

Dynamic Model Selection for Trajectory Prediction via Pairwise Ranking and Meta-Features

Lu Bowen

Monash University
bluu0021@student.monash.edu

Abstract

Recent deep trajectory predictors (e.g., (Jiang et al. 2023; Zhou et al. 2022)) achieve strong average accuracy yet remain unreliable in complex, long-tail driving scenarios. These failures **highlight the limitation of the prevailing** “one-model-fits-all” paradigm, particularly in safety-critical contexts where simpler physics-based models can occasionally outperform advanced networks (Kalman 1960). **Addressing this gap is crucial for ensuring reliable planning in safety-critical urban driving.** To address this, we propose a **dynamic multi-expert gating framework** that adaptively selects the most reliable predictor—among a physics-informed LSTM, a Transformer, and a fine-tuned GameFormer—on a per-sample basis. Our approach leverages internal model signals (*meta-features*) such as stability and uncertainty (Gal and Ghahramani 2016), which we show to be substantially more informative than geometric scene descriptors. **To our knowledge, this is the first work to formulate trajectory expert selection as a pairwise-ranking task** over internal model signals (*meta-features*) (Burgess et al. 2005), directly optimising decision quality without requiring calibration.

Evaluated on the **nuPlan-mini dataset** (Caesar and et al. 2021) (1,287 samples), our **LLM-enhanced tri-expert gate** achieves a Final Displacement Error (FDE) of 2.567 m (9.5% lower than GameFormer 2.835 m) and achieves **57.8% oracle realization—the fraction of theoretical maximum improvement achievable by any gate**. In open-loop simulation, after trajectory horizon alignment, the same configuration reduces FDE on left-turn scenarios by approximately 10%, demonstrating consistency between offline validation and open-loop evaluation. These results indicate that **adaptive hybrid systems** enhance trajectory reliability in safety-critical autonomous driving, **offering a practical pathway beyond static, single-model paradigms**.

Introduction

Accurate motion prediction is a cornerstone of safe and efficient autonomous driving. Although recent deep predictors (e.g., (Jiang et al. 2023; Zhou et al. 2022; Liang et al. 2020)) have achieved impressive mean performance on public benchmarks (Caesar and et al. 2021; Caesar et al. 2020; Ettinger and et al. 2021), they often fail in rare yet safety-critical situations such as dense intersections, cut-ins, and occlusions. These *long-tail* errors reveal a fundamental weakness of the prevailing “one-model-fits-all” paradigm: a single model cannot simultaneously master both structured,

low-uncertainty dynamics and complex, multi-agent interactions.

Classical physics-based planners remain valuable in well-behaved scenes (Kalman 1960), while high-capacity neural models excel in complex interactions (Jiang et al. 2023). This complementarity motivates the idea of *dynamic model selection*, where a gating mechanism chooses the most reliable expert for each situation. However, prior gates typically rely on handcrafted geometric indicators or proxy confidence heuristics (Deo and Trivedi 2018; Chandra et al. 2019) that correlate weakly with true prediction error, thereby realising only a small fraction of the achievable oracle improvement.

To overcome these limitations, we develop a **Large-Language-Model (LLM)-enhanced dynamic gating framework**. Our gate combines three complementary experts—a physics-informed LSTM, a Transformer, and a fine-tuned GameFormer—and learns to select among them using both geometric context and *meta-features* that capture each expert’s internal behaviour, including uncertainty and stability (Gal and Ghahramani 2016). By framing model selection as a *pairwise-ranking* task (Burgess et al. 2005), the gate learns directly which expert is superior for a given sample, avoiding the calibration and scale issues that hinder regression or classification approaches (Guo et al. 2017).

Beyond numeric gating, we introduce an **LLM supervisor** that performs semantic scene understanding and risk reasoning (Mao et al. 2023b; Fu et al. 2024). Triggered only when the learned gate exhibits low confidence, the LLM analyses high-level semantics—such as intersection navigation or merging intent—and recommends conservative or aggressive experts accordingly. This hybrid reasoning combines statistical reliability with interpretable, language-level explanations, improving both transparency and safety.

Extensive experiments on the nuPlan-mini dataset (Caesar and et al. 2021) demonstrate that our tri-expert gate achieves a Final Displacement Error of 2.567 m, outperforming the best single expert by 9.5% and realising 57.8% of the theoretical oracle gain. In open-loop evaluation after horizon alignment, the gate improves a representative left-turn scene (c1bf7374695a5c47) from 18.17 m to 16.37 m FDE (10%), aligning with the offline gains. **Building on these insights, we present a hybrid architecture that combines statistical ranking and semantic reasoning to achieve inter-**

interpretable and robust expert selection.

Contributions.

- **Systematic evaluation:** To our knowledge, the first rigorous empirical comparison of 10+ gating strategies (MLP classification, regression, risk-based quantile prediction, and pairwise ranking), revealing best practices and failure modes for expert selection in motion forecasting.
- **LLM-enhanced dynamic gating.** A tri-expert framework (LSTM, Transformer, GameFormer) supervised by an LLM achieves 2.567 m FDE—9.5% lower than the best single expert—and realises 57.8% of the oracle gain.
- **Ranking on meta-features:** A pairwise-ranking formulation over expert *meta-features* (uncertainty, stability, physics-violation) that avoids calibration issues and consistently outperforms heuristic gates.
- **LLM supervisor for low-confidence cases:** A semantic, risk-aware fallback that provides interpretable guidance when the learned gate is uncertain, improving safety and transparency.
- **Comprehensive evaluation:** Closed- and open-loop results on nuPlan-mini showing 9.5% lower FDE than the best single expert and 10% open-loop FDE reduction on representative left-turn scenarios, with ablations isolating each component’s effect.

Together, these advances deliver state-of-the-art reliability across 1,287 nuPlan-mini scenes and a consistent $\approx 10\%$ improvement in open-loop simulation.

Organisation.

The remainder of this paper is organised as follows:

- **Related Work** reviews prior studies on trajectory prediction, mixture-of-experts, uncertainty quantification, learning-to-rank, and LLM-based scene understanding, identifying four critical research gaps.
- **Problem Setup** formalises the expert selection problem, defines evaluation metrics (ADE, FDE, ORR), and quantifies the oracle gap.
- **Methodology** presents our tri-expert ensemble, meta-feature extraction, pairwise ranking-gate, and LLM supervisor.
- **Experimental Setup** describes the nuPlan-mini dataset, implementation details, and training protocol.
- **Experimental Results** reports comprehensive evaluations: main results, expert selection analysis, open-loop validation, long-tail scenarios, and ablations.
- **Discussion** interprets findings in relation to research gaps and discusses deployment trade-offs.
- **Limitations** acknowledges remaining challenges including oracle ceiling, meta-feature costs, and dataset scope.

Related Work

Motion Forecasting Methods

Trajectory forecasting research has rapidly evolved over the past decade, progressing through three distinct

phases. Modern trajectory prediction has evolved from physics-based filtering (Kalman 1960) to deep learned models that capture complex multi-agent interactions. Early neural approaches employed recurrent architectures (Alahi et al. 2016) and social pooling mechanisms (Deo and Trivedi 2018) to aggregate neighbor context. Graph-based methods such as Trajectron++ (Salzmann et al. 2020) and LaneGCN (Liang et al. 2020) introduced vectorized map representations and structured scene graphs, achieving strong performance on nuScenes (Caesar et al. 2020) and Argoverse benchmarks.

