Trajectory Prediction Meets Large Language Models: A Survey

Yi Xu Ruining Yang Yitian Zhang Yizhou Wang
Jianglin Lu Mingyuan Zhang Lili Su Yun Fu
Department of Electrical and Computer Engineering, Northeastern University
{xu.yi,yang.ruini,lu.jiang,zhang.mingyua,l.su}@northeastern.edu
{markcheung9248,wyzjack990122}@gmail.com
yunfu@ece.neu.edu

Abstract

Recent advances in large language models (LLMs) have sparked growing interest in integrating language-driven techniques into trajectory prediction. By leveraging their semantic and reasoning capabilities, LLMs are reshaping how autonomous systems perceive, model, and predict trajectories. This survey provides a comprehensive overview of this emerging field, categorizing recent work into five directions: (1) Trajectory prediction via language modeling paradigms, (2) Direct trajectory prediction with pretrained language models, (3) Language-guided scene understanding for trajectory prediction, (4) Language-driven data generation for trajectory prediction, (5) Language-based reasoning and interpretability for trajectory prediction. For each, we analyze representative methods, highlight core design choices, and identify open challenges. This survey bridges natural language processing and trajectory prediction, offering a unified perspective on how language can enrich trajectory prediction.

https://github.com/colorfulfuture/Awesome-Trajectory-Motion-Prediction-Papers

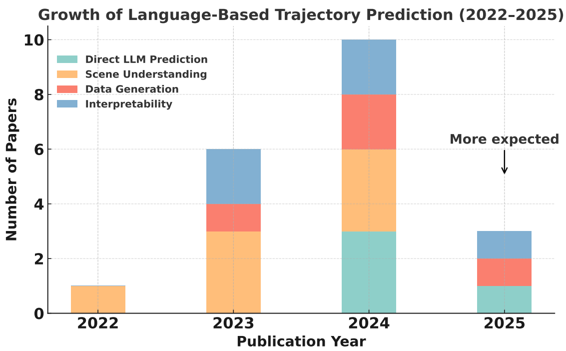

Figure 1: Number of published papers from 2022 to 2025 across four categories of language-based trajectory prediction. We exclude works that are merely inspired by language modeling architectures (e.g., Transformer-based baselines without language inputs) to focus on methods that explicitly incorporate natural language or pretrained language models.

1 Introduction

Trajectory prediction is a critical task in autonomous systems, with applications in autonomous driving (Huang et al., 2022; Madjid et al., 2025), robot navigation (Rösmann et al., 2017; Bhaskara et al., 2024), interaction modeling (Huang et al., 2019; Zhu et al., 2021), and multi-agent coordination (Zhao et al., 2019; Xu and Fu, 2025). Traditional approaches rely on numerical models grounded in geometry, physics, or patterns extracted from historical data. A core challenge lies in accurately modeling the complex interactions among agents. While modern data-driven methods aim to learn these interactions implicitly from large-scale datasets, their performance remains highly sensitive to data quality and diversity, limiting generalization in diverse or unseen scenarios.

Recent advances in large language models (LLMs) (Zhao et al., 2023; Chang et al., 2024) and multimodal reasoning (Wei et al., 2022; Plaat et al., 2024) have opened new avenues for bridging high-level semantic understanding with low-level trajectory prediction. Language is inherently expressive, abstract, and compositional, making it a natural medium for conveying context, intent, and interaction in dynamic environments. Through language, one can describe scenes, articulate goals, reason about causality, and even speculate about alternative futures. These capabilities have motivated a growing interest in incorporating language understanding, particularly via LLMs, into trajectory prediction pipelines. Existing works vary in how they incorporate language: some use LLMs for scene and context understanding, others reformulate trajectory prediction as a language modeling task, and some leverage language for data generation, control, and interpretability.

In this survey, we present a comprehensive overview of recent advances in leveraging LLMs for trajectory prediction. Although still in its early stages, this interdisciplinary field is rapidly expanding, as evidenced by the sharp increase in related publications over the past three years (Figure 1). This surge highlights both the timeliness and necessity of a dedicated synthesis. By focusing on the emerging integration of LLMs into trajectory prediction, we aim to provide a valuable reference for both academics and practitioners interested in deploying LLMs for trajectory prediction. Our contributions are summarized as follows:

- To the best of our knowledge, this is the first survey on the growing intersection of LLMs and trajectory prediction, bridging perspectives from both NLP and trajectory prediction communities.
- We systematically organize existing work into five categories and analyze each in terms of modeling paradigms, architecture, and the role of language across different prediction stages.
- We discuss open challenges and outline future directions aimed at better aligning language understanding with trajectory prediction objectives.

2 Background

In this section, we briefly introduce the trajectory prediction task and then summarize relevant Large Language Models (LLMs) concepts.

2.1 Trajectory Prediction Formulation

Trajectory prediction aims to forecast the future positions of agents based on their observed status and contextual cues. Formally, given observed states $\mathbf{X} \in \mathbb{R}^{N\times H\times F}$ for N agents over H time steps, the goal is to predict future trajectories $\mathbf{Y} \in \mathbb{R}^{N\times T\times F}$ over a horizon of T steps. Here, F denotes the feature dimension, typically 2 for 2D coordinates. Contextual information $\mathbf{I}$, such as maps or scene layouts, is often included. A common approach learns a probabilistic model $p_\theta(\mathbf{Y} \mid \mathbf{X}, \mathbf{I})$ to account for multiple plausible futures.

However, trajectory prediction remains challenging due to its multimodal nature, the need for long-range temporal reasoning, and the integration of heterogeneous inputs such as scene context. Recent work has begun to explore LLMs as a promising tool to address these challenges.

2.2 Key Concepts in Large Language Models

LLMs are deep neural networks, typically based on the Transformer architecture (Vaswani et al., 2017), trained on internet-scale data, and have demonstrated remarkable capabilities in language understanding, generation, and reasoning. These properties make them appealing for trajectory prediction, where semantic abstraction, intent reasoning, and multi-agent interaction modeling are essential. In particular, the autoregressive nature of LLMs allows them to model sequences of spatial-temporal states, while their pretraining enables zero-shot or few-shot generalization (Wei et al., 2021; Kojima et al., 2022) in novel scenarios.

To apply LLMs in trajectory tasks, spatial and contextual data are often transformed into token sequences or descriptive prompts aligned with the model's input format. Some approaches directly use pretrained models like GPT (Floridi and Chiriatti, 2020; Achiam et al., 2023), while others adapt them via fine-tuning (Ziegler et al., 2019), instruction tuning (Zhang et al., 2023), or lightweight methods (e.g., LoRA (Hu et al., 2022)). These adaptations allow LLMs to process structured inputs, generate diverse trajectories, and provide rationales, all through the lens of language modeling.

