Commun Transp Res

Research Article

CogDrive: Cognition-driven multimodal prediction-planning fusion for safe autonomy

Heye Huang[1,2,†], Yibin Yang[3,†], Mingfeng Fan[4], Haoran Wang[5], Xiaocong Zhao[5,✉], Jianqiang Wang[3]

Cite this article: https://doi.org/10.26599/COMMTR.2026.9640016

ABSTRACT: Safe autonomous driving in mixed traffic requires a unified understanding of multimodal interactions and dynamic planning under uncertainty. Existing learning-based methods often fail to capture rare but safety-critical behaviors, while rule-based systems lack adaptability in complex interactions. To address these limitations, we propose CogDrive, a cognition-driven multimodal prediction–planning fusion framework that integrates explicit modal reasoning with safety-aware decision optimization. The prediction module introduces cognitive representations of interaction modes based on topological motion semantics and nearest-neighbor relational encoding. By incorporating a differentiable modal loss and multimodal Gaussian decoding, CogDrive effectively learns sparse and unbalanced interaction behaviors, improving long-tail trajectory prediction accuracy. The planning module builds upon an emergency-response concept and develops a safety-stabilized trajectory tree optimization. Short-term consistent root trajectories ensure safety within replanning cycles, while long-term branches provide smooth and collision-free avoidance under low-probability or rapidly switching modes. Experiments on Argoverse2 and INTERACTION datasets show that CogDrive achieves state-of-the-art performance, reducing minADE and miss rate while maintaining smoothness. Closed-loop simulations further confirm stable and adaptive behavior across strong-interaction scenarios such as merging and intersections. By coupling cognitive multimodal prediction with safety-oriented planning, CogDrive establishes an interpretable and reliable paradigm for safe autonomy in complex traffic.

KEYWORDS: Safe autonomy, Cognitive modeling, Multimodal prediction, Dense traffic

1 Introduction

In mixed traffic, autonomous vehicles must interact with uncontrollable human-driven agents, making interactive trajectory planning fundamentally challenging. The ego vehicle must generate dynamically feasible and safe trajectories in real time to pass potential conflict zones (Yang et al., 2024; Huang et al., 2024; Han et al., 2025). Overly conservative strategies lead to the "freezing robot" phenomenon, while overly aggressive behaviors increase collision risk.

This dilemma stems from two challenges. First, strongly interactive scenarios are sparse and imbalanced, limiting learning-based methods in capturing rare but safety-critical behaviors (Liu et al., 2025). Second, trajectory prediction is inherently multimodal and uncertain, with modes that may shift or conflict (Ngiam et al., 2021). Conventional planners thus struggle to reason over all modes in real time, often producing unstable trajectories. Since surrounding agents are uncontrollable, the ego vehicle must adapt to decentralized multimodal interactions rather than rely on centralized planning (Hagedorn et al., 2024). The key challenge is safe and efficient conflict traversal.

Existing methods generally fall into two categories. Learning-based approaches leverage large-scale data for prediction (Jiang et al., 2023; Yang et al., 2025) but lack interpretability and robustness in rare cases. Rule-based planners ensure feasibility and transparency (Pek & Althoff, 2020; De Luca et al., 2025) but often lack adaptability in complex interactions. To bridge this gap, we propose CogDrive, a cognition-driven multimodal prediction–planning fusion framework. CogDrive integrates differentiable modal reasoning with safety-stabilized planning, unifying prediction and planning into a coherent process that perceives, anticipates, and acts under multimodal uncertainty (Huang et al., 2025a). The main contributions are as follows: (1) We propose CogDrive, a cognition-driven multimodal prediction–planning fusion framework that unifies learning-based adaptability and rule-based safety stability within a unified cognitive decision process. (2) We design an interaction-aware multimodal prediction module, which encodes inter-agent semantics via modal classification and differentiable modal loss, improving the learning of sparse yet safety-critical behaviors. (3) We evaluate CogDrive on the Argoverse2 and INTERACTION datasets, achieving state-of-the-art performance in prediction accuracy, planning stability, and driving safety.

† Heye Huang and Yibin Yang contributed equally to this work.

[1] Singapore-MIT Alliance for Research and Technology, Singapore 138602, Singapore. [2] Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge MA 02139, USA. [3] School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China. [4] Department of Mechanical Engineering, National University of Singapore, Singapore 117576, Singapore. [5] Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai 200092, China.

✉ Corresponding author. E-mail: zhaoxc@tongji.edu.cn.

Received: October 31, 2025; Revised: December 10, 2025; Accepted: March 3, 2026.

2 Related Works

Interactive single-vehicle trajectory planning methods can be broadly classified into end-to-end and non-end-to-end approaches (Chib & Singh, 2023). End-to-end methods directly map sensor inputs to trajectories or control commands using neural networks (Prakash et al., 2021). Despite performance gains from attention mechanisms and large-scale datasets, these methods depend heavily on high-fidelity sensing and simulation, limiting their ability to capture multimodal uncertainty and exacerbating sim-to-real gaps. Non-end-to-end approaches mitigate these issues by leveraging detection, tracking, and high-definition maps to generate feasible ego trajectories, and can be further categorized into learning-based and rule-based planning.

2.1. Learning-based Methods

Learning-based methods are generally categorized into reinforcement learning (RL) and imitation learning (IL). RL optimizes policies through reward signals but often struggles with real-world transfer, whereas IL learns ego motions from large-scale naturalistic data with subsequent safety post-processing. These approaches can be further grouped into three categories.

Embedding and Feature Representation. Embedding methods map heterogeneous inputs, such as agent states and road centerlines, into latent representations essential for interaction modeling and trajectory planning (Shi et al., 2022). Two main approaches are rasterized and vectorized embeddings. Rasterized inputs, typically derived from bird's-eye-view images or LiDAR grids, enable multi-source fusion but are computationally expensive and less effective for long-range interactions, and their reliance on handcrafted features and limited receptive fields restricts performance in highly interactive scenarios. Vectorized embeddings have become the dominant paradigm by directly encoding agent trajectories and map elements as polylines. VectorNet pioneered this design using graph neural networks to capture pairwise and group-level interactions (Gao et al., 2020). Building on this, MTR (Sun et al., 2024) applies PointNet (Qi et al., 2017) with MLPs to encode polylines, followed by max-pooling for feature aggregation. Such vectorized representations are now widely adopted in trajectory prediction and form a highly effective embedding strategy.

Coordinate Systems and Normalization. Coordinate system design is closely tied to normalization, which stabilizes training by reducing gradient explosion or vanishing and improving optimization efficiency (Carion et al., 2020). In single-agent prediction, ego-centric coordinates, where the ego vehicle's position and heading define the frame, naturally provide translation and rotation invariance, enabling more efficient learning. Scene-centric coordinates instead align inputs to a global frame, ensuring consistency but reducing flexibility for local interactions. For multi-agent prediction, unifying different agent frames is challenging. Recent work therefore adopts an instance-centric design, where each agent or map element defines its own local frame. Road centerlines use midpoints and tangents, while vehicles use their current pose as the origin. This allows normalization and symmetry across agents, but requires mechanisms to transform information between frames (Zhang et al., 2024). Representative methods include MTR++, which applies sinusoidal relative position encodings (Shi et al., 2022), HPTR, which encodes relative coordinates with sine–cosine functions (Zhang et al., 2023), and HiVT (Zhou et al., 2022), which construct local frames with rotation matrices to model inter-agent interactions. Instance-centric coordinates have thus become a robust compromise for interaction-aware trajectory prediction and planning.