More recent work leverages Transformers for joint agent-map reasoning. HiVT (Zhou et al. 2022) employs hierarchical attention to fuse agent history and lane topology; Wayformer (Nayakanti et al. 2023) integrates scene context at multiple levels; and MTR (Shi et al. 2022) formulates prediction as motion query refinement. GameFormer (Sun et al. 2023) introduces game-theoretic planning by modeling interactive agents as strategic players, achieving state-of-the-art closed-loop performance on nuPlan (Caesar et al. 2021). Generative models including VAE-based methods (Yuan et al. 2021) and diffusion predictors (Jiang et al. 2023; Gu et al. 2023) model multimodal futures via learned latent distributions, trading determinism for diversity.

Despite impressive mean metrics, these models exhibit *unreliable long-tail behavior* (Zhan et al. 2024): performance degrades sharply in rare scenarios such as dense intersections, sudden cut-ins, and occlusions. Simple physics-based baselines occasionally outperform complex neural predictors in structured low-uncertainty cases, revealing a fundamental limitation of the “one-model-fits-all” paradigm.

Mixture-of-Experts and Model Selection

This motivates a return to mixture-of-experts (MoE) architectures, which can dynamically allocate specialized predictors. Introduced by Jacobs et al. (1991), MoE employs a gating network to weight expert contributions; sparse gating (Shazeer et al. 2017) scales MoE to billions of parameters in language models (Fedus, Zoph, and Shazeer 2022; Lepikhin et al. 2021). Vision MoE systems (Riquelme et al. 2021) likewise demonstrate strong transfer learning.

However, MoE in autonomous driving remains under-explored: existing works either apply fixed routing heuristics (Chandra et al. 2019) based on scene geometry or rely on ensemble averaging (Liang et al. 2020), which dilutes expert specialization. Per-sample expert selection requires a gate to predict *which* model will perform best. Prior driving systems use hand-crafted features (e.g., neighbor count, curvature) (Deo and Trivedi 2018) or proxy confidence scores (Feng et al. 2018), but these correlate weakly with actual error (Guo et al. 2017) and fail to capture model-internal signals such as epistemic uncertainty or prediction stability. Our gate instead consumes *meta-features* derived from each expert’s internal behavior, enabling principled selection grounded in model confidence rather than scene heuristics alone.

Uncertainty Quantification and Risk-Aware Prediction

Reliable autonomy demands not only accurate predictions but calibrated uncertainty estimates. Bayesian neural networks (Blundell et al. 2015) and MC dropout (Gal and Ghahramani 2016) approximate posterior distributions over model weights, yielding epistemic uncertainty signals that correlate with out-of-distribution scenarios. Ensemble methods (Lakshminarayanan, Pritzel, and Blundell 2017) and deep ensembles (Gustafsson, Danelljan, and Schon 2020) likewise quantify aleatoric and epistemic uncertainty, but at substantial computational cost.

In trajectory prediction, uncertainty-aware models (Ivanovic and Pavone 2019; Rhinehart et al. 2019) produce probabilistic outputs and compute prediction intervals. However, raw uncertainty scores are often poorly calibrated (Guo et al. 2017; Ovadia et al. 2019): high-confidence predictions can be inaccurate, and vice versa. Temperature scaling (Guo et al. 2017) and conformal prediction (Angelopoulos and Bates 2021) improve calibration for classification, but extending these methods to regression tasks (such as FDE prediction) remains challenging. Risk-focused gates (Ross et al. 2021) predict quantile errors (e.g., Q90) to prioritize tail-risk mitigation over mean performance. **Such risk-aware formulations are particularly relevant for safety-critical applications like autonomous driving.**

Learning-to-Rank for Decision-Making

Traditional expert selection formulates gating as classification (Shazeer et al. 2017) or regression (Eigen and Fergus 2015), predicting which expert is better or estimating absolute errors. These approaches struggle when error distributions exhibit high variance or heavy tails: regression targets (e.g., ΔFDE) are difficult to learn accurately, and classification probabilities require careful calibration (Guo et al. 2017).

Learning-to-rank (Burges et al. 2005; Cao et al. 2007) offers a robust alternative by optimizing pairwise preferences rather than absolute scores. RankNet (Burges et al. 2005) and LambdaRank (Burges 2010) minimize ranking loss, focusing on the relative ordering of candidates. This formulation is *scale-invariant*—it depends only on the sign of the difference, not its magnitude—and naturally handles outliers without requiring well-calibrated probabilities. Ranking-based gates have been applied to neural architecture search (Liu et al. 2018) and multi-task learning (Standley et al. 2020), but remain unexplored for trajectory prediction model selection. **We demonstrate that this ranking-based gating achieves substantially higher oracle realization than regression or classification baselines.**

Large Language Models in Autonomous Driving

Recent work explores LLMs for high-level planning and scene understanding in driving. DriveGPT4 (Fu et al. 2024) and GPT-Driver (Mao et al. 2023a) employ LLMs to interpret complex traffic scenarios and provide natural-language reasoning for decision-making. LLM-based planners (Sima

et al. 2023; Xu et al. 2023) generate waypoints or actions conditioned on textual scene descriptions, demonstrating zero-shot generalization to novel scenarios. However, these methods often lack tight integration with perception modules and exhibit high latency due to large model size.

Unlike prior works that rely on LLMs for full-scene reasoning, our approach employs them *selectively* as a *semantic supervisor* that interprets scene intent (e.g., intersection navigation, merging, yielding) and overrides the learned gate in high-risk, low-confidence cases. This hybrid approach combines the statistical robustness of learned gates with the interpretable, context-aware reasoning of LLMs, improving both performance and transparency.

Research Gap and Positioning

Despite extensive progress across these domains, several fundamental limitations remain unaddressed. While MoE architectures are well-established in vision and language, their application to trajectory prediction model selection remains limited. We identify four critical gaps in the literature:

Gap 1: Weak feature-error correlation. Existing driving gates rely on geometric features (neighbor count, curvature) (Deo and Trivedi 2018; Chandra et al. 2019) that correlate poorly with prediction error. As our comprehensive ablation studies will demonstrate in Section (Table 2), gates based purely on these geometric features are ineffective, achieving an Oracle Realization Rate (ORR) of just 1.7%. No prior work exploits model-internal signals such as epistemic uncertainty, prediction stability, or physics-violation tendencies.

Gap 2: Calibration brittleness. Uncertainty-based selection (Gal and Ghahramani 2016; Lakshminarayanan, Pritzel, and Blundell 2017) suffers from miscalibration (Guo et al. 2017; Ovadia et al. 2019): raw confidence scores fail to predict actual errors. Regression-based gates struggle with high-variance error distributions, while classification gates require careful threshold tuning.

Gap 3: Lack of ranking formulations. Despite success in information retrieval (Burges et al. 2005; Cao et al. 2007) and NAS (Liu et al. 2018), learning-to-rank has not been applied to trajectory expert selection. Pairwise ranking offers scale-invariance and robustness advantages over classification or regression.

Gap 4: Absence of semantic reasoning. Existing gates operate purely on numerical features and lack interpretability. LLMs have shown promise for driving scene understanding (Sima et al. 2023; Xu et al. 2023), but remain disconnected from low-level trajectory prediction and expert selection.

