Differentiable Semantic Meta-Learning Framework for Long-Tail Motion Forecasting in Autonomous Driving

Bin Rao[1], Chengyue Wang[1], Haicheng Liao[1], Qianfang Wang[2], Yanchen Guan[1], Jiaxun Zhang[1], Xingcheng Liu[1], Meixin Zhu[3], Kanye Ye Wang[4], Zhenning Li[4]*

[1]State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau SAR, China
[2]School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, China
[3]School of Transportation, Southeast University, Nanjing, China
[4]State Key Laboratory of Internet of Things for Smart City and Departments of Civil and Environmental Engineering and Computer and Information Science, University of Macau, Macau SAR, China
zhenningli@um.edu.mo

Abstract

Long-tail motion forecasting is a core challenge for autonomous driving, where rare yet safety-critical events—such as abrupt maneuvers and dense multi-agent interactions—dominate real-world risk. Existing approaches struggle in these scenarios because they rely on either non-interpretable clustering or model-dependent error heuristics, providing neither a differentiable notion of "tailness" nor a mechanism for rapid adaptation. We propose SAML, a Semantic-Aware Meta-Learning framework that introduces the first differentiable definition of tailness for motion forecasting. SAML quantifies motion rarity via semantically meaningful intrinsic (kinematic, geometric, temporal) and interactive (local and global risk) properties, which are fused by a Bayesian Tail Perceiver into a continuous, uncertainty-aware Tail Index. This Tail Index drives a meta-memory adaptation module that couples a dynamic prototype memory with an MAML-based cognitive set mechanism, enabling fast adaptation to rare or evolving patterns. Experiments on nuScenes, NGSIM, and HighD show that SAML achieves state-of-the-art overall accuracy and substantial gains on top 1–5% worst-case events, while maintaining high efficiency. Our findings highlight semantic meta-learning as a pathway toward robust and safety-critical motion forecasting.

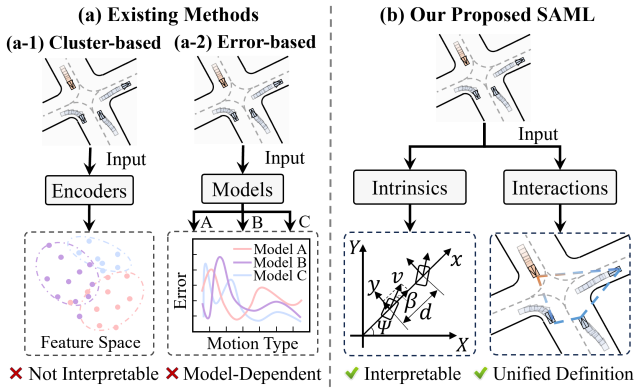

Figure 1: Conceptual comparison of (a) existing methods and (b) the proposed SAML framework. Existing methods detect long-tail events indirectly via non-interpretable clustering (a-1) or model-dependent error signals (a-2). In contrast, SAML (b) offers a principled, interpretable framework that quantifies a motion's tailness from its intrinsic (dynamics, geometry, temporality) and interactive (local and global risk) properties, enabling robust long-tail forecasting.

Introduction

Motion forecasting is a foundational component of autonomous driving pipelines, yet current systems remain fragile in low-frequency, high-risk **long-tail** scenarios (Li et al. 2023; Wang et al. 2023b). These rare events—characterized by abrupt kinematic changes, complex multi-agent interactions, and atypical scene structures—are precisely the ones that most threaten safety and public trust. Despite steady progress on common cases, state-of-the-art predictors still degrade disproportionately on the tail, where data scarcity, semantic heterogeneity, and distributional shift collide (Salzmann et al. 2020).

A key reason for this brittleness is foundational: the community lacks a **principled**, **differentiable**, and **semantically** grounded definition of the "long tail". Existing practices typically (i) cluster motions with non-interpretable heuristics (Makansi et al. 2021; Wang et al. 2023b) or (ii) backsolve "hard cases" from model-specific forecasting errors (Zhang, Pourkeshavarz, and Rasouli 2024). Both strategies are problematic. Clustering-based labels are hard to interpret and highly sensitive to hyperparameter choices, while error-based labels inherit the biases of a specific model, offering neither a generalizable notion of tailness nor a label that can guide end-to-end training (Figure 1). Crucially, both yield discrete, non-differentiable outputs, hindering gradient-based learning targeted at the tail.

To address these issues, we propose **SAML**, a **S**emantic-**A**ware **M**eta-**L**earning framework that redefines the long tail in a differentiable manner and learns to adapt to it. SAML rests on two pillars:

1) **Differentiable semantic definition of tailness**. We

*Corresponding author: zhenningli@um.edu.mo

quantify a motion's "tailness" through a principled set of intrinsic (kinematic dynamism, geometric complexity, temporal irregularity) and interactive (local interaction risk, global scene risk) metrics—each fully differentiable. A Bayesian Tail Perceiver aggregates these metrics into a Tail Index that is both uncertainty-aware and continuously optimizable.

2) **Tail-index–guided meta-adaptation**. The Tail Index steers a Meta-Memory Adaptation module that couples a dynamic prototype memory bank with an MAML-driven cognitive set mechanism. This design enables rapid, few-shot adaptation to emerging or sparsely observed long-tail patterns while mitigating bias toward majority behaviors.

The main contributions of this paper are threefold:

- We operationalize tailness via semantically meaningful intrinsic and interactive metrics, aggregated by a Bayesian perceiver into a continuous Tail Index that unlocks end-to-end optimization for rare events.
- We introduce a Tail-Index–guided Meta-Memory Adaptation module that integrates a dynamic memory bank with a cognitive set mechanism inside a MAML framework, enabling fast adaptation to novel tail patterns.
- Extensive experiments on nuScenes, NGSIM, and HighD show that SAML achieves SOTA overall accuracy, with substantial gains in long-term horizons and worst-case (top 1–5%) subsets, alongside competitive inference efficiency. Comprehensive ablations verify the necessity of each component, with the cognitive set mechanism yielding the largest impact on worst-case robustness.

Related Work

Motion forecasting, a core component of autonomous driving, is formulated as a time-series prediction problem. Early methods used Recurrent Neural Networks (RNNs) and variants like Gated Recurrent Units (GRUs) for temporal dependencies (Huang et al. 2021; Liao et al. 2024b). Social LSTM introduced "social forces" to model agent interactions (Alahi et al. 2016; Liao et al. 2024d), inspiring graph-based approaches where agents are nodes and interactions are edges (Mohamed et al. 2020; Xu et al. 2022; Liao et al. 2024c; Wang et al. 2025b). Attention-based models later dominated, capturing non-local dependencies across agents and maps (Salzmann et al. 2020; Liao et al. 2025c, 2024a). However, benchmark gains have plateaued, as routine success fails to ensure robustness in rare, safety-critical events, motivating long-tail research.

