Trajectory Prediction Meets Large Language Models: A Survey

Anonymous ACL submission

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 language-based predictors, (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.

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 geometric or physics-based formulations or patterns extracted from past data. A core challenge lies in accurately modeling the interactions among agents. Although recent data-driven models learn these dependencies from large datasets, their generalization still hinges on data diversity and quality.

Large language models (LLMs) (Zhao et al., 2023; Chang et al., 2024) and multimodal reasoning (Wei et al., 2022; Plaat et al., 2024) open new avenues for connecting high-level semantics with low-level motion prediction. Language naturally conveys context, intent, and causality, allowing descriptions of scenes, goals, and potential futures. This expressiveness has driven growing interest in integrating LLMs into trajectory prediction, whether for contextual reasoning, reformulating trajectories as language sequences, or enabling data generation and interpretability.

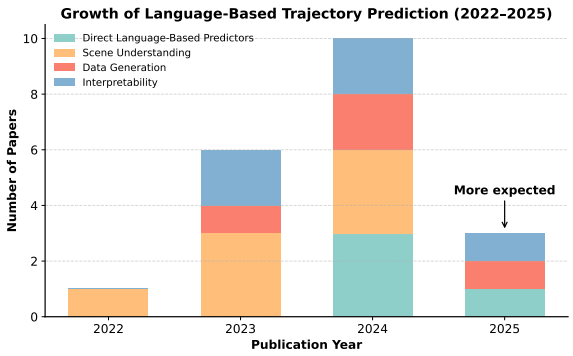

Figure 1: Growth of language-based trajectory prediction papers (2022–2025) across four categories. Only methods explicitly incorporating natural language or pretrained language models (PLMs/LLMs) are counted.

This survey provides the first systematic overview of how LLMs are reshaping trajectory prediction. The field, though young, is expanding rapidly (Figure 1), underscoring the need for a unified synthesis. We summarize existing studies, identify modeling patterns, and discuss open challenges to guide future research. 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.

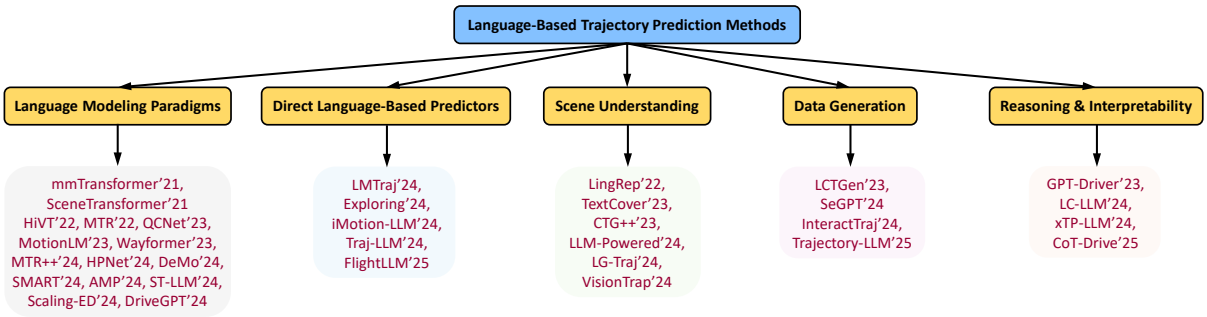

Figure 2: Taxonomy of language-based trajectory prediction methods. We categorize representative works (2022–2025) into five groups based on the functional role of language: *language modeling paradigms*, *direct language-based prediction*, *scene understanding*, *data generation*, and *reasoning & interpretability*.

- We discuss open challenges and outline future directions aimed at better aligning language understanding with trajectory prediction objectives.

2 Background

2.1 Trajectory Prediction Formulation

Trajectory prediction forecasts the future positions of agents from their observed states and scene context. Given observations $\mathbf{X} \in \mathbb{R}^{N \times H \times F}$ for N agents over H steps, the task is to predict $\mathbf{Y} \in \mathbb{R}^{N \times T \times F}$ over a horizon T, where F is typically 2 for 2D coordinates. Additional context $\mathbf{I}$ (e.g., maps or layouts) is often included, and models usually learn a probabilistic distribution $p_\theta(\mathbf{Y} \mid \mathbf{X}, \mathbf{I})$ to capture multiple plausible futures.