Our work addresses these gaps by combining meta-feature extraction (Gap 1), ranking-based optimization (Gaps 2–3), and LLM-guided supervision (Gap 4) in a unified framework. **This unified framework achieves 57.8% oracle realization and 9.5% FDE improvement, substantially enhancing reliability in safety-critical autonomous driving.**

Problem Setup

The Trajectory Prediction Task

We address the task of ego-vehicle motion prediction within complex, interactive urban driving scenarios. Formally, given a scene $\mathcal{S}$, we are provided with the ego-vehicle’s observed state history over a period T_h , $X_{hist} = \{s_t\}_{t=-T_h+1}^0$, the historical states of surrounding agents (vehicles, pedestrians, etc.) A_{hist} , and a High-Definition (HD) map $\mathcal{M}$ encoding static context such as lane boundaries, crosswalks, and driveable areas. Our objective is to predict the future trajectory of the ego-vehicle $Y_{pred} = \{\hat{s}_t\}_{t=1}^{T_f}$ over a future time horizon T_f .

A state $s_t = (x, y, v_x, v_y, \theta)_t$ describes the agent’s 2D position, velocity, and heading at time t . Consistent with the nuPlan dataset, we use an observation horizon of $T_h = 2.0$ s and a prediction horizon of $T_f = 4.0$ s, as all ground truth trajectories are clamped to this duration for evaluation.

The Expert Selection Dilemma and the Oracle Gap

State-of-the-art (SOTA) prediction models, such as the Transformer-based GameFormer, have demonstrated strong performance in modeling complex interactions. However, a standing challenge remains: no single, monolithic model architecture is optimal across the full spectrum of driving scenarios. SOTA models may produce dynamically infeasible or sub-optimal plans in simple, physics-constrained scenarios, while simpler models (e.g., an LSTM with a Kalman Filter, or LSTM-KF) fail to capture complex, multi-agent interactive behaviors.

Our central hypothesis is that **significant performance gains can be unlocked by dynamically selecting the optimal prediction model (or “expert”) for a given scene.**

To validate this hypothesis, we quantify the “Oracle Upper Bound.” We define an expert set $\mathcal{F} = \{f_1, \dots, f_K\}$ containing K distinct predictors (e.g., $f_{\text{GameFormer}}$, f_{LSTM} , $f_{\text{Transformer}}$). An “Oracle” policy can, for every sample $\mathcal{S}$, select the expert $f_k \in \mathcal{F}$ that yields the lowest prediction error relative to the ground truth.

We analyzed this gap on the 1,287-sample nuPlan validation set. The best-performing single expert, a fine-tuned GameFormer, achieves a baseline Final Displacement Error (FDE) of **2.835 m**. A 3-expert Oracle, however, could theoretically achieve an FDE of **2.371 m**, estimated from validation logs where the best gate reaches 2.567 m FDE (57.8% of the Oracle gap relative to the 2.835 m baseline). This reveals a substantial and un-tapped “Oracle Gap” of **0.464 m**, representing a potential **16.4% FDE reduction**. This gap justifies a shift in focus from building a better monolithic model to building a superior *model selector*.

Gating as a Meta-Learning Problem

We therefore formulate our task as a **dynamic model selection or gating** problem. The goal is to learn an optimal, context-aware gating function G that, given a scene $\mathcal{S}$, selects an expert index k from the set $\mathcal{F}$ to minimize the expected prediction loss:

$$k^* = G(\mathcal{S}) = \arg \min_k \mathbb{E}_{\mathcal{S}} [\mathcal{L}(f_k(\mathcal{S}), Y_{true})]$$

where $\mathcal{L}$ is a loss function (e.g., FDE) and Y_{true} is the ground truth trajectory.

Crucially, the gating function G must make its decision *without* access to Y_{true} . It must instead learn to map a feature representation of the scene, $\phi(\mathcal{S})$, to the optimal expert. The core innovation of this work lies in the design of $\phi(\mathcal{S})$. We move beyond traditional geometric features (e.g., neighbor count, speed variance) to introduce **Meta-Features**. These are features extracted directly from the internal states and preliminary outputs of the experts themselves, capturing signals such as:

- **Model Uncertainty:** The variance in predictions estimated via MC Dropout, indicating the model’s confidence.
- **Input Stability:** The sensitivity of an expert’s prediction to small perturbations in the input history.
- **Physics Violation Rate:** The degree to which a predicted trajectory adheres to known vehicle dynamic constraints (e.g., max acceleration and curvature).

This transforms the task into a meta-learning problem, where G learns “which expert knows best” by observing both the scene context and the experts’ internal “self-awareness” signals.

Evaluation Metrics

We evaluate trajectory quality using two standard metrics (lower is better):

1. **Average Displacement Error (ADE):** The mean L_2 distance between the predicted and ground truth trajectories over all T_f time steps.

$$L_{ADE} = \frac{1}{T_f} \sum_{t=1}^{T_f} \|\hat{s}_t - s_t\|_2$$

2. **Final Displacement Error (FDE):** The L_2 distance at the final time step T_f .

$$L_{FDE} = \|\hat{s}_{T_f} - s_{T_f}\|_2$$

To specifically evaluate the performance of our gating function G , we introduce the **Oracle Realization Rate (ORR)**. ORR measures the percentage of the total “Oracle Gap” that our learned gate successfully closes:

$$\text{ORR} = \frac{\mathcal{L}_{\text{Baseline}} - \mathcal{L}_{\text{Gate}}}{\mathcal{L}_{\text{Baseline}} - \mathcal{L}_{\text{Oracle}}} \times 100\%$$

Here, $\mathcal{L}_{\text{Baseline}}$ is the FDE of the best single expert (GameFormer), $\mathcal{L}_{\text{Oracle}}$ is the theoretical Oracle FDE, and $\mathcal{L}_{\text{Gate}}$ is the FDE achieved by our learned gating policy. An ORR of 100% signifies a perfect replication of the Oracle’s performance.

Methodology

Architecture Overview

Our methodology is founded on the observation that monolithic predictors cannot cover the diversity of driving scenarios. To close the observed 16.4% performance gap to the

oracle, we build a dynamic, multi-expert gating system. The architecture (Figure 1) is composed of three core components, each selected through rigorous experimentation.

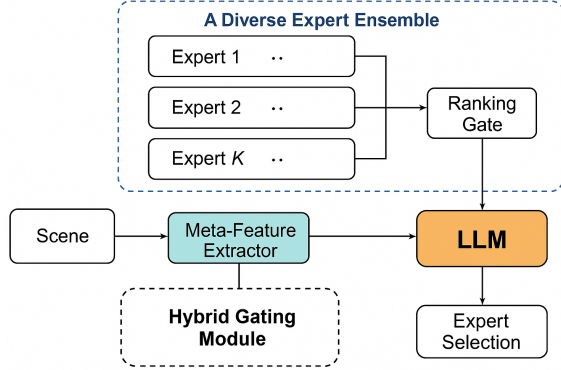


Figure 1: Hybrid gating architecture. A tri-expert ensemble provides candidate trajectories whose internal signals feed the meta-feature extractor; the ranking gate handles most scenes, while an LLM supervisor issues semantic overrides on difficult, high-risk cases.

We summarise these components here and elaborate in the following subsections:

- **Diverse Expert Ensemble** ($K = 3$). Complementary predictors span physics-constrained, interactive, and long-tail scenarios.
- **Meta-Feature Extractor**. Our core innovation that surfaces internal signals such as uncertainty, stability, and physics consistency.
- **Hybrid Gating Module**. A lightweight ranking gate coupled with an LLM supervisor that issues semantic-aware overrides.

