CUSP: Cross-domain Uncertainty-aware Driving Style Posterior

Xinyue Liu
Department of Computer Science
University of Manchester
Manchester, United Kingdom
xinyue.liu-9@postgrad.manchester.ac.uk

Fanlin Meng*
University of Exeter Business School
University of Exeter
Exeter, United Kingdom
f.meng2@exeter.ac.uk
* *Corresponding author*

Xiao-Jun Zeng
Department of Computer Science
University of Manchester
Manchester, United Kingdom
x.zeng@manchester.ac.uk

***Abstract*—Inferring driving style from real-world trajectories is challenging under domain shift: datasets collected across regions or sites differ in speed distributions, congestion regimes, sensing noise, and annotation quality, causing learned "styles" to absorb domain-specific shortcuts such as absolute speed. We propose CUSP, an uncertainty-aware framework that represents driving style with interpretable mechanism parameters of the Intelligent Driver Model (IDM), such as desired speed and time headway, and infers a posterior from short interaction windows under domain shift. CUSP performs scalable amortised posterior inference and learns the parameters via a physics-grounded likelihood with variational regularisation. To keep style semantics comparable across datasets, we stabilise the inferred parameters by discouraging domain leakage and absolute-speed shortcuts through adversarial training, optionally aided by masking speed-related channels and using simple neighbourhood context. Experiments under cross-location and cross-site evaluations on highD and NGSIM show that CUSP preserves mechanistic fidelity while reducing speed and domain information in the inferred parameters, yielding more stable style semantics and uncertainty signals that support downstream interaction modelling.**

***Index Terms*—Driving style, Intelligent Driver Model, Domain shift, Uncertainty**

I. Introduction

Autonomous driving decision-making relies on understanding and predicting the behaviour of surrounding vehicles. Yet drivers differ in how they interact, motivating the characterisation of such behavioural differences. Driving style captures stable preferences and risk attitudes in traffic interactions, reflected in choices such as headway, acceleration, and yielding. As a result, driving style systematically affects gap acceptance, acceleration and braking responses, and the trade-off between safety, efficiency, and comfort [1]. Despite its importance for interaction modelling, driving style still lacks a unified and operational definition [2].

In real-world trajectory data, driving style labels are rarely available, so style is usually inferred indirectly from trajectories [3]. Because observed behaviour is shaped both by what a driver prefers and intends and by what the environment allows, such inference is inherently ambiguous [4]. The same outward behaviour, such as a lane change or close following, may reflect a driver's preferences and intent, yet it may be largely driven by external factors [5]. These factors include traffic opportunities such as available gaps and the feasible action space, local regulations, and dataset-specific collection conditions. Moreover, datasets collected in different regions and road segments often exhibit systematic differences in speed distributions, congestion patterns, and traffic rules [6].

Without supervision, a model may therefore learn styles that absorb these domain-specific factors, such as speed limits or congestion levels, as if they were the style itself [7]. This can create semantic inconsistency across datasets and lead to semantic drift. It also reduces transfer performance, precisely in cross-domain settings where robust interaction modelling is most needed and most challenging [8]. Therefore, we aim to make driving style operational by representing it with interpretable parameters of interaction mechanisms. From short interaction windows, we infer a posterior over these parameters and suppress domain- and opportunity-induced shortcuts to keep their semantics stable under domain shift. Existing efforts fall into three categories: mechanistic calibration of microscopic driving models, probabilistic calibration, and learning-based style representations.

The first category characterises style by binding it to interpretable parameters of microscopic driving models, such as car-following and lane-changing models [9]. These parameters correspond to behavioural notions such as desired time headway, comfortable deceleration, and yielding or politeness tendencies. Classical calibration typically fits parameters per vehicle or per trajectory window via nonlinear iterative optimisation [10], which is often non-convex and sensitive to initialisation. As data scale increases and cross-dataset evaluation becomes necessary, the cost of per-window optimisation quickly accumulates [11]. The second category, probabilistic and Bayesian calibration, treats parameters as random variables to provide uncertainty estimates and capture individual differences [12], [13]. Existing sampling or approximate methods are computationally expensive at scale. Thus, a scalable probabilistic framework for interpretable style inference and cross-dataset evaluation remains elusive. The third category learns style representations using latent embeddings or discriminative features such as speed, accel-