3 Taxonomy

We categorize the literature into five distinct yet interconnected directions, as illustrated in Figure 2. These categories reflect the methodological diversity and highlight the different roles that language plays in enhancing predictive capabilities. The five categories are as follows:

1. **Trajectory Prediction via Language Modeling Paradigms (Section 4):** This category formulates trajectory prediction as a sequence modeling task inspired by language modeling. By representing trajectories as discrete tokens or sequences, these methods adopt autoregressive or masked modeling strategies, analogous to those used in NLP. This paradigm enables models to capture long-range temporal dependencies using architectures such as Transformers.

2. **Direct Trajectory Prediction with Pretrained Language Models (Section 5):** In this category, pretrained LLMs are directly employed, often via prompting or lightweight adaptation, to generate future trajectories. These approaches

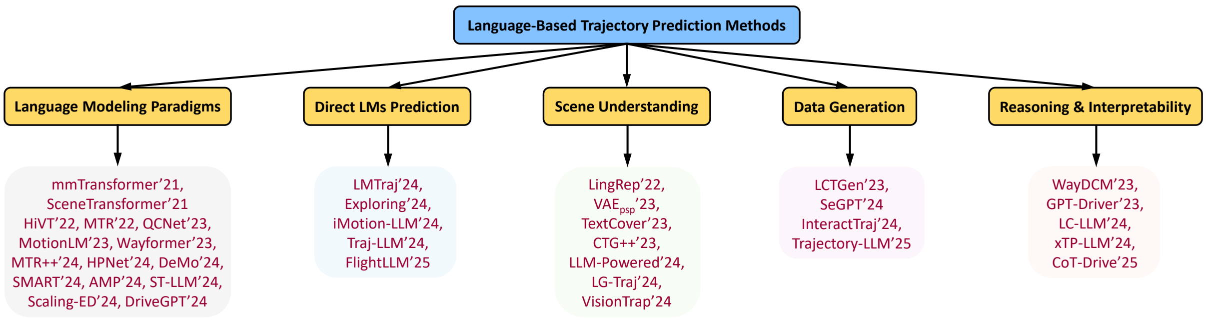

Figure 2: An overview of language-based trajectory prediction methods, categorized into five directions. Representative methods are shown with their names and publication years.

leverage the reasoning and generalization capabilities of LLMs without modifying core model weights, frequently incorporating textual prompts or multimodal inputs. They are particularly effective in zero-shot or low-data regimes.

3. **Language-Guided Scene Understanding for Trajectory Prediction (Section 6):** Language is used here to enrich context modeling. Scene elements, such as traffic rules, map layouts, or agent attributes, are expressed or augmented in natural language, enabling models to reason more effectively about intent, interaction, and semantics. Language serves as a complementary modality to sensor-based perception.

4. **Language-Driven Data Generation for Trajectory Prediction (Section 7):** This category focuses on using language to synthesize diverse trajectory data or driving scenarios. Approaches include simulating agent behaviors from textual descriptions or generating controllable training samples to improve data coverage. These methods enhance scalability and support model development in rare or long-tail conditions.

5. **Language-Based Reasoning and Interpretability for Trajectory Prediction (Section 8):** These approaches aim to enhance transparency by generating natural language explanations alongside predicted trajectories. By enabling models to explain or justify their decisions, this category strengthens the interpretability and trustworthiness of prediction systems in real-world deployments.

4 Trajectory Prediction via Language Modeling Paradigms

The trajectory prediction landscape has been significantly influenced by advances in natural language processing (NLP) (Vaswani et al., 2017; Devlin et al., 2019), particularly through the introduction of Transformer architectures and attention mechanisms. Originally designed for language sequences, these models have proven highly effective in multi-agent motion forecasting due to their ability to capture long-range temporal dependencies, model uncertainty, and represent complex interactions. Motivated by these strengths, recent works have increasingly adapted language modeling paradigms, especially Transformer-based sequence models, for representing and generating agent trajectories. We first summarize key representative methods inspired by language modeling architectures, presented chronologically to illustrate their evolution.

One early contribution is mmTransformer (Liu et al., 2021), which employs a stacked Transformer framework for multimodal motion prediction. It uses a two-stage design to first generate diverse trajectory candidates and then refine them via attention-based decoding. A region-based training strategy encourages coverage over multiple plausible futures, laying foundational work for using attention in multimodal prediction. Another notable work, SceneTransformer (Ngiam et al., 2021), introduces a unified Transformer-based model that jointly predicts multiple agent trajectories, employing a masking strategy inspired by NLP to flexibly condition on goals. It integrates information across road elements, agent interactions, and time steps using attention, achieving strong performance on both marginal and joint prediction benchmarks.

To improve representational efficiency and incorporate geometric invariance, HiVT (Zhou et al., 2022) proposed a hierarchical vectorized Transformer that disentangles local and global interactions through structured inputs and attention. Its hierarchical design helps disentangle local versus global dynamics and sets a strong benchmark on

Argoverse (Chang et al., 2019; Wilson et al., 2023). Around the same time, Wayformer (Nayakanti et al., 2023) presents a modular and scalable attention-based architecture that unifies scene encoding and trajectory decoding through flexible fusion strategies. It explores early, late, and hierarchical fusion of input modalities (e.g., map, agents, signals), and introduces latent query attention to balance efficiency with predictive performance.

Motion Transformer (MTR) (Shi et al., 2022) introduces a two-branch architecture that separates global intention localization from local motion refinement. Instead of generating dense candidate trajectories, MTR employs learnable motion queries, each dedicated to predicting and refining a specific mode. This design stabilizes training and improves coverage over diverse future outcomes, yielding strong performance on the Waymo Open Motion Dataset (WOMD) (Ettinger et al., 2021). Extending this framework, MTR++ (Shi et al., 2024) adapts MTR to multi-agent settings by introducing symmetric scene modeling and mutually guided intention querying. Through structured attention, agents can coordinate predictions to produce scene-consistent trajectories, further enhancing interaction modeling in complex environments.

To better capture agent-centric behavior, QCNet (Zhou et al., 2023) re-centers the prediction process around a single target agent's query. This targeted formulation enables selective context extraction and improves both predictive accuracy and interpretability, especially in dense traffic scenarios. In parallel, HPNet (Tang et al., 2024) enhances temporal modeling by incorporating past predictions into its attention mechanism. Its Historical Prediction Attention module learns dependencies between prior forecasts and future motion, improving stability over extended horizons. Pushing this line of disentangled modeling further, DeMo (Zhang et al., 2024) decomposes multi-agent trajectory prediction into two complementary components: directional mode queries for encoding intent and dynamic state queries for capturing motion evolution. By integrating attention with state-space modeling (e.g., Mamba (Gu and Dao, 2023)), DeMo enables fine-grained control over both the diversity and physical plausibility of predicted futures.

While earlier works primarily drew architectural inspiration from NLP, leveraging Transformers and attention mechanisms to model agent interactions, recent advances have gone further by adopting the modeling paradigms of language directly. In particular, the rise of large language models (LLMs) has motivated a shift toward treating trajectory prediction as an autoregressive token generation problem, analogous to text generation. We next highlight several representative works that embrace this generative modeling strategy, marking a transition from architecture-level borrowing to full paradigm-level alignment with language modeling.