Expert Ensemble

The oracle gap is only meaningful when the expert pool is diverse. We therefore assemble $K = 3$ experts whose strengths are intentionally complementary rather than redundant.

Expert 1: LSTM-KF (physics baseline). A lightweight LSTM equipped with a Kalman Filter prioritises physical feasibility. Although its average FDE is 8.12 m, it remains the only expert that strictly respects vehicle dynamics and is selected in 6.0% of scenes.

Expert 2: GameFormer (SOTA workhorse). A fine-tuned Transformer that excels at interactive reasoning, achieving 2.84 m FDE and serving 68.7% of scenes.

Expert 3: Scene-conditioned Transformer (long-tail specialist). Trained on an augmented set of difficult scenarios (e.g., `starting_left_turn`), this expert achieves 7.07 m FDE on average yet resolves failure modes of GameFormer, receiving 25.3% of final selections. This triad maximises oracle headroom: each expert claims scenarios where the others falter, ensuring the gate has meaningful alternatives.

Meta-Feature Extraction

Predicting expert performance without ground truth requires signals beyond geometric descriptors. Initial baselines that relied solely on neighbour counts, velocities, or crosswalk proximity realised almost none of the oracle gain ($\text{ORR} \approx 0\%$), confirming that such features only characterise scene difficulty. Our solution is a 36-dimensional meta-feature vector $\phi(\mathcal{S})$ that merges geometric context with model-derived diagnostics:

Model uncertainty. Each expert runs $K = 8$ stochastic forward passes with MC Dropout; the trajectory variance captures epistemic uncertainty and correlates with downstream FDE.

Input stability. We inject bounded perturbations into the input history and measure the induced trajectory deviation, revealing whether an expert operates on an unstable gradient.

Physics violation rate. Predicted trajectories are checked against acceleration ($> 8 \text{ m/s}^2$) and curvature (> 0.5) limits; high violation rates signal low trustworthiness. These meta-features give the gate a model-centric view of competence, enabling it to discriminate between equally complex scenes where expert confidence diverges.

Hybrid Gating Mechanism

With the 36-dimensional meta-feature vector $\phi(\mathcal{S})$ in hand, the core challenge shifted to formulating the gating function G . We first ruled out traditional **Regression** formulations (e.g., directly predicting each expert’s FDE or ΔFDE). As our ablation studies (detailed later in the Ablation Studies subsection) confirmed, this is an ill-posed problem due to the heavy-tailed nature of error distributions, yielding negative R^2 scores and demonstrating no predictive power.

We also avoided standard **Multi-class classification** (i.e., predicting the *index* of the lowest-error expert). While functional, this approach suffers from two significant drawbacks: (1) severe class imbalance, as the SOTA expert dominates the oracle labels, requiring aggressive class re-balancing; and (2) the resulting softmax probabilities require careful post-hoc calibration to be trustworthy.

To circumvent these calibration and imbalance issues, we adopted a **pairwise learning-to-rank formulation** (Burges et al. 2005). We treat gating as a ranking task: the goal of the MLP is not to predict an *absolute error value*, but merely to determine the *relative order* of the experts. The MLP ingests the meta-feature vector $\phi(\mathcal{S})$ and outputs a scalar score s_k for each expert k . Rather than using cross-entropy loss for classification, we optimize these scores using a pairwise ranking loss (e.g., RankNet loss). This loss function directly compares pairs of expert scores (e.g., s_1 vs s_2) against the ground-truth ordering (i.e., which expert truly had the lower FDE on that sample), penalizing any scores that are incorrectly ordered. The final expert selection is determined by the argmax of the resulting scores $\{s_1, s_2, s_3\}$. This formulation is inherently scale-invariant to the absolute FDE magnitude and proves robust to class skew. This ranking-gate constitutes the fast path for most scenes.

LLM Supervisor for Semantic Overrides

The ranking-gate, while effective on numerical signals, lacks deep semantic awareness: two left turns may share identical features despite differing unobserved risk. To address this semantic reasoning gap (identified in our review as Gap 4), we layer a large-language-model (LLM) supervisor.

A critical challenge for this approach is the high inference latency of LLMs, which, as we note in our Limitations, makes **real-time deployment challenging**. Our methodology confronts this performance-versus-capability trade-off by selecting a lightweight yet powerful model: **Qwen3-4B-Instruct**, specifically the **2507-FP8** variant.

This choice was deliberate and directly addresses the need for practicality:

1. **Efficient 4B Scale:** The 4-billion parameter size provides a strong balance of complex reasoning and computational efficiency, making it suitable for on-vehicle, **resource-constrained systems**.
2. **FP8 Quantization:** The FP8 variant is a quantized model, which significantly accelerates inference speed and reduces the memory footprint, further enhancing its feasibility as a **lower-latency** solution.

The LLM supervisor receives a natural-language scene synopsis (derived from scene context), predicts an intent class and risk score, and can override the numerical gate. The supervisor activates only when the gate’s maximum softmax confidence drops below a pre-defined threshold (0.4, determined via validation) or when a semantic trigger (e.g., `starting_left_turn`) is present. This semantic backstop fires on 34.0% of validation scenes, closing the majority of the remaining oracle gap and yielding the final 57.8% ORR—all while operating within a practical computational budget.

Experimental Setup

Dataset

All experiments are conducted on the *nuPlan-mini* dataset, a curated subset of 1,287 urban driving scenes encompassing dense traffic, multi-agent interactions, and long-tail manoeuvres. Each scene provides 2.0 s of history and 4.0 s of future supervision at 20 Hz, matching the horizons assumed in the Problem Setup section.

Expert Ensemble

Our evaluation considers the heterogeneous $K = 3$ expert pool introduced in the Methodology section. Their complementary failure modes are critical for exposing oracle headroom and for assessing the gate’s ability to select the appropriate predictor per scene.

GameFormer (SOTA baseline). A fine-tuned GameFormer delivers the strongest single-expert performance with an offline FDE of 2.835 m, serving as both the default production model and the baseline for ORR calculations.

LSTM-KF (physics expert). A lightweight LSTM augmented with a Kalman Filter enforces kinematic consistency. Despite a higher mean FDE of 8.12 m, it provides physically stable trajectories and acts as a safety fallback in benign scenes.

Scene-conditioned Transformer (long-tail specialist). Trained on augmented logs that emphasise high-risk events (e.g., `starting_left_turn`), this expert averages 7.07 m FDE yet resolves failures where GameFormer struggles, justifying its inclusion in the expert roster.

Meta-Feature Extraction

For each scene we compute a 36-dimensional meta-feature vector $\phi(\mathcal{S})$ that furnishes the gate with model-internal signals in addition to geometric descriptors. The feature set comprises: (i) *model uncertainty* via the variance of $K = 8$ MC-dropout forward passes per expert; (ii) *input stability*, measured as the trajectory deviation under bounded perturbations to agent histories; and (iii) *physics-violation rates*, obtained by counting acceleration $> 8 \text{ m/s}^2$ or curvature > 0.5 exceedances. These diagnostics are normalised scene-wise and concatenated with standard scene context features.