eration, headway, and jerk [14]. These methods scale well and integrate naturally with prediction or planning in an end-to-end manner. Yet under domain shift, they are prone to exploiting domain-specific shortcuts that correlate with the training distribution but do not reflect stable preferences [15]. In particular, absolute speed and closely related cues can be highly predictive in the source domains, even though they largely reflect speed limits and traffic regimes rather than driver intent. When speed limits, congestion regimes, or road geometry change, speed distributions can shift across domains, causing style semantics to drift. In constrained traffic, observed behaviour is often dominated by what the environment allows rather than what a driver prefers. Although some methods output uncertainty, few provide diagnostics to identify uninformative windows and avoid overconfident but distorted estimates when opportunities are limited. In addition, this semantic drift raises a practical issue: ensuring that each learned style dimension preserves the same meaning across datasets remains difficult when style labels are unavailable [16]. Trajectory datasets can differ substantially in their distributions, and domain shift can severely affect safety-critical learning components [17]. Measurement issues in speed and acceleration may further distort model-based inference. Yet existing driving style methods often lack mechanisms to prevent domain-specific factors from contaminating the style representation. As a result, variations induced by speed limits, rule differences, and congestion can be absorbed into the inferred style rather than attributed to context. Without style labels, it remains difficult to diagnose whether each inferred style dimension preserves a consistent mechanism-level interpretation across datasets [18].

To address these challenges, we propose CUSP, an uncertainty-aware driving style inference framework under domain shift. CUSP represents style as a vector of interpretable parameters of a microscopic interaction model, instantiated with the Intelligent Driver Model (IDM) [9], and infers a posterior from short interaction windows with optional context. It performs scalable amortised posterior inference and learns mechanism parameters through a physics-grounded likelihood with variational regularisation. To preserve cross-domain semantic stability, CUSP reduces domain and absolute-speed leakage via adversarial training, optionally aided by simple input interventions such as masking speed-related channels and leveraging neighbourhood context features. We evaluate on highD [19] and NGSIM [20] with cross-location splits, showing more stable style semantics, lower leakage, and useful uncertainty for downstream interaction modelling. The main contributions of this paper are:

- We propose a scalable amortised framework for window-level probabilistic inference of interpretable mechanism-based style parameters.
- We introduce an anti-shortcut and diagnostic protocol that reduces speed and domain leakage via adversarial training, with masking of speed-related channels and neighbourhood context.
- We evaluate under cross-location and cross-site splits on real-world trajectory datasets, demonstrating improved mechanistic fidelity, reduced leakage, and informative uncertainty for downstream interaction modelling.

II. Methodology

A. Problem Setup

We study window-level driving-style inference under domain shift across datasets and sites with different traffic regimes and measurement noise. We represent driving style with interpretable parameters of a microscopic interaction model, instantiated with IDM. We extract fixed-length windows of horizon T_w, where each window is represented by a feature sequence $\tau = \{x_t\}_{t=1}^{T_w}$ and an optional context vector c. Each x_t concatenates interaction features (e.g., kinematic variables and gap-related signals), while c summarises window-level or slowly varying information (e.g., mean speed and simple neighbourhood descriptors). Given (τ, c), our goal is to infer a posterior over IDM parameters, $q_\phi(\theta \mid \tau, c)$. Training uses mechanistic supervision from observed trajectories via an IDM likelihood and a KL regulariser, and optionally a cross-domain stabilisation term. A domain index d is used only during training for regularisation (and domain-dependent noise modelling); it is not required at inference, where the model takes only (τ, c) as input.

B. Overview of CUSP

We introduce CUSP, an uncertainty-aware framework for inferring an interpretable driving style under domain shift from short interaction windows. CUSP has two components:

- **Window-level posterior inference:** Given (τ, c), an amortised inference network predicts a posterior over mechanism parameters. It is trained with a physics-grounded likelihood and variational regularisation.
- **Cross-domain semantic stabilisation:** We impose a cross-domain regularisation on the inferred parameters to discourage shortcut cues and preserve semantic consistency across domains.

Fig. 1 summarises the pipeline: CUSP maps each interaction window (τ, c) to a posterior $q_\phi(\theta \mid \tau, c)$, where θ parameterises the underlying interaction model.

C. Window-level Posterior Inference

Given (τ, c), CUSP outputs a posterior over IDM parameters $\theta = (v_0, T, s_0, a, b)$ [9], where v_0 is desired speed, T time headway, s_0 jam distance, a maximum acceleration, and b comfortable deceleration. Here $\tau = \{x_t\}_{t=1}^{T_w}$ is a fixed-length interaction feature sequence, and c summarises window-level context. Fig. 2 summarises the inference.