Encoder–Decoder Architectures and Multimodal Generation. To address multimodality in interactive scenarios, various encoder–decoder architectures have been developed. Probabilistic generative models, such as GANs (Gupta et al., 2018), VAEs (Cai et al., 2025), and diffusion models (Jiang et al., 2023), generate multiple candidate trajectories via sampling, but often face mode collapse, limited interpretability, and high computational cost. Anchor-based approaches, exemplified by the TNT family (Huang et al., 2020), predefine endpoints or behaviors as anchors to produce trajectories with clearer modal distinctions, though their performance is constrained by the trade-off between accuracy and anchor count. More recently, Transformer-based encoder–decoder frameworks with learnable query embeddings have emerged as a strong alternative for multimodal prediction (Zhou et al., 2022). DETR-style architectures leverage attention to jointly model spatial relations and behavioral diversity, offering improved scalability and interpretability over sampling- or anchor-based methods. Each learnable query encodes a potential behavioral mode, enabling unified modeling of both common and rare interactions within a shared attention space. By replacing stochastic sampling with deterministic relational reasoning, query-driven designs achieve higher training stability, controllable diversity, and stronger mode–behavior alignment, making them a leading paradigm for learning-based multimodal prediction.

2.2. Rule-based Methods

In contrast, rule-based methods emphasize interpretability and safety, typically generating trajectories through predefined models or rules. The core challenge lies in making safe decisions under uncertainty. (a) **Maximum Likelihood Planning.** These methods assume surrounding agents will follow their most probable behaviors and plan ego trajectories accordingly. Methods include Monte Carlo tree search (Lenz et al., 2016), finite-state machines (Meng et al., 2021), raster or optimizations (Huang et al., 2025b). They are computationally efficient but often fail to handle rare yet dangerous behaviors, which can result in discontinuous or unsafe decisions under unexpected interactions. (b) **Partially Observable Markov Decision Processes.** POMDP frameworks explicitly model uncertainty and have been applied in interactive planning systems such as EPSILON, which integrates behavior planning with optimization-based motion planning (Ding et al., 2021). EPSILON employs guided branching in action–observation spaces and a spatio-temporal semantic corridor to generate safe, smooth trajectories. These methods provide clear interpretability under uncertainty but remain computationally demanding in large-scale dynamic traffic. (c) **Defensive and Contingency Planning.** Defensive planning approaches generate trajectory trees with shared initial segments and branches to hedge against different predicted futures (Huang et al., 2024a; Liu et al., 2019). Fail-safe motion planning or contingency MPC generate such trees to cover multiple modalities (Pek & Althoff, 2020). They ensure short-term safety but scale poorly: trajectory tree size grows exponentially with prediction horizon and number of agents, limiting real-time use. They provide strong safety guarantees and interpretability, but often lack adaptivity and flexibility when confronted with highly interactive, multimodal uncertainties.

3 Framework Overview

CogDrive follows the principle of cognition-driven autonomy, in which cognitive mechanisms bridge perception, reasoning, and control to model the human–vehicle–road system and its interactions. Fig. 1 provides a unified overview of the end-to-end workflow, from coordinate embedding and multi-agent interaction encoding to multimodal prediction and safety-stabilized planning. By combining the interpretability of rule-based reasoning with the adaptability of data-driven learning, CogDrive enables generalizable and reliable decision-making. Building on this foundation, interaction-aware motion generation in mixed traffic is formulated as a unified process that couples multi-agent prediction with ego-vehicle planning.

Multi-Agent Joint Trajectory Prediction. In interactive traffic, each agent's future motion depends on its history and surrounding behaviors. Given observed agent trajectories $S_A \in \mathbb{R}^{M\times T_h\times C_a}$ and HD map features $S_R \in \mathbb{R}^{N_r\times N_p\times C_r}$, CogDrive estimates the joint predictive distribution $P(Y \mid S_A, S_R)$. Each agent's motion is modeled with K behavioral modes using a Gaussian mixture, yielding $P(Y_{i,t} \mid S_A, S_R) = \sum_{k=1}^{K} p_k \mathcal{N}(\mu_{i,t}^k, \Sigma_{i,t}^k)$, where $\sum_k p_k = 1$. This formulation captures multimodal uncertainty, including rare but safety-critical behaviors.

Ego-Vehicle Safety-Stabilized Planning. Conditioned on the multimodal prediction, the planner generates a feasible and safe ego trajectory $\mathcal{T}$. CogDrive adopts a cognition-inspired hierarchical strategy by first computing a short-horizon root trajectory $\mathcal{T}_{0:T_b}^{\text{root}}$ to guarantee immediate safety, and then extending mode-specific branches $\mathcal{T}_{T_b+1:T}^{k}$. The branching time T_b exceeds the replanning cycle, ensuring dynamically feasible and collision-free execution. Through this bidirectional coupling, multimodal prediction provides interpretable intent cues, while planning enforces safe realization of each mode, forming the cognitive backbone of CogDrive.

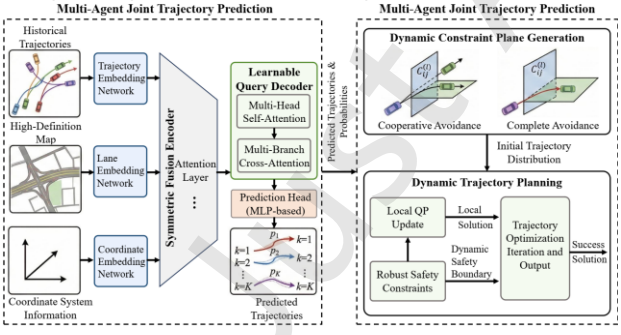


Fig. 1. Unified framework of CogDrive. The system comprises two tightly coupled components: (1) Multiagent joint trajectory prediction, (2) Safety-stabilized trajectory planning.

4 Multimodal Joint Trajectory Prediction

The proposed CogDrive framework formulates multimodal joint trajectory prediction as a cognition-driven learning process, enabling interpretable reasoning over motion intentions among interacting agents (Fig. 2). The network integrates scene geometry, agent dynamics, and coordinate information into a unified vectorized representation to capture diverse behavioral modalities.

4.1. Behavioral Modality Modeling

To represent diverse and controllable interaction patterns among agents, CogDrive introduces a cognitive behavioral modality representation inspired by topological motion equivalence. Traditional single-vehicle planning assumes that trajectories with identical start and end points can be smoothly deformed without crossing obstacles, forming a topological homotopy class. Extending this idea to interactive driving, CogDrive encodes agent-to-agent behaviors using continuous deformation relationships that distinguish different motion intentions, such as yielding, merging, or overtaking.

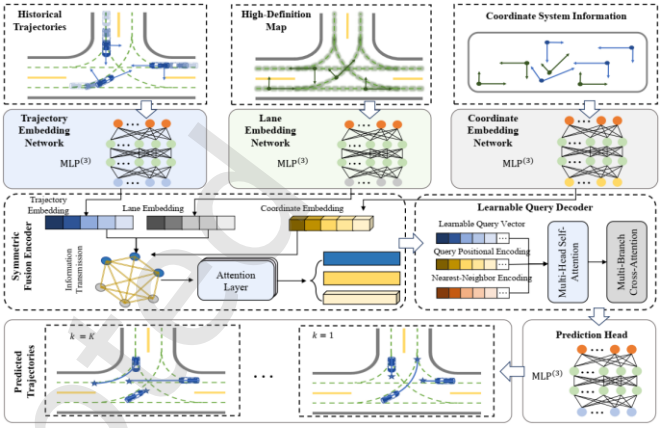


Fig. 2. Cognition-driven multimodal prediction network in CogDrive. Trajectories, maps, and local coordinates are encoded and symmetrically fused to model agent interactions. Learnable queries decode multimodal joint trajectories, each corresponding to a distinct ego–agent interaction mode.

Given two agents i and j with trajectories $\mathcal{T}_i$ and $\mathcal{T}_j$, the relative angular displacement between them is quantified by the cumulative change in their relative bearing:

$$\Delta\theta_m(\mathcal{T}_i,\mathcal{T}_j) = \sum_{t=0}^{T-1} f_{norm}\left(\arctan\frac{y_i^{t+1}-y_j^{t+1}}{x_i^{t+1}-x_j^{t+1}} - \arctan\frac{y_i^{t}-y_j^{t}}{x_i^{t}-x_j^{t}}\right), \quad (1)$$

where $f_{norm}(\cdot)$ normalizes the angular difference to the interval $[-\pi,\pi]$. By applying a threshold $\hat{\theta}$, interaction modes are categorized into three discrete types:

$$m(\mathcal{T}_i,\mathcal{T}_j) = \begin{cases} -1, & \Delta\theta_m < -\hat{\theta}, \\ 0, & -\hat{\theta} \le \Delta\theta_m \le \hat{\theta}, \\ 1, & \Delta\theta_m > \hat{\theta}. \end{cases} \quad (2)$$

Here, $m = -1$, 0, and 1 respectively represent yielding, neutral, and aggressive behaviors between two vehicles. This compact encoding allows the network to classify and predict interaction modes directly from spatial relationships.