The problem remains challenging due to multimodal uncertainty, long-range temporal dependencies, and heterogeneous inputs. Recent studies explore LLMs as a means to address these issues.

2.2 Key Concepts in Large Language Models

Large Language Models (LLMs) are Transformer-based networks (Vaswani et al., 2017) trained on internet-scale data, excelling at language understanding and reasoning. Their autoregressive design models sequential dependencies, and pretraining enables zero- or few-shot generalization (Wei et al., 2021; Kojima et al., 2022). These traits align naturally with trajectory prediction, which requires semantic abstraction and intent reasoning among interacting agents.

To leverage LLMs for trajectory tasks, spatial or contextual data are converted into token sequences or descriptive prompts. Some works directly employ pretrained models such as 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 adapters (e.g., LoRA (Hu et al., 2022)). These adaptations enable LLMs to interpret structured inputs, generate diverse trajectories, and produce linguistic rationales within a unified framework.

3 Taxonomy

We categorize the literature into five distinct yet interconnected directions, as illustrated in Figure 2. These categories capture the methodological diversity across recent studies and highlight the different roles that language plays in enhancing predictive capability. The survey methodology, including search criteria and selection pipeline, is described in Appendix B.

- **Language Modeling Paradigms:** Trajectories are tokenized and modeled autoregressively, enabling long-range temporal reasoning.
- **Direct Language-Based Predictors:** Pretrained LLMs are directly used via prompting or fine-tuning for trajectory generation.
- **Scene Understanding:** Language augments spatial context by encoding semantics such as traffic rules or agent intent.
- **Data Generation:** LLMs synthesize diverse and controllable trajectory datasets from text.
- **Reasoning & Interpretability:** Models generate natural-language rationales to explain or justify their predictions.

We further provide a quantitative summary of representative works in Table 1, highlighting datasets, metrics, and evaluation conventions across the five categories. The appendix (Appendix D) expands this summary with detailed per-category comparisons and model configurations.

Category	Representative Methods	Common Datasets	Typical Metrics
Language Modeling Paradigms	mmTransformer, MTR, MotionLM	Argoverse, WOMD	ADE / FDE / MR
Direct Language-Based Predictors	LMTraj, iMotion-LLM, Traj-LLM	WOMD, nuScenes	ADE / FDE
Language-Guided Scene Understanding	CTG++, LLM-Powered, VisionTrap	nuScenes, BDD100K	ADE / FDE / Qualitative
Language-Driven Data Generation	LCTGen, InteractTraj, Trajectory-LLM	L2T, nuScenes-Sim	– (Synthetic)
Reasoning and Interpretability	LC-LLM, CoT-Drive, GPT-Driver	WOMD, xTP	ADE / FDE + Text Eval

Table 1: Summary-level comparison across representative works.

4 Trajectory Prediction via Language Modeling Paradigms

Trajectory prediction has been strongly shaped by advances in NLP (Vaswani et al., 2017; Devlin et al., 2019), particularly Transformer architectures and attention mechanisms. Originally designed for language sequences, Transformers effectively capture long-range dependencies and multi-agent interactions. Recent works thus adapt these sequence modeling paradigms for representing and generating trajectories, reflecting an ongoing convergence between NLP and motion forecasting.

One early contribution is mmTransformer (Liu et al., 2021), which employs a stacked Transformer framework for multimodal motion prediction. It follows a two-stage design that first generates diverse trajectory candidates and then refines them through attention-based decoding. A region-based training strategy encourages coverage over multiple plausible futures, laying the groundwork for applying attention to multimodal trajectory modeling. Another representative work, SceneTransformer (Ngiam et al., 2021), introduces a unified Transformer model that jointly predicts multiple agent trajectories, using a masking strategy inspired by NLP to flexibly condition on goals. By integrating information across road elements and agent interactions, it demonstrates the strength of attention for capturing agent dependencies, though at high computational cost and with limited semantic grounding. Collectively, these early architectures demonstrate the promise of attention for modeling interactions, while revealing two key limitations, computational inefficiency and lack of explicit semantics, that later works address through hierarchical and geometry-aware modeling.

HiVT (Zhou et al., 2022) introduces a hierarchical vectorized Transformer that separates local and global interactions through structured attention, improving representational efficiency and geometric consistency. Wayformer (Nayakanti et al., 2023) further unifies scene encoding and trajectory decoding via modular fusion and latent query attention, exploring early, late, and hierarchical integration of input modalities such as maps, agents, and signals. Together, these architectures emphasize structured fusion and query-based attention as effective biases for interaction modeling, though they still lack explicit semantic grounding from language.