1) Posterior inference: We infer a diagonal-Gaussian posterior in a latent space and map it to physically bounded IDM parameters. An encoder maps the window τ to a representation h, and a posterior head outputs $(\mu, \log \sigma^2)$:

$$\begin{aligned} &h = \mathrm{Enc}_\phi(\tau), \qquad (\mu, \log \sigma^2) = f_\phi([h, c]), \\ &q_\phi(z \mid \tau, c) = \mathcal{N}\big(\mu, \mathrm{diag}(\sigma^2)\big), \end{aligned} \tag{1}$$

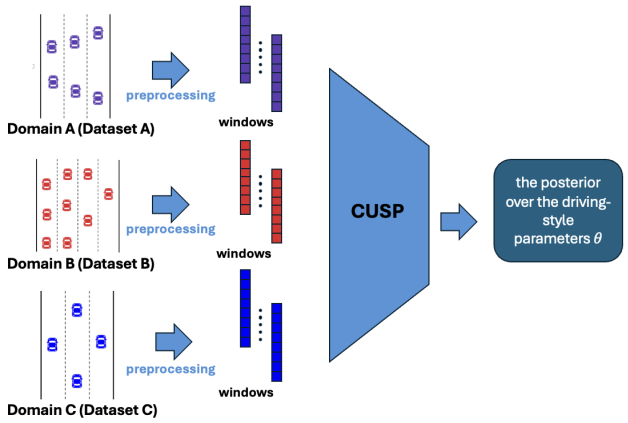

Fig. 1. CUSP maps interaction windows to a posterior over θ.

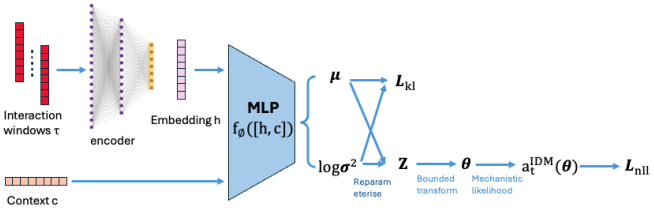

Fig. 2. Window-level posterior inference.

where ϕ denotes parameters and $\mathrm{diag}(\sigma^2)$ denotes a diagonal covariance matrix. We sample with reparameterisation:

$$\epsilon \sim \mathcal{N}(0, I), \qquad z = \mu + \sigma \odot \epsilon, \tag{2}$$

where $\odot$ denotes element-wise multiplication and I is the identity matrix. To enforce physical plausibility, we map z to bounded parameters by an element-wise squashing transform:

$$\theta = \theta_{\min} + (\theta_{\max} - \theta_{\min}) \odot \mathrm{sigmoid}(z), \tag{3}$$

where $\mathrm{sigmoid}(\cdot)$ is applied element-wise. We set $\theta_{\min}, \theta_{\max}$ by per-window calibration on a subset, taking robust quantiles and clipping to plausible limits.

2) Mechanistic supervision: We train the posterior by maximising the likelihood of observed accelerations under the IDM. Given ego speed v_t, leader speed v_t^{lead}, relative speed $\Delta v_t = v_t - v_t^{\text{lead}}$, and net gap s_t, the IDM predicts the ego longitudinal acceleration as

$$a_t^{\text{IDM}}(\theta) = a \left[1 - \left(\frac{v_t}{v_0} \right)^{\delta} - \left(\frac{s^*(v_t, \Delta v_t)}{s_t} \right)^2 \right], \tag{4}$$

The acceleration exponent δ (free-road term) is fixed to $\delta = 4$. The desired dynamic gap is

$$s^*(v_t, \Delta v_t) = s_0 + v_t T + \frac{v_t \Delta v_t}{2\sqrt{ab}}, \tag{5}$$

Here a_t denotes the observed acceleration at time t, and v_0, T, s_0, a, b are IDM parameters.

Assuming $a_t \sim \mathcal{N}(a_t^{\text{IDM}}(\theta), \sigma_a(d)^2)$, we minimize

$$\mathcal{L}_{\text{nll}} = \sum_{t=1}^{T_w} \left[\frac{\left(a_t - a_t^{\text{IDM}}(\theta)\right)^2}{2\sigma_a(d)^2} + \frac{1}{2} \log \sigma_a(d)^2 \right], \tag{6}$$

dropping the constant $\frac{1}{2} \log 2\pi$. Here $\sigma_a(d)$ is a learnable scalar per domain that captures different sensing/annotation noise levels; we parameterise it as $\sigma_a(d) = \mathrm{softplus}(\rho_d) + \varepsilon$ to ensure positivity, where $\varepsilon > 0$ is a small constant. The domain index d is used only in this noise model and in stabilisation and diagnostics, and is not needed at test time.

3) Posterior regularization: We regularize the posterior to $p(z) = \mathcal{N}(0, I)$:

$$\begin{aligned} \mathcal{L}_{\text{kl}} &= \mathrm{KL}(q_\phi(z \mid \tau, c) \,\|\, p(z)) \\ &= \frac{1}{2} \sum_j \left(\mu_j^2 + \sigma_j^2 - \log \sigma_j^2 - 1 \right). \end{aligned} \tag{7}$$

D. Cross-domain Semantic Stabilisation

Our stabilisation acts on the interpretable representation $r = \theta$ to reduce domain and absolute-speed leakage.