In multi-agent settings, the ego vehicle's modality vector is defined as

$$\mathbf{m}(\mathcal{T}_{ego},\mathcal{T}_1,\ldots,\mathcal{T}_{M-1}) = [m(\mathcal{T}_{ego},\mathcal{T}_1),\ldots,m(\mathcal{T}_{ego},\mathcal{T}_{M-1})], \quad (3)$$

where M denotes the number of surrounding agents. This vector compactly encodes the ego's behavioral relations with multiple neighbors, capturing the combinatorial nature of interactive driving. To maintain tractability, CogDrive considers only the nearest neighbors in interaction space, enabling efficient learning and real-time inference. During training, a differentiable surrogate of the modality function is adopted to support gradient-based optimization, yielding smooth and interpretable transitions between interaction behaviors.

The interaction modality thresholds are learned in a data-driven manner from the multimodal probability distribution, where distinct modes emerge as probability peaks under a max-margin modality objective, rather than being manually specified or heuristically clustered.

4.2. Scene Representation and Relative Feature Encoding

To ensure geometric consistency and spatial invariance, CogDrive adopts an instance-centric coordinate representation. For each predicted agent, a local frame is defined with its current position as the origin and heading aligned with the x-axis. Surrounding agents and lane polylines are transformed from the global map frame into this local coordinate system, preserving scene structure while improving learning efficiency and generalization.

In multi-agent settings, agents reside in distinct local frames, which obscures mutual geometric relations. To recover this information, CogDrive introduces relative positional encoding to model pairwise spatial relationships. Let $\mathbf{z}_i = (x_i, y_i)$ denote the origin of instance i and θ_i its heading. For any pair (i, j), their relative orientation and distance are computed as

$$\theta_{i,j} = \theta_j - \theta_i, \beta_{i,j} = \arctan\frac{y_j - y_i}{x_j - x_i} - \theta_i, d_{i,j} = \| \mathbf{z}_i - \mathbf{z}_j \|. \quad (4)$$

These terms encode heading difference, bearing angle, and distance. The resulting relative feature is defined as

$$\mathbf{r}_{p,i,j} = [\sin\theta_{i,j}, \cos\theta_{i,j}, \sin\beta_{i,j}, \cos\beta_{i,j}, d_{i,j}]. \quad (5)$$

This formulation ensures rotational invariance and smooth continuity, enabling interaction learning independent of absolute positions. All pairwise features are aggregated into $\mathbf{R}_p \in \mathbb{R}^{N\times N\times 5}$ and fed into the attention-based fusion encoder, supporting compact and interpretable spatial reasoning across agents and road elements.

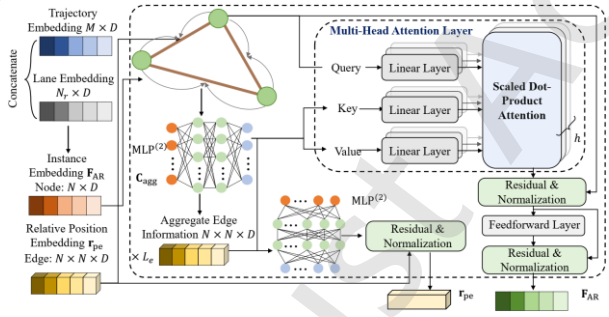


Fig. 3. Architecture of the symmetric fusion encoder. The encoder introduces explicit relational encoding across instance-centric coordinates while preserving viewpoint and ordering invariance. Instances are modeled as nodes in a fully connected self-looped graph, enabling consistent bidirectional fusion of multimodal features.

4.3. Vectorized Embedding and Symmetric Context Encoding

At the input stage, CogDrive transforms heterogeneous scene elements, including historical trajectories S_A, lane segments S_R, and relative positional features $\mathbf{r}_p$, into a unified latent space through vectorized embedding. Inspired by VectorNet and PointNet architectures, each trajectory polyline or map segment is first transformed to its corresponding instance-centric coordinate system, followed by vector-wise feature extraction using MLPs. The embedding process can be written as

$$\mathbf{F}_A = \rho\left(\mathrm{MLP}\left(\Phi(S_A)\right)\right), \mathbf{F}_R = \rho(\mathrm{MLP}(\Phi(S_R))), \mathbf{r}_{pe} = \mathrm{MLP}(\mathbf{r}_p), \quad (6)$$

where Φ denotes coordinate transformation and $\rho(\cdot)$ represents max-pooling along the temporal dimension. This design yields compact trajectory features $\mathbf{F}_A \in \mathbb{R}^{M\times D}$, lane features $\mathbf{F}_R \in \mathbb{R}^{N\times D}$, and relative positional embeddings $\mathbf{r}_{pe} \in \mathbb{R}^{N\times N\times D}$ that preserve both spatial topology and motion continuity.

The encoded features are concatenated as $\mathbf{F}_{AR} = [\mathbf{F}_A, \mathbf{F}_R]$ and passed to a symmetric fusion encoder based on a Transformer architecture, as illustrated in Fig. 4. Unlike conventional attention mechanisms, the symmetric fusion encoder explicitly maintains reciprocal relationships among different instance-centric coordinate systems. For each pair of instances (i, j), the model aggregates directional and relational features using MLPs:

$$\mathbf{C}'_{agg,i,j} = \mathrm{MLP}(\mathrm{Concat}(\mathbf{F}'_{AR,i}, \mathbf{F}'_{AR,j}, \mathbf{r}'_{pe,i,j})), \quad (7)$$

where $\mathbf{C}'_{agg}$ encodes bidirectional contextual dependencies. A multi-head attention (MHA) module then updates the feature representation:

$$\mathbf{F}'^{\,l+1}_{AR} = \mathrm{MHA}(\mathbf{F}'^{\,l}_{AR}, \mathbf{C}'_{agg}, \mathbf{C}'_{agg}), \quad (8)$$

allowing information to propagate symmetrically across agents and map segments. Residual connections and layer normalization preserve stability and gradient flow, while additional MLP layers refine relative embeddings:

$$\mathbf{r}'^{\,l+1}_{pe} = \mathrm{MLP}(\mathbf{C}'_{agg}) + \mathbf{r}'^{\,l}_{pe}. \quad (9)$$

After L_e layers of iterative encoding, the network outputs updated relational features $\mathbf{F}'_{AR}$ and $\mathbf{r}'_{pe}$ that capture bidirectional spatial dependencies and cross-coordinate consistency. This symmetric relational design ensures that all pairwise interactions are represented in an order-invariant and geometrically consistent manner, enabling CogDrive to reason over complex inter-agent dependencies. The fused relational features $\mathbf{F}'_{AR}$ are subsequently used in the decoding stage for multimodal trajectory generation and interaction-aware risk prediction.