A pioneering effort in this direction is MotionLM (Seff et al., 2023), which formulates multi-agent motion forecasting as a language modeling problem. It discretizes continuous trajectories into motion tokens and uses a temporally causal Transformer decoder to model joint futures across agents. Trained with a causal language modeling loss, MotionLM achieves state-of-the-art results on the WOMD interactive benchmark (Ettinger et al., 2021). Shortly after, SMART (Wu et al., 2024) introduces a GPT-style token-level simulation framework for real-time autonomous driving. By vectorizing the map and trajectories and modeling them as discrete tokens, SMART enables low-latency inference by avoiding repeated re-encoding of past observations, making it highly suitable for deployment in simulation and planning loops. To enhance structure-aware prediction, AMP (Jia et al., 2024) revisits autoregressive prediction with a focus on unified token space and factorized attention. It uses next-token prediction to model motion sequences while incorporating spatial, temporal, and semantic relations via carefully designed attention heads. AMP outperforms prior autoregressive approaches, including MotionLM, particularly on long-horizon and interactive benchmarks. Another recent effort, ST-LLM (Liu et al., 2024), extends this line of work by casting traffic forecasting as a spatial-temporal token generation problem. The model discretizes both spatial and temporal dimensions into tokens and employs a Transformer-based decoder to autoregressively predict future traffic states. By bridging spatial-temporal representation learning with language modeling paradigms, this approach further illustrates the flexibility and generality of LLM-inspired architectures in the prediction task.

While these models demonstrate strong predictive performance, their high computational cost poses challenges for practical deployment. To address this, Scaling-ED (Ettinger et al., 2024) explores the use of model ensembles and distillation to boost accuracy without incurring high compute costs. The authors show that large ensembles of optimized models can be distilled into compact student models, achieving a favorable trade-off between performance and efficiency, a practical solution for real-world deployment. The most recent effort, DriveGPT (Huang et al., 2024), pushes this paradigm further by scaling autoregressive Transformers on billions of driving frames. Treating behavior modeling as sequential token generation, it draws direct inspiration from the scaling laws observed in LLMs, systematically studying how model capacity and dataset size influence performance. The model demonstrates strong gains from large-scale pretraining, bridging the gap between foundation models and trajectory prediction.

Together, these works represent a natural evolution, from architectural borrowing to full integration of generative language modeling paradigms. By leveraging sequence modeling, token-level reasoning, and scaling, this line of work continues to redefine the frontier of trajectory prediction.

5 Direct Trajectory Prediction with Pretrained Language Models

While the previous section focused on trajectory models architecturally inspired by language modeling, those approaches do not directly employ pretrained language models in prediction. In contrast, the recent emergence of LLMs has enabled a new line of research where LLMs themselves serve as the predictive core. These methods aim to leverage LLMs' representational power, sequential reasoning, and generalization capabilities to address trajectory forecasting across diverse domains.

A pioneering effort in this direction is LMTraj (Bae et al., 2024), which reformulates trajectory prediction as a question-answering task. It converts numerical trajectories and scene features into structured natural language prompts and feeds them into pretrained language models, including T5 (Roberts et al., 2019) for supervised training and GPT-3.5/4 (Achiam et al., 2023) for zero-shot and few-shot evaluation. To bridge the modality gap between continuous values and discrete tokens, LMTraj introduces a decimal-aware tokenizer and auxiliary tasks designed to capture interaction cues. The model outperforms traditional regression baselines, demonstrating that language-based reasoning can enhance trajectory prediction. Shortly after, Munir et al. (2024) conducts a systematic evaluation of open-source LLMs, including GPT-2 (Radford et al., 2019), LLaMA-7B (Touvron et al., 2023), LLaMA-7B-Chat, Zephyr-7B (Tunstall et al., 2023), and Mistral-7B (Jiang et al., 2023), on trajectory prediction tasks. The study finds that while these models exhibit strong capabilities in capturing high-level intent and awareness, they struggle with numerical precision and grounding when applied out of domain. These findings underscore the importance of fine-tuning and domain-specific adaptations to fully realize the potential of LLMs in trajectory prediction.

To improve alignment with task-specific commands, iMotion-LLM (Felemban et al., 2024) introduces instruction tuning for trajectory generation. Using LoRA-based (Hu et al., 2022) fine-tuning on the WOMD dataset (Ettinger et al., 2021) augmented with natural language directives (e.g., "turn left at the next intersection"), the model learns to generate trajectories aligned with intent. This work demonstrates the potential of natural language as a flexible interface for intent-conditioned prediction. Extending the use of LLMs beyond driving, FlightLLM (Luo and Zhou, 2025) adapts this paradigm to the aviation domain. By serializing flight paths into token sequences and fine-tuning models such as LLaMA (Touvron et al., 2023), it achieves strong performance in both short- and long-term trajectory forecasting. However, the study also identifies practical concerns, such as inference latency, that may hinder real-time deployment without architectural optimization.

More recently, Traj-LLM (Lan et al., 2024) proposes a fully end-to-end framework that integrates LLMs into the prediction pipeline without relying on handcrafted prompts. It encodes scene and agent context into a format compatible with transformer-based language models, and employs a GPT-2 backbone fine-tuned via LoRA to capture high-level interactions. To improve spatial fidelity and multimodal reasoning, Traj-LLM introduces a lane-aware probabilistic head. Notably, Traj-LLM maintains strong performance even in low-data settings, underscoring LLMs' few-shot capabilities.

Together, these works represent a shift from language modeling as a source of architectural ideas to LLMs as active, trainable predictors in structured forecasting tasks. Whether through instruction tuning, question answering, or domain-specific adaptation, these approaches highlight both the flexibility and the current limitations of LLMs in trajectory prediction, particularly regarding grounding, precision, and deployment efficiency.

6 Language-Guided Scene Understanding for Trajectory Prediction

Traditional trajectory prediction methods primarily rely on structured inputs such as agent coordinates and map features. However, these signals often fall short in capturing scene semantics or intent-level interactions. To address this gap, recent methods explore natural language as a complementary modality, either through textual input, intermediate representations, or language-conditioned modules. By incorporating language-guided reasoning, these approaches aim to enrich perception and deepen understanding of complex environments.