The long-tail phenomenon, studied in machine learning (Liu et al. 2022), is acute in autonomous driving due to imbalanced naturalistic data (Zhou et al. 2022; Zhang, Pourkeshavarz, and Rasouli 2024). Routine behaviors dominate, while rare events like evasive maneuvers or intersections conflicts are underrepresented, biasing models toward common patterns and risking safety in long-tail cases (Li et al. 2023). Tackling this is essential for deployment.

Existing methods for long-tail forecasting can be grouped into data-centric and model-centric strategies:

1) Data-centric strategies rebalance datasets via oversampling rare events or undersampling frequent ones, risking overfitting or loss (Han, Wang, and Mao 2005; Wang et al. 2025a). Advanced techniques generate synthetic samples with VAEs (Makansi et al. 2021; Wang et al. 2023b), GANs (Yang et al. 2024; Liao et al. 2025a), or diffusion models (Bae, Park, and Jeon 2024), but they may add artifacts, require validation, and lack real-world guarantees.

2) Model-centric approaches enhance recognition without data changes, using loss re-weighting for rare samples (Ross and Dollár 2017), contrastive learning to distinguish behaviors such as Hi-SCL (Lan et al. 2024) and AMD (Rao et al. 2025), or meta-learning for few-shot adaptation (Li et al. 2024), though limited in multi-agent long-tail contexts.

While both data-centric and model-centric approaches provide partial solutions, there is still no framework that defines and adapts to long-tail events in a differentiable, semantically meaningful way—the motivation behind our proposed SAML framework.

Methodology

Problem Formulation

Given the motion histories of all agents in a scene $X = \{p, v, h\}$ and the static map context M, our objective is to forecast the future motion of a target agent $Y = \{p\}$. The history X comprises sequences of kinematic state vectors, each encoding an agent's position p, velocity v, and heading angle h. To capture the multi-modal nature of future behavior, we formulate the task as learning a model to estimate the conditional probability distribution $P(Y|X, M)$. In practice, this distribution is approximated by a set of K motion modes with associated confidence scores.

Overall Framework and Motivation

SAML, illustrated in Figure 2, jointly identifies and adapts to long-tail motions. An Interaction-Aware Encoder extracts scene features from motion histories and map context, while a Bayesian Tail Perceiver fuses intrinsic and interactive metrics into a continuous, uncertainty-aware Tail Index. Guided by this Tail Index, a meta-memory module with dynamic prototypes and a cognitive set mechanism emphasizes rare but safety-critical motions and mitigates bias toward common patterns. The adapted features are then decoded into multimodal motions, enabling SAML to handle long-tail cases efficiently even with limited data.

Differentiable Definition of the Long Tail

Existing methods for long-tail motion identification heavily rely on clustering techniques or model-specific forecasting errors. These approaches make it difficult to explain why a motion is long-tail, and error-dependent methods hinder end-to-end differentiable optimization. To address these issues, we define a motion's "tailness" by deconstructing long-tail events into semantically meaningful, fully differentiable metrics based on intrinsic and interactive properties. Specifically, we define three intrinsic properties—Kinematic Dynamism, Geometric Complexity, and Temporal Irregularity—and two interactive properties—Local Interaction Risk and Global Scene Risk. The conceptual definitions are as follows, with more details in the **Appendix A**.

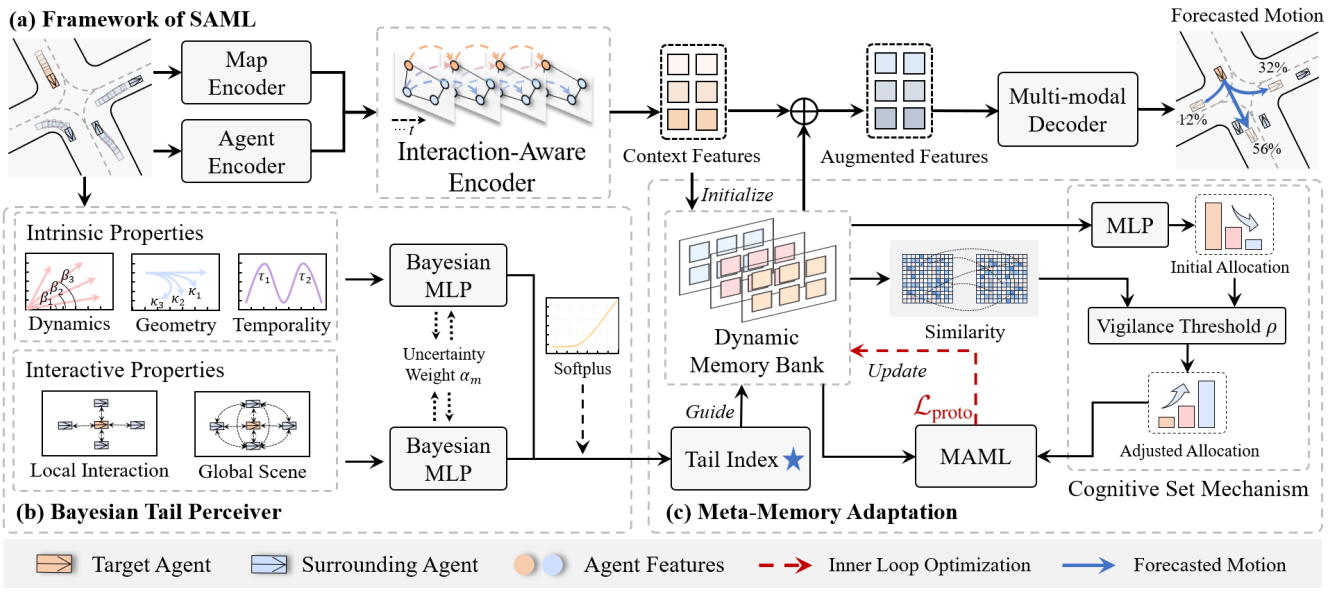

Figure 2: Overview of the proposed SAML framework. (a) The overall model architecture. The model processes motion histories of a target agent and surrounding agents, along with HD map data, using four key modules: the Interaction-Aware Encoder, Bayesian Tail Perceiver, Meta-Memory Adaptation, and Multi-modal Decoder to generate a multimodal motion forecast. (b) and (c) Detailed illustration of the Bayesian Tail Perceiver and the Meta-Memory Adaptation modules, respectively.