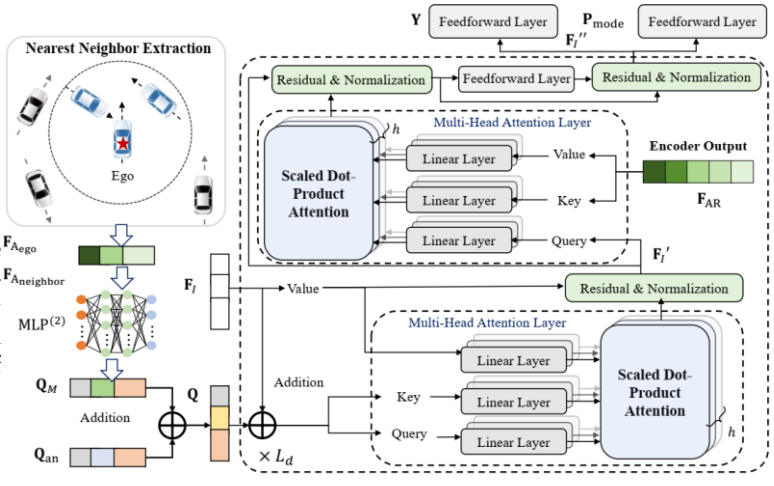


Fig. 4. Decoder for interaction-aware representation learning. The decoder transforms relational features into multimodal trajectory hypotheses via iterative self- and cross-attention. Learnable queries encode distinct behavior modes and are refined into trajectories with associated probabilities.

4.4. Learnable Decoding

To achieve multimodal and interpretable trajectory prediction, CogDrive adopts a learnable query-based decoding mechanism inspired by the DETR architecture. As illustrated in Fig. 4, the decoder generates a set of learnable queries in latent space and associates them with behavioral mode embeddings that represent distinct interaction intentions, such as yielding, merging, or overtaking. Each query interacts with encoded contextual features through multi-head and cross-attention layers, gradually refining its spatial hypothesis into a trajectory mode with corresponding probability.

The query embedding $\mathbf{Q}$ is composed of two components: a learnable anchor query $\mathbf{Q}_{an}$ and a modality-guided query $\mathbf{Q}_M$. The modality-guided component is generated from the encoded features of the ego vehicle and its selected neighboring agents:

$$\mathbf{Q}_M = \mathrm{MLP}\big([\mathbf{F}'_{ego}, \mathbf{F}'_{neighbor}]\big), \tag{10}$$

where $\mathbf{F}'_{ego}$ and $\mathbf{F}'_{neighbor}$ denote the updated latent representations from the symmetric fusion encoder. To limit computational complexity, only the $n_{neighbor}$ most relevant agents are considered based on proximity or potential collision risk. The total query embedding is defined as

$$\mathbf{Q} = \mathbf{Q}_{an} + \mathbf{Q}_M, \tag{11}$$

where $\mathbf{Q} \in \mathbb{R}^{k_M \times D}$ contains k_M learnable queries, each representing a potential interaction mode.

Within each decoding layer, self-attention ensures the diversity of query embeddings, preventing mode collapse, while cross-attention enables each query to interact with the encoded feature map $\mathbf{F}_I$. The computation process follows

$$\mathbf{F}'_I = \mathrm{MHA}(\mathbf{F}_I + \mathbf{Q}, \mathbf{F}_I + \mathbf{Q}, \mathbf{F}_I), \tag{12}$$

where MHA denotes the standard multi-head attention operation. Subsequently, cross-attention aggregates contextual information between the query features and the encoder output $\mathbf{F}'_{AR}$:

$$\mathbf{F}''_I = \mathrm{MHA}(\mathbf{F}'_I + \mathbf{Q}, \mathbf{F}'_{AR}, \mathbf{F}'_{AR}). \tag{13}$$

After L_d decoding layers, the final output $\mathbf{F}''_I$ represents multimodal latent trajectories, which are projected through two independent fully connected layers to obtain the predicted trajectory coordinates X and the mode probability vector

$$\mathbf{P}_{mode} = (p_1, p_2, \ldots, p_K). \tag{14}$$

This learnable decoding mechanism allows CogDrive to autonomously discover and refine distinct interaction patterns through iterative attention updates, bridging the gap between cognition-driven behavior understanding and accurate multimodal trajectory forecasting. It further ensures that rare yet safety-critical interaction modes remain represented, supporting robust downstream planning and decision-making.

4.5. Network Training

CogDrive is trained under a Winner-Takes-All (WTA) strategy that jointly optimizes trajectory regression, classification, and modality consistency. The overall training objective combines three complementary losses:

$$\mathcal{L} = \mathcal{L}_{reg} + \alpha_1 \mathcal{L}_{cls} + \alpha_2 \mathcal{L}_{mode}, \tag{15}$$

where α_1 and α_2 are balancing coefficients set to 0.2 and 0.01, respectively. The WTA selects the predicted mode k^* with the minimum final displacement error and updates the corresponding mode probability and trajectory through

$$\mathcal{L}_{cls} = \sum_{k=1}^{K} \max(0, \epsilon_{margin} + p_k - p_{k^*}), \tag{16}$$

where p_k is the probability of mode k, p_{k^*} corresponds to the optimal mode, and ϵ_{margin} is a predefined margin (set to 0.2 in experiments). This margin-based constraint prevents mode collapse by maintaining sufficient separation between mode probabilities, encouraging the network to learn diverse interaction hypotheses.

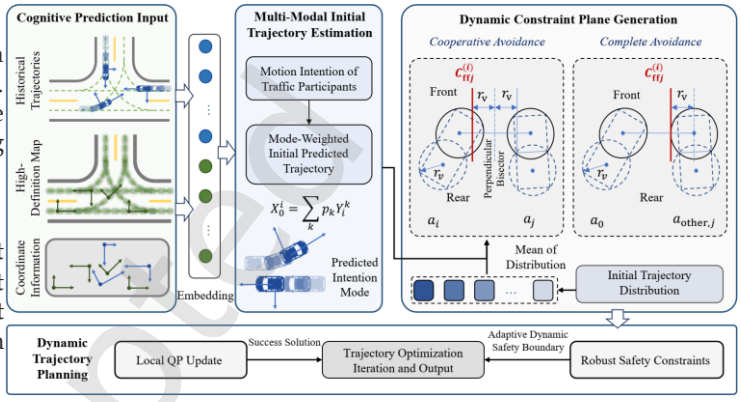


Fig. 5. Dynamic safety-aware trajectory planning in CogDrive. Mode-weighted predictions define dynamic constraints and an adaptive safety boundary. A local QP planner generates collision-free trajectories that realize cooperative yielding or avoidance behaviors.

The regression loss refines trajectory accuracy and motion smoothness through position and yaw-angle supervision:

$$\mathcal{L}_{reg} = \mathcal{L}_{pos}(\bar{Y}_{pos}, Y^*_{pos}) + \mathcal{L}_{yaw}(\bar{Y}_{yaw}, Y^*_{yaw}), \tag{17}$$

where $\bar{Y}$ denotes predicted trajectories of the best mode k^*, and Y^* represents ground-truth references. The positional loss $\mathcal{L}_{pos}$ is computed as the mean squared error of predicted coordinates, while the yaw loss captures angular consistency between predicted and true heading directions:

$$\mathcal{L}_{yaw}(\bar{Y}_{yaw}, Y^*_{yaw}) = \frac{1 - \phi_{CosSim}(\bar{Y}_{yaw}, Y^*_{yaw})}{2}, \tag{18}$$

where ϕ_{CosSim} measures cosine similarity between yaw vectors. This design penalizes orientation discontinuities and promotes physically plausible trajectories, especially under low-speed or near-collision scenarios where continuous steering control is critical.

For multimodal regularization, CogDrive introduces a differentiable surrogate to approximate discrete mode matching, enabling end-to-end optimization of mode consistency. The resulting $\mathcal{L}_{mode}$ encourages alignment between predicted and reference modes while preserving smooth gradients, improving multimodal coverage and inter-mode calibration. Combined with position and yaw objectives, this design balances accuracy and diversity, yielding precise yet well-separated trajectory predictions that reflect interpretable and interaction-aware driving behaviors.

5 Multimodal Safety-Aware Trajectory Planning

The multimodal trajectory prediction module outputs K probabilistic trajectory hypotheses $\mathbf{Y}$ representing distinct behavioral intentions of surrounding agents. Building on these predictions, CogDrive performs safety-aware emergency trajectory tree planning to ensure short-term collision avoidance and long-term stability under multimodal uncertainty. The framework integrates multimodal preparedness planning with single-vehicle trajectory optimization in a unified hierarchical process, as shown in Fig. 5. The planner (i) constructs a mode-weighted nominal trajectory for initialization, (ii) generates dynamic constraint planes between ego and neighbors, and (iii) maintains an adaptive safety boundary updated via local quadratic programming iterations to realize cooperative or full avoidance consistent with multimodal intents.