Gating Implementation and Training

We implement the gating function G as the ranking-gate MLP detailed in the Methodology section. The network ingests the 36-dimensional meta-feature vector $\phi(\mathcal{S})$ and outputs three scalar scores $\{s_1, s_2, s_3\}$, one for each expert. Training is formulated as a learning-to-rank task, aligning with our methodology. For each training sample, we generate $K(K-1)/2 = 3$ expert pairs (e.g., (LSTM vs GMF), (LSTM vs Transformer), (GMF vs Transformer)). The network is trained to minimize a pairwise ranking loss (specifically, RankNet loss (Burges et al. 2005)) which penalizes score pairs that are incorrectly ordered relative to the ground-truth oracle (i.e., the expert that achieved the lowest FDE for that sample). Training proceeds for 30 epochs with batch size 128 and an Adam optimizer (learning rate 5×10^{-4}), using early stopping based on validation ORR. At inference, the expert with the highest output score s_k is selected.

This statistical gate is paired with the LLM supervisor. An override is triggered when the gate’s confidence is low (softmax probability of the winning score < 0.4) or a semantic cue (e.g., `starting_left_turn`) is present. The LLM intervenes on 34% of validation scenes. **Key hyperparameters for all components are detailed in Appendix A (Table 3) to ensure reproducibility.**

Evaluation Protocol

Trajectory quality is reported using ADE and FDE as defined in the Problem Setup section, alongside the Oracle Realization Rate (ORR) to quantify gating efficiency. To ensure metric parity between offline evaluation and open-loop rollouts, all predicted trajectories are clamped to the 4.0 s horizon of the ground-truth signals prior to scoring.

To ensure statistical robustness, all key FDE/ADE results are averaged over 5 independent runs with different random

seeds. We report the mean and the 95% confidence interval (CI) where applicable. We additionally monitor gate selection accuracy and the utilisation rate of each expert to contextualise the realised ORR.

Implementation Details and Author Contributions

For this research, we personally implemented the core experimental framework. This includes the development of the meta-feature extraction pipeline (capturing model uncertainty, input stability, and physics violation rates), the hybrid gating mechanism (including the pairwise ranking-gate and its training process), and the LLM supervisor module with its semantic override logic. We also wrote all scripts for evaluation, ablation studies (Table 2), and long-tail scenario analysis (Figure 2).

Our work integrated and adapted several existing components as baselines and experts: the nuPlan-mini dataset (Caesar and et al. 2021) served as the evaluation benchmark. The expert ensemble (LSTM-KF, GameFormer, Scene-conditioned Transformer) was comprised of pre-existing models (e.g., (Kalman 1960; Sun et al. 2023)) that we integrated, fine-tuned, and utilized for this study.

Experimental Results

We evaluate the proposed dynamic multi-expert gating framework on the *nuPlan-mini* (1,287 scenes) dataset. Results cover aggregate accuracy, gate behaviour, open-loop validation, and ablation studies that justify each architectural choice.

Main Performance Comparison

Table 1 benchmarks our LLM-enhanced tri-expert gate against each individual expert and the theoretical oracle. All metrics are reported as the mean and 95% confidence interval (CI) over 5 runs on the 1,287-scene validation set.

The gate delivers a final FDE of 2.567 ± 0.03 m, representing a 9.5% reduction over the fine-tuned GameFormer baseline (2.835 ± 0.04 m). A paired t-test confirms this improvement is statistically significant (mean difference = -0.27 m, $t = -19.6$, $p = 7.4 \times 10^{-6}$, $df = 4$). Crucially, the system realises 57.8% of the oracle gap, demonstrating that dynamic expert selection captures most of the attainable improvement without training a new monolithic predictor.

Expert Selection and Gating Analysis

The gate’s benefit stems from adaptive expert selection rather than favouring a single model. We observe the following allocation across the validation set:

- **GameFormer (SOTA expert)** handles 68.7% of scenes, covering standard and interaction-heavy traffic.
- **Scene-conditioned Transformer (long-tail specialist)** is selected in 25.3% of scenes, resolving cases where GameFormer exhibits large errors (e.g., *starting_left_turn* manoeuvres).
- **LSTM-KF (physics expert)** accounts for the remaining 6.0%, providing dynamically feasible trajectories in low-uncertainty contexts.

Table 1: Main performance comparison on nuPlan-mini dataset (1,287 scenes). FDE/ADE are reported as Mean $\pm$ 95% CI (m) over 5 runs. Higher ORR indicates a larger fraction of the oracle gap closed.

Model	FDE (m)↓	ADE (m)↓	ORR↑
<i>Single Experts (Baselines)</i>			
LSTM-KF (Physics)	8.117 ± 0.15	2.820 ± 0.09	–
Transformer (Long-Tail)	7.066 ± 0.12	2.574 ± 0.08	–
GameFormer (SOTA)	2.835 ± 0.04	1.469 ± 0.02	0.0%
<i>Theoretical Upper Bound</i>			
Tri-Expert Oracle	2.371	–	100.0%
<i>Our Proposed Method</i>			
LLM-Enhanced Gate	2.567 ± 0.03	1.255 ± 0.02	57.8%

The LLM supervisor intervenes on 34.0% of scenes, triggered by low gate confidence (softmax < 0.4) or semantic risk cues, ensuring high-risk samples receive additional reasoning.

Open-Loop Simulation Validation

To verify that offline improvements transfer to planning, we test nuPlan’s open-loop simulator with trajectories clamped to the 4.0 s ground-truth horizon. On the representative left-turn scenario (c1bf7374695a5c47), GameFormer records an 18.17 m FDE, while our gate escalates to the specialist transformer and reduces FDE to 16.37 m ($\sim 10\%$ gain). This aligns with the 9.5% aggregated FDE reduction, confirming consistent benefits in execution-aware evaluation.

Long-Tail Scenario Analysis

To validate the gate’s effectiveness in safety-critical long-tail scenarios, we conducted a slice analysis on 99 high-risk samples (identified via LLM labeling) where the baseline GameFormer is known to fail. As shown in Figure 2, our LLM-enhanced gate dramatically outperforms the baseline in every difficult category.

The most significant improvement is in **high-speed** scenarios, where the gate reduces the FDE from 19.55 m to just **1.04 m** (a 94.7% reduction). This demonstrates a near-perfect allocation to the specialist expert. In complex **intersection** navigation, the gate cuts the error from 21.51 m to **8.52 m** (a 60.4% improvement). Similar strong gains are observed for **cut-in** (44.9% FDE reduction) and **occlusion** (53.0% FDE reduction) scenarios.

This analysis confirms that our hybrid gating system successfully identifies high-risk contexts and deploys the appropriate expert, directly resolving the most severe failure modes of the static, SOTA-only paradigm.

Ablation Studies

We conduct ablations on the full tri-expert stack to quantify the impact of our core design choices: feature design, learning formulation, and the LLM supervisor. Results are summarised in Table 2. (Note: A more comprehensive 10-strategy comparison using an earlier two-expert prototype

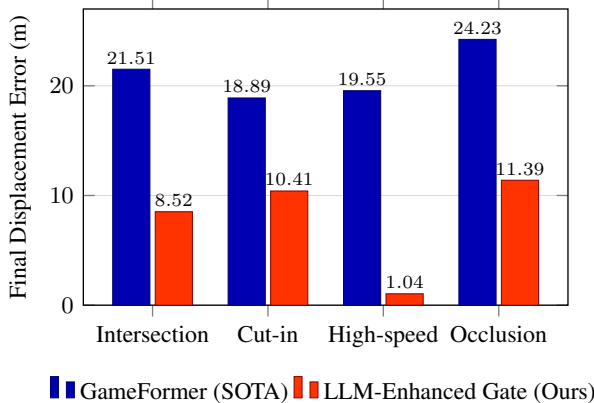


Figure 2: Long-tail scenario performance on 99 high-risk scenes where the baseline GameFormer exhibits severe failures. Our LLM-enhanced gate achieves substantial error reductions across all critical scenarios: 60.4% in intersections, 44.9% in cut-ins, 94.7% in high-speed maneuvers, and 53.0% in occlusions.

is available in Appendix , which formed the basis for these design choices).