Intrinsic Properties Intrinsic properties quantify the inherent complexity of an agent's individual motion, analyzed along kinematic, geometric, and temporal dimensions.

Definition 1 (Kinematic Dynamism)**.** *Kinematic dynamism assesses the magnitude of changes in motion states, reflecting external forces or abrupt transitions such as emergency braking or sharp turns. Key metrics include velocity volatility C_v and rotational instability C_α:*

$$C_v = \sqrt{\mathbb{E}_t\left[\left\|\frac{v(t) - v(t-\Delta t)}{\Delta t}\right\|_2^2\right]} \tag{1}$$

$$C_\alpha = \sqrt{\mathbb{E}_t\left[\left(\frac{\omega(t) - \omega(t-\Delta t)}{\Delta t}\right)^2\right]} \tag{2}$$

where $v(t)$ is velocity, $\omega(t)$ is angular velocity.

Other metrics include acceleration instability C_j, heading volatility C_ω, and movement direction volatility C_{vd}, with more details in the **Appendix**.

Definition 2 (Geometric Complexity)**.** *Geometric complexity measures the spatial curvature of the motion path, indicating turns or evasive maneuvers. A primary metric is curvature intensity C_κ:*

$$\kappa(t) = \frac{|v_x(t)a_y(t) - v_y(t)a_x(t)|}{[v_x(t)^2 + v_y(t)^2]^{3/2}}, \quad C_\kappa = \mathbb{E}_t[|\kappa(t)|] \tag{3}$$

where $\kappa(t)$ is instantaneous curvature and $v_x(t)$, $v_y(t)$, $a_x(t)$, $a_y(t)$ are velocity and acceleration components.

Another metric is curvature volatility $C_{\Delta\kappa}$, with more details in the **Appendix A**.

Definition 3 (Temporal Irregularity)**.** *Temporal irregularity evaluates the predictability of velocity patterns through time-series analysis. It captures aperiodic fluctuations such as stop-and-go traffic by measuring changes in the velocity autocovariance function. The autocovariance fluctuation $C_{\Delta\gamma}$ is defined as:*

$$\gamma(\tau) = \mathbb{E}_t\left[(v(t) - \bar{v}) \cdot (v(t+\tau) - \bar{v})\right] \tag{4}$$

$$C_{\Delta\gamma} = \frac{1}{T_h - 1}\sum_{\tau=1}^{T_h-1} |\gamma(\tau) - \gamma(\tau - 1)| \tag{5}$$

where $\gamma(\tau)$ is the autocovariance function at time lag τ, and $\bar{v}$ is the mean velocity over the observation window.

Interactive Properties Interactive properties quantify the potential risk and abnormality of a motion within a multi-agent context, evaluated at local and global levels.

Definition 4 (Local Interaction Risk)**.** *Local interaction risk assesses immediate threats from neighboring agents, indicating potential collisions in close proximity such as near-miss conflicts or aggressive tailgating. A Key metric is inverse time-to-collision (TTC) R_{ittc}, defined as:*

$$R_{ittc} = \mathbb{E}_t\left[\max_{j\in\mathcal{A}_i}\left(\frac{[-(v_j(t) - v_i(t)) \cdot (p_j(t) - p_i(t))]_+}{\|p_j(t) - p_i(t)\|_2^2}\right)\right] \tag{6}$$

where $[x]_+ = \max(0, x)$, $v_i(t), p_i(t)$ are velocity and position of agent i, and $\mathcal{A}_i$ is the set of neighboring agents to the target agent i.

Details of this definition, including longitudinal and lateral risks extending the Responsibility-Sensitive Safety (RSS) framework, are provided in the **Appendix A**.

Definition 5 (Global Scene Risk)**.** *Global scene risk evaluates overall threat from collective agent dynamics, capturing environmental complexity and density such as in dense traffic or chaotic intersections. A key metric is multi-agent conflict R_{mac}:*

$$R_{mac} = \mathbb{E}_t \left[\frac{2}{N(N-1)} \sum_{1 \le i < j \le N} ITTC_{ij}(t) \right] \tag{7}$$

where N is the number of agents, $ITTC_{ij}(t)$ is inverse TTC between agents i and j.

Other metrics include agent density Rad and neighborhood instability R_{ni}, with more details in the **Appendix A**.

Bayesian Tail Perceiver

The Bayesian Tail Perceiver aggregates semantic features into a continuous, differentiable Tail Index while modeling uncertainty to tackle sparsity and variability in long-tail scenarios. It employs a dual-path design to separately encode intrinsic and interactive properties into f_i and f_r, minimizing feature interference and capturing rare kinematic patterns alongside multi-agent risks. The Bayesian formulation yields a smoothed Tail Index, emphasizes rare cases via elevated uncertainty, and ensures reliable gradients for end-to-end training with limited data. Each path is handled by a dedicated Bayesian MLP to produce latent representations, which are fused to compute the Tail Index.

$$z_i = W_i^{(2)} \sigma(W_i^{(1)} F_i + b_i^{(1)}) + b_i^{(2)} \tag{8}$$

$$z_r = W_r^{(2)} \sigma(W_r^{(1)} F_r + b_r^{(1)}) + b_r^{(2)} \tag{9}$$

where σ is the ReLU activation function. The parameters for each path and layer, $\theta_m^{(l)} = \{W_m^{(l)}, b_m^{(l)}\}$ for $m \in \{i, r\}$ and $l \in \{1, 2\}$, are sampled from approximate posterior distributions $q(\theta_m^{(l)})$, parameterized as diagonal Gaussians.

To fuse the paths, we introduce an uncertainty-guided weighting based on the KL-divergence between posterior $q(\theta_m)$ and prior $p(\theta_m)$ distributions. The fusion weights α_m are computed as:

$$\alpha_m = \frac{\exp(\lambda \cdot \text{KL}(q(\theta_m) || p(\theta_m)))}{\sum_{n \in \{i,r\}} \exp(\lambda \cdot \text{KL}(q(\theta_n) || p(\theta_n)))} \tag{10}$$

where λ is a temperature parameter. The final Tail Index TI is obtained by linearly combining the latent representations and applying a Softplus activation to ensure non-negativity:

$$TI = \sigma_{\text{sp}} \left(w_o^\top (\alpha_i z_i + \alpha_r z_r) + b_o \right) \tag{11}$$

where $\sigma_{\text{sp}}(x) = \log(1 + e^x)$ is the Softplus activation function, and w_o, b_o are the learnable weight and bias of the output layer, respectively.