Building on these structured fusion designs, Motion Transformer (MTR) (Shi et al., 2022) introduces a two-branch framework that decouples global intention localization from local motion refinement. Instead of enumerating dense trajectory candidates, it employs learnable motion queries, each representing a distinct future mode. This query-based formulation improves diversity and training stability, achieving strong results on the Waymo Open Motion Dataset (WOMD) (Ettinger et al., 2021). MTR++ (Shi et al., 2024) extends this idea to multi-agent settings through symmetric scene modeling and mutually guided intention queries, enabling agents to coordinate predictions via structured attention and generate scene-consistent trajectories. These intention-query paradigms advance multimodal coverage and interaction reasoning, though they remain sensitive to mode granularity and candidate sampling.

Building on intention-query designs, QCNet (Zhou et al., 2023) shifts the focus to an agent-centric query formulation, where each target agent retrieves context most relevant to its own motion. This selective querying improves both prediction accuracy and interpretability, particularly in dense traffic scenarios. Complementing this direction, HPNet (Tang et al., 2024) enhances temporal reasoning by feeding past predictions back into the attention mechanism. Its historical prediction attention explicitly links prior forecasts with future motion, improving stability over long horizons. Extending disentangled modeling further, DeMo (Zhang et al., 2024) separates trajectory generation into two complementary query types, directional mode queries to encode intent and dynamic state queries to model motion evolution. By integrating attention with state-space modeling (Gu

and Dao, 2023), it achieves fine-grained control over trajectory diversity and physical plausibility. Together, these works demonstrate a progressive refinement from query-centric interaction modeling toward modular reasoning frameworks that unify attention and dynamics.

Building on Transformer-based architectures, recent research has moved beyond structural borrowing to directly adopt language modeling paradigms for trajectory generation. The rise of LLMs motivates viewing trajectory prediction as an autoregressive token-generation task, where continuous motion is discretized into symbolic tokens and generated step by step, much like words in a sentence. This shift marks a conceptual transition from attention-based interaction modeling to full language-style sequence generation.

MotionLM (Seff et al., 2023) first formulated motion forecasting as a language modeling problem, using a causal Transformer decoder to jointly predict multi-agent futures. SMART (Wu et al., 2024) extended this to real-time simulation, achieving low-latency inference through tokenized map-trajectory representations. Building on these ideas, AMP (Jia et al., 2024) unified token space design and factorized attention to enhance structure-aware prediction, while ST-LLM (Liu et al., 2024) further integrated spatial–temporal tokenization for large-scale forecasting. Together, these models show how discrete motion tokens enable unified sequence reasoning and improved long-horizon consistency.

Despite strong accuracy, the computational cost of autoregressive decoding remains high. Scaling-ED (Ettinger et al., 2024) addresses this by distilling large ensembles into compact student models, achieving a practical balance between performance and efficiency. Pushing scale further, DriveGPT (Huang et al., 2024) applies autoregressive Transformers to billions of driving frames, empirically validating scaling laws and bridging trajectory prediction with foundation model trends.

These works indicate the shift from architecture transfer to paradigm transfer, from using attention as a tool to adopting language modeling as a principle. While scaling and tokenization improve generalization, challenges remain in spatial grounding and physical realism, highlighting the next frontier for language-inspired prediction systems.

Discussion. Language-modeling paradigms have brought new structure and abstraction to trajectory prediction, introducing tokenized motion generation, scalable sequence reasoning, and cross-agent attention mechanisms. Yet, most remain architectural transfers from NLP rather than semantically grounded systems: they excel at temporal coherence but struggle with spatial precision and interpretability once trajectories are discretized. A core tension emerges between scalability and physical realism. Large autoregressive models generalize well with data scale but often overlook continuous dynamics and agent intent. Furthermore, evaluation protocols still center on geometric error (e.g., ADE/FDE) rather than semantic fidelity, leaving reasoning quality underexplored. Future directions may include hybrid formulations that integrate discrete sequence modeling with continuous spatial reasoning, enabling models to move beyond syntactic prediction toward genuine semantic understanding of motion and intent.