5.1. Multimodal Preparedness Planning

When interacting with uncertain agents, the ego vehicle must ensure safety across all modes. CogDrive therefore adopts a preparedness-based planning strategy that accounts for the distribution of predicted behaviors. Instead of following only the most probable mode, the planner generates a short-horizon emergency trajectory to ensure safety across modes and a long-horizon nominal trajectory to maintain continuity after mode resolution. This design mitigates overreaction and preserves decision consistency. As shown in Fig. 5, the emergency branch covers all predicted modes, while the nominal branch tracks the most consistent mode as the belief evolves.

The multimodal emergency planning problem is formulated as a constrained nonlinear optimization:

$$\min_{\mathbf{U}} \sum_{t=0}^{T_b} l_z(\mathbf{z}_t) + \sum_{t=1}^{T_b} l_u(\mathbf{u}_t) + \sum_{k\in K} \sum_{t=T_b}^{T} (l_z(\mathbf{z}_t^k) + l_u(\mathbf{u}_t^k)), \quad (19)$$

subject to

$$\begin{aligned} \mathbf{z}_t &= f(\mathbf{z}_{t-1}, \mathbf{u}_t), \forall t \in [1, T_b], \\ \mathbf{z}_t^k &= f(\mathbf{z}_{t-1}^k, \mathbf{u}_t^k), \forall t \in (T_b, T], \forall k \in [1, K], \\ h_i(\mathbf{z}_t) &\le 0, \qquad \forall t \in [0, T_b], \\ h_i^k(\mathbf{z}_t^k) &\le 0, \qquad \forall t \in (T_b, T], \forall k \in [1, K], \end{aligned}$$

where $\mathbf{z}_t \in \mathbb{R}^{n_z}$ and $\mathbf{u}_t \in \mathbb{R}^{n_u}$ denote the ego state and control input at time step t, $f(\cdot)$ represents discrete-time vehicle dynamics, and $h_i(\cdot)$ defines the safety and comfort constraints. The cost functions $l_z(\cdot)$ and $l_u(\cdot)$ penalize deviations from desired states and excessive control efforts. The first phase $[0, T_b]$ ensures immediate safety across all possible interactions, while the second phase $(T_b, T]$ maintains long-term stability and smoothness.

5.2. Single-Vehicle Trajectory Planning

Building on multimodal preparedness, CogDrive formulates a geometric and optimization-based single-vehicle planner that bridges cognition-driven prediction and low-level control. The planner enforces dynamic feasibility, spatial safety, and temporal smoothness within a unified framework.

Initialization from multimodal predictions. Given K predicted trajectory hypotheses $\{Y_i^k\}_{k=1}^K$ with mode probabilities p_k, the ego vehicle forms a weighted nominal trajectory

$$\bar{\mathbf{X}}_0 = \sum_{k=1}^K p_k \, Y_{\text{ego}}^k, \quad (20)$$

and generates candidate initializations $\{\bar{\mathbf{X}}_0^m\}_{m=1}^M$ via small perturbations. These mode-consistent guesses warm-start the optimizer and align planning with prediction (Fig.5).

Dynamic geometric constraints. Interactions with uncontrollable neighbors are modeled through time-varying geometric constraints. Vehicles are approximated by front and rear circular envelopes with radii r_F and r_R. For neighbor j relative to ego i, a separating hyperplane is defined using $\mathbf{d}_{ij} = \mathbf{p}_j - \mathbf{p}_i$:

$$\mathbf{A}_{c,ij}\mathbf{Y}_{c,i} \le \mathbf{b}_{c,ij}, \quad \mathbf{A}_{c,ij} = \frac{\mathbf{d}_{ij}^\top}{\|\mathbf{d}_{ij}\|}, \quad \mathbf{b}_{c,ij} = \| \mathbf{d}_{ij} \| - (r_F + r_R), \quad (21)$$

where $\mathbf{Y}_{c,i} = [x_{F,i}, y_{F,i}, x_{R,i}, y_{R,i}]^\top$. These constraints define a safe region that adapts over time, yielding cooperative or evasive behaviors under different modes.

Static and robust corridor constraints. To handle static obstacles and prediction uncertainty, a convex safety corridor bounds the ego trajectory:

$$\mathbf{Y}_{c,\min} \le \mathbf{Y}_{c,t+1} \le \mathbf{Y}_{c,\max}, \quad (22)$$

where bounds are adaptively expanded using the predicted uncertainty Σ_{ego}^k, ensuring robustness to bounded errors (Fig. 5).

Quadratic programming formulation. The planning problem is cast as a constrained QP:

$$\min_{\mathbf{X},\mathbf{U}} \sum_{t=0}^{T} \| \mathbf{X}_t - \bar{\mathbf{X}}_t \|_Q^2 + \| \mathbf{U}_t - \bar{\mathbf{U}}_t \|_R^2, \quad (23)$$

subject to

$$\begin{aligned} \mathbf{X}_{t+1} &= f_d(\mathbf{X}_t, \mathbf{U}_t), \quad \forall t \in [0, T], \\ \mathbf{A}_{c,ij}\mathbf{Y}_{c,i,t} &\le \mathbf{b}_{c,ij}, \quad \forall j \in \mathcal{N}(i), \\ \mathbf{Y}_{c,\min} &\le \mathbf{Y}_{c,t} \le \mathbf{Y}_{c,\max}, \forall t \in [0, T], \end{aligned}$$

where $\mathbf{X}_t = [x_t, y_t, v_t, \psi_t]^\top$, $\mathbf{U}_t = [a_t, \delta_t]^\top$, $f_d(\cdot)$ is the discrete kinematic model, and Q, R are positive-definite weights. Local QP updates iteratively refine a feasible solution.

Execution and replanning. Only the first segment is executed, while the remainder provides a predictive reference. At each replanning cycle, new predictions are incorporated and the solver is warm-started with the previous solution, enabling real-time adaptation and stable closed-loop behavior under multimodal uncertainty. By unifying probabilistic prediction, geometric safety reasoning, and optimization-based refinement, CogDrive generates trajectories that are dynamically feasible, spatially safe, and cognitively consistent with multimodal interactions. In highly uncertain scenarios such as unprotected turns or roundabout merges, the prediction module may alternate between modes, for example yielding or passing. Rather than switching discretely and inducing control oscillation, the Trajectory Tree Planner searches for a common safe subspace. By constraining the short-term root trajectory to remain valid across probable semantic branches, the planner produces a smooth and defensive motion that tolerates mode shifts without abrupt corrections.

The multimodal trajectory prediction module outputs K probabilistic trajectory hypotheses $\mathbf{Y}$ representing distinct behavioral intentions of surrounding agents. Based on these, CogDrive performs safety-aware emergency trajectory tree planning to ensure short-term collision avoidance and long-term stability under multimodal uncertainty. As shown in Fig. 5, the framework unifies preparedness planning with single-vehicle trajectory optimization in a hierarchical process. The planner (i) initializes a mode-weighted nominal trajectory, (ii) constructs dynamic constraint planes between the ego vehicle and neighbors, and (iii) updates an adaptive safety boundary via local quadratic programming to achieve cooperative or full avoidance consistent with multimodal intents.