Interaction-Aware Encoder

To generate rich, context-aware representations for motion histories, we design the Interaction-Aware Encoder. The process begins with independent encoding of scene elements: the target agent f_t, surrounding agents f_n, and map lanes f_l are processed using separate GRUs, supplemented by a temporal Transformer for the target to capture long-range dependencies, yielding initial features F_t, F_n, and F_l.

Subsequently, agent interactions are modeled by constructing a graph G_a from $\{F_t, F_n\}$ and applying self-attention $\mathcal{A}_{\text{self}}$ to produce intermediate interaction-aware features G'_a, which are aggregated into F'_t:

$$G'_a = \mathcal{A}_{\text{self}}(G_a + P_a), \quad F'_t = \mathbb{E}[G'_a] \tag{12}$$

where P_a is learnable positional encoding, and $\mathbb{E}$ denotes average pooling over G'_a. Multi-modal features are then decoded using learnable queries Q through chained cross-attention $\mathcal{A}_{\text{cross}}$, first attending to the agent context G'_a and then to the map context F_l:

$$F_m = \mathcal{A}_{\text{cross}}(\mathcal{A}_{\text{cross}}(Q + F'_t, G'_a + P'_a, G'_a), F_l + P_l, F_l) \tag{13}$$

where P'_a and P_l are learnable positional encodings. The resulting F_m provides context-aware, multi-modal features for subsequent meta-learning. Detailed steps and derivations are provided in **Appendix B**.

Meta-Memory Adaptation

The Meta-Memory Adaptation module enables SAML to handle sparse and evolving long-tail motion patterns. Guided by the Tail Index, it combines a dynamic prototype memory with an adaptive cognitive set mechanism to reduce bias toward frequent behaviors, and leverages a MAML framework to achieve rapid few-shot adaptation to rare or emerging motions.

Cognitive Set Mechanism To leverage the dynamic memory bank M, which stores C class-prototypes representing distinct motion categories as detailed in **Appendix C**, we design a mechanism to assess the relevance of each prototype to the current motion history. Relying solely on a data-driven gating network may induce a cognitive set, favoring frequent patterns and overlooking novel or long-tail events. To mitigate this bias, we propose a cognitive set mechanism that acts as a vigilance system. It computes a base category allocation distribution g using a MLP on a concatenated input $h = [F_m, F_i, F_r, TI]$, and calculates the normalized similarity s between feature F_m and prototype bank M:

$$g = \phi_S(\phi_M(h)) \tag{14}$$

$$s = \frac{F_m \cdot M^\top}{\|F_m\|_2 \|M\|_2} \cdot \tau, \quad \tau > 0 \tag{15}$$

where ϕ_M denotes the MLP, ϕ_S denotes the softmax function, and τ is a temperature parameter.

We then introduces a learnable vigilance threshold $\rho \in \mathbb{R}$ which dynamically modulates the initial allocation based on the maximum similarity score. This produces an adjusted allocation g' enabling the model to override its initial judgment when a strong match with a rare prototype is detected:

$$\lambda = \sigma(\gamma(\max\{s\} - \rho)) \tag{16}$$

$$g' = \lambda \cdot g + (1 - \lambda) \cdot b_{\text{tail}} \tag{17}$$

where σ denotes the sigmoid function, γ controls the steepness of the transition, and b_{tail} is a bias vector that amplifies long-tail categories. This process equips the model with adaptive vigilance, mitigating cognitive fixation and ensuring that novel patterns receive appropriate attention.

Model	Venue	minADE_{10}	minADE_5	minFDE_5	minFDE_1	MR_5
Trajectron++ (Salzmann et al. 2020)	ECCV	1.51	1.88	5.63	9.52	0.70
MultiPath (Chai et al. 2020)	CoRL	-	1.44	4.83	7.69	0.76
AgentFormer (Yuan et al. 2021)	ICCV	1.31	1.59	3.14	<u>6.45</u>	-
Autobot (Girgis et al. 2022)	ICLR	<u>1.03</u>	1.37	3.40	8.19	0.62
THOMAS (Gilles et al. 2022b)	ICLR	1.04	1.33	-	6.71	0.55
PGP (Deo, Wolff, and Beijbom 2022)	CoRL	1.03	1.30	2.52	7.17	0.61
GoHome (Gilles et al. 2022a)	ICRA	1.15	1.42	-	6.99	0.57
ContextVAE (Xu, Hayet, and Karamouzas 2023)	RAL	-	1.59	3.28	8.24	-
EMSIN (Ren et al. 2024)	TFS	1.36	1.77	3.56	9.06	0.54
SeFlow (Zhang et al. 2024)	ECCV	1.04	1.38	-	7.89	0.60
AMD (Rao et al. 2025)	ICCV	1.06	1.23	2.43	6.99	<u>0.50</u>
NEST (Wang et al. 2025b)	AAAI	-	<u>1.18</u>	<u>2.39</u>	6.87	0.50
SAML (Ours)	-	**1.01**	**1.18**	**2.34**	**6.33**	**0.48**

Table 1: Comparison of the performance of various models across all samples on nuScenes dataset. **Bold** and <u>underlined</u> text represent the best and second-best results, respectively. Cases marked with ('-') indicate missing values.

	Model	Forecasting Horizon (s)				
		1	2	3	4	5
NGSIM	CF-LSTM (Xie et al. 2021)	0.55	1.10	1.78	2.73	3.82
	iNATran (Chen et al. 2022a)	0.39	0.96	1.61	2.42	3.43
	STDAN (Chen et al. 2022b)	0.39	0.96	1.62	2.51	3.65
	WSiP (Wang et al. 2023a)	0.56	1.23	2.05	3.08	4.34
	HiT (Liao et al. 2025b)	0.38	0.90	1.42	2.08	2.87
	DEMO (Wang et al. 2025c)	<u>0.36</u>	<u>0.86</u>	1.48	2.10	2.88
	CITF (Liao et al. 2025c)	**0.30**	**0.81**	<u>1.42</u>	<u>2.04</u>	<u>2.82</u>
	SAML (Ours)	0.39	0.90	**1.36**	**1.81**	**2.41**
HighD	CF-LSTM (Xie et al. 2021)	0.18	0.42	1.07	1.72	2.44
	STDAN (Chen et al. 2022b)	0.19	0.27	0.48	0.91	1.66
	DRBP (Gao et al. 2023)	0.41	0.79	1.11	1.40	2.58
	WSiP (Wang et al. 2023a)	0.20	0.60	1.21	2.07	3.14
	HiT (Liao et al. 2025b)	0.08	0.13	0.22	0.39	0.61
	DEMO (Wang et al. 2025c)	<u>0.06</u>	0.14	0.25	0.44	0.70
	CITF (Liao et al. 2025c)	**0.04**	**0.09**	<u>0.18</u>	<u>0.30</u>	<u>0.43</u>
	SAML (Ours)	0.08	<u>0.12</u>	**0.15**	**0.21**	**0.33**

Table 2: Comparison of model performance on NGSIM and HighD datasets. Metric: RMSE.