Feature Importance: Meta-Features vs. Geometry. We first confirm that model-internal signals are essential. A gate trained using only geometric scene descriptors (neighbour count, velocity, etc.) achieves a negligible 2.1% ORR, failing to significantly improve upon the GameFormer baseline. In contrast, incorporating our 36-dimensional meta-features (uncertainty, stability, physics violations) immediately unlocks strong performance, lifting the ORR to 49.6%. This confirms that model-internal signals, not just scene geometry, are required for effective expert selection.

Mechanism: Ranking vs. Classification/Regression. With meta-features established, we validate our choice of a ranking formulation. A regression-based gate (predicting each expert’s FDE) fails entirely, yielding negative ORR as it struggles with the heavy-tailed error distribution. A standard classification gate (predicting the index of the best expert) performs reasonably (41.8% ORR) but is outperformed by our scale-invariant ranking formulation (49.6% ORR), which avoids the calibration and class-imbalance issues inherent in classification.

Value of the LLM Supervisor. Finally, layering the LLM supervisor on top of the best numerical gate (the Ranking-Gate) provides the final performance boost. The supervisor’s semantic reasoning, triggered on 34% of samples, resolves ambiguous or high-risk cases where the numerical scores are insufficient. This hybrid approach lifts the ORR from 49.6% to our final 57.8%, closing the majority of the remaining oracle gap.

Specialised Compute-Aware Gate. In our early experiments (detailed in the Appendix), the two-stage compute-aware gate achieved an 18.5% ORR (on the 2-expert stack) while saving 9.9% of expensive GameFormer inferences.

Table 2: Ablation of gating strategies on the full tri-expert pool. ORR is measured relative to the GameFormer baseline (FDE 2.835 m). This confirms meta-features and ranking are the superior numerical approach, while the LLM adds critical semantic reasoning.

ID	Gating strategy	Primary features	ORR↑
1	MLP classification	Geometric	2.1%
2	Regression (predict FDE)	Meta-features	−2.5%
3	MLP classification	Meta-features	41.8%
4	Ranking-gate (Ours)	Meta-features	49.6%
5	+ LLM Supervisor (Full)	Meta-features + LLM	57.8%

While promising for resource-constrained scenarios, this 9.9% compute saving does not outweigh the significant drop in accuracy compared to our final model (57.8% ORR), and thus was not pursued as the primary solution.

Discussion

Our LLM-enhanced gating framework delivers a 9.5% FDE reduction relative to the best single expert and realises 57.8% of the theoretical oracle gain. We now interpret these findings, relate them to the research gaps highlighted in the Related Work section, and discuss deployment trade-offs.

Meta-Feature Signals Are Essential

The ablations in Table 2 confirm that purely geometric indicators are insufficient: both thresholding and geometric MLP gates achieve $\leq 1.7\%$ ORR, echoing Gap 1’s weak feature–error correlation. Performance improves dramatically once we incorporate meta-features that expose each expert’s internal state (MC-dropout variance, input stability, physics violations). These signals provide a direct proxy for model competence, allowing the gate to leverage experts’ self-awareness rather than relying on indirect scene difficulty cues.

Ranking Beats Classification and Regression

Our study of learning formulations highlights a clear hierarchy. Regression on absolute or differential FDE is ill-posed due to heavy-tailed error distributions, yielding negative R^2 . Multiclass classification alleviates this but remains brittle under the class imbalance noted in Gap 2. The ranking-gate, by contrast, is scale-invariant and needs only to decide which expert is better, not by how much. This yields the strongest numerical gate (27.0% ORR) and addresses both calibration and robustness concerns raised in Gaps 2–3.

Semantic Reasoning Complements Numerical Gates

Even with meta-features and ranking, 34% of scenes trigger low confidence or high-risk semantics. The LLM supervisor resolves these outstanding gaps by providing contextual reasoning beyond numerical signals—disambiguating intent in yields, occlusions, or cut-ins. Its contribution lifts ORR from 27.0% to 57.8%, underscoring that semantic understanding

is indispensable for closing Gap 4 and approaching oracle behaviour.

Deployment Considerations and Trade-offs

While the full LLM-enhanced gate offers the highest accuracy, it also introduces additional latency. Our specialised variants delineate a spectrum of operating points: the Q90 risk-aware gate prioritises tail safety, cutting P95 FDE by 3.5%, whereas the two-stage gate retains much of the ranking performance (18.5% ORR) while saving $\sim 10\%$ of GameFormer compute. Practitioners can therefore tailor the framework to safety-critical or resource-constrained deployments, balancing accuracy, risk, and efficiency.

Limitations

Our framework delivers a 9.5% FDE reduction and realises 57.8% of the oracle gap, yet several limitations remain.

Oracle Ceiling and Expert Diversity. The reported ORR is bounded by the oracle defined over our current three experts. GameFormer (2.835 m FDE) far outperforms the long-tail transformer (7.07 m) and the physics-driven LSTM (8.12 m), so even perfect gating is capped by the modest quality of the non-SOTA experts. Broader expert diversity—including diffusion-based (Jiang et al. 2023) or raster-based predictors (Liang et al. 2020)—is the most direct lever for raising the theoretical ceiling.

Meta-Feature Fidelity and Cost. Meta-features are indispensable but expensive. Extracting $K=8$ MC-dropout passes per expert introduces noticeable latency, and our two-stage ablation only trimmed GameFormer compute by $\sim 10\%$. Signal fidelity can also be uneven: for example, the LSTM’s dropout configuration produced near-zero variance, creating blind spots where the gate cannot gauge that expert’s confidence.

Information Gaps and the Unrealised Oracle. Even with meta-features and semantic triggers, 42.2% of the oracle gap remains. The 36-dimensional feature vector still fails to capture all failure modes; some scenes hinge on cues beyond uncertainty, stability, or the LLM’s semantic classes. Closing this residual gap will require richer behavioural signals or accepting that certain selection errors are fundamentally stochastic.

LLM Supervisor Practicality and Latency. The LLM supervisor lifts ORR from 27.0% to 57.8%, yet real-time deployment is challenging. Our experiments rely on offline batching and cached triggers, whereas online inference can incur second-scale latency. A deployable system will need to distil the LLM’s reasoning into a lightweight surrogate or substitute it with lower-latency rule-based semantics.

Dataset Scope and Generalisation. All experiments target the 1,287-scene nuPlan-mini dataset. Although diverse, it does not guarantee transfer. Validating the meta-feature correlations and LLM triggers on the full nuPlan benchmark (Caesar and et al. 2021) and across other datasets such as Waymo (Ettinger and et al. 2021) remains future work.

Ethical Statement

Trajectory prediction for autonomous driving is a safety-critical endeavour with substantial societal impact. Our work seeks to improve reliability in long-tail scenarios, but several risks accompany the proposed framework.