5 Direct Trajectory Prediction with Language-Based Predictors

Beyond cataloging representative systems, this section examines how language conditioning aids trajectory prediction, in intent control, few-shot transfer, and cross-domain generalization, and where it still falls short, such as numerical precision and spatial grounding. Unlike Section 4, which focused on architectures inspired by NLP, the methods here employ pretrained language models (PLMs/LLMs) directly, either fine-tuned or used in zero-/few-shot modes, as the predictive core for trajectory generation. This shift reflects a move from architectural borrowing to language models as active predictors, leveraging their reasoning and generalization strengths across domains and agent types. In this section, we refer to both fine-tuned pretrained language models (PLMs) such as T5 and large-scale generative models like GPT-3.5/4 as language-based predictors. While they differ in scale, both function as pretrained, language-conditioned modules within trajectory pipelines.

LMTraj (Bae et al., 2024) reformulates trajectory prediction as a question-answering task, converting numerical and scene features into structured prompts. Through a decimal-aware tokenizer and auxiliary interaction objectives, it bridges the gap between continuous values and discrete tokens, outperforming regression baselines. This shows that linguistic prompting can encode high-level reasoning, though it remains sensitive to token granularity. Expanding this direction, Munir et al. (2024) benchmarks multiple 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 the trajectory prediction task. They find strong intent reasoning but weak numerical grounding, underscoring the need for domain-specific adaptation. To strengthen intent alignment, iMotion-LLM (Felemban et al., 2024) applies instruction tuning via LoRA fine-tuning (Hu et al., 2022) on WOMD, pairing language directives (e.g., "turn left at the next intersection") with motion data. This demonstrates language as a flexible control interface, though template-based commands may limit generalization. FlightLLM (Luo and Zhou, 2025) extends this paradigm to aviation, serializing flight paths into tokens and fine-tuning LLaMA for short- and long-term forecasting, while noting inference latency as a key barrier. Finally, Traj-LLM (Lan et al., 2024) integrates LLMs directly into the trajectory pipeline, encoding multimodal scene context into transformer-compatible sequences. Using a GPT-2 backbone with LoRA fine-tuning and a lane-aware probabilistic head, it maintains high accuracy even with limited data, highlighting LLMs' few-shot adaptability.

Together, these methods illustrate the transition from using language as inspiration to using it as computation, LLMs serving as trainable predictors. Despite clear gains in interpretability and cross-domain transfer, persistent issues remain in grounding, precision, and efficiency, calling for hybrid or hierarchical designs that reconcile reasoning power with spatial fidelity.

Discussion. Direct language-based predictors embed reasoning directly into motion forecasting, excelling in few-shot generalization and intent conditioning. Yet their reliance on tokenization and autoregressive decoding leads to numerical imprecision and latency bottlenecks, complicating deployment. Most current systems trade spatial control for interpretability: they can explain intent but struggle to anchor it geometrically. A promising path lies in hybrid models that couple continuous spatial embeddings with linguistic prompts, or multi-stage fine-tuning schemes that retain reasoning while improving precision and efficiency.

6 Language-Guided Scene Understanding for Trajectory Prediction

Traditional trajectory predictors mainly rely on structured inputs, agent coordinates, map layouts, or categorical signals, that often fail to capture scene semantics or intent-level interactions. To bridge this gap, recent studies introduce natural language as a complementary modality, either through textual inputs, intermediate representations, or language-conditioned reasoning modules. Language-guided cues enrich perception, enabling more interpretable and socially aware prediction in complex environments.

LingRep (Kuo et al., 2022) first used template-based descriptions (e.g., "crossing at intersection", "following another car") as intermediate representations to interpret agent behavior. While effective for interpretability, its rule-based design limits scalability and domain transfer. Moving beyond handcrafted templates, TextCover (Keysan et al., 2023) integrates DistilBERT embeddings into a Transformer backbone, fusing textual and spatial representations for joint reasoning over context and motion. Results show that pretrained textual embeddings can enhance urban-scene prediction, but only when spatial–text alignment is well learned.

Pushing this idea further, CTG++ (Zhong et al., 2023) adopts diffusion-based generation guided by natural-language prompts. GPT-4 converts free-form user descriptions (e.g., "a group of pedestrians crossing while a car waits") into differentiable control functions, enabling scenario-level synthesis. Similarly, LLM-Powered (Zheng et al., 2024) embeds GPT-4V-derived semantic cues into the attention layers of a motion-prediction network, improving focus on socially relevant agents. These efforts highlight the emerging role of multimodal LLMs as social reasoners, though their computational overhead and dependency on prompt quality remain barriers to deployment.