6 Experiments and Comparative Analysis

6.1. Experimental Setup

Datasets. The proposed CogDrive framework is evaluated on two large-scale real-world benchmarks, INTERACTION and Argoverse 2 , both of which provide detailed trajectories and high-definition (HD) maps for fine-grained motion prediction and multi-agent reasoning. The INTERACTION dataset contains naturalistic multi-agent driving data collected in China, Germany, and the United States using drone and roadside sensors. It covers diverse complex scenarios, including highway ramps, urban intersections, and roundabouts. Each scenario provides centimeter-level lanelet2 maps encoding lane topology, traffic rules, and connectivity. The dataset includes 18 representative scenarios with vehicle, cyclist, and pedestrian interactions. Each sample consists of 1 s of observed trajectories and a 3 s prediction horizon, yielding approximately 40,000 annotated motion sequences. Argoverse 2 consists of over 250,000 vehicle-centric trajectories collected across multiple U.S. cities at 10 Hz with accurate localization and map alignment. Compared to its predecessor, it features longer sequences (5 s observation, 6 s prediction) and richer multimodal behaviors. Its HD maps provide detailed lane geometry, drivable areas, and intersection semantics, supporting evaluation of long-horizon prediction accuracy and multimodal consistency. Both datasets are sampled at 10 Hz, with INTERACTION requiring 3 s future prediction from 1 s history, and Argoverse 2 requiring 6 s prediction from 5 s history.

Evaluation Metrics. Following standard motion forecasting benchmarks, we adopt both single-agent and joint multi-agent metrics. For Argoverse 2, four metrics are used: minimum Average Displacement Error (minADE), minimum Final Displacement Error (minFDE), Miss Rate (MR), and Brier-minFDE (b-minFDE). For INTERACTION, we report minimum joint metrics, including minimum joint Average Displacement Error (minJointADE) and minimum joint Final Displacement Error (minJointFDE). Specifically, minADE computes the average ℓ_2 distance between predicted and ground-truth positions, while minFDE focuses on the terminal displacement error. MR measures the proportion of predictions with a final error exceeding 2.0 m, and b-minFDE integrates confidence weighting to reflect both accuracy and reliability. The number of predicted modes is fixed at $K = 6$.

6.2. Results and Comparative Analysis

Tables 1 and 2 report quantitative results on the INTERACTION and Argoverse 2 datasets, respectively. CogDrive achieves competitive or superior performance across all key metrics compared with state-of-the-art methods. Owing to the distinct evaluation protocols, results are analyzed separately: Table 1 presents joint metrics for the multi-agent INTERACTION benchmark, while Table 2 reports standard single-agent metrics for Argoverse 2.

Results on INTERACTION. As summarized in Table 1, CogDrive achieves the best minimum joint Final Displacement Error (minJointFDE=0.914 m) and a strong minimum joint Average Displacement Error (minJointADE=0.301 m), surpassing most baselines such as FJMP and HDGT. Compared with learning-based models like Trai-MAE and HDGT, CogDrive demonstrates higher consistency in dense multi-agent scenes, indicating its stronger capability to capture complex social interactions and avoid long-tail mispredictions.

The improvement originates from its cognition-driven multimodal reasoning, which explicitly models behavioral intentions and adapts planning responses to diverse interaction patterns. Overall, CogDrive maintains accurate spatial alignment and interpretable trajectory diversity across heterogeneous and highly interactive driving conditions. Specifically, CogDrive achieves a minJointFDE of 0.914 m, outperforming the strong baseline FJMP (0.922 m) by approximately 0.9% and HDGT (0.958 m) by 4.6%. Although the improvement in minJointADE (0.301 m) is marginal compared to the FDE gain, this aligns with our cognition-driven objective: maximizing the accuracy of long-term intent identification (reflected by FDE) is more critical for safety-aware planning than minimizing average geometric path deviations (ADE).

To clarify this connection, we emphasize that long-tail prediction failures in multi-agent interactions often manifest as amplified deviations at the trajectory horizon endpoint, particularly when rare behaviors such as late yielding or aggressive merging are misinterpreted. Because minJointFDE is highly sensitive to such endpoint errors, CogDrive's lower minJointFDE already reflects improved robustness to these long-tail cases without requiring additional tail-specific statistics. This improvement stems from CogDrive's cognition-driven multimodal reasoning, which preserves semantically distinct interaction modes and prevents rare behaviors from collapsing into majority patterns.

Table 1. Trajectory prediction results on the INTERACTION dataset.

Method	minJointFDE (m)↓	minJointADE (m)↓
AutoBot	1.015	0.312
THOMAS	0.968	0.416
Trai-MAE	0.966	0.307
HDGT	0.958	0.303
FJMP	0.922	**0.275**
CogDrive (ours)	**0.914** (-0.87% ↓)	0.301 (+9.45% ↑)

Note: Teal values denote improvements over the best baseline, and magenta values denote trade-offs.

Results on Argoverse 2. Table 2 presents quantitative results on the Argoverse 2 dataset. CogDrive achieves the lowest Miss Rate (MR = 0.120), indicating superior safety consistency, and attains the best displacement accuracy (b-minFDE = 1.833 m, minFDE = 1.209 m) with a competitive minADE of 0.803 m. Compared with DenseTNT and SceneTransformer, it consistently reduces long-horizon errors, demonstrating improved temporal stability under multimodal uncertainty. To further validate the contribution of the Cognitive Modality Reasoning module, we analyze the performance gap between CogDrive and standard Transformer-based baselines such as SceneTransformer. The key distinction lies in the use of modality-guided queries derived from topological semantics, which mitigates the mode collapse commonly induced by randomly initialized queries. This architectural advantage is quantitatively

reflected as follows. (1) Lower Miss Rate: Explicit modeling of interaction modes (yielding, neutral, and passing) preserves rare but safety-critical behaviors as distinct hypotheses, reducing long-tail miss errors (0.120 vs. 0.126). (2) Improved minFDE: By anchoring predictions to well-defined semantic intentions, the cognitive module alleviates endpoint deviation caused by intent ambiguity (1.209 m vs. 1.232 m).

It is worth noting that the symmetric fusion encoder and the modality-guided decoder in CogDrive are structurally coupled. The encoder processes instance-centric coordinates and relational features, while the decoder leverages these representations to construct topological queries. As a result, the cognitive components operate synergistically rather than additively, and standard baselines serve as comparative references for architectures lacking this integrated cognition-driven design. Overall, these improvements stem from unified prediction–planning coupling, where multimodal intent reasoning enables anticipatory interaction modeling and adaptive trajectory refinement, leading to smooth, reliable, and human-like motion behaviors in complex urban environments.

Table 2. Trajectory prediction results on Argoverse 2 dataset.

Method	b-minFDE (m)↓	minFDE (m)↓	MR↓	minADE (m)↓
LaneGCN	2.054	1.362	0.162	0.870
mmTransformer	2.033	1.338	0.154	0.844
DenseTNT	1.976	1.282	0.126	0.882
TPCN	1.929	1.244	0.133	0.815
SceneTransformer	1.887	1.232	0.126	**0.803**
CogDrive (ours)	**1.833** (- 2.9% ↓)	**1.209** (- 1.9% ↓)	**0.120** (- 4.8% ↓)	**0.803**

Qualitative analysis. Figure.6 illustrates representative multimodal predictions across diverse driving scenarios. Red and blue vehicles denote human-driven agents, with the red vehicle as the ego agent exhibiting multiple potential behaviors. The examples include intersections and roundabouts with yielding, merging, and overtaking interactions. Historical trajectories are shown as dots, while solid and dashed lines indicate predicted futures. CogDrive effectively differentiates behavior modes while maintaining spatial smoothness and probabilistic consistency.

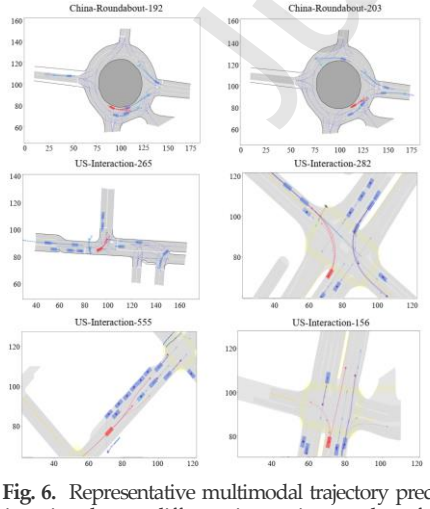


Fig. 6. Representative multimodal trajectory prediction results. Colored trajectories denote different interaction modes of the ego vehicle. Dots represent past motion, and solid or dashed lines indicate multimodal future predictions.