An early example of this direction is LingRep (Kuo et al., 2022), which generates natural language descriptions of agent behaviors, such as "crossing at intersection" or "following another car", and uses them as interpretable intermediate representations for downstream prediction. Notably, this approach relies on manually crafted templates and predefined rules rather than pretrained language models. The resulting linguistic abstractions serve as semantically meaningful priors, helping structure trajectory generation and providing insights into model behavior. Pushing further into the integration of language and perception, VAE$_{psp}$ (Syed and Morris, 2023) investigates the impact of semantic segmentation and textual scene descriptions on pedestrian trajectory prediction. While the textual cues are informative, the approach does not involve language models; instead, text is used to complement visual inputs and improve generalization across diverse environments. In contrast to these earlier methods, TextCover (Keysan et al., 2023) explicitly incorporates pretrained language models, specifically DistilBERT (Sanh et al., 2019), into a Transformer-based prediction pipeline. By fusing textual scene representations with spatial features, the model jointly reasons over context and agent dynamics, showing that linguistic embeddings derived from pretrained models can significantly improve trajectory quality, especially in complex urban scenarios.

Beyond augmenting scene representations, CTG++ (Zhong et al., 2023) introduces a diffusion-based trajectory generator conditioned on high-level natural language inputs. It leverages GPT-4 (Achiam et al., 2023) to translate free-form user descriptions, such as "a group of pedestrians crossing while a car is waiting," into differentiable guidance functions that control the generation of socially plausible multi-agent trajectories. This approach demonstrates the potential of LLMs as a flexible interface for scenario-level control. In parallel, LLM-Powered (Zheng et al., 2024) leverages GPT-4V, a vision-enabled large language model (Achiam et al., 2023), to derive high-level semantic cues from structured scene prompts and embed them into the attention layers of the motion prediction backbone. This LLM-enhanced architecture shows improved focus on social features, especially under dense multi-agent interaction.

Focusing specifically on pedestrian dynamics, LG-Traj (Chib and Singh, 2024) summarizes prior motion and visual context into a text-based scene embedding using an LLM. This embedding is aligned with visual features and used to guide trajectory generation, leading to better intent disambiguation in socially dense scenarios. Most recently, VisionTrap (Moon et al., 2024) leverages vision-language models (VLMs) to generate scene-level annotations, such as "child near crosswalk." These cues serve as soft guidance, helping trajectory predictors attend to salient visual regions and improving both performance and interpretability.

Together, these works illustrate a clear progression from using language as an interpretive layer to fully integrating it as a semantic control signal in trajectory prediction. By leveraging natural language for scene abstraction, relational reasoning, and controllable generation, these approaches demonstrate the growing utility of language in enriching context-aware motion forecasting. As language models improve in precision and alignment, their integration into prediction pipelines may play a central role in enabling more robust, interpretable, and socially aligned autonomous behavior.

7 Language-Driven Data Generation for Trajectory Prediction

As trajectory prediction models become more data-hungry and scenario-specific, recent research has begun to explore using LLMs as generative tools to create synthetic data. These methods leverage natural language as a flexible and high-level interface to describe, simulate, or construct diverse trajectory scenarios, offering a powerful alternative to traditional data collection and manual scene design.

An early exploration in this direction is LCT-Gen (Tan et al., 2023), a generative framework that synthesizes traffic scenarios from natural language prompts using GPT-4 (Achiam et al., 2023) as its interpreter module. Given inputs like "a vehicle is overtaking on the left while a pedestrian crosses," the model uses a Transformer-based decoder to produce realistic vehicle distributions and dynamics grounded in map topology. This work lays the foundational groundwork for scene-level control via high-level linguistic instructions. Expanding on this idea, SeGPT (Li et al., 2024) introduces a prompt-driven framework that uses chain-of-thought (Wei et al., 2022) reasoning with ChatGPT to generate driving scenarios. Free-form textual descriptions are translated into structured simulation configurations. While not producing physically realistic trajectories, it demonstrates the feasibility of adapting pretrained LLMs as controllable scenario generators for autonomous driving.

To support finer-grained control over agent behavior, InteractTraj (Xia et al., 2024) leverages GPT-4 to translate abstract language descriptions into structured motion codes, which are subsequently decoded into realistic multi-agent trajectories. This enables the generation of complex interactions such as yielding, merging, or aggressive acceleration, supporting simulation diversity and interoperability. Most recently, Trajectory-LLM (Yang et al., 2025) employs LLaMA-7B as its core language model to convert compact language descriptions of traffic interactions into multimodal trajectory data. The framework adopts a three-stage generation process, Interaction, Action, and Trajectory, and integrates map layout encoding to produce the L2T dataset with over 240K text-map-trajectory samples. This offers a scalable and semantically rich pathway to trajectory-specific data generation.

Collectively, these works highlight a paradigm shift toward using language not only as a control interface but also as a source of structured supervision for trajectory prediction. By enabling the generation of richly annotated, semantically coherent, and diverse data, language-driven approaches open new avenues for training and evaluating models across domains, scenes, and agent types.

8 Language-Based Reasoning and Interpretability for Trajectory Prediction

As trajectory prediction models grow in complexity, understanding how and why they produce certain outputs has become critical for ensuring safety, trust, and enabling effective human-AI interaction. Recent research has begun integrating LLMs into the trajectory prediction pipeline not only to generate predictions but also to explain and reason about them. These approaches aim to improve transparency through language by exposing high-level goals, inferring intermediate intentions, or articulating step-by-step decision chains using chain-of-thought (CoT) prompting (Wei et al., 2022).

An early step in this direction is Way-DCM (Ghoul et al., 2023), which predicts future trajectories via intermediate waypoint goals. These interpretable sub-goals provide insight into long-horizon planning intent, enabling downstream modules and human observers to better understand the rationale behind an agent's movement along a given path. Shortly after, GPT-Driver (Mao et al., 2023) treats trajectory prediction as a language modeling task. Given structured traffic scenes, the model uses GPT-style autoregressive decoding to produce both future maneuvers and natural language rationales. This unified output structure enables both action generation and transparent explanation, demonstrating that pretrained LLMs can serve as interpretable motion planners when prompted appropriately.

Building on intention-level explanation, LC-LLM (Peng et al., 2024) introduces a dual-task model that forecasts both lane-change decisions and their textual justifications. The model is built on GPT-2, which is fine-tuned to generate explanatory narratives such as "the vehicle merges left due to a slower vehicle ahead," using chain-of-thought prompting. This setup enhances user trust by offering interpretable outputs alongside predictions. Taking a broader perspective, Guo et al. (2024) proposes xTP-LLM, a traffic flow prediction framework based on LLaMA2-7B-Chat, fine-tuned via LoRA. The model converts multi-modal traffic inputs (e.g., weather, PoIs, past traffic flow) into structured textual prompts, enabling the LLM to forecast future traffic patterns while generating natural language explanations. More recently, CoT-Drive (Liao et al., 2025) explicitly frames trajectory prediction as a reasoning task. Instead of directly predicting trajectories, it uses LLMs to verbalize intermediate decision steps, e.g., "the car is approaching a stop sign, slowing down is expected," before generating the final output. This technique yields both more accurate and more interpretable forecasts, simulating human-like decision chains.

Together, these works demonstrate the growing role of language-based reasoning and interpretability in trajectory prediction. Through waypoint prediction, autoregressive explanation, high-level causal modeling, and CoT-driven decision chains, LLMs are emerging as powerful tools for making model behavior more transparent and understandable, laying the groundwork for safer and more trustworthy autonomous systems.