Focusing on pedestrian interactions, LG-Traj (Chib and Singh, 2024) summarizes prior motion and visual context into a language-derived scene embedding, improving intent disambiguation in crowded settings. VisionTrap (Moon et al., 2024) goes a step further by using vision–language models (VLMs) to auto-generate scene-level cues (e.g., "child near crosswalk") that guide attention toward salient regions. Despite such gains, current VLM-based pipelines still face limited visual grounding and high inference cost.

Together, these works trace a progression from language as description to language as control signal. By encoding scene semantics and relational context in natural language, they illustrate language's growing value for context-aware and so-

cially aligned prediction. As alignment improves, language-conditioned models may become key to bridging perception, reasoning, and motion generation in complex environments.

Discussion. Language-guided approaches show how linguistic cues can link low-level perception with high-level intent reasoning, enhancing interpretability and contextual awareness in dense or ambiguous scenes. Yet, their success hinges on multimodal alignment quality, imbalanced training often leads to misplaced attention or incoherent goal inference. The main trade-off lies between semantic richness and efficiency: stronger language integration yields better reasoning but higher inference cost. Future work could explore lightweight adapters for linguistic grounding, joint spatial–text training, and benchmarks that quantify semantic alignment. Reporting text-to-region grounding metrics (e.g., referring-expression accuracy) alongside ADE/FDE would make progress on reasoning and perception more measurable.

7 Language-Driven Data Generation for Trajectory Prediction

As trajectory predictors grow increasingly data-hungry and domain-specific, recent work explores LLMs as generative engines for creating synthetic data. Natural language serves as a high-level interface to describe, simulate, and construct diverse trajectory scenarios, offering a scalable alternative to data collection and rule-based scene design.

LCTGen (Tan et al., 2023) first proposed using GPT-4 (Achiam et al., 2023) as an interpreter to synthesize traffic scenes from natural-language prompts (e.g., "a vehicle overtakes on the left while a pedestrian crosses"). A Transformer-based decoder then generates agent distributions and motions grounded in map topology, establishing early feasibility for scene-level linguistic control, albeit with limited physical fidelity and diversity. Building on this idea, SeGPT (Li et al., 2024) employs chain-of-thought (Wei et al., 2022) prompting with ChatGPT to convert free-form text into structured driving configurations. Although trajectories remain simplified, the framework illustrates how pretrained LLMs can act as controllable scenario generators for simulation and planning research, while also revealing the difficulty of enforcing physical plausibility from purely textual reasoning.

To enhance behavioral realism, InteractTraj (Xia et al., 2024) translates abstract descriptions into structured motion codes via GPT-4, later decoded into multi-agent trajectories capturing interactions such as yielding, merging, or aggressive acceleration. This pipeline balances linguistic expressiveness with physical realism, though each translation stage introduces uncertainty. Most recently, Trajectory-LLM (Yang et al., 2025) adopts LLaMA-7B to convert compact textual descriptions of interactions into multimodal trajectory data. Its three-stage Interaction–Action–Trajectory process, paired with map encoding, yields the L2T dataset of 240K text–map–trajectory samples, demonstrating a scalable route to semantically rich data generation. However, such large-scale synthesis raises concerns about annotation quality, scenario balance, and cross-domain reproducibility.

Together, these works redefine language as not only a control interface but also a source of structured supervision. By producing richly annotated and semantically coherent synthetic data, language-driven frameworks open new directions for cross-domain training, evaluation, and simulation in trajectory prediction.

Discussion. Language-driven data generation offers a powerful yet underexplored frontier for trajectory modeling. Using language as a generative interface simplifies scenario specification and broadens data diversity, but realism and physical consistency remain key limitations, most frameworks depend on textual priors rather than explicit dynamics. Critical constraints include enforcing speed and acceleration bounds, collision avoidance, and map compliance (lane adherence, right-of-way). A recurring trade-off arises between controllability and fidelity: linguistic prompts enable intuitive design but complicate physically plausible synthesis and balanced coverage. To assess practical utility, we suggest reporting synthetic-to-real transfer deltas (ADE/FDE) and scenario-coverage metrics over taxonomy categories. Future work may integrate physics-aware constraints, human-in-the-loop validation, and standardized benchmarks to evaluate realism, diversity, and generalization of language-generated data.