Unlike purely kinematic models, the cognition-driven framework captures both cooperative and competitive interactions by adapting to temporal context and suppressing implausible transitions. In contrast to conventional Transformer decoders with random or static queries that often suffer from mode collapse on imbalanced data, CogDrive adopts modality-guided queries initialized from topological interaction semantics. This design enforces explicit coverage of distinct behavioral modes, such as yielding or accelerating, rather than implicitly fitting average trajectory clusters. Together with a differentiable modality loss, rare yet safety-critical long-tail behaviors are preserved as explicit hypotheses. As a result, CogDrive generalizes robustly across heterogeneous traffic environments and driving cultures, enabling interpretable and safety-consistent multimodal reasoning.

Interactive Decision-Making Case Study. To evaluate the decision-making capability of CogDrive, we conduct closed-loop simulations using naturalistic driving data within the MIND framework , which reproduces real-world geometry and traffic flow consistent with the Argoverse map specifications. In these simulations, surrounding agents follow their recorded trajectories, while the ego vehicle executes online prediction and planning based on CogDrive. This setup enables realistic evaluation of adaptive safety control under multimodal traffic conditions. We focus on intersection scenarios where potential vehicle conflicts occur. Surrounding agents replay their ground-truth motions, and the ego vehicle continuously updates its trajectory as new observations arrive, demonstrating responsive and stable interaction behavior. The total system latency is approximately 40–50 ms, which fully satisfies the real-time requirements (10 Hz) for dynamic urban driving. Note that we do not quantitatively compare with traditional search-based planners (e.g., A* or Hybrid A*) in this section, as they typically lack the capability to explicitly model probabilistic multimodal interactions, often resulting in overly conservative 'freezing' behaviors in such dense traffic scenarios. For this reason, we compare CogDrive with data-driven prediction-and-planning methods designed for interactive scenarios rather than classical rule-based planners.

7 Conclusions

This paper presents CogDrive, a cognition-driven multimodal prediction–planning fusion framework for safe and adaptive autonomy in complex mixed traffic. By embedding cognitive reasoning into motion forecasting and decision-making, CogDrive bridges data-driven adaptability with rule-based reliability, enabling interpretable and generalizable behavior under rare or unseen conditions. The multimodal prediction module captures interaction modes through cognition-inspired representations that link behavioral semantics with topological agent relationships. Combined with learnable query decoding and differentiable modality learning, it robustly models sparse yet safety-critical behaviors. The planning module incorporates an emergency-aware mechanism and safety-stabilized trajectory tree optimization, ensuring short-term safety and long-term smoothness across multimodal outcomes. Together, prediction and planning are unified into a coherent cognitive process of reasoning, anticipation, and control. Experiments on the Argoverse 2 and INTERACTION datasets show that CogDrive improves accuracy, stability, and multimodal consistency, while closed-loop simulations demonstrate safe, interpretable, and

human-like driving behaviors across diverse interactions. Future work will extend CogDrive toward large-scale deployment and human-centered cooperative driving, advancing cognition-aligned autonomy for trustworthy and interpretable safety.

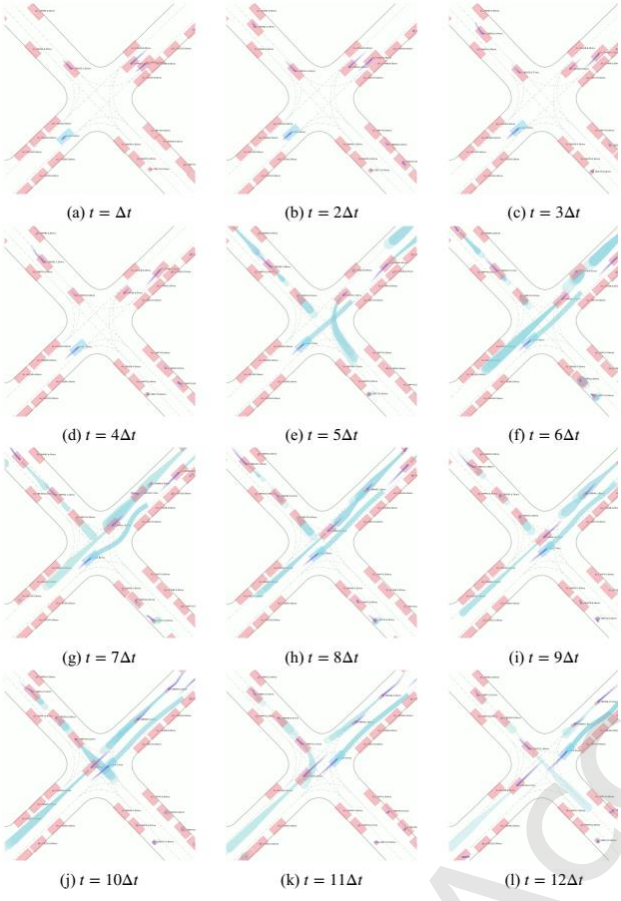


Fig. 7. Time-sequenced visualization of interactive decision-making at an urban intersection.

Author contributions

Heye Huang: Conceptualization, formal analysis, writing – original draft, visualization. Yibin Yang: Software, writing – review & validation. Mingfeng Fan: Investigation, visualization. Haoran Wang: Formal analysis, review & editing. Xiaocong Zhao: Data curation, investigation, review & editing. Jianqiang Wang: Supervision, conceptualization, review & editing.

Replication and data sharing

The source codes and replication package are available on ETS data at https://doi.org/10.26599/ETSD.2026.9190005.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Cai, Y., Lu, Z., Wang, H., Lian, Y., Chen, L., Liu, Q., et al., 2025. Mlig: Scene-level multimodal motion prediction based on multi-layer interaction graph. *IEEE Trans Intell Transp Syst*.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. *In Proc Eur Conf Comput Vis*, 213–229. Springer.

Chen, H., Wang, J., Shao, K., Liu, F., Hao, J., Guan, C., et al., 2023. Traj-MAE: Masked autoencoders for trajectory prediction. *In Proc IEEE/CVF Int Conf Comput Vis*, 8351–8362.

Chib, P.S., Singh, P., 2023. Recent advancements in end-to-end autonomous driving using deep learning: A survey. *IEEE Trans Intell Veh*, 9, 103–118.

De Luca, V., Pascarelli, C., Colucci, M., Afrune, P., Corallo, A., Avanzini, G., 2025. A platform for safe operations of unmanned aircraft systems in critical areas. Engineering, 49, 314–331.

Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., et al., 2020. VectorNet: Encoding HD maps and agent dynamics from vectorized representation. *In Proc IEEE/CVF Conf Comput Vis Pattern Recognit*, 11525–11533.

Gilles, T., Sabatini, S., Tsishkou, D., Stanciulescu, B., Moutarde, F., 2021. THOMAS: Trajectory heatmap output with learned multi-agent sampling. *arXiv preprint*, arXiv:2110.06607.

Girgis, R., Golemo, F., Codevilla, F., Weiss, M., D'Souza, J.A., Kahou, S.E., et al., 2021. Latent variable sequential set transformers for joint multi-agent motion prediction. *arXiv preprint*, arXiv:2104.00563.

Gu, J., Sun, C., Zhao, H., 2021. DenseTNT: End-to-end trajectory prediction from dense goal sets. *In Proc IEEE/CVF Int Conf Comput Vis*, 15303–15312.

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A., 2018. Social GAN: Socially acceptable trajectories with generative adversarial networks. *In Proc IEEE Conf Comput Vis Pattern Recognit*, 2255–2264.

Hagedorn, S., Hallgarten, M., Stoll, M., Condurache, A.P., 2024. The integration of prediction and planning in deep learning automated driving systems: A review. *IEEE Trans Intell Veh.*

Han, J., Liu, K., Li, W., Zhang, F., Xia, X.G., 2025. Generating inverse feature space for class imbalance in point cloud semantic segmentation. *IEEE Trans Pattern Anal Mach Intell*, in press.