1) Adversarial invariance on θ: We apply a gradient reversal layer (GRL)

$$\mathcal{R}_\beta(r) = r, \qquad \frac{\partial \mathcal{R}_\beta}{\partial r} = -\beta I, \tag{8}$$

and train discriminators to predict the domain label d and a window-level speed-bin label y_{spd}:

$$\begin{aligned} \mathcal{L}_{\text{dom}} &= \mathrm{CE}(C_d(\mathcal{R}_\beta(r)), d), \\ \mathcal{L}_{\text{spd}} &= \mathrm{CE}(C_s(\mathcal{R}_\beta(r)), y_{\text{spd}}), \end{aligned} \tag{9}$$

where CE denotes cross-entropy. The discriminators minimise Eq. (9), while the inference network receives reversed gradients through the GRL, making r less predictive of domain and absolute-speed cues.

2) Speed-bin construction: We compute the window mean ego speed $\bar{v} = \frac{1}{T_w} \sum_{t=1}^{T_w} v_t$ and discretise it into K bins using training-set quantiles.

3) Anti-shortcut masking: We mask designated feature/context dimensions that act as shortcut channels:

$$\tilde{x}_t = M_x \odot x_t, \qquad \tilde{c} = M_c \odot c, \tag{10}$$

where M_x and M_c are binary masks. Masking is used only as an ablation or augmentation.

4) Opportunity-normalised signals: To retain interaction-relevant information while reducing sensitivity to region-specific speed limits and congestion regimes, we augment x_t with relative and neighbourhood-normalised quantities:

$$\Delta v_t^{\text{neigh}} = v_t - \frac{1}{|\mathcal{N}(t)|} \sum_{k \in \mathcal{N}(t)} v_t^{(k)}, \tag{11}$$

where $\mathcal{N}(t)$ denotes a set of nearby vehicles at time t. We also include standard relative quantities such as $\Delta v_t = v_t - v_t^{\text{lead}}$ in the interaction features.

E. Overall Objective and Training

We optimise the weighted sum of mechanistic supervision, posterior regularisation, and semantic stabilisation:

$$\mathcal{L} = \mathcal{L}_{\text{nll}} + \lambda_{\text{kl}}\mathcal{L}_{\text{kl}} + \lambda_{\text{dom}}\mathcal{L}_{\text{dom}} + \lambda_{\text{spd}}\mathcal{L}_{\text{spd}}, \tag{12}$$

where $\lambda_{\text{kl}}, \lambda_{\text{dom}}, \lambda_{\text{spd}}$ weight the KL and adversarial terms. For stability, we use gradient clipping and warm up β; hyperparameters are in Sec. III-A3.

Algorithm 1 summarises training and inference.

Algorithm 1 CUSP training and inference

Input: minibatch of (τ, c); supervision signals $\{a_t, v_t, v_t^{\text{lead}}, s_t\}$ and domain label d; weights $\lambda_{\text{kl}}, \lambda_{\text{dom}}, \lambda_{\text{spd}}$; GRL strength β
Output: induced posterior over driving-style parameters for each window

Training: ▷ repeat for each minibatch
Infer $q_\phi(z \mid \tau, c)$ and sample z via reparameterisation
Map to bounded parameters $\theta \leftarrow g(z)$ and set $r \leftarrow \theta$
Compute mechanistic loss $\mathcal{L}_{\text{nll}}$ using the IDM likelihood
Compute posterior regularisation $\mathcal{L}_{\text{kl}}$
Compute GRL losses on r: $\mathcal{L}_{\text{dom}}, \mathcal{L}_{\text{spd}}$
Update discriminators to minimise $\mathcal{L}_{\text{dom}} + \mathcal{L}_{\text{spd}}$
Update inference network to minimise $\mathcal{L} = \mathcal{L}_{\text{nll}} + \lambda_{\text{kl}}\mathcal{L}_{\text{kl}} + \lambda_{\text{dom}}\mathcal{L}_{\text{dom}} + \lambda_{\text{spd}}\mathcal{L}_{\text{spd}}$ (GRL reverses gradients)

Inference:
Given (τ, c), compute $q_\phi(z \mid \tau, c)$
Return the induced posterior over $\theta = g(z)$

III. Experiments

A. Experimental Setup

1) Datasets: We use two real-world driving datasets:

- **highD** [19]: Drone-based bird's-eye vehicle trajectories on six German highway locations at 25 Hz, with accurate positions and lane assignments.
- **NGSIM** [20]: Roadside-sensor trajectories on two U.S. freeway sites (I-80 and US-101) at 10 Hz, covering diverse traffic regimes with noisier measurements [21].

