick, J.W.: End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In: AAAI. pp. 3387–3395 (2019)
5. Cui, H., Nguyen, T., Chou, F.C., Lin, T.H., Schneider, J., Bradley, D., Djuric, N.: Deep kinematic models for kinematically feasible vehicle trajectory predictions. In: IEEE Int. Conf. Robot. Autom. (ICRA). pp. 10563–10569 (2020)
6. Cusumano-Towner, M., Hafner, D., Hertzberg, A., Huval, B., Petrenko, A., Vinit-sky, E., Wijmans, E., Killian, T., Bowers, S., Sener, O., Koltun, V., Krähenbühl, P.: Robust autonomy emerges from self-play. In: ICML. Proceedings of Machine Learning Research (PMLR), vol. 267, pp. 11710–11737 (2025)
7. Dalal, G., Dvijotham, K., Vecerik, M., Hester, T., Paduraru, C., Tassa, Y.: Safe exploration in continuous action spaces. arXiv preprint arXiv:1801.08757 (2018)
8. Elbanhawi, M., Simic, M., Jazar, R.: In the passenger seat: Investigating ride comfort measures in autonomous cars. IEEE Intelligent Transportation Systems Magazine **7**(3), 4–17 (2015)
9. Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C.R., Zhou, Y., Yang, Z., Chouard, A., Sun, P., Ngiam, J., Vasudevan, V., McCauley, A., Shlens, J., Anguelov, D.: Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In: ICCV. pp. 9690–9699 (2021)
10. Feng, F., Bao, S., Sayer, J.R., Flannagan, C., Manser, M., Wunderlich, R.: Can vehicle longitudinal jerk be used to identify aggressive drivers? an examination using naturalistic driving data. Accident Analysis & Prevention **104**, 125–136 (2017). <https://doi.org/10.1016/j.aap.2017.04.012>
11. Gulino, C., Fu, J., Luo, W., Tucker, G., Bronstein, E., Lu, Y., Harb, J., Pan, X., Wang, Y., Chen, X., Co-Reyes, J.D., Agarwal, R., Roelofs, R., Lu, Y., Montali, N., Mougin, P., Yang, Z., White, B., Faust, A., McAllister, R., Anguelov, D., Sapp, B.: Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. In: NeurIPS (2023)
12. Kazemkhani, S., Pandya, A., Cornelisse, D., Shacklett, B., Vinitsky, E.: GPUDrive: Data-driven, multi-agent driving simulation at 1 million FPS. In: ICLR (2025)
13. Kong, J., Pfeiffer, M., Schildbach, G., Borrelli, F.: Kinematic and dynamic vehicle models for autonomous driving control design. In: IEEE Intelligent Vehicles Symposium (IV). pp. 1094–1099 (2015)
14. Krasowski, H., Wang, X., Althoff, M.: Safe reinforcement learning for autonomous lane changing using set-based prediction. In: IEEE Int. Conf. Intell. Transp. Syst. (ITSC) (2020)

15. Martin, D., Litwhiler, D.: An investigation of acceleration and jerk profiles of public transportation vehicles. In: ASEE Annual Conference & Exposition (2008)
16. Montali, N., Lambert, J., Mougin, P., Kuefler, A., Rhinehart, N., Li, M., Gulino, C., Emrich, T., Yang, Z., Whiteson, S., White, B., Anguelov, D.: The waymo open sim agents challenge. In: NeurIPS (2023)
17. Polack, P., Altché, F., d’Andréa Novel, B., de La Fortelle, A.: The kinematic bicycle model: A consistent model for planning feasible trajectories for autonomous vehicles? In: IEEE Intelligent Vehicles Symposium (IV). pp. 812–818 (2017)
18. Rajamani, R.: Vehicle Dynamics and Control. Mechanical Engineering Series, Springer, 2nd edn. (2012). <https://doi.org/10.1007/978-1-4614-1433-9>
19. Salzmann, T., Ivanovic, B., Chakravarty, P., Pavone, M.: Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In: ECCV. pp. 683–700 (2020)
20. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
21. Suárez, J.: PufferLib: Making reinforcement learning libraries and environments play nice. In: Agent Learning in Open-Endedness Workshop at the Conference on Neural Information Processing Systems (NeurIPS) (2023)
22. Theile, M., Dirnberger, L., Trumpp, R., Caccamo, M., Sangiovanni-Vincentelli, A.L.: Action mapping for reinforcement learning in continuous environments with constraints. Reinforcement Learning Journal **6** (2025)
23. Vinitsky, E., Lichtlé, N., Yang, X., Amos, B., Foerster, J.: Nocturne: A scalable driving benchmark for bringing multi-agent learning one step closer to the real world. In: NeurIPS (2022)
24. Wang, X., Krasowski, H., Althoff, M.: CommonRoad-RL: A configurable reinforcement learning environment for motion planning of autonomous vehicles. In: IEEE Int. Conf. Intell. Transp. Syst. (ITSC) (2021)
25. de Winkel, K.N., Irmak, T., Happee, R., Shyrokau, B.: Standards for passenger comfort in automated vehicles: Acceleration and jerk. Applied Ergonomics **106**, 103881 (2023)
26. Xu, D., Chen, Y., Ivanovic, B., Pavone, M.: BITS: Bi-level imitation for traffic simulation. In: IEEE Int. Conf. Robot. Autom. (ICRA). pp. 2929–2936 (2023)
27. Zheng, L., Son, S., Liang, J., Wang, X., Clipp, B., Lin, M.C.: Deep stochastic kinematic models for probabilistic motion forecasting in traffic. arXiv preprint arXiv:2406.01431 (2024)