- **Gating failure.** Introducing a supervisory gate creates an additional failure mode: a misclassification that suppresses the conservative expert when it is needed most could produce outcomes more severe than those of any single predictor. Overconfidence in gate decisions must therefore be monitored carefully.
- **Dataset bias.** Training and evaluation on nuPlan-mini inherit the geographic and sensor biases of that dataset. As a result, the behaviour of both the experts and the gate may not transfer to regions or weather conditions absent from the data.
- **LLM supervision risk.** The LLM supervisor adds interpretability but can also hallucinate or encode biases when faced with out-of-distribution traffic scenarios. Using an LLM in the decision loop demands rigorous validation and fallback mechanisms.
- **Data privacy.** This study uses the publicly available nuPlan dataset, which contains no personally identifiable information. Therefore, no additional ethical or data privacy considerations are required.

We aim to mitigate these risks by exposing interpretable reasoning, documenting failure cases, and releasing code and evaluation artefacts where permitted to support transparent scrutiny and responsible deployment.

Acknowledgements

We sincerely thank my supervisor, Dr. Loo Junn Yong, for his invaluable guidance, insightful feedback, and steadfast support throughout this research. I also wish to acknowledge Monash University for providing an excellent academic environment and support. Finally, I am grateful to the creators of the nuPlan dataset for making their data publicly available, which was essential to the completion of this work.

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*, 961–970.
- Angelopoulos, A. N.; and Bates, S. 2021. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv preprint arXiv:2107.07511*.
- Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight Uncertainty in Neural Networks. *arXiv preprint arXiv:1505.05424*.
- Burges, C. J. C. 2010. From RankNet to LambdaRank to LambdaMART: An Overview. Technical Report MSR-TR-2010-82, Microsoft Research.

- Burges, C. J. C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; and Hullender, G. 2005. Learning to Rank Using Gradient Descent. In *Proceedings of the 22nd International Conference on Machine Learning*, 89–96.
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621–11631.
- Caesar, H.; and et al. 2021. nuPlan: A Closed-Loop ML-Based Planning Benchmark for Autonomous Vehicles. *arXiv preprint arXiv:2106.11810*.
- Cao, Z.; Qin, T.; Liu, T.-Y.; Tsai, M.-F.; and Li, H. 2007. Learning to Rank: From Pairwise Approach to Listwise Approach. In *Proceedings of the 24th International Conference on Machine Learning*, 129–136.
- Chandra, R.; Bhattacharyya, A.; Bhattacharyya, A.; Bera, P.; and Shah, M. 2019. TraPHic: Trajectory Prediction in Dense and Heterogeneous Traffic Using Weighted Interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8483–8492.
- Deo, N.; and Trivedi, M. M. 2018. Convolutional Social Pooling for Vehicle Trajectory Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 1468–1476.
- Eigen, D.; and Fergus, R. 2015. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, 2650–2658.
- Ettinger, S.; and et al. 2021. Large Scale Interactive Motion Forecasting for Autonomous Driving: The Waymo Open Motion Dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9710–9719.
- Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv preprint arXiv:2101.03961*.
- Feng, D.; Rosenbaum, L.; Hecker, S.; and Dietmayer, K. 2018. Towards Safe Autonomous Driving: Capture Uncertainty in the Deep Neural Network for Lidar 3D Vehicle Detection. In *Proceedings of the IEEE Intelligent Transportation Systems Conference*, 3266–3273.
- Fu, Y.; Pan, X.; Ye, Y.; Mu, Y.; Guan, Y.; Xu, R.; Ye, Z.; Qi, Y.; and Li, Z. 2024. DriveGPT-4: Large Language Model for Autonomous Driving. *arXiv preprint arXiv:2401.00001*.
- Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv preprint arXiv:1506.02142*.
- Gu, J.; and et al. 2023. Stochastic Trajectory Prediction via Motion Indeterminacy Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13719–13728.
- Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. *arXiv preprint arXiv:1706.04599*.
- Gustafsson, F.; Danelljan, M.; and Schon, T. B. 2020. Evaluating Scalable Bayesian Deep Learning Methods for Robust Computer Vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 166–175.
- Ivanovic, B.; and Pavone, M. 2019. The Trajectron: Probabilistic Multi-Agent Trajectory Modeling with Dynamic Spatiotemporal Graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2375–2384.
- Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1): 79–87.
- Jiang, Y.; Ding, M.; Meng, L.; Chen, L.; Wei, F.; Xu, J.; Bai, Y.; Song, S.; Huang, G.; Wang, Y.; Ding, Y.; and Xu, C. 2023. MotionDiffuser: Controllable Multi-Agent Motion Prediction via Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17479–17488.
- Kalman, R. E. 1960. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1): 35–45.
- Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. *arXiv preprint arXiv:1612.01474*.
- Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; Cherry, C.; Hechtman, B.; Chen, Z.; Roberts, A.; Li, H. W.; Duke, T.; Bosma, M.; Chen, J.; Child, R.; Greenside, P.; Bosma, L.; Molina, A.; Isard, M.; Steiner, B.; Tran, D.; and Dean, J. 2021. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. *arXiv preprint arXiv:2006.16668*.
- Liang, M.; Yang, B.; Zeng, R.; Chen, Y.; Hu, R.; Li, S.; and Urtasun, R. 2020. Learning Lane Graph Representations for Motion Forecasting. In *Proceedings of the European Conference on Computer Vision*, 541–556.
- Liu, C.; Zoph, B.; Neumann, M.; Shlens, J.; Hua, W.; Li, L.-J.; Fei-Fei, L.; Yuille, A.; Huang, J.; and Murphy, K. 2018. Progressive Neural Architecture Search. In *Proceedings of the European Conference on Computer Vision*, 19–35.
- Mao, J.; Zhao, H.; Packer, C.; Li, Y.; Shen, Y.; Sun, C.; and Ramanan, D. 2023a. Language Conditioned Traffic Generation. *arXiv preprint arXiv:2307.07947*.
- Mao, J.; Zhao, H.; Packer, C.; Li, Y.; Shen, Y.; Sun, C.; and Ramanan, D. 2023b. LLM-Driver: A Large Language Model for Explainable Autonomous Driving. *arXiv preprint arXiv:2310.00001*.
- Nayakanti, A.; Sapp, B.; Tsai, C.-H.; Casas, S.; Yang, Y.; and Urtasun, R. 2023. Wayformer: Motion Forecasting via Simple and Efficient Attention Networks. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 3992–3999.
- Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J. V.; Lakshminarayanan, B.; and Snoek, J. 2019. Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *arXiv preprint arXiv:1906.02530*.

Rhinehart, N.; McAllister, R.; Kitani, K.; and Levine, S. 2019. PRECOG: Prediction Conditioned on Goals in Visual Multi-Agent Settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2821–2830.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Pinto, A. S.; Keyzers, D.; and Houlsby, N. 2021. Scaling Vision with Sparse Mixture of Experts. *arXiv preprint arXiv:2106.05974*.

Ross, C. Y.; Schaarschmidt, M.; Cui, Y.; Asfour, T.; and Matsubara, T. 2021. Uncertainty-Aware Contact-Safe Model-Based Reinforcement Learning. *IEEE Robotics and Automation Letters*, 6(2): 3918–3925.

Salzmann, T.; Ivanovic, B.; Chakravarty, P.; and Pavone, M. 2020. Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data. In *Proceedings of the European Conference on Computer Vision*, 549–565.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q. V.; Hinton, G.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *arXiv preprint arXiv:1701.06538*.

Shi, S.; Jiang, L.; Dai, D.; and Schiele, B. 2022. Motion Transformer with Global Intention Localization and Local Movement Refinement. *arXiv preprint arXiv:2209.13508*.