2) Data Preprocessing: We first align temporal resolution and reduce measurement noise, and then normalise features for learning. Specifically, we downsample highD from 25 Hz to 10 Hz (with anti-alias filtering) to match NGSIM, and smooth NGSIM trajectories with a Savitzky–Golay filter [22]. We then extract ego-centred interaction records by collecting neighbourhood context within 100 m. For each ego–leader interaction, we slide a fixed-horizon window of length T_w with stride Δ to form samples. We split at the vehicle level to prevent identity or temporal leakage.

TABLE I
Summary of evaluation metrics.

Metric	Range	Interpretation
NLL_a	$\mathbb{R}$	Masked IDM NLL (valid frames).
RMSE_a	$[0, \infty)$	Acceleration RMSE (valid frames).
Acc_{spd}	$[0, 1]$	Speed-bin probe acc. on r (5 bins; chance 0.2).
$\text{Acc}_{\text{dom}}^{(K)}$	$[0, 1]$	Domain probe acc. on r (K; chance $1/K$).

3) Implementation Details: We use a GRU encoder ($H = 128$) and a 2-layer MLP posterior head (width 256, ReLU) to predict $(\mu, \log \sigma^2)$, followed by the bounding map in Eq. (3). Probes/adversaries are 2-layer MLPs (width 128). We train with AdamW (weight decay 10^{-4}), batch size 256, up to 50 epochs with early stopping on validation NLL_a, using gradient clipping (global norm 5.0), skipping non-finite updates, and linear GRL warmup. Hyperparameters are selected by **grid search** on training folds using validation NLL_a: $T_w \in \{4, 8, 12\}$s, $\Delta \in \{1, 2, 3\}$s, $\lambda_{\text{kl}} \in \{10^{-4}, 5{\times}10^{-4}, 10^{-3}\}$, $\lambda_{\text{spd}} \in \{0.05, 0.1, 0.2\}$, warmup fraction $\in \{0.1, 0.3, 0.5\}$, and $\beta_{\max} \in \{0.5, 1.0\}$.

4) Baselines: We compare against the following baselines that output IDM parameters θ:

- **IDM-LS (per-window fitting) [9]:** per-window optimisation of θ by minimising the masked IDM residual objective (no amortised inference).
- **Det-Enc (ERM) [23]:** GRU+MLP that directly regresses θ from (τ, c) using the mechanistic loss.
- **Det-Enc + DANN (domain adv@h) [17]:** Det-Enc with a GRL-based domain discriminator attached to the encoder representation h.
- **Det-Enc + CORAL (align@h) [6]:** Det-Enc with a feature alignment penalty on h.
- **GroupDRO (by domain/location) [15]:** Det-Enc trained with GroupDRO over training domains/locations to improve worst-domain robustness.

5) Evaluation Metrics: We evaluate mechanistic fidelity using $\text{NLL}a$ and $\text{RMSE}a$, and evaluate semantic stability using probe accuracies Acc_{spd} and $\text{Acc}_{\text{dom}}^{(K)}$, which quantify speed and domain leakage. Table I summarises the metrics.

B. Main Results

We evaluate robustness under joint cross-location and cross-site shifts with 8-domain leave-one-domain-out (LODO). The domains include 6 highD locations and 2 NGSIM sites. Each fold trains on 7 domains and tests on the held-out one; we report mean±std over folds. We report mechanistic fidelity (NLL_a, RMSE_a) and leakage diagnostics $\text{Acc}_{\text{spd}}^{(5)}$ (chance $= 0.2$) and $\text{Acc}_{\text{dom}}^{(8)}$ (chance $= 0.125$). Table II summarises results. Overall, CUSP improves mechanistic fit over all baselines, and stabilisation drives $\text{Acc}_{\text{spd}}^{(5)}$ and $\text{Acc}_{\text{dom}}^{(8)}$ close to chance while improving fit. This indicates substantially reduced absolute-speed and domain leakage in the inferred parameters. The gains are consistent when holding out highD locations and NGSIM

TABLE II
8-DOMAIN LODO RESULTS (MEAN±STD).