MAML-driven Memory Adaptation To enable rapid adaptation to emerging long-tail distributions, we employ a MAML framework, optimizing memory prototypes for few-shot generalization in data-sparse scenarios. The prototype memory M is refined using a contrastive loss to align features with class prototypes:

$$\mathcal{L}_{\text{proto}} = -\frac{1}{B}\sum_{i=1}^{B}\log\sigma(\sum_{k=1}^{C} g'_{i,k}s_{i,k} - \sum_{k=1}^{C}(1-g'_{i,k})s_{i,k}) \tag{18}$$

where $g'_{i,k}$ denotes the element for category k in the adjusted category allocation g'_i of sample i, and $s_{i,k}$ is the element for category k in the similarity vector s_i of sample i.

In the inner loop, M is updated via gradient descent:

$$M' = M - \alpha\nabla_M\mathcal{L}_{\text{proto}} \tag{19}$$

where α is the inner-loop learning rate. The outer loop optimizes the model parameters for cross-task generalization. This meta-learning approach ensures rapid alignment with long-tail patterns, such as sudden evasive actions, enhancing forecasting robustness. The refined memory M' is then used to generate augmented features:

$$F_v = F_m + \sigma(\phi_M(h)) \cdot (g' \cdot M') \tag{20}$$

where F_v is the augmented feature, σ denotes the sigmoid function, ϕ_M is a MLP, and g' is the adjusted category allocation from the cognitive set mechanism.

Multi-modal Decoder

The Multi-modal decoder transforms the augmented features F_v into future motion forecasts. We employ a GRU and MLP to generate multi-modal future motion representations, mapped to a Laplace distribution to capture uncertainty in long-tail scenarios. The Laplace distribution, with its peaked and heavy-tailed nature, is well-suited for modeling central tendencies and extreme deviations in long-tail motions. This approach produces diverse and robust motion forecasts, enhancing sensitivity to rare driving events. The model is trained end-to-end using a composite loss function, detailed in the **Appendix D**.

Experiments

Experimental Setup

We conduct a comprehensive evaluation of our framework on three diverse, large-scale datasets: the urban-centric nuScenes (Caesar et al. 2020), and the highway-focused NGSIM (Deo and Trivedi 2018) and HighD (Krajewski et al. 2018). To ensure fair comparison, we adhere to the established evaluation protocol for each dataset. For the multi-modal nuScenes benchmark, we report the standard metrics of minADE_K, minFDE_K, and MR_K over K modes. For the NGSIM and HighD datasets, we report the RMSE. Further details on datasets, metric definitions, and implementation are provided in the **Appendix E**.

Model	Top 1%	Top 2%	Top 3%	Top 4%	Top 5%	All
PGP (Deo, Wolff, and Beijbom 2022)	8.86/21.92	7.21/17.90	6.24/15.68	5.52/13.77	5.02/12.44	1.28/2.52
Q-EANet (Chen et al. 2024)	7.55/18.78	6.15/15.58	5.44/13.76	4.94/12.49	4.55/11.49	1.20/2.45
LAformer (Liu et al. 2024)	8.19/19.03	6.73/15.81	5.89/13.90	5.33/12.60	4.90/11.61	1.19/2.42
UniTraj (MTR) (Feng et al. 2024)	7.84/21.69	6.44/18.06	5.69/15.95	5.18/14.49	4.78/13.37	**1.15**/2.61
AMD (Rao et al. 2025)	7.50/18.47	6.37/15.71	5.65/13.99	5.08/12.45	4.62/11.36	1.23/2.39
SAML (Ours)	**6.21/14.72**	**5.36/12.36**	**5.09/11.50**	**4.48/10.07**	**4.21/9.41**	1.18/**2.34**
SAML (Ours) (50%)	7.81/19.33	6.48/16.02	5.75/14.07	5.13/12.57	4.73/11.52	1.30/2.56

Table 3: Worst-case performance comparison on the nuScenes dataset, reported in minADE_5 / minFDE_5. The Top 1-5% subsets are defined for each model individually based on its own worst-performing samples, ranked by minFDE_5.

Quantitative Results

Overall Performance Comparison We evaluate the overall performance of our model against a range of state-of-the-art methods on three diverse datasets. The results on the urban-centric nuScenes benchmark, presented in Table 1, demonstrate that our model achieves top performance across all metrics. Notably, our model achieves a 5.7% error reduction in the minFDE_1 metric compared to the second-best method. Table 2 details the performance on the highway datasets NGSIM and HighD, where our model exhibits exceptional capabilities. It consistently outperforms all other models in long-term forecasting (3s-5s horizons) on both datasets, with a 14.5% improvement over the second-best in 5s RMSE on NGSIM and a 23.3% improvement on HighD. These results affirm the effectiveness and robustness of our proposed model in both urban and highway environments.

Worst-Case Performance Comparison Traditional long-tail evaluations often define hard cases based on the errors produced by a single, fixed baseline model, which can introduce inherent biases. To enable a more equitable comparison, we assess each model's performance in its respective worst-case scenarios. Specifically, for every model, we rank all test samples according to the minADE_5 errors it generates and select the top 1% to 5% of samples where that model exhibits the poorest performance. This protocol evaluates the upper bound of Forecasting error for each method. As illustrated in Table 3, our proposed model exhibits superior performance. On the top 1% of the most challenging samples, SAML achieves a minADE_5 of 6.21 m, representing a 17.2% reduction relative to the second-best baseline. The advantage in minFDE_5 is even more substantial, with a 20.3% reduction in error. These results indicate that SAML's worst-case performance is markedly superior to that of other state-of-the-art methods. Notably, our model variant trained on only 50% of the data, SAML (50%), still surpasses several fully trained baselines, such as LAformer and UniTraj (MTR), thereby demonstrating competitive efficacy even with reduced data volumes. This underscores the targeted and efficient enhancements to long-tail samples afforded by our semantic meta-learning framework.