9 Challenges and Future Directions

Despite recent progress, LLM-based trajectory prediction still faces fundamental challenges that represent opportunities for future research. We discuss key challenges and future directions as follows.

Tokenization and Representation. A foundational challenge lies in converting continuous, multi-agent trajectories into discrete sequences suitable for LLM processing. Naive numeric tokenization often results in bloated or lossy representations. Recent work like (Philion et al., 2023) discretizes vehicle motion into compact token vocabularies while preserving centimeter-level accuracy, enabling realistic multi-agent simulation via next-token prediction. However, these token schemes are often hand-crafted and may not generalize across tasks or agent types. Future efforts may explore learned tokenizers, hierarchical representations, or semantic abstractions (e.g., maneuver-level tokens) to better bridge the gap between spatial-temporal data and language modeling.

Prompt Design and Alignment. Effectively guiding LLMs to perform trajectory-related tasks requires carefully crafted prompts that encode scene semantics, agent states, and prediction objectives. Despite early progress, prompt alignment remains challenging and often task-specific. Scalable approaches such as soft prompts, instruction tuning, or reinforcement learning from feedback may enable LLMs to generalize better across diverse driving contexts and output formats.

Reasoning and Prediction Coherence. LLMs offer a unique opportunity to embed commonsense and causal reasoning into trajectory prediction, yet their reasoning quality in spatial and temporal domains is still limited. While Peng et al. (2024) demonstrates the benefit of integrating Chain-of-Thought reasoning to explain and improve long-term forecasts, many LLMs struggle with consistency or generate physically implausible outcomes (Munir et al., 2024). Enhancing reasoning may require integrating physics-informed priors, using structured scene graphs, or coupling LLMs with verifiable planning modules. Developing reasoning-aligned datasets could further encourage robust decision-making.

Multimodal Context Integration. Trajectory forecasting in real-world environments demands multimodal understanding, including visual perception, map topology, and language instructions. Several frameworks tackle this by projecting scene features into token-compatible embeddings (Felemban et al., 2024) or encoding environmental elements as text (Chen et al., 2024). However, tightly coupling perception and prediction remains difficult due to modality mismatches and context window limitations. Emerging architectures that unify vision and language modeling (e.g., vision-language transformers) or graph-enhanced scene prompts offer a promising direction for fusing heterogeneous context into a coherent predictive policy.

Transparency and Interpretability. One of the promising byproducts of using LLMs in trajectory prediction is the ability to generate natural language rationales for predicted behaviors. LC-LLM (Peng et al., 2024) and xTP-LLM (Guo et al., 2024) provide explanations alongside trajectory outputs, enhancing human interpretability and trust. Still, ensuring explanation fidelity remains a challenge. Future models could embed internal alignments, enable counterfactual Q & A, or leverage multi-task training where explanation quality directly influences loss. This line of work will be necessary for deploying predictors in safety-critical applications.

10 Conclusion and Discussion

In this survey, we reviewed the growing research at the intersection of language modeling and trajectory prediction. By organizing recent work into five categories, we highlight how LLMs and natural language representations can be used to enhance prediction performance, enable data generation, improve interpretability, and support multimodal reasoning. This trend indicates a shift from traditional numeric or vision-only models toward more semantically enriched and cognitively aligned systems. While the language integration offers promise in generalization, transparency, and user interaction, it also introduces new challenges in grounding and evaluation. We hope this survey provides a useful foundation for advancing research at the intersection of language and trajectory prediction.

AI Disclosure. We used ChatGPT solely for grammar correction and language polishing. All research content, literature analysis, and writing were conducted independently by the authors.

Potential Risks. We note that emphasizing LLM-based approaches could amplify existing data or representation biases. We encourage readers to interpret the reviewed works critically and consider their broader implications before deploying them in real-world settings.

Limitations

Given the rapid development of natural language processing and trajectory prediction, it is possible that some recent approaches may not have been included in this survey. Nonetheless, we have made every effort to ensure that the most representative and impactful works are covered. In Section 4, we do not attempt to exhaustively review all Transformer-based trajectory prediction models, as Transformer architectures and attention mechanisms have become foundational components in the field. Instead, we selectively highlight well-recognized, open-source, and influential works that best illustrate the evolution of language modeling paradigms in trajectory prediction. Finally, some references could reasonably belong to multiple taxonomic categories. In such cases, we categorize each work according to what we consider to be its primary contribution, while acknowledging that many methods operate at the intersection of several themes.

Ethics Statement

We confirm that this survey adheres to ethical research standards. All analyzed works are publicly available, and we do not involve any human subjects or personally identifiable data. The purpose of this survey is to facilitate academic understanding and responsible development of language-based trajectory prediction. We have acknowledged prior work accurately and its original contributions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Inhwan Bae, Junoh Lee, and Hae-Gon Jeon. 2024. Can language beat numerical regression? language-based multimodal trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 753–766.

Rashmi Bhaskara, Hrishikesh Viswanath, and Aniket Bera. 2024. Trajectory prediction for robot navigation using flow-guided markov neural operator. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 15209–15216. IEEE.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631.

Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. 2021. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*.

Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and 1 others. 2019. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8748–8757.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2024. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 14093–14100.

Pranav Singh Chib and Pravendra Singh. 2024. Lg-traj: Llm guided pedestrian trajectory prediction. *arXiv preprint arXiv:2403.08032*.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, and 1 others. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979.

Nachiket Deo and Mohan M Trivedi. 2018. Convolutional social pooling for vehicle trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1468–1476.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, and 1 others. 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*, pages 9710–9719.

Scott Ettinger, Kratarth Goel, Avikalp Srivastava, and Rami Al-Rfou. 2024. Scaling motion forecasting models with ensemble distillation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 4812–4818.

Abdulwahab Felemban, Eslam Mohamed Bakr, Xiaoqian Shen, Jian Ding, Abduallah Mohamed, and Mohamed Elhoseiny. 2024. imotion-llm: Motion prediction instruction tuning. *arXiv preprint arXiv:2406.06211*.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Amina Ghoul, Itheri Yahiaoui, and Fawzi Nashashibi. 2023. Interpretable long term waypoint-based trajectory prediction model. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, pages 3162–3167.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Xusen Guo, Qiming Zhang, Junyue Jiang, Mingxing Peng, Meixin Zhu, and Hao Frank Yang. 2024. Towards explainable traffic flow prediction with large language models. *Communications in Transportation Research*, 4:100150.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Xin Huang, Eric M Wolff, Paul Vernaza, Tung Phan-Minh, Hongge Chen, David S Hayden, Mark Edmonds, Brian Pierce, Xinxin Chen, Pratik Elias Jacob, and 1 others. 2024. Drivegpt: Scaling autoregressive behavior models for driving. *arXiv preprint arXiv:2412.14415*.

Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. 2018. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Rattern Recognition Workshops*, pages 954–960.

Yanjun Huang, Jiatong Du, Ziru Yang, Zewei Zhou, Lin Zhang, and Hong Chen. 2022. A survey on trajectory-prediction methods for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 7(3):652–674.

Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. 2019. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6272–6281.

Xiaosong Jia, Shaoshuai Shi, Zijun Chen, Li Jiang, Wenlong Liao, Tao He, and Junchi Yan. 2024. Amp: Autoregressive motion prediction revisited with next token prediction for autonomous driving. *arXiv preprint arXiv:2403.13331*.

Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, and 1 others. 2023. Mistral 7b. arxiv. *arXiv preprint arXiv:2310.06825*, 10.

Ali Keysan, Andreas Look, Eitan Kosman, Gonca Gürsun, Jörg Wagner, Yu Yao, and Barbara Rakitsch. 2023. Can you text what is happening? integrating pre-trained language encoders into trajectory prediction models for autonomous driving. *arXiv preprint arXiv:2309.05282*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. 2018. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *Proceedings of the International Conference on Intelligent Transportation Systems*, pages 2118–2125.

Yen-Ling Kuo, Xin Huang, Andrei Barbu, Stephen G McGill, Boris Katz, John J Leonard, and Guy Rosman. 2022. Trajectory prediction with linguistic representations. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2868–2875.

Zhengxing Lan, Lingshan Liu, Bo Fan, Yisheng Lv, Yilong Ren, and Zhiyong Cui. 2024. Traj-llm: A new exploration for empowering trajectory prediction with pre-trained large language models. *IEEE Transactions on Intelligent Vehicles*.

Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. 2007. Crowds by example. In *Computer Graphics Forum*, pages 655–664.

Xuan Li, Enlu Liu, Tianyu Shen, Jun Huang, and Fei-Yue Wang. 2024. Chatgpt-based scenario engineer: A new framework on scenario generation for trajectory prediction. *IEEE Transactions on Intelligent Vehicles*, 9(3):4422–4431.

Haicheng Liao, Hanlin Kong, Bonan Wang, Chengyue Wang, Wang Ye, Zhengbing He, Chengzhong Xu, and Zhenning Li. 2025. Cot-drive: Efficient motion forecasting for autonomous driving with llms and chain-of-thought prompting. *IEEE Transactions on Artificial Intelligence*.

Haicheng Liao, Zhenning Li, Huanming Shen, Wenxuan Zeng, Dongping Liao, Guofa Li, and Chengzhong Xu. 2024. Bat: Behavior-aware human-like trajectory prediction for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10332–10340.

Chenxi Liu, Sun Yang, Qianxiong Xu, Zhishuai Li, Cheng Long, Ziyue Li, and Rui Zhao. 2024. Spatial-temporal large language model for traffic prediction. In *Proceedings of the IEEE International Conference on Mobile Data Management*, pages 31–40.

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. 2023. Largest: A benchmark dataset for large-scale traffic forecasting. In *Advances in Neural Information Processing Systems*, pages 75354–75371.

Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. 2021. Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7577–7586.

Kaiwei Luo and Jiliu Zhou. 2025. Large language models for single-step and multi-step flight trajectory prediction. *arXiv preprint arXiv:2501.17459*.

Nadya Abdel Madjid, Abdulrahman Ahmad, Murad Mebrahtu, Yousef Babaa, Abdelmoamen Nasser, Sumbal Malik, Bilal Hassan, Naoufel Werghi, Jorge Dias, and Majid Khonji. 2025. Trajectory prediction for autonomous driving: Progress, limitations, and future directions. *arXiv preprint arXiv:2503.03262*.

Kumar Manas and Adrian Paschke. 2025. Knowledge integration strategies in autonomous vehicle prediction and planning: A comprehensive survey. *arXiv preprint arXiv:2502.10477*.

Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. 2023. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*.

Seokha Moon, Hyun Woo, Hongbeen Park, Haeji Jung, Reza Mahjourian, Hyung-gun Chi, Hyerin Lim, Sangpil Kim, and Jinkyu Kim. 2024. Visiontrap: Vision-augmented trajectory prediction guided by textual descriptions. In *Proceedings of the European Conference on Computer Vision*, pages 361–379.

Farzeen Munir, Tsvetomila Mihaylova, Shoaib Azam, Tomasz Piotr Kucner, and Ville Kyrki. 2024. Exploring large language models for trajectory prediction: a technical perspective. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 774–778.

Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S Refaat, and Benjamin Sapp. 2023. Wayformer: Motion forecasting via simple & efficient attention networks. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2980–2987.

Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, and 1 others. 2021. Scene transformer: A unified architecture for predicting multiple agent trajectories. *arXiv preprint arXiv:2106.08417*.

Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc J. Van Gool. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 261–268.

Li Peng, Hui Zhang, Yao Wang, and Ming Liu. 2024. Lc-llm: Explainable lane-change intention and trajectory predictions with large language models. *arXiv preprint arXiv:2403.18344*.

Jonah Philion, Xue Bin Peng, and Sanja Fidler. 2023. Trajeglish: Traffic modeling as next-token prediction. *arXiv preprint arXiv:2312.04535*.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J Liu, Sharan Narang, Wei Li, and Yanqi Zhou. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *Google Research*.

Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. 2016. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Proceedings of the European Conference on Computer Vision*, pages 549–565.

Christoph Rösmann, Malte Oeljeklaus, Frank Hoffmann, and Torsten Bertram. 2017. Online trajectory prediction and planning for social robot navigation. In *Proceedings of the IEEE International Conference on Advanced Intelligent Mechatronics*, pages 1255–1260.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. 2023. Motionlm:

Multi-agent motion forecasting as language modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8579–8590.

Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. 2022. Motion transformer with global intention localization and local movement refinement. In *Advances in Neural Information Processing Systems*, pages 6531–6543.

Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. 2024. Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3955–3971.

Arsal Syed and Brendan Tran Morris. 2023. Semantic scene upgrades for trajectory prediction. *Machine Vision and Applications*, 34(2):23.

Shuhan Tan, Boris Ivanovic, Xinshuo Weng, Marco Pavone, and Philipp Kraehenbuehl. 2023. Language conditioned traffic generation. *arXiv preprint arXiv:2307.07947*.

Xiaolong Tang, Meina Kan, Shiguang Shan, Zhilong Ji, Jinfeng Bai, and Xilin Chen. 2024. Hpnet: Dynamic trajectory forecasting with historical prediction attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15261–15270.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, and 1 others. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, and 1 others. 2023. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*.

Wei Wu, Xiaoxin Feng, Ziyan Gao, and Yuheng KAN. 2024. SMART: Scalable multi-agent real-time motion generation via next-token prediction. In *Advances in Neural Information Processing Systems*.

Junkai Xia, Chenxin Xu, Qingyao Xu, Yanfeng Wang, and Siheng Chen. 2024. Language-driven interactive traffic trajectory generation. In *Advances in Neural Information Processing Systems*, volume 37, pages 77831–77859.