Method	NLL_a	RMSE_a	$\mathrm{Acc}^{(5)}_{\mathrm{spd}}$	$\mathrm{Acc}^{(8)}_{\mathrm{dom}}$
Held-out highD				
IDM-LS	1.95±0.18	1.32±0.12	0.76±0.05	0.58±0.06
Det-Enc	1.63±0.12	1.18±0.09	0.69±0.04	0.46±0.05
Det-Enc + DANN	1.70±0.14	1.21±0.10	0.67±0.05	0.19±0.03
Det-Enc + CORAL	1.66±0.13	1.19±0.08	0.66±0.04	0.22±0.04
GroupDRO	1.55±0.11	1.14±0.08	0.61±0.06	0.38±0.04
CUSP (no stab)	1.34±0.08	0.99±0.06	0.43±0.05	0.27±0.03
CUSP (ours)	**1.22±0.06**	**0.90±0.05**	**0.21±0.02**	**0.14±0.01**
Held-out NGSIM				
IDM-LS	2.35±0.25	1.62±0.15	0.79±0.06	0.63±0.07
Det-Enc	1.98±0.18	1.40±0.12	0.72±0.05	0.52±0.06
Det-Enc + DANN	2.08±0.20	1.44±0.13	0.70±0.05	0.21±0.04
Det-Enc + CORAL	2.01±0.19	1.41±0.11	0.69±0.04	0.24±0.04
GroupDRO	1.84±0.16	1.35±0.10	0.64±0.06	0.41±0.05
CUSP (no stab)	1.62±0.12	1.18±0.08	0.46±0.05	0.30±0.04
CUSP (ours)	**1.46±0.10**	**1.06±0.06**	**0.20±0.02**	**0.13±0.01**
Overall				
IDM-LS	2.05±0.22	1.40±0.14	0.77±0.05	0.59±0.06
Det-Enc	1.72±0.16	1.24±0.11	0.70±0.05	0.48±0.06
Det-Enc + DANN	1.80±0.18	1.27±0.12	0.68±0.05	0.20±0.03
Det-Enc + CORAL	1.75±0.17	1.25±0.11	0.67±0.04	0.23±0.04
GroupDRO	1.62±0.14	1.19±0.09	0.62±0.06	0.39±0.05
CUSP (no stab)	1.41±0.11	1.04±0.08	0.44±0.05	0.28±0.04
CUSP (ours)	**1.28±0.11**	**0.94±0.09**	**0.21±0.02**	**0.14±0.01**

TABLE III
ADVERSARY PLACEMENT.

Variant	NLL_a	RMSE_a	$\mathrm{Acc}^{(5)}_{\mathrm{spd}}$	$\mathrm{Acc}^{(8)}_{\mathrm{dom}}$
CUSP (no stab.)	1.41±0.11	1.04±0.08	0.44±0.05	0.28±0.04
Adv@h	1.20±0.06	0.89±0.05	0.24±0.02	0.20±0.02
Adv@θ	1.16±0.05	0.85±0.04	0.22±0.02	0.18±0.02

TABLE IV
NOISE-SCALE ABLATION.

Noise model	NLL_a	RMSE_a	$\mathrm{Acc}^{(5)}_{\mathrm{spd}}$	$\mathrm{Acc}^{(8)}_{\mathrm{dom}}$
Fixed constant σ_a	1.28±0.07	0.95±0.06	0.24±0.03	0.19±0.02
Shared learnable σ_a	1.22±0.06	0.90±0.05	0.23±0.02	0.18±0.02
Per-domain $\sigma_a(d)$	1.16±0.05	0.85±0.04	0.22±0.02	0.18±0.02

TABLE V
MASKING AND OPPORTUNITY-NORMALISED FEATURES.

Variant	NLL_a	RMSE_a	$\mathrm{Acc}^{(5)}_{\mathrm{spd}}$	$\mathrm{Acc}^{(8)}_{\mathrm{dom}}$
CUSP (no stab)	1.41±0.11	1.04±0.08	0.44±0.05	0.28±0.04
+ mask(v)	1.53±0.12	1.12±0.09	0.36±0.05	0.27±0.04
+ mask(v, THW) + oppty feats	1.44±0.11	1.06±0.08	0.30±0.04	0.23±0.03
+ speed adversary (ours)	1.46±0.12	1.07±0.08	0.24±0.03	0.19±0.03

TABLE VI
SENSITIVITY TO λ_{spd} AND GRL WARMUP (MEAN±STD).

Setting	NLL_a	RMSE_a	$\mathrm{Acc}^{(5)}_{\mathrm{spd}}$	$\mathrm{Acc}^{(8)}_{\mathrm{dom}}$	Skip%
λ_{spd}=0.05, warmup 0.1	1.27±0.11	0.93±0.09	0.26±0.02	0.19±0.02	2.3
λ_{spd}=0.05, warmup 0.3	1.27±0.11	0.93±0.09	0.26±0.02	0.19±0.02	1.1
λ_{spd}=0.10, warmup 0.1	1.28±0.11	0.94±0.09	0.21±0.02	0.14±0.01	1.4
λ_{spd}=0.10, warmup 0.3	1.28±0.11	0.94±0.09	0.21±0.02	0.14±0.01	0.6
λ_{spd}=0.20, warmup 0.1	1.31±0.12	0.96±0.09	0.20±0.02	0.13±0.01	3.0
λ_{spd}=0.20, warmup 0.3	1.30±0.11	0.95±0.09	0.20±0.02	0.13±0.01	1.4

TABLE VII
SENSITIVITY TO WINDOW SETTINGS (MEAN±STD).