8 Language-Based Reasoning and Interpretability for Trajectory Prediction

As trajectory predictors become increasingly complex, understanding how and why they produce specific outputs is essential for safety, trust, and

effective human–AI collaboration. Recent studies integrate LLMs not only to generate trajectories but also to explain and reason about them—linking motion forecasting with natural language reasoning. By articulating goals, intermediate intentions, or decision chains via chain-of-thought (CoT) prompting (Wei et al., 2022), these approaches aim to expose the logic behind autonomous decisions.

GPT-Driver (Mao et al., 2023) represents one of the first unified frameworks to generate both future maneuvers and textual rationales using GPT-style decoding. This joint formulation provides interpretable motion planning but struggles with causal grounding, often yielding explanations that sound reasonable yet lack mechanistic accuracy.

LC-LLM (Peng et al., 2024) extends this idea with a dual-task setup, predicting lane-change actions and their textual justifications (e.g., "the vehicle merges left due to a slower car ahead"). Such narrative explanations improve user trust but risk post-hoc rationalization, where language aligns with outcomes rather than underlying decision logic. Taking a broader view, xTP-LLM (Guo et al., 2024) adapts LLaMA2-7B-Chat for traffic flow forecasting, translating multimodal signals, weather, points of interest, and historical flows into structured prompts. It demonstrates the scalability of LLM reasoning for aggregate prediction, though extending interpretability to agent-level behavior remains unsolved.

More recently, CoT-Drive (Liao et al., 2025) explicitly treats trajectory prediction as a reasoning task. Instead of direct regression, the model verbalizes intermediate steps (e.g., "the car approaches a stop sign, slowing down is expected") before final trajectory generation. This structured reasoning improves both accuracy and interpretability but incurs high computational cost and limited quantitative evaluability, as reasoning traces often diverge from ground-truth dynamics.

Collectively, these works shift language from a descriptive layer to a reasoning mechanism, using verbalized logic to reveal model intent and decision processes. Through waypoint explanation, rationale generation, and CoT-based decision chains, LLMs are emerging as a bridge between predictive performance and human-centered transparency.

Discussion. Language-based reasoning offers a path toward interpretable and trustworthy motion forecasting by verbalizing internal decision logic. Yet current explanations are mostly descriptive rather than causal—LLMs often produce plausible narratives without faithfully mirroring internal computations. Evaluating faithfulness remains difficult, practical probes include counterfactual sensitivity (does the rationale change under scene edits?) and rationale–action consistency (does removing rationale tokens affect predicted motion?). A persistent tension arises between interpretability and reliability: detailed rationales improve human understanding but risk inconsistency or ambiguity. Future work should align textual rationales with latent decision representations, develop faithfulness metrics, and explore causal reasoning modules that ensure both correctness and clarity. A lightweight direction is to distill CoT traces into compact latent rationales, preserving decision-critical cues while mitigating decoding cost.

9 Challenges and Future Directions

Despite remarkable progress, LLM-based trajectory prediction still faces intertwined challenges across representation, reasoning, and evaluation. These are not isolated technical bottlenecks but manifestations of a deeper tension between linguistic abstraction and spatial grounding. Below, we summarize key challenges and outline promising directions toward more scalable, interpretable, and physically consistent predictors.

Tokenization and Representation. A core challenge lies in converting continuous, multi-agent trajectories into discrete sequences suitable for LLM processing. Naïve numeric tokenization often leads to bloated or lossy representations. Recent work such as (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, such token schemes are typically hand-crafted and may not generalize across agents or domains. Future directions include learned or hierarchical tokenizers, semantic abstractions (e.g., maneuver- or intent-level tokens), and hybrid representations that couple discrete language tokens with continuous latent embeddings. The central difficulty remains balancing spatial fidelity with linguistic tractability, a fundamental trade-off that defines this research frontier.

Prompt Design and Alignment. Effectively steering LLMs for trajectory reasoning depends on prompts that encode scene semantics, agent states, and prediction objectives. Despite early success, prompt alignment remains fragile and often task-

specific. Scalable strategies such as soft prompting, instruction tuning, or reinforcement learning from feedback (RLHF/RLAIF) may enhance cross-domain generalization and adaptive control. Ultimately, progress hinges on moving beyond handcrafted templates toward self-aligned interfaces, where models autonomously interpret and refine instructions based on context and feedback.