Yi Xu and Yun Fu. 2025. Sports-traj: A unified trajectory generation model for multi-agent movement in sports. In *International Conference on Learning Representations*.

Kairui Yang, Zihao Guo, Gengjie Lin, Haotian Dong, Zhao Huang, Yipeng Wu, Die Zuo, Jibin Peng, Ziyuan Zhong, Xin Wang, and 1 others. 2025. Trajectory-llm: A language-based data generator for trajectory prediction in autonomous driving. In *International Conference on Learning Representations*.

Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. 2023. Llm4drive: A survey of large language models for autonomous driving. *arXiv preprint arXiv:2311.01043*.

Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kummerle, Hendrik Konigshof, Christoph Stiller, Arnaud de La Fortelle, and 1 others. 2019. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps. *arXiv preprint arXiv:1910.03088*.

Bozhou Zhang, Nan Song, and Li Zhang. 2024. Demo: Decoupling motion forecasting into directional intentions and dynamic states. In *Advances in Neural Information Processing Systems*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and 1 others. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris Baker, Yibiao Zhao, Yizhou Wang, and Ying Nian Wu. 2019. Multi-agent tensor fusion for contextual trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12126–12134.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Xiaoji Zheng, Lixiu Wu, Zhijie Yan, Yuanrong Tang, Hao Zhao, Chen Zhong, Bokui Chen, and Jiangtao Gong. 2024. Large language models powered context-aware motion prediction in autonomous driving. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 980–985.

Ziyuan Zhong, Davis Rempe, Yuxiao Chen, Boris Ivanovic, Yulong Cao, Danfei Xu, Marco Pavone, and Baishakhi Ray. 2023. Language-guided traffic simulation via scene-level diffusion. In *Proceedings of the Conference on Robot Learning*, pages 144–177.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C Knoll. 2024. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*.

Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. 2023. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17863–17873.

Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. 2022. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8833.

Yanliang Zhu, Dongchun Ren, Yi Xu, Deheng Qian, Mingyu Fan, Xin Li, and Huaxia Xia. 2021. Simultaneous past and current social interaction-aware trajectory prediction for multiple intelligent agents in dynamic scenes. *ACM Transactions on Intelligent Systems and Technology*, 13(1):1–16.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Background on Trajectory Prediction

Trajectory prediction seeks to forecast the future positions $\{\mathbf{p}_{t,i}\}_{t=1}^{T} \subset \mathbb{R}^2$ of N agents (indexed by $i = 1, \ldots, N$), given their past positions $\{\mathbf{p}_{t,i}\}_{t=-\tau+1}^{0}$.

Problem Formulation. We denote the observed history of length τ for agent i as

$$\mathbf{X}_{\text{hist},i} = (\mathbf{p}_{-\tau+1,i}, \ldots, \mathbf{p}_{0,i}),$$

and the target future as

$$\mathbf{Y}_i = (\mathbf{p}_{1,i}, \ldots, \mathbf{p}_{T,i}).$$

Most modern methods learn a conditional distribution

$$p\big(\mathbf{Y}_i \,\big|\, \mathbf{X}_{\text{hist},i}, \mathcal{C}_i\big),$$

where $\mathcal{C}_i$ may include scene context (HD-maps, images), social interaction graphs, or goal priors.

Prediction Outputs. For each test sample i, the model typically generates K hypotheses $\{\hat{\mathbf{p}}_{1:T,i}^{(k)}\}_{k=1}^{K}$.

Metrics. Let $\mathbf{p}_{1:T,i} = (\mathbf{p}_{1,i}, \ldots, \mathbf{p}_{T,i})$ denote the ground-truth, and $\hat{\mathbf{p}}_{1:T,i}^{(k)}$ the k-th predicted trajectory. Then:

- **Average Displacement Error (ADE)** for sample i, hypothesis k:

$$\text{ADE}_i^{(k)} = \frac{1}{T} \sum_{t=1}^{T} \big\| \hat{\mathbf{p}}_{t,i}^{(k)} - \mathbf{p}_{t,i} \big\|_2.$$

- **Final Displacement Error (FDE)** for sample i, hypothesis k:

$$\text{FDE}_i^{(k)} = \big\| \hat{\mathbf{p}}_{T,i}^{(k)} - \mathbf{p}_{T,i} \big\|_2.$$

- **Minimum ADE (minADE)** for sample i:

$$\text{minADE}_i = \min_{k=1,\ldots,K} \text{ADE}_i^{(k)}.$$

- **Minimum FDE (minFDE)** for sample i:

$$\text{minFDE}_i = \min_{k=1,\ldots,K} \text{FDE}_i^{(k)}.$$

- **Miss Rate (MR)** at threshold δ:

$$\text{MR}(\delta) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\Big[\min_{k=1,\ldots,K} \big\| \hat{\mathbf{p}}_{T,i}^{(k)} - \mathbf{p}_{T,i} \big\|_2 > \delta \Big],$$

  where $\mathbf{1}[\cdot]$ is the indicator function.

Benchmarks and Datasets. Commonly used publicly available benchmarks include:

- **ETH/UCY** (Pellegrini et al., 2009; Lerner et al., 2007).
- **SDD** (Robicquet et al., 2016).
- **INTERACTION** (Zhan et al., 2019).
- **nuScenes** (Caesar et al., 2020).
- **Argoverse 1 & 2** (Chang et al., 2019; Wilson et al., 2023).
- **Waymo Open Motion Dataset (WOMO)** (Ettinger et al., 2021).
- **Sports-Traj** (Xu and Fu, 2025).

B Relation with Other Surveys

Several recent surveys have explored the integration of large language models (LLMs) and vision-language models (VLMs) into autonomous driving systems. These efforts primarily focus on perception, planning, and decision-making, highlighting applications such as multimodal scene understanding (Zhou et al., 2024), instruction following (Cui et al., 2024), and end-to-end control stacks (Yang et al., 2023). However, trajectory prediction, a core module in the autonomy pipeline, has received relatively limited attention within this emerging paradigm.

In contrast, our survey provides the first comprehensive review dedicated to LLM-based trajectory prediction, focusing on how language modeling, prompting, data generation, and interpretability mechanisms are reshaping this subfield. We emphasize the use of LLMs beyond high-level planning or user interaction, instead targeting their application in modeling, guiding, or generating fine-grained future trajectories. To our knowledge, this is the first taxonomy to structure these advances explicitly around language as a modeling axis.

We also differentiate our work from traditional trajectory prediction surveys (Huang et al., 2022; Madjid et al., 2025), which extensively cover geometric, learning-based, and multimodal methods, but largely overlook the recent surge of language-driven approaches. Our survey fills this gap by highlighting how natural language is used as both input and supervision to enhance prediction. Finally, one recent survey on knowledge integration in prediction and planning (Manas and Paschke, 2025) summarizes rule-based and symbolic reasoning frameworks, they do not systematically cover the use of pretrained LLMs or generative prompting techniques. Our work complements theirs by showcasing how foundation models introduce new forms of semantic grounding, controllability, and reasoning into the trajectory prediction.