Huang, H., Liu, J., Zhang, B., Zhao, S., Li, B., Wang, J., et al., 2025b. LEAD: Learning-enhanced adaptive decision-making for autonomous driving in dynamic environments. *IEEE Trans Intell Transp Syst.*

Huang, H., Cheng, H., Zhou, Z., Wang, Z., Liu, Q., Li, X., et al., 2025a. REACT: Runtime-enabled active collision-avoidance technique for autonomous driving. *arXiv preprint*, arXiv:2505.11474.

Huang, H., Liu, Y., Liu, J., Yang, Q., Wang, J., Abbink, D., et al., 2024a. General optimal trajectory planning: enabling autonomous vehicles with the principle of least action. *Engineering*, 33, 63–76.

Huang, H., Wang, J., Fei, C., Zheng, X., Yang, Y., Liu, J., et al., 2020. A probabilistic risk assessment framework considering lane-changing behavior interaction. *Sci China Inf Sci*, 63, 190203.

Jia, X., Wu, P., Chen, L., Liu, Y., Li, H., Yan, J., et al., 2023. HDGT: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encoding. *IEEE Trans Pattern Anal Mach Intell*, 45, 13860–13875.

Jiang, C., Cornman, A., Park, C., Sapp, B., Zhou, Y., Anguelov, D., et al., 2023. MotionDiffuser: Controllable multi-agent motion prediction using diffusion. *In Proc IEEE/CVF Conf Comput Vis Pattern Recognit*, 9644–9653.

Lenz, D., Kessler, T., Knoll, A., 2016. Tactical cooperative planning for autonomous highway driving using Monte-Carlo tree search. *In Proc IEEE Intell Veh Symp (IV)*, 447–453.

Li, T., Zhang, L., Liu, S., Shen, S., 2024. Multi-modal integrated prediction and decision-making with adaptive interaction modality explorations. *arXiv preprint*, arXiv:2408.13742.

Liang, M., Yang, B., Hu, R., Chen, Y., Liao, R., Feng, S., et al., 2020. Learning lane graph representations for motion forecasting. *In Proc Eur Conf Comput Vis*, 541–556. Springer.

Liu, Q., Huang, H., Zhao, S., Shi, L., Ahn, S., Li, X., et al., 2025. RiskNet: Interaction-aware risk forecasting for autonomous driving in long-tail scenarios. *arXiv preprint*, arXiv:2504.15541.

Liu, S., Fan, Y.,

Belabbas, M.-A., 2019. Affine geometric heat flow and motion planning for dynamic systems. *IFAC-PapersOnLine*, 52, 168–173.

Liu, Y., Zhang, J., Fang, L., Jiang, Q., Zhou, B., 2021. Multimodal motion prediction with stacked transformers. *In Proc IEEE/CVF Conf Comput Vis Pattern Recognit*, 7577–7586.

Meng, F., Liu, A., Jing, S., Zu, Y., 2021. FSM trajectory tracking controllers of OB-AUV in the horizontal plane. *In Proc IEEE Int Conf Intell Safety Robot (ISR)*, 204–208.

Ngiam, J., Caine, B., Vasudevan, V., Zhang, Z., Chiang, H.-T. L., Ling, J., et al., 2021. Scene transformer: A unified architecture for predicting multiple agent trajectories. *arXiv preprint*, arXiv:2106.08417.

Pek, C., Althoff, M., 2020. Fail-safe motion planning for online verification of autonomous vehicles using convex optimization. *IEEE Trans Robot*, 37, 798–814.

Prakash, A., Chitta, K., Geiger, A., 2021. Multi-modal fusion transformer for end-to-end autonomous driving. *In Proc IEEE/CVF Conf Comput Vis Pattern Recognit*, 7077–7087.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. PointNet: Deep learning on point sets for 3D classification and segmentation. *In Proc IEEE Conf Comput Vis Pattern Recognit*, 652–660.

Rowe, L., Ethier, M., Dykhne, E.-H., Czarnecki, K., 2023. FJMP: Factorized joint multi-agent motion prediction over learned directed acyclic interaction graphs. *In Proc IEEE/CVF Conf Comput Vis Pattern Recognit*, 13745–13755.

Shi, S., Jiang, L., Dai, D., Schiele, B., 2022. Motion transformer with global intention localization and local movement refinement. *Adv Neural Inf Process Syst*, 35, 6531–6543.

Sun, J., Yuan, C., Sun, S., Wang, S., Han, Y., Ma, S., et al., 2024. ControlMTR: Control-guided motion transformer with scene-compliant intention points for feasible motion prediction. *In Proc IEEE Intell Transp Syst Conf (ITSC)*, 1507–1514.

Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., et al., 2023. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint*, arXiv:2301.00493.

Yang, L., Yuan, M., Liu, Y., Qu, X., Hu, Z., Zhao, X., Fang, S., 2026. Optimization of task scheduling and resource allocation for autonomous vehicle tests in vehicle-road-cloud collaborative systems. Expert Syst Appl, 129943.

Yang, Y., Xu, S., Yan, X., Jiang, J., Wang, J., Huang, H., 2024. CSDO: Enhancing efficiency and success in large-scale multi-vehicle trajectory planning. *IEEE Robot Autom Lett.*

Ye, M., Cao, T., Chen, Q., 2021. TPCN: Temporal point cloud networks for motion forecasting. *In Proc IEEE/CVF Conf Comput Vis Pattern Recognit*, 11318–11327.

Zhan, W., Sun, L., Wang, D., Shi, H., Claussee, A., Naumann, M., et al., 2019. INTERACTION dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps. *arXiv preprint*, arXiv:1910.03088.

Zhang, L., Li, P., Liu, S., Shen, S., 2024. SIMPL: A simple and efficient multi-agent motion prediction baseline for autonomous driving. *IEEE Robot Autom Lett*, 9, 3767–3774.

Zhang, Z., Liniger, A., Sakaridis, C., Yu, F., Gool, L.V., 2023. Real-time motion prediction via heterogeneous polyline transformer with relative pose encoding. *Adv Neural Inf Process Syst*, 36, 57481–57499.

Zhou, Z., Ye, L., Wang, J., Wu, K., Lu, K.H., 2022. Hierarchical vector transformer for multi-agent motion prediction. *In Proc IEEE/CVF Conf Comput Vis Pattern Recognit*, 8813–8823.

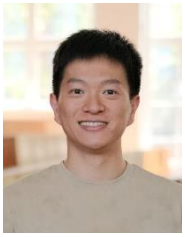
Author biography

Heye Huang received the Ph.D. degree in mechanical engineering from Tsinghua University, Beijing, China. She is currently a Postdoctoral Associate with the Singapore-MIT Alliance for Research and Technology (SMART), Massachusetts Institute of Technology. Her research interests include safe and trustworthy autonomy, generative AI, risk-sensitive decision making and LLM reasoning.

Yibin Yang received the bachelor's degree and the Ph.D. degree in mechanical engineering from Tsinghua University, Beijing, China. His research interests include multi-agent systems and autonomous driving.

Mingfeng Fan received the Ph.D. degree in traffic and transportation engineering from Central South University, Changsha, China, in 2019 and 2024, respectively. She is a Research Fellow with the Department of Mechanical Engineering, National University of Singapore. Her research interests include deep reinforcement learning, combinatorial optimization, and robotic control and scheduling.

Haoran Wang received the bachelor's degree in transportation engineering and the Ph.D. degree from Tongji University, Shanghai, China, in 2017 and 2022, respectively. He is currently an Associate Professor with the College of Transportation, Tongji University. He is a researcher on vehicle engineering, majoring in intelligent vehicle control and cooperative automation.



Xiaocong Zhao received the Ph.D. degree in transportation engineering from Tongji University, Shanghai, China. He is currently a postdoctoral research fellow with the College of Transportation, Tongji University. His research interests include human–machine interaction, social driving behavior modeling, and interactive decision-making.

Jianqiang Wang received the Ph.D. degree from Jilin University, Changchun, in 2002. He is currently a Professor with the School of Vehicle and Mobility, Tsinghua University, Beijing, China. His active research interests include intelligent vehicles, driving assistance systems, and driver behavior.