Reasoning and Prediction Coherence. LLMs offer an opportunity to embed commonsense and causal reasoning into motion forecasting, yet their temporal and physical reasoning remain limited. While Peng et al. (2024) demonstrates the benefit of chain-of-thought reasoning for explaining long-term forecasts, most models still produce inconsistent or physically implausible outputs (Munir et al., 2024). Enhancing reasoning may require physics-informed priors, structured scene graphs, or verifiable planning modules that constrain generation. Developing reasoning-aligned datasets with explicit causal annotations could further encourage robust decision-making. The broader challenge is to infuse linguistic reasoning with causal, physically grounded logic, while maintaining the flexibility that makes LLMs powerful.

Multimodal Context Integration. Real-world trajectory forecasting demands understanding of visual perception, map topology, and language instructions. Several works address this by projecting scene features into token-compatible embeddings (Felemban et al., 2024) or encoding environmental cues as text (Chen et al., 2024). However, coupling perception with prediction remains difficult due to modality mismatches and limited context windows. Emerging architectures that unify vision and language modeling (e.g., VLMs, multimodal Transformers) or use graph-enhanced scene prompts show promise for integrating heterogeneous signals. The ultimate goal extends beyond simple fusion, it is representational unification, where vision, map, and language share a coherent reasoning space.

Transparency and Interpretability. One of the most promising byproducts of incorporating LLMs is their ability to generate natural language rationales for predicted behaviors. Models such as LC-LLM (Peng et al., 2024) and xTP-LLM (Guo et al., 2024) enhance human trust by coupling predictions with textual explanations. However, ensuring fidelity, that explanations reflect actual decision processes, remains challenging. Future research may integrate internal alignment losses, counterfactual reasoning, or multi-task learning, where explanation quality directly influences prediction. A persistent tension exists between interpretability and autonomy: systems that explain well may not always act optimally. Achieving causal, faithful, and efficient interpretability frameworks will be crucial for safety-critical deployment.

Evaluation and Benchmarking. As LLM-based predictors generate both trajectories and rationales, traditional metrics such as ADE, FDE, and MR (see Appendix A) are no longer sufficient. Existing benchmarks rarely assess multi-agent coordination, reasoning coherence, or language-grounded objectives. Moreover, textual outputs lack standardized evaluation protocols. Future benchmarks should jointly evaluate trajectory accuracy, semantic plausibility, and explanation faithfulness, supported by language-augmented datasets with aligned ground truth. Next-generation evaluations must capture holistic decision quality, covering reasoning correctness, behavioral safety, and linguistic reliability, rather than isolated positional errors.

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.

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.

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.

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.

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*.

Appendix

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 Survey Methodology

To ensure both comprehensive coverage and methodological transparency, we followed a structured process to select and organize relevant literature at the intersection of large language models (LLMs) and trajectory prediction.

Time frame and data sources. Our survey focuses on works published between 2022 and 2025, reflecting the rapid development of LLMs in recent years. We sourced papers from arXiv, ACL Anthology, Semantic Scholar, and Google Scholar. These repositories were chosen to cover both peer-reviewed publications and influential preprints.

Search process. We used keyword-based queries including *"trajectory prediction"*, *"trajectory forecasting"*, *"motion forecasting"*, and *"large language model"* to identify relevant works. Boolean combinations of terms (e.g., "trajectory prediction and LLM") were used to refine results. We also

performed backward and forward citation tracking on several seed papers to ensure broader coverage.

Selection criteria. We included papers that use language models, either through prompting, fine-tuning, or alignment, to guide, improve, or explain trajectory prediction tasks. Both agent-level motion forecasting (e.g., vehicles, pedestrians) and simulation-oriented prediction were considered. We excluded studies that focus solely on traffic flow modeling, origin-destination estimation, or aggregate-level forecasting without agent-specific predictions.

Scope. After initial screening, we reviewed over 70 papers and retained 35 representative works for in-depth analysis. These were selected based on relevance, technical novelty, diversity of modeling approaches, and influence (e.g., venue, citations, or novelty). Preprints from arXiv were included when they presented novel approaches or had gained significant attention within the community.

C 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.

D 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 Language-Based Predictors, (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 2, Table 3, Table 4, and Table 5, 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 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 2: 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 3: 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 4: 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 5: Methods focusing on reasoning and interpretability in trajectory prediction via language-based explanations and LLM-generated semantics.