C Comparative Overview of Language-Based Methods

To provide a clearer comparison of existing work, we present a summary covering the four categories that explicitly utilize language models: (1) Direct Trajectory Prediction with Pretrained Language Models, (2) Language-Guided Scene Understanding for Trajectory Prediction , (3) Language-Driven Data Generation for Trajectory Prediction, and (4) Language-Based Reasoning and Interpretability for Trajectory Prediction. These categories are summarized in Table 1, Table 2, Table 3, and Table 4, respectively, highlighting how different methods incorporate LLMs, their prompt design, fine-tuning strategies, and datasets. We do not include a table for the first taxonomy, *Trajectory Prediction via Language Modeling Paradigms*, as the methods in this category do not explicitly employ pretrained language models, but rather draw structural inspiration from language modeling paradigms.

Method	Year	LLM Usage	Prompt Design	Fine-tuning Strategy	Dataset
LMTraj (2024)	2024	GPT-style Seq2Seq LMs (e.g., T5) with a custom numerical tokenizer and QA-style sentence generation	QA format with image caption and social questions	Fully supervised fine-tuning QA objective	ETH/UCY, SDD
Exploring (2024)	2024	GPT-2, LLaMA, Zephyr, LLM predicts future trajectory tokens	GPT-Driver style prompt with structured tokens	PEFT (LoRA, P-Tuning)	nuScenes
iMotion-LLM (2024)	2024	LLM (GPT-like) used to interpret scene embeddings and human instructions for generating trajectories	Natural language instructions + scene embedding tokens	LoRA fine-tuning on LLM projection and decoding layers	InstructWaymo (Waymo-augmented)
Traj-LLM (2024)	2024	Frozen LLM (GPT-2, LLaMA), used for joint scene-agent representation	No handcrafted prompt, implicit in tokenization	Frozen LLM + learnable modules	nuScenes
FlightLLM (2025)	2025	LLaMA 3.1, Mistral, etc., LLM outputs next-step tokens	Structured prompt of flight context	PEFT (LoRA, QLoRA, Adapter)	ADS-B (OpenSky)

Table 1: Methods for direct trajectory prediction using pretrained language models, highlighting how LLMs are integrated into the prediction pipeline.

Method	Year	LLM Usage	Prompt Design	Fine-tuning Strategy	Dataset
LingRep (2022)	2022	LLM (e.g., GPT2) generates intermediate linguistic descriptions, used as latent features for prediction	Generates short descriptions from heuristic templates	LLM is co-trained with the trajectory decoder	Argoverse 1
TextCover (2023)	2023	Uses DistilBERT as text encoder for traffic scene representations	Structured text prompt describing target agent, past positions, and lanes	DistilBERT fine-tuned, joint encoder uses frozen weights	nuScenes
CTG++ (2023)	2023	Uses GPT-4 to generate differentiable loss functions from language queries	Few-shot prompting with helper functions and paired query-loss examples	LLM is frozen, only the diffusion model is trained	nuScenes
LLM-Powered (2024)	2024	Use GPT-4-V (vision-language model) to interpret TC-Maps and output traffic context	Combined image and structured text prompt, includes intention, affordance, and scenario description	LLM frozen, only the motion predictor is trained	WOMD
LG-Traj (2024)	2024	Uses LLM to generate motion cues from observed trajectories (e.g., linear/curved/stationary)	Instruction-style prompt using trajectory coordinates in chat template format	LLM is frozen, only the encoder-decoder model is trained	ETH/UCY, SDD
VisionTrap (2024)	2024	Uses BLIP-2 and LLM to generate and refine scene-level textual descriptions	Vision-language captions auto-generated and refined (e.g., "expected to continue straight")	LLM/VLM frozen, only the trajectory predictor is trained	nuScenes-Text

Table 2: Methods leveraging large language models to enhance scene and context understanding for trajectory prediction.

Method	Year	LLM Usage	Prompt Design	Fine-tuning Strategy	Dataset
LCTGen (2023)	2023	Uses GPT-4 in the Interpreter to convert natural language into structured representations	In-context learning and Chain-of-Thought prompting, YAML-like structure + map/actor specification	LLM is frozen; only the generator module is trained WOMO	
SeGPT (2024)	2024	Uses ChatGPT (GPT-4) to generate complex and diverse trajectory prediction scenarios	Prompt engineering includes persona-setting, reference text, task decomposition, CoT, and zero-shot CoT strategies	ChatGPT frozen, no fine-tuning	INTERACTION, and SeGPT-generated dataset
InteractTraj (2024)	2024	Uses GPT-4 to convert language descriptions into structured interaction, vehicle, and map codes	Multi-part structured prompt (interaction/vehicle/map) with detailed rules and examples	LLM is frozen, only the code-to-trajectory decoder is trained	WOMO, nuPlan (2021)
Trajectory-LLM (2025)	2025	Uses LLaMA-7B for two-stage "interaction - behavior - trajectory" generation	Generate behavior text from interaction descriptions, and then translate behavior text into trajectories	LLM is frozen, trainable regression heads on top	L2T, WOMD, Argoverse

Table 3: Methods that use large language models to generate synthetic data or augment training sets for trajectory prediction.

Method	Year	LLM Usage	Prompt Design	Fine-tuning Strategy	Dataset
GPT-Driver (2023)	2023	Uses GPT-3.5 for motion planning as a language modeling task	Parameterized perception, ego-state, and mission goal, and prompt with explicit chain-of-thought reasoning	Prompting - reasoning - fine-tuning strategy	nuScenes
LC-LLM (2024)	2024	Reformulates lane change prediction as language modeling using LLaMA-2-13B	Natural language prompt describing vehicle/map context, and CoT-style explanatory reasoning	Supervised fine-tuning with LoRA on LLaMA-2	highD (2018)
xTP-LLM (2024)	2024	Uses LLaMA2-7B-chat and LoRA fine-tuned for prediction and explanation generation	Multi-modal prompt with spatial, temporal (historical traffic), and external factors, and CoT examples	LoRA-based supervised fine-tuning	CATraffic (based on LargeST (2023))
CoT-Drive (2025)	2025	Uses GPT-4 Turbo as teacher, distilled into lightweight edge LMs (e.g., Qwen-1.5, Phi-1.5, TinyLLaMA)	4-step CoT prompt: Background - Interaction - Risk Assessment - Prediction, structured like human reasoning	Two-stage: (1) LLM distillation on CoT-labeled text; (2) supervised trajectory prediction	NGSIM (2018), highD (2018), MoCAD (2024), ApolloScape (2018), nuScenes

Table 4: Methods focusing on reasoning and interpretability in trajectory prediction via language-based explanations and LLM-generated semantics.