Efficiency Analysis

To evaluate the efficiency of our proposed model, we conduct a comparative analysis of inference time and accuracy using the nuScenes dataset. All models are tested on a single NVIDIA RTX 3090 GPU. As presented in Table 4, SAML demonstrates a distinct advantage in computational efficiency. The results indicate that our model achieves state-of-the-art accuracy while operating at a significantly faster inference speed of 21 ms compared to other high-performing methods, such as LAformer (Liu et al. 2024), making it a practical and effective solution for real-world deployment.

Model	Time	minADE_5	minFDE_1
Trajectron++ (Salzmann et al. 2020)	38	1.88	9.52
AgentFormer (Yuan et al. 2021)	107	1.59	6.45
PGP (Deo, Wolff, and Beijbom 2022)	215	1.30	7.17
LAformer (Liu et al. 2024)	115	1.19	6.95
VisionTrap (Moon et al. 2024)	53	1.35	8.72
SAML (Ours)	**21**	**1.18**	**6.33**

Table 4: Efficiency comparison on the nuScenes dataset.

Comp.	Ablation Model				
	A	B	C	D	E
BTP	×	✓	✓	✓	✓
IAE	✓	×	✓	✓	✓
CSM	✓	✓	×	✓	✓
MAML	✓	✓	✓	×	✓
Top 1%	7.87/19.17	7.86/19.18	7.89/19.21	7.39/18.15	**6.21/14.72**
Top 2%	6.29/15.51	6.45/15.72	6.47/15.78	6.22/15.30	**5.36/12.36**
Top 3%	5.34/13.43	5.62/13.73	5.84/14.16	5.58/13.75	**5.09/11.50**
All	1.26/2.46	1.36/2.64	1.33/2.56	1.30/2.52	**1.18/2.34**

Table 5: Ablation results (minADE_5/minFDE_5) on nuScenes dataset. BTP: Bayesian Tail Perceiver; IAE: Interaction-Aware Encoder; CSM: Cognitive Set Mechanism; MAML: MAML-driven Memory Adaptation.

Ablation Studies

To validate the effectiveness of each key component in our framework, we conducted ablation studies on the nuScenes dataset, with results in Table 5. The full model (Model E) achieves the best performance, while ablated variants show degradation that highlights each module's contribution. Most notably, removing the Cognitive Set Mechanism

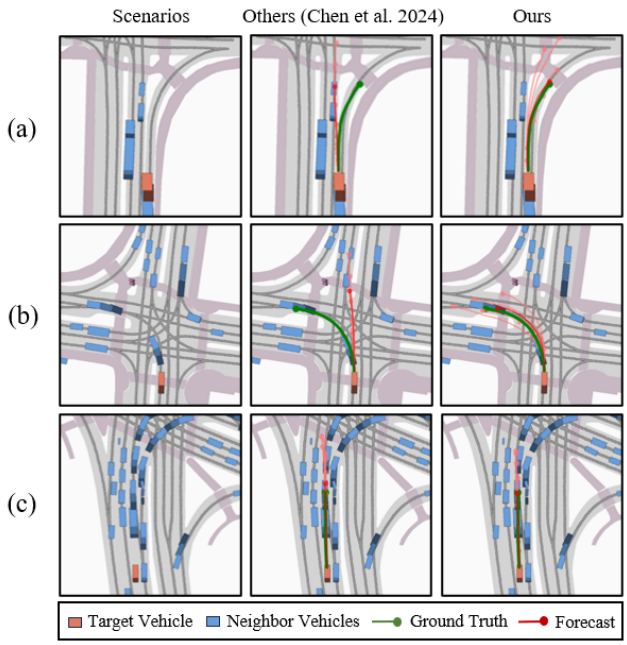

Figure 3: Visualization of long-tail performance on the nuScenes dataset. Red denotes the highest-probability forecast, and pink represents other multimodal options.

(Model C) causes the most significant drop, with minADE_5 and minFDE_5 on the Top 1% set increasing by 27.1% and 30.5% compared to the full model. This underscores CSM's critical role in mitigating bias towards common patterns and enhancing sensitivity to rare, long-tail events. Additionally, removing the Bayesian Tail Perceiver (Model A), replacing the Interaction-Aware Encoder (Model B), or ablating MAML-driven adaptation (Model D) leads to noticeable declines, particularly on challenging long-tail subsets, confirming the necessity and synergy of our core components.

Qualitative Results

Long-tail Performance Figure 3 compares our SAML model with another model (Chen et al. 2024) on multimodal motion forecasting in various long-tail scenarios from the nuScenes dataset. Panels (a), (b), and (c) illustrate right-turn, left-turn, and deceleration maneuvers in congested intersections, respectively. The other model, biased towards common straight-line patterns, struggles with abrupt changes, leading to deviations from ground truth and limited multimodal diversity. In contrast, SAML achieves excellent performance in these long-tail scenarios. In panel (a), the target motion exhibits high velocity volatility (**top 10%** in kinematic dynamism), an intrinsic property reflecting abrupt deceleration, enabling SAML to forecast the right turn precisely by learning the deceleration feature preceding right turns. Panel (b) shows elevated lateral risk (**top 1%** in local interactive risk), arising from numerous conflict points typical in left turns at intersections, while panel (c) features high scene density and longitudinal risk (**top 3%** in local interactive risk), due to dense traffic impeding forward motion. SAML effectively learns these distinct long-tail features through its semantic-aware meta-learning framework, achieving precise forecasting. Additional visualizations of long-tail scenarios are provided in **Appendix F**.

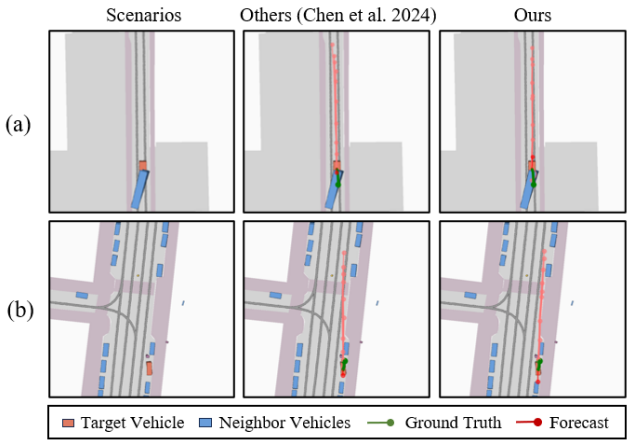

Figure 4: Visualization of failure cases on the nuScenes dataset. Red denotes the highest-probability forecast, and pink represents other multimodal options.