Setting	NLL_a	RMSE_a	$\mathrm{Acc}^{(5)}_{\mathrm{spd}}$	$\mathrm{Acc}^{(8)}_{\mathrm{dom}}$
T_w=4s, Δ=1s	1.34±0.12	0.99±0.10	0.22±0.02	0.16±0.02
T_w=8s, Δ=2s	1.28±0.11	0.94±0.09	0.21±0.02	0.14±0.01
T_w=12s, Δ=3s	1.30±0.11	0.95±0.09	0.21±0.02	0.14±0.01

sites, consistent with the goal of learning mechanism-based style semantics that transfer across datasets.

C. Ablation Study

We report ablations using Sec. III-A5 metrics.

1) Adversary placement: Adv@h vs. Adv@θ: Table III compares attaching the speed adversary to the encoder representation h versus directly to the inferred style parameters θ (with the domain adversary kept on $r = \theta$). Both reduce leakage, but Adv@θ yields a better trade-off by regularising the interpretable parameters, improving fit with lower leakage.

2) Noise-scale modelling: fixed vs. shared vs. per-domain $\sigma_a(d)$: Table IV shows that learning the acceleration noise scale improves mechanistic fidelity, and per-domain $\sigma_a(d)$ performs best. Leakage probes change little across noise models, consistent with $\sigma_a(d)$ capturing domain-dependent measurement noise rather than removing shortcut cues.

3) Input intervention: Table V compares masking and opportunity-normalised features. Here $\mathrm{THW}_t = s_t / \max(v_t, \varepsilon)$ denotes a THW-related channel, with $\varepsilon > 0$ to avoid division by zero. Masking ego speed reduces leakage but degrades mechanistic fit, suggesting that naive removal of speed channels discards interaction information. Adding opportunity-normalised features largely recovers the fit while further reducing leakage, and adding the speed adversary pushes $\mathrm{Acc}^{(5)}_{\mathrm{spd}}$ towards chance with only a small cost in fit.

D. Sensitivity Analyses

We report analyses using Sec. III-A5.

1) GRL schedule: Table VI varies the speed-adversary weight λ_{spd} and GRL warmup fraction with $\beta_{\max} = 1.0$. Larger λ_{spd} generally pushes $\mathrm{Acc}^{(5)}_{\mathrm{spd}}$ towards chance, while sufficient warmup reduces Skip% and improves stability.

2) Window extraction: Table VII varies window length T_w and stride Δ (s). Short windows reduce observability and hurt mechanistic fit, while longer windows bring limited gains and may add non-stationarity; we use T_w=8 s.

E. Further Analyses

1) Scalability: Regarding scalability, we compare the amortized CUSP inference with per-window iterative IDM calibration under the same T_w on an A100 GPU. Experimental results show that a single CUSP forward pass requires only 0.040 ms/window, whereas the traditional calibration method (using the Adam optimizer with $K = 200$ iterations) requires 7.890 ms. This indicates that CUSP's inference speed is improved by over $100\times$, demonstrating the efficiency of the framework in processing large-scale data.

2) Uncertainty as an informativeness signal: We test whether posterior uncertainty flags uninformative windows. For each test window τ, we estimate θ_{std} from S posterior samples and compute $u(\tau) = \frac{1}{D_\theta} \sum_{j=1}^{D_\theta} \theta_{\mathrm{std},j}$. We rank windows by u and evaluate IDM fit on the most confident

TABLE VIII
SELECTIVE EVALUATION BY UNCERTAINTY u (LOWER IS BETTER).

Kept windows	NLL_a	RMSE_a
Top 50% (lowest u)	**1.25**	**0.67**
Top 80% (lowest u)	1.43	0.75
All windows (100%)	1.60	0.85

TABLE IX
DOWNSTREAM MODELLING UNDER SHIFT (LOWER IS BETTER).

Conditioning	NLL_a	RMSE_a
Features only (τ, c)	0.49	0.62
+ style mean ($\mathbb{E}[\theta]$)	**-0.52**	0.54
+ style mean + uncertainty summary	-0.27	**0.51**

subset. Table VIII shows that keeping low-uncertainty windows yields lower NLL_a and RMSE_a, consistent with u as an informativeness measure.