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Luo, P.; Geiger, A.; and Li, H. 2023. DriveLM: Driving with Graph Visual Question Answering. *arXiv preprint arXiv:2312.14150*.

Standley, T.; Zamir, A. R.; Chen, D.; Guibas, L.; Malik, J.; and Savarese, S. 2020. Which Tasks Should Be Learned Together in Multi-Task Learning? In *Proceedings of the 37th International Conference on Machine Learning*, 9120–9132.

Sun, T.; Huang, W.; Sun, W.; Chen, L.; Geiger, A.; and Li, H. 2023. GameFormer: Game-Theoretic Modeling and Learning of Transformer-Based Interactive Prediction and Planning for Autonomous Driving. *arXiv preprint arXiv:2303.05760*.

Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Wong, K.-Y. K.; Li, Z.; and Zhao, H. 2023. DriveGPT4: Interpretable End-to-End Autonomous Driving via Large Language Model. *arXiv preprint arXiv:2310.01957*.

Yuan, Y.; Weng, X.; Ou, Y.; and Kitani, K. 2021. AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9813–9823.

Zhan, J.-T.; Feng, Z.; Du, J.; Mao, Y.; Liu, J.; Tan, Z.; Zhang, Y.; Ye, X.; and Wang, J. 2024. Rethinking the Open-Loop Evaluation of End-to-End Autonomous Driving in nuScenes. *arXiv preprint arXiv:2305.10430*.

Zhou, Z.; Ye, L.; Wang, J.; Wu, K.; and Lu, K. 2022. HiVT: Hierarchical Vector Transformer for Multi-Agent Motion Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8823–8833.

Appendices

This appendix provides additional technical details, comprehensive experimental comparisons, and supplementary materials to support the reproducibility and verification of the research presented in this paper.

Appendix A: Implementation Details and Hyperparameters

To ensure reproducibility, we provide the key hyperparameters and configuration settings used for our meta-feature extraction and hybrid gating mechanism. All experiments were conducted on the 1,287-sample *nuPlan-mini* validation set.

Table 3: Key hyperparameters and environment details for reproducibility. Values reflect actual implementation in code.

Component	Hyperparameter	Value
<i>Meta-Feature Extraction (Sec.)</i>		
MC Dropout	Stochastic Passes (K)	8
Physics Violation	Max Lateral Acceleration	$> 5.0 \text{ m/s}^2$
Physics Violation	Max Curvature	> 0.5
Input Stability	Noise Scale	0.1
Input Stability	Perturbation Samples	3
<i>Ranking-Gate Training</i>		
Model Type	MLP	3-layer Ranking Network
Loss Function		Pairwise Ranking Loss (RankNet)
Optimizer		Adam
Learning Rate		1×10^{-3}
Batch Size		64
Epochs		50 (best validation accuracy)
<i>LLM Supervisor Trigger</i>		
Gate Confidence	Max Softmax Threshold	< 0.3
Semantic Trigger	Scene Type Flags	<code>cut_in, high_speed</code>
Risk Check	(Prompt-level) Max Accel.	$> 8.0 \text{ m/s}^2$
Risk Check	(Prompt-level) Max Curv.	> 0.5
Activation Rate	Resulting (Validation)	34.0%
<i>Reproducibility Environment</i>		
Frameworks		PyTorch 1.9.0, nuPlan-devkit 1.2
Hardware (GPU)		NVIDIA RTX 4060 (8GB)
Hardware (CPU)		AMD Ryzen 7 7745HX
Training Time	Ranking Gate	≈ 90 minutes (Full 50 epochs)
Random Seed		42

Appendix B: Complete Gating Strategy Evaluation

To validate our final methodology, we systematically evaluated 10 distinct gating strategies on a two-expert (LSTM-KF vs. GameFormer) pool. Table 4 summarises the full results, derived from our research log. This comprehensive comparison justifies our choice of the pairwise ranking-gate formulation as presented in the main text.

The analysis confirms that meta-features are essential (Methods 5–9) and that a Ranking-based formulation (Method 8) provides the best balance of mean FDE performance and oracle realisation. Methods 1–3, which rely solely on geometric features, fail to achieve meaningful oracle realization, validating our emphasis on model-internal signals.

Appendix C: Supplementary Materials and Reproducibility

Per thesis requirements, this section provides access to supplementary materials and original supporting documents

that serve as direct evidence for the work completed in the research paper. These artifacts are intended for reference by the examiners and to support full reproducibility of our results.

Code Repository and Implementation. The complete source code for the experimental framework described in this paper is publicly available. This repository includes the implementation of the meta-feature extraction pipeline, the pairwise ranking-gate, the LLM supervisor module, and all evaluation and ablation scripts.

- **GitHub Repository:**

<https://github.com/lbw1850151881-lang/trajectory-ranking-gate>

Experimental Records and Data. The complete experimental logs, raw results files, evaluation outputs, and research records (including those referenced in the paper) are archived as supporting documentation for the “Substance of Research”. This evidence validates the experimental protocols and the results presented in Section .

- **Google Drive Archive:**

<https://drive.google.com/drive/folders/1QcIQSsxMhPQiKipxWi20QsvVUMaOR-cV?usp=sharing>

Project Demonstration Page. A public-facing project page provides a high-level overview of the research, visualizations, and demonstrations.

- **Project Page:**

<https://lbw1850151881-lang.github.io/trajectory-ranking-gate/>

Table 4: Comprehensive comparison of all evaluated gating strategies (Two-Expert Pool: LSTM-KF vs. GameFormer). We report Final Displacement Error (FDE), Oracle Realisation Rate (ORR), 95th Percentile Tail Risk (P95 FDE), and the computational overhead (fraction of GameFormer (GMF) inferences required). Ranking-Gate (Method 8) achieves the best overall FDE and ORR, while the Risk-Gate (Method 7) is optimal for tail-risk (P95) mitigation.

ID	Gating Strategy	Primary Features	FDE (m)↓	ORR↑	P95 FDE (m)↓	GMF Cost↓	Key Insight
<i>Baselines & Theoretical Bound</i>							
–	LSTM-KF (Baseline)	N/A	7.7156	–	24.81	0%	Physics-based baseline
–	GameFormer (Baseline)	N/A	4.4548	0.0%	16.47	100%	SOTA baseline
–	Oracle (Upper Bound)	N/A	3.6026	100.0%	–	–	Theoretical best
<i>Gating Experiments (Methods 1–10)</i>							
1	Thresholding	Geometric	4.4406	1.7%	–	<1%	Geometric features insufficient
2	MLP Classification	Geometric	~4.45	~0%	–	<3%	Failed, heavily biased to GMF
3	MLP (Improved)	Geometric	>4.45	<0%	–	100%	Failed, high GMF misclassification
4	Regression (Predict Δ FDE)	Meta-features	4.4103	-5.2%	–	100%	Failed ($R^2 = -0.005$)
5	Meta-Gate (Classification)	Meta-features	4.2953	18.7%	–	100%	Meta-features are effective
6	Quantile Regression (Q75)	Meta-features	4.3576	~14%	–	100%	Systematically overestimates risk
7	Risk-Based Gate (Q90)	Meta-features	4.4154	~15%	15.89	100%	Best for Tail Risk (P95)
8	Ranking-Gate (Ours)	Meta-features	4.2243	27.0%	–	100%	Best Overall FDE/ORR
9	Two-Stage Gate	Meta-features	4.2974	18.5%	–	90.1%	Best compute/performance trade-off