Failure Cases To investigate the limitations of SAML, we analyze two contrasting failure cases shown in Figure 4, which highlight a deeper challenge: resolving ambiguity within extreme long-tail scenarios. In the panel (a), a vehicle unexpectedly reverses on a two-way road—a highly rare maneuver—that SAML fails to anticipate, instead defaulting to a high-probability forward motion forecast. Conversely (b), another vehicle is angled in a manner suggestive of backing into a parking spot; SAML detects the anomalous orientation and forecasts a reversal, yet the vehicle moves forward slightly to adjust its position. These cases collectively illustrate that while SAML can identify a scenario's rarity (its "tailness"), it struggles to disambiguate the driver's intent when the cues from these rare events are contradictory. This limitation highlights the need for future research to focus on resolving the semantic ambiguity arising from conflicting cues, a defining challenge of extreme long-tail events.

Conclusion

In this paper, we introduce SAML, a novel framework that features a differentiable semantic meta-learning approach for long-tail motion forecasting in autonomous driving. Our method establishes the first principled way to quantify a motion's tailness, enabling interpretable, end-to-end optimization for rare events. The effectiveness of SAML is validated through extensive experiments that demonstrate superior performance in diverse long-tail scenarios. Our analysis of failure cases reveals a deeper challenge beyond simple rarity detection: while SAML excels at identifying a scenario's tailness, it struggles to resolve the ambiguity of contradictory cues within extreme events. This limitation highlights the need for future research to focus on resolving semantic ambiguity from conflicting cues.

Acknowledgments

This work was supported by the Science and Technology Development Fund of Macau [0129/2022/A, 0122/2024/RIB2, 0215/2024/AGJ, 001/2024/SKL], the Research Services and Knowledge Transfer Office, University of Macau [SRG2023-00037-IOTSC, MYRG-GRG2024-00284-IOTSC], the Shenzhen-Hong Kong-Macau Science and Technology Program Category C [SGDX20230821095159012], the Science and Technology Planning Project of Guangdong [2025A0505010016], National Natural Science Foundation of China [52572354], the State Key Lab of Intelligent Transportation System [2024-B001], and the Jiangsu Provincial Science and Technology Program [BZ2024055].

References

Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 961–971.

Bae, I.; Park, Y.-J.; and Jeon, H.-G. 2024. Singulartrajectory: Universal trajectory predictor using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17890–17901.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 11621–11631.

Chai, Y.; Sapp, B.; Bansal, M.; and Anguelov, D. 2020. MultiPath: Multiple Probabilistic Anchor Trajectory Hypotheses for Behavior Prediction. In *Conference on Robot Learning*, 86–99. PMLR.

Chen, J.; Wang, Z.; Wang, J.; and Cai, B. 2024. Q-EANet: Implicit social modeling for trajectory prediction via experience-anchored queries. *IET Intelligent Transport Systems*, 18(6): 1004–1015.

Chen, X.; Zhang, H.; Zhao, F.; Cai, Y.; Wang, H.; and Ye, Q. 2022a. Vehicle trajectory prediction based on intention-aware non-autoregressive transformer with multi-attention learning for Internet of Vehicles. *IEEE Transactions on Instrumentation and Measurement*, 71: 1–12.

Chen, X.; Zhang, H.; Zhao, F.; Hu, Y.; Tan, C.; and Yang, J. 2022b. Intention-aware vehicle trajectory prediction based on spatial-temporal dynamic attention network for internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(10): 19471–19483.

Deo, N.; and Trivedi, M. M. 2018. Convolutional social pooling for vehicle trajectory prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1468–1476.

Deo, N.; Wolff, E.; and Beijbom, O. 2022. Multimodal trajectory prediction conditioned on lane-graph traversals. In *Conference on Robot Learning*, 203–212. PMLR.

Feng, L.; Bahari, M.; Amor, K. M. B.; Zablocki, É.; Cord, M.; and Alahi, A. 2024. Unitraj: A unified framework for scalable vehicle trajectory prediction. In *European Conference on Computer Vision*, 106–123. Springer.

Gao, K.; Li, X.; Chen, B.; Hu, L.; Liu, J.; Du, R.; and Li, Y. 2023. Dual transformer based prediction for lane change intentions and trajectories in mixed traffic environment. *IEEE Transactions on Intelligent Transportation Systems*, 24(6): 6203–6216.

Gilles, T.; Sabatini, S.; Tsishkou, D.; Stanciulescu, B.; and Moutarde, F. 2022a. Gohome: Graph-oriented heatmap output for future motion estimation. In *2022 international conference on robotics and automation (ICRA)*, 9107–9114. IEEE.

Gilles, T.; Sabatini, S.; Tsishkou, D.; Stanciulescu, B.; and Moutarde, F. 2022b. THOMAS: Trajectory Heatmap Output with learned Multi-Agent Sampling. In *International Conference on Learning Representations*.

Girgis, R.; Golemo, F.; Codevilla, F.; Weiss, M.; D'Souza, J. A.; Kahou, S. E.; Heide, F.; and Pal, C. 2022. Latent Variable Sequential Set Transformers for Joint Multi-Agent Motion Prediction. In *International Conference on Learning Representations*.

Han, H.; Wang, W.-Y.; and Mao, B.-H. 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, 878–887. Springer.

Huang, M.; Zhu, M.; Xiao, Y.; and Liu, Y. 2021. Bayonet-corpus: a trajectory prediction method based on bayonet context and bidirectional GRU. *Digital Communications and Networks*, 7(1): 72–81.

Krajewski, R.; Bock, J.; Kloeker, L.; and Eckstein, L. 2018. 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)*, 2118–2125. IEEE.

Lan, Z.; Ren, Y.; Yu, H.; Liu, L.; Li, Z.; Wang, Y.; and Cui, Z. 2024. Hi-SCL: Fighting long-tailed challenges in trajectory prediction with hierarchical wave-semantic contrastive learning. *Transportation Research Part C: Emerging Technologies*, 165: 104735.

Li, X.; Huang, F.; Fan, Z.; Mou, F.; Hou, Y.; Qian, C.; and Wen, L. 2024. MetaTra: Meta-learning for generalized trajectory prediction in unseen domain. *arXiv preprint arXiv:2402.08221*.

Li, X.; Liu, J.; Li, J.; Yu, W.; Cao, Z.; Qiu, S.; Hu, J.; Wang, H.; and Jiao, X. 2023. Graph structure-based implicit risk reasoning for Long-tail scenarios of automated driving. In *2023 4th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 415–420. IEEE.