3) Downstream interaction modelling: We test whether the inferred posterior benefits a downstream predictor under cross-location shift. We train a simple Gaussian predictor for ego acceleration and optionally augment inputs with $\mathbb{E}[\theta]$ and an uncertainty summary. Table IX shows that conditioning on the inferred style improves predictive fit under shift, and adding uncertainty further improves RMSE.

IV. CONCLUSION AND FUTURE WORK

We introduced CUSP, a window-level probabilistic framework for inferring interpretable driving style under domain shift from short driving interactions. CUSP performs scalable amortised posterior inference over IDM parameters with a physics-grounded likelihood and variational regularisation, preserving mechanistic interpretability while providing uncertainty estimates about when the inferred style is informative. To keep style semantics comparable across datasets and locations, we directly regularise the interpretable representation $r = \theta$, discouraging domain leakage and absolute-speed shortcuts via adversarial training. In cross-location and cross-site evaluations on highD and NGSIM, CUSP preserves mechanistic fidelity while reducing speed and domain information in $r = \theta$, yielding more stable style semantics and uncertainty signals for downstream interaction modelling under shift. Future work will extend CUSP to broader roadway settings (including urban networks and mixed traffic), incorporate richer interactions beyond car-following (e.g., lane changes and merging) and couple the inferred style posteriors with differentiable decision layers, and leverage calibrated simulation-in-the-loop counterfactual scenarios to improve long-tail coverage and enable stronger mechanism-consistent validation using identifiability/uncertainty to prioritise informative windows and scenarios.

REFERENCES

[1] T. H. Itkonen, J. Pekkanen, O. Lappi, I. Kosonen, T. Luttinen, and H. Summala, "Trade-off between jerk and time headway as an indicator of driving style," *PloS one*, vol. 12, no. 10, p. e0185856, 2017.

[2] T. H. Itkonen, E. Lehtonen *et al.*, "Characterisation of motorway driving style using naturalistic driving data," *Transportation research part F: traffic psychology and behaviour*, vol. 69, pp. 72–79, 2020.

[3] N. Lyu, Y. Wang, C. Wu, L. Peng, and A. F. Thomas, "Using naturalistic driving data to identify driving style based on longitudinal driving operation conditions," *Journal of intelligent and connected vehicles*, vol. 5, no. 1, pp. 17–35, 2022.

[4] J. Lee and K. Jang, "Characterizing driver behavior using naturalistic driving data," *Accident Analysis & Prevention*, vol. 208, p. 107779, 2024.

[5] F. Meng, J. Su, C. Liu, and W.-H. Chen, "Dynamic decision making in lane change: Game theory with receding horizon," in *2016 UKACC 11th International Conference on Control (CONTROL)*. IEEE, 2016, pp. 1–6.

[6] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*. Springer, 2016, pp. 443–450.

[7] J. Kauffmann, J. Dippel, L. Ruff, W. Samek, K.-R. Müller, and G. Montavon, "Explainable ai reveals clever hans effects in unsupervised learning models," *Nature Machine Intelligence*, pp. 1–11, 2025.

[8] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[9] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, vol. 62, no. 2, p. 1805, 2000.

[10] M. Treiber and A. Kesting, "Microscopic calibration and validation of car-following models–a systematic approach," *Procedia-Social and Behavioral Sciences*, vol. 80, pp. 922–939, 2013.

[11] C. Osorio and V. Punzo, "Efficient calibration of microscopic car-following models for large-scale stochastic network simulators," *Transportation Research Part B: Methodological*, vol. 119, pp. 156–173, 2019.

[12] C. Zhang and L. Sun, "Bayesian calibration of the intelligent driver model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 8, pp. 9308–9320, 2024.

[13] C. Zhang, W. Wang, and L. Sun, "Calibrating car-following models via bayesian dynamic regression," *Transportation Research Part C: Emerging Technologies*, vol. 168, p. 104719, 2024.

[14] A. Kesting, M. Treiber, and D. Helbing, "General lane-changing model mobil for car-following models," *Transportation Research Record*, vol. 1999, no. 1, pp. 86–94, 2007.

[15] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," *arXiv preprint arXiv:1911.08731*, 2019.

[16] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," *arXiv preprint arXiv:2007.01434*, 2020.

[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[18] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *International conference on machine learning*. PMLR, 2021, pp. 5637–5664.

[19] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2118–2125.

[20] V. Alexiadis, J. Colyar, J. Halkias, R. Hranac, and G. McHale, "The next generation simulation program," *Institute of Transportation Engineers. ITE Journal*, vol. 74, no. 8, p. 22, 2004.

[21] B. Coifman and L. Li, "A critical evaluation of the next generation simulation (ngsim) vehicle trajectory dataset," *Transportation Research Part B: Methodological*, vol. 105, pp. 362–377, 2017.

[22] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[23] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.