Liao, H.; Kong, H.; Wang, B.; Wang, C.; Ye, W.; He, Z.; Xu, C.; and Li, Z. 2025a. Cot-drive: Efficient motion forecasting for autonomous driving with llms and chain-of-thought prompting. *IEEE Transactions on Artificial Intelligence*.

Liao, H.; Li, X.; Li, Y.; Kong, H.; Wang, C.; Wang, B.; Guan, Y.; Tam, K.; and Li, Z. 2024a. Cdstraj: Characterized diffu-

sion and spatial-temporal interaction network for trajectory prediction in autonomous driving. In *IJCAI*, 7331–7339.

Liao, H.; Li, Z.; Shen, H.; Zeng, W.; Liao, D.; Li, G.; and Xu, C. 2024b. Bat: Behavior-aware human-like trajectory prediction for autonomous driving. volume 38, 10332–10340.

Liao, H.; Li, Z.; Wang, C.; Shen, H.; Liao, D.; Wang, B.; Li, G.; and Xu, C. 2024c. MFTraj: Map-Free, Behavior-Driven Trajectory Prediction for Autonomous Driving. In *IJCAI*.

Liao, H.; Li, Z.; Wang, C.; Wang, B.; Kong, H.; Guan, Y.; Li, G.; and Cui, Z. 2024d. A Cognitive-Driven Trajectory Prediction Model for Autonomous Driving in Mixed Autonomy Environments. In *IJCAI*.

Liao, H.; Li, Z.; Zhang, G.; Li, K.; and Xu, C. 2025b. Toward Human-Like Trajectory Prediction for Autonomous Driving: A Behavior-Centric Approach. *Transportation Science*.

Liao, H.; Wang, C.; Zhu, K.; Ren, Y.; Gao, B.; Li, S. E.; Xu, C.; and Li, Z. 2025c. Minds on the move: Decoding trajectory prediction in autonomous driving with cognitive insights. *IEEE Transactions on Intelligent Transportation Systems*.

Liu, M.; Cheng, H.; Chen, L.; Broszio, H.; Li, J.; Zhao, R.; Sester, M.; and Yang, M. Y. 2024. Laformer: Trajectory prediction for autonomous driving with lane-aware scene constraints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2039–2049.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Stella, X. Y. 2022. Open long-tailed recognition in a dynamic world. 46(3): 1836–1851.

Makansi, O.; Cicek, Ö.; Marrakchi, Y.; and Brox, T. 2021. On exposing the challenging long tail in future prediction of traffic actors. 13147–13157.

Mohamed, A.; Qian, K.; Elhoseiny, M.; and Claudel, C. 2020. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14424–14432.

Moon, S.; Woo, H.; Park, H.; Jung, H.; Mahjourian, R.; Chi, H.-g.; Lim, H.; Kim, S.; and Kim, J. 2024. Visiontrap: Vision-augmented trajectory prediction guided by textual descriptions. In *European Conference on Computer Vision*, 361–379. Springer.

Rao, B.; Liao, H.; Guan, Y.; Wang, C.; Wang, B.; Zhang, J.; and Li, Z. 2025. AMD: Adaptive Momentum and Decoupled Contrastive Learning Framework for Robust Long-Tail Trajectory Prediction. *arXiv preprint arXiv:2507.01801*.

Ren, Y.; Lan, Z.; Liu, L.; and Yu, H. 2024. EMSIN: Enhanced Multistream Interaction Network for Vehicle Trajectory Prediction. *IEEE Transactions on Fuzzy Systems*, 33(1): 54–68.

Ross, T.-Y.; and Dollár, G. 2017. Focal loss for dense object detection. 2980–2988.

Salzmann, T.; Ivanovic, B.; Chakravarty, P.; and Pavone, M. 2020. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, 683–700. Springer.

Wang, C.; Liao, H.; Li, Z.; and Xu, C. 2025a. WAKE: Towards Robust and Physically Feasible Trajectory Prediction for Autonomous Vehicles With WAvelet and KinEmatics Synergy. *PAMI*.

Wang, C.; Liao, H.; Wang, B.; Guan, Y.; Rao, B.; Pu, Z.; Cui, Z.; Xu, C.-Z.; and Li, Z. 2025b. Nest: A neuromodulated small-world hypergraph trajectory prediction model for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 808–816.

Wang, C.; Liao, H.; Zhu, K.; Zhang, G.; and Li, Z. 2025c. A dynamics-enhanced learning model for multi-horizon trajectory prediction in autonomous vehicles. *Information Fusion*, 118: 102924.

Wang, R.; Wang, S.; Yan, H.; and Wang, X. 2023a. Wsip: Wave superposition inspired pooling for dynamic interactions-aware trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4685–4692.

Wang, Y.; Zhang, P.; Bai, L.; and Xue, J. 2023b. Fend: A future enhanced distribution-aware contrastive learning framework for long-tail trajectory prediction. 1400–1409.

Xie, X.; Zhang, C.; Zhu, Y.; Wu, Y. N.; and Zhu, S.-C. 2021. Congestion-aware multi-agent trajectory prediction for collision avoidance. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 13693–13700. IEEE.

Xu, P.; Hayet, J.-B.; and Karamouzas, I. 2023. Context-aware timewise VAEs for real-time vehicle trajectory prediction. *IEEE Robotics and Automation Letters*, 8(9): 5440–5447.

Xu, Y.; Wang, L.; Wang, Y.; and Fu, Y. 2022. Adaptive trajectory prediction via transferable gnn. 6520–6531.

Yang, B.; Yan, K.; Hu, C.; Hu, H.; Yu, Z.; and Ni, R. 2024. Dynamic subclass-balancing contrastive learning for long-tail pedestrian trajectory prediction with progressive refinement. *IEEE Transactions on Automation Science and Engineering*.

Yuan, Y.; Weng, X.; Ou, Y.; and Kitani, K. M. 2021. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9813–9823.

Zhang, J.; Pourkeshavarz, M.; and Rasouli, A. 2024. Tract: A training dynamics aware contrastive learning framework for long-tail trajectory prediction. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, 3282–3288. IEEE.

Zhang, Q.; Yang, Y.; Li, P.; Andersson, O.; and Jensfelt, P. 2024. Seflow: A self-supervised scene flow method in autonomous driving. In *European Conference on Computer Vision*, 353–369. Springer.

Zhou, W.; Cao, Z.; Xu, Y.; Deng, N.; Liu, X.; Jiang, K.; and Yang, D. 2022. Long-tail prediction uncertainty aware trajectory planning for self-driving vehicles. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 1275–1282. IEEE.